IT UNIVERSITY OF COPENHAGEN

Software Design
KISPECI1SE - Thesis

Master's Thesis

Network-based urban analysis using OpenStreetMap data

Prepared by: Zuzanna Emilia Derylo (zude@itu.dk)
Supervised by: Michele Coscia
Date: January 1, 2026

Abstract

Urban livability is a widely used but inconsistently defined concept, which is often measured using aggregated socio-economic indicators that overlook spatial structure and everyday accessibility. This thesis proposes a network-based approach to measuring livability that focuses on the physical built environment and the spatial distribution of urban amenities. Using OpenStreetMap data, cities are represented as networks in which urban blocks form nodes and their spatial adjacency forms edges, while amenities are assigned to blocks as functional attributes.

A reusable pipeline is developed to construct urban blocks, build block adjacency graphs and assign categorized Points of Interest. Livability is measured using a combination of two network-based components. The first component uses Generalized Euclidean distance to capture how accessible different amenity categories are within the urban network. The second component measures network variance, which evaluates how amenities are distributed across blocks relative to a randomized baseline. Together, these components capture both accessibility and spatial balance.

The proposed method is applied to two case studies, Copenhagen and Gdańsk, which differ in urban structure, density and planning history. Results show that Gdańsk achieves a higher livability score than Copenhagen. This outcome is a result of a more balanced spatial distribution of amenities rather than differences in overall city size or network distances. The analysis highlights the importance of amenity distribution in network-based livability.

The thesis also discusses limitations related to block definition, data quality and the absence of mobility networks, and outlines directions for future work, including integrating public transport and refining spatial units. Overall, the work demonstrates how network science and open geospatial data can be combined to study livability as a structural property of cities.

Contents

1 Introduction

Understanding how well cities support the needs of their residents is an important topic in urban research. As cities continue to face challenges related to quality of life, accessibility and sustainability, the concept of urban livability has become a goal for planners and researchers [1, 2, 3]. Livability broadly reflects how the built environment affects residents' wellbeing [4]. However, despite its importance, livability is difficult to define and measure [5].

Existing livability measures tend to rely on aggregated, high-level indicators. The most established indices, such as the Economist Intelligence Unit's Global Liveability Index or the Mercer Quality of Living Survey, rely on socio-economic, environmental and institutional factors [6, 7]. While these measures are useful for global comparisons, they typically ignore the spatial and functional relationships between amenities, the distribution of everyday services and their accessibility[8].

This thesis proposes the idea that a meaningful livability score should include the city's spatial structure and the way its streets and blocks connect. Cities can be represented as networks, where streets, blocks and amenities form connected systems that shape the way how people move and what they can access [9, 10]. From this point of view, livability depends not only on number of amenities, but also on where they are located, how close they are and how they connect to one another [11]. To study this, I use a block-level, network-base approach that represents the city through its physical layout and the types of amenities contained within each block.

The primary goal of this thesis is to develop a network-based livability score that considers both the physical layout of the city and distribution of amenities. The pipeline uses openly available data from OpenStreetMap (OSM)[12, 13]. The methodology includes creating urban blocks from the street network, constructing a graph where blocks act as nodes and their adjacency as edges, assigning categorized Points of Interest to each block, and generating amenity vectors that describe the composition of amenities in each block. Using these vectors, I calculate Generalized Euclidean (GE) distances and combine them with a network variance measure that reflects the diversity of amenities. Together, these two components form the proposed livability score. The method is applied to Copenhagen and Gdańsk to compare how the two cities perform when livability is measured through spatial accessibility and amenity diversity rather than aggregated socio-economic indicators.

The code developed for this project is provided in the `Thesis_Code.pdf` file submitted along with this thesis.

This thesis makes two main contributions. First, it provides a reusable pipeline for constructing and analyzing block-based urban networks and amenities from OSM data. Second, it proposes and evaluates a network-based livability score using OSM data and network science methods, that is GE distances and network variance measures. This method offers an alternative to traditional, non-spatial indices.

The results show that, using the proposed livability measure, Gdańsk achieves a higher livability score than Copenhagen. While this outcome may seem surprising, it is primarily determined by differences in amenity distributions and block-level accessibility, rather than overall city size or density. In particular, Gdańsk has a more balanced spatial distribution of amenities across blocks, resulting in shorter distances between services and lower network variance in amenity accessibility. These results suggest that a more even spatial distribution of amenities across the block network is the main factor increasing livability when evaluated using the proposed network-based measure

At the same time, the analysis highlights several limitations of the proposed method. The livability score captures only the physical built environment and represents potential accessibility rather than how residents experience the city in practice. Mobility-related factors such as public transport networks and travel behavior are not included, even though they shape everyday accessibility. In addition, the results also depend on the chosen block construction, which can vary across cities and affect comparability. Future work could address these limitations by integrating transport networks and exploring block definitions.

2 Literature review

2.1 Definition of livability

Livability is often described simply as "the degree to which a place is suitable or good for living in", as defined by the Cambridge Dictionary [14]. While this captures intuitive meaning of the term, academic literature shows that livability is more complex and lacks a clear definition. Many authors argue that the concept is highly context-specific and shaped by cultural, political and local priorities, which means that cities and researchers may define and measure it differently [2, 4, 5, 8].

Pacione [2] is one of the key authors who describes the absence of a single, agreed-upon definition. Pacione also argues that livability is not an attribute of a place, but rather comes from the interaction between environmental characteristics and human well-being, and that this interaction differs across individuals and contexts [2, 4]. Van Kamp et al. note that concepts such as quality of life, livability, environmental quality and sustainability are often used interchangeably, even though they refer to different ideas [4]. Their review shows that researchers apply these concepts inconsistently, which increases the ambiguity around livability. At the same time, their work points out that livability includes both objective conditions, such as access to services and environmental quality, and subjective experiences, including satisfaction, perceived safety and overall well-being.

Other authors similarly argue that livability cannot be reduced to a single, fixed definition. Ruth and Franklin describe livability as rooted in "here and now," shaped by local expectations, cultural norms and demographic differences [8]. They emphasize that what counts as "livable" varies between cities and even within the same city, depending on the needs of different groups.

Higgs et al. expand this by emphasizing social and health-related dimensions of livability [5]. They describe livable communities as "safe, attractive, socially cohesive and inclusive, and environmentally sustainable", with diverse housing options and convenient access to jobs, open spaces, education, shops, health and community services, leisure facilities, and cultural opportunities. This definition highlights how livability is closely tied to factors that influence people's health and well-being.

There is a shared theme across these definitions: livability is multidimensional. Authors such as Pacione and Van Kamp identify several dimensions,

including the physical environment, social environment, economic opportunities, access to essential services and subjective well-being. Variation in any of these factors can lead to differences in livability even within the same city [2, 3, 4].

Recent studies also show that many aspects of livability are closely connected to the built environment and residents' mobility. Liang et al.[15] show that walkability and urban attractiveness are important in creating vibrant and socially active neighborhoods. Their study shows that places with good pedestrian environments, accessible amenities and high-quality public spaces, tend to attract more people and encourage social interaction and local activity.

Overall, the literature shows that livability is not a single attribute but it comes from the variety of places, people, services, and experiences that shape everyday life. While definitions differ, a conclusion is that more accessible, walkable and amenity-rich environments tend to support higher livability.

The literature reviewed in this section highlights that livability is a complex concept shaped by many factors. In this thesis, this complexity is acknowledged, but the scope is intentionally narrowed to focus on a specific component: the physical built environment and the spatial organization of amenities. Other dimensions discussed in the literature, such as social relations, mobility patterns and subjective experience, are recognized but not modeled here and are considered directions for future work.

2.2 Livability scores

Traditional livability scores are widely used tools for comparing cities at global scale. Two of the most influential examples are The Economist Intelligence Unit's Global Liveablity Index and Mercer's Quality of Living Ranking. They aim to capture how well cities support a quality of life by using standardized sets of indicators.

The Economist Intelligence Unit (EIU) evaluates how difficult it would be for an individual to live in a given city. Its Global Liveability Index rates 140 cities around the world on over 30 qualitative and quantitative factors across five broad categories: stability, healthcare, culture and environment, education and infrastructure. The index is used mainly by businesses and international organizations, and provides information about political, social, and service-related conditions [6, 16].

Mercer's Quality of Living Ranking follows a similar structure. Mercer

evaluates cities using 39 criteria grouped into ten categories, covering political stability, socio-economic conditions, public services, housing, recreation, availability of consumer goods and environmental quality. Mercer describes its assessment as "objective, neutral, and unbiased,". The ranking is designed to evaluate the living conditions experienced by expatriates, using New York City as the baseline for comparison [7, 16].

Despite their influence, traditional livability scores face several important limitations. First, they often prioritize economic indicators such as employment or income levels. While these factors contribute to overall quality of life, research shows that non-economic and spatial aspects also play a major role in shaping everyday life and wellbeing, but are often overlooked [17]. For example, characteristics such as walkability, access to services and opportunities for social interaction have been identified as important to improve livability, but are rarely captured in global indices [18].

Livability is also partly subjective. Pacione argues that understanding livability requires paying attention to both "the city on the ground and the city in the mind," which means that people's perceptions and experiences matter as much as objective urban conditions. Traditional rankings rarely capture this subjective dimension [2, 3].

Overall, while traditional livability scores provide a useful overview for large-scale comparisons, they are limited in their ability to reflect how residents experience the city in daily life. They focus on what cities have in general terms, but do not consider how amenities are spatially distributed, how accessible they are, or how urban form and neighborhood characteristics influence livability.

In response to these limitations, this thesis uses a spatial and network-based perspective to study livability. Using OpenStreetMap data, cities are represented as networks of urban blocks enriched with amenity information. This allows the analysis to capture how amenities are distributed and how accessible they are within the urban structure. This approach shifts focus from the presence of services to their spatial organization and connectivity.

2.3 Livability in urban studies

While traditional indices evaluate livability based on socio-economic indicators, urban researchers argue that what makes the city livable lies in the spaces people use every day. They show that proximity and accessibility to essential services, such as shops, schools, green spaces, healthcare or other

daily services, shape how people navigate and experience the city. These factors influence their wellbeing and quality of life.

The importance of proximity in shaping livability is introduced in 15-minute city concept, introduced by Carlos Moreno. The idea is that people should be able to access their daily needs, whether shops, schools, healthcare, workplaces or leisure spaces, in not more that 15 minute on foot or by bicycle. Minimizing travel distances reduces time spent in traffic, encourages active mobility, and improves overall quality of life through more sustainable everyday routines. The COVID-19 pandemic increased interest in this concept as cities recognized the value of providing essential services close to home, to avoid long-distance travel. From this perspective, proximity is a key component for achieving livability, because it makes it possible for residents to access amenities and move between them efficiently [11].

Ideas about proximity and accessibility also connect to human-centered views of the city, especially work of Jane Jacobs and Jan Gehl. Jacobs argued that cities work best when neighborhoods are lively, mixed-use and easy to walk through, with short blocks and a variety of buildings. She believed that diversity - of functions, buildings and people - is what creates safety, community and street life. She saw walkability and everyday interactions between residents as core elements of urban vitality [19]. Jan Gehl approached livability from a similar angle. He observed that many cities had been shaped around cars instead of human needs. In his work, he argues that livable cities are designed for people, that is with walkable distances, comfortable streets and inviting public spaces [1].

More urban researches see the accessibility of amenities as an important component of livability. Residents depend on nearby shops, healthcare, parks, schools and public transport, and easy access to these services improves daily comfort. Studies show that having these amenities close by shapes not only practical convenience, but also supports social interaction and physical activity [20]. Lee's findings from Seoul highlight that accessibility has the greatest impact on perceived livability from all environmental factors, with residents reporting higher satisfaction when essential services and green spaces are within reach [18]. Walkable streets and local amenities therefore play important role in how positively people experience their neighborhood.

Another important aspect of livability is how amenities are distributed within the city. Neighborhoods tend to be more livable when they offer a mix of functions so that daily needs can be met in the same area. Research on

walkability shows that higher-density, mixed-use areas give residents better access to services and more chances to meet others in public spaces, which strengthens the social environment [21]. Similarly, Zhang and Yan's work shows that neighborhoods with varied and accessible amenities function as "enabling places" where the availability of amenities helps people stay active, connect with others, and feel comfortable in their surroundings. Studies also find that features like public services, transportation, safety, retail options, and especially green spaces all have impact onto residents' sense of well-being. In dense areas, green and ecological areas are particularly valuable, because they offer spaces for rest and recreation [20].

Taken together, these studies show that livability is tied to the spatial organization of the city - how close essential services are, how walkable streets feel, how dense neighborhoods are, and how public spaces support everyday life. These factors shape people's ability to meet daily needs, connect with others and stay active. Because of this, current research on livability pays greater attention to neighborhood-level differences and to how amenities are distributed, a perspective that fits well with network-based approaches.

Building on this perspective, this thesis applies a block-based network representation to analyze livability through spatial structure and amenity accessibility. Urban blocks are represented as nodes and their adjacency as edges, which allows the computation of distances between amenities. These are used to capture spatial organization of the city - how amenities are connected and distributed across the city.

2.4 OpenStreetMap

OpenStreetMap (OSM) is a free, global, collaborative geospatial database, created and maintained by a large community of volunteers [12]. They map geographic data, including roads, buildings, land use, railways, and many types of amenities. Because it is a free and open-access platform, OSM has become widely used in academic research [13].

OSM's strengths are tied to its large and active community of contributors. Urban areas are especially well represented because of the higher number of contributors. It results in detailed coverage of streets, public spaces, buildings and POIs. Some studies showed that OSM has a very high level of completeness and spatial accuracy, compared to the data from other sources [22]. Contributors continuously edit and update the map, even before official datasets are released, which makes it highly responsive to a changing

environment. Another strength is the open access of OSM data, which enabled its use to a wide range of applications, from transportation modelling to environmental analysis and urban planning [23].

At the same time, OSM has several limitations that are important in the context of spatial analysis. First, data completeness varies between regions because of differences in number of contributors. Urban areas tend to be more detailed, rural or disadvantaged neighborhoods often remain incomplete [13, 24]. Second, OSM does not offer systematic measures of data quality, such as standardized accuracy or completeness metrics. Contributions can be made without moderation at the point of entry, which introduces the possibility of mistakes and inconsistencies. Also, there is a lack of strict specifications when it comes to tagging practices, which can lead to inconsistent classification, as the contributors do not always follow the same conventions [13].

OSM is both a valuable and complex dataset. For urban studies, where detailed information on street networks, buildings and amenities is essential, it is one of the most useful open datasets available. At the same time, its limitations must be taken into account, as some preprocessing and validation steps have to be applied. In particular, issues related to data completeness and tagging consistency require additional care, which is addressed in Section 4.3 of this thesis.

2.5 Network science in urban analysis

Network science provides a useful way to understand cities by treating them as networks of connected elements. This way the city can be analyzed as graphs, where elements such as streets, intersections, building and blocks can be represented as nodes and edges. This allows researchers to examine how individual elements relate to one another and how these relationships shape patterns [9, 10].

Cities can be viewed as spatial networks, which means that their elements are embedded in geographic space and shaped by physical distance. Many forms of urban infrastructure, including roads, railways and utilities, are spatial, which makes network methods well-suited for understanding their structure [10]. Using these tools, researchers can measure properties such as centrality, clustering, reachability, which allows them to understand how central or connected different areas are, how easily they can be reached, or how movement is distributed across the network.

The idea of representing cities as networks has been used in urban studies

for decades [9]. A common representation is known as the primal graph. It models intersections as nodes and streets as edges, which allows researchers to study connectivity, traffic flow and accessibility. An alternative, the dual graph, treats streets as nodes and intersections as edges, which helps to understand pedestrian movement patterns, safety and social activity [9, 25].

Research in urban science, particularly the work of Luis Bettencourt, shows that cities function as complex systems composed of many interacting elements embedded in physical space [26]. From this perspective, what matters is not only the presence of urban components, but how these components interact across space and scale. Urban properties arise from the way elements are organized and connected, meaning that two cities with the same number of amenities may function very differently depending on how accessible those amenities are and how they are distributed across neighborhoods. In this view, urban outcomes emerge from patterns of interaction and spatial organization rather than from individual components.

Relationships between blocks, buildings, streets and public spaces shape how people move in the city and how livable they feel [27]. Because cities operate across multiple scales and contain many nonlinear interactions, network science offers tools for capturing these complexities and patterns. For this reason, network-based methods are more often used in urban research, especially for studying accessibility, mobility and urban form [9, 10, 27].

Building on this perspective, this thesis adopts a network-based approach to study livability as a structural property of the built environment. By representing urban blocks as nodes and their spatial adjacency as edges, the proposed model captures how amenities are connected across the city and how accessible they are.

2.6 Summary

The literature reviewed in this section shows that livability is a complex and multifaceted concept, influenced by social, economic, environmental and spatial factors. While many studies emphasize the importance of accessibility, walkability and amenity availability, these aspects are often addressed through aggregated indicators that do not capture how people experience the city in everyday life. At the same time, network-based urban research provides tools for analyzing spatial structure and connectivity, but is typically applied to mobility or infrastructure rather than livability itself.

This thesis addresses this gap by focusing on a specific dimension of liv-

ability: the physical built environment and its structural properties. By combining a block-based representation of urban space with network-based measures of accessibility and amenity diversity, the proposed approach offers a way to study how livability-related properties emerge from urban structure rather than from aggregated indicators alone. In doing so, the thesis brings together livability research and network science in a way that has not been systematically explored, providing a reusable method for analyzing urban livability. Other dimensions of livability are acknowledged as important but remain outside the scope of this work and are considered directions for future research.

3 Data and case studies

3.1 Data source

This study uses geospatial data from OpenStreetMap (OSM) [28], an open-source mapping platform, which is built by a global community of volunteers who update and maintain the data. It provides detailed information on physical and functional elements of cities, such as roads, paths, buildings, land use, and a wide range of Points of Interest (POIs), including shops, parks, schools and public services.

In OSM, geographical information is represented as nodes, ways and relations, where each is tagged with descriptive tags (e.g., `"highway=primary"`, `"amenity=school"`) that describe its function and category. Table 1 provides an example of such a tagged object. It only includes a selection of relevant columns, and fields with `None` values have been omitted.

...	highway	...	width	id	...	osm_type	geometry
...	residential	...	8	2796	...	way	`MULTILINESTRING ((12.54873 55.71641, 12.54879 ...))`

Table 1: Example record from the street network dataset

3.2 Data retrieval

For both Copenhagen and Gdańsk, the OSM datasets were retrieved using the Python library Pyrosm, which provides a way to read OSM data from Protocolbuffer Binary Format (`.pbf`) files into Geopandas GeoDataFrames. The raw datasets extracted for each city include:

- the administrative boundary of each city,
- the street network (all road segments),
- the railway network,
- water features,

- and a full set of Points of Interest (POIs)

3.3 Study areas

This study focuses on two cities – Copenhagen (Denmark) and Gdańsk (Poland). They differ in urban structure, density, planning history and spatial organization. Using two differing cities provides an opportunity to test how the proposed method for analyzing cities perform across different urban contexts. Importantly, both cities have high OSM data coverage, so data availability does not limit the analysis.

Copenhagen has a population of around 672,000 and covers approximately 90,9 km^2 , which results in relatively high population density [29, 30]. The city is known for its compact urban form, strong cycling culture, public transport network (such as buses, metro and trains) and well-developed pedestrian infrastructure. Many neighborhoods combine residential, commercial and recreational functions within short distances [31].

In this analysis, the municipality of Frederiksberg is included as part of Copenhagen, even though it is an independent administrative unit. Frederiksberg occupies an area of around 8,7 km^2 and has population of approximately 106,000 [30]. It is incorporated in the analysis because it is surrounded by Copenhagen and appears together in the OSM dataset.

Gdańsk has a population of about 489,000 and covers approximately 683 km^2, making it the largest city in Poland in terms of the area. However, this area includes part of the Baltic Sea; excluding the sea water, the area of Gdańsk is around 262 km^2 [32, 33]. The city has more polycentric urban structure, shaped by the coastal geography and historical development. The city contains a historic core, post-socialist housing areas, industrial zones and suburban districts. This creates more varied spatial pattern and more diverse distribution of amenities [34]. Gdańsk's transport system includes buses, tram lines and suburban trains. Compared to Copenhagen, it is more car-oriented city.

The most relevant differences between Copenhagen and Gdańsk for this study lie in density, block size and spatial distribution of amenities. Differences in public transport and cycling infrastructure also shape how residents move around the city, but this aspect is not included in this analysis. These are mentioned only to provide spatial context; their potential inclusion is discussed in Sections 7.4 and 7.5.

3.4 Raw dataset contents

The raw datasets extracted from OSM for each city contain several layers representing physical and functional components of the urban environment. At this stage, the data is unprocessed, reflecting the original structure of the OSM data.

3.4.1 Administrative boundary

The administrative boundary is used to determine the study area and ensure that only blocks within the city boundary are included in the network.

Administrative boundary is a set of `Polygons` and `MultiPolygons`, and defines the boundaries of study areas - *Københavns Kommune* for municipality of Copenhagen (including Frederiksberg), and *Gdańsk* for city of Gdańsk.

For Copenhagen, the correct administrative boundary was automatically identified. For Gdańsk, however, multiple OSM boundaries share the same name but differ by `admin_level`, so it was necessary to specify it manually. The choice was made based on the official OSM template for admin levels [35]. Following this guideline, the boundary `admin_level=8` was selected for Gdańsk.

The administrative boundary is used as a `bounding_box`, to make sure that the data is extracted only within the city area.

3.4.2 Street network

The street network is used to construct urban blocks, which form the nodes in the network representation of the city.

The raw street dataset consists of all `driving` roads extracted from OSM. Initially, it contained 20 027 road segments for Copenhagen and 32 120 road segments for Gdańsk.

Each road is represented as a `MultiLineString` geometry. Road segments are represented by various attributes that describe their characteristics. The most important among them is the `highway` tag, which categorizes each segment by its functional type (e.g., `residential`, `primary`, `service`, etc.). Not all road types contribute to the formation of urban blocks, so only a selected subset of road categories was selected for construction of blocks.

Additional columns in street network dataset include, for example, the id (`id`), number of lanes (`lanes`), speed limits (`maxspeed`), surface material

(`surface`), and access by transport mode (e.g., `bicycle`, `foot`, `sidewalk`). However, most of these attributes were not used in this project.

3.4.3 Railway network

The railway network is included as an additional boundary element to better define urban blocks in areas where rail infrastructure creates physical separation.

The railway network data contains information about rail infrastructure, such as suburban rails, mainline tracks and regional railway (depending on the city). It was extracted by filtering features tagged as `railway=rail`. Data is represented as `LineString` and `MultiLineString` geometries. Each segment includes attributes such as type of railway (`railway`), whether it is operational or under development (`construction`), and standard OSM metadata (e.g., `id`, `tags`, `geometry`).

For Copenhagen the railway dataset consisted of 615 objects. For Gdańsk there was 1 150 geometries.

3.4.4 Water features

Water features are included as unpassable obstacles that constrain block formation and influence the structure of the resulting network. The water features dataset includes information about water bodies, such as coastline, canals, rivers, lakes or ponds. For Copenhagen there was 473 water geometries, and 800 for Gdańsk.

Water features were added using a custom filter to select `natural=water` and `natural=coastline`. The geometries are a mix of `Polygon`, `MultiPolygon`, and `MultiLineString` types. The water dataset includes attributes describing, whether the water is natural object (`natural`) or what kind of water body it is (`water`). It also has standard OSM metadata (e.g., `id`, `tags`, `geometry`).

A combined overview of the extracted raw data, including the administrative boundary, road network, railway lines, and water features, is shown in Figure 1 for Copenhagen and Figure 2 for Gdańsk.

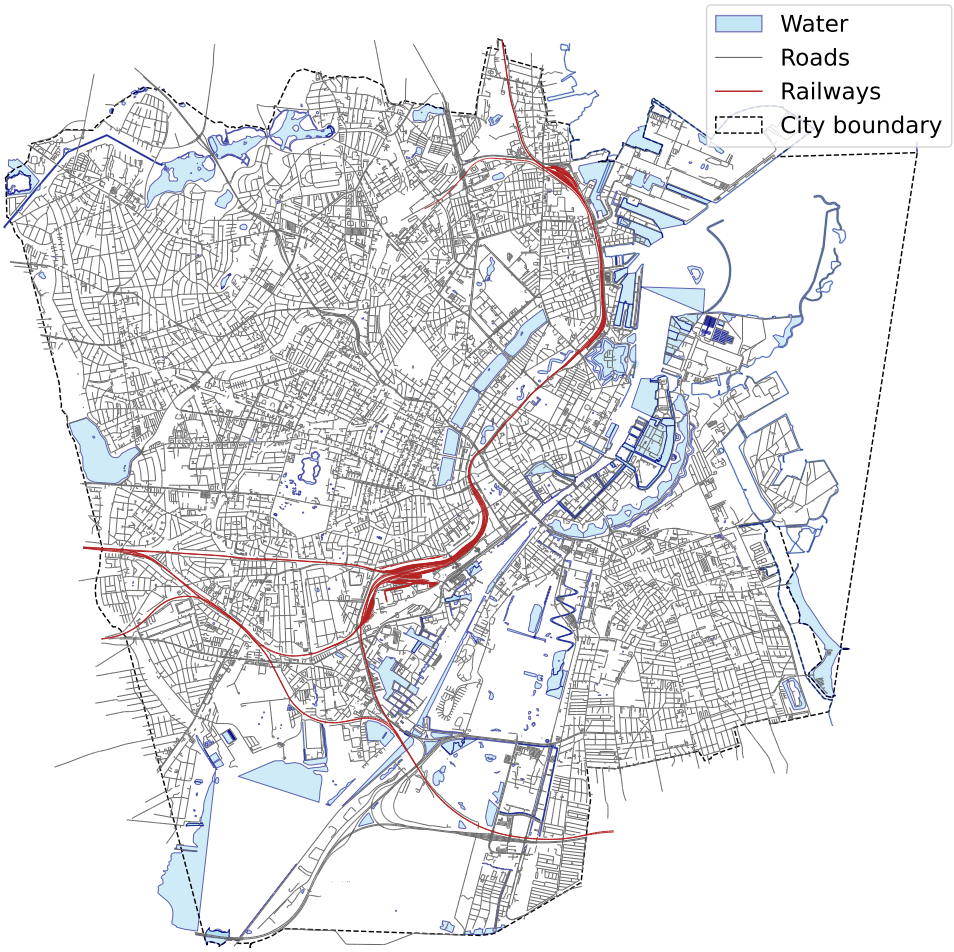


Figure 1: Overview of raw spatial data for Copenhagen.

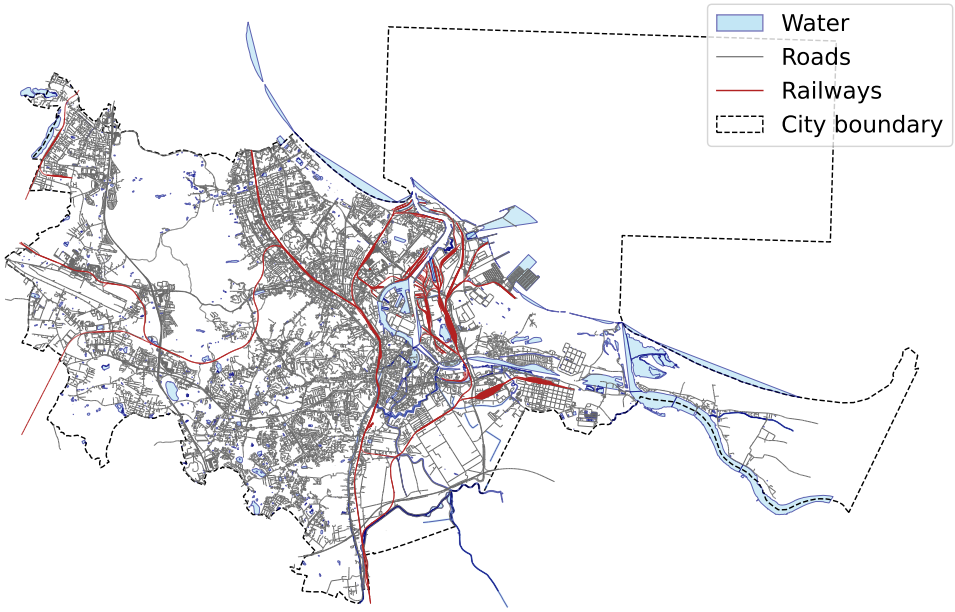


Figure 2: Overview of raw spatial data for Gdańsk.

3.4.5 Points of interest (POIs)

POIs dataset provides information about urban amenities and are lated assigned to blocks as node attributes, where they form the basis for analyzing accessibility and livability.

The Copenhagen's POIs dataset contains 30 119 entries across a wide range of OSM tag categories. The Gdańsk's POI dataset consists of around 33 140 features representing a comparable range of categories. These appear as `Point`, `Polygon` and `MuultiPolygon` geometries, depending on how they are mapped in OSM.

These POIs datasets include general tags such as `amenity`, `shop`, and `tourism`, as well as more specific tags like `parking`, `bicycle_parking`, `museum`, `bench` and `zoo`. For most features, only a few key tags contain values, whereas most available tag fields are missing or empty.

The POI dataset is rich, messy and highly tagged. The biggest problem is tagging inconsistency. For example, `parking` and `parking_space` refer to

the same type of POI but are tagged differently. To make this dataset more usable for this analysis, the POIs were cleaned (Section 4.3.5) and categorized into nine categories (Section 4.3.6)

An overviews of POI distribution across Copenhagen and Gdańsk are shown in Figure 3 and 4.

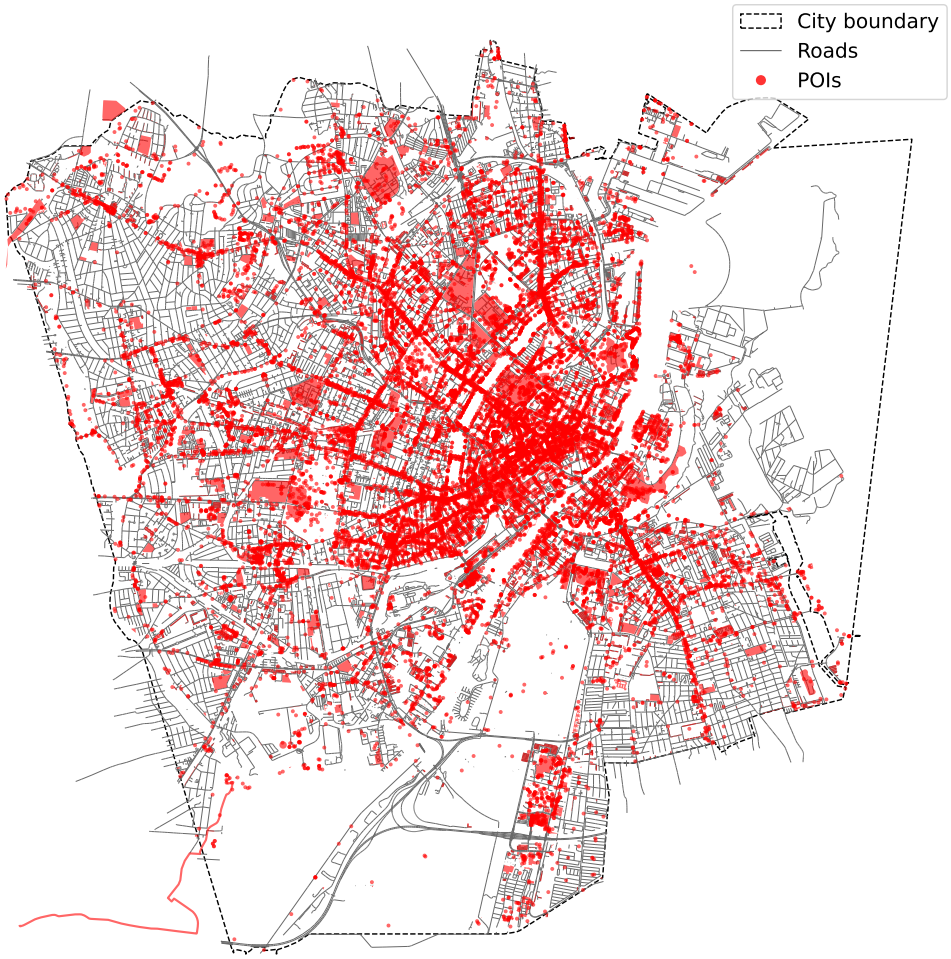


Figure 3: Distribution of initial POIs in Copenhagen.

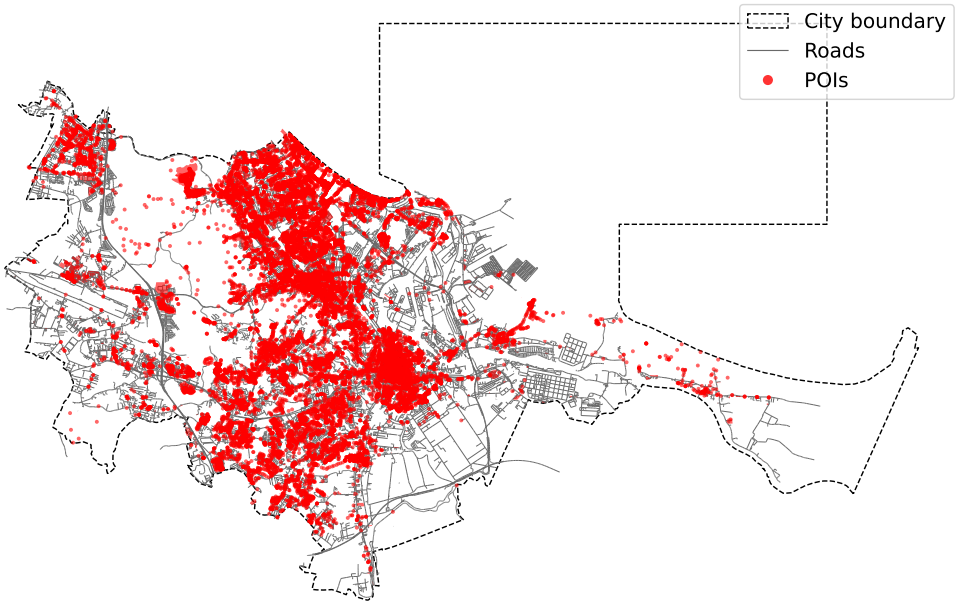


Figure 4: Distribution of initial POIs in Gdańsk.

3.4.6 Data summary

A summary of the raw dataset contents for each city is shown in Table 2.

City	Road segments	Railway segments	Water features	POIs
Copenhagen	20 027	615	473	30 119
Gdańsk	32 120	1150	800	33 140

Table 2: Summary of raw OpenStreetMap data for Copenhagen and Gdańsk. Each number represents number of geometries.

This data form the basis for the entire analysis, including construction of urban blocks, assigning POIs to each blocks, creation of block-based networks and the computation of the final livability scores.

4 Methodology

4.1 Pipeline overview

The pipeline developed in this thesis transforms raw OpenStreetMap (OSM) data into a livability score. The process begins by extracting city's administrative boundary and collecting relevant data: the street network, railways, water features and Points of Interest (POIs). All data is clipped to the city boundary to form the initial datasets for analysis.

The next step is preprocessing. POIs are cleaned and grouped into broad functional categories. Roads, railways and water geometries are combined to generate urban blocks, which are then filtered and merged to accurately represent the city's structure.

These blocks are used to build a block adjacency graph, where each block becomes a node, with edges connecting blocks that share a border. POIs are assigned to their corresponding blocks. This graph represents the spatial structure of the city and provides the base for measuring livability.

The livability score is calculated using two components, which capture different aspects of accessibility. The first applies Generalized Euclidean (GE) distances to reflect how difficult it is to travel from a block to various amenities (POI-to-home). The second is a network variance-based measure, which captures how amenities relate to each other spatially, indicating how difficult it is to move between different types of services (POI-to-POI). Together, these measures describe both how close amenities are and how they are spread in the city.

The overall pipeline is designed to be reusable for other cities. There are few elements that require manual adjustment, such as boundary selection, POI categorization or block filtering thresholds.

The full implementation of the pipeline described in this thesis is provided in a separate file (`Thesis_Code.pdf`), which contains the complete source code.

4.2 Logical structure of the model

Before describing the data preprocessing steps in detail, it is useful to clarify the logical structure of the data model constructed in this thesis. The goal is to represent the city as a spatial network that captures both its physical layout and the distribution of urban amenities.

The city is modeled as an undirected graph in which each node represents an urban block. Blocks are defined as spatial units formed by physical boundaries such as streets and other linear barriers. An edge is created between two nodes that are spatially adjacent, meaning they share a boundary.

Each node has attributes, which describe the amenities located within the block. Specifically, the number of POIs in each amenity category is counted for every block, describing what types of services are available. This information is later used to analyze accessibility, amenity distribution and livability across the network.

4.3 Data preprocessing

Data preprocessing prepares the raw OSM data for the livability analysis, specifically construction of blocks, the block graph and livability score. This involved three main stages: defining and clipping the data to the administrative boundary, cleaning and categorizing POIs, and constructing and filtering urban blocks.

4.3.1 Boundary selection and filtering

The first step was to define the administrative boundary for each city. This boundary is important because all datasets are clipped to it, as it acts as a bounding box. This ensures that only data located within the city is included in the analysis. It is important to note that the city used internally in the code is in English (e.g., *Copenhagen*, *Gdansk*), but the corresponding administrative boundary must use its original name, such as *Københavns Kommune* for Copenhagen or *Gdańsk* for Gdańsk. This distinction is crucial because OSM uses local naming conventions for administrative units.

For Copenhagen, the correct administrative boundary (*Københavns Kommune*, including Frederiksberg) was identified automatically. However, the OSM boundary contains an internal "hole" in the area of the city's canals. This likely results because water is not considered part of the administrative unit, so the boundary is drawn around it. This caused several problems: important features such as water and bridges were missing, and the block construction and filtering processes did not work as intended. To resolve this, instead of using the raw boundary polygon directly, only the outer boundary edge was extracted and used. This step ensured that the boundary had no holes or gaps. Figure 5 shows before and after.

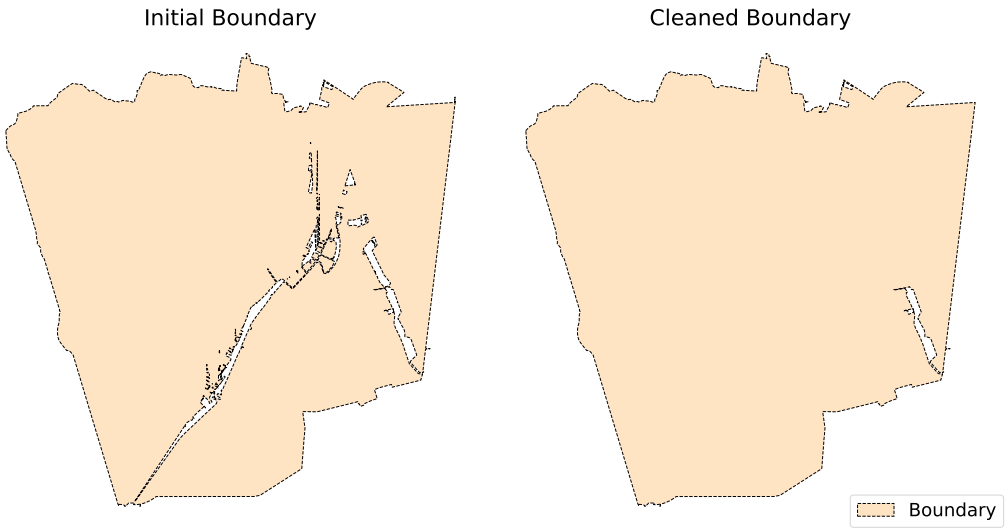


Figure 5: Copenhagen's boundary before and after preprocessing.

For Gdańsk, several administrative boundaries share the same name, but have different `admin_level`. This reflects the Polish administrative system, in which a large city can simultaneously function as a district, commune, and city, sharing the same boundary. To resolve this, the correct level was selected manually based on the OSM's official administrative template for `admin_level` [35]. According to this convention, cities in Poland correspond to `admin_level=8`, so the boundary with this level was chosen. All spatial layers were then clipped to this boundary.

4.3.2 Road network filtering

Initially, all driving roads within the bounding box were extracted from OSM, but many included types unsuitable for forming meaningful urban blocks (e.g., proposed roads, steps, bus lanes). The first filter kept the road categories that typically define block structure - `primary`, `secondary`, `tertiary`, and `residential` - as these wide, two-way streets generally act as boundaries.

However, after generating the first block layout and comparing it to the actual OSM map, it became clear that this selection was too limited. Many blocks, especially in dense or pedestrian-oriented areas, were also shaped by smaller or non-vehicular roads such as `living_street`, `service`,

`pedestrian`, and even `cycleway` or `path`. Including these categories produced a block structure that more accurately reflected cities' real urban form.

The final road types used for block construction were: `primary`, `secondary`, `tertiary`, `residential`, `motorway`, `trunk`, `unclassified`, `living_street`, `service`, `pedestrian`, `cycleway`, and `path`.

Since road geometries were stored as `MultiLineString`, I split them into individual `LineString` geometries while keeping attributes. This made the geometries easier to work with and more consistent for merging with other layers later. Also, it allowed for later block formation, which would be difficult to handle if the geometry types were different.

The number of road segments before and after preprocessing, for both Copenhagen and Gdańsk, is shown in Table 3.

	Copenhagen	**Gdańsk**
Before preprocessing	20 027 (geometries)	32 120 (geometries)
After preprocessing	77 373 (LineStrings)	138 385 (LineStrings)

Table 3: Summary of roads preprocessing for Copenhagen and Gdańsk.

4.3.3 Railway network preprocessing

Railway networks were included, because in some areas they acted as additional boundaries for blocks.

To make the railway dataset usable for block construction step, it needed to be preprocessed. As with the road data, these geometries were converted into single `LineString` objects to match the rest of the dataset.

The number of railway segments before and after this preprocessing step, for both Copenhagen and Gdańsk, is shown in Table 4.

	Copenhagen	**Gdańsk**
Before preprocessing	615 (geometries)	1 150 (geometries)
After preprocessing	4 600 (LineStrings)	14 541 (LineStrings)

Table 4: Summary of railway preprocessing for Copenhagen and Gdańsk.

4.3.4 Water features preprocessing

Water features were included for the same reason as railway network - because in some areas they acted as boundaries in the block construction.

Water geometries came in many forms - `Polygon`, `MultiPolygon`, and `MultiLineString`. Since only the edges of water bodies were needed (not the full areas), the `.boundary` method was appplied to convert them into line-based geometries. It resulted in `LineString`, `MultiPoint` and `MultiLineString` geometries. `MultiPoint` types were removed and the rest was splitted into individual `LineString` objects, in order to have a consistent dataset.

The number of water geometries before and after this preprocessing for both cities is shown in Table 5.

	Copenhagen	**Gdańsk**
Before preprocessing	473 (geometries)	799 (geometries)
After preprocessing	3 150 (LineStrings)	1 843 (LineStrings)

Table 5: Summary of water preprocessing for Copenhagen and Gdańsk.

4.3.5 POI dataset cleaning

Points of Interest (POIs) are a important component of this livability analysis because they represent the everyday functions that residents rely on. Their spatial distribution forms the basis for measuring accessibility across the block network in this thesis.

The raw OSM POI dataset, while rich in detail, is not suitable for urban analysis due to two main limitations. First, it contains many irrelevant rags, noise, and amenities that do not meaningfully contribute to livability. Second, OSM uses thousands of detailed tags, many of which are highly specific and do not correspond well to broader functional amenity types. To make the POI dataset usable for this analysis, I applied several preprocessing steps.

The initial POI dataset extracted using `load_pois` was enriched with green-area features (e.g. parks, gardens, recreation grounds) obtained from `leisure=park`, `landuse=recreation_ground`. The combined dataset was then cleaned by:

- removing entries with missing or empty geometries,

- discarding geometries in unsupported formats (that is `MultiPolygon` and `MultiLineString`); these cases represented $< 0.23\%$ of POIs (69 for Copenhagen, 19 for Gdańsk),
- converting all `Polygon` and `LineString` geometries to centroid points (`Point`) to have only a point-based representation.

This produced a dataset of POIs represented as `Point` geometries within each city boundary.

4.3.6 POI categorization

To make the dataset more meaningful for the analysis, POIs were grouped into broad functional categories representing essential urban functions. The categories were based on insights from the literature, including 15-minute city [11], Jan Gehl's "Cities for People" [1] and global livability indices [6, 7]. POIs that did not fit any functional categories or did not have any meaningful impact on the livability, such as `bench`, `parking_entrance`, `waste_basket`, were discarded.

I ended up with nine categories:

- Food – places where people can buy or eat food. It includes restaurants, cafés, bars, fast-food, bakeries, canteens and similar.
- Retail – wide variety of shops and stores, where residents can purchase goods, ranging from groceries to clothing and specialty items.
- Education – learning facilities of all types, including schools, universities, kindergartens and other educational institutions.
- Healthcare – all services providing medical care, from hospitals and clinics to doctors' offices, dental practices and pharmacies.
- Infrastructure & transport – amenities related to mobility and transport infrastructure, such as bus stops, bike parking and car parking.
- Culture & leisure – spaces that support recreation, sports, entertainment and cultural experiences. This includes museums, theaters, cinemas, libraries, sports centers, swimming pools and fitness facilities.
- Green spaces - green areas for recreation, such as parks or gardens.

- Public services – institutions that provide administrative, civic or community functions, such as post offices, police stations or government buildings.
- Other daily utilities – practical amenities that support everyday routines but do not fit into the categories above. Examples include ATMs, public toilets, recycling points, and more.

An expanded list of all categories is provided in Appendix A.

The categorization was performed in two steps, because of the tagging inconsistencies in the POIs dataset.

The first step was amenity-based categorization, which was based on the `amenity` column. I defined a dictionary (`amenity_to category`) that mapped each amenity tag (e.g., `restaurant`, `school`, `pharmacy`) to one of the functional categories. The mapping was created manually after reviewing all amenity tags, so this method would be reusable for other cities with minimal adjustments. Categories were assigned to the POIs using `.map()`, producing a new `category` column.

A large number of POIs lacked values in the `amenity` column but appeared in other columns such as `shop`, `tourism`, `office` or yes/no columns (e.g., `atm=yes`). So it was necessary to introduce the next categorization step - column-based categorization. To categorize remaining POIs, a second dictionary (`other_columns_to_category`) was created - it was created in the same way as for the amenity-based categorization, but for other columns. Rows without a category after the previous step were collected into a temporary dataframe (`pois_no_cat`). A helper function (`normalize()`) was used to clean inconsistent values (e.g. converting `None`, `NaN`, `no` or empty strings as missing). Each remaining POI was then assigned to a category with `assign_category()` method, which goes through each relevant column and assigns the first matching category.

Finally, I updated the main dataframe `pois` with new categories from `pois_no_cat`. The number of POIs before and after categorization, along with the percentage of categorization coverage, is shown in Table 6. The resulting categorized POIs serve as inputs in the next steps of the analysis, where they are assigned to urban blocks and used to compute livability measures.

Step	**Copenhagen**	**Gdańsk**
Raw POIs	30 119	33 140
Raw POIs after adding green spaces	31 103	33 391
Categorized POIs	23 314	19 019
Categorization coverage (%)	75,2%	57,0%

Table 6: Summary of POI preprocessing and categorization for Copenhagen and Gdańsk.

The POI categorization was initially defined based on the Copenhagen dataset and then systematically reviewed and expanded for Gdańsk to ensure that all relevant POI tags present in both cities were included. A similar review step would be required when applying the method to other cities.

Differences in categorization coverage between the two cities therefore reflect differences in the composition of POI types rather than missing tags. Since OpenStreetMap includes many objects unrelated to everyday accessibility, such as street furniture and technical infrastructure, coverage is not expected to be complete. This is illustrated by the most frequent POI tags in the `amenity` column listed in Appendix B, which show that a large proportion of mapped objects are not directly related to accessibility.

4.4 Construction and filtering of urban block

4.4.1 Construction of urban blocks

As described in the *Logical structure of the model* (Section 4.2), urban blocks form the fundamental spatial units in this analysis, as they are represented as nodes in the block-based network. They were constructed by combining three types of features from OSM: the street network, railways, and water features. These datasets divide the urban area into enclosed spaces (blocks).

First, roads and railways were merged into a single network. Water bodies (rivers, lakes, coastlines) were added because they also act as natural barriers and form boundaries between blocks. The combined network was then polygonised, using the `polygonize()` function, to produce a set of polygons representing potential urban blocks. Figures 6 and 7 show initial blocks for Copenhagen and Gdańsk.

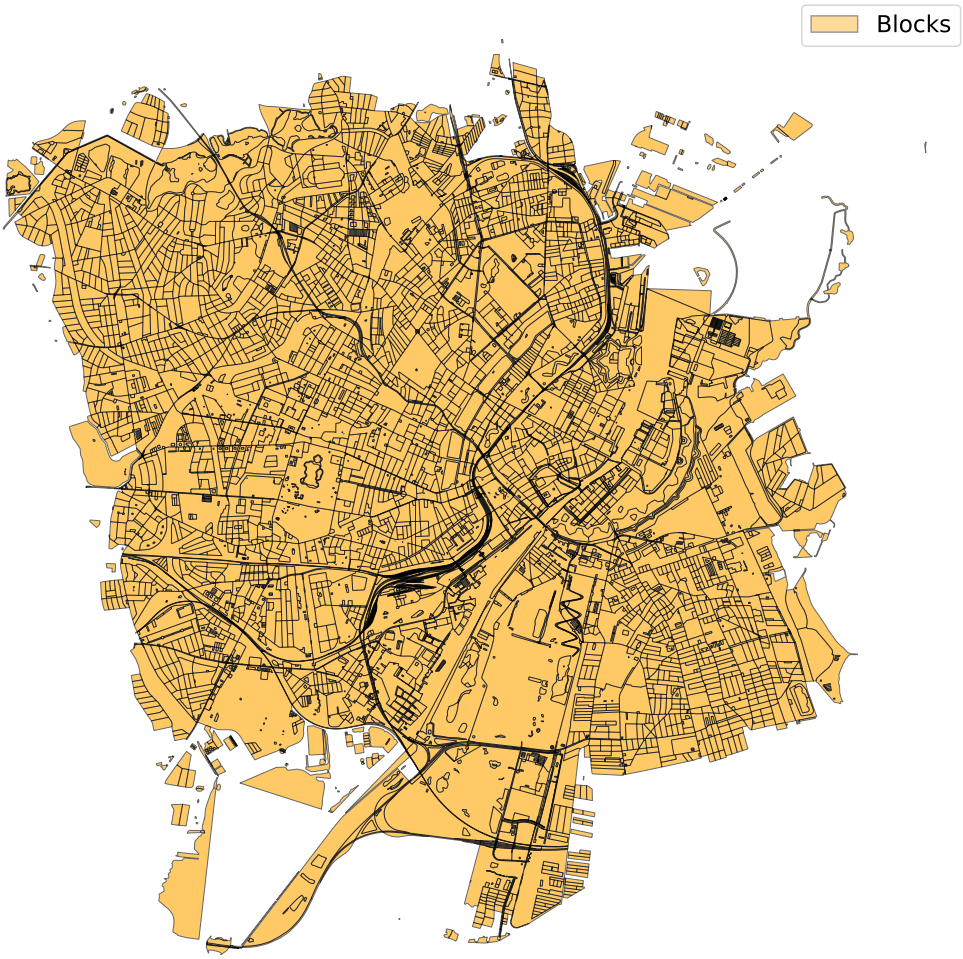


Figure 6: Initial blocks for Copenhagen.

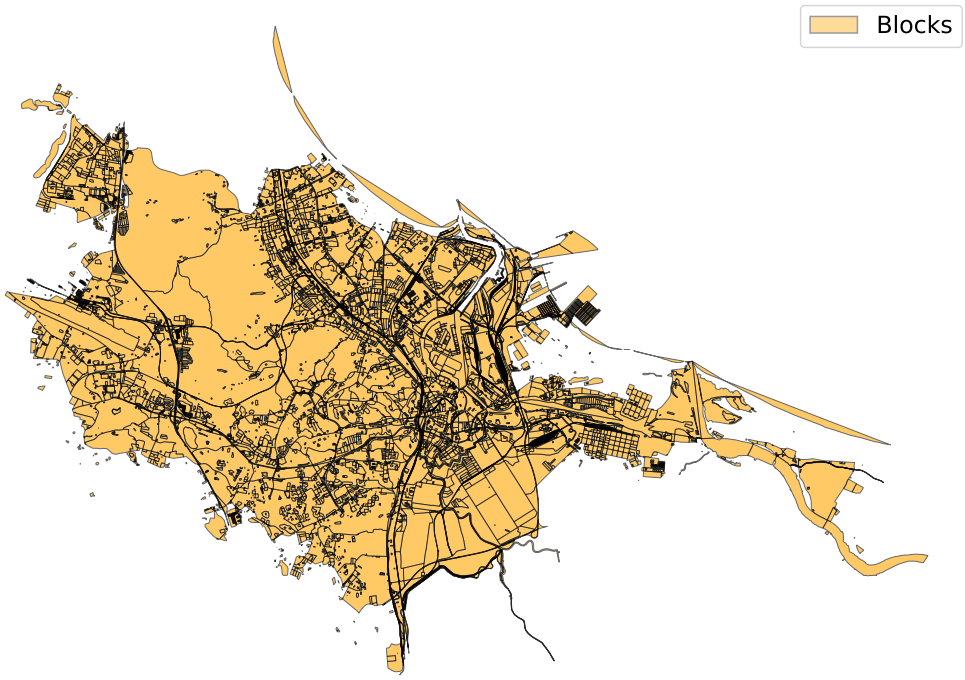


Figure 7: Initial blocks for Gdańsk.

However, the polygons created in this step contained many geometries that do not represent real-world urban blocks. Inconsistencies in OSM data and the structure of the street network generate polygons with unrealistic shapes and sizes. These blocks needed to be filtered out before analysis to avoid misleading results.

4.4.2 Filtering of urban blocks

To ensure that the final block layer represents meaningful urban units, several filtering steps were applied. The filtering consisted of four stages:

1. **Removing water blocks**
   Some polygons produced by polygonization represented water, such as lakes, rivers or coastline. While water boundaries were necessary for creating blocks, water bodies themselves should not count as blocks. To remove them, each polygon was compared with the water polygons

from OSM. Using a spatial join with a `within` predicate, blocks whose geometry fell entirely inside a water polygon were flagged as water blocks. These flagged polygons were then excluded from the dataset, leaving only non-water blocks .

2. **Filtering out small blocks**
   Because polygonization operates on detailed OSM geometries, it often produces many very small polygons, especially along road edges, railway lines and water boundaries. These small polygons do not correspond to real-world urban blocks and would create unnecessary nodes in the block graph.

   To identify small blocks, the area of each polygon was computed. The distribution of block areas was then analysed separately for Copenhagen and Gdańsk. In both cities, the distributions were highly right-skewed, with a long tail of very small blocks. Since a linear-scale histogram compressed the lower end of the distribution, the same data was visualised on a logarithmic scale to better show the structure of the tail of small blocks. Figures 8 and 9 show the resulting block area distributions in logarithmic scale for Copenhagen and Gdańsk.

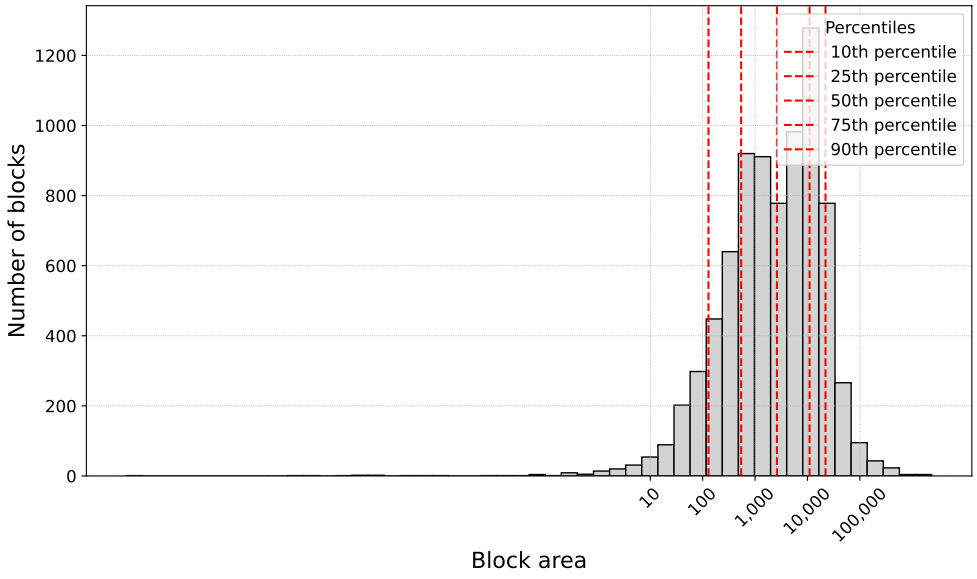


Figure 8: Distribution of block areas for Copenhagen (log scale).

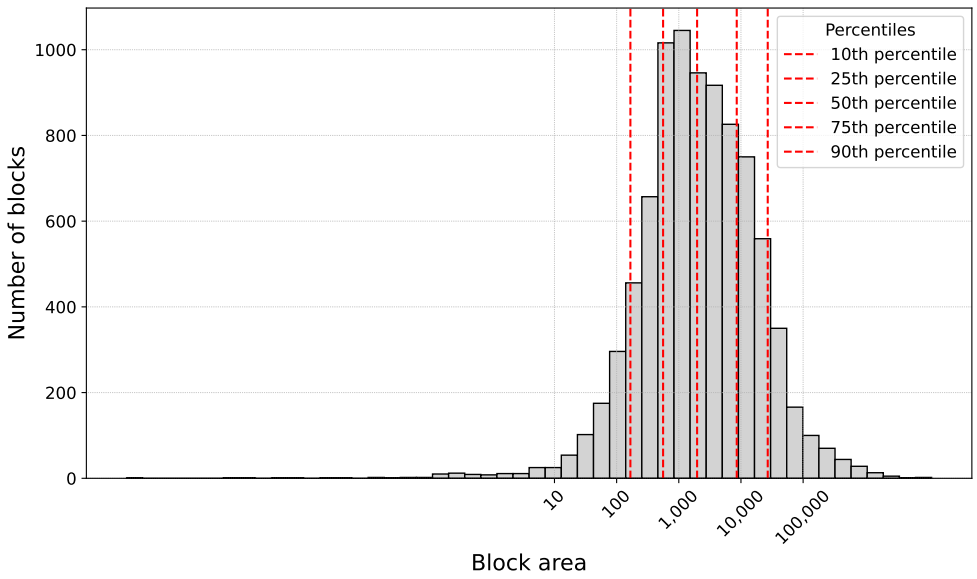


Figure 9: Distribution of block areas for Gdańsk (log scale).

Rather than defining a fixed area threshold, a percentile-based approach was used. Based on the observed distributions of block area distributions, the 25th percentile was selected as the threshold in both cities. Blocks with areas below this value were considered too small to represent meaningful urban blocks.

To better understand what types of polygons were captured by the threshold, blocks from different area percentiles were visualized over the full block layout. This analysis showed that blocks below the 25th percentile were typically fragmented or small polygons, whereas blocks above this threshold corresponded more closely to the urban blocks. Figures 10 and 11 illustrate the spatial distribution of blocks below the selected threshold for Copenhagen and Gdańsk. Additional visualisations of block area percentiles are provided in Appendix C.

All blocks below the selected threshold were iteratively merged into larger neighbouring blocks. In each iteration, the algorithm identified all blocks smaller than this percentile, computed distances between their centroids and the centroids of all valid (non-small) blocks, and merged each small block into its nearest valid neighbour. After each merging round, block geometries, areas, and IDs were recalculated. The

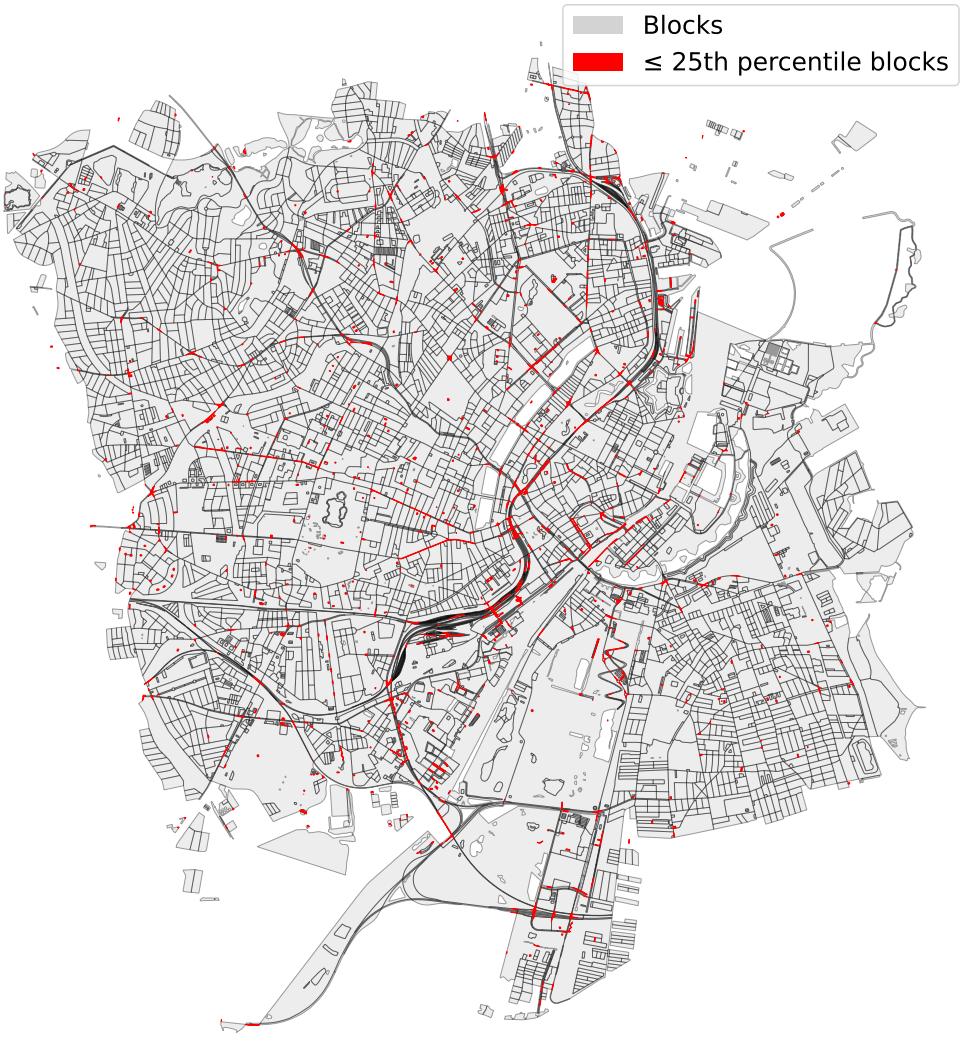


Figure 10: Spatial distribution of blocks below the selected threshold (25th percentile) for Copenhagen.

process repeated until either no blocks remained below the threshold or the maximum number of iterations was reached.

3. **Removing false water blocks**
   Due to boundary issues and incomplete water mapping, some polygons

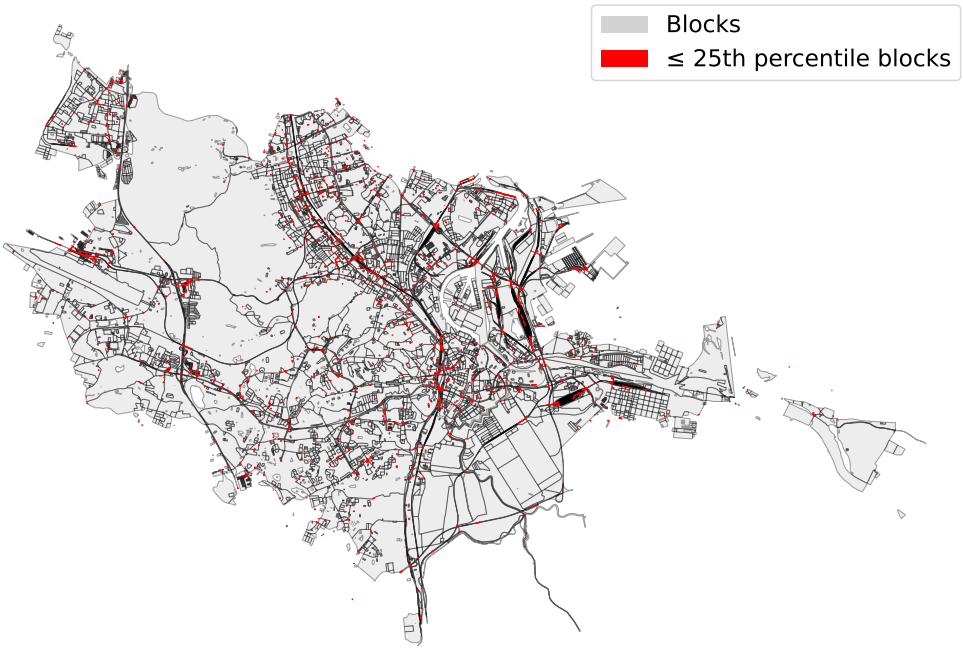


Figure 11: Spatial distribution of blocks below the selected threshold (25th percentile) for Gdańsk.

representing water were not captured as water blocks, especially in Copenhagen's canals area. This was a problem, because they were creating big blocks that were distorting the analysis.
These "false-water" polygons tended to be both very large and highly irregular. Therefore, to deal with them, blocks were evaluated using two combined criteria:

- area above a high quantile threshold (e.g., top 1%),
- compactness (irregular shape) below a low quantile threshold (based on the Polsby-Popper score [36])

Blocks that met both criteria were removed. Threshold values were selected manually after testing several combinations and identifying those that captured the most problematic blocks, which were Copenhagen's canals. This step may require manual adjustment for other cities with different boundary or water geometries. Figures 12 and 13 illustrate

the blocks identified as false-water blocks in Copenhagen and Gdańsk, respectively.

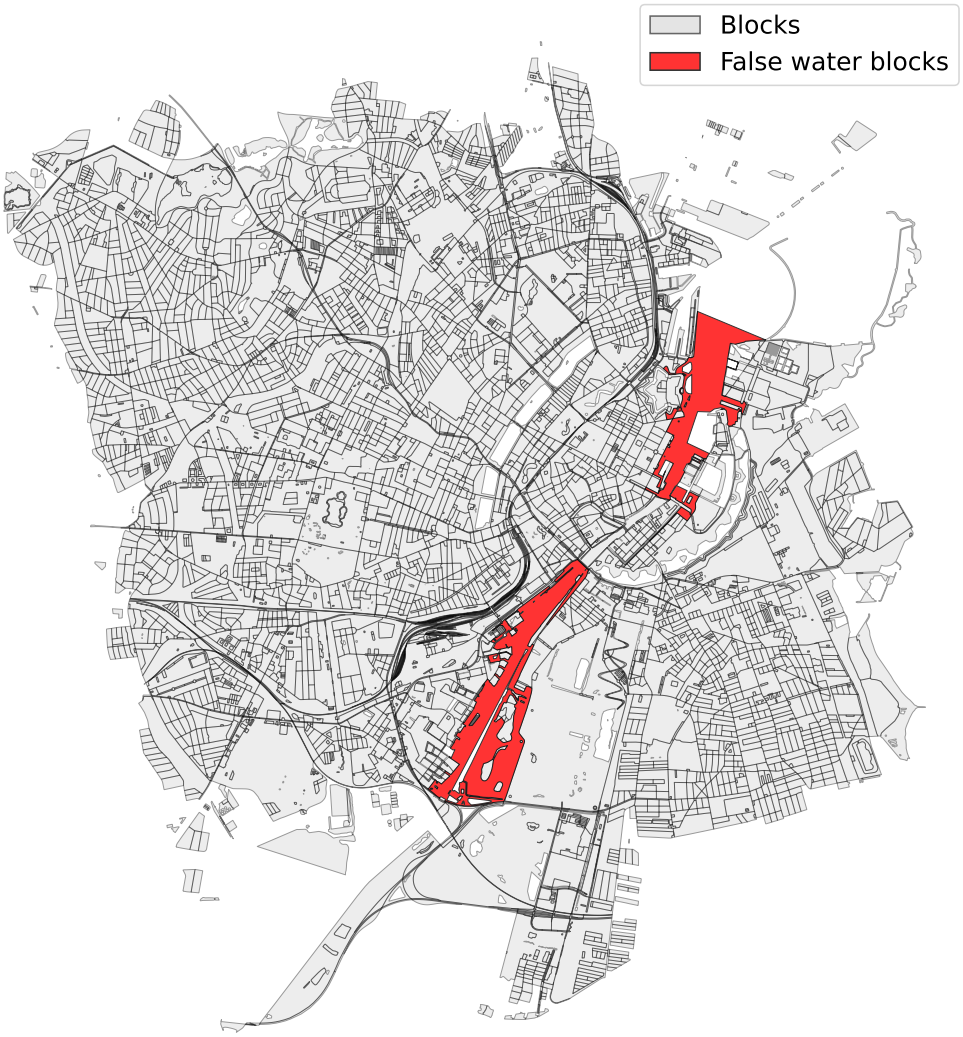


Figure 12: Blocks identified as false water for Copenhagen.

4. **Filtering out irregular blocks**
   Even after removing water polygons and merging small blocks, there were still a number of blocks that did not represent real-world urban

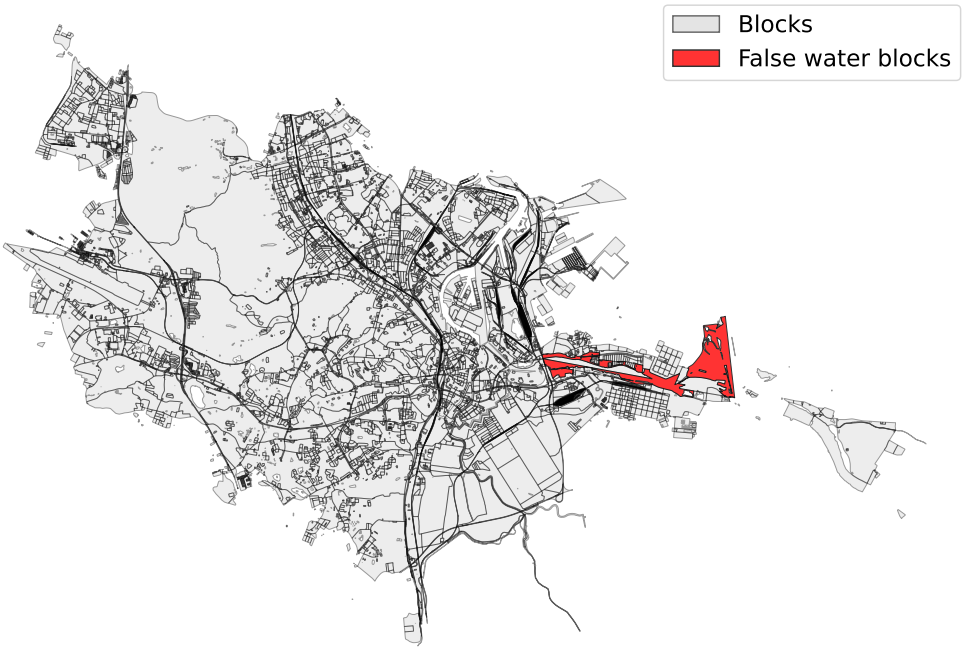


Figure 13: Blocks identified as false water for Gdańsk.

blocks. These blocks were typically long, thin or highly irregular. While some degree of irregularity is expected in urban structure, extreme cases distort the block representation and introduce noise.

To identify such blocks, a geometric compactness measure was applied to the remaining blocks, using the Polsby-Popper [36] formula, defined as:

$$\text{Compactness} = \frac{4\pi \cdot \text{Area}}{\text{Perimeter}^2}$$

This measure ranges from 0 to 1, where values close to 1 indicate more compact shapes (e.g., circular or square polygons), while values closer to 0 correspond to highly irregular geometries. Figure 14 shows blocks at three levels of compactness, demonstrating how extremely low values correspond to geometries that do not represent meaningful urban blocks.

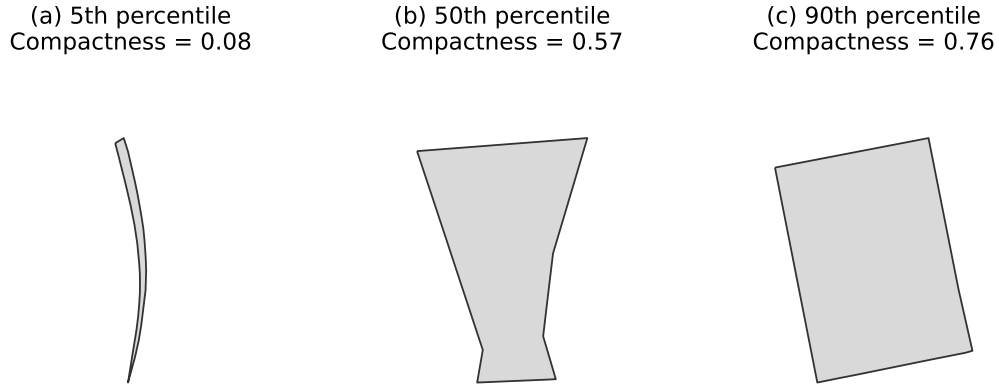


Figure 14: Examples of urban blocks at low (5th percentile), medium (50th percentile) and high (90th percentile) levels of compactness, measured using the Polsby-Popper score.

The distribution of compactness values was computed separately for Copenhagen and Gdańsk (Figures 15 and 16). In both cities, the distributions were right-skewed, with most blocks having moderate to high compactness and a small tail of extremely irregular blocks. Based on these distributions, a compactness threshold of 0.05 was selected for both cities. Figures 17 and 18 show the spatial distribution of blocks below the compactness threshold.

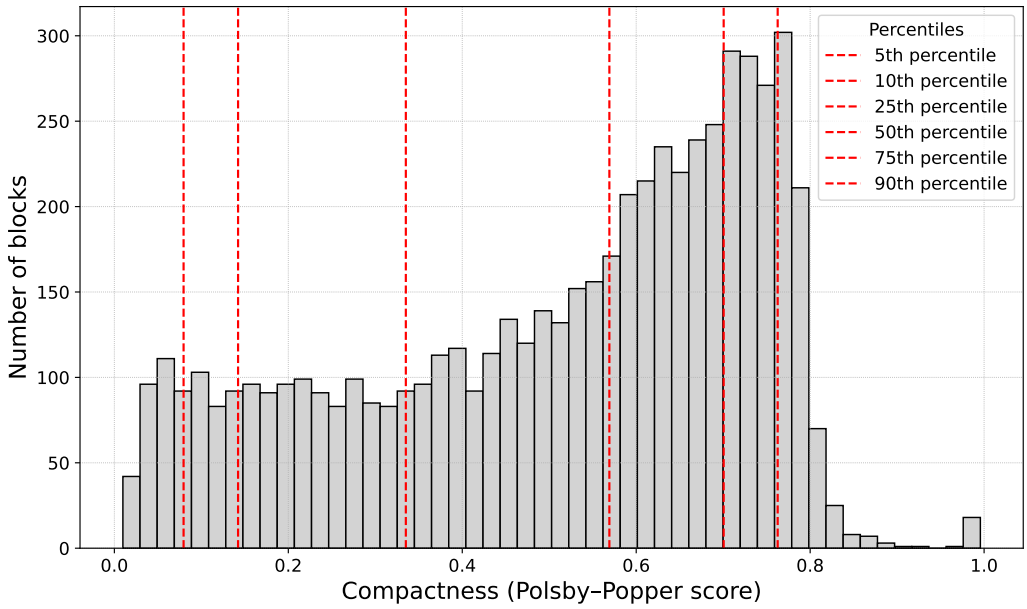


Figure 15: Distribution of compactness scores for all blocks in Copenhagen.

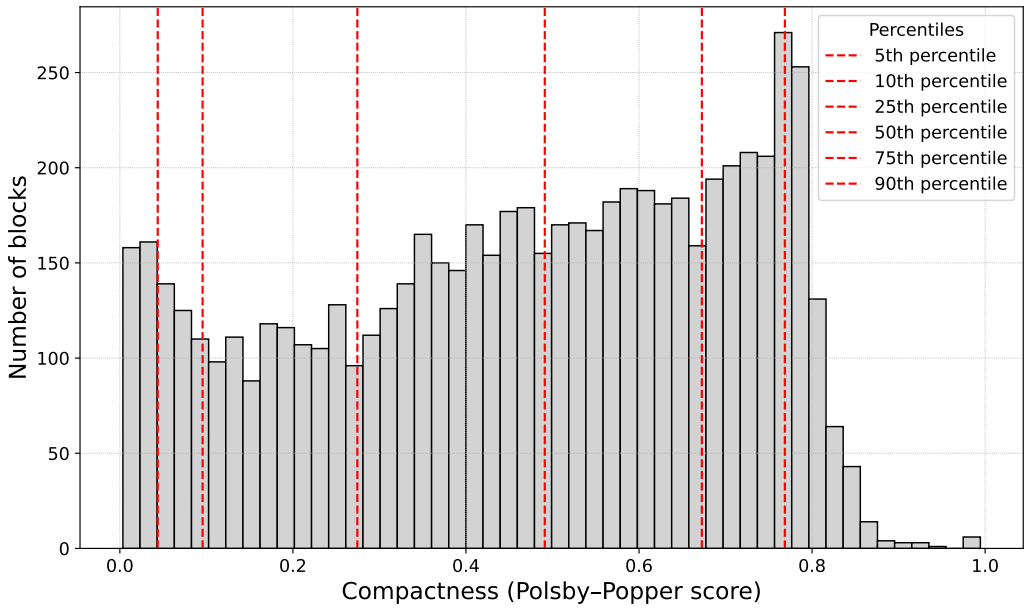


Figure 16: Distribution of compactness scores for all blocks in Gdańsk.

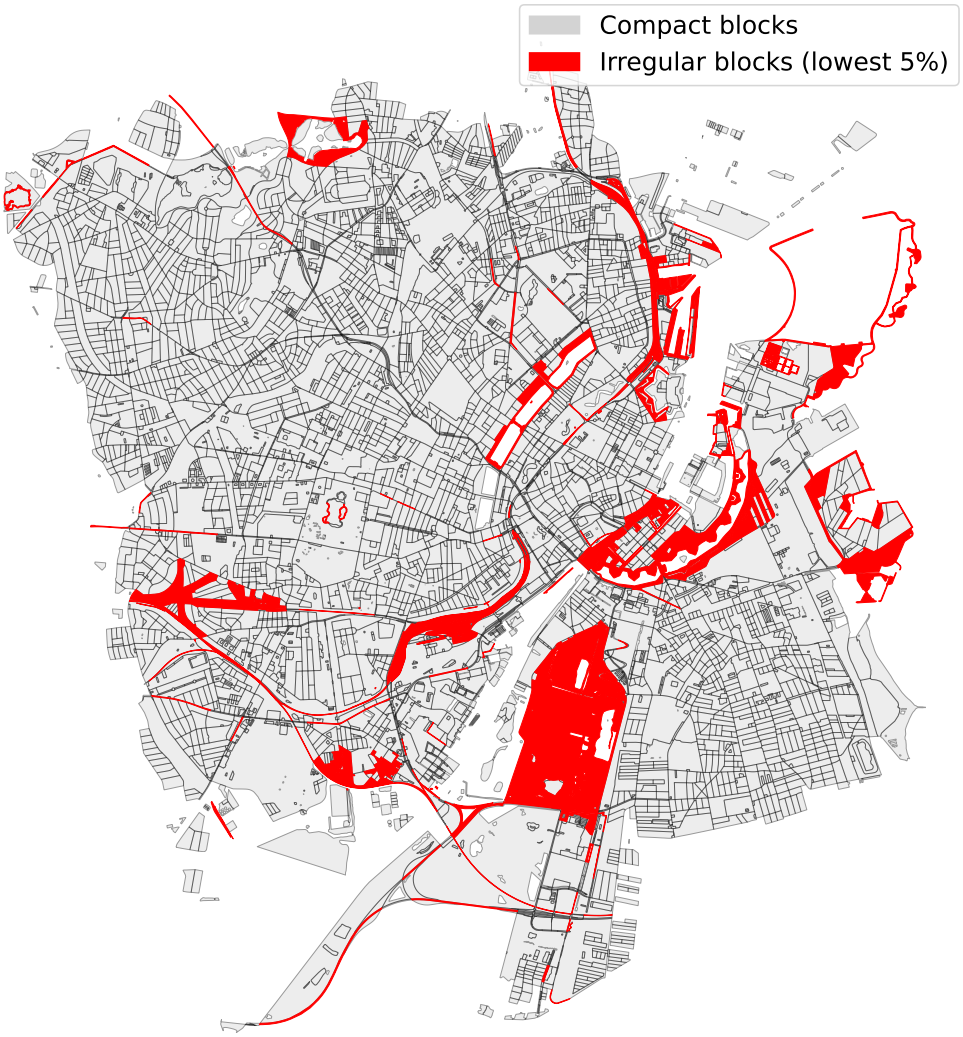


Figure 17: Blocks below the compactness threshold in Copenhagen.

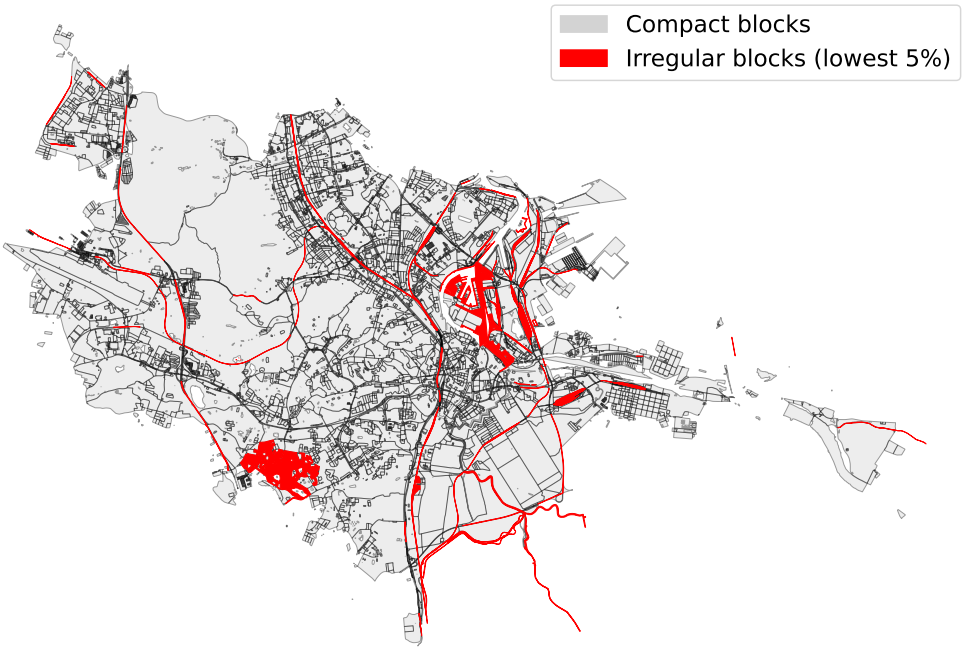


Figure 18: Blocks below the compactness threshold in Gdańsk.

Blocks with compactness values below this threshold were classified as irregular. Rather than removing these blocks entirely, they were iteratively merged into neighboring blocks with higher compactness. For each irregular block, the nearest valid neighboring block was identified using centroid distance, and the geometries were merged. After each iteration, block geometries and compactness values were recalculated. This process continued until no blocks remained below the compactness threshold. This produced a more realistic set of block shapes.

After applying these filtering steps, the final block layer consists of polygons that more closely represent real urban blocks. These blocks form the basis for assigning POIs to spatial units, constructing the block adjacency graph, and computing accessibility and livability measures. The final blocks for Copenhagen and Gdańsk are shown in Figures 19 and 20.

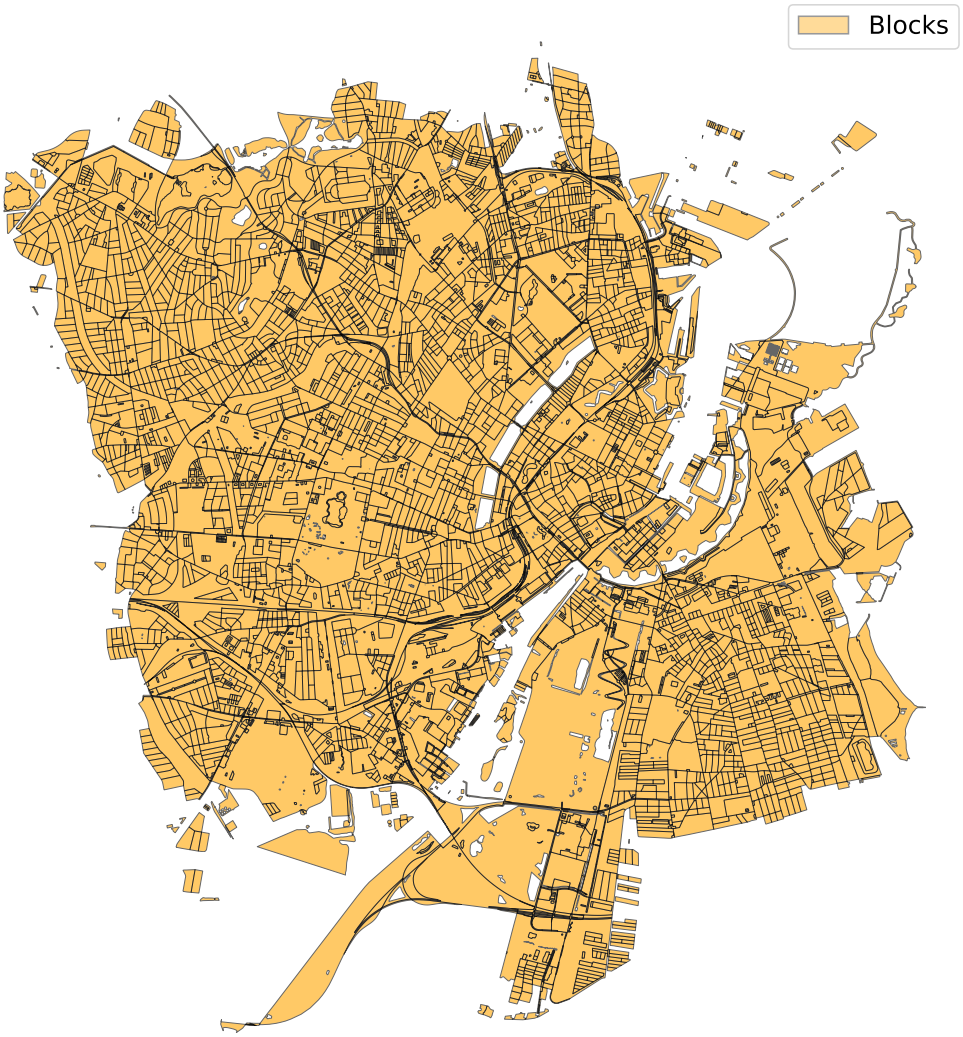


Figure 19: Final blocks for Copenhagen.

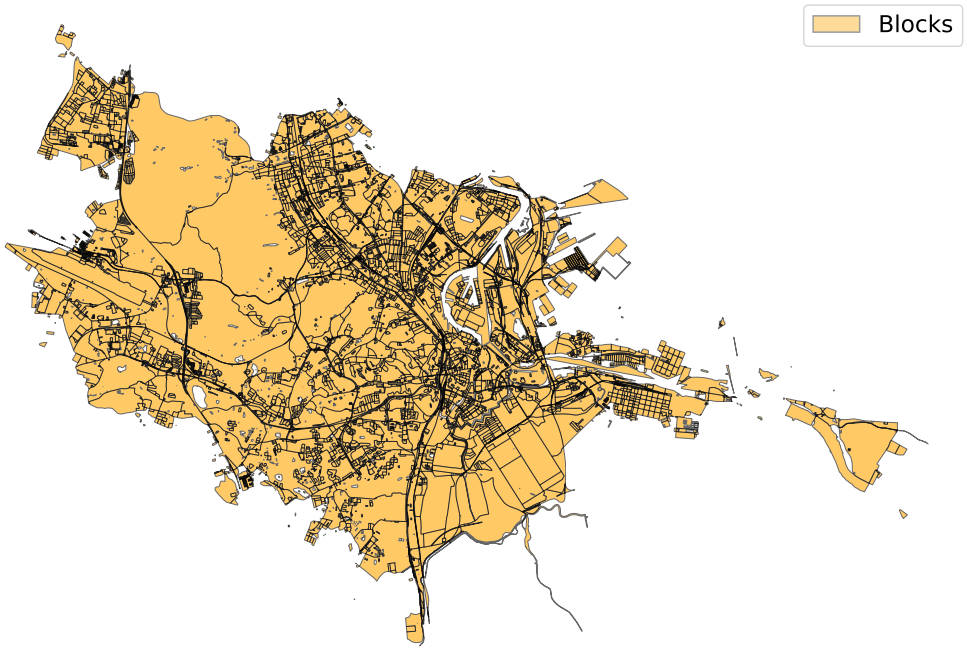


Figure 20: Final blocks for Gdańsk.

A summary of block counts before and after each filtering step is presented in Table 7.

Filtering step	**Copenhagen**	**Gdańsk**
Initial blocks	8 485	10 127
After removing water blocks	7 911	8 703
After filtering small blocks	5 933	6 527
After removing false-water blocks	5 931	6 526
After filtering irregular blocks	5 338	6 003
Number of final blocks	5 338	6 003

Table 7: Summary of block filtering steps for Copenhagen and Gdańsk.

4.5 Assigning POIs to blocks

Once the set of final blocks had been prepared, each POI was assigned to the block in which it is located, using a spatial join. Before doing that, the POI dataset was cleaned: rows with completely missing values were removed, only POIs with a functional category were kept, and IDs were renamed to maintain consistency. The spatial join (using a `within` spatial predicate) then assigned each POI to the block whose polygon contained it. POIs that did not fall inside any block (e.g., those exactly on boundaries or outside the filtered block set) were excluded.

After the join, POIs were aggregated in two ways. First, each block got a simple count of how many POIs it contained (`poi_count`). Second, POIs were grouped by category and counted, creating a set of category-specific counts stored as block attributes. Blocks without POIs received a count of zero.

Not every block has amenities. In Copenhagen, 2829 out of 5338 blocks include at least one POI, and in Gdańsk, 2286 out of 6003 blocks do.

This POI-to-block assignment step produces a block–level dataset in which each block is associated with the amenities located within it. This dataset is then used to build the block adjacency graph and to calculate livability measure. Figures 21 and 22 illustrate the resulting spatial distribution of POI density per block for Copenhagen and Gdańsk, highlighting areas with higher and lower concentrations of amenities.

4.6 Graph construction

The next step was to construct a graph that represents how the urban blocks connect to one another. This graph provides the spatial structure through which accessibility and livability measures are computed.

In the graph, each block corresponds to a single node. The POI count, which is an attribute associated with blocks, is used in the graph structure as a node property. Edges in the graph represent adjacency: two blocks are connected if they share a common boundary.

To determine which blocks are neighbors, comparing every pair of blocks was avoided, and instead a spatial index built from the block geometries was used. For each block, neighboring blocks were identified by checking for intersections between their bounding boxes. A bounding box is the smallest rectangle that fully contains a geometry. If two bounding boxes do not

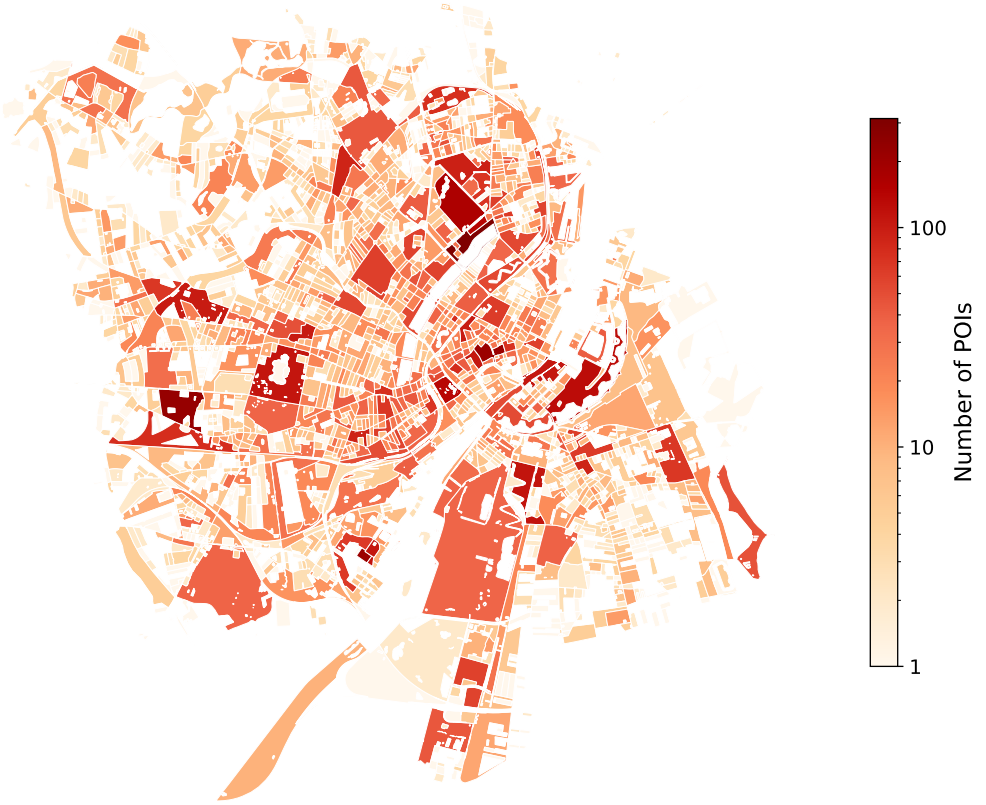


Figure 21: Spatial distribution of POI density per block for Copenhagen.

overlap, their geometries are guaranteed not to touch - except in rare cases involving perfect alignment or minimal precision errors, which are unlikely in real-world urban data. Then, among the candidate neighbors with overlapping bounding boxes, the `touches()` method was used to check whether the actual geometries shared a boundary. If they did, an undirected edge was added between the corresponding nodes in the graph.

This results in an undirected, unweighted adjacency graph, where edges represent spatial proximity. In this way, the graph captures how blocks relate to each other within the urban structure.

For both Copenhagen and Gdańsk, the block graphs contained multiple

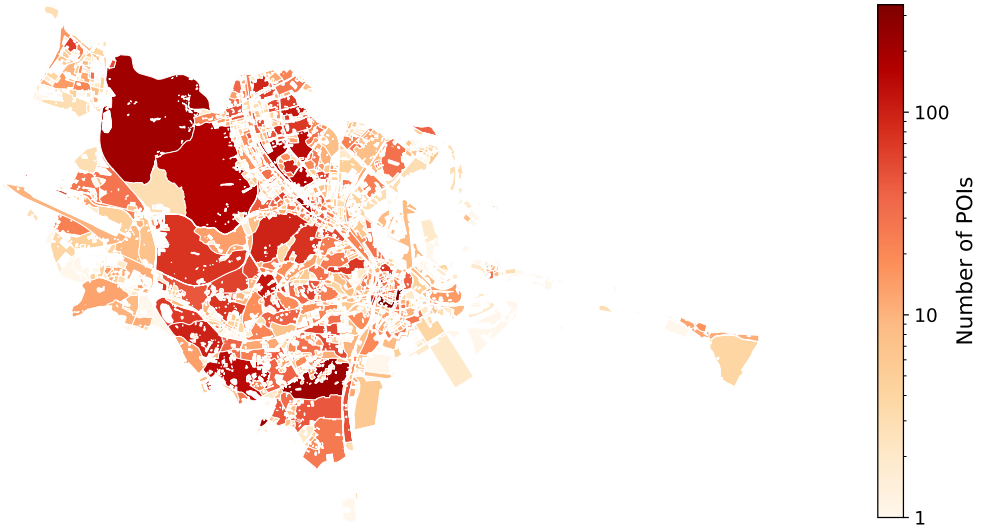


Figure 22: Spatial distribution of POI density per block for Gdańsk.

connected components. This can occur for several reasons:

- water bodies, such as canals or rivers, separate groups of blocks;
- industrial, coastal or port areas may be isolated by infrastructure;
- filtering steps can leave groups of blocks without disconnected;
- islands naturally form separate subgraphs;
- occasional data inconsistencies may prevent some blocks from being connected even though they should be.

Because livability is defined through accessibility across blocks, only the largest connected component of the graph is used in the analysis. Smaller isolated components do not meaningfully contribute to the livability measures, so they are excluded. This approach also ensures that accessibility measures, such as GE distance and network variance, are computed on a network in which all nodes represent reachable parts of the city.

Table 8 shows the number of connected components, nodes and edges. The largest connected components for both cities are shown in Figures 23 and 24.

	Copenhagen	**Gdańsk**
Connected components	81	92
Nodes (all)	5338	6003
Edges (all)	15 802	16 520
Nodes (largest connected component)	5187	5789
Edges (largest connected component)	15 710	16 351

Table 8: Summary of components, nodes and edges for Copenhagen and Gdańsk.

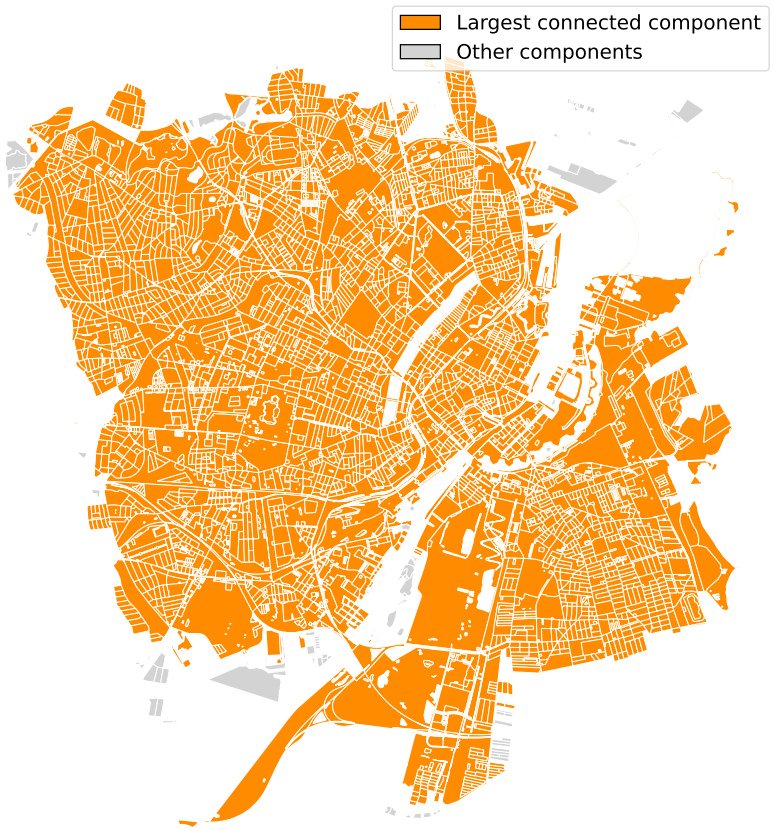


Figure 23: Largest connected component for Copenhagen

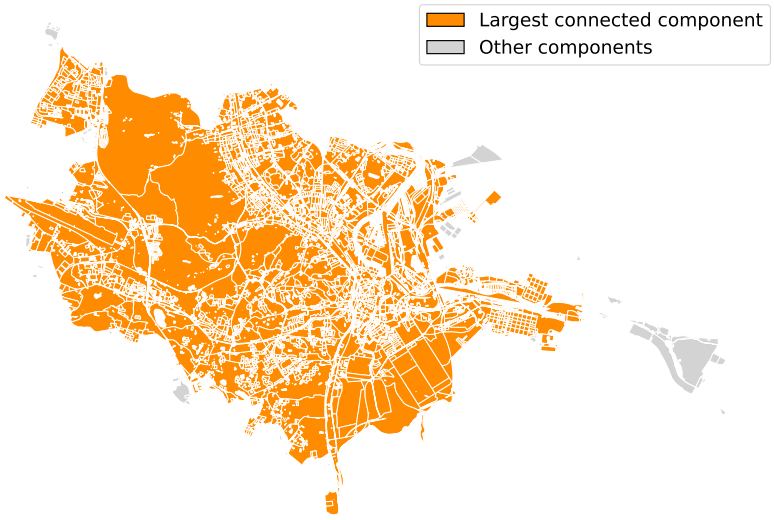


Figure 24: Largest connected component for Gdańsk

The final graph is a representation of the city's block structure. It forms the basis for computing livability measures in the next stages of the methodology. Figures 25 and 26 present the final block graphs for Copenhagen and Gdańsk.

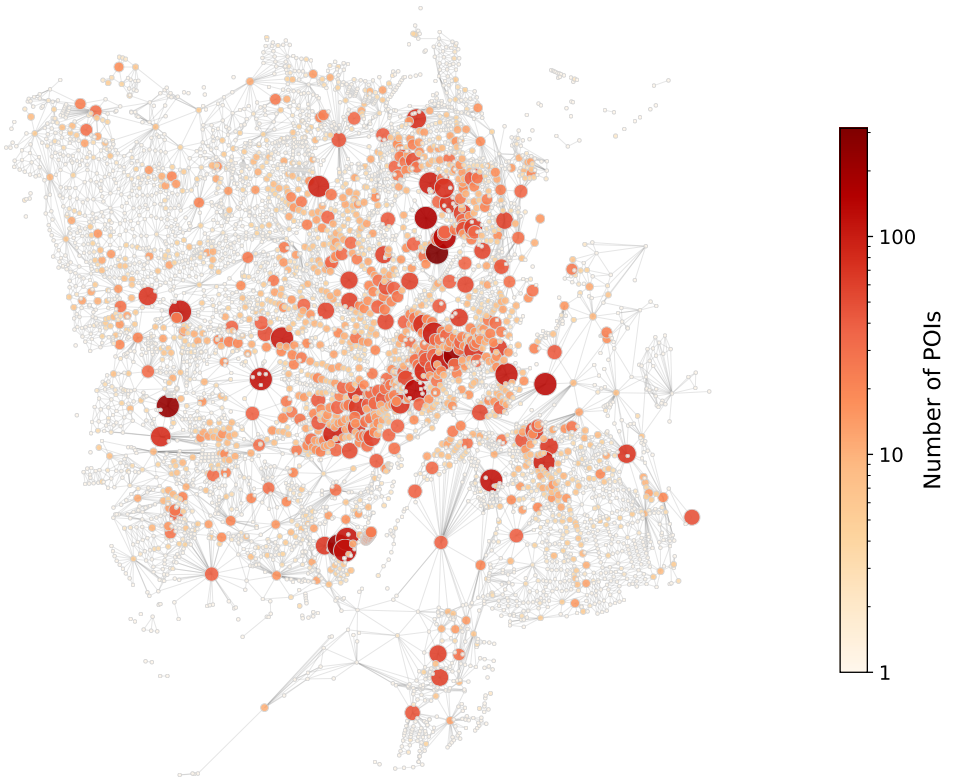


Figure 25: Final block graph for Copenhagen

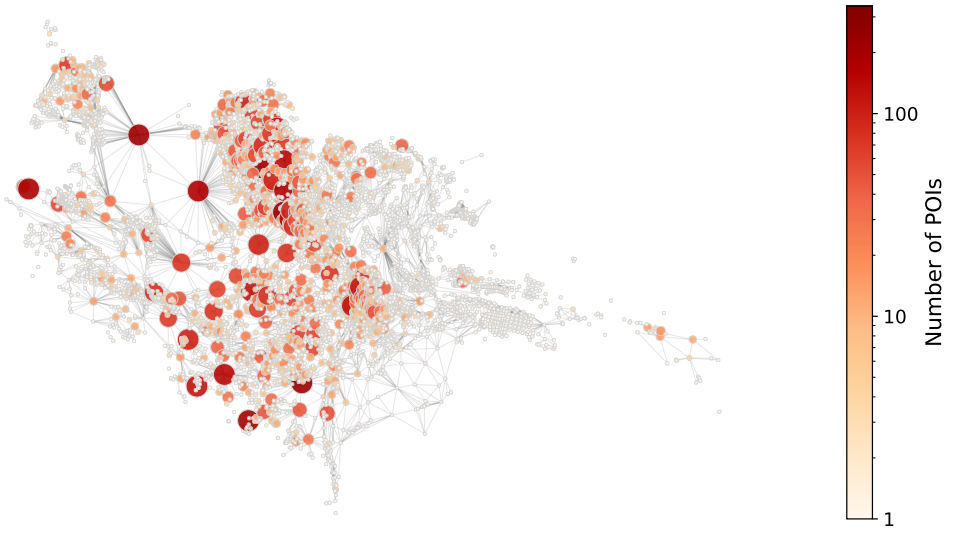


Figure 26: Final block graph for Gdańsk

4.7 Livability score

In this thesis, livability is understood as a measure of how easily residents can access everyday amenities within the urban structure. This livability score reflects the idea that cities are more livable when essential functions, such as shops, services, education, leisure or healthcare, are both close and reasonably distributed across the city. In other words, livability depends not only on how many amenities a city has, but also from how they are spatially organized.

The measure is computed on the block adjacency graph constructed earlier, where each block represents a spatial unit and amenities are assigned to the blocks in which they are located. In this framework, livability is determined by how accessible these amenities are across the block network. Blocks that can reach diverse set categories of amenities within a short distance are considered more livable than those that are poorly connected or located far away. At the same time, livability also depends on how accessibility is distributed across the entire city. A city is not considered livable if only a small cluster of centrally located blocks has excellent access while peripheral areas have very little.

When creating the livability score, there were two scenarios I wanted to model:

- Case A - POI-to-POI.
  This case represents the situation where a person is already at one amenity and wants to reach another. It focuses on how amenities are arranged in relation to each other - the distances between them. On its own, this perspective is not enough. It is limited because it does not capture the general accessibility across the city - it only considers the case, when person is at the POI already.

- Case B - POI-to-home.
  This case looks at accessibility from the perspective of someone living in each block. It captures how amenities are spread out across the city and how easy they are to reach from home. This is more general case than Case A.

The goal of the livability score is to combine both perspectives. Together, they measure how accessible and how well distributed essential amenities are within the city. A high score indicates that amenities are both close and reasonably distributed across the city, so that the travel effort is reduced. In the same way, a low score suggests that amenities are too fart apart, overly centralized or scattered, in the way that requires long or inconsistent travel.

In the final formula, each scenario contributes one component:

- Case A is captured by Generalized Euclidean distance, which measures distances between categories of amenities;

- Case B is captured by network variance, which measures how balanced accessibility is across blocks.

Both components are described in the following sections.

4.8 Generalized Euclidean distance

Generalized Euclidean distance (GE) is used in this thesis to measure how close different amenity categories are to one another within the city structure (Case A: POI-to-POI). In practice, GE produces a distance matrix between amenity categories:

- If two categories tend to appear in the same blocks or areas, their GE distance is small.
- If they appear in very different parts of the graph (e.g. one mostly in the center, the other mostly on the edge), their GE distance is large.

Unlike standard Euclidean distance, which measures straight-line distance between points, GE calculates the distance based on the block adjacency graph. Distances are computed along the network of blocks, which means that the metric reflects how blocks are connected rather than geometric distances.

This distinction is important because geographic distance alone does not capture how people experience accessibility in cities. Physical barriers, street layouts and network structure can make nearby locations difficult to reach, while distant locations may be functionally close if they are well connected.

Although in the current implementation blocks are connected through direct adjacency, the future work includes adding transportation modes to the model (Sections 7.4 and 7.5), such as public transport or cycling networks. With this addition, two spatially distant blocks may be close due to fast or direct connections (e.g., metro or train lines). Therefore, using GE distances provides a foundation for extending the model toward more realistic representations of urban accessibility.

4.8.1 How is GE applied

For each amenity category (food, retail, education, healthcare, infrastructure & transport, culture & leisure, green spaces, public services, daily utilities - Section 4.3.6), a block-level vector is created showing how many POIs of that category appear in each block. Each vector is then normalized to sum to one, so that categories can be compared over the network. This way GE compares spatial distributions rather than absolute counts, so that categories with many POIs do not dominate the distance. In this implementation, GE compares these category vectors by weighting their differences by how the blocks are connected.

Formally, the distance between two categories p and q is:

$$\mathrm{d}_{GE}(p, q) = \sqrt{(p - q)^\top Q (p - q)}$$

This measures how different the two categories are across the block graph. If two categories appear in similar blocks or areas, their difference vector

$(p - q)$ is small, and the resulting GE distance is low. If they appear in different or poorly connected parts of the network, the GE distance becomes larger.

Q is the pseudoinverse of the graph Laplacian. This matrix encodes how strongly blocks are connected: differences between nearby blocks contribute less to the distance, while differences across distant or weakly connected blocks contribute more. As a result, GE is small when two categories occupy similar areas of the city, and large when they occur in different regions. This computation is performed for all pairwise combinations of categories.

4.8.2 Implementation

The computation of GE consists of four steps:

1. **Creating category distributions**
   For each category, the number of POIs in each block is counted. This produces one block-by-block distribution per category. Conceptually, each element of `category_dicts[category]` is a dictionary of the form: {`block_id: number_of_POIs_in_this_block`}
   These dictionaries are then converted into vectors:

   $$p = [p_1, p_2, \ldots, p_n]$$

   where p_i is the number of POIs of that category in block i.

   After normalization (to sum to one), each amenity category becomes a spatial distribution over blocks. For each category, this distribution is represented as a mapping from block IDs to POI counts, describing how amenities of a given type are distributed across the block network.

2. **Computing Laplacian pseudoinverse Q**
   To compare these spatial distributions in a way that reflects the structure of the city, GE uses the pseudoinverse of the graph Laplacian, denoted Q. Differences in the nearby blocks matter less, while differences in far-apart or weakly connected blocks matter more.
   For each city the largest connected component `G_largest` of the block graph is extracted. In the implementation, Q is computed once per city using:

```
Q = _ge_Q(G_largest)
```

For efficiency, Q is computed once per city and stored:

```
Q = compute_and_store_Q(G_largest, CITY_NAME)
```

This avoids recomputing a large matrix for every pair of categories.

3. **Computing GE distances between categories**

   Given two amenity category distributions p and q, their GE distance is computed as in the formula given above.

   In the code, this calculation is performed by:

   ```
   d = ge(category_dicts[c1], category_dicts[c2], G,Q=Q)
   ```

   where c_1 and c_2 are pairs of categories.

   The `ge()` function automatically:

   - converts the dictionaries into aligned vectors over blocks,
   - normalizes them (unless disabled),
   - computes their difference,
   - applies the pseudoinverse Q,
   - returns the GE distance.

4. **Constructing the GE matrix**

   To understand how all categories relate to one another, the GE distance is computed for every pair of amenity categories. This produces a symmetric matrix in which each entry represents how similar two categories are in their spatial distribution across the city.

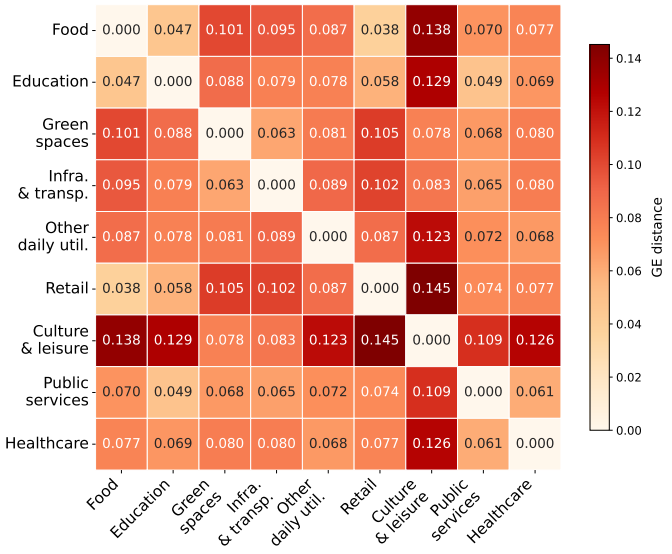
   In the implementation, this is done via:

   ```
   df_ge = compute_generalized_euclidean_matrix(
   G_largest, category_dicts_largest, ge_func=ge, Q_func=
   lambda G: Q)
   ```

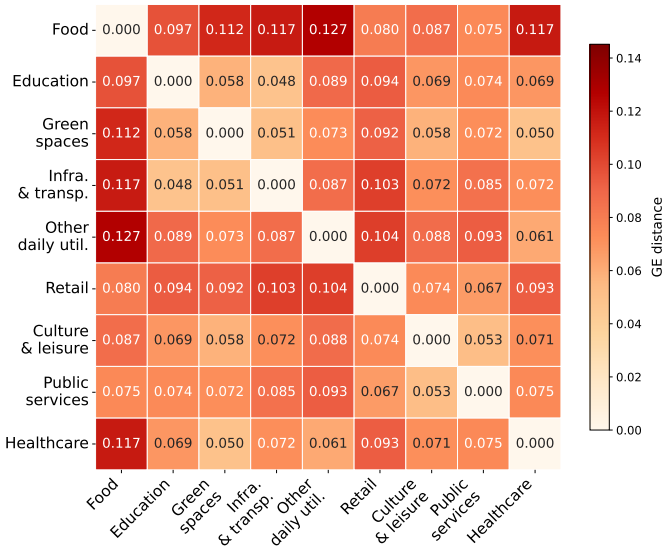
   The resulting matrix is saved and visualized as a heatmap.

4.8.3 Results

Figure 27 shows the GE heatmaps for Copenhagen and Gdańsk.



(a) Copenhagen



(b) Gdańsk

Figure 27: Generalized Euclidean distances between amenity categories for (a) Copenhagen and (b) Gdańsk.

Because lower GE values indicate that different amenity categories tend to occur in the same or nearby blocks, GE contributes to the livability score in an inverted way: lower GE increases the livability score, while higher GE decreases it.

However, GE alone does not capture whether accessibility is well distributed across blocks. Two cities may have similar average GE values but very different spatial patterns: one may be evenly accessible everywhere, while another may be strongly centralized and surrounded by poorly served areas. For this reason, the livability score also includes second component, network variance, which is explained in the next section.

4.9 Network variance

Network variance provides a complementary perspective on livability and accessibility. While the GE measures how far residents must travel to reach amenities, network variance evaluates how evenly amenities are distributed across the city network. This analysis is based on Case B: POI-to-home, which assumes that the resident starts at home - that is, from a block that does not necessarily contain any POIs. Since most residents start their journeys at home rather than at a POI, Case B reflects the general accessibility.

4.9.1 How network variance is applied

Network variance is used to quantify how much we expect any random observation to differ from the average (squared) [37]. Network variance applies the same idea to a graph structure and, in this project, to categories of amenities. It measures how a category is spatially distributed across network.

Network variance is defined as:

$$\mathrm{var}(x) = \frac{1}{2} \sum_{u,v} x_u \, x_v \, d_{uv}$$

where x is the vector assigning to each block u and v the number of POIs of a given category, and d_{uv} is the graph distance between blocks u and v.

In the classical variance formula, distance is measured using Euclidean distance. Here, Euclidean distance is replaced with graph distance, specifically resistance distance, a metric from network theory that captures connectivity between nodes [37]. If POIs of a category are located close together, the pairwise distances d_{uv} between blocks with positive x values will be small,

resulting in low network variance. In the same way, if POIs are spread out across the urban network, network variance becomes high.

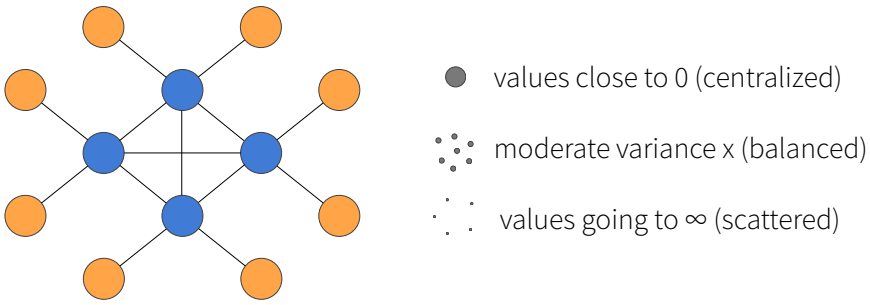


Figure 28: Interpretation of network variance. Blue nodes represent very low network variance (centralized amenities), and orange nodes very high network variance (scattered amenities).

The interpretation of this measure and why extremely low and extremely high network variance is undesirable, is illustrated in Figure 28, which shows two extreme cases:

- (a) Very low network variance (close to 0): when amenities are only in the center (only in blue nodes). POIs are highly centralized. This scenario is not ideal for residents, who will have to constantly travel to the center and stay there.
- (b) Very high network variance (towards infinity): amenities are located only on the periphery (only in orange nodes). POIs are highly scattered, meaning that different types of services are spatially separated rather than clustered, which indicates a fragmented urban structure.

Both extremes are problematic. Desirable urban structure lies in the middle, where network variance is neither too low nor too high - moderate network variance (x). In this case, the spatial structure is balanced, so that amenities are neither centralized nor scattered. In order to find this "moderate network variance", in later steps the randomization process is introduced.

For each POI category, a vector x is computed that assigns to every block the number of POIs of that type, using exactly the same vector structure as

in the GE measure. Combined with the precomputed resistance distances, network variance quantifies how POIs are distributed across the block structure:

- Low network variance - amenities are centralized;
- High network variance - amenities are spread;
- Moderate network variance - balanced spatial structure.

4.9.2 Resistance distance

To compute distances d_{uv} in the graph, a resistance distance is used. It is a graph metric based on the analogy between network and electrical circuits [37].

In this interpretation, each block-to-block adjacency is treated as a resistor, and the distance between two blocks corresponds to the effective electrical resistance between them. Unlike shortest-path distance, resistance distance incorporates all possible paths between two nodes:

- Blocks connected by many short alternative routes have low resistance.
- Blocks connected only by long paths have high resistance.

Effective resistance uses random walks and not shortest paths, which makes is less sensitive to small changes, such as adding or removing a single edge. The resistance distance matrix for the largest connected component is computed once and reused in network variance calculations.

4.9.3 Estimating the moderate network variance using randomization

In order to measure how spread out category is, first it is necessary to find a baseline value of network variance - a value that represents a balanced level of distribution. This makes it possible to judge whether an observed network variance is more centralized or more scattered.

Figure 29 shows a conceptual example of this idea. The grey curve shows the distribution of network variance values that we might get if the POIs were places randomly across the block network. The arrow marks the randomized mean, which is treated as the moderate (ideal) network variance. The two

dashed lines show one standard deviation on each side, giving a sense of what "normal" random variation looks like.

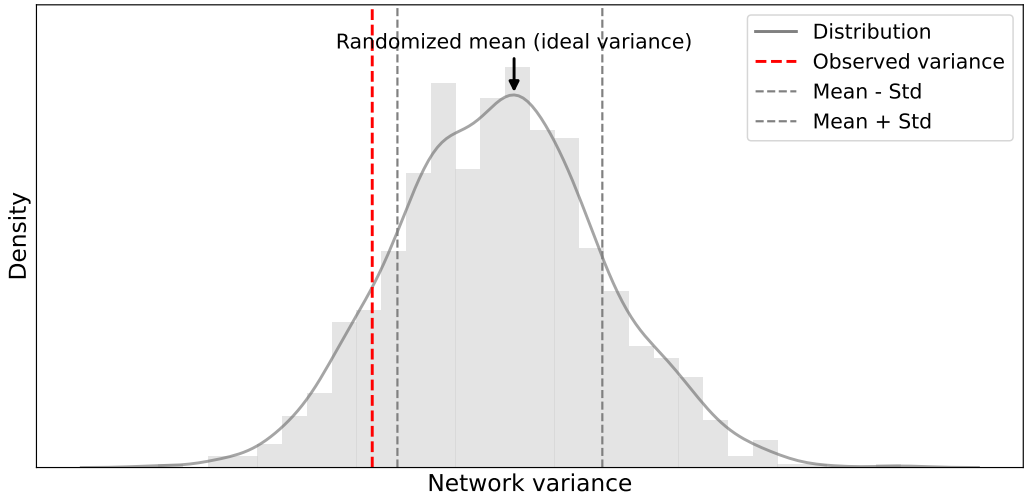


Figure 29: Conceptual distribution of randomized network variances distribution.

By comparing the observed network variance (the red line) to this distribution, we can see:

- If V_{real} is lower than the randomized mean - the category is more centralized than expected.
- If V_{real} is higher than the randomized mean - the category is more spread out than expected.

To find the moderate network variance for both cities, the randomization procedure is used. For each category, the steps are as follows:

1. Keep the block IDs fixed.
2. Randomly shuffle the POI values in the category vector x.
3. Recompute network variance using the same resistance matrix.
4. Repeat the shuffle 1000 times.

5. Collect the resulting variances into a distribution.

This gives a clear reference distribution for each category, which is then used to compute the z-scores.

4.9.4 From network variance to z-score

After computing the network variance for a category, the real variance is compared with the distribution of randomized variances. This comparison shows whether the observed value is lower, higher or close to what would be expected under random placement.

This comparison is expected as a z-score:

$$z = \frac{V_{\text{real}} - \mu_{\text{rand}}}{\sigma_{\text{rand}}}$$

where V_{real} is real variance computed for the city; μ_{rand} is the moderate variance (randomized mean variance); and σ_{rand} is the standard deviation of the randomized variances.

The z-score tells how many standard deviations the real layout is above or below the moderate network variance:

- $z < 0$ - the real variance is lower than what we see in random layouts; the category is more centralized than expected;
- $z > 0$ - the real variance is higher than random; the category is more spread out than expected;
- $z \approx 0$ - the real variance is close to random; the spatial distribution is balanced.

4.9.5 Implementation

The variance and z-score calculations were implemented on the largest connected component of the block graph. The process follows the steps described below:

1. **Precomputing resistance distances**
   Resistance distances are computed between all pairs of blocks in the largest connected component using the `_resistance()` function. Because it is expensive to compute, it is done once, stored and reused.

```
resistance_matrix = _resistance(G_largest)
```

2. **Computing observed network variance**
   For each POI category, a dictionary assigns to every block the number of POIs of that type. This category vector is passed to `variance()`, which computes the network variance using the precomputed resistance distances.

   ```
   variance(v_dict, G_largest, shortest_path_lengths=
   resistance_matrix, kernel="resistance")
   ```

3. **Generating randomized variances**
   To evaluate whether the observed network variance is low or high, a null model is constructed by generating many randomized versions of the category vector. In each iteration the block IDs remain fixed, the POI values are randomly shuffled and reassigned to blocks, and the network variance is recomputed using the same resistance matrix.

   This process is implemented in the `shuffled_variances()` function:

   ```
   random.shuffle(values)
   shuffled_v = dict(zip(nodes, values))
   var_random = variance(shuffled_v, ...)
   ```

   This is repeated 1000 times, creating a distribution of randomized variances. That represents what network variance would look like if POIs of that category were placed randomly, while keeping the number of POIs, the block network structure, and the form of the category vector.

   The mean of this distribution represents the moderate network variance.

4. **Computing z-scores**
   Using the function `compute_z_scores()`, following values are calculated: calculate observed network variance, moderate network variance, standard deviation of randomized variances, and the final z-score for each category.

   The function returns these results for all categories in a DataFrame, together with their randomized variance distributions.

4.9.6 Results

For each POI category, the procedure produces four outputs:

- real network variance,
- moderate network variance,
- standard deviation of the randomized variances,
- the final z-score for each category.

Figure 30 shows an example of the randomized variance distribution on the real data with the observed network variance marked as a vertical line.

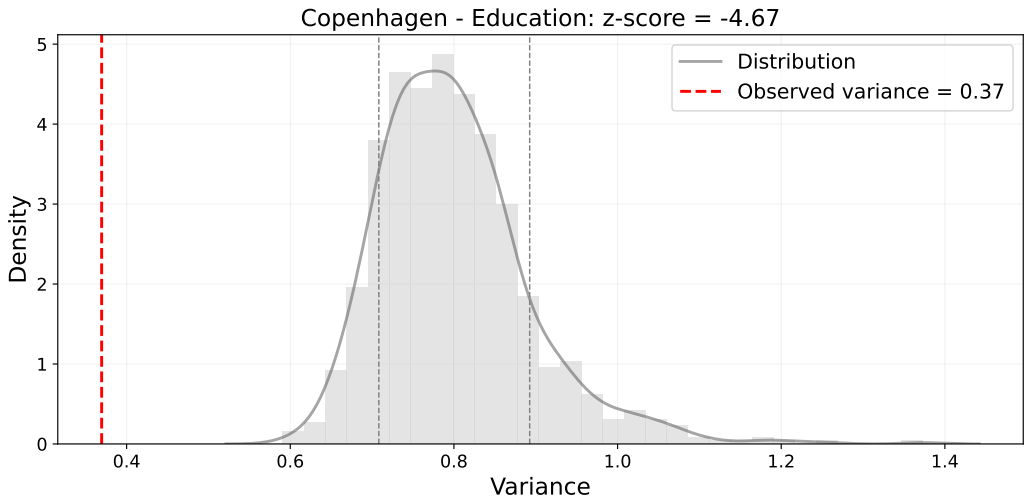


Figure 30: Example of the randomized network variance distribution (in here category "Education" for Copenhagen).

Table 9 shows summary of all z-scores for both Copenhagen and Gdańsk.

Category	**Copenhagen**	**Gdańsk**
Food	-7.22	-3.09
Retail	-8.30	-4.12
Education	-4.79	-4.20
Healthcare	-3.96	-4.60
Infrastructure & transport	-4.49	-4.63
Culture & leisure	-6.08	-2.59
Green spaces	-3.64	-4.02
Public services	-6.26	-4.46
Other daily utilities	-4.50	-5.62

Table 9: Summary table of z-scores for all categories, for both cities.

4.10 Livability score - first results

The first integrated livability score is the combination (specifically multiplication) of the two accessibility components developed in the previous sections:

- Generalized Euclidean distance (GE), which captures how easily each block can access amenities of a given category;
- Network variance, expressed through the z-scores that measure how strongly each category is centralized or scattered, relative to randomized baseline.

These two components reflect different aspects of urban structure. GE measures accessibility, while network variance measures spatial balance. A livable city should provide both short distances to amenities and a reasonable distribution of amenities (neither centralized nor scattered).

The final livability score is in the form:

$$\text{Livability} = \frac{1}{\sum_{c \in C} |z_c| \cdot \sum_{i \in B} GE_{c,i}}$$

Where:

- C = set of POI categories
- B = set of blocks

- $GE_{c,i}$ = Generalized Euclidean value for category c at block i
- z_c = variance z-score for category c
- $|z_c|$ = absolute deviation from balanced distribution

Higher livability score L corresponds to better livability. The score increases when city has shorter GE distances (good accessibility) and smaller network variance imbalances (categories distributed more evenly).

This formulation directly matches the implementation used in the analysis, where z-scores and GE values are combined as:

```
scaled = |z_c| * GE
total = scaled.sum()
livability = 1 / total
```

Table 10 shows the resulting livability score for each city.

	Copenhagen	**Gdańsk**
Livability score	0.030	0.044

Table 10: First livability score for each city.

4.11 Normalization variants

When calculating the first version of livability score, it became clear that the values that go into the formula can vary a lot between cities. Some of these differences reflect the actual structure of the city, but some of them are simply side effect of the city's scale. Factors such as the number of POIs, the number of blocks (nodes), the distribution of GE distances, or the distribution of amenities have a strong impact on the final score.

Because of this, it was important to think about whether the livability score should allow comparisons between cities. Without that, the score loses much of its meaning. A larger city should not appear more livable simply because it has more blocks or more amenities, and a small city should not seem less livable just because it is physically smaller. The score should reflect accessibility and spatial organization, not the city's size.

Without any normalization, however, the score tends to reflect these scale differences more than the actual spatial patterns. This is why exploring

normalization methods became necessary - to see which variants could help fix this issue and make the score more comparable across cities.

Different normalization techniques address different problems. Some reduce city-size bias, others try to correct for skewed GE distance distributions, and others introduce weights that better reflect block properties. Since each method solves only part of the problem, it can sometimes be helpful to combine multiple variants.

Exploring these normalization methods made it possible to understand what the score is sensitive to, what it ignores, and which parts of the formula change when the city changes size. This helped justify why the final version of the livability score remains unnormalized in this project, but also why it needs future improvements.

Four normalization variants were tested to understand how they affect the livability score.

4.11.1 Replacing sum with average

The original livability score uses a sum over all block-amenity pairs. This means that larger cities automatically produce larger values simply because they have more blocks (nodes) and more pairs of distances.

Replacing the sum with an average tries to remove this size effect. Instead of measuring "total accessibility" across the whole city, the score becomes closer to an "average accessibility per block". This helps reduce the direct "more blocks $\rightarrow$ bigger score" bias, which is useful for comparing cities of different sizes.

However, using the average changes the interpretation of the score. It reduces the differences between sparse and dense cities. High-accessibility blocks no longer dominate the result, since their values are balanced by many lower-accessibility blocks.

4.11.2 Normalized GE (total and row-wise)

GE distances can also be normalized, either globally (dividing all values by the maximum GE in the city) or row-wise (scaling each block's distances to a 0-1 range). This idea is to make GE values comparable across cities.

In practice, however, this introduces a different problem: it removes the sensitivity to city size almost completely. Since all distances are in the same fixed range, large and small cities end up looking structurally similar even if

one is physically ten times larger. The score then loses the ability to judge real spatial differences, because the original distance scale has been removed. In extreme cases (row-wise normalization), all blocks behave as if they were located in cities of the same size.

Because accessibility is fundamentally tied to physical distance, GE normalization ends up eliminating an important part of the livability measure. The livability score should consider the size of the city. Although it should not be the most influential factor, it still should have some effect the results.

4.11.3 Logarithm of GE

GE distances have a distribution that most distances are short, but a few very large ones dominate the sum. To reduce the influence of these extreme values, it was tested how the result changes when logarithm of GE is used.

Applying a logarithm helps to stabilize the score by preventing a few large distances from overshadowing other ones. It also makes the score less sensitive to outliers and unusual parts of the network (for example, those near the city's boundary).

However, the log transformation changes the meaning of "distance" in a non-linear way. Distances no longer behave proportionally the way they did before. The logarithm compresses distances so that far-away blocks no longer appear dramatically less accessible than nearby ones. This makes the interpretation less intuitive.

4.11.4 Weighted graph version

Another idea was to incorporate weights into the block adjacency graph. The intuition was that not all blocks contribute equally to the the functioning of the city. Some blocks are more connected, while others are more isolated.

To capture this, weights were added to edges based on inverse node degree:

$$w(u, v) = \frac{1}{2}\left(\frac{1}{\deg(u)} + \frac{1}{\deg(v)}\right)$$

The weighting scheme was based on inverse node degree. The degree of the block represents the number of neighboring blocks it shares an edge with in the graph. Blocks with many neighbours (high degree) typically belong to dense, central areas of the city. Blocks with few neighbours (low degree) are more peripheral or constrained by infrastructure. This means that edges

connected to high-degree blocks receive smaller weights, and edges connected to low-degree blocks receive larger weights.

When computing GE distances on this weighted graph, paths that pass through isolated or peripheral blocks become "longer," while paths through well-connected areas become "shorter." While this captures structural differences of the cities, it also makes GE distances harder to interpret.

4.11.5 Small-scale test

To verify the behavior of the normalization variants, each method was also tested on a small, controlled example: a 3×3 matrix and a 3-element POI vector. All normalization functions were adapted to work on this simplified example. The goal of this was to observe the mathematical effect of each normalization method.

While the following examples provide intuition about the livability variants, their behavior is explored more thoroughly in Section 5 using random graph.

The following cases were checked:

- a base city
- a "larger" city (all GE values x10)
- a "smaller" city (all GE values ÷10)
- denser amenities (POI vector decreased)
- perfectly balanced amenities ($z = 0$)
- no amenities at all

In the unnormalized variant, when distances increased, the livability score dropped by the same factor, and when distances decreased, the score increased proportionally. Also, adding more amenities produces a small increase in the score. This confirmed that the unnormalized score is tied to both city size and amenity volume.

The averaged version showed the same behavior, only with smaller numerical values.

When GE was normalized, it produced identical scores for the base, larger, and smaller cities, and even adding more POIs did not change the result. It

shows that once GE is normalized, the livability score loses sensitivity to physical scale.

The logarithmic version reduced the impact of very large distances, so the bigger city scored slightly lower and the smaller city slightly higher. The effect of additional amenities was still visible.

Overall, the small-scale test showed that unnormalized and averaged versions remain sensitive to true physical distances, GE normalization removes this sensitivity entirely, and logarithmic transformation modifies but does not eliminate it.However, it does not check how the score acts when there is more or less number of blocks.

4.11.6 Comparison of results

The livability scores for all normalization variants are summarized in Table 11. The results show that each method affects the score in different way. Although for almost (except for logarithm of GE) every variant the livability score is higher for Gdańsk than for Copenhagen.

Variant	**Copenhagen**	**Gdańsk**
Unnormalized	0.030	0.044
Replacing sum with average	2.181	3.139
Normalized GE (total)	0.182	0.251
Normalized GE (row-wise)	0.020	0.028
Logarithm of GE	-0.0005	-0.0007
Weighted graph	0.007	0.015

Table 11: Summary of livability scores for all normalization variants.

At this point, the unnormalized and averaged versions behave in the most interpretable way. Normalizing GE causes the score to ignore the actual size of the city, which is not desirable. The logarithmic variant behaves inconsistently: it produces negative scores and reacts to changes in distance and amenity distribution in less predictable ways.

4.12 Pipeline reusability

The pipeline developed in this project was designed to be reused for other cities - to calculate their livability scores and to compare them with each

other. Most of the steps operate on OSM data and follow a fixed logic, but some parts depend on city's characteristics, tagging conventions or spatial structure. Because of this, pipeline is mostly reusable, with some parts that are city-specific and require manual adjustments.

4.12.1 Components that require manual adjustments

To make the pipeline work as intended, several elements must be adapted for each new city.

- **City and boundary names**
  To download the `.pbf` files for each city, Pyrosm uses English names such as *Copenhagen* and *Gdansk*. However, when reading the tags inside the file, OpenStreetMap uses local naming for administrative boundaries - e.g., *Københavns Kommune* instead of *Copenhagen*, *Gdańsk* instead of *Gdansk*.

  Therefore, the correct boundary must be identified manually and added to the name-mapping.

- **Differences in `admin_level`**
  Along with the city name, in order to extract city boundary it is necessary to define correct `admin_level`. This tag defines which administrative unit is selected. It is used early in the pipeline to filter all spatial data to the extent of the city.

  Administrative hierarchies differ across countries, so the `admin_level` representing a city boundary in Denmark is not necessarily the same as in Poland or other regions. Some cities also appear multiple times in OSM under different `admin_level` values.

  Therefore, it is necessary to check the available administrative levels for each country and city, and manually adjust it if needed. OSM provides a template for checking possible admin levels [35].

- **POI categorization**
  The set of amenities in each city is different, and OSM tagging practices are not standardized across countries. The categorization mapping in this pipeline is based on POI data from Copenhagen and Gdańsk.

A new city may contain POI tags that do not appear in these datasets, ot use uncommon tag combinations. For this reason, the POI categorization should be reviewed and, if necessary, expanded or cleaned to include missing but important types of amenities.

- **Block filtering thresholds**
  Block filtering thresholds for both area and compactness were defined based on distributions observed in Copenhagen and Gdańsk. Although percentile-based thresholds remove the same proportion of blocks in each city, the blocks included below a given percentile may differ between them. Differences in urban structure, block size distributions and street patterns mean that a threshold corresponding, for example, to the lowest 25th percentile in one city may represent different sizes and types of blocks in another. Therefore, when applying the filtering step to a new city, it is necessary to verify the resulting blocks and adjust thresholds so that they match the local urban structure.

- **Water filtering**
  Cities also vary in terms of presence of water bodies: some have coastlines and harbors, others have rivers or canals, and some have no water at all. Because of this, the water-handling steps may require modifications.

  In this pipeline, water blocks are removed, but for both Copenhagen and Gdańsk some blocks containing water were not flagged correctly as water blocks. It can happen for many reasons, such inaccurate boundaries or incomplete OSM water data. For some cities there might be similar issues, which are not solved by this pipeline.

  Therefore, water-related blocks should be checked manually. If water blocks are not removed correctly, it is necessary to find and handle the issue.

4.12.2 Components that are reusable

Many steps of the pipeline can be applied directly to any city without modifications.

- **Block construction**
  The approach of combining roads, railway lines, and water edges into a

single boundary layer and polygonizing it into blocks works for all cities. This part is independent of local naming, local tagging conventions, or city size.

- **Graph construction**
  The method for creating the block adjacency graph - using spatial indexing and the `touches()` relationship - remains identical across cities. Nodes always represent blocks, and edges represent shared boundaries.
- **Network variance and z-score calculation**
  The method for generating randomized distributions, computing moderate network variance, and calculating z-scores does not depend on the location or type of city.
- **Livability score computation**
  All livability score variants are fully reusable. Once the block graph and POI attributes are ready, the computation does not require any city-specific changes.
- **Normalization methods**
  All normalization variants operate on matrices and vectors in the same way regardless of the city.

5 Testing

5.1 Purpose of testing

The goal of testing is to understand how the livability score behaves when the structure of the graph changes. The tests focus on two main properties of the network: the number of nodes and the number of POIs assigned to them. Changing these properties makes it possible to observe how the livability scores scales with city size and amenity density.

The tests evaluate all livability score variants developed in this project, that is:

- Unnormalized version,
- Average version (replacing the sum with a mean),
- GE normalized, so that the full matrix sums to 1,
- GE normalized row-wise, so that each row of the matrix sums to 1,
- Logarithmic transformation of GE,
- Weighted graph version.

Each variant was evaluated on the same synthetic graphs and POI assignments. This allows for comparison how different variants affect the livability score and how do they behave. It helps to later decide which variant to choose as the final livability score - the variant that behaves in the way the measure is intended to work.

The code for testing is provided in the `Testin_Code.pdf` file submitted along with this report, which has been exported from the Jupyter Notebook.

5.2 Generation of random graph

The tests are performed on synthetic random graphs to analyze how the livability score responds to changes in network size and amenity density under controlled conditions. Using random graphs makes it possible to control specific parameters while keeping everything else constant, so any changes in the livability score come directly from the chosen inputs.

Graphs are generated using the $G_{n,m}$ random graph model, where a fixed number of nodes n and a fixed number of edges m are specified (using

`gnm_random_graph`). In this model, edges are placed uniformly at random between node pairs. This results in graphs without predefined spatial or hierarchical structure. Edge density is kept fixed across different graph sizes to allow meaningful comparisons.

This choice keeps the structure of the network simple and unbiased, making it suitable for checking how the livability score responds to changes in graph size and amenity density, rather than to specific urban patterns. For each selected graph size, only the largest connected component is used, and its Laplacian matrix Q is precomputed.

Each test varies two parameters:

- number of nodes $n \in \{100, 200, 400, 800, 1600\}$
- average number of POIs per node $\{1, 2, 4, 8, 16\}$

This creates 25 parameter combinations. the test was repeated 10 000 times, each time generating new POI distributions and computing all livability variants. For every combination, POI counts are generated from a normal distribution with the chosen mean, assigned to all nodes, and used to compute z-scores and GE matrices. All livability score variants are then evaluated on the same data to allow direct comparison.

Because the structure of the graph does not change between iterations, matrices such as Q and GE are precomputed once per graph size, which reduces computation time.

5.3 Test results

Here are the results for all the normalization variants:

- **Unnormalized version**
  The number of nodes has a strong impact on the livability score. Graphs with more nodes produce higher scores, meaning that having more blocks leads to a higher livability value. POI density also affects the score, but the effect is more subtle: adding more POIs increases the score at first, and then it leads to nearly linear increase. Overall, this variant behaves in a predictable and interpretable way.
- **Average version**
  The average variant behaves almost identically to the unnormalized

version. Larger graphs still produce visibly higher scores, and the score increases smoothly with POI density. This suggests that dividing by the number of categories does not change the general behavior of the score.

- **GE normalized to total sum = 1**
  Normalizing the GE matrix, so that its total sum equals one, removes most differences between graph sizes. Larger graphs no longer produce higher scores, and the score changes only slightly when POI density increases. As a result, the method no longer distinguishes between small and large networks.

  The resulting scores are also difficult to interpret. Intuitively, a graph with more POIs per node on average should give a better livability score, because there is more services. But this is not reflected here. For example, for some POI means, the score for $n = 400$ is lower than for $n = 100$ which contradicts the intended meaning of livability.

- **Row-wise GE normalization**
  The row-wise normalization behaves similarly to the total-sum normalization shows the same issues. The scores stays very similar across different graph sizes and POI densities, and the resulting values are hard to interpret.

- **Logarithmic GE**
  The logarithmic variant produces only negative scores, because all GE values lie below 1 and their logarithms are negative. The score changes slightly with POI density, but the overall variation remains extremely small. But there is still visible difference between different graph sizes, similarly to unnormalized version.

- **Weighted graph version**
  The weighted version behaves similarly to the unnormalized and average variants. The score increases with graph size, and shows a slight increase with POI density. Although differences between values are smaller due to the weighting, the overall behavior is consistent and interpretable.

The unnormalized and average variants show clear, stable, and intuitive behaviour. From all versions, these two match the intended interpretation of livability the best - scores increase with both graph size and POI density.

Based on all the observations, the unnormalized livability score variant was chosen as the final variant in this thesis. It is the most suitable and interpretable formulation from all the proposed variants. Figure 31 shows the behavior of this unnormalized version. Figures for all other normalization variants are provided in Appendix D. Although the unnornalized variant has some issues that will need future improvements, which are described in Section 7, it is a good starting point.

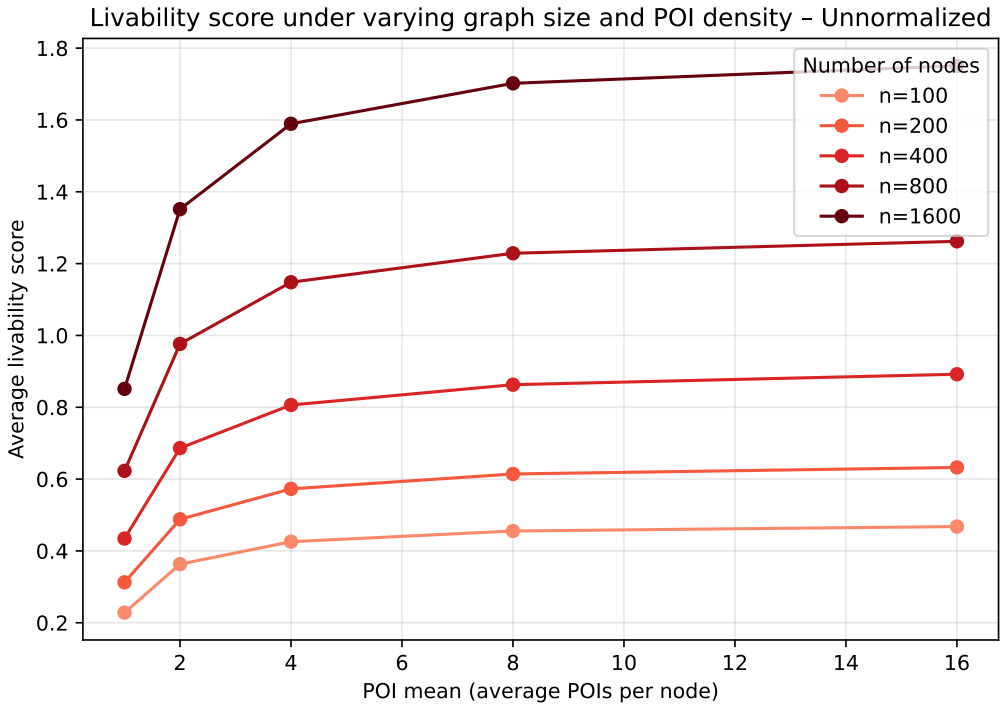


Figure 31: Behaviour of the final livability score under varying graph size and POI density.

6 Results

The final livability scores were computed for both Copenhagen and Gdańsk using the unnormalized version of the livability measure. The resulting values indicate that Gdańsk achieves a higher livability score than Copenhagen.

In this thesis, a higher score corresponds to higher livability. It is constructed in the way, that it increases when amenities are more evenly distributed across the network and decreases when amenities are strongly concentrated in specific areas. So according to the proposed network-based measure, the results suggest that Gdańsk is assessed as more livable than Copenhagen.

To better understand why these values differ between the two cities, this section presents a comparative analysis of Copenhagen and Gdańsk. Specifically, the main components contributing to the livability score. Key summary statistics for both cities are shown in Table 12.

6.1 Comparative analysis of score components

6.1.1 Distribution of block sizes

Gdańsk contains a larger number of blocks than Copenhagen, as shown in Table 12.

The cities also differ significantly in terms of average block size and the distribution of block sizes. Figure 32 illustrates this difference. It shows that Gdańsk has a wider range of block sizes, with many small blocks alongside some larger ones. In contrast, Copenhagen has a more uniform distribution, with blocks that are generally medium-sized.

Because each block corresponds to a node in the network, differences in block structure directly influence the livability score. A higher number of blocks increases the number of nodes over which amenities can be distributed. It affects how POIs are aggregated and how network-based measures behave. As a result, livability score is higher for cities with larger number of nodes, which may not fully reflect the intended focus of the measure.

6.1.2 GE distance distribution

The GE distance distributions were compared to check whether differences in network distances could explain the results. Despite differences in overall

	Copenhagen	**Gdańsk**
Structure		
Number of nodes	5167	5798
Number of edges	15 710	16 351
Average degree	6.055	5.646
Density	0.00117	0.00098
Blocks		
Number of blocks	5338	6003
Average block area	16456.35	30688.07
Number of blocks with POIs	2829 out of 5338 (53%)	2286 out of 6003 (38%)
Average compactness	0.53	0.49
POIs		
Number of POIs	31 021	33 368
POIs per block	5.81	5.56
Graph		
Largest connected component size	5167 nodes out of 5338 (97%) 15 710 edges out of 15 802 (99%)	5798 nodes out of 6003 (97%) 16 351 edges out of 16 520 (99%)
Average GE distance	0.075	0.072
Average z-score	-5.429	-4.182

Table 12: Key summary statistics.

city layout and block structure, the GE distance distributions for Copenhagen and Gdańsk are similar.

This similarity is illustrated in Figure 33, where the distributions largely overlap. The average GE values in Table 12 support this observation.

To compare the distributions, a Mann–Whitney U test was performed, resulting in $U = 685.0$ and $p = 0.681$, indicating no statistically significant difference between the two distributions.

These results suggest that GE distances are not the reason of the difference in livability scores. Instead, the similarity of GE distributions suggests that the underlying structure of the block networks is comparable in terms

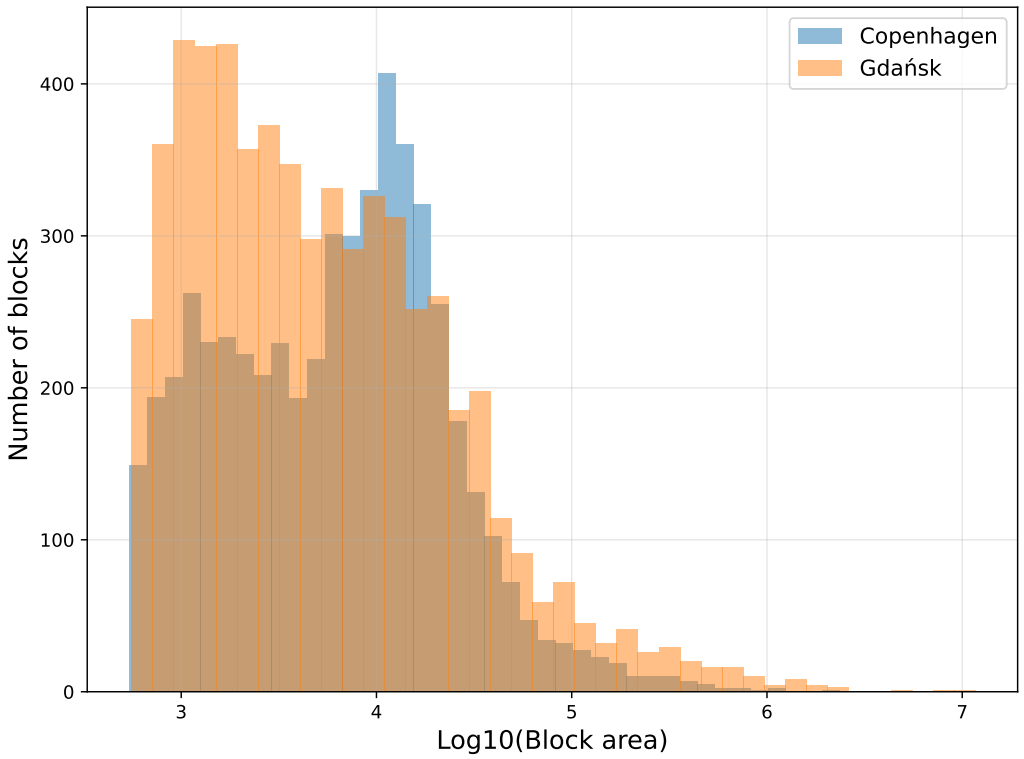


Figure 32: Distribution of block sizes for Copenhagen and Gdańsk

of network distances.

6.1.3 Z-score distribution of network variance

A contrast between the two cities appear in the distribution of z-scores. Copenhagen has more extreme z-score values, indicating that amenities are more concentrated in some areas of the city. In contrast, Gdańsk shows z-scores closer to the randomized baseline, reflecting a more even distribution of amenities. This difference is illustrated in Figure 34.

These differences in z-score distributions have a direct impact on the livability score. The measure penalizes strong spatial concentration of amenities and favors more evenly distributed patterns. More balanced amenity distribution observed in Gdańsk explains its higher livability score. So z-score differences explain most of the score gap between two cities.

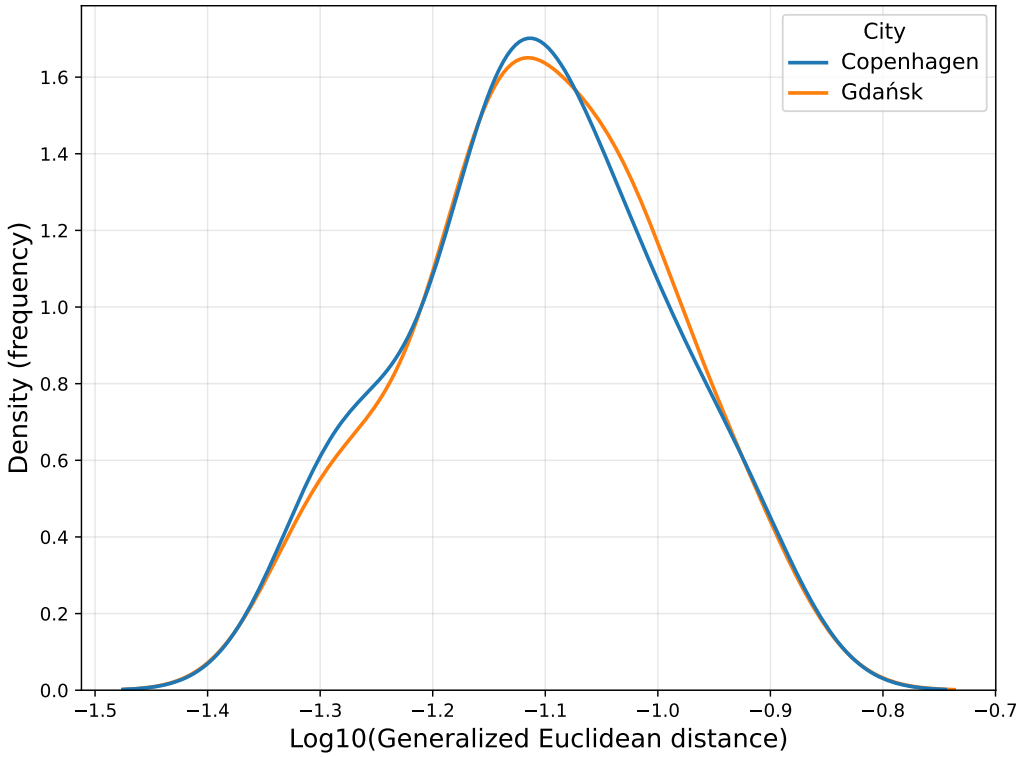


Figure 33: Distribution of GE distances for Copenhagen and Gdańsk

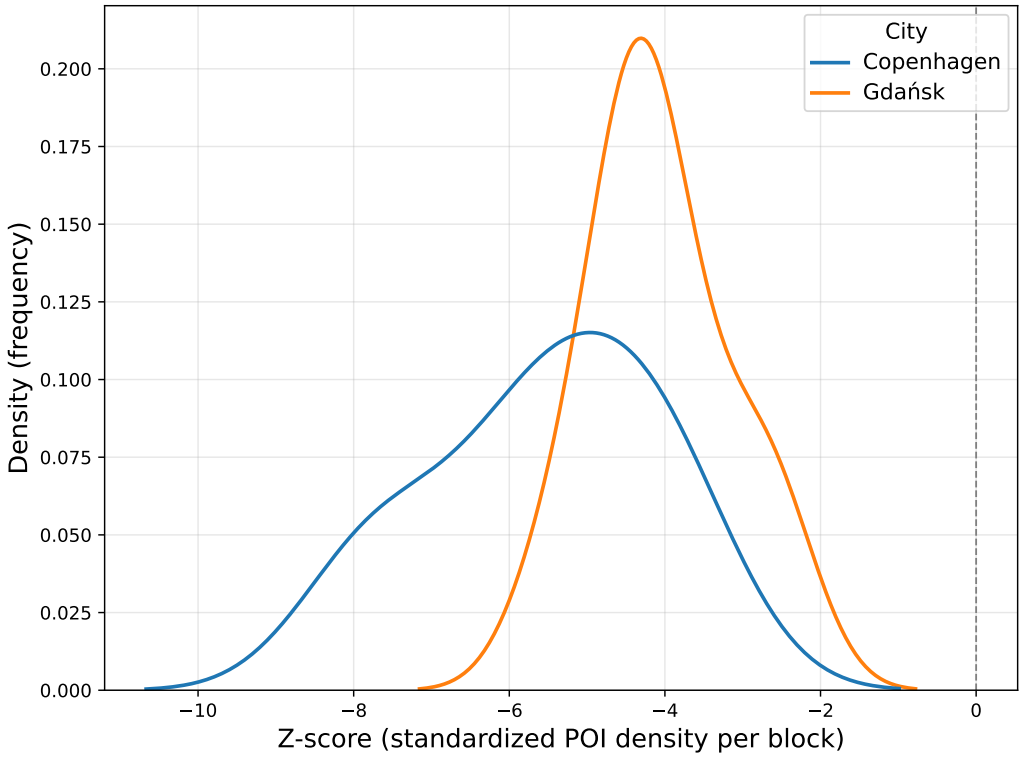


Figure 34: Distribution of Z-scores for Copenhagen and Gdańsk

7 Discussion and review

7.1 Interpretation of results

It may be surprising that, when applying the livability measure developed in this thesis, Gdańsk achieves a higher livability score than Copenhagen. This might suggest that this method is incorrect, as Copenhagen is among the most livable cities according to existing global livability indices [6].

However, the proposed livability measures behave consistently with its design. Higher scores are assigned to cities where amenities are more evenly distributed across the block network, while strong spatial concentration is penalized. With this definition, the result that Gdańsk achieves a higher livability score than Copenhagen follows directly from the observed distributions of amenities and network variance.

At the same time, the livability score captures only a limited set of urban characteristics. It reflects structural properties of the city, such as the spatial distribution of amenities and their organization within a block-based network. However, it does not measure the overall quality of life. Broader aspects of livability, including social, economic, and political conditions, are outside the scope of the measure, which is focused only on urban structure.

An important limitation is that the current livability score considers only the built structure of the city. The data that the measure is based on, does not include information about public transport systems such as metro, trains, trams, or buses. It also does not consider cycling infrastructure. As a result, the network captures spatial adjacency between blocks but does not reflect how people actually move through the city or access amenities. These factors can change the livability score a lot, but are not considered in the present analysis. This limitation is particularly relevant for Copenhagen, which is a city with well-developed cycling infrastructure and public transport.

From this perspective, the finding that Gdańsk achieves a higher livability score than Copenhagen is not necessarily contradictory. It does exactly what it is told to do. It evaluates the spatial structure of amenity distribution rather than overall urban quality. Most of the observed limitations come from the data and the resulting incomplete network representation, rather than the measure itself.

The proposed score should therefore be seen as a starting point for further development. Future work should focus on improving the network representation by incorporating information about mobility, such as public transport,

cycling infrastructure, and road categorization. Including these elements would allow the model to better represent how people access amenities in practice. It would lead to a more comprehensive assessment of urban livability.

7.2 Conceptual scope of the livability score

As discussed in the literature review (Section 2), livability is not a property of a place but emerges from the interaction between environmental characteristics and human well being, which can vary across individuals and contexts. This thesis acknowledges this understanding of livability, but the analysis is intentionally limited to the physical built environment and the spatial distribution of amenities.

As a result, the proposed livability score should not be interpreted as a universal measure of livability. Instead, it represents a structural indicator of potential accessibility and amenity organization, rather than overall quality of life. The score reflects properties of urban structure, not subjective perceptions or lived experience.

Subjective and social dimensions of livability, such as personal well-being, social interactions and individual mobility choices, are not included in the model. These aspects remain outside the scope of the thesis and represent directions for future research.

7.3 Block definition and cross-city comparability

An important issue in current analysis is the effect of the number of blocks on the livability score. Since each block corresponds to a node in the network, an increase in the number of blocks directly increases the number of nodes over which amenities can be distributed. As a consequence, the more blocks there are, the higher the score.

This effect was observed across all tested normalization variants of the measure. Regardless of how distances or aggregations were normalized, the livability score remained sensitive to the number of nodes in the network. This raises an important question about whether this is what we want - whether the livability score should depend on the number of blocks.

This issue becomes particularly important in comparisons between cities. The measure is intended to be used for different cities, so the measure should be made a way that allows for it. Cities might substantially differ in their

block structure: one city may consist of many small blocks, while another may be composed of fewer but larger blocks. In such cases, the livability scores may reflect those differences rather than livability.

The measure should allow for comparison of different cities. This leads to a broader question: what does it mean for a block in one city to be comparable to a block in another city? Without a clear definition of comparability, differences in block structure can dominate the results of the livability score.

There are several possible ways to interpret or address this issue. One option is to accept block structure as characteristic of a city, in which a higher number of blocks would be interpreted as a positive feature. As a result, cities with fewer, larger blocks would be considered less livable. Another approach would be to standardize blocks across cities, for example, by merging or dividing blocks. This way, the blocks would be more comparable, and then the livability scores computed using those blocks would be more comparable across cities.

Each approach leads to different interpretations of the livability score. The results presented in this thesis show that the score is sensitive to block definition and spatial granularity. Whether this sensitivity should be minimized or treated as a meaningful feature remains an important topic for future work.

7.4 Data and network limitations

The resulting livability scores are influenced by several limitations related to the used data. While the livability measure itself behaves consistently with its design, the quality and completeness of the data directly affect how urban structure and accessibility are represented.

A key limitation of the proposed network representation is the incomplete modeling of accessibility. As discussed in the interpretation of results, the network captures spatial adjacency between blocks but does not incorporate information about public transport systems, such as metro lines, train networks, or bus routes, nor does it represent cycling infrastructure. As a result, areas with spatially concentrated amenities may appear less accessible than they actually are. This limitation is particularly relevant for Copenhagen, where concentrated amenity areas are often well connected through public transport and cycling networks. In this context, the concentration of amenities reflected in the z-scores is likely due to the incomplete representation of mobility rather than the actual inaccessibility of services.

Integrating public transport systems as a logical transportation network, rather than relying solely on physical adjacency, would modify the network structure and directly affect the livability score. Such improvement would introduce additional connections (new edges) between blocks that are not physically adjacent. For example, two far away points in the city will be much closer to each other if there are two connected metro stations between them.

This would increase network connectivity and reduce effective distances between blocks, leading to lower GE values. At the same time, a denser and better-connected network would reduce differences in accessibility across blocks. The network variance would be decreased, pulling the corresponding z-scores toward zero. As a result, areas with well-developed transport infrastructure would appear more accessible than in the current adjacency-based model.

Related to this issue is the lack of additional information on network edges. In the current model, edges indicate adjacency between blocks, but they do not distinguish between different types of connections. Information about road type, transport mode or accessibility is not included. As a result, all edges are treated equally, regardless of their actual capacity to support movement, which limits the model's ability to represent how people actually move through the city.

Including traffic or transport capacity data would allow adding weights to the edges, according to their ability to move people. Edges with higher weights would correspond to higher capacity or faster travel conditions. These would reduce effective resistance in the network, leading to lower GE distances between blocks. This would result in a more realistic representation of accessibility without changing the overall network structure.

Further limitations come from the quality of the OSM data. Even though OSM provides detailed and openly accessible geospatial data, it also contains inaccuracies and inconsistencies. In this analysis, such issues were especially visible in the administrative boundary and water features data, which play a crucial role in block construction. Additionally, the POI dataset is messy, with many tags and limited standardization. All these issues affected block geometry and connectivity, and also reduced the reusability od the pipeline across cities.

Finally, the constructed block networks include disconnected components. These do not reflect true urban structure and fragmentation within it. This issue is probably caused by missing or incomplete data, such as gaps in road

geometries or boundary artifacts. To make the measure work, the analysis focuses on the largest connected component. Nevertheless, the presence of disconnected components highlights limitations of the underlying data.

Overall, these limitation show that the results are highly dependent on the available data and its quality. The livability measure is conceptually valid, but more detailed data would lead to a more accurate representation of urban accessibility.

7.5 Future work

The results and limitations discussed in previous parts show several directions for future improvements for this livability measure.

One of the key improvements would be to include information about mobility in the network. Especially incorporate public transport systems, such as metro lines, train networks, trams and bus routes. It would allow the model to better reflect how people access amenities in practice. Also, adding some additional information to network edges, such as road type, would improve the network by distinguishing between different types of connections.

Another area for improvement concerns the representation of points of interest. In the current implementation, POIs are grouped into nine broad categories. While this choice was intentional to group amenities into meaningful categories and to capture functional differences between them, the categories may be too broad. Future work could include adding subcategories to better capture functional diversity.

The definition of blocks is another area for future development. As shown in this thesis, the livability score is sensitive to the number and size of blocks. Future work could explore methods for standardizing block size across cities, for example, by merging or subdividing blocks to create comparable spatial units. Alternatively, some analysis could be performed to better understand how changes in block size affect the livability score and when this effect is meaningful.

References

[1] J. Gehl, *Cities for People*. Island Press, 2010.

[2] M. Pacione, "Urban liveability: A review," *Urban Geography*, vol. 11, no. 1, pp. 1–30, 1990.

[3] ——, "Urban environmental quality and human wellbeing - a social geographical perspective," *Landscape and Urban Planning*, vol. 65, pp. 19–30, 09 2003.

[4] I. van Kamp, K. Leidelmeijer, G. Marsman, and A. Hollander, "Urban environmental quality and human well-being: Towards a conceptual framework and demarcation of concepts; a literature study," *Landscape and Urban Planning*, vol. 65, pp. 5–18, 09 2003.

[5] C. Higgs, H. Badland, K. Simons, L. Knibbs, and B. Giles-Corti, "The urban liveability index: developing a policy-relevant urban liveability composite measure and evaluating associations with transport mode choice," *International Journal of Health Geographics*, vol. 18, 06 2019.

[6] T. E. I. Unit, "The global liveability index 2025," 2025. [Online]. Available: https://www.eiu.com/n/campaigns/global-liveability-index-2025/

[7] M. LLC, "Quality of living city ranking 2024," 2024. [Online]. Available: https://www.mercer.com/insights/total-rewards/talent-mobility-insights/quality-of-living-city-ranking/

[8] M. Ruth and R. S. Franklin, "Livability for all? conceptual limits and practical implications," *Applied Geography*, vol. 49, pp. 18–23, 2014.

[9] S. Porta, P. Crucitti, and V. Latora, "The network analysis of urban streets: A dual approach," *Physica A: Statistical Mechanics and its Applications*, 11 2004.

[10] M. Barthelemy, "Spatial network," *Physics Reports-review Section of Physics Letters*, vol. 499, 10 2010.

[11] C. Moreno, Z. Allam, D. Chabaud, C. Gall, and F. Pratlong, "Introducing the "15-minute city": Sustainability, resilience and place identity in

future post-pandemic cities," *Smart Cities*, vol. 4, no. 1, pp. 93–111, 2021.

[12] O. contributors, "About openstreetmap – openstreetmap wiki." [Online]. Available: https://wiki.openstreetmap.org/wiki/About_OpenStreetMap

[13] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *Haklay, M. and Weber, P. (2008) OpenStreetMap: user-generated street maps. IEEE Pervasive Computing, 7 (4). pp. 12-18. ISSN 15361268*, vol. 7, 10 2008.

[14] C. U. Press., "liveability," n.d., [Online; accessed 27 November 2025]. [Online]. Available: https://dictionary.cambridge.org/dictionary/english/liveability

[15] Y. Liang, D. D'Uva, A. Scandiffio, and A. Rolando, "The more walkable, the more livable? – can urban attractiveness improve urban vitality?" *Transportation Research Procedia*, vol. 60, pp. 322–329, 01 2022.

[16] H. A. Khalil, "Enhancing quality of life through strategic urban planning," *Sustainable Cities and Society*, vol. 5, pp. 77–86, 12 2012.

[17] A. Okulicz-Kozaryn, "A geography of european life satisfaction," *Social Indicators Research*, vol. 101, no. 3, pp. 435–445, 2011.

[18] K.-Y. Lee, "Factors influencing urban livability in seoul, korea: Urban environmental satisfaction and neighborhood relations," *Social Sciences*, vol. 10, p. 138, 04 2021.

[19] J. Jacobs, *The Death and Life of Great American Cities*. Random House, 1961.

[20] K. Zhang and D. Yan, "Enhancing the community environment in populous residential districts: Neighborhood amenities and residents' daily needs," *Sustainability*, vol. 15, p. 13255, 09 2023.

[21] H.-J. Jun and M. Hur, "The relationship between walkability and neighborhood social environment: The importance of physical and perceived walkability," *Applied Geography*, vol. 62, pp. 115–124, 08 2015.

[22] G. Foody, S. Fritz, C. Fonte, L. Bastin, A.-M. Olteanu Raimond, P. Mooney, L. See, V. Antoniou, H.-Y. Liu, M. Minghini, and R. Vatseva, *Mapping and the Citizen Sensor*. Ubiquity Press, 09 2017, pp. 1–12.

[23] A. Grinberger, M. Minghini, L. Juhász, G. Yeboah, and P. Mooney, "Osm science—the academic study of the openstreetmap project, data, contributors, community, and applications," *ISPRS International Journal of Geo-Information*, vol. 11, p. 230, 03 2022.

[24] J.-F. Girres and G. Touya, "Quality assessment of the french openstreetmap dataset," *T. GIS*, vol. 14, pp. 435–459, 08 2010.

[25] P. Crucitti, V. Latora, and S. Porta, "Centrality in network of urban streets," *Chaos (Woodbury, N.Y.)*, vol. 16, p. 015113, 04 2006.

[26] L. Bettencourt, *Introduction to Urban Science: Evidence and Theory of Cities as Complex Systems*. The MIT Press, 08 2021.

[27] N. Martino, C. Girling, and Y. Lu, "Urban form and livability: socioeconomic and built environment indicators," *Buildings and Cities*, vol. 2, pp. 220–243, 03 2021.

[28] OpenStreetMap contributors, "Openstreetmap," https://www.openstreetmap.org, 2025, data accessed: 01-12-2025.

[29] S. Denmark, "Befolkningstal," https://www.dst.dk/da/Statistik/emner/borgere/befolkning/befolkningstal, 2025, accessed: 01-12-2025.

[30] ——, "Statbank denmark," https://www.statistikbanken.dk/, 2025, accessed: 01-12-2025.

[31] S. Gössling, "Urban transport transitions: Copenhagen, city of cyclists," *Journal of Transport Geography*, vol. 33, p. 196–206, 12 2013.

[32] C. of Gdańsk, "Gdańsk w liczbach," https://www.gdansk.pl/gdansk-w-liczbach, 2025, accessed: 01-12-2025.

[33] Statistics Poland (GUS), "Ludność: stan i struktura ludności oraz ruch naturalny w przekroju terytorialnym w 2025 r." https://stat.gov.pl/obszary-tematyczne/ludnosc/ludnosc/

ludnosc-stan-i-struktura-ludnosci-oraz-ruch-naturalny-w-przekroju-terytorialnym-w-2025-r-st
6,39.html, 2025, accessed: 01-12-2025.

[34] L. Bugalski, *Historic centre of Gdansk as a unique example of postwar socialist city creation // Historyczne srodmiescie Gdanska jako unikalny przyklad powojennej kreacji miasta socjalistycznego.* International Society of City and Regional Planners ISOCARP, 01 2015, pp. 221–227.

[35] OpenStreetMap contributors, "Admin_level — openstreetmap wiki," https://wiki.openstreetmap.org/wiki/Template:Admin_level, 2025, accessed: 01-12-2025.

[36] D. D. Polsby and R. D. Popper, "The third criterion: Compactness as a procedural safeguard against gerrymandering," *Yale Law & Policy Review*, vol. 3, pp. 301–353, 1981.

[37] M. Coscia, "The atlas for the aspiring network scientist," 2025. [Online]. Available: https://arxiv.org/abs/2101.00863

A POI categorization

This appendix describes the categorization of Points of Interest (POIs) based on OpenStreetMap (OSM) tags. The categorization was primarily based on values in the `amenity` column. However, many POIs did not contain an `amenity` value and were instead described using other OSM tag columns. Additional columns were reviewed and mapped to POI categories. The complete set of reviewed tags and categorization rules is listed below.

A.1 Categorization based on the `amenity` column

Values from the `amenity` column were reviewed first and grouped into nine POI categories. Tags not representing access to everyday amenities were excluded from the final POI dataset.

Food

- bar
- restaurant
- cafe
- ice_cream
- fast_food
- pub
- hookah_lounge
- food_court
- internet_cafe
- food_sharing
- pastry
- community_centre;cafe
- canteen
- biergarten

Infrastructure & transport

- parking
- parking_space
- bicycle_parking
- motorcycle_parking
- charging_station
- taxi
- ferry_terminal
- car_rental
- car_wash
- bicycle_rental
- bus_station
- car_sharing
- scooter_parking
- traffic_park
- motorcycle_rental
- kick-scooter_parking

Education

- school
- kindergarten
- childcare
- university
- college
- language_school
- research_institute
- music_school
- prep_school

Culture & leisure

- social_facility
- events_venue
- theatre
- library
- cinema
- gambling
- music_venue
- arts_centre
- casino
- nightclub
- stripclub
- gallery
- monastery
- dojo
- dive_centre
- exhibition_centre
- planetarium
- public_bath
- festival_grounds
- climbing_wall
- dancing_school
- surf_school
- convent

Public services

- place_of_worship
- community_centre
- bank
- post_office
- police
- courthouse
- fire_station
- social_centre
- conference_centre
- funeral_hall
- crematorium
- townhall
- parliament
- lost_property_office
- animal_shelter
- local government unit
- ranger_station

Healthcare

- pharmacy
- clinic
- dentist
- doctors
- veterinary
- hospital
- nursing_home
- fysioterapy
- healthcare

Retail

- marketplace

Green spaces

- playground

Other daily utilities

- recycling
- toilets
- drinking_water
- atm
- fuel
- parcel_locker
- bicycle_repair_station
- coworking_space
- bureau_de_change

- luggage_locker
- locker
- left_luggage
- self_storage

Discarded amenity tags

The following `amenity` values were observed in the dataset but were excluded from the analysis because they do not represent access to everyday urban amenities.

Discarded tags

- waste_basket
- bench
- parking_entrance
- post_box
- vending_machine
- shelter
- waste_disposal
- public_bookcase
- fountain
- bbq
- compressed_air
- driving_school
- student_accommodation
- studio
- prison
- shower
- clock
- boat_rental
- smoking_area
- photo_booth
- ticket_validator
- table
- stage
- grave_yard
- publisher
- dressng_room
- vacuum_cleaner
- money_tranfer
- sentry_box
- loading_dock
- dormitory
- del
- kayak_storage
- reception_desk
- bik
- hammock
- give_box
- traffic_school
- vehicle_inspection
- printer
- electrical
- binoculars
- wine_storage
- games
- photo
- erotic
- kictchen
- user_defined
- bath
- elevator
- outdoor_seating
- water_point
- watering_place
- waterpoint
- waste_transfer_station
- baggage_reclaim

- sanitary_dump_station
- greengrocer
- waxing
- device_charging_station
- security_control
- payment_centre
- lifeboat
- baby_hatch
- grit_bin
- animal_training
- driver_training
- library_dropoff
- dog_parking
- hitching_post
- reception
- weighbridge
- hunting_stand
- trolley_bag
- lounger

A.2 Categorization using non-`amenity` columns

Approximately 5,000 POIs did not contain any value in the `amenity` column. These POIs were described using other OSM tag columns, often with boolean or category-specific values (e.g. `atm=yes`). To avoid discarding these POIs, additional columns were systematically reviewed and mapped to the same POI categories.

A subset of POIs did not contain any value in the `amenity` column. These POIs were described using other OSM tag columns. For each POI without value in the `amenity` column, selected tag columns are checked and if a column contains a meaningful value (i.e., not missing and not a negative/empty), the POI is assigned to the category associated with that column.

In this step, a column value was treated as *present* if it was not `NaN`, not `none`, and not `no`. Boolean tags such as `atm=yes` were therefore interpreted as indicating the presence of that POI type.

If multiple tag columns were present for the same POI, the category was assigned using the first matching column according the order of columns in the mapping dictionary.

Column-based mapping used when `amenity` is missing

- `office` → Public services
- `post_office` → Public services
- `charity` → Public services
- `police` → Public services
- `attraction` → Culture & leisure
- `camp_site` → Culture & leisure

- `information` → Culture & leisure
- `museum` → Culture & leisure
- `tourism` → Culture & leisure
- `caravan_site` → Culture & leisure
- `zoo` → Culture & leisure
- `swimming_pool` → Culture & leisure
- `bar` → Food
- `tea` → Food
- `pastry` → Food
- `restaurant` → Food
- `books` → Retail
- `butcher` → Retail
- `clothes` → Retail
- `confectionery` → Retail
- `craft` → Retail
- `furniture` → Retail
- `gift` → Retail
- `massage` → Retail
- `model` → Retail
- `music` → Retail
- `outdoor` → Retail
- `pet` → Retail
- `second_hand` → Retail
- `wholesale` → Retail
- `shop` → Retail
- `shoes` → Retail
- `medical_supply` → Retail
- `bicycle_rental` → Infrastructure & transport
- `green_spaces` → Green & spaces

Green space POIs tags were not taken directly from the OSM dataset. Instead, green areas (e.g., parks and recreation grounds) were extracted from OSM landuse/leisure features and added to the POI dataset in as a separate column, as a new tag (`green_spaces`). This additional tag was created manually as part of the preprocessing and then treated as an input to the same categorization.

B Distribution of frequent `amenity` tags

This appendix presents the distribution of the most frequent OSM values in the `amenity` column for Copenhaggen and Gdańsk. The figure highlights that many of the most common tags refer to object that are not directly related to everyday accessibility, for example street furniture or technical infrastructure. This explains why the categorization coverage reported in Section 4.3.6 is not expected to be complete.

Amenity tag	**Count**
None	5495
bench	5042
parking	3337
bicycle_parking	3047
parking_space	2838
waste_basket	1795
fast_food	1360
cafe	858
restaurant	797
recycling	673
charging_station	649
bar	524
toilets	263
post_box	256
parking_entrance	249
social_facility	236
school	169
vending_machine	166
kindergarten	157
place_of_worship	154

Table 13: Top 20 most frequent values in the OpenStreetMap `amenity` column for Copenhagen.

While the specific frequency distributions differ between the two cities, both tables show that many of the most frequent tags represent objects such as benches, waste bins, or other forms of street furniture. They were intentionally excluded from the POI categorization as they do not represent access to everyday urban amenities.

Amenity tag	**Count**
bench	8002
None	6108
parking	4240
waste_basket	3075
bicycle_parking	1656
parking_entrance	1121
parking_space	1073
parcel_locker	703
recycling	623
shelter	620
restaurant	562
waste_disposal	559
vending_machine	462
bicycle_rental	419
fast_food	315
atm	285
kick-scooter_parking	269
kindergarten	178
cafe	164
toilets	156

Table 14: Top 20 most frequent values in the OpenStreetMap `amenity` column for Gdańsk.

C Block area percentiles and small-block filtering

This appendix provides additional visualizations of urban blocks grouped by area percentiles for Copenhagen and Gdańsk. The figures complement the analysis presented in Section 4.4 and illustrate how block size varies across the distribution.

C.1 Copenhagen

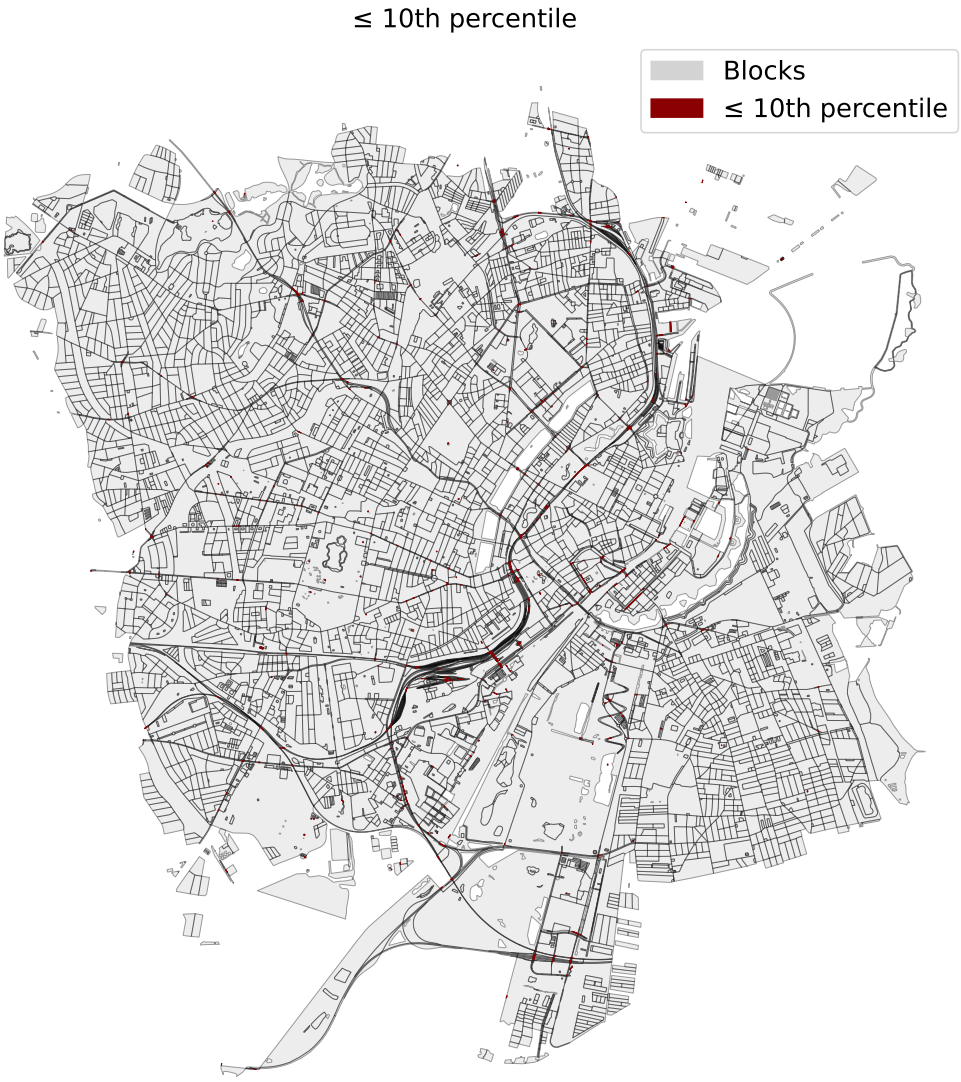


Figure 35: Urban blocks in Copenhagen with areas in the lowest 10% of the block area distribution, highlighted over the full block layout.

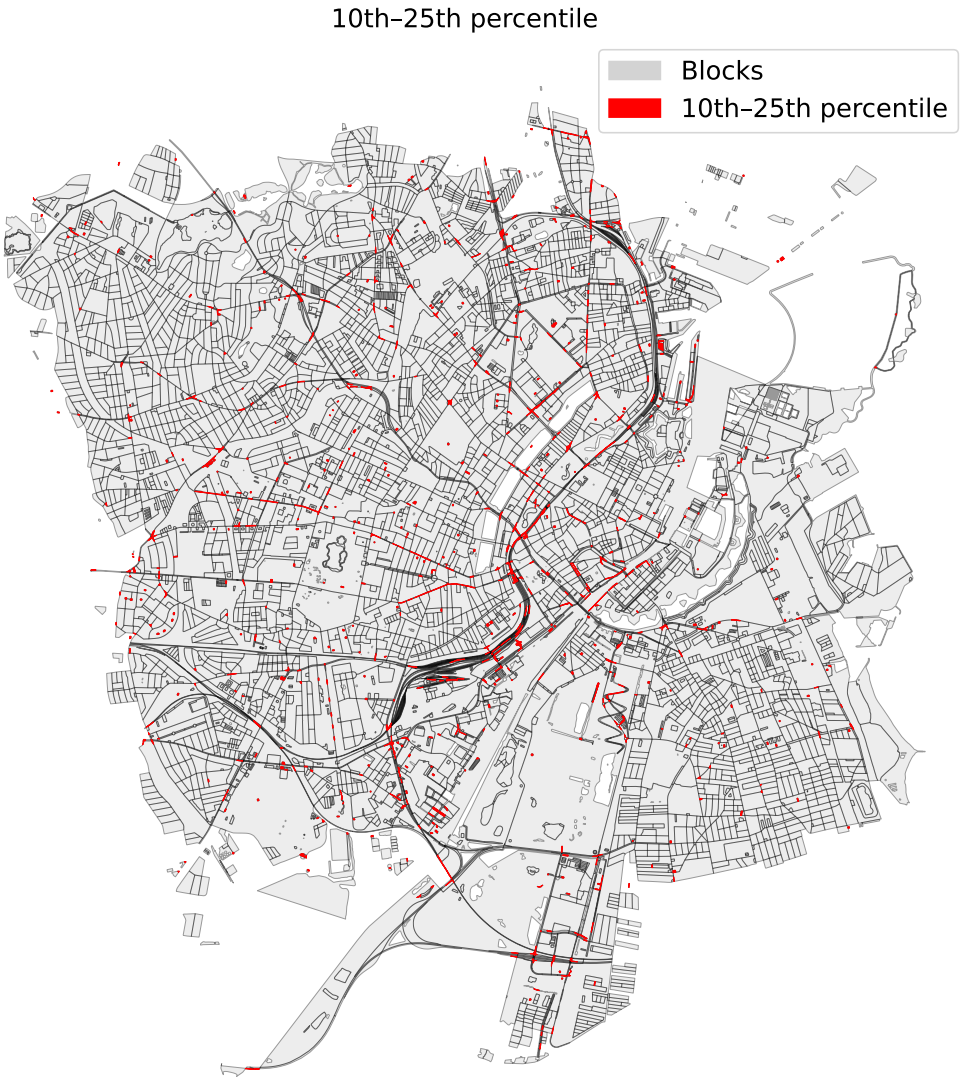


Figure 36: Urban blocks in Copenhagen with areas between the 10th and 25th percentiles of the block area distribution.

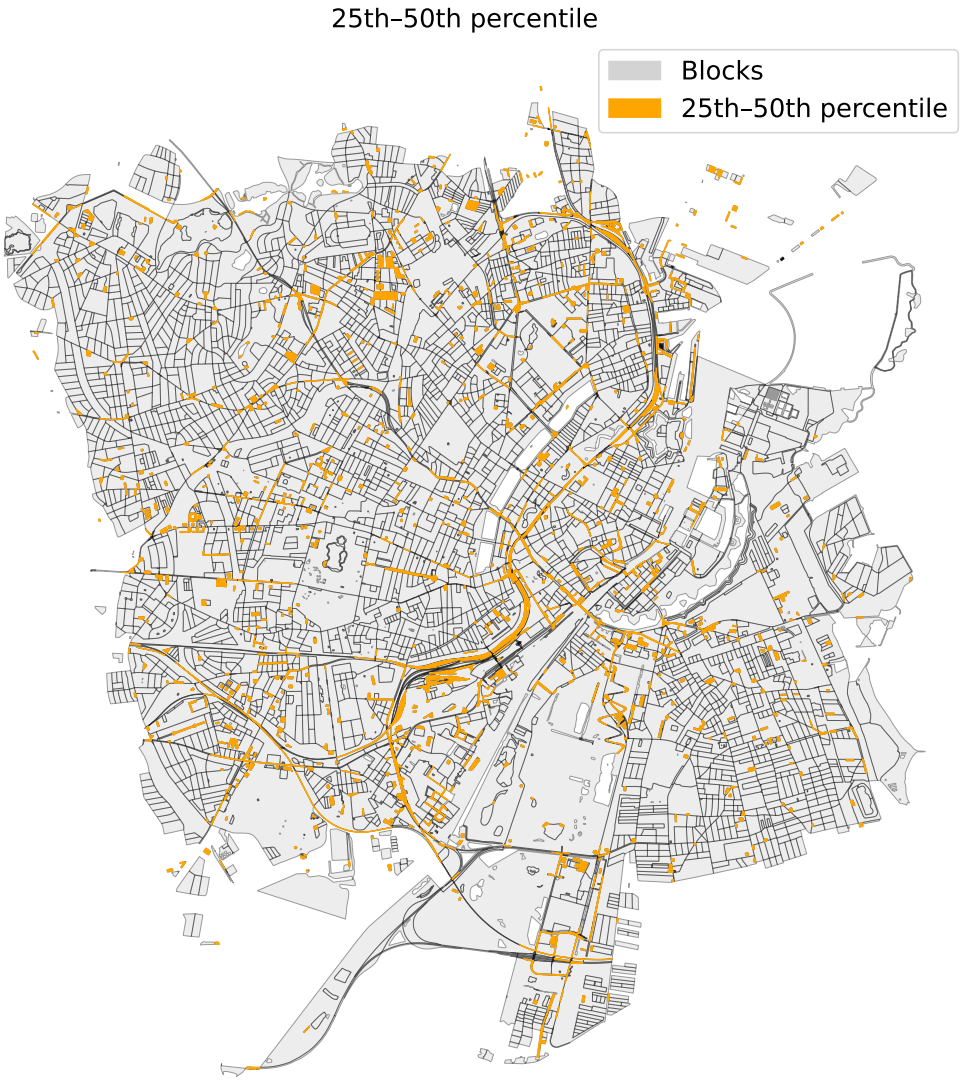


Figure 37: Urban blocks in Copenhagen with areas between the 25th and 50th percentiles of the block area distribution.

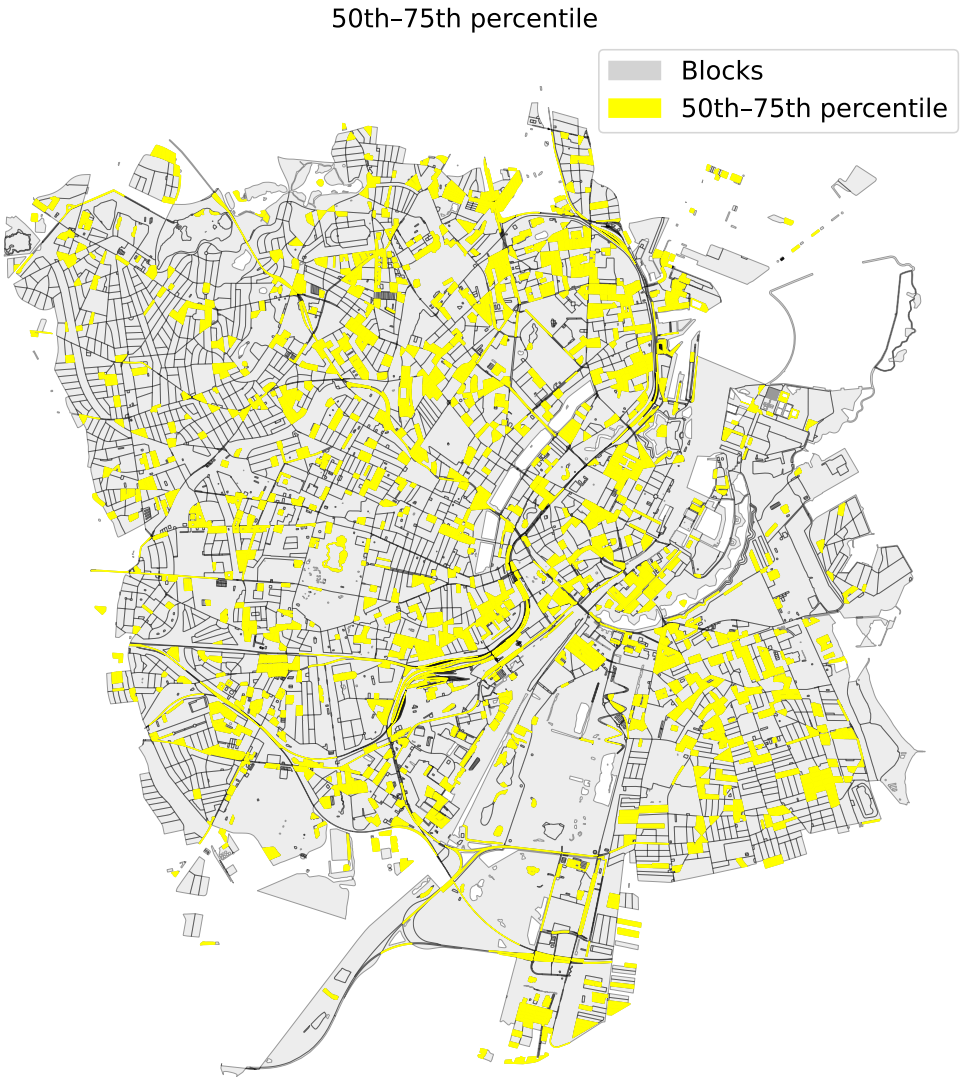


Figure 38: Urban blocks in Copenhagen with areas between the 50th and 75th percentiles of the block area distribution.

C.2 Gdańsk

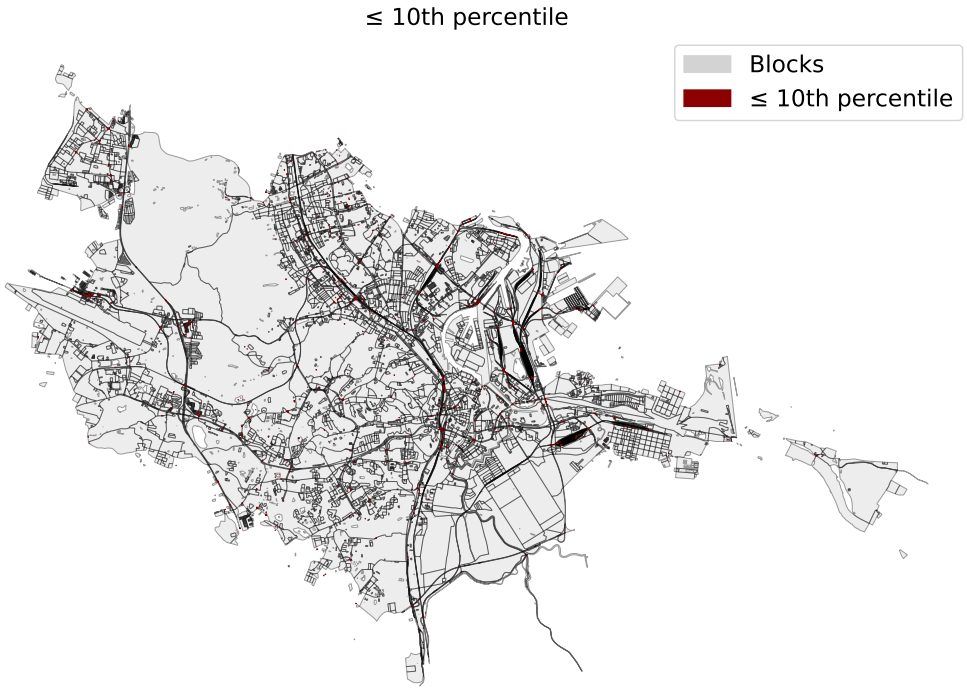


Figure 39: Urban blocks in Gdańsk with areas in the lowest 10% of the block area distribution, highlighted over the full block layout.

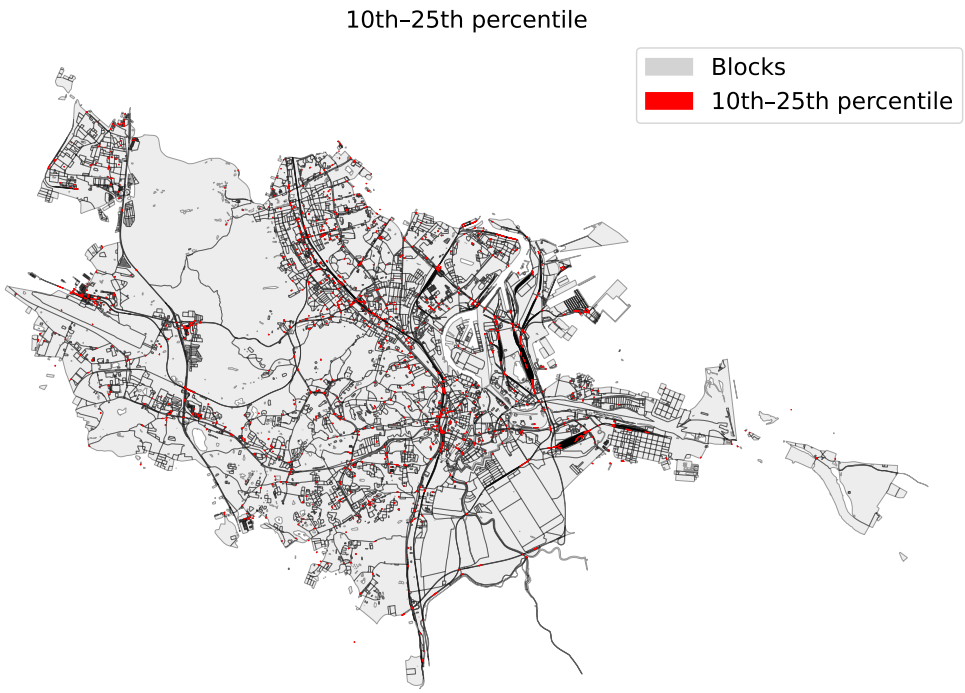


Figure 40: Urban blocks in Gdańsk with areas between the 10th and 25th percentiles of the block area distribution.

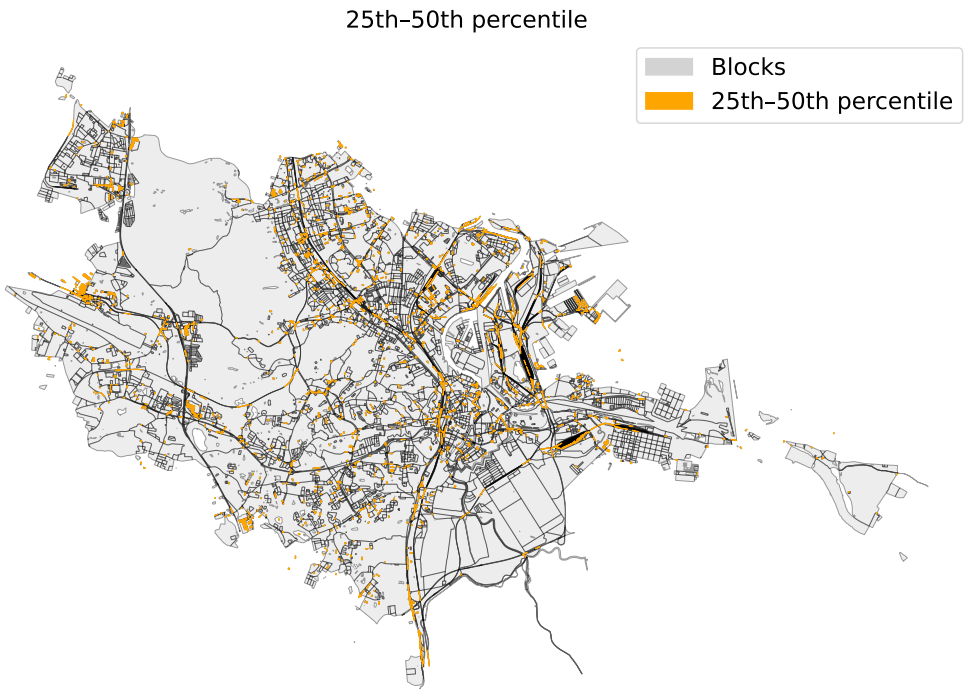


Figure 41: Urban blocks in Gdańsk with areas between the 25th and 50th percentiles of the block area distribution.

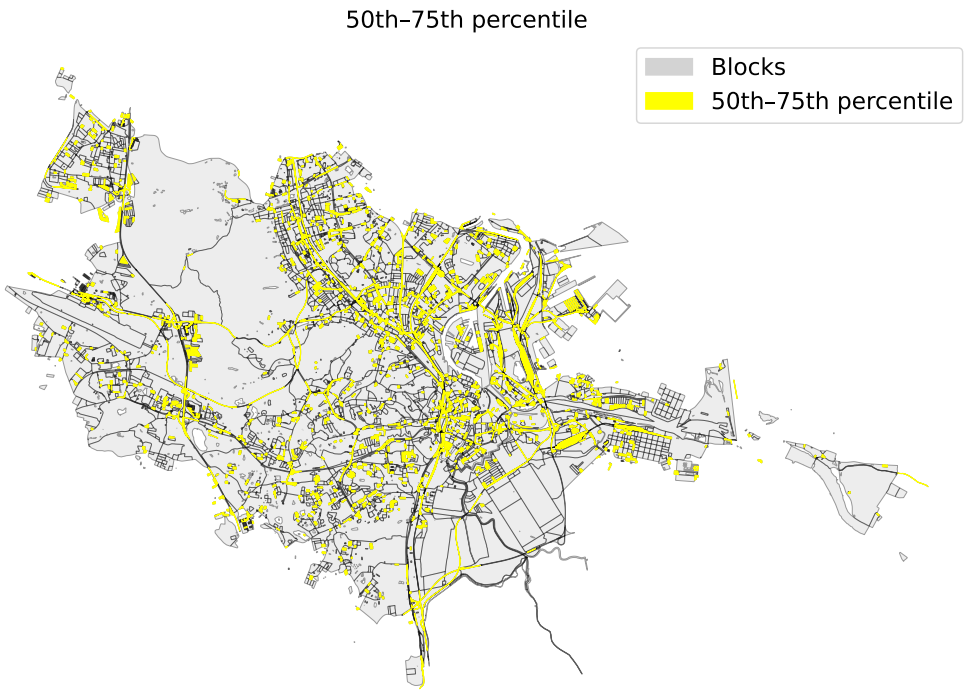


Figure 42: Urban blocks in Gdańsk with areas between the 50th and 75th percentiles of the block area distribution.

D Testing results for livability score normalization variants

This appendix presents supplementary testing results for the livability score. The figures illustrate the behavior of the score under different normalization variants across varying graph sizes and POI densities. Test results are discussed in Section 5.3. All the figures are provided here for completeness.

Unnormalized livability score

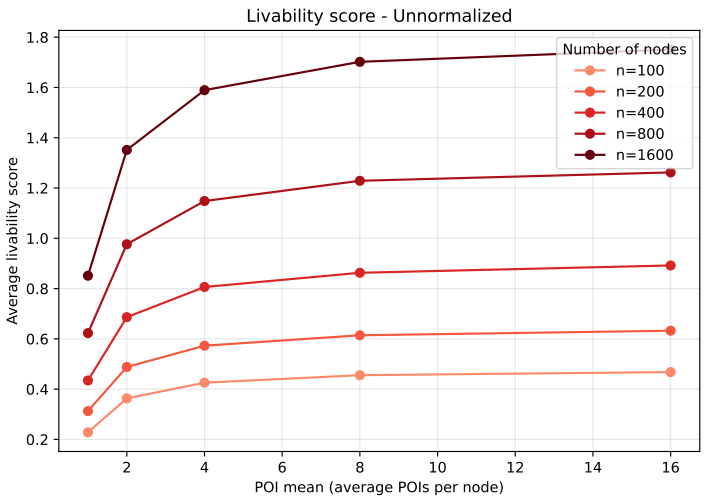


Figure 43: Livability score using unnormalized variant.

Average-normalized livability score

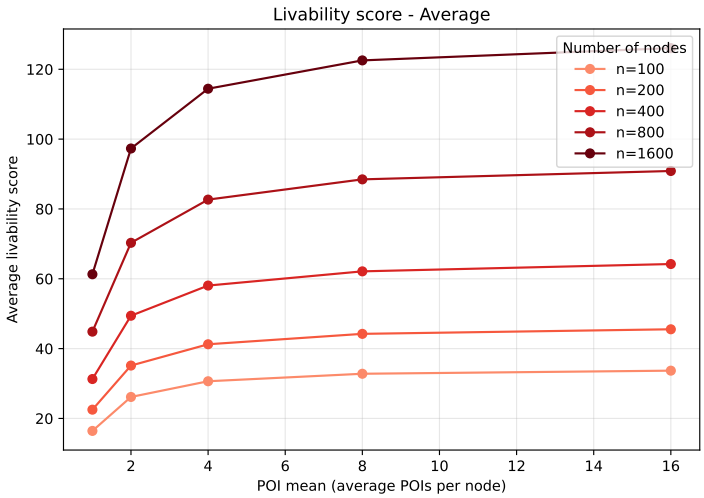


Figure 44: Livability score using average variant.

GE-normalized livability score

GE total normalization

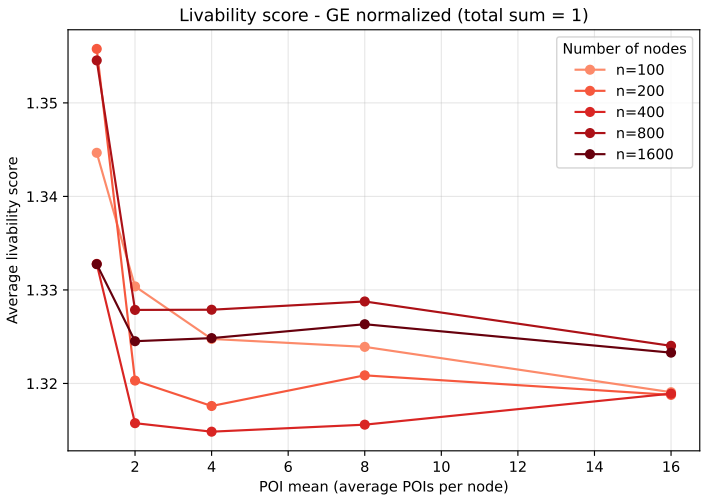


Figure 45: Livability score using GE total normalization.

GE row-wise normalization

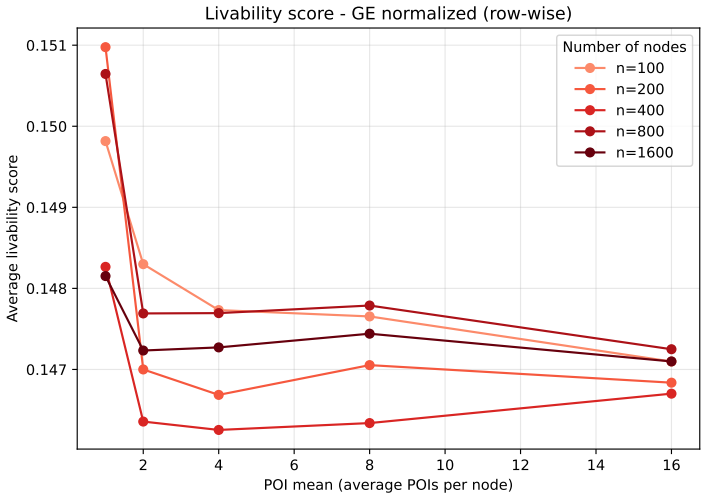


Figure 46: Livability score using GE row-wise normalization.

Log-transformed livability score

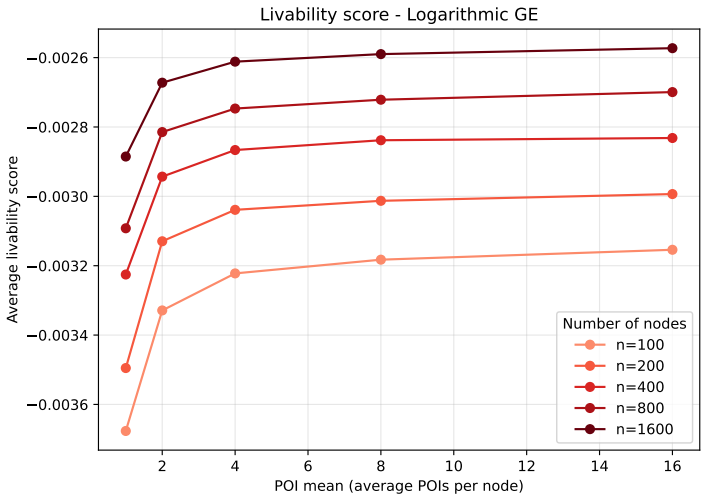


Figure 47: Livability score using logarithmic variant.

Weighted-graph livability score

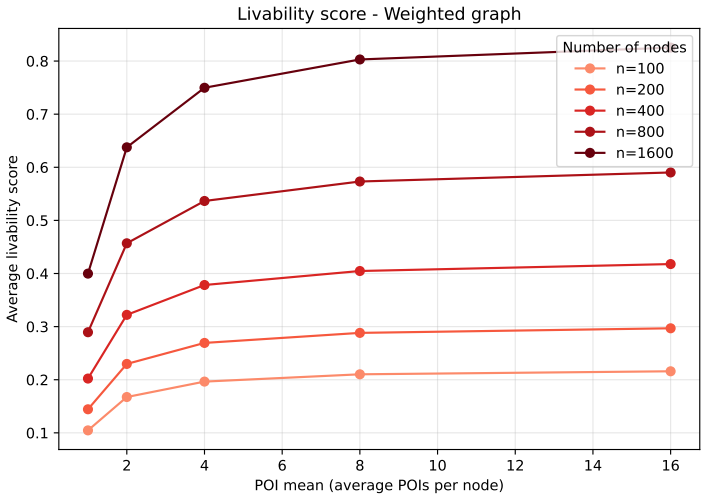


Figure 48: Livability score using weighted-graph variant.