IT UNIVERSITY OF COPENHAGEN

Data-driven bicycle networks

Kristof Gasior

Master thesis
Software Design
Kristof Gasior

March 31, 2023
Supervisor: Maria Sinziiana Astefanoaei

ABSTRACT

This thesis aims to explore a data-driven approach for growing urban bicycle networks in the city of Copenhagen, taking into account network structure, bicycle traffic flow, and population density. The study will utilize existing street and bicycle network data, as well as traffic flow data and population density data. The proposed approach will include weighted distance calculations for defining edge weights in a network and propose data-aware betweenness-centrality measures. From these, bicycle networks can be built that are both accessible to the public and connect the areas with the highest traffic flow. The results will be compared to existing research on bicycle network growth, and the limitations of growing data-driven urban bicycle networks will be discussed. The goal of this study is to provide urban planners with computational tools for building more sustainable urban transport systems, reducing the reliance on vehicle-centric transport systems, and in general, informing decision-making about bike lane infrastructure.

CONTENTS

1 INTRODUCTION

Climate change is a growing priority on political agendas worldwide, and one of the goals is for cities to find sustainable solutions for transport systems. Cycling is one way of reducing reliance on already established inefficient vehicle-centric transport systems, providing a sustainable solution to most intra-urban trips (Banister (2005); Nieuwenhuijsen et al. (2016)). The transition to cycling will ideally reduce the level of co2, the number of traffic injuries, and the demand for parking spaces, generating more space for people to enjoy the city Miljøforvaltningen (2021). However, few cities have well-connected bicycle networks, and prevailing bicycle network development is a very slow process, as many cities have a deep-rooted complexity with car dependence Mattioli et al. (2020). Computational methods from urban network science can be utilized to synthetically grow bicycle networks to inform decision-making about bike-lane infrastructure Szell et al. (2022). These methods are developed by combining computer science and data analysis with urban planning to develop tools for designing cities. The computational methods involve graph theory, optimization algorithms, and machine learning. To synthetically grow networks, spatial information (street network structure) and place-specific empirical data can be used. In academic literature related to this, the approaches that are data-driven tend to be mainly based on survey data (Lovelace et al. (2015); Larsen et al. (2013)), and not sufficiently evidence based. The approaches concerned with growing networks based on spatial information tend to be utilizing these methods only on single cities at a time, such as Seattle Lowry and Loh (2016), Montreal Boisjoly et al. (2019), and Berlin Palominos and Smith (2020). Therefore, a combination of these types of data has recently been requested.

Szell et al. (2022) address the need for a more global analysis and explore the topological limitations of growing urban bicycle infrastructure based on spatial information obtained from OSM (OpenStreetMap). Olmos et al. (2020) and Folco et al. (2022) grow bicycle infrastructure while also including data-driven approaches; computational generalizations are made to account for traffic data spareness, and optimization algorithms are utilized to define customized edge weight measures in the graphs from which bicycle networks are grown. These findings will provide the basis for the analysis proposed in this thesis

Networks can be grown with different growth strategies. In cities such as Copenhagen, with well-developed bicycle infrastructure, growing networks with betweenness-centrality (see section 2.1.4 for definition) show high overlaps with existing bicycle and bikeable infrastructure Szell et al. (2022). Betweenness-centrality can be a useful tool to identify key locations in transportation networks that may have the potential for high traffic flow Kazerani and Winter (2009). That being said, it cannot be used as a direct proxy for actual traffic flow when only based on network topology, and other factors must be considered. Vybornova (2021) defines a customized betweenness centrality measure based on network topology and actual empirical observations of bicycle flow, allowing for identifying important missing links in the bicycle network in Copenhagen

My research extends the growth framework developed by Szell et al. (2022), to which I will contribute with a data-driven approach. I will be embedding population density data and bicycle count data from Copenhagen in a metric space that defines the street network of Copenhagen. Some of the data is very sparse and I will therefore make generalizations to account for data spareness. I will grow and

analyze all bicycle networks with betweenness-centrality as a growth strategy, with customized weighted distance measures for graph edges. These are defined for each road section as a trade-off between the length, the population density, and the generalized bicycle flow. The grown networks will be evaluated and compared to the baseline (the network grown by Szell et al. (2022)) based on several evaluation metrics (see section 4.2.1 for a description of chosen metrics). I will contribute to the analysis by defining a metric to measure the accessibility of these networks to the people of Copenhagen and explore how population density can be used as a proxy for traffic flow estimation (see section 5 for analysis results). First, I will describe the background of this research; how data-driven and data-agnostic methods are used for synthetically growing bicycle infrastructure. This includes studying customized distance measures in graphs, centrality measures, and edge weight completion. Then follows a description of the process of acquiring and preprocessing place-specific empirical data and how this is used to inform the street network of Copenhagen about population density and bicycle traffic flow. Finally, I will describe my network growth model for optimized bicycle networks, as well as the analysis and evaluation of the grown networks.

2 BACKGROUND

2.1 LITERATURE REVIEW: A SURVEY OF HOW TO SYNTHETICALLY GROW BICYCLE INFRASTRUCTURE

2.1.1 Introduction

In this section, I will describe the challenges with growing urban infrastructure and how computational approaches can help optimize these processes. I have chosen a set of papers describing data-driven and data-agnostic methods for synthetically growing urban infrastructure. I will explain these methods, discuss the different approaches, and reflect on how to develop an optimized bicycle infrastructure in Copenhagen. My research mainly extends the findings of Szell et al. (2022), exploring the topological limitations of growing urban bicycle infrastructure, to which I will contribute with a data-driven approach, analyzing how population density can be used as a proxy for generalized traffic flow.

2.1.2 Towards climate-friendly urban infrastructure in cities worldwide

In recent decades, there has been a focus on finding ways of meeting the needs of humans, which is not at the expense of future generations Commission (1987). One of these ways is creating sustainable forms of transportation. Cycling is one of the most promising and widely adopted forms of sustainable transportation, and many governments and organizations worldwide are investing in bicycle infrastructure and bicycle-friendly roads. Copenhagen is among one of the most bicycle-friendly cities in the world, with very well-developed bicycle infrastructure and various bicycle-sharing programs, as well as a strong cultural leaning towards this form of transport, resulting in over 60% of all residents in Copenhagen biking to work cycling magazine (2022).

Olmos et al. (2020) address a global paradigm shift towards non-motorized transport alternatives in cities worldwide, from which a boost in cycling rates can be observed. Because of this boost, Olmos et al. (2020) encourage the need for novel data sources, methods, and tools to help identify and choose the best locations to build optimal cycling infrastructure. Olmos et al. (2020) emphasize that already existing urban planning, in this case in Bogotá, relies on mobility surveys, opinion polls, and bicycle traffic counts. This process is usually very slow, and data quickly get outdated. Like Copenhagen, Bogotá is among these cities where large investments in bicycle infrastructure have been made, and today approx, 5% of all trips in the city of Bogotá are made on a bicycle.

Folco et al. (2022) point to micro-mobility, such as cycling, as one of the most economical and promising solutions to the climate crisis. Folco et al. (2022) point out that urban planners need automated assistance for developing safe and climate-friendly infrastructure. The studies of Folco et al. (2022) give us insight into how data of micro-mobility trips and crashes can shape and automatize such network planning processes in the city of Turin.

Szell et al. (2022) are motivated by climate change being a growing priority on political agendas, and making cities invest in the right strategy for growing bicycle networks, is essential. Szell et al. (2022) point out bicycling as a sustainable form of intra-urban transport, which can help reduce the reliance on inefficient vehicle-centric transport systems in cities worldwide.

I choose these as my primary papers for studying, as all authors propose different methods for using network science to synthetically grow bicycle infrastructure. They aim to help cities build sustainable transport systems and combat climate change.

2.1.3 Fields of study in network science concerned with growing urban bicycle infrastructure

From network science studies on developing urban bicycle infrastructure, data-driven and data-agnostic methods are used and sometimes combined.

Data-driven methods rely on empirical data to inform analysis and development, such as gathering and analyzing data on real-world phenomena. On the other hand, data-agnostic methods do not rely on this empirical data. Instead, they rely mainly on assumptions about how a particular system behaves, e.g., that cyclists always try to find the shortest path to their destination.

Data-driven methods

Folco et al. (2022) propose a data-driven approach for optimizing the placement of electric scooters and bicycles in a given city. The goal is to increase the accessibility of vehicles while also maintainging safety. The method presented in the paper starts by collecting data from a real-world deployment of electric scooters in Turin over several months. This data includes information on the scooters location, the duration of their use, and finally, if taken place, any accident the scooter was in. This data is used to build a model of demand for electric scooters in different parts of the city. The model considers spatial and temporal factors influencing vehicle demand, such as population density and time of day. Folco et al. (2022) use the data to identify the locations where collisions are most likely to occur and propose a set of safety measures, such as reducing the speed limit in certain areas and adding bike lanes.

Olmos et al. (2020) state that the use of personal tracking devices and network science for understanding urban dynamics has already been seen in recent years, and coinciding with this, Olmos et al. (2020) propose a data-analysis framework that integrates trajectories from biking applications with already validated OD's (origin-destination matrices), to prioritize bicycle paths at an urban scale. The framework proposed by Olmos et al. (2020) is reproducible for many cities if the following data requirements are fulfilled:

- socio–referenced census population.
- socio-economic information.
- bicycle trajectories.
- empirical estimates.
- bicycle facilities.
- road network infrastructure.

Data-agnostic methods

Szell et al. (2022) explore strategies for growing urban bicycle networks. Szell et al. (2022) seek to answer whether there are fundamental topological limitations to developing bicycle networks and whether optimal growth policies can be replicated in several cities worldwide. Szell et al. (2022) initially point to several locally based data-driven approaches of analyzing bicycle networks, such as the analysis of Olmos et al. (2020), underlining how invaluable these studies are, but also argues for the need for a more global analysis.

Szell et al. (2022) propose an analysis exploring the fundamental topological limitations of bicycle network development. This analysis uses quantitative computational methods of urban data science and is concerned with growing synthetic bicycle networks from scratch. The study of Szell et al. (2022) is open source, made with minimum ingredients, and applicable for even a data-scarce environment. The analysis consists of only two ingredients, making it more easily reproducible than the analysis of Olmos et al. (2020):

- Manually sampled street networks of 62 diverse cities. This data is downloaded and processed from OpenStreetMap and embedded in a metric space.
- An arbitrary set of seed points representing points of interest and implemented as nodes in planar graphs

My analysis aims to draw these concepts from data-driven and data-agnostic studies to generate optimized bicycle networks in Copenhagen.

2.1.4 Growing networks in OSMnx

Folco et al. (2022) and Szell et al. (2022) use the OSMnx Boeing (2017) library for their analysis, and parts of the framework of Folco et al. (2022) is based on the framework Szell (2022) of Szell et al. (2022).

OSMnx is a Python package for working with OpenStreetMap (OSM), a crowdsourced platform providing free global map data, also utilized by Olmos et al. (2020). It allows for downloading and analyzing OSM data (data about streets, paths, sidewalks, bike lanes, etc.) and uses graph theory to represent this data as edges (street segments) and nodes (intersections/endpoints of streets); Custom queries are used to download CSV files and encode these OSM data structures; Polygons (a filled area constructed of several coordinates forming an outer ring) are extracted for a chosen city, and transformed into graphs. Cities are hereby represented as networks on which various network statistics can be calculated, such as shortest paths and centrality measures.

This library is built on top of more common Python packages such as Pandas, Matplotlib, and NetworkX, often used in other GIS (Geographic Information Systems) and data science applications.

Triangulation

Folco et al. (2022) and Szell et al. (2022) use the OSMnx library to download and analyze OSM data, and they compose their graphs from both already-established bike lanes and lanes that could be potential bike lanes (streets with a speed limitation of 30 km/h or below). In these graphs, they use triangulation to generate the basis for potential network bicycle routes.

Szell et al. (2022) use greedy triangulation to build a planar network. For this, a set of seed points is needed, and these seed points/nodes in the graph are ordered by route distance and connected stepwise, with no crossing links. In the case of Szell et al. (2022), seed points are generated from railway stations/halts and bus stops, and in the case of Folco et al. (2022), seed points are based on population density counts. Greedy triangulation is argued to be an easily computational and satisfying solution for investors and travelers and is therefore chosen for this analysis.

Network growth strategies

On top of this planar network, Szell et al. (2022) explore different growth strategies to order the greedy triangulation to synthetically grow bicycle networks. All growth strategies satisfy different network metrics, and both existing and newly generated networks are analyzed to compare the chosen metrics. From these metrics, the network quality can be analyzed. Szell et al. (2022) point out that these strategies

resemble reality, as several pronounced overlaps of synthetically grown networks are found with well-developed existing bicycle networks. The strategies are:

- Betweenness Freeman (1977): Connecting points of interest through the most important streets. It is a path-based measure, computing the fraction of paths that pass through a given node i

$$C_B(i) = \frac{1}{N} \sum_{s \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

  σ_{st} is the amount of shortest paths going from nodes s to t, and $\sigma_{st}(i)$ is the amount of these paths that go through i.

- Closeness: Locally connecting points of interest based on closeness. This is a measure of the length of all the shortest paths from a node i to all the rest of the nodes in the network.

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

- Random: Randomly connecting points of interest.

Szell et al. (2022) choose to use 40 growth quantiles, making it possible to analyze the network at each of the 40 growth stages. Figure 1 shows the stepwise growth process of a bicycle network from the analysis of Szell et al. (2022). Here we see 1) seed points are generated, 2) greedy triangulation is performed from these seed points, 3) the growth is illustrated for five chosen quantiles with each growth strategy, 4) the routing of bicycle networks is illustrated for five chosen quantiles with each growth strategy.

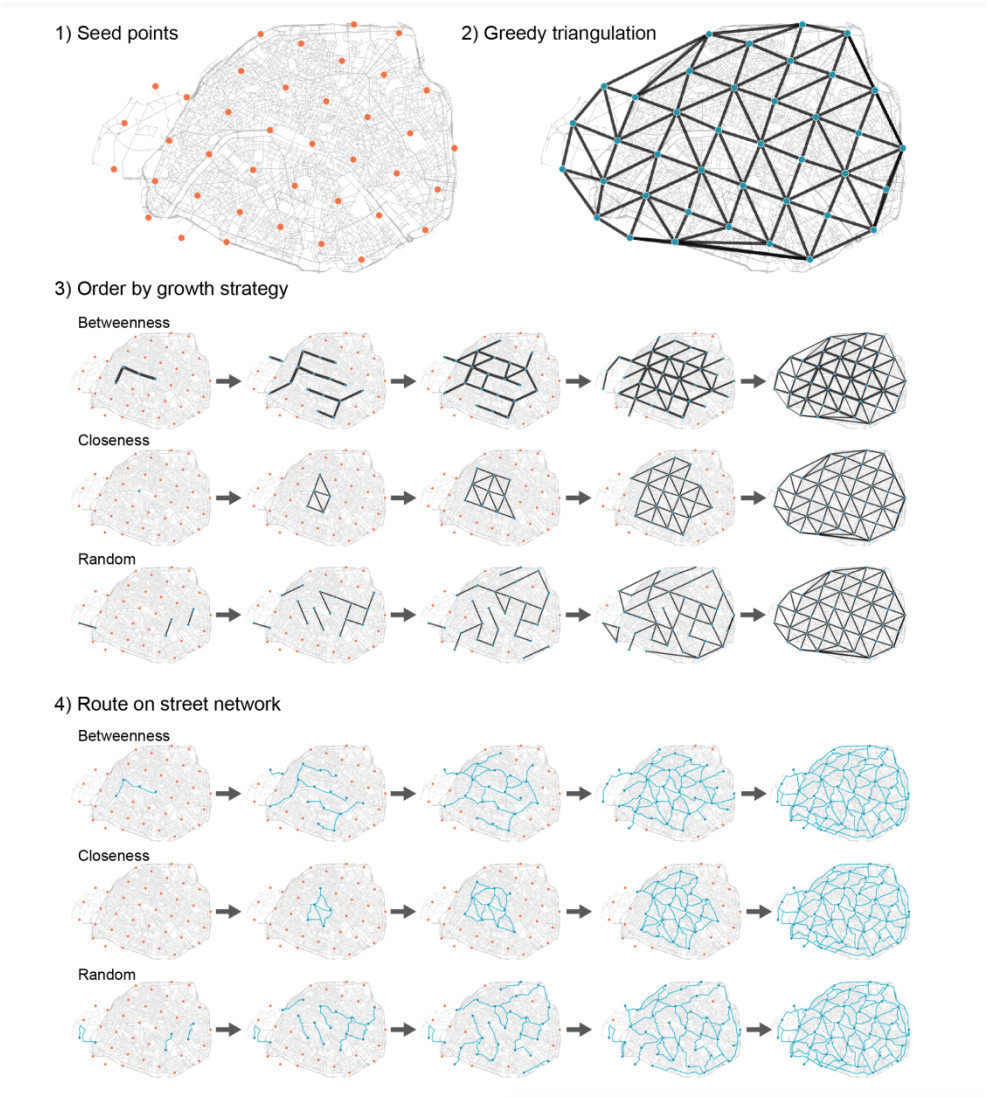


Figure 1: Stepwise network growth illustrated for Paris. Step 1) Seed points are snapped to intersections in the road network. Step 2) Greedy triangulation. Step 3) Order by growth strategies visualized for five chosen growth quantiles. Step 4) Routing links in the growth stages on the street network.

Evaluation metrics

Each of the three growth strategies mentioned optimize different quality metrics. Both existing and newly generated networks are analyzed to compare the chosen metrics from which the quality of the network is defined. I have chosen a few main metrics which I aim to use for the evaluation of the proposed networks as well.

Directness

The main results of Szell et al. (2022) focus on the 'Directness' of bicycle networks. The directness between two nodes in a graph i and j is defined as the ratio $\frac{d_E(i,j)}{d_G(i,j)}$ between euclidean distance $d_E(i,j)$ and distance of the shortest path $d_G(i,j)$. The directness of the network is the average of this ratio for all pairs of nodes:

$$D = \left\langle \frac{d_E(i,j)}{d_G(i,j)} \right\rangle_{i \neq j}$$

Directness is argued to be the most important metric for bicycle network planning, apart from connectivity.

Coverage
Network coverage is determined by measuring the combined area of all network elements within a certain buffer zone of εm around nodes and links. Szell et al. (2022) use a value of $\varepsilon = 500$ m, ensuring complete coverage of a city area for grid triangulation and an average distance being 167 m to the network. Furthermore, Szell et al. (2022) define the metric 'POI coverage', to evaluate how fast network elements cover all seed points as the network grows. Szell et al. (2022) argue that a comprehensive bicycle network should cover most of the city.

Overlap with existing infrastructure
Szell et al. (2022) measure, for each growth stage, the overlap of the grown network with the already existing bicycle and bikeable infrastructure. The results of Szell et al. (2022) show that betweenness tends to have a high overlap at the beginning of the network growth, especially for a city such as Copenhagen, which has an already well-developed bicycle infrastructure, and argues that this indicates that cities take into account traffic flow when building new bicycle infrastructure.

Global efficiency
A network's global efficiency Latora and Marchiori (2001) measures how efficiently information is transferred. It is measured by computing the harmonic mean of the lengths of shortest paths between all node pairs in a network:

$$E_{glob} = \frac{\sum_{i \neq j} \frac{1}{d_G(i,j)}}{\sum_{i \neq j} \frac{1}{d_E(i,j)}}$$

The numerator is the sum of inverse shortest path lengths between all node-pairs in the network, where $d_G(i,j)$ defines the length of shortest paths between nodes i and j in a network G. The denominator is the sum of inverse shortest path lengths between all node-pairs in a network that is fully connected, where $d_E(i,j)$ defines the length of the shortest path between nodes i and j in a complete graph where the number of nodes is the same as in G. The global efficiency is the ratio of these two values. Szell et al. (2022) underline that synthetic networks tend to have three times the global efficiency.

From evaluating the networks, some of the results of Szell et al. (2022) show that these urban bicycle network planning strategies would NOT instantly improve the quality of a bicycle network. Systematic commitment is often needed to surpass a 'critical threshold,' from which dramatic quality improvement can be seen. From a geometric perspective, this critical threshold is described by Szell et al. (2022) as: 'the emergence of a well-connected giant component'. Cities must be consistent with overcoming this fundamental topological limitation to gain benefit from these growth strategies.

Szell et al. (2022) point out that traditional slow-paced urban planning resembles the random growth strategy. Random growth, although having the fastest growth of coverage, has a low directness, connectedness, and efficiency. Random growth is often the result of cities following common urban planning strategies based on local stepwise refinement. Szell et al. (2022) therefore encourage city planners to implement bicycle networks in a more holistic way, rather than a piece-wise local approach, and, e.g., argue how betweenness-strategy can reduce the cost of growing bicycle networks by three times, compared to the random-like strategy often observed.

2.1.5 Growing weighted networks and data spareness

The analysis of Szell et al. (2022) would be categorized as data-agnostic, as most data is obtainable through spatial information (OSMnx). Edge weights in the graphs of Szell et al. (2022) are defined as the length of a given edge/street.

In the analysis of Olmos et al. (2020), a weighted network is created, where edge weights are defined as the volume of users that pass through a given edge. For Bogota, Olmos et al. (2020) address socio-demographic biases of the data from the BIKO app since only 18.3% of the app users are low-income (SES 1 and 2). In comparison, 51% of the population of Bogotá belongs to the low-income group. Olmos et al. (2020) argue that this is most likely because of this population group's low access to smartphones. This means that Olmos et al. (2020) dont have exact data for each edge in the network, and generalizations about these edge weights are made.

Olmos et al. (2020) integrate trajectories from biking applications and generalize edge weights using percolation analysis (a mathematical technique to study random network behavior, focusing on the emergence of large-scale connectivity). This is done to ensure global connectivity and identify a minimal connected network of bicycle paths that covers the whole city and is composed of links with the highest population density. Stepwise, the process is as follows:

First, raw GPS data from the BIKO app (a smartphone app encouraging the use of bicycles BIKO (2019)) is parsed, and stay points are extracted from a sequence of traces of an anonymous individual user. From these, speeds, durations, and distances are analyzed, and a potential bicycle demand is defined. The potential demand is a distance-based measure of trips that could be done on a bicycle and is obtained from filtering data from the BIKO app using an rejection-sampling algorithm (a statistical method to generate samples from an unknown, or difficultly computable distribution).

Existing infrastructure is downloaded with OSM, and the potential bicycle flow is map-matched. Shortest paths are calculated between nodes from the potential demand (nodes of all trips and links after rejection), generating a weighted bicycle-path network.

The potential flow and the OSM-bike path networks are joined, and percolation analysis is performed to ensure global connectivity. This allows for identifying the minimal connected bike-path network, which both covers the whole city and is composed of the links with the highest bicycle flow.

From the resulting network, some geo-processing steps are done, and edges are subtracted which do not have an existing or projected bicycle facility; if their end nodes are both not further away than 50 meters. Otherwise, they would be categorized as proposed links.

Finally, they implement an algorithm that collects street-level imagery of road sections to intervene to provide decision-makers with a tool to evaluate the proposed interventions' viability visually.

Like Olmos, I aim to find computational methods for edge weight completion as the data I will be working with is likewise sparse.

2.1.6 Traffic prediction and edge weight completion for road networks

Like Olmos et al. (2020), the bicycle count data I will be using, consisting of 122 bicycle count points holding just a single yearly measure pr counter-point, is incomplete for all (approx 65000) edges in the street network of Copenhagen. With this minimal data, I aim to create generalizations about bicycle counts (See the section 4.1.1). However, more precise predictions can be made with more extensive temporal data and the help of machine learning techniques.

Spatial data, combined with sufficient temporal data, can allow for using more complex models to account for data spareness in road networks. More recently, machine learning techniques, such as deep learning and neural graph networks, have been used for traffic prediction in road networks.

The LSM model (Latent Space Model) Deng et al. (2016) is a state-of-the-art method for traffic prediction. Where previous models did not consider the underlying structure of the network, the LSM model incorporates both spatial and temporal information for more accurate prediction of time-varying patterns in traffic. The road network is represented as a set of latent variables from observed traffic data, assuming that traffic flow on a given road segment is described as a combination of these variables and the latent variables of its neighbors in this network. They use an extensive data set of traffic data from the New York City area, collected over several years. The data holds 5-minute interval traffic flow measures for more than 20.000 road segments, including topological information about the network (how roads are connected). The model expects its variables to vary over time, and they use a combination of Gaussian (statistical models used for continuous functions) and auto-regressive processes (mathematical models used to describe the relationship between current and previous observations) to model the temporal variation.

Inspired by this, Hu et al. (2019) are also concerned with traffic prediction. Still, in contrast, they are using graph convolution networks (CGNns), which is a way of using semi-supervised learning on graph structure data. They propose a stochastic weight-completion algorithm on these networks to account for missing edge weights. Like Deng et al. (2016), their method is based on spatial and temporal data, including 5 - minute interval traffic flow measures. However, their data also holds a lot of missing edge weights (in this case absent measures of traffic flow), and they use CGNns to calculate the stochastic weights for the edges not covered by traffic data. Deng et al. (2016) point out that existing studies addressing the data spareness problem only consider deterministic weights. All employ linear models to cope with these correlations among edge weights, even though these correlations can be highly non-linear. Therefore they propose a framework that enables stochastic weight annotation, considering non-linear weight correlations.

These studies are highly in line (and a slight step ahead) with my reflections about the potential of generalizing edge weights for bicycle counts. My initial thought was also to generalize edge weights in connection with their neighborhood, but just because Nørrebrogade in Copenhagen (the World's busiest bike lane copenhagensize (2011)) is busy does NOT imply that the nearest crossing streets of Nørrebrogade are busy too. Therefore a further study of the network topology is needed (see section 4.1.1 for generalizations of bicycle counts). If temporal data, such as the data used by Deng et al. (2016) and by Hu et al. (2019), were available for bicycle counts, it would have made it possible to make a lot more accurate predictions/generalizations for bicycle flow.

2.1.7 Extending graph edges with weighted distance calculations

As mentioned, in the analysis of Szell et al. (2022), we see that all edge weights are provided by spatial information of the network, in this case, the edge lengths.

As observed in the analysis of Olmos et al. (2020), edge weights can also be defined by an application of external data, in this case, potential bicycle flow.

As observed in the analysis of Folco et al. (2022), there might also be a need to define edge weights based on several factors, such as the length of edges, demand for scooters, and safety constraints.

If several factors are being accounted for, one also needs to decide the importance of each factor. Folco et al. (2022) solve this problem by formulating it as a mathematical optimization problem and use an algorithm to find the best scooter locations based on demand and safety constraints. Their model consists of six steps:

Step 1: Identify potential scooter locations by creating a grid of evenly spaced cells covering the entire city. The probability of starting a trip at each grid cell is calculated and used as a weight in the next step.
Step 2: Determining the potential links between the grid cells by connecting each cell to its neighboring cells within a certain distance threshold. This distance threshold is determined on the assumption that scooters will not drive more than a certain distance between recharging their batteries or maintenance.
Step 3 - Applying weighted distances: using a mathematical optimization algorithm to find the best scooter locations based on demand and safety constraints. The weighted distance metric considers the distance between the grid cells and the demand for scooters (step 1) to ensure that scooters are placed in areas with high demand. It also considers the safety constraints and a weighting is applied to grid cells where collisions are more likely to occur to ensure that scooters are placed in areas with low safety risk.
Step 4: Optimization algorithm. The fourth step involves solving the optimization problem to find the best scooter locations. The objective is to minimize the total weighted distance of all potential links with a certain amount of scooters. The weighted distance metrics between two grid cells, A and B, are calculated, considering the trade-off between demand and safety. The formula is a linear combination of the two distances, and the α -parameter is used to balance the trade-off between demand and safety. This value is a scalar value between 0 and 1. E.g., if α is 1, only the demand will be considered in the final metric. The overall formula (see section 4.1.2 for my customization of the formula) looks as such:

$$d_W(A, B) = \alpha \times d_{\text{trip}} + (1 - \alpha) \times d_{\text{crash}}$$

Step 5: Analysis of the solution to the optimization problem to understand the distribution of scooters across the city and the characteristics of areas where scooters are located.
Step 6: Implementation and evaluation of the solution in the real world and evaluation of the effectiveness by monitoring both the usage of scooters and the safety. From this evaluation, adjustments and improvements to the scooter network can be made if needed.

The approach used by Folco et al. (2022) focuses on scooters and micro-mobility in general and would be categorized as partly data-driven. Like Olmos et al. (2020), the focus is on data-science methods to inform infrastructure planning decisions.

From the analysis of Folco et al. (2022), I will be fine-tuning the weighted distance formula from step 4 to fit my weighted networks. I will define a weighted distance measure that considers the trade-off between edge lengths, population density, and bicycle flow.

2.1.8 Modifying edge-betweenness-centrality

Betweenness centrality is a widely used tool in network statistics, measuring to which extent a node or edge is on the shortest path between other nodes or edges in the network.

As argued by Szell et al. (2022), using betweenness as a growth strategy can reduce the cost of growing bicycle networks by three times, compared to the random-like strategy often observed in traditional slow-paced urban planning.

This strategy is also chosen by Folco et al. (2022) as a simple proxy for flow. This is due to its effective way of quickly building a functional network based on the weighted distance measure and ranking potential links. Figure 2 shows an excerpt of the network growth process of Folco et al. (2022), where in this case, population density counts are defined as seed points. The points are connected, and betweenness-centrality is calculated for each link. We notice how the edge betweenness-centrality is visualized to the outer right.

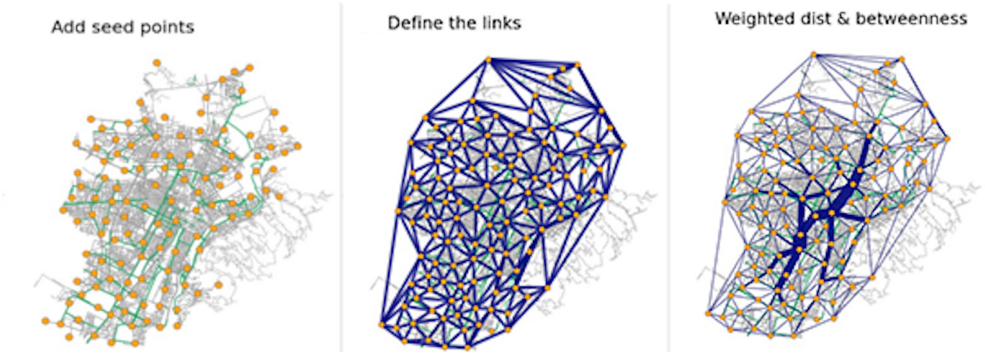


Figure 2: First seed points are added to the existing bicycle network. Second, triangulation is performed, linking the abstract network. Third, for each link, the weighted distance dW is defined, and edge betweenness (represented as the width of the links) Folco et al. (2022) is calculated.

Vybornova (2021) uses topological network analysis to detect gaps in urban bicycle networks, and in this case, for the city of Copenhagen. Vybornova (2021)'s methods are highly in line with Szell et al. (2022) but also include several methodical considerations for betweenness-centrality, using edge betweenness-centrality as a proxy for bicycle traffic flow (the number of cyclists expected on a particular link of the network). This is likewise to inform decision-making about building new bike lanes. As mentioned by Vybornova (2021): "indeed, if there was no bicycle path yet, you'd better place one there".

In Vybornova (2021)'s analysis, for choosing which gaps to prioritize, betweenness-centrality is calculated for each gap and weighed by length, as the shortest gaps are preferred. Each gap is assigned a new length measure; the bicycle count is multiplied by edge length to get how many meters have been cycled. An example of this, as mentioned by Vybornova (2021): "Lets assume that gap A has a length of 10 m and a traffic volume of 50 cyclists in a time unit (e.g., during one day); and gap B has a length of 20 m and a traffic volume of 15 cyclists. Then, by multiplying the numbers, we obtain a total of 500 m for gap A and 300 m for gap B, with the values indicating how many meters less would be cycled in mixed traffic if this gap was to be provided with a protected bicycle path. In other words, by "closing" gap A, we would avoid more meters cycled in mixed traffic. Gap A is therefore ranked as more relevant than gap B".

Figure 3 shows the Copenhagen network, where the red edges are the gaps, the number is their rank, and 1 is the busiest gap (Knippelsbro).

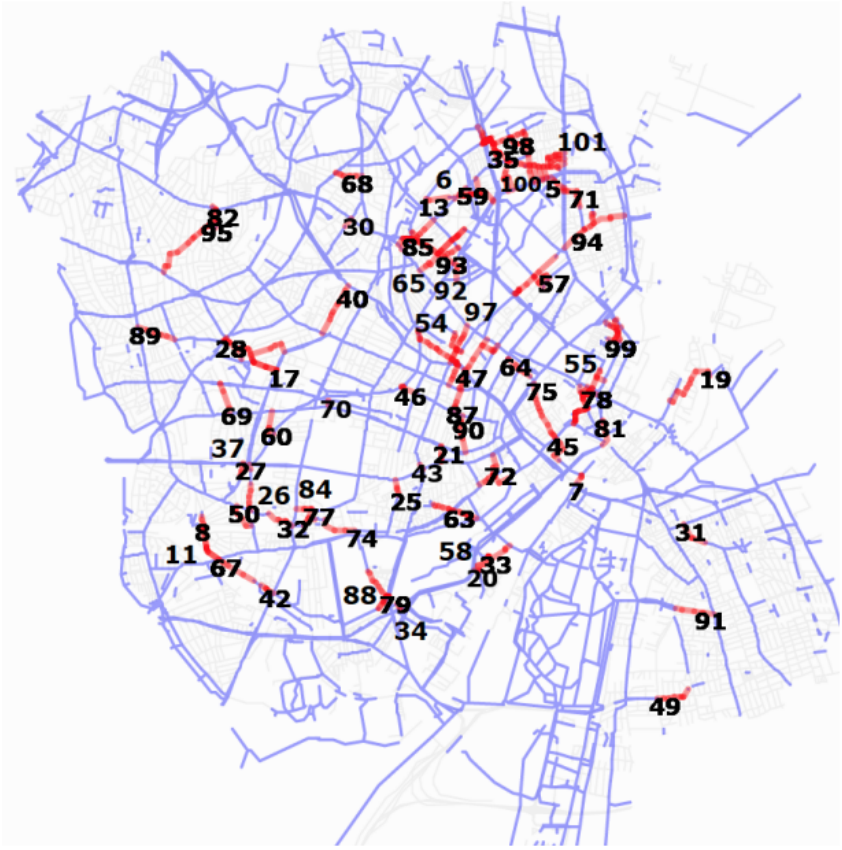


Figure 3: Vybornova (2021): "Overview map of the 67 missing links. The street network is grey, the bicycle network is blue, and the missing links are red. Numbers correspond to gap ranking."

Vybornova (2021) furthermore addresses several challenges with using betweenness-centrality, such as the 'network edge effect', related to the impact that changes in network boundaries might have on centrality measures; a city's physical streets rarely stop at municipal boundaries, and these boundaries might also change over time. The impact of this effect is reviewed, and studies are mentioned that propose a cut-off radius for shortest paths for calculating betweenness-centrality to smooth this effect. Vybornova (2021) is applying a cut-off radius of 2500 m to calculate the shortest paths for this purpose. This results in a much narrower distribution of edge betweenness-centralities because of less extreme outliers, reducing the previous bias towards the city center when calculating betweenness-centrality with no cut-off radius.

Like Vybornova (2021), I aim to modify edge betweenness-centrality calculations to express connectivity from a connection between edge length and a given attribute. I also aim to find an aesthetically pleasing way of visualizing edge betweenness for the street network of Copenhagen.

2.1.9 Conclusion

To explore how population density can be used as a proxy for generalized traffic flow in Copenhagen, my process has consisted of researching data-driven and data-agnostic methods for growing urban bicycle infrastructure. This also included a survey of the potentials in the combination of both spatial and temporal data and

some machine-learning techniques (as described by both Deng et al. (2016) and by Hu et al. (2019)) to generalize traffic flow.

Having no prerequisites in data science, an equally significant part of my process has been studying geospatial data science and getting the necessary coding skills to understand and extend previous work in the field.

I aim to extend the growth framework developed by Szell et al. (2022), and as Szell et al. (2022) encourage, I will include place-specific data in this analysis and apply spatial generalizations for traffic flow.

My grown networks will be based on network structure (as described by Szell et al. (2022)), traffic flow data (as described by Olmos et al. (2020)), and population density data. I will include weighted distance calculations (as described by Folco et al. (2022)) and implement an algorithm for assigning edge weights to the road network in Copenhagen. This includes calculating data-aware edge-betweenness-centrality measures (as mentioned by Vybornova (2021)) and using this as a proxy for traffic flow.

3 DATA ACQUISITION AND PROCESSING

3.1 INCLUDING BOTH SPATIAL AND PLACE-SPECIFIC EMPIRICAL DATA

When growing synthetic bicycle networks, both spatial and place-specific empirical data can be used.

Spatial data refers to any data linked to a geographic location, such as GPS coordinates, boundaries, addresses, etc. In my case, it refers to network structure, meaning the network's physical infrastructure, such as bike lanes, roads, and sidewalks, for a given geographic location. This data can be analyzed using Geographic Information Systems (GIS) and is typically collected from satellite imagery and ground surveys. OSMnx is a Python package that uses GIS techniques to work with OpenStreetMap (OSM), enabling me to download and preprocess such spatial data.

While spatial information can sometimes be more based on assumptions about how a particular system behaves and sometimes a little unreliable (see section 3.2 for explanation), place-specific empirical data, such as bicycle counts and population densities, is gathered from real-world phenomena and can provide valuable insight into the actual behavior of a system Burrough and Frank (1995).

In this section, I will describe my process of gathering and preprocessing spatial data and place-specific empirical data for bicycle counts and population densities.

3.2 DATA PREPROCESSING

3.2.1 OSM datasets and OSMnx

The first part of the data acquisition process for this study is gathering spatial data from OpenStreetMap (OSM). The Python library OSMnx Boeing (2017) is used for downloading the CSV files and encoding the OSM data structures, (see section 2.1.4 for further explanation of OSMnx), and for our case, only the street network of Copenhagen is being utilized.

The bicycle network of Copenhagen is composed as a union of both on-street and off-street protected bicycle infrastructure. Each node in the graph represents an intersection, and each edge represents a street connecting two intersections. The graphs are simplified to clean up nodes that are not intersections while retaining edge geometry. Figure 4 shows the entire street network of Copenhagen, and figure 5 shows the overlapping of bicycle infrastructure with car infrastructure after a few processing steps.

Figure 4: The street network of Copenhagen, from OSM data (downloaded February 2023)

Figure 5: The largest connected component in the Copenhagen street network. Car links are shown in grey, bicycle links are in green, and the black links show where these two types of links overlap

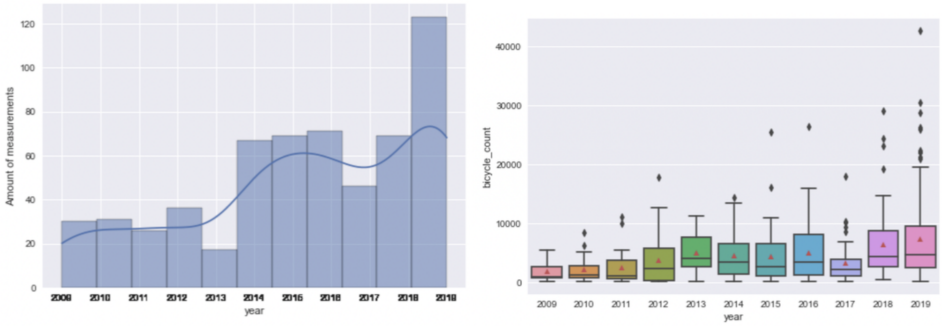
When working with OSMnx, addressing some data quality issues is essential. While having the benefit of enabling the provision of open-source data, there might, unfortunately, be some uncertainty of correct labeling in OSM, especially for less common types, like bike lane infrastructure; bike lanes and streets where biking is possible are not always distinguished. A more optimal solution would be downloading bicycle networks from a more approved source. Though, for street networks, OSM is proved to be rather accurate.

3.2.2 Bicycle count data

Besides spatial information, I will include place-specific empirical data in this study to inform network structure and improve network quality. The data I use provides bicycle counts from Copenhagen (average counts of bicycles passing through locations in the bicycle network on a full day) and is downloaded from Opendata.dk (2014). The data portal Opendata.dk provides collective data from a union of municipalities in Denmark.

The data set initially contained 36 attributes and 1077 entries, including counts for pedestrians, bicycles, electric scooters, buses, trucks, and other vehicles. Each traffic count also had an associated date of the measurement, the address, and the geographic coordinates of the measurement. This data was preprocessed and cleaned by Ditte M. Hjorth, and Sidsel Rindom Koch Ditte M. Hjorth (2021), in several steps, before being handed over to me; first of all, all irrelevant attributes were removed (vehicles that are not bicycles) and from these, only bicycle counts from weekdays were selected to make it as illustrative for everyday traffic as possible. Furthermore, the data set initially contained some duplicate values and some counts with zero values. These values were considered errors and removed.

The final data, handed over to me, includes 585 data points with bicycle counts measured over ten years, from 2009 to 2019. Basic statistics show that 100 is the minimal bicycle count, and 42600 is the maximum. Figure 6 is obtained from the analysis of Ditte M. Hjorth (2021) and shows on the left (a) the yearly amount of measurements on the y-axis. We note that the amount of measurements, meaning the number of counter-points placed in the city, has increased. This is highly in line with the general trend in Copenhagen that more and more people bike Kommune (2022). On the right is a further investigation of the data, showing the distribution of counts in box plots, with the mean displayed as red triangles. The distributions are skewed positively, indicating many outliers (e.g., the very busy Dronning Louises bro), but the growth in the mean value conforms to a growing bicycle trend.



(a) Amount of measurements made per year

(b) *bicycle_count* per year with mean

Figure 6: Yearly measurements and bicycle counts. To the left: a histogram showing yearly measurements, to the right: boxplots for each year.

Initially, I aimed to include all 585 data entries for my analysis to have as many data points as possible. Still, after application to the Copenhagen network, visualizations made it clear that some counts were placed very close to each other, indicating that the measurement was done at the relatively same spot for each year. Therefore these could be considered duplicate values in this particular case. I have chosen to consider only data from 2019, as this was the most recent data and the data with the maximum number of measurements. The data consists of 122 measures for several locations around Copenhagen, with a minimum count of 100 and a maximum of 42600 bicycles, representing the whole range of bicycle counts. The mean value is 7398.4. The distribution of bicycle counts is displayed in figure 7

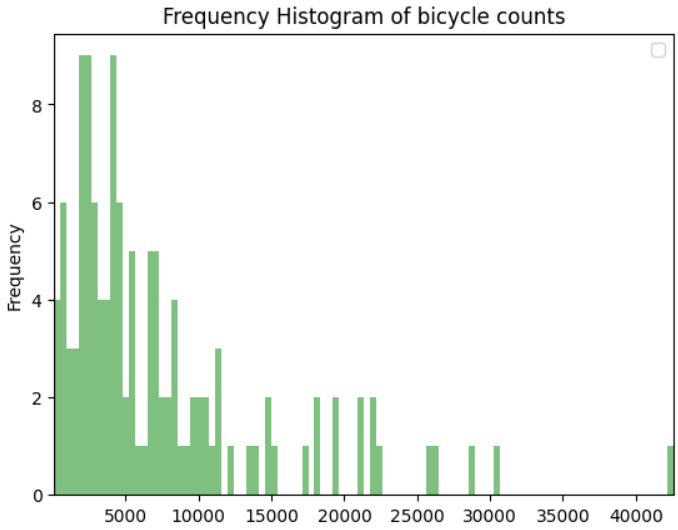


Figure 7: Distribution of bicycle counts from the year 2019

3.2.3 Population density data

Besides bicycle counts, I also choose to include data about Copenhagen's spatial distribution of population density. Population density data is geospatial data that describes the number of people living in a given area and can be derived from various sources, such as satellite imagery, census data, or mobile phone data. This data type is used for urban planning and in various other fields, e.g., public health Dulin et al. (2010) and resource allocation in general.

The data I am using is obtained from WorldPop (data.humdata.org (2019)), which can provide various gridded population density counts for cities worldwide. I choose to include a data set for 2019, as the bicycle counts are also from 2019. The data holds population density in ASCII XYZ format for Denmark, with a resolution of 30 arc seconds (roughly equivalent to 1 km at the equator). In other words, the data represents the number of individuals per square kilometer for each pixel in the grid.

Basic statistics show that for the whole country of Denmark, the minimum population density is 0,04 (approximately the twentieth of a single person), and the maximum is 35463,5 people per square kilometer. The mean is 134.2. This is quite a range, also displayed in the frequency distribution in figure 8, and the heat map in figure 9 of population densities explains how the high values seem to be centered around large cities in Denmark.

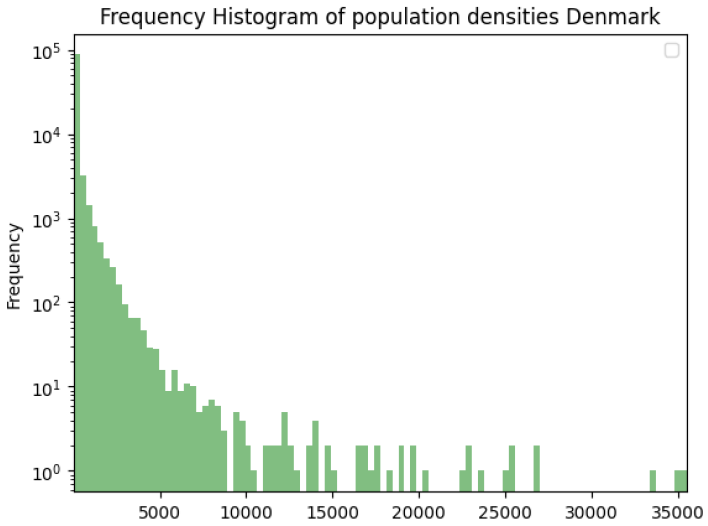


Figure 8: Frequency distribution of population densities in Denmark for the year 2019 (log scale on y-axis)

Figure 9: Geographical visualization of population densities for the whole country of Denmark in the year 2019

To obtain the data for Copenhagen, some geoprocessing steps had to be done, as the data was only in XYZ format (see section 4.1.1 for data description). We see on the frequency distribution in figure 10 that for Copenhagen only, the data appears a lot less dispersed. Basics statistics show that the minimum density is 2.8,

the maximum (also being the global maximum for Denmark) is 35462.5, and the mean value is 6893.5. This is a relatively larger mean value than the whole country of Denmark, most likely due to people often being more densely populated around cities UNCTAD (2022). The heat map in figure 11 shows where in the city people are most populated.

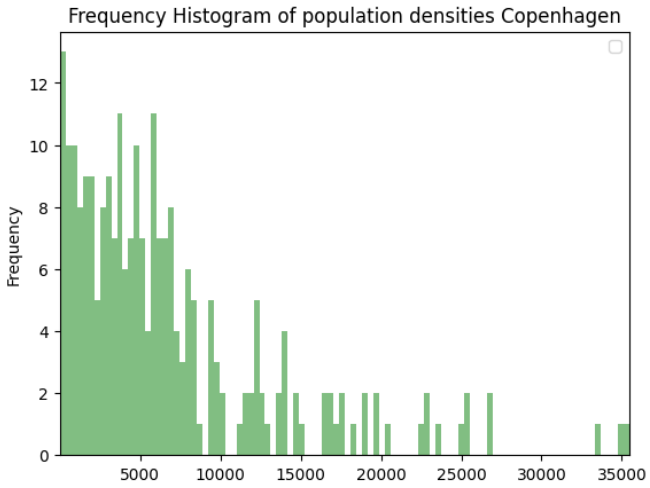


Figure 10: Frequency distributions of population density in Copenhagen for the year 2019

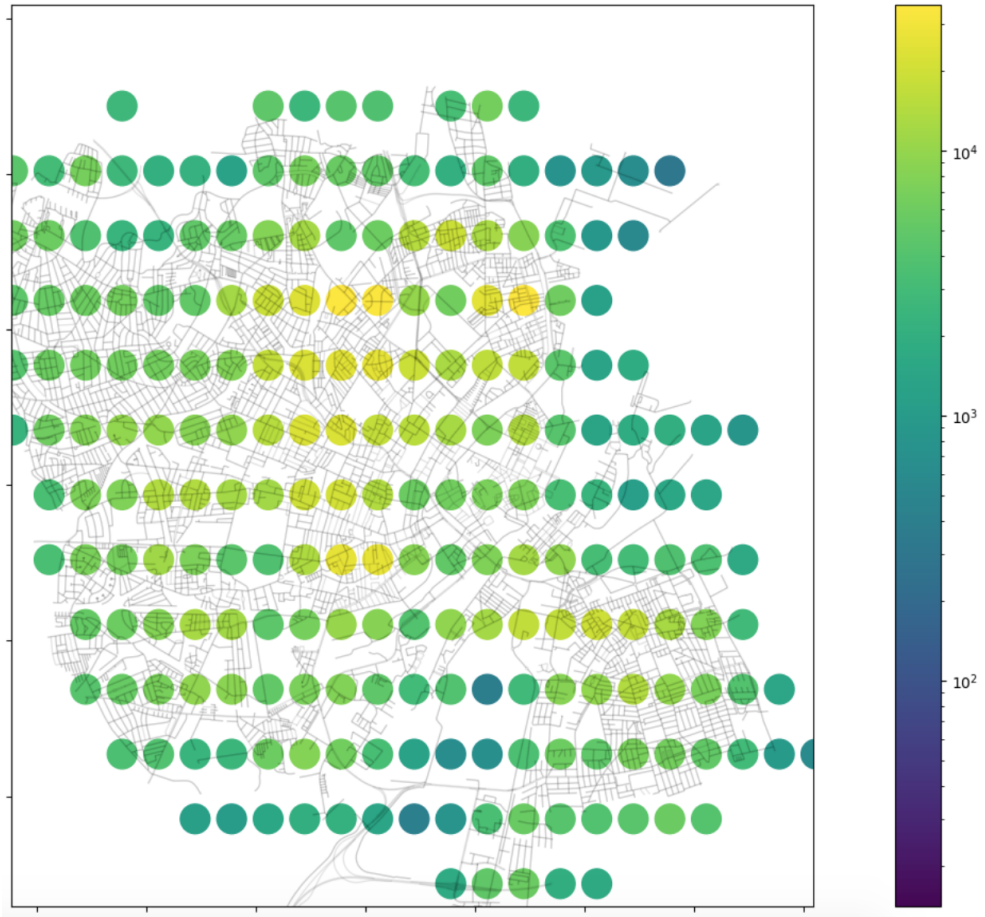


Figure 11: Geographical visualization of population densities for the city of Copenhagen in the year 2019, visualized on top of the street network of Copenhagen.

4 DISCUSSION

4.1 BUILDING A DATA-DRIVEN NETWORK

4.1.1 Model description

The OSMnx library allows downloading and analyzing OSM data (spatial data) for specific locations, from which networks can be grown (see 2.1.4 for description of how networks are grown) and visualized. As an extension of this, yet another powerful tool is associating place-specific empirical data to these networks to inform actual observed patterns (see section 5.1 for description of the observed patterns from the analysis proposed in this thesis).

The way I aim to grow networks is based on Szell et al. (2022) and the basis of this network growth, including downloading existing infrastructure, seed point generation, and greedy triangulation, is explained in section 2.1.4, and the ingredients used for the network growth of Szell et al. (2022) are described in section 2.1.3. As an extension of this network growth, based on spatial information (network structure), I will include place-specific empirical data from the city of Copenhagen, resulting in networks that are also partly data-driven. The source code is provided on Gasior (2023).

Using OSMnx, we can obtain spatial information from OSM, such as the length of road segments, represented as edge lengths in a graph. The bicycle count and population density data I use holds coordinates for each data entry, and with a few geoprocessing steps, this data can be associated to a metric space by assigning these counts as edge weights in graphs. I choose to assign data to edges in the graph, as these represent road sections in the street network of Copenhagen. Each road section will therefore be informed with the inferred bicycle flow and population density. My final networks are built in OSMnx using four ingredients:

- Street-and bicycle networks from the city of Copenhagen. This data is downloaded and processed from OpenStreetMap and embedded in a metric space.
- An arbitrary set of seed points representing points of interest and implemented as nodes in planar graphs. (Szell et al. (2022))
- The spatial distribution of population density in Copenhagen. Assigned as edge weights to graphs.
- Bicycle counts from Copenhagen, representing bicycle flow. Generalized and assigned as edge weights to graphs.

As informed by Folco et al. (2022), edge weights can be even further specified using weighted distance measures. With this additional empirical data, we can define edge weights as a trade-off between bicycle flow, population density, and the length of a given road segment (see section 4.1.2 for calculation of edge weights). With a weighted distance measure, we will grow networks from the seed points given by Szell et al. (2022) and analyze whether these networks are optimal (see section 4.2 for analysis description). Growing networks from graphs with these customized edge weights will allow us to explore the possibilities of building bicycle infrastructure that is both accessible to the people living in the city, as well as connecting areas with a high bicycle flow.

- Step 1: Downloading and processing existing network data from OSM.
- Step 2: Embedding existing network data in a metric space
- Step 3: Generating seed points and implementing these as nodes in planar graphs.
- Step 4: Preprocessing and associating population density data to the nearby edges in the graphs.
- Step 5: Preprocessing and associating bicycle count data to the nearby edges in the graphs.
- Step 6: Using spatial information to generalize about bicycle count data to account for data spareness.
- Step 7: Applying weighted distance measures as edge weights in the graph, with various trade-offs between bicycle flow, population density, and the length of each given road segment.
- Step 8: Growing networks with data-aware betweenness-centrality as a growth strategy.
- Step 9: Analysing and evaluating the quality of the networks and comparison to the baseline (the network grown by Szell et al. (2022)).

In the background section, steps 1-3 are explained (however, not denoted in steps), as Szell et al. (2022) already did this. This section will focus on steps 4-8.

Embedding population densities (Step 4)

The gridded population density data (see section 3.2.3 for data description) represents the whole country of Denmark and provides the number of individuals per square kilometer for each pixel in the grid. The data is in an XYZ format, representing coordinates; X = longitude, Y = latitude, Z = population density. No information is provided about which region or address the count belongs to, and therefore a few geoprocessing steps are needed to filter only the necessary entries from this data.

OSMnx has built-in functions for finding the nearest edges and nodes in a graph for a specific (latitude, longitude) coordinate. This means that we can find the nearest road segment in a street network for a given specific location. OSMnx also has a built-in function for calculating the haversine distance (the distance between two points on a sphere, like the Earth). This distance can specify a threshold based on the max haversine distance between the location and the road segment; how close does the nearest edge have to be, to be assigned a value. I've chosen a threshold of 500m, meaning that a given point has to be within 500m of reach to the road segment we wish to attach its data to. If this wasn't utilized, there might be a risk of, e.g., a population density from Western Jutland being attached to a road segment somewhere in Valby (Copenhagen), depending on how population densities were iterated in the assignment process.

Initially, we iterate through the whole grid of population densities and assign each count to the Copenhagen street network, but only if it is within 500m of reach to the nearest road segment in Copenhagen. This reduces our population density data set from 94677 (the whole country of Denmark) to only 239 entries (Copenhagen). Figure 12 illustrates the network and how each of the 239 gridded data points could be assigned to the very nearest edge in the graph. Highlighted edges are the ones containing a population density count. Notice how the evenly spaced gridded placement of the nearest edges corresponds to the gridded placement of the data points (see figure 11).

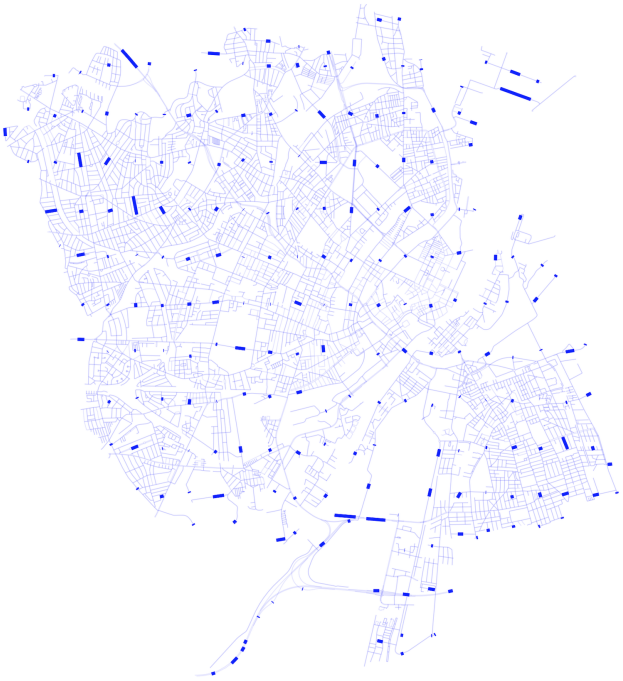


Figure 12: Gridded population density data from 2019 associated to the single nearest edges in the Copenhagen street network. The nearest edges are highlighted.

However, assigning population density counts to the single nearest edge in the street network graphs would only leave us with 239 assigned edges out of 66053 edges in all. Therefore, to account for this spareness, we choose to iterate through all edges in the graph and assign the nearest population density count to each edge. This is highly in line with the fact that each data point is a measure of a whole square kilometer (which would most likely include more than just one single edge in the graph), justifying assigning a population density count to a whole neighborhood of road sections.

The population density counts are applied to the graph as edge weights, and figure 13 illustrates the street network of Copenhagen, where both edge thickness and edge opacity are determined by the population density count, enhancing the more populated areas. Notice the connection between regions highlighted in this figure and the yellow (densely populated) areas in figure 11, illustrating the population densities prior to being associated to edges in the graph. This indicates correct assignment of counts to the nearest edges.

Figure 13: The street network of Copenhagen, where all edges are assigned the nearest population density count from the year 2019. Densely populated areas are enhanced.

Embedding bicycle counts and generalizations for bicycle flow (Steps 5 and 6)

Like population densities, I will associate bicycle counts (the average counts of bicycles passing through specific locations in the bicycle network on a full day) to the Copenhagen street network. These can inform bicycle flow and recognize the areas with the most bicycle traffic. Bicycle counts are likewise assigned to the nearest edges in the graph, meaning the closest road segment in the street network. After preprocessing, it was natural to visualize both population density and bicycle flow together to see if there were any overlaps between these two; can we say anything about the bicycle-traffic flow by looking at the population density? Figure 14 shows the placement of bicycle counts, with the size of each red circle relatively describing the count (the bigger the red circle, the higher the bicycle flow), and the blue edge-enhancement describing the population density of each edge. Note how densely populated areas seem to be located around areas with a high bicycle flow, as marked with the yellow circle.

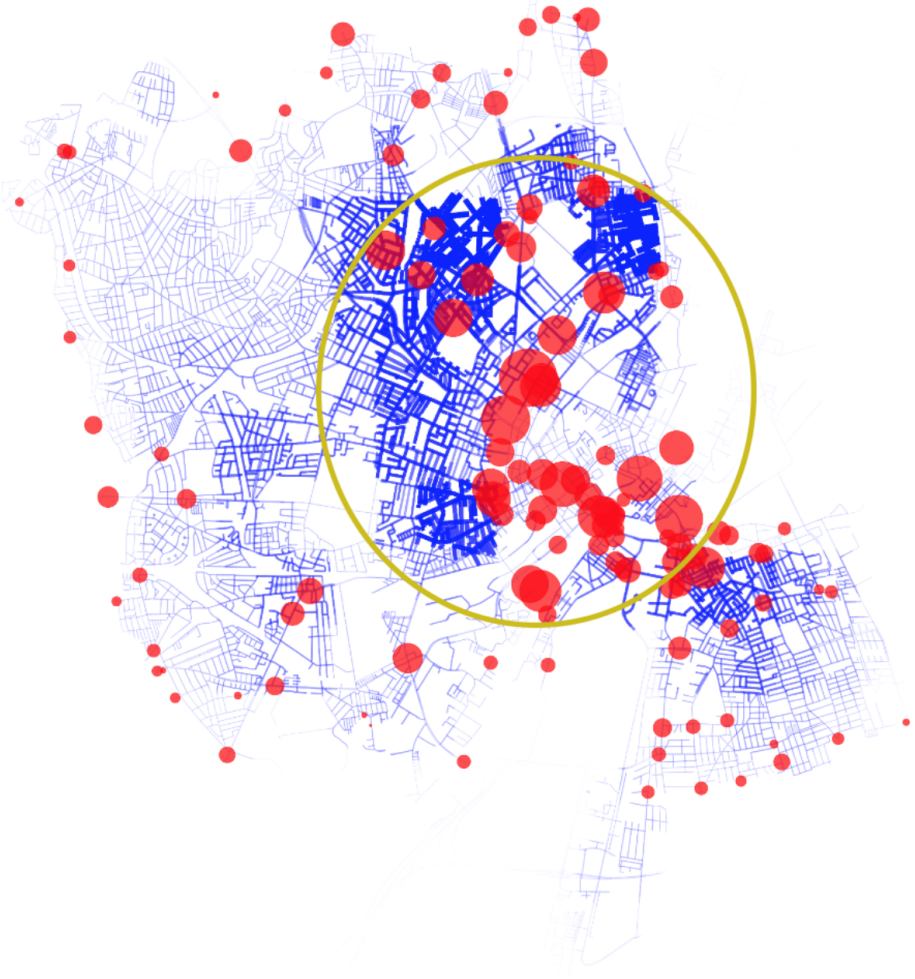


Figure 14: Street network of Copenhagen, where all edges are assigned the nearest population density count (from the year 2019), highlighted in blue, and bicycle counts (from the year 2019), highlighted in red, relative to the bicycle flow.

For population densities, it was given that each counter-point was describing the population density pr. square kilometer, justifying generalizing these counts to a whole neighborhood of edges. For bicycle counts, the number of data entries was even more sparse, consisting of only 122 data entries, and each of these were indeed only describing a single road segment (a single edge in the graph). Therefore, assigning a whole neighborhood of edges to the nearest bicycle count in the graph would not be very descriptive of the actual bicycle flow; e.g., the fact that Dronning Louises Bro has a very high bicycle flow would not necessarily imply that a neighboring small street would have a high bicycle flow. Therefore, other generalizations need to be made.

When using OSMnx to assign a count to the nearest road segment, we observe that in many cases, these assigned road segments appear quite short, probably due to how OSMnx builds graphs. We can also observe in the data, and from the visualization of the data, that in many cases, several counts exist for a given street, e.g., to the left in figure 15 showing the bicycle counts on Amagerfælledvej. Only three edges (out of 202 edges that Amagerfælledvej is constructed of) are assigned a bicycle count. One could reflect on whether the bicycle flow of the remaining edges of this street could have a similar bicycle flow?

I wish to inform as large a part of the street network as possible about bicycle flow,

and I, therefore, choose to generalize about these counts. The generalizations are computed by finding the mean of bicycle counts for a given street, grouping edges in the graph on street name and assigning this group's mean value to the rest of the street.

Suppose counts are placed ideally in different sections of the street. In that case, I will claim that the probability of the other unassigned edges on this given street having a similar bicycle flow is relatively high. To the right, in figure 15, is depicted how the mean of bicycle counts for Amagerefælledvej (8600) is associated to the remaining edges of this whole road. Also, we see at the top how the count on Langebro (thick and enhanced blue line) is generalized to the rest of Langebro. Langebro is a bridge, and one could assume that the count belongs to all edges that represent this bridge in the graph.

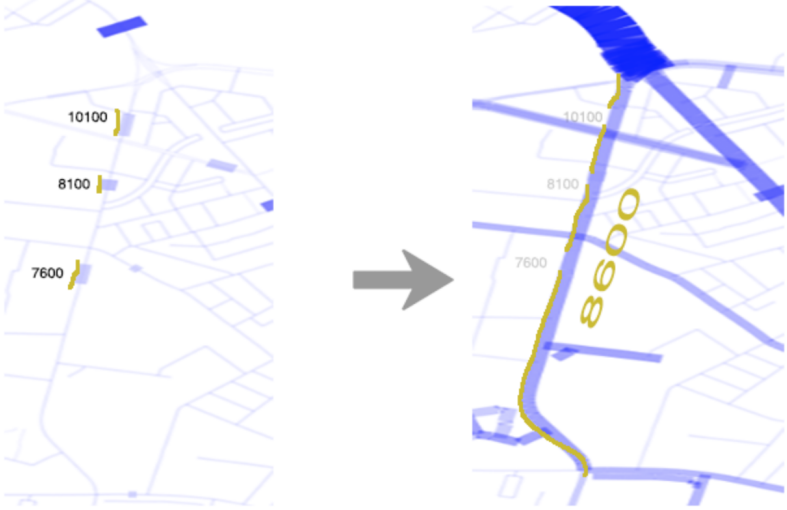


Figure 15: An excerpt of the street network of Copenhagen, showing the mean of bicycle counts for the three assigned edges ((7600 + 8100 + 10100)/3 = 8600) on Amagerfælledvej, being associated to the remaining (199 out of 202) edges, that this street is built of. To the left: bicycle flow (prior generalization). To the right: generalised bicycle flow. The bicycle count determines edge thickness and edge opacity for the given edge in the graph

This generalization of edges gives us 10329 assigned edges out of 66053, roughly 15% of the edges, which is a large improvement from the 0,2% (122 edges) of edges obtainable from the empirical data associated to the street network. Figure 16 shows the complete Copenhagen street network before (left) and after (right) the generalization of bicycle counts. The bicycle count determines edge opacity and edge thickness to illustrate bicycle flow.

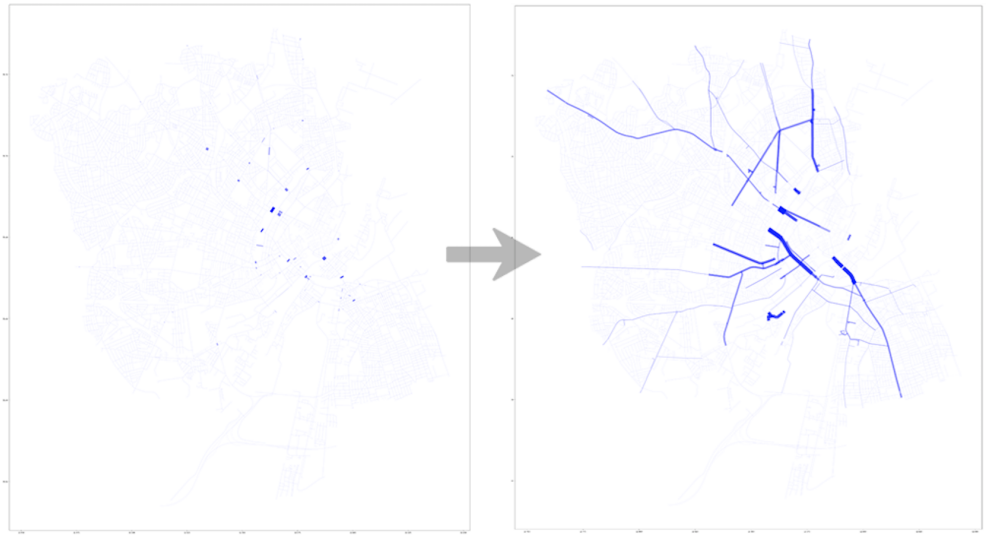


Figure 16: The street network of Copenhagen. To the left: bicycle flow (prior generalization). To the right: Generalised bicycle flow. The bicycle count determines edge thickness and edge opacity for the given edge in the graph

The remaining 85% of edges in the street network are assigned the mean bicycle count value, being 7398.37. This can be argued to be a relatively high bicycle flow, only slightly smaller than Amagerfælledvej, which could be considered a street with a high traffic flow, and therefore perhaps not very descriptive for some smaller streets away from the city center.

This value for unassigned street segments could be further investigated, as it might be influenced by the placement of bicycle counters. Furthermore, if further temporal measures were available for bicycle counts, machine-learning techniques could have been utilized for more accurate generalizations, but data spareness made this out of scope.

4.1.2 Transforming edge data to weighted distance measurements (Step 7)

In the existing code provided by Szell et al. (2022), the edge length (length of the road segment) is used as an edge weight for calculating shortest paths and network growth. Figure 17 shows the distribution of edge lengths to the left. To the right is the street network of Copenhagen, where streets are highlighted based on length.

We have already obtained the edge length from OSM, and with a few geo-processing steps in OSMnx, we will also be embedding population density, and bicycle counts, as edge data. Each edge in the street network graph is assigned the nearest population density count, and each bicycle count is assigned to the nearest edge in the graph and generalized on the street name. This leaves us a graph where each edge holds information about the length of the street segment, the population density, and the generalized bicycle flow.

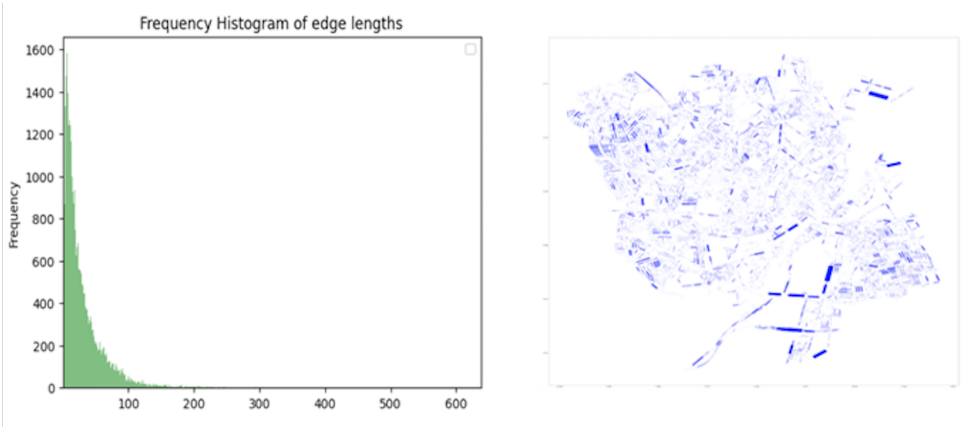


Figure 17: To the left: Frequency distribution of edge-lengths in the street network of Copenhagen. To the right: Visual representation of edge lengths, where the length measure determines edge-thickness and opacity.

With both the edge length and the additional associated edge data, it is now also possible to provide a customized edge weight as described by Folco et al. (2022). This edge weight will be described as a trade-off between the length of a road segment, the population density, and the generalized bicycle flow; a path has to be considered shorter if its edges have a high bicycle flow and/or high population density, and it shall be possible by utilizing the formula, to decide how bicycle flow or population density should be prioritized.

I will introduce a weighted distance measure. This measure will be calculated from a mathematical optimization formula, where the edge length is multiplied by the fraction of each given attribute (either bicycle count or population density count), allowing for balancing the trade-off between both entities. This formula is fine-tuning the formula provided by Folco et al. (2022) (see section 2.1.7 for description). The weighted distance will be calculated as d_W for each edge/(link between two nodes A and B) in the existing infrastructure. For an edge $e(A, B)$ we define $d_W(A, B)$ as follows:

$$d_W(A, B) = \alpha \times d_{\text{bicycle-flow}} + (1 - \alpha) \times d_{\text{population-density}}$$

Where:

(i) $\alpha \in [0, 1]$
(ii) $d_{\text{bicycle-flow}} = d \times N_{\text{bicycle-flow}}$ with $N_{\text{bicycle-flow}} = \frac{1}{n_{\text{bicycle-flow}}}$
(iii) $d_{\text{population-density}} = d \times N_{\text{population-density}}$ with $N_{\text{population-density}} = \frac{1}{n_{\text{population-density}}}$

The variable d denotes the length of the edge $e(A, B)$ in meters, $n_{\text{population-density}}$ is the population density in the neighbourhood that the edge $e(A, B)$ is situated in, and $n_{\text{bicycle-flow}}$ is the average number of bicycles that pass through edge $e(A, B)$ in a 24 hour period. Both the $N_{\text{bicycle-flow}}$ and $N_{\text{population-density}}$ values are normalized, meaning that the range of the denominator of $d_{\text{population-density}}$ and $d_{\text{bicycle-flow}}$ is the interval $[0, 1]$. In this way, the range of variability of $d_{\text{population-density}}$ and $d_{\text{bicycle-flow}}$ is the same.

As an example, let's say we apply the weighted distance measure with $\alpha = 0$ (trade-off lies solely on population density), and wish to calculate edge weights. Then each edge length in the graph gets multiplied by the fraction of the population density, meaning that a road segment with a high population density would be "shortened"; prioritized for shortest paths when betweenness-centrality is

calculated. This is visualized in figure 18, and notice how edges to the left that are more 'visible', meaning high-population density, have now become almost invisible in the graph to the right. Also, notice these extreme outliers in the upper right of the graph to the right and the lower middle of the map. These can be explained by being both relatively long road segments (see in figure 17 to the right, where the same edges appear as highlighted) with a low population density (noticeable to the left as these road segments are almost invisible). These road segments will have a lower betweenness-centrality and are less likely to be prioritized as bicycle paths when networks are grown.

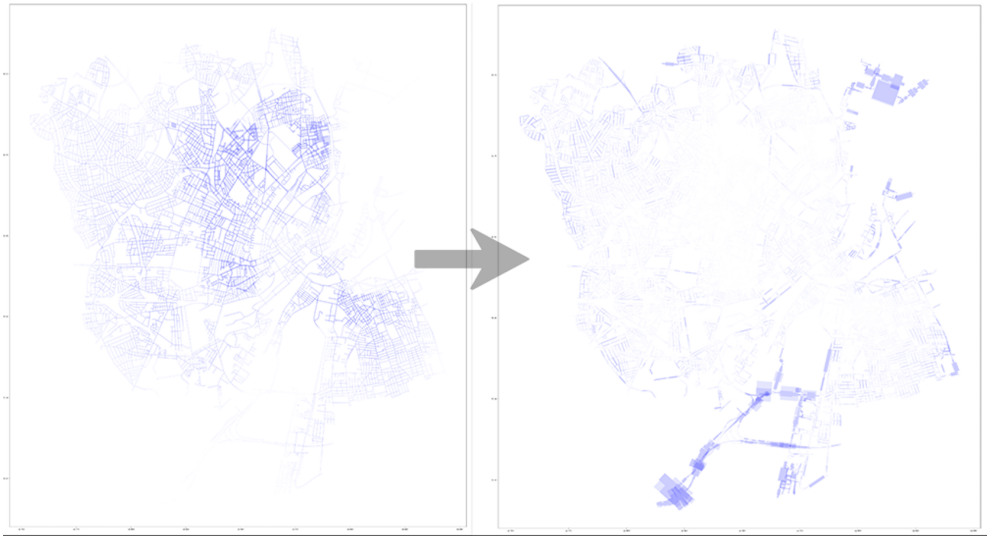


Figure 18: To the left: population density from 2019 visualized on the street network of Copenhagen. To the right: weighted distance measure d_W with $\alpha = 0$ (meaning that the trade-off lies solely on population density) applied to edges. Edges with a high value are enhanced.

4.1.3 The effects of weighted distances on edge betweenness-centrality

Edge betweenness-centrality is derived from an all-pair shortest paths algorithm. The calculations involve computing shortest paths between all pairs of nodes in a graph and counting how many times an edge appears on a shortest path. Vybornova (2021) shows the basic formula for edge betweenness-centrality of an edge/link l, as the fraction of two values; the amount of all shortest paths between all possible node pairs (A, B) in the network, and $\sigma_l(A, B)$ the number of all shortest paths that contain the edge/link l:

$$C_B(l) = \sum_{A,B} \frac{\sigma_l(A, B)}{\sigma(A, B)}$$

Figure 19 shows the edge betweenness-centrality visualized on the street network of Copenhagen, where the shortest paths are calculated based on the length of each road segment (Szell et al. (2022)). Edges with a higher betweenness-centrality are enhanced. Note how Langebro, which continues into Åboulevarden, Is especially highlighted.

Figure 19: Edge betweenness-centrality of the street network in Copenhagen. In this case, the length of each road segment is considered for shortest path calculations (Szell et al. (2022)) inside the formula for edge betweenness-centrality. Edges with a high betweenness-centrality are enhanced.

For weighted graphs, which I aim to build, the formula for edge betweenness-centrality is essentially the same. However, shortest paths are calculated differently. Instead of solely distance, the weighted distance measure $d_W(A, B)$ is applied as an edge weight for calculating shortest paths inside the formula.

Our weighted distance measure is data-aware, implying that calculations of edge-betweenness-centrality will also be data-aware. Applying weighted distance measures will affect betweenness-centrality, as calculations of shortest paths will change. As shown in the right side of figure 18, these extreme outliers in, e.g., top right of the plot, which have a low betweenness-centrality, now most likely will have an even lower betweenness-centrality after the application of weighted distances.

Figure 20 shows the application of weighted distances, on all edges in the Copenhagen street network graph, with respectively $\alpha = 0$ (trade-off solely on population density), $\alpha = 1$ (trade-off solely on bicycle flow) and $\alpha = 0.5$ (equal trade-off between bicycle-flow and population density). Notice a change in edge betweenness-centrality for different α values used as a weighted distance and how certain areas are highlighted.

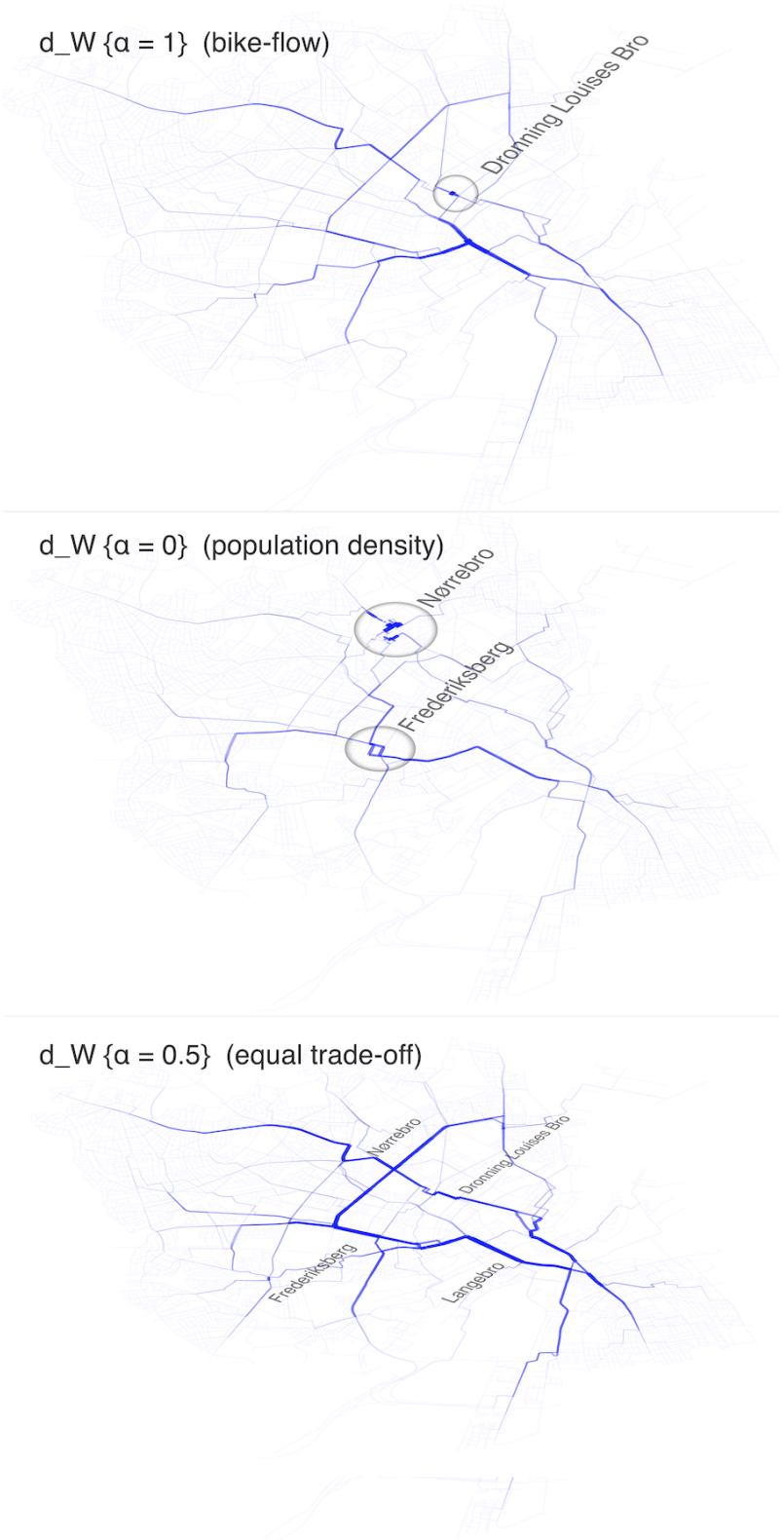


Figure 20: Edge betweenness-centrality visualized for the street network of Copenhagen. The edge weights are defined as a weighted distance d_W for each edge/(link between two nodes A and B), with respectively $\alpha = 0$ (trade-off solely on population density), $\alpha = 1$ (trade-off solely on bicycle flow) and $\alpha = 0.5$ (equal trade-off between bicycle-flow and population density)

When assigning α to 1, road sections such as Dronning Louises Bro show a higher betweenness-centrality. Indeed, this is known to be the World's busiest bike lane.

When assigning α to 0, especially road sections in the Nørrebro area show a high betweenness-centrality. Indeed, Nørrebro is known for the highest population density in Copenhagen. Also, edges around Frederiksberg show a high betweenness-centrality due to a known high population density. Langebro shows a slightly lower population density, which is naturally due to the bridge being uninhabitable.

When assigning α to 0.5, all areas and road sections mentioned show a high betweenness-centrality. This is due to the equal trade-off between population density and bicycle flow.

Further, notice that the distribution of betweenness-centralities appears much more uniform in figure 19, where betweenness-centrality is calculated solely on distances, than for e.g., $\alpha = 0.5$. This might indicate that the trade-off between the length of road sections and our chosen attributes could be further studied.

4.1.4 Growing networks based on data-aware betweenness centrality and weighted distance measures (Step 8)

We have downloaded and preprocessed both spatial information and empirical place-specific data. All data is embedded in a metric space, and our weighted distance algorithm shows how edge-betweenness centrality depends on the trade-off between bicycle flow and population density. This measure is now ready to be utilized for growing networks. The process of growing a network is explained in section 2.1.4, and figure 1 by Szell et al. (2022) illustrates the basic growth process. Instead of growing networks based on different growth strategies, I will grow networks based on three additional edge attributes, with data-aware betweenness-centrality as a through-going growth strategy. The first network is similar to the network grown by Szell et al. (2022) and is grown from edges being weighed by length (the length of a road section). The remaining three networks are grown from edges being weighed with a customized weighted distance measure $d_W(A, B)$, respectively with $\alpha = 0$, $\alpha = 1$, and $\alpha = 0.5$. Figure 21 illustrates the greedy triangulation for each edge attribute based on the seed points. Figure 22 shows how these are routed on the street network. In both figures, five different growth stages for each network are illustrated to visualize the process. The growth process differs for each network, as betweenness centrality, depending on the α value, will also vary, as depicted in figure 20. This will be the basis of our analysis and evaluation, which I will be explaining in the next section.

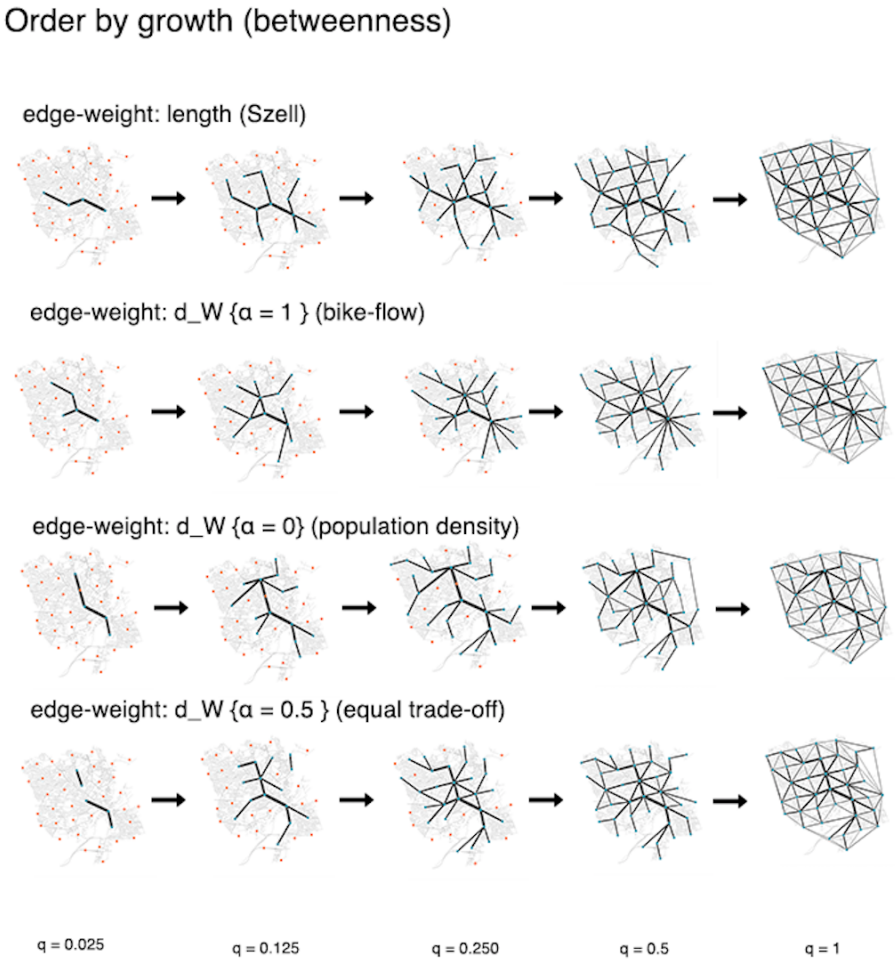


Figure 21: Order by growth and the process of greedy triangulation using betweenness-centrality for all grown networks. 5 (out of 40) different growth quantiles are chosen for visualization. The edge weight 'length' is our baseline (the network grown by Szell et al. (2022)). The weighted distance formula $d_W(A,B) = \alpha \times d_{\text{bicycle-flow}} + (1-\alpha) \times d_{\text{population-density}}$ is used to calculate edge weights for the proposed networks with three different alpha values: $\alpha = 0$ (trade-off solely on population density), $\alpha = 1$ (trade-off solely on bicycle-flow) and $\alpha = 0.5$ (equal trade-off between bicycle-flow and population density)

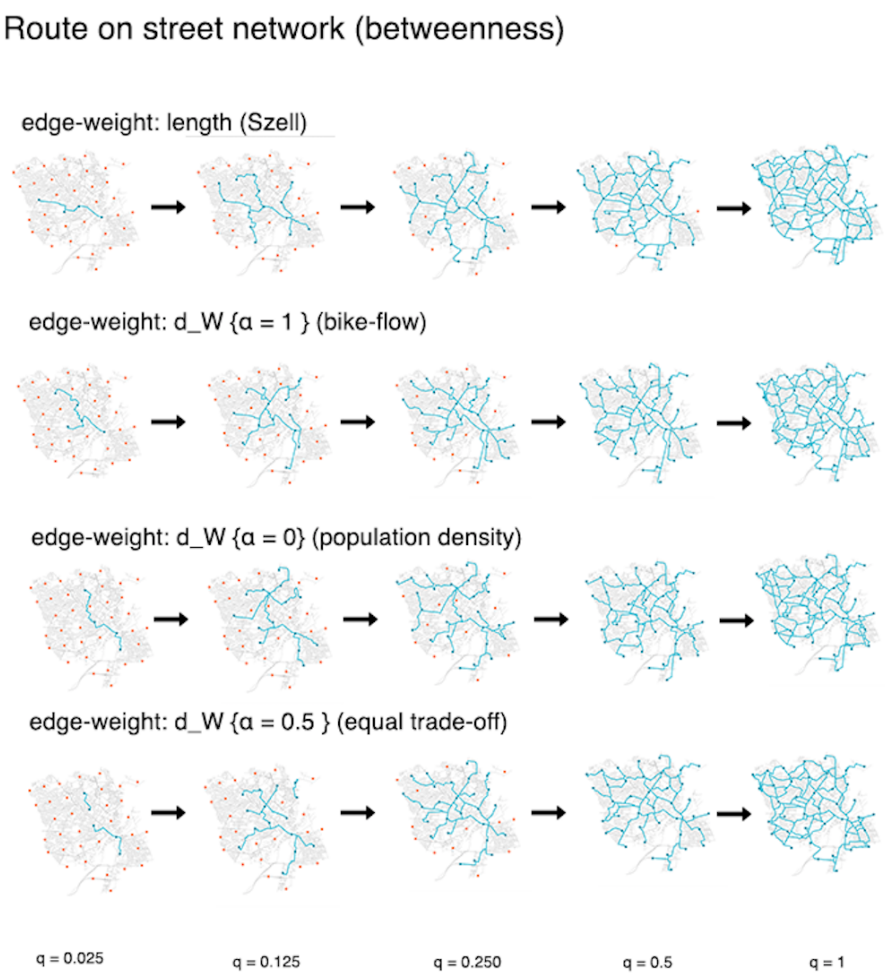


Figure 22: Routing of bicycle infrastructure, using betweenness-centrality, for all grown networks. 5 (out of 40) different growth quantiles are chosen for visualization. The edge weight 'length' is our baseline (the network grown by Szell et al. (2022)). The weighted distance formula $d_W(A,B) = \alpha \times d_{\text{bicycle-flow}} + (1-\alpha) \times d_{\text{population-density}}$ is used to calculate edge weights for the proposed networks with three different alpha values: $\alpha = 0$ (trade-off solely on population density), $\alpha = 1$ (trade-off solely on bicycle-flow) and $\alpha = 0.5$ (equal trade-off between bicycle-flow and population density)

4.2 BICYCLE ANALYSIS: POPULATION DENSITY AS A PROXY FOR GENERALIZED TRAFFIC FLOW

4.2.1 Network evaluation metrics

We will explore how population density can be used as a proxy for traffic flow by growing bicycle networks with data-aware betweenness-centrality as a growth strategy. These are represented as graphs, with edge weights defined as a weighted distance measure determined by the trade-off between population density and bicycle flow.

To compare these new networks with the networks grown by Szell et al. (2022), we need to be able to extract the same information and apply the same evaluation metrics on all networks. For this to be possible, changes are made in the way that we convert the Copenhagen street network graph (a NetworkX graph) to an Igraph.

The function `csv_to_ig()` converts these CSV files generated in OSMnx, holding all edge data, into Igraphs. Adjustments are made inside this particular function to enable graph generation from customized edge weight measures (the weighted distance measure in this case). For these networks to be analyzed and compared with the baseline (Szell et al. (2022)), this function was also customized to store information about the edge lengths as well as the edge-population densities to analyze how accessible these networks are to the population of Copenhagen (see sections 4.2.1 for analysis description).

All networks are analyzed and evaluated on different evaluation metrics (see section 2.1.4 for the definition of these metrics). First, we will analyze directness, which Szell et al. (2022) argue to be the most important metric for bicycle planning, as well as global efficiency, meaning; how well information gets transferred in the network. Then we will analyze how these networks overlap with existing bicycle and bikeable infrastructure. We will calculate the overall length of the network in kilometers, the area covered by the network in square kilometers, and POIs coverage, meaning how fast seed points are covered. Finally, we will define our custom metric, 'Population Coverage', to examine how accessible these networks are to the population of Copenhagen.

Exploring population coverage

For each network, both edge lengths and edge population densities are stored. By multiplying the length of the edge (in meters) and the population density of the area in which the edge is placed (population per square kilometer), we can obtain a measure of the coverage of an edge. By summing this value for all edges in a grown network, we can obtain a score of how well this network covers the population of Copenhagen.

As networks are grown based on weighted distance measures, and this trade-off varies for each network, we can examine how population density can be a proxy for generalized bicycle flow.

Further note that instead of multiplying edge length with population density, it would have been more optimal to include a way of obtaining population densities by extracting them from the shape (Polygon) covered by each edge. This calls for further studies.

4.2.2 Analysis visualization

The results of the networks will be visualized in one figure, where we can compare all evaluation metrics for each network in each plot. The results will be visualized as line plots for all 40 growth quantiles. From these results, we can reflect on which network is optimal at each growth stage.

5 RESULTS

5.1 PATTERNS

In this section, we will look at some observed patterns from network growth plots, data plots, edge-betweenness-centrality plots, as well as findings from generalizing bicycle flow.

5.1.1 Densely populated areas showing high bicycle flow

By downloading and processing spatial information from OSMnx, and acquiring and preprocessing place-specific empirical data, we can observe a general pattern in highly-populated areas being very close to areas with a high bicycle flow, as depicted in figure 14.

5.1.2 A relative improvement of generalizing bicycle flow

The spatial information and the empirical data are embedded in a metric space in graphs, where population densities and bicycle counts are assigned to edges. Generalizations have been applied to account for data spareness so that each edge in the Copenhagen street network graph is informed. We observe how generalizations about bicycle flow have shown an increase from covering only 0.02% to covering 15% of the network graph. However, there is still a uniform pattern among these, as the remaining 85% of edges are just assigned the mean of bicycle counts.

5.1.3 A connection between bicycle network routing, edge-betweenness centrality, and street-level knowledge

The addition of empirical data has allowed for the assignment of customized weighted distance measures to edges. These measures are defined for each edge as a trade-off between bicycle flow, population density, and the length (length of a road segment). When calculating edge-betweenness-centrality (considered as a proxy for flow) with these optimized measures, we see a pattern in the betweenness centrality being affected by the given trade-off. This is visualized in figure 20. The pattern we see is a high correlation between street-level knowledge about bicycle flow/population densities, and the betweenness-centrality; e.g., when the trade-off lies on bicycle flow, sections such as Dronning Louises Bro (the World's busiest bike lane) show a high betweenness-centrality, and likewise, when the trade-off lies on population density, places in the Nørrebro area (highest populated area in Copenhagen) show a high betweenness-centrality.

All weighted bicycle-networks are grown with our data-aware betweenness-centrality as a growth strategy. Each growth stage is visualized as depicted in figure 22. From these grown networks, we can observe a pattern in how networks are routed and the edge-betweenness-centrality in the street network for each weighted distance measure. Notice, e.g., in figure 23 how Dronning Louises Bro and a large part of Nørrebrogade get prioritized for routing when the trade-off is on bicycle-flow, compared to the network grown by Szell et al. (2022).

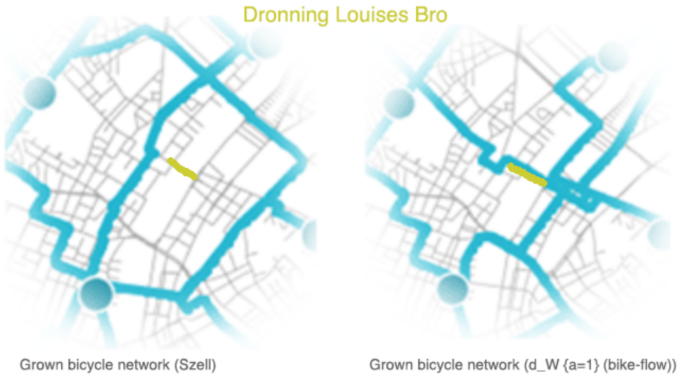


Figure 23: Excerpt from 2 grown bicycle networks. To the left: the network grown by Szell et al. (2022). To the right: weighted distance network where the trade-off lies on bicycle flow. Dronning Louises Bro (the World's busiest bike lane) is marked with yellow and chosen for routing when prioritizing bicycle flow.

We can also observe that when the trade-off is on population density (how accessible these networks are to the population), Sjællandsbroen is not chosen for routing. We know from figure 11 that this area is one of the least populated in Copenhagen, but does this mean that it is a good idea to disconnect the whole Amager Vest region from the southern region of Copenhagen? Would it improve connectivity if this section had a slightly higher betweenness centrality? This calls for further studies.

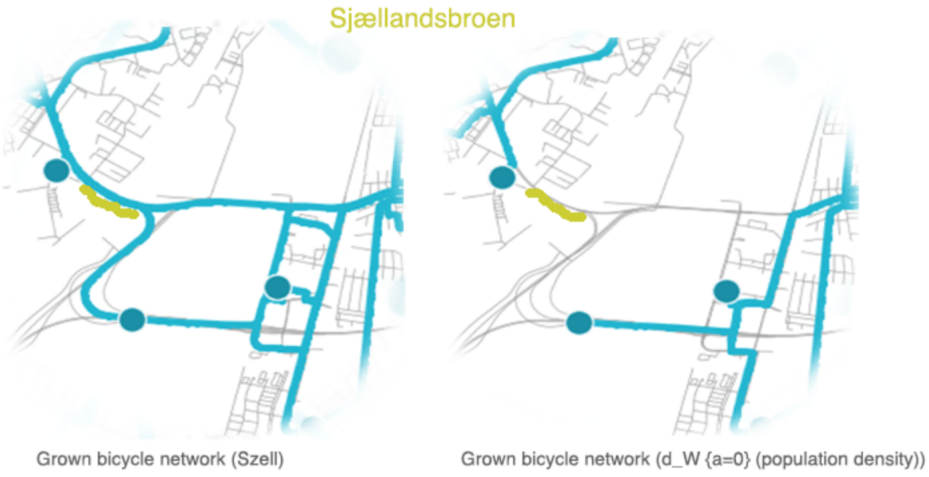


Figure 24: Excerpt from 2 grown bicycle networks. To the left: the network grown by Szell et al. (2022). To the right: weighted distance network where the trade-off lies on population density. Sjællandsbroen (connecting the whole Amager Vest region and the Southern part of Copenhagen) is marked with yellow and not chosen for routing when population density is solely prioritized.

5.2 RESULTS

From data analysis of the grown networks and visualizations of data-aware edge-betweenness-centrality, we can visually observe several patterns. We can furthermore analyze the quality of these networks from a set of evaluation metrics from which we can compute and plot results for comparison. The evaluation metrics are described in section 4.2.1, and figure 25 illustrates the analysis results from the four grown networks.

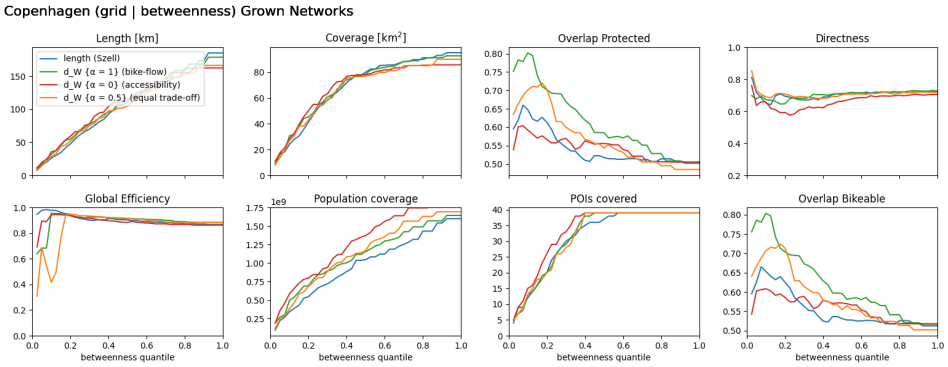


Figure 25: Analysis results of chosen metrics of all grown networks, illustrating the stepwise analysis for each growth quantile in the network growth process.

Basic observations from this plot show that all networks are similar in length and coverage until the final growth stages, where the networks that prioritize accessibility appear to win on being the shortest while having the lowest coverage.

All networks grown with a weighted distance measure appear to cover all points of interest (seed points) faster and have a high population coverage.

From observing directness, we see that the networks seem to differ significantly in the first growth stages but then seem to flatten out. Prioritizing only accessibility shows a relatively low directness until 2/3 of the way through the growth process. At the same time, it appears that prioritizing bicycle flow and accessibility equally shows the highest directness until the last growth stages.

From networks that prioritize bicycle flow, we see a high overlap with existing protected and bikeable infrastructure in almost the first half of the growth process, especially in the network having the trade-off solely on bicycle flow, showing a through-going high overlap.

For global efficiency, the networks differ significantly in the first stages of the growth process and then quickly flatten out with similar values. The initial network (Szell et al. (2022)) generally has a high global efficiency. We notice radical changes in the network with an equal trade-off between bicycle flow and population density, showing a sudden dip followed by a quick increase.

From observing bicycle network plots for all four networks and observing the analysis results, and with our understanding of network science, we can make the following claims:

5.2.1 Prioritizing bicycle flow for better accessibility and higher overlaps with existing bicycle infrastructure

Compared to the network grown by Szell et al. (2022), we see that prioritizing solely bicycle-flow shows improvements in population coverage almost all the way through the growth stages. This points to a connection between areas with a lot of bicycle traffic and areas with a high population density, indicating that population density can indeed be used as a proxy for this generalized bicycle flow. We also see

a very high overlap between this network and existing bicycle and bikeable infrastructure, especially in the first growth stages. This indicates that the Copenhagen bicycle network, which is known to be very well developed compared to many other cities in the world, might have been built with traffic flow in mind.

5.2.2 Prioritizing accessibility for optimal population coverage at the cost of directness and connectivity

Compared to the network grown by Szell et al. (2022), we see that prioritizing accessibility solely, shows the most significant improvements in population coverage; this bicycle network is most accessible to the population living in Copenhagen. This might appear trivial as this was indeed the priority for growth. One could also reflect on whether an accessible network where people don't have to make any hazardous crossings of roads with a lot of car traffic to reach the bike lane could improve safety? could population density be used as a proxy for safety? This calls for further studies.

For many growth stages, this network shows a lower directness (argued by Szell et al. (2022) as the most important metric), indicating that building this network would require some patience, as improvements in directness are first to be seen in later growth stages. From looking at the figure 24, and the disconnection between Amager Vest and the southern part of Copenhagen, we might also reflect on the connectivity of this network and whether other sources of street-level knowledge could contribute to optimize this connectivity.

5.2.3 Prioritizing both bicycle flow and accessibility, and persistent investment, for optimal bicycle networks

Compared to the network grown by Szell et al. (2022), prioritizing both accessibility and bicycle flow shows a high population coverage and good directness for many growth stages. It also shows a high overlap with existing protected bicycle and bikeable infrastructure halfway through the growth process but then takes a dip at the finish line, indicating an alternative choice of routes. For global efficiency, an important notice is a sudden dip followed by a radical increase in the first growth quantiles. This growth process looks like the betweenness-centrality network grown by Szell et al. (2022) for the city of Boston. Szell et al. (2022) argue that this curve for global efficiency has important implications for policy and planning, as cities need to surpass this critical threshold to see gains in global efficiency, meaning; cities need to invest persistently in these networks to see improvements. Szell et al. (2022) refer to this sudden increase, in graph-theoretical terms, as the emergence of a well-connected giant component. Compared to the network grown by Szell et al. (2022), having a through-going high global efficiency, we see these sudden increases for all networks grown with weighted distances. This indicates that for these networks to be optimal, there would be a need to surpass a critical threshold in the growth process.

6 CONCLUSION

In these studies, I have been researching data-driven and data-agnostic methods for growing urban infrastructure and exploring how population density can be used as a proxy for generalized traffic flow in Copenhagen. It appears that spatial data about network structure is a key element for network growth, and to this, we see many different ways that data-driven methods are being utilized. One of these data-driven methods is associating place-specific empirical data with spatial data. We have encountered that place-specific empirical data, such as bicycle flow data, can vary in accessibility, it can be sparse and scattered, and measurements can sometimes even be determined by socio-economic factors, depending on the city. I knew from the start that the bicycle-flow data I would be working with was very sparse, and a significant part of my research has been researching how to account for this data spareness in traffic networks. Among my findings were percolation analysis and machine-learning techniques, such as the state-of-the-art LSM model, combining spatial and temporal data for traffic prediction. These techniques were, however, not possible to utilize as my data only holds one single measurement for each location per year. I also knew from the start that if I were to combine spatial information and place-specific empirical data, some calculations had to be made to express a relationship between these entities. I discovered weighted distance measurements, as defined by Folco et al. (2022), which I modified to fit my case.

My network growth model is based on the findings and the framework provided by Szell et al. (2022), which took care of the initial steps; downloading and processing existing network data from OpenStreetMap (OSM) and embedding this data in a metric space using OSMnx. Seed points (used for triangulation) are generated and implemented as nodes in planar graphs.

I contributed by preprocessing population density data and bicycle count data as edge weights in the graphs. Spatial information was used to generalize bicycle flow data to account for data spareness; the mean of bicycle counts for each street was associated to the rest of the street (the remaining unassigned edges in the graph holding the same street name). The population density counts described population density per square kilometer and could simply be assigned to the nearest edges in the graph based on a haversine distance threshold.

After acquiring and preprocessing both spatial information and place-specific empirical data and embedding this in a metric space, a graph with three independent edge weights (length, population density, bicycle flow) was built. Having these additional edge weights would allow me to understand and visualize where in the street network roads are more and less populated or have a high traffic flow. Still, there was a need for a more customized measure combining these three data sources to grow a meaningful network.

A significant part of the process has been designing intelligent edge weight measures, as described by Folco et al. (2022), that both take into account spatial information, such as the length of a road segment, and the place-specific empirical data embedded in the street network as edge weights. I discovered how weighted distance measures can be assigned to the edges in the graphs from a mathematical optimization formula. This formula allowed for balancing the trade-off between the edge length (the length of a given road segment), the edge population density (the

population per square kilometer assigned to an edge and its neighboring edges), and the edge bicycle flow (the average amount of bicycles passing through an edge in 24 hours). From adjusting the trade-off in this formula, it became possible to adjust how we wish to prioritize connecting the areas with the most bicycle traffic or growing networks that are most accessible to the people living in the city.

The framework provided by Szell et al. (2022) also consisted of an analysis part where all networks are analyzed and evaluated based on a set of quality metrics. I re-used many of these quality metrics, adapting the framework to fit my weighted networks to be able to compare my grown networks to Szell et al. (2022). I contributed to this analysis by defining a customized metric to evaluate how accessible these networks are to the people living in Copenhagen.

All networks are grown with betweenness centrality as a growth strategy. I discovered from my background research that several methodical considerations have to be made when using betweenness-centrality as a proxy for traffic flow. As my grown networks are weighted, having road distances defined as a customized spatial and data-driven distance measure, it became possible to make customized data-aware edge-betweenness-centrality calculations. I calculated and visualized this data-driven edge-betweenness-centrality from my optimized weighted distance measurements, and from the visualizations, we see that depending on the trade-off (how bicycle traffic flow and population density are balanced in the formula), there is a very high connection between street-level knowledge about bicycle flow and population density, and the betweenness-centrality for a given road sections. In the routing of these resulting networks, we can visually observe how growing networks with these different trade-offs between population density and bicycle flow shows similar connections to this street-level knowledge.

From analyzing the quality of the networks, we see several patterns. Prioritizing bicycle flow solely, shows improvements in accessibility, indicating a connection between areas with a high traffic flow and densely populated areas. This could motivate us to use population density as a proxy for generalized bicycle flow.

We notice a radical dip followed by an increase in global efficiency for our grown networks, indicating the emergence of a giant well-connected component, meaning; if these networks were being built in Copenhagen, there would be a need to surpass a critical threshold to see actual benefits. In general, we see a high overlap between the grown networks and the existing bicycle infrastructure in Copenhagen, as we would also expect from using betweenness-centrality as a growth strategy in a city with already well-developed bicycle infrastructure Szell et al. (2022). We see that prioritizing population density solely, increases the accessibility of the network but, unfortunately, at the cost of directness (argued by Szell et al. (2022) as the most important metric) and perhaps also connectivity, as vital areas are suddenly disconnected. We could also reflect on whether accessibility could be used as a proxy for safety, as cyclists could maybe avoid hazardous crossings of streets with a lot of car traffic in order to reach the bike lane? With deductive reasoning by the transitive property of logical consequence, if prioritizing bicycle flow implies better accessibility and prioritizing accessibility implies better safety, bicycle flow could potentially be used as a proxy for safety. This calls for further studies.

This research was limited by the low accessibility of bicycle flow data for the city of Copenhagen, and some assumptions had to be made. Even though our generalizations of bicycle flow showed a significant increase in assigned edges (an increase from 0.02% to 15%), 85% of edges were assigned the mean of bicycle counts. We could assume that many counters are placed in areas with a high

traffic flow, resulting in a relatively high mean value which is most likely not very representable for small streets outside the city center. A deeper understanding of geoprocessing analysis would also improve our customized analysis metric for population coverage as these would ideally be calculated from polygons, allowing for similar coverage plots as made by Szell et al. (2022) but with a modified measure. We also observed in the plots that the distribution of betweenness centralities appeared more extreme with weighted distance measures, pointing towards further studies in how the length of road segments and the empirical data is being weighed inside the weighted distance measure. Perhaps improvements could also be seen in the distribution if applying a cut-off radius for shortest paths, as described by Vybornova (2021). The findings provided by Szell et al. (2022) propose a global analysis, and with this being the skeleton of the network growth being proposed in this thesis, ideally, an automated way of obtaining bicycle flow data and population density data from cities worldwide would open up for easily making this analysis global. This calls for further studies (the day that this data becomes better accessible).

BIBLIOGRAPHY

Banister, D. (2005, 06). Unsustainable transport: City transport in the new century. *Unsustainable Transport: City Transport in the New Century*, 1–292.

BIKO (2019). Biko app. Biking app, https://bikoapp.com/.

Boeing, G. (2017). Osmnx. https://github.com/gboeing/osmnx.

Boisjoly, G., U. Lachapelle, and A. El-Geneidy (2019, 04). Bicycle network performance: Assessing the directness of bicycle facilities through connectivity measures, a montreal, canada case study. *International Journal of Sustainable Transportation 14*, 1–15.

Burrough, P. and A. Frank (1995, 03). Concepts and paradigms in spatial information: Are current geographical information systems truly generic? *International Journal of Geographical Information Systems 9*, 101–116.

Commission, U. N. B. (1987). Report of the world commission on environment and development: Our common future.

copenhagensize (2011). The life-sized city blog: Bicycle rush hour copenhagen.

cycling magazine, C. (2022). More than 60 per cent of people in copenhagen commute to work or school by bike.

data.humdata.org (2019). Worldpop population density for denmark. data retrieved in december 2022, https://data.humdata.org/dataset/worldpop-population-density-for-denmark?

Deng, D., C. Shahabi, U. Demiryurek, L. Zhu, R. Yu, and Y. Liu (2016). Latent space model for road networks to predict time-varying traffic.

Ditte M. Hjorth, S. L. R. K. (2021). Algorithms for data-aware cycling network expansion a data-driven approach.

Dulin, M. F., T. Ludden, H. Tapp, J. A. Blackwell, B. U. de Hernandez, H. A. Smith, and O. J. Furuseth (2010). Using geographic information systems (gis) to understand a community's primary care needs. *The Journal of the American Board of Family Medicine 23*, 13 – 21.

Folco, P., L. Gauvin, M. Tizzoni, and M. Szell (2022, October). Data-driven micromobility network planning for demand and safety. *Environment and Planning B 0*(0), 1–16.

Freeman, L. (1977, 03). A set of measures of centrality based on betweenness. *Sociometry 40*, 35–41.

Gasior, K. (2023). A data driven approach for growing bicycle networks, based on network structure, population density and bicycle traffic flow. https://github.itu.dk/krga/bicycle-networks-thesis.git.

Hu, J., C. Guo, B. Yang, and C. S. Jensen (2019). Stochastic weight completion for road networks using graph convolutional networks. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 1274–1285.

Kazerani, A. and S. Winter (2009, 01). Can betweenness centrality explain traffic flow.

Kommune, K. (2022). The bicycle account 2022 copenhagen city of cyclists.

Larsen, J., Z. Patterson, and A. El-Geneidy (2013, 06). Build it. but where? the use of geographic information systems in identifying locations for new cycling infrastructure. *International Journal of Sustainable Transportation 7*.

Latora, V. and M. Marchiori (2001, 12). Efficient behavior of small-world networks. *Physical review letters 87*, 198701.

Lovelace, R., A. Goodman, R. Aldred, N. Berkoff, A. Abbas, and J. Woodcock (2015, 09). The propensity to cycle tool: An open source online system for sustainable transport planning. *Journal of Transport and Land Use 10*.

Lowry, M. and T. Loh (2016, 12). Quantifying bicycle network connectivity. *Preventive Medicine 95*.

Mattioli, G., C. Roberts, J. Steinberger, and A. Brown (2020, 08). The political economy of car dependence: A systems of provision approach. *Energy Research Social Science 66*, 101486.

Miljøforvaltningen (2021). Mobilitetsredegørelsen 2021.

Nieuwenhuijsen, M., H. Khreis, M. Triguero-Mas, M. Gascon, and P. Dadvand (2016, 08). Fifty shades of green: Pathway to healthy urban living. *Epidemiology 28*, 1.

Olmos, L. E., M. S. Tadeo, D. M. Vlachogiannis, F. Alhasoun, X. E. Alegre, C. Ochoa, F. Targa, and M. C. González (2020). A data science framework for planning the growth of bicycle infrastructures. *Transportation Research Part C-emerging Technologies 115*, 102640.

Opendata.dk (2014). Traffic numbers. data retrieved in december 2021, https://www.opendata.dk/city-of-copenhagen/trafiktal#resource-trafiktal.geojson.

Palominos, N. and D. Smith (2020, 05). Identifying and characterising active travel corridors for london in response to covid-19 using shortest path and streetspace analysis.

Szell, M. (2022). Growing urban bicycle networks. https://github.com/mszell/bikenwgrowth.

Szell, M., S. Mimar, T. Perlman, G. Ghoshal, and R. Sinatra (2022, apr). Growing urban bicycle networks. *Scientific Reports 12*(1).

UNCTAD (2022). Handbook of statistics 2022.

Vybornova, A. (2021). Identifying and classifying gaps in the bicycle network of copenhagen.