University of Padova

Department of Mathematics

Master Thesis in Master in Data Science

Data-Driven Prediction of Missing Links in Padova's Bike Lane Network

Supervisor
Professor Francesco Silvestri
University of Padova

Master Candidate
Alejandra Olivia Cruces Andrews

Academic Year
2023-2024

This thesis is dedicated:

To my husband, for being part of this journey and sharing every step of this adventure.

To my parents and siblings, for their unwavering support and encouragement throughout this experience.

Abstract

Cycling infrastructure is crucial for urban mobility, yet many cities struggle to keep up with the increasing demand for well-connected bike paths. This research, focused on Padova, Italy, proposes a data-driven method to identify and prioritize new bike routes. By integrating data from the Bike Sharing System, RideMovi, and modifying traditional centrality measures, a novel weighted benefit metric is introduced. Grounded in stress centrality, this metric considers both cyclist demand and the strategic importance of network components to prioritize infrastructure improvements effectively.

The results indicate that central areas of Padova, especially around the train station and city center, need substantial upgrades. Key routes connecting the university and hospital are also prioritized, while peripheral areas are considered lower priority. This study provides actionable recommendations for enhancing bike network connectivity in high-demand zones. The analysis reveals that while the Component approach identifies broad infrastructure gaps, the Routing approach offers a more efficient solution by adding fewer kilometers of new paths, better aligning with actual cyclist demand. By combining centrality measures with real-world trip data, the research supports a targeted and effective strategy for urban bike network planning.

Contents

Listing of figures

Listing of tables

Listing of acronyms

BSS .... Bike-Sharing Sytem

LCC .... Largest Connected Component

OSM .... Open Street Map

1
Introduction

The primary objective of this thesis is to tackle the challenges of urban biking infrastructure by developing a data-driven approach to predict and prioritize new bike lanes in Padova, Italy. The research focuses on identifying areas within the city where cycling infrastructure is lacking, despite high demand from cyclists and other users reliant on biking.

To achieve this goal, data from the RideMovi Bike Sharing system is analyzed to identify gaps in the current bike lane network. By examining cycling flow and demand, the study identifies areas where existing infrastructure does not meet user needs. A key contribution of this work is the introduction of a weighted benefit measure, which modifies traditional centrality metrics to incorporate flow demand, helping prioritize which gaps to close to improve network connectivity and efficiency.

In addition to identifying key areas in need of improvement, the study refines the prioritization process by focusing on specific street segments crucial for linking the most important sections of the city. This approach ensures that infrastructure investments are targeted where they will have the greatest impact, enhancing connectivity throughout Padova.

The results show that central areas, particularly around the train station and city center, are in the greatest need of infrastructure upgrades. These are followed by areas near the city center, while zones in the periphery are identified as lower priorities for development.

This introductory section is structured as follows: the initial sections present the context and motivation for this research, emphasizing the importance of a connected bike infrastructure in fostering healthy urban living and addressing environmental concerns. The evolution of public policies supporting cycling infrastructure and the role of shared mobility systems for short trips are also discussed, followed by an overview of the methodology and structure of the research.

1.1 Context

The transportation sector is one of the most challenging areas to address in terms of emission reductions [1]. However, there is substantial potential to mitigate these challenges by "greening" the transportation sector, primarily through a shift towards more sustainable mobility modes such as cycling and walking [2]. In this context, cycling significantly contributes to promoting sustainable and healthy urban living [3].

Additionally, cycling offers multiple benefits for public health, including the promotion of physical activity and the reduction of air pollution [4]. Statistical analyses have shown that cities with higher rates of cycling have lower rates of cardiovascular diseases and respiratory illnesses [5]. Authorities and organization worldwide recognize the importance of cycling, with two main focuses to highlight in the promotion of micromobility: planning transportation networks to include micromobility [6, 7], and promoting the use of shared e-scooters and bikes [8].

In the context of actions taken by authorities and organizations, they seek to enhance cycling infrastructure through improved planning and design of transportation networks. This includes better definitions of bike paths/bike tracks, ensuring safety conditions, and planning future improvements in cities' bike infrastructure [9, 10]. To achieve this, various guidelines and strategies have been proposed to integrate cycling more effectively into urban landscapes. Authorities and organizations provide a range of quantification guidelines and metrics for measuring and guiding bike network developments. Key sources include the Federal Highway Administration (FHWA), the European Cyclists' Federation (ECF), and the National Association of City Transportation Officials (NACTO). These guidelines cover aspects such as network coverage, connectivity, usage, demand, safety, accessibility, quality, and maintenance. Common metrics include total length of bike lanes, bike lane density, connectivity index, cycling mode share, cyclist counts, crash data, surface quality, and maintenance records. Data collection methods involve automated counters, surveys, GPS tracking, and manual counts. Analytical frameworks such as GIS and network analysis are utilized to assess the spatial distribution and structural properties of bicycle networks, while cost-benefit analysis helps to assess the economic impacts of investments in bike infrastructure [11, 12, 13, 14, 15, 16].

The effect of public policies is already reflected in an increase in cycling has been already reflected in the data. In a survey conducted by [17], an increase in cycling was observed in 7 out of 13 countries when comparing the first quarter of 2023 with the first quarter of 2024 (Figure 1.1). The countries with the largest increase in bicycle traffic were Poland, Belgium and Austria.

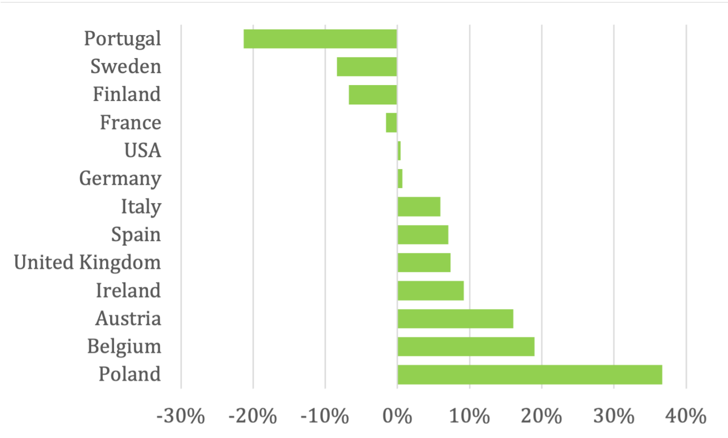


Figure 1.1: Evolution of bicycle traffic Q1 2024 vs. Q1 2023. Source: Eco Counter.

Additionally, there is an increasing preference for using bikes for short trips. In 2022, the Netherlands led Europe with 45% of people using bikes for short trips, followed by Germany (21%) and Belgium (20%). Italy had a 13% share, highlighting its potential for growth. These variations underscore the need to enhance cycling infrastructure across Europe (Figure 1.2).

In the realm of shared micromobility systems, such as e-scooters and bike-sharing systems (BSS), have gained traction as effective solutions for urban transportation challenges. Originating in Amsterdam in 1965 [18], BSS have experienced widespread adoption, offering an alternative mode of transportation for short trips. These systems present several advantages over individual bicycle ownership, including increased accessibility, flexibility, and integration with public transportation networks [18, 9].

The growth of BSS underscores the importance of comprehensive urban planning and policy interventions to support their integration into transportation networks. This includes designing infrastructure that accommodates micromobility vehicles, implementing regulations to ensure safety and accessibility, and promoting multi-modal transportation solutions that encourage cycling and shared micromobility adoption.

In Figure 1.3, we observe that between the third quarter of 2022 and the third quarter of 2023, the ridership of shared bikes in Europe increased, with dockless schemes experiencing the strongest growth at 24%.

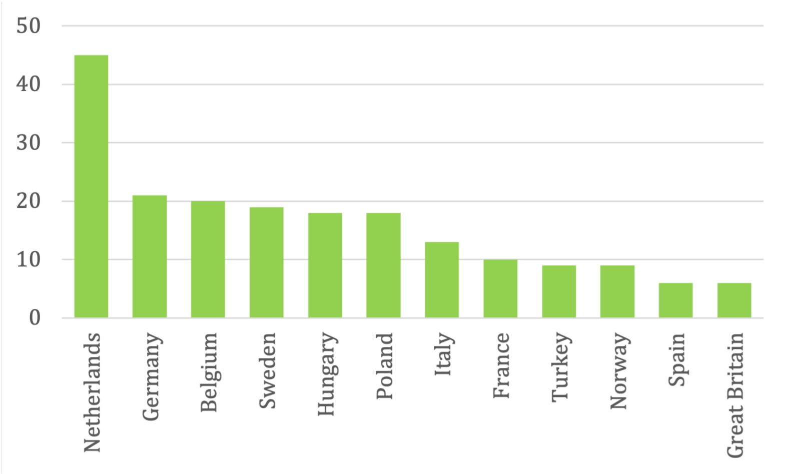


Figure 1.2: Cycling as primary form of transport for short journeys in Europe 2022. Note: Share of population who ride a bicycle as their primary mode of transport for short journeys in Europe in 2022, by selected countries. Europe; March 25 to April 8, 2022; 16 years and older; 20,507 respondents.

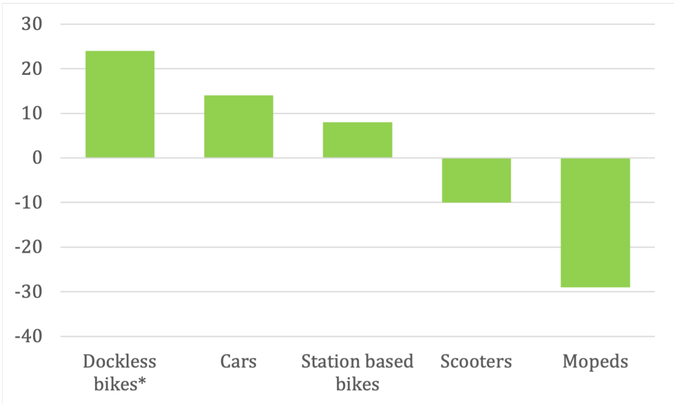


Figure 1.3: European shared mobility ridership growth 2022 to 2023. Note: Shared mobility ridership growth in Europe in Q3 2023 as compared to Q3 2022, by mode. Norway, Switzerland, United Kingdom, EU; Q3 2023; excludes ride-hailing services, car-pooling, and long-term and multi-day hires. Source: Statista.

In the particular case of Italy (Figure 1.4), the number of users in the "Bike-sharing" segment of the shared mobility market has been increasing in recent years and is forecasted to continue rising by a total of 0.4 million users between 2024 and 2028.

However, despite these efforts, there is still a gap in many cities. The development of efficient and interconnected bicycle networks remains a challenge, with many cities struggling to overcome fragmented networks [19]. This underscores the need for comprehensive urban planning and policy interventions to support the integration of cycling into transportation networks.

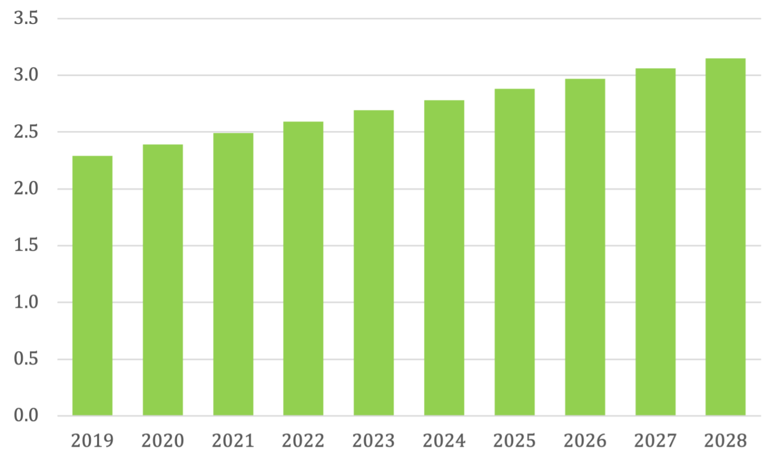


Figure 1.4: Number of users of bike-sharing in Italy from 2019 to 2028 (in millions). Source: Statista.

1.2 Thesis overview and Structure

This thesis focuses on identifying the missing links within the bicycle network of Padova, with the goal of contributing to the development of a more cohesive and user-friendly cycling infrastructure. By addressing the fragmentation in the network, this research aims to predict and fill in these gaps using a data-driven approach. Leveraging data from existing micromobility systems, the study seeks to create a more connected and efficient bike network in the city.

The thesis is structured into four main chapters:

- **Chapter 2**: This section discusses the related work and centrality measures that support this research, offering a foundation for analyzing urban cycling infrastructure. It also introduces the primary data source, the RideMovi dataset, which contains detailed information on bike trips in Padova from July 2020 to September 2023. The dataset includes origin and destination data, making it crucial for constructing the city's flow dynamics.

- **Chapter 3**: This section outlines the methodology used to tackle the problem of enhancing Padova's bike network. It describes the steps involved in constructing the network, transforming trip data into flow information, and modifying centrality measures to incorporate this flow data. The methodology prioritizes zones in need of connectivity improvements through a weighted benefit metric that integrates traffic flow data and centrality measures.

  The approach is divided into key components: identifying disconnected segments in the network, prioritizing these segments using the weighted benefit metric, and conducting refined route analysis within each segment. Two complementary strategies—the Component Approach and the Routing Approach—are applied. The Component Approach identifies critical gaps at a macro level, while the Routing Approach refines the analysis by determining the most efficient routes within each segment. This dual framework ensures comprehensive network planning, combining high-level gap identification with focused, route-specific optimization.

- **Chapter 4**: This chapter presents the study's results, comparing the two methodologies discussed and evaluating the metrics with and without modifications. The findings confirm that the bike network in the city is fragmented, highlighting a significant need for improvements not only in the city center and near the train station but also in other critical areas such as near the university and the hospital. The analysis demonstrates that prioritizing network components based on weighted benefit metrics results in a distinct convergence pattern, underscoring the strategic importance of specific areas for enhancing overall network connectivity and efficiency.

- **Chapter 5**: Concludes the thesis by summarizing the main findings, offering additional remarks, and suggesting directions for future research.

1.3 Specific Perspective of This Work

This research seeks to address the fragmentation of Padova's bike network by predicting and strategically filling in missing links through a data-driven approach. By utilizing data from micromobility sharing systems, the goal is to create a more cohesive, efficient, and user-centric bike network. The specific objectives of the study are to:

- Assess the current state of Padova's bike network, identifying key areas of disconnection and fragmentation.

- Leverage micromobility system data to analyze travel patterns and determine areas of high demand.

- Develop a data-driven methodology to identify the most critical missing links in the bike network.

- Modify and implement a metric that incorporates flow dynamics to better prioritize network improvements.

- Provide actionable recommendations for prioritizing the most critical areas for infrastructure development, focusing on zones with the highest usage and demand.

2
Preliminaries

This chapter establishes the foundational elements necessary for understanding the methodology applied in this research. It begins by reviewing relevant literature, providing insights into both network topology approaches and data-driven strategies. Network topology approaches focus on using centrality measures to evaluate and identify critical network elements, highlighting their role in understanding network structure and identifying key edges. Although these methods are effective in comparing existing networks to theoretical models, they often overlook user behavior and demand patterns.

In contrast, data-driven strategies integrate empirical flow data to address these gaps, emphasizing the importance of incorporating actual usage patterns into network analysis. This approach provides a more detailed understanding of bike lane demand and network performance. The chapter further delves into the technical aspects of centrality measures, such as betweenness and stress centrality, which are pivotal for analyzing network connectivity and edge importance. Additionally, it offers a detailed description of the RideMovi dataset, including data cleaning, preprocessing, and key variables, which form the foundation of the subsequent analysis and methodology.

2.1 Related Work

Addressing the gaps in bike lane networks is essential to fully harness the potential of cycling to promote sustainable urban living. Academic literature has also contributed to improving bike network planning, focusing on constructing efficient and connected bicycle networks from a more quantitative point of view.

Different approaches to constructing optimal networks have been explored, ranging from developing frameworks focusing solely on network topology to incorporating data-driven strategies. These strategies also consider different metrics and information sources, while aligning with guidelines and manuals mentioned previously [11, 12, 13, 14, 15, 16].

2.1.1 Network Topology Approaches

When considering the development of methodologies that focus solely on network topology, the primary goal is to create generalizable approaches using centrality measures of graph networks to identify critical edges and network quality measures to evaluate the evolving network. These methodologies often rely on different centrality measures, such as betweenness and closeness, to guide the construction and improvement of the network.

In this context, [20] present a methodology for strategic network growth from scratch across 62 cities. They explore various growth strategies for greedy triangulation, such as betweenness, closeness, and random approaches, and measure the quality of the resulting network using metrics like length, largest connected component, coverage, and directness. Their objective is to construct the shortest and locally dense planar networks.

Similarly, [21] and [19] focus on identifying critical missing links by considering the existing bike path network as the starting point for network construction.

[19] propose a greedy algorithm applied to 14 world cities, prioritizing bike path construction based on topology. Their approach involves making strategic connections between components by implementing two greedy algorithms that identify the largest connected component in the bicycle infrastructure network and connect it to the second largest or closest component, assessing quality through metrics like connectedness and directness.

On the other hand, [21] developed the IPDC procedure (Identify, Prioritize, Decluster, and Classify) to locate the most crucial missing links in urban bicycle networks. This procedure was applied to Copenhagen but it starts from a generalize framework. It begins by identifying all possible gaps in the bicycle network through a multiplex network approach, calculating the shortest paths between nodes, and discarding parallel paths using a minimum detour factor. These gaps are then prioritized based on flow-based metrics, such as betweenness centrality, which evaluates the centrality of a missing link and the benefits of closing it.

Overall, these methodologies aim to improve network topology by strategically enhancing connectivity and directness. They provide a framework for systematically identifying and addressing gaps in the network, ensuring more efficient and effective bicycle infrastructure development.

The primary advantage of topology-focused methodologies is their ability to compare existing bike networks with optimal ones, offering insights into fundamental topological limitations. Valuable metrics are provided by comparing synthetic networks with real-world examples [20]. The authors offer a comprehensive perspective on the evolution of various indicators, such as connectivity and directness, when prioritizing bike path construction in urban mobility planning.

Additionally, topological methodologies allow for the incorporation of safety measures in bike lane design. Both [21] and [20] acknowledge the importance of ensuring cyclists' well-being by considering existing bike paths in their network construction. This integration helps address safety concerns while developing a comprehensive network.

Despite their advantages, these methodologies have significant limitations. Connectivity alone is insufficient to support demand, as routes along streets equipped with bike paths might be indirect and require large detours. Shortest path trees optimally support demand only from and to a single location, which does not address broader mobility needs [22].

Moreover, these methodologies often overlook critical factors such as user behavior and demand patterns, which are crucial in constructing bike networks [22, 23]. They also frequently disregard existing bicycle infrastructure within specific cities, necessitating adaptation to each unique urban context [21]. Most real-world bicycle infrastructures already have varying degrees of connectivity and accessibility [20]. The main challenge is connecting these fragmented networks, as a general approach that effectively considers this wide range of connectivity and cycling demand is difficult to develop. Ignoring existing infrastructure can lead to suboptimal expansion plans or redundant investments in areas with adequate infrastructure.

2.1.2 Data-Driven Approaches

Although incorporating flow dynamics is crucial, the literature review by [18] highlighted the limited previous research on identifying and analyzing flows. Several pioneering studies have made significant progress by incorporating flow dynamics into their research. For instance, [22] integrated information flow data from bike-sharing systems to understand demand and usage patterns within bike lane networks. They iteratively removed bike paths from an initially complete bike path network, continually updating cyclists' route choices to create a sequence of networks adapted to the cycling demand. They started with a fully connected network, identified current bike paths, and then applied a pruning approach considering demand distribution and a penalty factor if the path was not protected, continuing until all the bike paths had been removed.

Similarly, [24] extended the network growth model by [20] for the city of Turin, Italy, incorporating the existing bicycle network and empirical micromobility datasets of e-scooter trips and bicycle crashes. They used a weighted distance to account for crashes and trips, and calculated a betweenness centrality to prioritize links. The network process was evaluated with two metrics: crash coverage and trip coverage, with a maximum total distance of paths to add.

These studies highlight the benefits of incorporating demand and flow dynamics in bike network planning. For example, [22] obtained insights about differences in bike path coverage density, dictated by demand distribution, allowing for quantitative comparison. They demonstrated that different cities require different optimal investment strategies, with better bike network adaptation when considering flow measures. Additionally, [25] showed that compared to traditional approaches, incorporating traffic flow into a modified weighted betweenness measure produces better results in traffic flow prediction.

However, there are also notable disadvantages. The quality and type of input data, such as street network data and bike-sharing demand, are critical for the resulting networks. Poor data quality can lead to unreliable results, considering challenges in analyzing human-generated data due to its noisy characteristics [26]. Additionally, modeling cyclist behaviors and preferences is complex, and demand is place-specific, making generalization difficult.

Two main challenges arise when incorporating flow/demand in network transportation: modifying centrality measures to appropriately incorporate flow, and obtaining accurate flow data. Flow analysis can be addressed with surveys, GPS tracking, and other data sources. Recent advancements integrate various data sources, including bike-sharing system data, GPS tracking, and user surveys, to inform decision-making processes and develop more effective and user-centric network designs [27, 18, 24].

In this context, [22] utilized flow information from station-based Bike-Sharing Systems (BSS) as a proxy for demand. However, such BSS require users to pick up and drop off bikes at designated stations, potentially limiting the accuracy of demand estimation [27]. Conversely, free systems provide trajectories reflecting actual urban travel flow more accurately.

As the use of BSS has become widespread, data collected from these trips provide valuable insights into bike lane usage and flow dynamics. Various studies have focused on abstracting flow dynamics from Bike Share Systems [27, 18, 22, 24]. This dynamic flow serves as an approach to bike-lane demand, providing significant value in planning and expanding cycling networks, and allowing the identification of routes that would offer the greatest benefits to citizens [19]. In this regard, [22] pioneered the integration of information flow data from bike-sharing systems as a proxy for understanding demand and usage patterns within bike lane networks. In this work, data from Ridemovi, a non-station-based BSS, will be utilized, offering realistic travel demands [27].

2.2 Centrality Measures

In transportation networks, various centrality measures are employed to qualitatively understand the network structure and to assign importance to nodes and edges. This research focuses on two specific measures: betweenness centrality and stress centrality.

Betweenness centrality measures the extent to which an edge lies on the shortest paths between other nodes, thereby indicating its role in facilitating movement across the network. Similarly, stress centrality assesses the frequency with which a node or edge is traversed in all possible shortest paths, providing insight into potential load or congestion within the network.

While these measures are typically applied at the node level, this section presents betweenness centrality for nodes and adapts stress centrality for edge-level analysis. This section introduces the adapted edge-based centrality measure and highlights relevant studies where it has been applied in traffic flow analysis.

2.2.1 Betweenness Centrality

Betweenness centrality is a widely used metric in traffic flow analysis that quantifies the proportion of shortest paths that pass through a particular node [20, 28]. Introduced by Freeman [29], the formal definition for a given graph $G(V, E)$, where V is the set of vertices and E is the set of edges, is expressed as follows. The betweenness centrality $C(v)$ of a node v is defined by:

$$C(v) = \sum_{s,t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2.1}$$

where:

- $C(v)$ is the betweenness centrality of node v.
- σ_{st} is the total number of shortest paths from node $s \in V$ to node $t \in V$.

- $\sigma_{st}(v)$ is the number of those shortest paths that pass through v.

2.2.2 Modified Betweenness Centrality

Incorporating flow dynamics into centrality measures is a key consideration in network analysis. This research utilizes a modified betweenness centrality measure, as proposed by [25], which accounts for traffic flow:

$$C(v) = \sum_{s,t \neq v} \phi_{st} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2.2}$$

where:

- $C(v)$ is the modified betweenness centrality of node v.
- σ_{st} is the total number of shortest paths from node $s \in V$ to node $t \in V$.
- $\sigma_{st}(v)$ is the number of those shortest paths that pass through v.
- ϕ_{st} is the standardized number of trips from node $s \in V$ to node $t \in V$.

The standardized number of trips, denoted ϕ_{st}, ϕ_{ts} is calculated as:

$$\phi_{st} = \phi_{ts} = \frac{M_{st} + M_{ts}}{\max_{s,t \in P}(M_{st} + M_{ts})} \quad \text{for} \quad i \neq j \in V \tag{2.3}$$

where:

- M is the O-D demand distribution matrix.
- M_{st} is the number of trips from node $s \in V$ to node $t \in V$.
- M_{ts} is the number of trips from node $t \in V$ to node $s \in V$.
- i and j are indexes of two vertices (of which the set is V) of topology network.

These formulations serve as the foundation for the metric used in this study, incorporating flow dynamics as outlined in the Methodology chapter.

2.2.3 Stress Centrality

Stress centrality is another measure used to evaluate the importance of edges within a transportation network. For an edge l, the stress centrality is defined as:

$$c(l) = \sum_{i,j \in P} n_{i,j}(l) \tag{2.4}$$

where:

- $c(l)$ is the stress centrality of edge l.
- P is the set of all node pairs.
- $n_{i,j}(l)$ is the number of shortest paths from node $i \in P$ to node $j \in P$ that pass through edge l.

To prioritize edges when predicting missing links, stress centrality has been referred to as Link Closure Benefit by [21], used for one unit of flow.

This formulation will be the foundation for the modified measure proposed in this research, which accounts for multiple units of flow.

2.2.4 O-D Centrality

Origin-Destination (O-D) centrality evaluates the importance of nodes or edges in connecting specific O-D pairs, ensuring efficient travel between key locations. A modified version of stress centrality for an edge l that considers O-D pairs, as outlined by [28], is given by:

$$\text{O-D centrality}_l = \sum_{\substack{i \in I, j \in J \\ d_{ij} \leq \delta}} \sigma^*_{ij}(l) M_i M_j \tag{2.5}$$

In this equation:

- I represents the subset of origins in the network.
- J represents the subset of destinations in the network.
- σ^*_{ij} is the preferred bicycle path from node i to node j.
- M_i and M_j are multipliers for origin i and destination j, respectively.
- d_{ij} is the distance between origin i and destination j.

- δ is the reachable distance threshold for bicycles.

In summary, the centrality measures discussed provide the basis for the methodology developed in this research. Stress Centrality, as defined in Equation 2.4, will be used to assess the qualitative structure of the network. Following the approach outlined by [21], this measure will be termed Link Closure Benefit. These metrics will be adapted to account for flow dynamics, with flow quantified as the standardized number of trips according to Equation 2.3. To facilitate the analysis, the methodology will also incorporate a simplified Origin-Destination (O-D) centrality approach, as described in Equation 2.5.

2.3 Route Estimation Techniques

One main drawback of working with transportation data is the frequent absence of certain variables of interest. Specifically, detailed information about the routes followed by bicycle users is often missing or difficult to access due to privacy concerns, memory constraints, data collection infraestructure restrictions. To address this issue, various routing algorithms are employed to estimate the likely paths taken by cyclists. The most common approach is the shortest path algorithm, which minimizes metrics such as distance, travel time, or other weighted constraints. This method is widely used due to its simplicity and efficiency in predicting user behavior. For instance, [30] discuss different shortest path algorithms and their applications in real-world scenarios, emphasizing their effectiveness in transportation networks. .

In addition to the shortest path approach, more sophisticated methods have been developed to better capture the complexity of route choices. These methods include stochastic routing algorithms that incorporate variability in user preferences and behavior, as well as machine learning models that predict routes based on historical data and contextual factors. [31] illustrate the use of a multi-criteria routing approach that takes into account factors like safety, directness, and cyclist preferences, providing a more nuanced estimation of cyclist routes.

In our study, we employ the shortest path algorithm to estimate the routes followed by cyclists, minimizing the distance metric. This approach allows us to leverage existing network data efficiently while providing a reasonable approximation of actual routes. However, it is important to acknowledge the limitations of this method, as it may not capture all factors influencing route choice, such as road safety, traffic conditions, or personal preferences. Future research could explore integrating more complex routing models to enhance the accuracy of route estimations.

2.4 Data Description

The primary data source for this study is the RideMovi Bike Sharing System data, which includes all cycling trips made with bikes and e-bikes in Padova. This dataset is chosen due to its ability to reflect actual urban travel flow, as it comes from a non-station-based Bike Sharing System (BSS). Unlike station-based BSS, where bikes must be picked up and dropped off at specific locations, RideMovi provides trajectory data that offers a more realistic representation of travel demand. This dynamic flow data is crucial for modifying centrality measures to incorporate flow dynamics and for understanding the true demand for bike lanes.

The following subsection will detail the RideMovi dataset, including the data preprocessing and cleaning steps, as well as the key variables retained for the methodology proposed in this research.

2.4.1 RideMovi Data Description

Ridemovi provides a free-flow bike sharing system featuring "Lite" bicycles and eBikes, available for use 24/7. The bicycles are equipped with advanced technology, including GPS, a SIM card, and a smart lock controlled via the Ridemovi app. Users can access the service by downloading the free Ridemovi app from the App Store or Google Play. To ensure proper usage, there are designated "no parking zones" where bicycles cannot be parked. These zones are visible and marked within the app. Parking in these zones may result in a fine [32].

The RideMovi dataset comprises 1,348,415 rows with 57,573 unique trip IDs, spanning the period from July 2022 to November 2023. It includes details for each trip, such as start and end locations, trip duration, bike types, and other relevant attributes. This dataset encompasses all trips made with bikes and e-bikes in Padova, providing valuable insights into cycling patterns in the city. All dataset information is summarized in Table 2.1. For the methodology, we focus on the variables start_latitude, start_longitude, end_latitude, end_longitude. These variables contain the coordinates where the bikes are picked up (start variables) and dropped off (end variables)

2.4.2 Data cleaning and preprocessing

The data cleaning and preprocessing phase is crucial to ensure the integrity and reliability of the RideMovi dataset for analysis. This section outlines the handling of missing values, geospatial filtering to focus on trips within Padova, and the selection of relevant columns for the study.

Column	Type	Range/Classes
order_id	object	['16588199786175271 5IE12H09333' '16588200756026411 5IE12H04192'] (and 1,345,742 more)
country	object	['Italy']
City	object	['Padova']
bike_no	object	['IE12H09333' 'IE12H04192'] (and 2,855 more)
vehicle_type	object	['ebike' 'bike']
user_id	int64	(15, 1,955,598)
start_time	object	['26/7/2022 9:19:46' '26/7/2022 9:21:24'] (and 1,313,554 more)
end_time	object	['26/7/2022 9:20:13' '26/7/2022 13:01:31'] (and 1,313,364 more)
original_total_amount	int64	(0, 3,075)
original_currency	object	['EUR']
pay_total_amount	int64	(0, 47,597)
pay_amount	int64	(-100, 47,597)
pay_currency	object	[nan 'EUR'] (and 10 more)
owe_amount	int64	(0, 3,075)
promotion_deduction	float64	(-548.0, 47,597.0)
refund_amount	int64	(0, 3,000)
start_latitude	float64	(38.185941, 47.07438)
start_longitude	float64	(9.204108, 15.430411)
end_latitude	float64	(43.710991, 45.908418)
end_longitude	float64	(7.644125, 12.368157)
ride_time	int64	(0, 14,400)
ride_distance	int64	(0, 4,051,100)
soc_end	float64	(1.0, 100.0)
soc_start	float64	(1.0, 100.0)
penalty_order_id	object	['P16588199786175271 5IE12H09333' 'P16588200756026411 5IE12H04192'] (and 1,039,801 more)
penalty_type	object	['PARKING_PHOTO_NOT_TAKEN' nan] (and 3 more)
promotion_channel	float64	(0.0, 10.0)
Pass Group	object	['PAYG' 'Times Pass'] (and 4 more)
start_date	datetime64[ns]	(Timestamp('2022-06-30 19:10:39'), Timestamp('2023-11-22 23:56:45'))
end_date	datetime64[ns]	(Timestamp('2022-07-01 00:00:22'), Timestamp('2023-11-22 23:59:17'))

Table 2.1: RideMovi Data set Information.

Each step in this process refines the dataset, eliminating irrelevant or erroneous data and preparing it for subsequent analysis.

Missing Values

The columns containing missing values and zero values were analyzed, as displayed in Table 2.2. The columns with missing or zero values are not of primary interest to this study, so they will be addressed later in the methodology. Additionally, time gaps in the dataset were checked, and it was confirmed that no gaps are present.

Geospatial Filtering

To ensure the trips are within Padova's municipality, the city boundary from OpenStreetMap was used, defined as a polygon with the following bounding box coordinates:

- **North**: 45.457373
- **South**: 45.339567

Column	Missing Values	Zero Counts
pay_currency	55,352	0
promotion_deduction	55,354	480,670
ride_time	0	22,284
ride_distance	0	26,378
soc_end	663,685	0
soc_start	658,437	0
penalty_order_id	305,942	0
penalty_type	305,942	0
promotion_channel	525,064	234,557

Table 2.2: Columns with Missing and Zero Values.

- **East**: 11.976435
- **West**: 11.805412

Trips with start or end points outside these boundaries were dropped.

Subsetting Columns

For analysis, only the columns of interest for the flow analysis were kept: `order_id`, `start_latitude`, `start_longitude`, `end_latitude`, and `end_longitude`.

Data Cleaning and Filtering

The data cleaning and filtering process involved the following steps:

- Removing duplicate entries: 2,671 rows removed.
- Excluding trips with incorrect coordinates (outside Padova): 3,587 rows removed.

The resulting dataset provides a reliable basis for subsequent analyses, focusing on trips entirely within Padova's administrative boundaries.

The RideMovi dataset, featuring comprehensive data on bike trips in Padova, forms the foundation of this study. This dataset's unique attributes, including its coverage of all trips within the city and detailed trip information, provide a valuable basis for analyzing urban cycling patterns. The subsequent data cleaning and preprocessing phases ensure the dataset's integrity by addressing missing values, geospatial filtering, and column subsetting, thereby preparing it for accurate flow analysis and network evaluation.

3 Methodology

This study aims to enhance the urban biking infrastructure in Padova through a data-driven approach focused on identifying and prioritizing new bike lanes. The methodology is designed to systematically identify and address gaps in the existing bike network by analyzing cycling activity data and applying weighted centrality measures.

To address this challenge, the research introduces a two-tiered approach: Component Approach and Routing Approach. The distinction between the two approaches lies in their scope and detail. The Component Approach focuses on identifying large areas of the city where bike network connectivity is lacking, while the Routing Approach serves as a refinement step that drills down to the micro-level, pinpointing specific streets within each component that require attention. This distinction allows for a comparison between results with and without the application of the Routing Approach, providing insights into the effectiveness of the refined methodology.

Central to this methodology is the Component Approach, which identifies disconnected sub-networks, or components, within the broader bike infrastructure. These components represent areas of the city where bike paths are absent or poorly connected. By analyzing these components, the methodology provides a macro-level overview of the network's connectivity, which is essential for city planners and policymakers. This approach allows for the visualization of the network's structure across different city sectors, highlighting zones with significant connectivity gaps that need attention. The approach is structured into key steps: network creation, flow acquisition, route assignment, gap identification, and gap prioritization.

Once components are identified, the next step involves prioritizing them according to a weighted benefit metric. This metric takes into account factors such as route volume, bike trip frequency, and component length. This prioritization process reveals that smaller, high-traffic components may be more critical to address than larger, less-used ones, emphasizing the value of a data-driven approach.

While the Routing Approach is not strictly necessary, it offers a refinement by focusing on the micro-level connections within each component. After prioritizing components, the Routing Approach is employed to connect "flow nodes" or "contact nodes"—points where transitions occur between bike paths and non-bike paths. A routing algorithm is proposed to optimize these connections, effectively bridging identified gaps within the network. This micro-level routing solution complements the macro-level component analysis by ensuring that the most frequently traveled and critical routes within each component are prioritized.

Finally, the enhanced network's quality is assessed using various metrics to evaluate the effectiveness of the proposed methodology. The primary objective of this research is to identify areas within Padova's bike network that lack adequate infrastructure and to prioritize interventions that will generate the greatest impact for cyclists. By combining the Component and Routing Approaches, the methodology provides a comprehensive framework for improving the city's bike infrastructure in a targeted and efficient manner.

3.1 Component Approach

As previously mentioned, this procedure forms the core of the thesis, serving to identify key components within Padova's bike network that lack adequate infrastructure but are critical for improving overall connectivity in terms of demand. The approach is organized into several key steps: network creation, flow acquisition, route assignment, gap identification, and gap prioritization.

The network creation process starts with obtaining the transportation network, focusing on identifying existing bike paths and potential routes suitable for cyclists. In the flow acquisition and route assignment steps, the process involves identifying key origin and destination points within the city and analyzing flow data to determine which streets and paths are currently used or have the potential to be used by cyclists. This results in a refined subnetwork that focuses exclusively on bike paths and routes most relevant to cycling activity.

Next, the methodology identifies the disconnected components within this subnetwork—specific areas that consist of potential bike paths but are not yet connected to the larger network.

These disconnected components represent zones where infrastructure improvements could significantly enhance network connectivity.

Finally, the prioritization of these components is conducted using a modified centrality measure that incorporates demand-side considerations. Specifically, a flow-weighted stress centrality measure is used to assess the benefit of closing each component, helping to prioritize areas that will offer the greatest impact in terms of improving bike network connectivity and meeting cyclist demand.

3.2 Methodological Framework

The methodology for identifying and prioritizing gaps in Padova's bike network is structured around a series of key steps, each integral to the overall process. This framework leverages both the Component and Routing Approaches to systematically enhance the city's cycling infrastructure by addressing both macro and micro-level connectivity issues.

- **Network Creation:** The first step involves acquiring transportation network data using tools like OSMnx to map out existing and potential bike paths. This phase focuses on constructing a subnetwork that captures the relevant routes for cyclists, forming the basis for subsequent analysis.

- **Flow Acquisition:** This stage involves collecting and analyzing cycling activity data to identify key origin and destination points. Understanding the flow of cyclists is crucial for identifying the most frequently used parts of the network, thereby informing priority areas for infrastructure development.

- **Route Assignment:** This step involves assigning routes within the identified subnetwork based on flow data. The goal is to determine which paths are most frequently used or are likely to be used by cyclists, helping to identify areas where infrastructure improvements are needed.

- **Network Simplification:** After route assignment, the network is simplified to focus on the core paths that are most significant for cyclists. This involves retaining only the existing and potential paths that play a crucial role in the network, ensuring that resources are concentrated where they will have the most impact.

- **Component Identification:** The methodology then identifies disconnected sub-networks, or "components," within the simplified network. These components represent areas with potential bike paths that are not yet integrated into the broader cycling infrastructure, highlighting critical gaps in the network.

- **Component Prioritization:** The identified components are then prioritized using a modified centrality measure that incorporates flow data. By calculating a flow-weighted stress centrality metric, the methodology assesses the benefit of closing each gap. This step focuses on prioritizing components that, when connected, will significantly enhance network efficiency and meet cyclist demand.

In the Routing Approach, an additional step focuses on prioritizing streets within each component:

- **Route Prioritization:** This step involves optimizing the selection of routes to connect "flow nodes" or "contact nodes"—critical transition points between bike paths and non-bike paths within each component. By prioritizing these routes, the methodology ensures that the most heavily used and strategically important paths within each component are effectively connected, further improving the overall network's connectivity and efficiency.

The steps outlined above form the core of the methodological framework for enhancing Padova's bike network. Each step will be explained in detail in the following sections, providing a comprehensive guide to the approach and its implementation.

3.2.1 Network Creation

In this step, the transport infrastructure network is acquired, identifying both current bike paths and potential bike paths.

The transport network creation process involves three key activities: data acquisition, network definition, and integration. These steps are essential for establishing a comprehensive network that will serve as the foundation for analyzing cycling routes in Padova.

Padova's Biking Infraestructure

Padova's biking infrastructure has witnessed significant development in recent years, reflecting the city's dedication to promoting sustainable transportation options. Between 2018 and 2022, Padova expanded its bike lane network from 169 kilometers to 178 kilometers, with the goal of a further 20-kilometer extension by 2024, demonstrating a tangible commitment to improving cycling infrastructure. The city's proactive approach to expanding bike lanes aligns with its goal of fostering sustainable mobility and reducing reliance on traditional modes of transportation. Furthermore, Padova's recognition as the 9th most bikeable city in Italy in the 2015 Rapporto Statistica highlights its commitment to cycling infrastructure.

Despite these advancements, challenges and limitations persist within Padova's biking infrastructure. The current network may still exhibit gaps that hinder optimal connectivity. Identifying these missing links and addressing them effectively is crucial for maximizing the utility of the biking infrastructure and ensuring a safe and convenient cycling experience for residents and visitors alike. Therefore, this study aims to contribute to Padova's objective of enhancing its biking infrastructure by predicting missing links within the existing network and incorporating information on flow.

Network Definition

The initial step involves acquiring map and network data using the Python library OSMnx [33], sourced from OpenStreetMap.

The geographical area of interest is defined as the administrative boundary of the Comune di Padova, Italy. The OSMnx library allows downloading nodes and edges with additional tags specified to enhance the detail of the network data. These tags include attributes related to cycling infrastructure, such as cycleways, bicycle permissions, and road surface types. The settings of the OSMnx library are updated to include these extra tags:

```
['cycleway', 'bicycle', 'cycleway:left', 'foot',
'cycleway:right:segregated', 'tracktype', 'cycleway:lane',
'cycleway:left:lane', 'cycleway:width', 'oneway:bicycle',
'parking:lane:right', 'parking:lane:left', 'cycleway:right:lane',
'cycleway:right', 'cycleway:surface', 'surface', 'parking',
'bus', 'hgv', 'smoothness', 'cycleway:both', 'level',
'cycleway:left:segregated', 'cycleway:both:lane']
```

Then, two separate networks are created:

1. **Drive Network**: This network represents roads accessible to motor vehicles and is obtained by specifying the drive network type. This network includes all roads and street segments used by vehicles.

2. **Bike Network**: This network focuses on paths and roads accessible to bicycles, ensuring that cycling infrastructure is included. The bike network is obtained by specifying the bike network type.

Both networks are obtained by specifying the city location, the type of network, with no simplification, and keeping all the nodes and edges.

Next, a composite network is created by combining the drive and bike networks using the `nx.compose` function from the NetworkX library. This composite network integrates both types of infrastructure, providing a unified graph for route analysis. The order of combining networks is important to maintain the attributes of the nodes and edges. However, the bike network does not accurately identify the current bike paths in Padova, including many paths that are not actual bike paths but can be used as such. For the analysis, it is necessary to correctly identify the current bike paths.

Current Bike Paths

To accurately identify current bike paths, a set of filters is applied to the network data. These filters use additional tags to isolate paths that are specifically designated for cycling. As mentioned earlier, the bike network includes many paths that are bikeable but not currently designated as biketracks or bikepaths. This process identified several extra bike paths that were not considered in the initial bike network but are identifiable in the drive network.

It is important to note that many of the filters may appear similar, but their specificity is necessary due to the potential for missing values or inconsistencies in attribute columns. Identifying these paths often requires cross-referencing multiple tag columns.

- **Filter 1: Cycleway Track**
  This filter identifies segments with cycleways designated by various attributes such as "yes," "track," "share_busway," "opposite_lane," "opposite_share_busway," "advisory_lane," and "opposite." These attributes help in specifying paths that are specifically intended for cycling along roadways.

- **Filter 2: Highway Cycleway**
  This filter targets segments tagged as "cycleway" within the highway attribute. It focuses on detecting dedicated cycling infrastructure along highways, ensuring the identification of safe and efficient routes for cyclists.

- **Filter 3: Designated Bicycle Paths**
  This filter searches for paths where the "bicycle" attribute is designated, focusing on paths specifically marked for bicycle use within the highway infrastructure. It helps in identifying exclusive cycling paths.

- **Filter 4: Cycleway Right Track**
  This filter examines the presence of cycleways on the right side of highways. It identifies segments where the "cycleway:right" attribute includes values like "track," "share_busway," "opposite_lane," "opposite_share_busway," "advisory_lane," and "opposite" .

- **Filter 5: Cycleway Left Track**
  Similar to the previous filter, this one focuses on the left side of highways, identifying seg-

ments where the "cycleway:left" attribute includes values such as "track," "share_busway," "opposite_lane," "opposite_share_busway," "advisory_lane," and "opposite."

- **Filter 6: Cycle Street**
  This filter identifies segments tagged as "cyclestreet," which are streets designed primarily for cycling traffic. It highlights areas where cycling is prioritized, contributing to the overall network analysis.

- **Filter 7: Bicycle Road**
  This filter searches for roads where bicycles are designated or where there are optional or permissible side paths. It focuses on identifying roads specifically marked for bicycle use, enhancing the understanding of the bicycle road network.

- **Filter 8: Living Street**
  This filter targets segments tagged as "living_street" within the highway attribute. It identifies areas where cycling is integrated with living streets, providing insights into cyclist-friendly urban areas.

- **Filter 9: Both Sides Cycleway**
  This filter examines the presence of advisory lanes on both sides of the roadway. By analyzing the "cycleway:both:lane" attribute, it provides insights into the availability of dual-side cycling infrastructure.

- **Filter 10: Tertiary Highway and Pedestrian Paths**
  This filter identifies segments within tertiary highways and pedestrian paths, focusing on areas where cycling is integrated with pedestrian infrastructure. It targets segments tagged as "tertiary," "path," and "pedestrian."

- **Filter 11: Tertiary Highways**
  This filter focuses on tertiary highways, identifying segments where the "highway" attribute is tagged as "tertiary." It provides insights into cycling infrastructure within these types of roads.

- **Filter 12: Pathways**
  This filter targets pathways within the network, identifying segments where the "highway" attribute is tagged as "path."

The network edges were examined using these filters, and those that met the criteria were collected to create a new network of biketracks.

Potential Bike Paths

To identify potential bike paths within the composite network, a series of filters based on various attributes such as highway types, bicycle permissions, and specific cycleway configurations are applied. This selection process ensures that the identified bike paths are suitable for future bike lanes while excluding high-speed highways and streets primarily for buses and trucks. The main process involves only keeping the edges that are not in the current bike paths network and excluding high-speed highways and streets primarily for buses and trucks. The filtered paths were then marked within the network to facilitate further analysis and visualization.

These identified bike paths are consolidated into the edges dataset, and a new attribute, `type`, is set to indicate the presence of cycling infrastructure. This attribute is added to the composite network's edges using the `set_edge_attributes` function from `networkx`. Similarly, nodes are also identified, adding a new attribute to the nodes dataset to indicate those in the potential, bike, and both networks.

Finally, the updated network, now enriched with bike path information, is converted back into GeoDataFrames and Multidigraph for further analysis and visualization.

Figure 3.1 displays the transportation network of Padova and the existing cycling infrastructure (in green), as of May 2024 based on OSM data. While this network offers a comprehensive representation of the cycling infrastructure in Padova, it may not exactly match the current bike network due to potential errors or missing information in the OSM data. Nevertheless, this representation serves as a valuable approach for identifying missing links and proposing a methodology that can be refined in future iterations. The subsequent sections will leverage this network for route analysis and gap identification, providing a foundation for enhancing the city's biking infrastructure.

Graph Representation of the Network

A network can be formally represented as a graph $G = (V, E)$, where V is the set of vertices (nodes) and E is the set of edges (links). In the particular case of the transportation Network we have:

- **Nodes (Intersections)**: Represented as points with geographic coordinates.
- **Edges (Street Segments)**: Represented as sequences of points.

For the network presented in the previous section, a simplification process was needed. For this, the OSMnx library provides a simplify function. This function reduces the complexity of

Figure 3.1: Transportation Network in Padova.

the network by consolidating consecutive nodes and edges that are not actual intersections or endpoints into single edges, thus streamlining the graph.

The network is formed by disconnected components, which are subgraphs where any two nodes are connected to each other by paths and are not connected to any additional nodes in the supergraph. The number of disconnected components in the previous network is 179, with the largest component containing 58,412 nodes (of the total of 59,723 in the main graph). For the analysis, only the subgraph formed by the Largest Connected Component (LCC) was kept, while all other disconnected components were dismissed as negligible for the sake of simplicity. Figure 3.2 represents the network of the Largest Connected Component. In the real street network of the city, disconnected components, i.e., street segments that are not accessible from any other street segment, are quite rare [21].

The network representation in this study combines current bike paths and potential bike paths, ensuring that the final graph G incorporates nodes and edges with appropriate attributes to identify their types and other relevant properties. The final network includes 58,412 nodes and 63,037 edges, fully connected and undirected. The network is characterized by the following attributes for nodes and edges:

- **Node Attributes**:
  - **ID**: Unique identifier for the node.
  - **Coordinates**: Geographic coordinates $(x_v : Longitude, y_v : Latitude)$.
  - **Type**: Type of node (e.g., bike, potential, both).

- **Edge Attributes**:
  - **ID**: Unique identifier for the edge.
  - **Start Node**: The node where the edge begins.
  - **End Node**: The node where the edge ends.
  - **Length**: Length of the street segment.
  - **Type**: Type of edge (e.g., current bike path, potential bike path).

Figure 3.2: Transportation Network in Padova considering the Largest Connected Component.

In parallel, a Pandana network is constructed with a focus on the essential attribute required for calculating the nearest node—the length of street segments. The nearest node calculation is crucial in both flow acquisition and route assignment steps. By prioritizing segment length, the Pandana graph is optimized to facilitate efficient nearest node calculations, thereby enhancing the overall effectiveness of the network analysis.

To ensure consistency between Pandana and NetworkX, the previously loaded OSM graph from OSMnx is utilized as the base for creating the Pandana graph. This ensures that both graphs share the same underlying network structure and attributes.

Geodataframes that contain nodes and lengths serve as the basis for constructing the Pandana graph, providing the necessary spatial and attribute data for accurate node assignment and distance calculations.

3.2.2 Flow Acquisition

Demand-side influences, represented by the number of trips between origin-destination (O-D) pairs, are critical for understanding traffic flows on the bike network. Analyzing these flows provides insights into potential future bike paths and informs the enhancement of cycling infrastructure in Padova. By examining trip data, high-demand routes that could benefit from dedicated bike paths can be identified.

To incorporate demand, the number of trips between O-D pairs is analyzed using data from RideMovi, as introduced in a previous section. The essential variables include the start and end points, defined by their respective latitude and longitude coordinates. The initial step involves mapping these coordinates to the nearest nodes in the network (Figure 3.3). This can be achieved using either the NetworkX or Pandana libraries, both of which offer functions to locate the nearest neighboring node on the graph. Pandana [34] is preferred for its efficiency in calculating the nearest node, providing faster performance. A sample check ensures that the node assignment is consistent between both methods.

After identifying the nearest nodes, trips with the same start and end nodes are aggregated to form an Origin-Destination (O-D) Matrix. The number of trips sharing the same origin and destination is recorded, representing the flow between these locations. To simplify the O-D matrix and reduce computational complexity, trips in reverse directions are collapsed into a single undirectional entry. This approach standardizes the route representation for trips between two locations, irrespective of the direction of travel, thus reducing computational time and memory requirements. Additionally, only trips with a minimum frequency over the year are retained in the analysis. This threshold ensures that the matrix focuses on frequently traveled routes, which are more relevant for infrastructure planning and development.

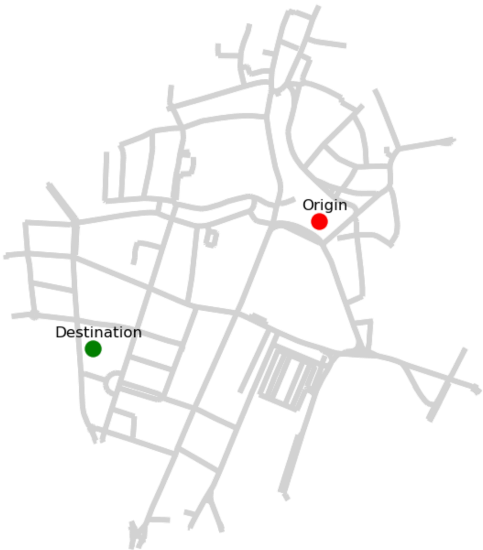


Figure 3.3: Nearest Node Calculation Step. The figure illustrates the process of mapping trip coordinates to the nearest nodes in the network.

3.2.3 Route Assigment

In this section, the process involves calculating the most probable routes for cyclists between each origin-destination pair, aiming to emulate their actual travel paths. This approach aligns with existing literature on route prediction methodologies [30], which commonly assume that cyclists aim to minimize distance or time during their journeys.

To achieve this objective, the shortest path algorithm is employed, considering the entire bikeable network rather than solely the bike track network. The NetworkX library, which provides a function for calculating shortest paths using Dijkstra's algorithm, facilitates this process efficiently [35]. Dijkstra's algorithm operates by iteratively selecting the node with the lowest cumulative cost from a set of candidate nodes, progressively building the shortest path from the origin to the destination [36].

The function returns a list of nodes that make up the shortest path between the given origin and destination points. Subsequently, the associated edges related to these nodes are retrieved. This process iterates for each origin-destination pair, enabling the assignment of routes to each

O-D pair based on the NetworkX shortest path algorithm, as illutrated in Figure3.4 This comprehensive approach ensures that the predicted routes closely reflect the actual travel paths of cyclists, capturing the nuances of the bikeable network.

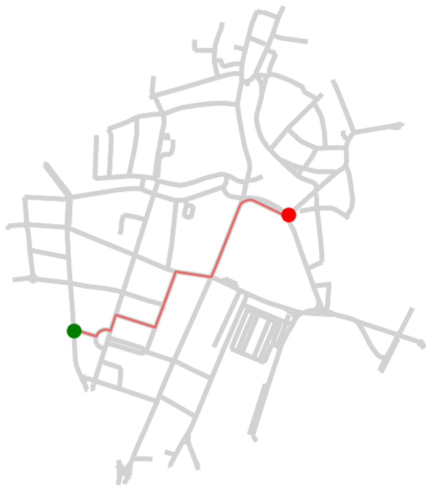


Figure 3.4: Route Assignment Step. The figure illustrates the process of mapping the shortest path.

3.2.4 Network Simplification

The network is simplified by creating the Route Network, which is formed by extracting all edges that belong to designated routes from the Largest Connected Component (LCC) of the original network. This subnetwork includes only the bike paths and potential bike paths that are part of the routes, excluding other, less relevant network edges. The Route Network is crucial for identifying disconnected components within the bike network and streamlines the analysis by focusing on the most pertinent edges. Each edge in the Route Network is assigned attributes such as the number of trips that pass through it and whether it is classified as a current bike path or a potential bike path.

3.2.5 Components Identification

Identifying gaps in the bike transportation network is essential for enhancing connectivity and overall network efficiency. Not all gaps have the same impact, so it's crucial to systematically identify the components—disconnected sub-networks—that contribute to significant connectivity issues. By breaking down the bike network into these components, we can focus on areas that have the potential to substantially improve cyclist experience when connected. This section details the process of identifying these critical components, which will later be evaluated and prioritized for intervention. This methodology helps in reducing computational memory and processing time by focusing on subgraphs of components rather than the entire network.

Identification of Disconnected Components

To identify disconnected components within the bike network, the Route Network created in the previous step is utilized. This network comprises both existing bike paths and potential paths that form part of the designated routes.Initially, components within Padova are identified as groups of edges within routes that lack bike paths. These zones, referred to as "Components," consist solely of potential bike paths and are detected using the connected component in-built function of NetworkX, which identifies connected subgraphs within the network. These areas can be visually identified in Figure 3.5: Step 1, which highlights areas formed exclusively by light gray lines. This process yields a set of nodes representing each component, facilitating the visualization and enumeration of the total number of identified components. This method aligns with approaches used in previous studies, such as [21] and [20], focusing on detecting subgraphs within larger networks. Further refinements are made to eliminate dead-end components.

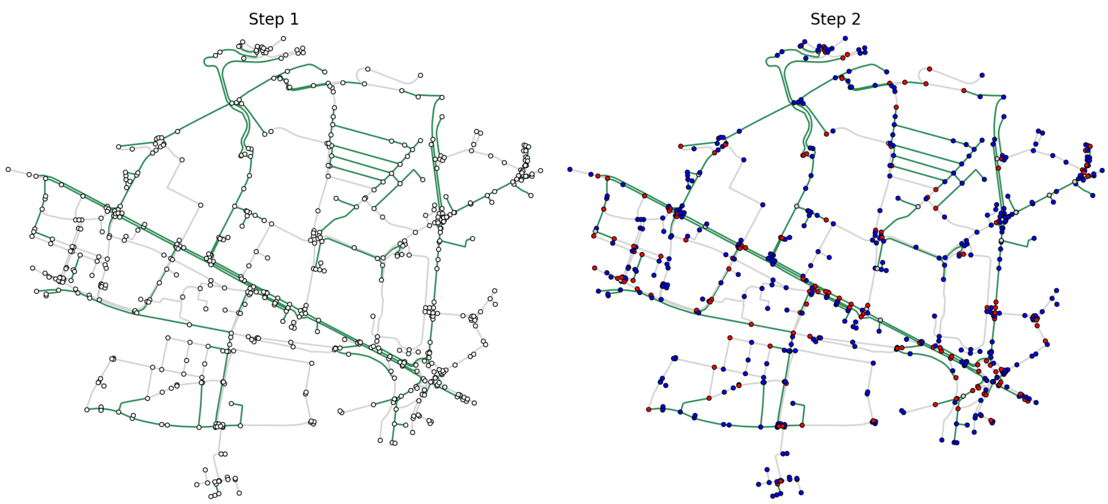
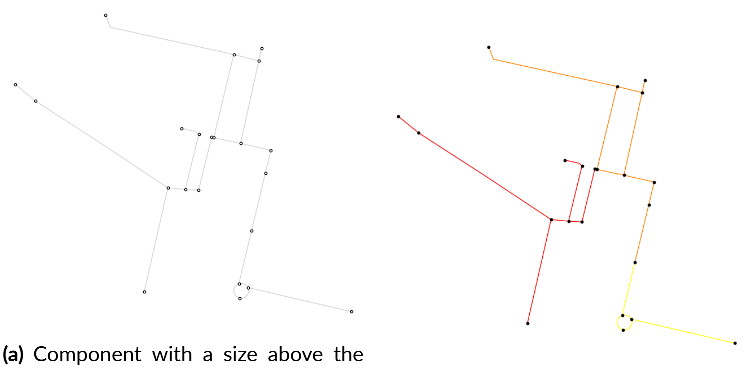


Figure 3.5: Component Identification Step and Classification of nodes. Step 1: Component Identification. Step 2: Classification of nodes. Edges legend: Green lines represent bike paths. Light gray lines represent potential bike paths. Nodes legend: Red dots indicate contact nodes. Blue dots indicate nodes that are part of bike path edges only. Light gray dots indicate nodes that are part of potential edges only.

To improve the granularity of the analysis, especially in large sections of the city that form substantial components, a subdivision approach is employed. This involves dividing these large components into smaller, more manageable sections using a distance-limited exploration technique. The method starts from a node and explores all reachable nodes within a specified maximum distance, known as the cutoff distance. The NetworkX function single source shortest path length facilitates this by generating a set of nodes considered a subgraph. This process is iteratively repeated until all nodes and edges are covered. Key parameters in this approach include the size threshold, which determines the definition of a large component to be subdivided, and the cutoff distance, which limits the exploration range. These hyperparameters significantly impact the network's structure; a higher size threshold results in larger, more complex components, while a greater cutoff distance reduces the number of resulting subcomponents. This proccess is presented in Figure 3.6.

The identification of components enables targeted improvements within specific zones. Components are prioritized based on a strategy that will be detailed later, considering the demand and potential benefits of closing gaps.

For a comprehensive understanding of each component, nodes are categorized into "contact nodes," "bike nodes," and "potential nodes." Contact nodes connect a component with the main bike path network, bike nodes connect multiple bike paths, and potential nodes connect multiple potential paths (Figure 3.5, Step 2). This classification allows for strategic planning and prioritization of network improvements.



(a) Component with a size above the threshold.

(b) Sub-divisions in the component.

(c) Sub-divisions with contact nodes.

Figure 3.6: Subdivision of Large Network Components using a distance-limited exploration technique.

3.2.6 Components Prioritization

Once the key components of the bike transportation network have been identified, the next step is to prioritize them based on their potential impact. Not all gaps are equally important; some, when addressed, can significantly enhance network efficiency and better meet cyclist needs. To quantify the significance of each component, we employ a weighted benefit metric that combines centrality measures with demand flow data, reflecting actual cyclist behavior. This section outlines the prioritization process, focusing on the components that, when closed, will provide the greatest benefit to the cycling network in Padova.

Modified Centrality for Traffic Demand

As discussed in the subsection on Related Work, stress centrality quantifies the importance of an edge based on the number of shortest paths passing through it, highlighting edges that play a crucial role in connecting different parts of the network [37]. In comparison, betweenness centrality also considers shortest paths but weights them by the inverse of their redundancy, offering a more detailed perspective on edge importance [20, 28].

In this study, we adapt the gap benefit measure from [21], following a similar approach to the modified betweenness centrality proposed by [25], which incorporates flow data. To streamline the computation, we also draw on the simplified calculation method used in the origin-destination (O-D) centrality framework by [28]. This section outlines the development of the new metric.

This section presents two key metrics: Weighted Link Closure and Weighted Benefit. These metrics are weighted adaptations of those proposed by [21] for single-unit flow scenarios. The Weighted Link Closure metric assesses the importance of a single edge within the network, while the Weighted Benefit evaluates the benefit of closing a group of edges, or a component, in terms of the expected meters cycled in mixed traffic per unit of investment, accounting for bike traffic demands.

In the Preliminaries chapter, essential centrality measures were introduced, forming the basis of the metrics used in this study. The steps are as follows:

1. **Stress Centrality as Link Closure Benefit:** The foundational metric is stress centrality, represented as link closure benefit in Equation 2.4.

2. **Incorporation of Flow Dynamics:** Flow dynamics are integrated by weighting the number of shortest paths based on the standardized number of trips, using a framework

similar to [25] in Equation 2.2. The standardized trip count is derived similarly to Equation 2.3, but with adjustments to consolidate reverse trips into a single direction.

3. **Weighted Benefit:** This metric accounts for the modified link closure benefit, normalized by the total length of the gap to reflect the importance of closing the gap relative to its size.

4. **Simplified O-D Centrality Approach:** A simplified Origin-Destination (O-D) centrality approach, akin to Equation 2.5, is adopted to streamline the calculations.

Standardized Trip Flow Calculation

The standardized number of trips in Equation 2.3 is adjusted to consolidate reverse trips into a single direction. This adjustment stems from the simplification of the O-D matrix, where trip direction is disregarded. As a result, $\phi_{st} = \phi_{ts}$, and for each pair i, j, the standardized number of trips between node i and node j is now denoted $\phi_{i,j}$ and is calculated as:

$$\phi_{i,j} = \frac{t_{i,j}}{\max_{i,j \in P}(t_{i,j})} \quad \text{for} \quad i \neq j \tag{3.1}$$

where:

- $t_{i,j}$ is the number of trips from node i to node j.
- $\max_{i,j \in P}(t_{i,j})$ is the maximum number of trips between any two nodes in the set P.

Modification to Stress Centrality

The stress centrality measure is adapted to account for traffic demand, similar to the modifications made by [25] and [28]. The modified stress centrality, incorporating traffic demand, is expressed as:

$$tc(l) = \sum_{i,j \in P} \phi_{i,j} \cdot n_{i,j}(l) \tag{3.2}$$

where:

- $tc(l)$ represents the weighted stress centrality of edge l.

- $\phi_{i,j}$ is the standardized number of trips from node $i \in P$ to node $j \in P$.
- $n_{i,j}(l)$ is the number of shortest paths from node $i \in P$ to node $j \in P$ that pass through edge l.

Weighted Benefit

The Weighted Benefit metric quantifies the value of closing a gap, referred to as a Component in this research, in the bike network by considering both traffic demand and the physical length of the edges within that component. This metric is adapted from an edge-level benefit measure introduced by [21], which calculates the avoided distance cycled in mixed traffic per unit length of an edge.

The weighted benefit for a component is calculated by summing the weighted stress centrality of every edge l in the component c, and normalizing by the total length of the component, using the following equation:

$$WB^*(c) = \frac{\sum_{l \in c} tc(l) \cdot L(l)}{\sum_{l \in c} L(l)} \tag{3.3}$$

where:

- $WB^*(c)$ represents the weighted benefit of closing component c.
- $tc(l)$ is the weighted stress centrality of edge l within the component.
- $L(l)$ is the length of edge l.

For components consisting of multiple edges, this weighted benefit is determined by summing the contributions of each edge within the component and then normalizing by the total length of the component. This method effectively provides a "benefit per unit of investment," which reflects how beneficial closing the gap would be relative to the component's size.

This metric ensures that the prioritization process accounts for both the demand (through traffic centrality) and the utilization of edges, aiding in the strategic planning and targeted improvement of the bike network.

Additionally, a baseline benefit metric without considering traffic flow is also utilized for comparison. This simpler benefit metric is given by:

$$B^*(c) = \frac{\sum_{l \in c} c(l) \cdot L(l)}{\sum_{l \in c} L(l)} \tag{3.4}$$

where:

- $B^*(c)$ represents the benefit of closing a component c.
- $c(l)$ is the stress centrality of edge l.
- $L(l)$ is the length of edge l.

By comparing the results of these two metrics—one considering traffic demand and one that does not—planners can gain insights into the added value of incorporating cyclist behavior into network improvement strategies.

Mathematical Intuition and Numerical Example

To provide mathematical intuition, consider a network with nodes A, B, C, and D, and edges e_1 (between A and B), e_2 (between B and C), and e_3 (between C and D). Assume there are 100 trips between A and D, distributed equally among all paths. The standardized number of trips $\phi_{A,D}$ from A to D is 1 if it is the maximum, otherwise scaled accordingly.

If the number of shortest paths passing through edge e_2 is 50, with $\phi_{A,D} = 0.5$, then:

$$tc(e_2) = \phi_{A,D} \cdot n_{(A,D)}(e_2) = 0.5 \cdot 50 = 25$$

If edge e_2 has a length $L(e_2) = 10$ meters, and the total length of the gap g consisting of e_1, e_2, and e_3 is 30 meters, then:

$$WB^*(g) = \frac{tc(e_2) \cdot L(e_2)}{\text{Total Length of Component}} = \frac{25 \cdot 10}{30} = \frac{250}{30} \approx 8.33 \text{ meter-benefit units}$$

This calculation is repeated for each edge within the gap, and the total benefit is normalized by the total length of the gap, ensuring that the measure accurately reflects the weighted benefit of closing the gap relative to its total length.

Edge Prioritization Simplification

To streamline the edge prioritization process, several simplifications are introduced:

- **Node Pairs for Centrality Calculation:** The set P of all node pairs is restricted to only those pairs present in the Origin-Destination (OD) matrix. This focuses the centrality calculation on the routes that are of actual interest to cyclists, rather than considering every possible node pair in the network. The set of all pairs includes all the edges that belong to a route; thus, one route can span several edges. This subnetwork, known as the Routes Network, includes only those edges that are part of any route, filtering out edges not associated with the routes.

- **Preferred Bicycle Routes:** The shortest path between each node pair (i, j) is defined as the preferred bicycle route. This definition aligns the routing with realistic cyclist preferences, ensuring that the routes prioritized for improvement reflect actual travel behavior.

- **Subset of Edges:** The weighted centrality is calculated only for edges that are part of these preferred bicycle routes and are classified as potential bike paths. This restricts the analysis to the most relevant edges, excluding those not involved in the preferred routes.

This approach ensures that the centrality measures and component benefits are concentrated on the most pertinent areas of the network, aligning with actual cyclist demand and behavior.

Numeric Example of Simplification

To illustrate the effectiveness of this simplification, consider a small network with four nodes A, B, C, and D, and the following edges with their lengths:

- Edge e_1 between A and B with a length of 5 meters.

- Edge e_2 between B and C with a length of 10 meters.

- Edge e_3 between C and D with a length of 15 meters.

- Edge e_4 between A and D with a length of 20 meters.

Assume the OD matrix indicates that cyclists frequently travel between nodes A and D (50 trips), and between A and C (30 trips).

Without simplification:

All edges e_1, e_2, e_3, and e_4 would be considered for centrality calculations, regardless of their relevance to the actual travel routes.

With simplification:

- **Node Pairs for Centrality Calculation:** Focus on pairs (A, D) and (A, C) from the OD matrix. The Routes Network consists of the edges involved in these routes: e_1, e_2, and e_3, while e_4 is excluded as it does not belong to any of the preferred routes.
- **Preferred Bicycle Routes:**
  - Shortest path from A to D is $e_1 \rightarrow e_2 \rightarrow e_3$ with a total length of $5+10+15 = 30$ meters.
  - Shortest path from A to C is $e_1 \rightarrow e_2$ with a total length of $5 + 10 = 15$ meters.
- **Subset of Edges:** The edges considered for weighted centrality calculations are e_1, e_2, and e_3, excluding e_4 as it is not part of the preferred routes.

By focusing on the subset of edges involved in the preferred routes, the centrality measures and gap benefits are concentrated on the most relevant parts of the network. For example, if we calculate the weighted centrality for these edges, we focus on how critical e_1, e_2, and e_3 are for the preferred routes (A, D) and (A, C), rather than considering all edges in the network. This targeted approach ensures that the improvements made will have the highest impact on the most utilized routes, thereby effectively addressing cyclist demand.

3.2.7 Routing Approach

The Routing Approach focuses on efficiently constructing optimal routes within a component to enhance connectivity. Instead of evaluating every possible route, which would be computationally expensive, this method prioritizes routes connecting "contact nodes"—key transition points between bike paths and non-bike paths. The approach emphasizes significant routes to maximize the benefit of the network enhancements.

The process involves the following key steps:

- **Identify Contact Nodes:** Begin by identifying all contact nodes within the component. Contact nodes are points where transitions occur between existing bike paths and potential bike paths.
- **Calculate Potential Routes:** Determine all possible routes between each pair of contact nodes within the component. This involves calculating paths that traverse the component's edges, taking into account the connectivity between these nodes.

- **Compute Route Benefits:** For each potential route, compute the benefit of incorporating that route into the network. This benefit is evaluated based on the impact of the route on overall network connectivity and the volume of bike traffic it supports.

- **Select Optimal Route:** Choose the route with the highest benefit score. This route is prioritized for inclusion in the network as it offers the greatest improvement in connectivity.

- **Update Route Set:** Remove the selected route from the set of potential routes. Recalculate the benefit values for the remaining routes, considering the impact of the newly added bike path.

- **Iterate:** Repeat the process of selecting the highest benefit route, updating the route set, and recalculating benefits until all contact nodes within the component are connected by optimal routes. This iterative process ensures that the network is progressively improved with each added route.

This iterative method ensures that the most significant and beneficial routes are constructed first, leading to an efficient enhancement of the bike network. By focusing on routes between contact nodes and updating the network incrementally, the approach maximizes the impact of each added route while managing computational complexity.

3.3 Metrics for Network Quality Assessment

To assess the quality of the network throughout the growing process, different metrics were calculated and reported at each iteration level.

Network Metrics assess the overall Bike Network growing process, providing an aggregate view of network coverage and performance.

- **Bike Path Coverage:** Measures the extent to which the required cycled network is covered with bike paths.
$$\text{BPC} = \frac{\sum L_{bp}}{\sum L_{total}} \tag{3.5}$$

- **Bike Trips Coverage:** Assesses how much of the total demand for trips is covered with bike paths.
$$\text{BTC} = \frac{\sum T_{bp}}{\sum T_{total}} \tag{3.6}$$

- **Routes Coverage:** Calculates the proportion of all cyclist routes covered with bike paths.

$$\text{RC} = \frac{\sum R_{bp}}{\sum R_{total}} \tag{3.7}$$

- **Gap Length:** The total length of bike paths that need to be constructed to cover the chosen routes.

$$\text{GL} = \sum L_{gp} \tag{3.8}$$

Where:

- L_{bp}: Length of edges that are bike paths.

- L_{total}: Total length of all edges that are bike paths or potential bike paths and belong to a route.

- T_{bp}: Number of trips covered by bike paths.

- T_{total}: Total number of trips.

- R_{bp}: Number of routes covered by bike paths.

- R_{total}: Total number of routes.

- L_{gp}: Length of edges in chosen routes that are potential bike paths.

4
Results

In this section, the results of the study are presented, building on the previously discussed Network Creation step, which focused on maintaining the Largest Connected Component.

The analysis continues with Flow Acquisition, providing a detailed examination of key origin and destination points, potential routes, and zones with high demand for cycling infrastructure.

Next, the primary missing components within the network are identified and prioritized using two metrics for comparison: Weighted Benefit and Benefit.

Finally, the results from the algorithm for detecting missing links are highlighted, emphasizing the most critical missing components and the primary routes that should be constructed to enhance Padova's bike network.

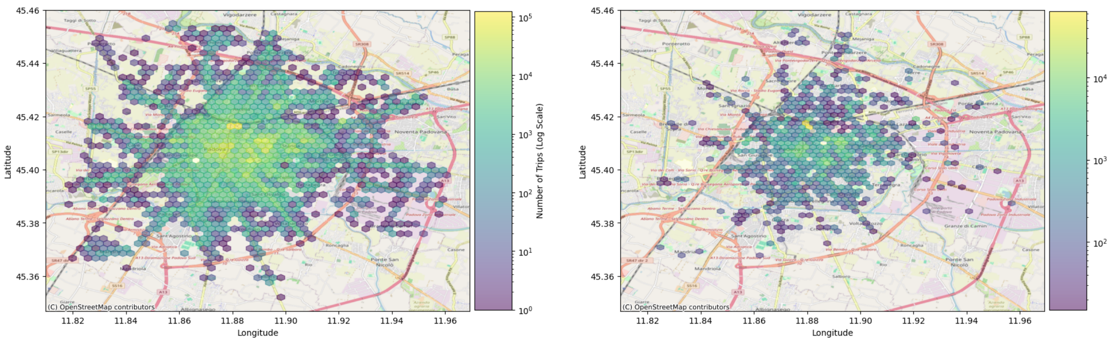
4.1 Flow Acquisition and Route Assigment

Following the steps outlined in the methodology section, the Origin-Destination (O-D) Matrix was obtained. Table 4.1 presents a subset of the final O-D Matrix, detailing the shortest paths between origin nodes (from_node_id) and destination nodes (to_node_id) along with their respective lengths. As described in the methodology, only routes with more than fifteen trips were retained. The final O-D matrix includes 7,565 routes, accounting for a total of 261,682 trips.

from_node_id	to_node_id	n_trips	route_nx	route_nx_length
3899062047	5752101515	1426	[3899062047, 6449412215, 3899062054, 330049257...]	1484.905
5752101515	3899062047	1211	[5752101515, 1904462791, 5809457838, 334237404...]	1484.905
246791092	250092661	511	[246791092, 540999696, 540999675, 540999673, 5...]	1683.026
330049257	5752101515	448	[330049257, 2193230671, 1101896768, 2193230680...]	1374.593
5752101515	2193230680	388	[5752101515, 1904462791, 5809457838, 334237404...]	1194.976

Table 4.1: Summary of trips and routes with their lengths. Note: This table provides a snapshot of the Origin-Destination matrix, highlighting key routes and their lengths.

To analyze bike trip origins and destinations within the bike-sharing system, map visualizations are used due to the large dataset. Hexbin plots are particularly suitable for visualizing this information. Figures 4.1a and 4.1b illustrate the distribution of bike-sharing trip origins and destinations in Padova, both before and after the routes were filtered. Before filtering, a significant number of low-frequency trips can be observed in the city's periphery. After filtering, most of these trips are removed, revealing a concentration of trip origins and destinations around the main city zones and connection points. High-density areas, depicted in brighter colors, include the train and bus stations, city center, and hospital. In contrast, areas farther from the city center show lower trip densities, possibly due to fewer bike parking spots, less connectivity, or reduced transportation demand.



(a) Before the Filtering Process.

(b) After the Filtering Process.

Figure 4.1: Distribution of Bike-Sharing Trip Origins and Destinations. Note: Hexbin maps show the spatial distribution of bike-sharing trips in Padova. The left figure represents the distribution before the filtering process, and the right figure represents the distribution after the filtering process. The color intensity represents the logarithmically scaled number of trips.

To gain deeper insights into the connections between origin-destination points, flow data is visualized using Kepler.gl [38], as illustrated in Figure 4.2. In this visualization, node sizes correspond to the number of trips starting or ending at each point, while colors distinguish nodes as origins (purple) or destinations (green). The lines between nodes represent the trips between these points. This visualization, similar to a hexbin map, clearly shows that most trips are concentrated around key areas such as the train station, city center, and university, with substantial flow patterns connecting these central locations.

After applying route assignment to each O-D pair, the network is simplified, resulting in the creation of the Route Network, which forms the foundation for the subsequent analysis. This subnetwork includes only the bike paths and potential bike paths that are part of the calculated routes. The Route Network was further segmented into potential and existing bike edges. Of the approximately 6,500 edges involved in these routes, around 4,119 edges were classified as "potential," indicating they are not currently part of Padova's existing bike network. The Route Network is shown in Figure 4.3.

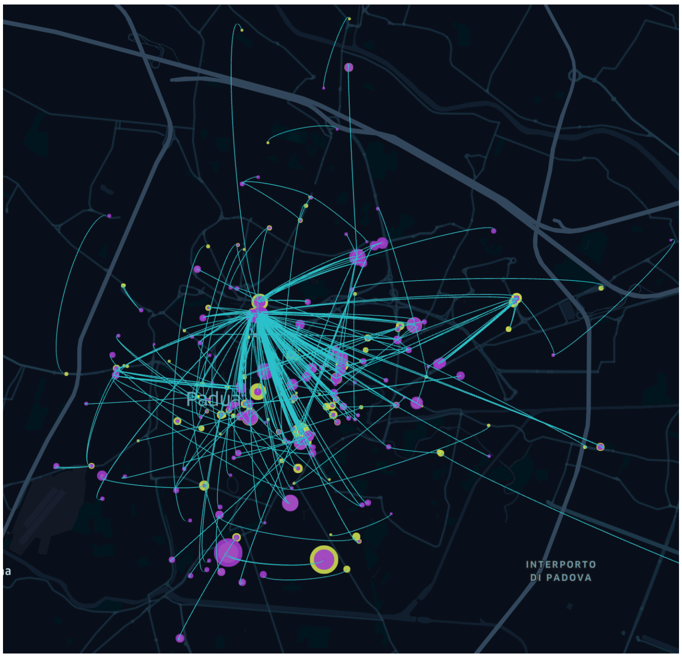


Figure 4.2: Cyclist Flow Representation in Padova. Note: Node size represents the number of trips starting or ending at that point. Colors indicate the role of the node: origin (purple) or destination (green). Lines represent the connections between these points.

Figure 4.3: Route Network. Note: Green lines are bike paths, light gray lines are potential bike paths.

4.2 Component Identification and Prioritization

In this section, the results of the Component Identification and Component Prioritization steps are presented. Both numerical results and visualizations will be provided to offer a comprehensive understanding of the analysis. The combination of quantitative data and visual representations is crucial for effectively conveying the distribution, significance, and spatial context of the identified components. This dual approach ensures that the findings are not only analytically robust but also easily interpretable, facilitating informed decision-making for network improvements.

4.2.1 Identification of Disconnected Components

The identification of disconnected components focused on assessing gaps and zones of disconnection within the network. The Route Network served as the foundation for this analysis. Initially, the process involved removing existing bike paths and using a built-in function of NetworkX to identify connected components composed solely of potential bike paths.

This analysis began with the identification of 209 components, with 19 components requiring further subdivision. By applying a distance-limited exploration technique, the total number of disconnected components identified across the city increased to 619. Many of these were single-node components or dead-ends, which led to further refinement. Only those components through which routes passed were retained, resulting in a final count of 355 components. The analysis revealed significant variation in the size of these gaps, with the largest connected component (LCC) consisting of 29 edges and the smallest comprising only 2 edges. The median number of edges per component was two, indicating that many sections of the bike network require only minor adjustments.

To present this information in an interactive manner, maps were generated to provide both detailed views of individual components and an overview of all disconnected components. Figure 4.4 shows the identified components within the Padova Network, with many situated in the city center, near the train station, and in other critical areas. The complexity of the city center, characterized by restricted streets for cars and bikes and main streets shared with trams and buses, explains the lack of continuous routes. However, this analysis highlights that these areas are crucial for cyclists to navigate efficiently.

Figure 4.4: Components Identification. Legend: Light gray lines represent routes with bike paths. Blue lines represent component edges.

Figure 4.5 illustrates the concept of components and contact nodes, with red nodes highlighting the points where components connect to existing bike paths.

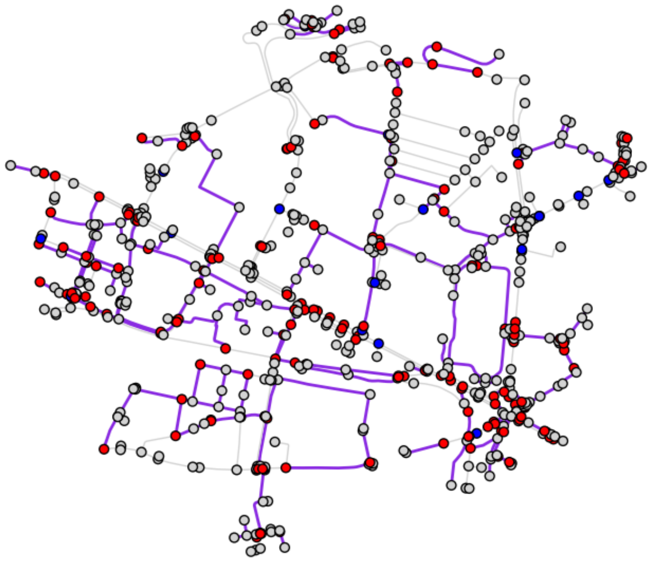


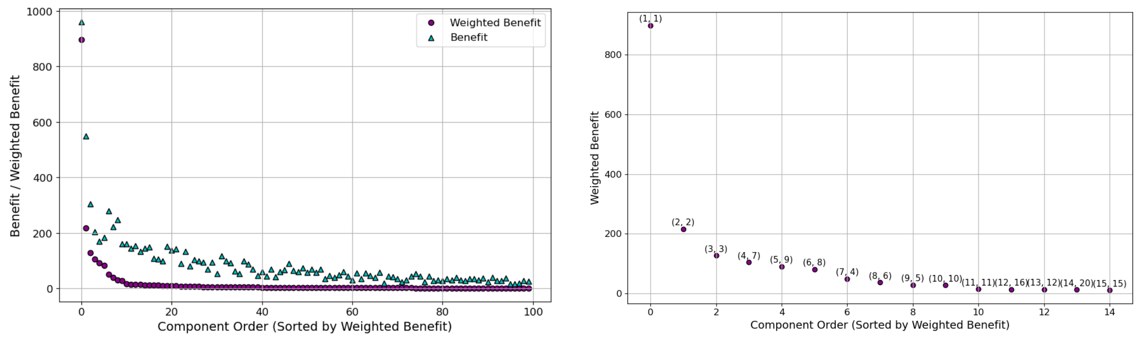
Figure 4.5: Components Identification. Legend: Edges are color-coded as follows: violet edges represent those that are part of a component, while light gray edges depict bike paths or potential bike paths not included in the component analysis. The nodes are also color-coded: contact nodes are shown as red dots, nodes part of only bike path edges are shown in blue, and nodes part of only potential edges are shown in light gray.

4.2.2 Component Prioritization Results

This section presents the results of the prioritization process, using both benefit metrics discussed in the Methodology. These metrics take into account the length of the components, the number of routes passing through them, and, in the case of the weighted benefit, the number of trips through each edge relative to the maximum number of trips expected within the network. The analysis provides a clear ranking of components based on their potential impact on improving the bike network, helping to identify the most critical gaps to address.

The first step in prioritizing the components involves calculating the metrics for each one as outlined in Equations 3.3 and 3.4. Figure 4.6a displays the values for both benefit metrics, with components ordered according to their weighted benefit. It is observed that the weighted benefit metric generally decreases when flow weight is considered, likely due to its ratio-based nature. This observation suggests that the relative importance of components changes based on whether or not flow is factored into the analysis.

To highlight this shift, Figure 4.6b allows for a direct comparison of the component rankings according to the different benefit metrics. For instance, while the weighted benefit metric ranks component number 12 higher, the benefit metric places component number 16 in a more prominent position. This variation in ranking is further illustrated by visualizing components 12 and 16 in Figures 4.7a and 4.7b, respectively. Component 12 is distinguished by a higher flow of trips, while component 16 stands out due to its greater length, as detailed in Table 4.2.



(a) Weighted Benefit and Benefit for the First 100 Components.

(b) Weighted Benefit and Benefit for the First 15 Components.

Figure 4.6: Comparison of Weighted Benefit and Benefit for Components.

(a) Number 12.

(b) Number 16.

Figure 4.7: Comparison of Components 12 and 16.

To better illustrate the metrics used for prioritizing components, Figures 4.8a and 4.8b display the component identification based on the Weighted Benefit and Benefit metrics, respectively. In these figures, red zones indicate areas with the highest priority, while yellow highlights those with lower significance according to each metric. The distribution of these high-value areas appears similar across both metrics, with only minor variations. The similarity in the distribution of high-value areas across both metrics can be attributed to the centrality measures, which heavily influence zones where more routes converge, such as areas near the train station and city center. These central hubs naturally see higher traffic and route convergence.

Size	Contact	Weighted Benefit	Benefit	Rank WB	Rank B
0	6	897.74	961.74	1.0	1.0
1	6	216.30	548.70	2.0	2.0
2	12	127.67	304.01	3.0	3.0
3	16	105.17	202.73	4.0	7.0
4	5	90.55	169.75	5.0	9.0
5	8	81.43	183.11	6.0	8.0
6	6	49.22	280.03	7.0	4.0
7	3	37.83	222.38	8.0	6.0
8	2	29.25	247.00	9.0	5.0
9	12	28.28	160.61	10.0	10.0

Table 4.2: Comparison of Component Metrics.

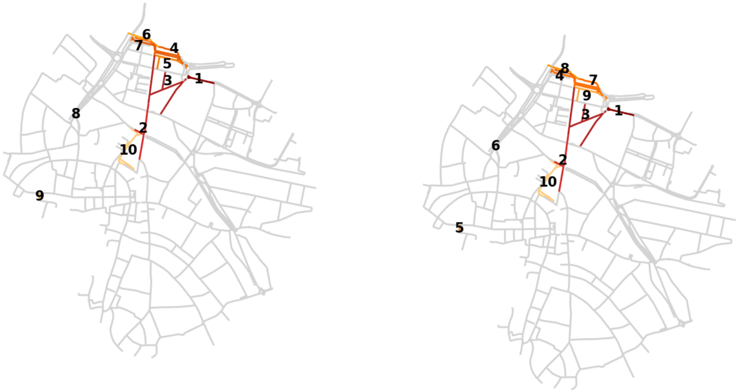
The figures highlight that the critical areas for bike network improvements are those connecting major city hubs, including the city center and train station. Both maps show that the importance of these connections is key to optimizing the bike network.

(a) Components Importance by Weighted Benefit. **(b)** Components Importance by Benefit.

Figure 4.8: Comparison of Components by Weighted Benefit and Benefit. Legend: Edges are color-coded from red to yellow, with red representing the most important components according to the benefit metric, and yellow indicating the least important ones.

To delve deeper into the data, Figure 4.9 focuses on the top 10 components identified by each metric. The zoomed-in map of the city center, particularly around the train station, shows that while the top 10 components are consistent, their ranking changes based on the metric used. This shift reflects the influence of demand considerations in the Weighted Benefit metric, highlighting areas with higher bike traffic.

Figures 4.10 and 4.11a provide a more detailed view of the components using street-level maps created with Folium.



(a) Ordered by Weighted Benefit Metric.

(b) Ordered by Benefit Metric.

Figure 4.9: Top 10 Components in Padova's Bike Network Based on Benefit Metrics. Legend: Numbers indicate the component order.

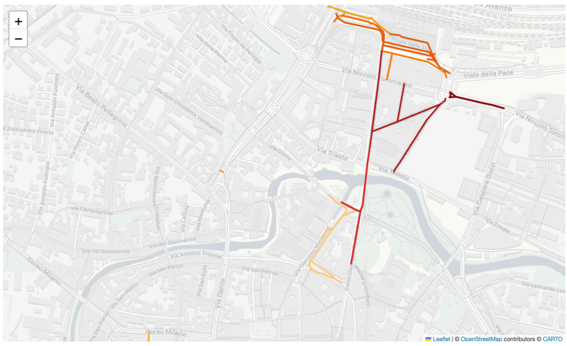
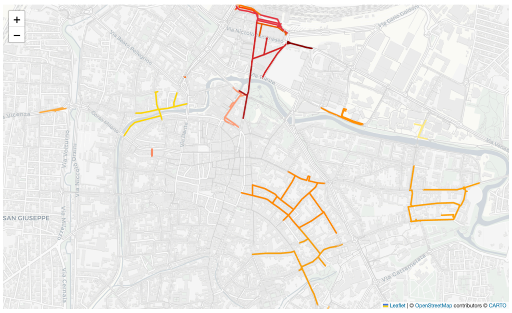


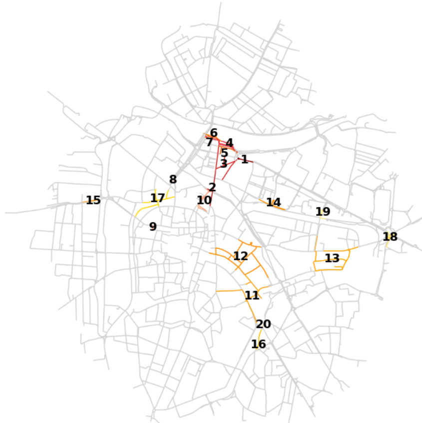
Figure 4.10: Top 10 Components Based on Weighted Benefit Metric.

Figure 4.10 showcases the top 10 components, highlighting their concentration around major locations such as the train station and city center. Figures 4.11a and 4.11b provide an extended view of the next 10 components, demonstrating that while many components remain centered around key areas, there are also significant needs for improvements in other parts of Padova. For instance, component numbers 13 (Figure 4.13) and 17 facilitate connections between the university and the city center, while component number 12 links the hospital with the city center. This broader view underscores the necessity for enhancing bike infrastructure across the entire city.

Figure 4.12 presents a series of plots highlighting the precise locations of various components identified using the weighted benefit metric. Components 1 and 3 are located near the train station, addressing key connectivity gaps in this high-traffic area. Components 2 and 10



(a) Top 20 Components



(b) Top 20 Components Labeled According to Prioritization

Figure 4.11: Comparison of the Top 20 Components Based on Weighted Benefit Metric and Their Labeled Visualization.

represent important streets within the city center, while Component 17 connects key areas of the university.

A limitation of this methodology is its heavy reliance on data accuracy, particularly from OpenStreetMap (OSM). Some areas identified as lacking bike infrastructure, such as Component 1, already have bike paths, but these are painted on the sidewalk and not properly captured in OSM data. This issue also affects other components, underscoring the need for improved data quality in urban infrastructure planning.

4.2.3 Routing Components

The process begins by identifying the best component based on the benefit metric. After pinpointing the component, a routing process between contact nodes is carried out. This process utilizes edge-level information to determine the optimal edges and routes for fully connecting the component. Unlike previous approaches that identified gaps and performed routing without specific criteria for route construction, this method provides a detailed analysis.

However, due to the reduced size of most components, the routing approach does not significantly enhance the results for smaller components. The decision to reduce component size was influenced by time and computational constraints when initially implementing the routing approach. This reduction in size leads to smaller components where the routing process offers limited additional insights. For future work, it will be necessary to find a balance between reduc-

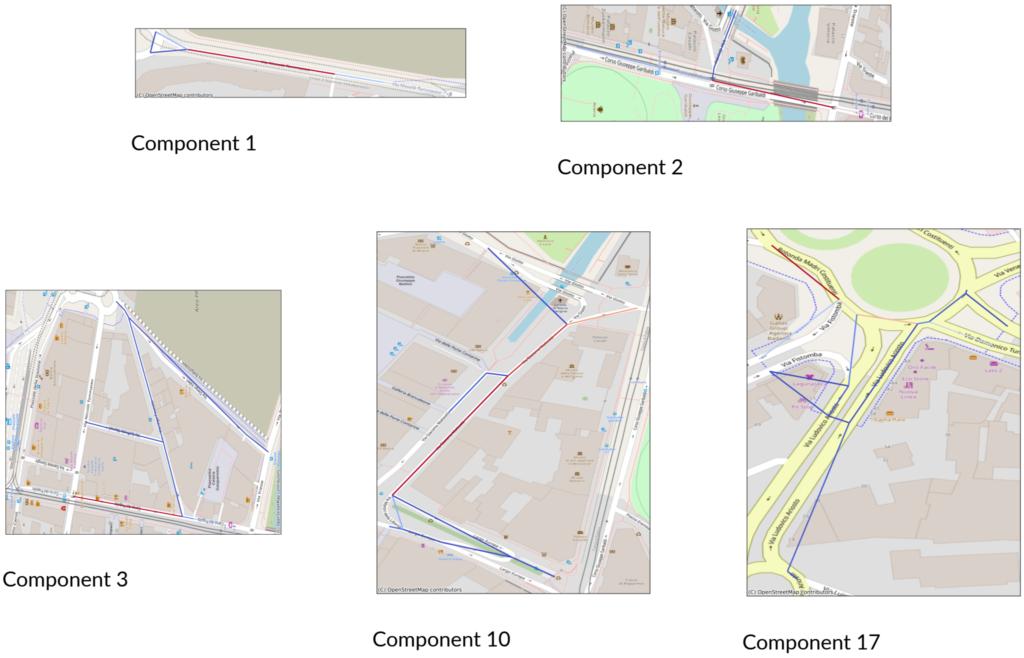


Figure 4.12: Comparison of Selected Components. Each image represents a different component with its respective visual details.

ing component size and providing effective routing within larger components. The purpose of having components instead of smaller sections is to maintain a macro-level perspective on where efforts should be concentrated. The routing process then provides a high-level solution, minimizing the number of routes needed to better connect zones where bike path availability does not meet demand and to reduce the cost of fully connecting a large component.

Despite the limitations with smaller components, the routing approach is still applied to refine each component according to the prioritization based on the weighted benefit metric. While many components are small, the routing approach proves valuable in larger components, improving the quality metrics by reducing the number of meters needed to close gaps in the city's bike network.

Figure 4.13 illustrates an example of how the routing approach functions within Component Number 13. This component has a size of 26 edges and includes 10 contact nodes. When applying the routing approach, many edges are eliminated, leaving only the central ones to connect all contact nodes. In this case, instead of adding 1,255.087 meters of new paths, only 1,055.064 meters are required, demonstrating the efficiency of the routing approach.

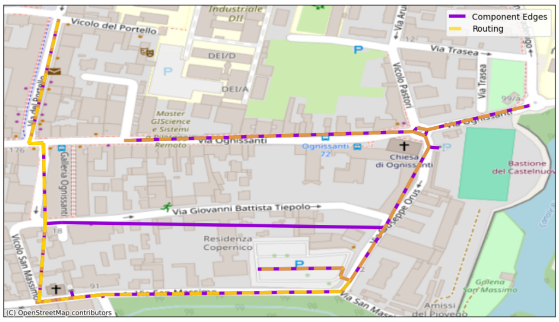


Figure 4.13: Highlighted component number 13 for the Routing Approach Result.

4.3 Quality Metrics Analysis

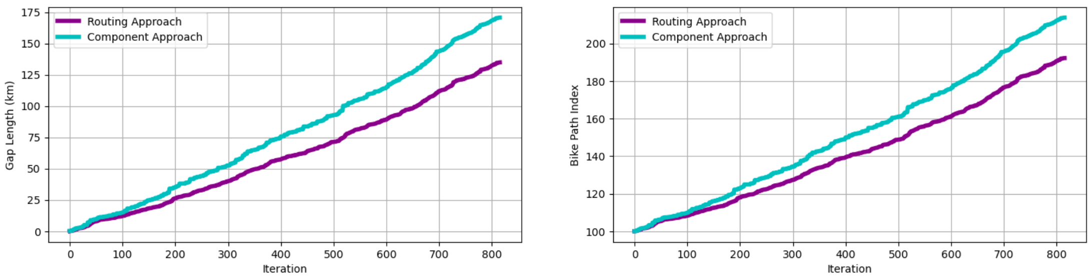
This section presents metrics evaluating the quality of the network under both the Component Approach and its refinement via the Routing Approach. The analysis includes prioritization based on both weighted benefit and standard benefit metrics for the Component Approach. The discussion begins by examining the results from the weighted benefit prioritization.

4.3.1 Comparison between Approaches

The comparison in Figure 4.14 reveals that the Component Approach adds a total of 175 km to the network to close all gap components, whereas the Routing Approach requires nearly 125 km for the same task. In both cases, the total bike path length nearly doubles to close all gaps in the network. The Routing Approach demonstrates greater efficiency in reducing gap length, as it results in fewer kilometers being added to the network. This efficiency suggests a more selective and optimized process in connecting nodes with minimal additional infrastructure.

Figure 4.15 provides a focused analysis on the first 30 components. Initially, both approaches exhibit similar trends in gap length and bike path length index. However, as more components are added, differences emerge, with the Routing Approach becoming more efficient in minimizing gap length while selectively adding bike paths. This efficiency is reflected in the more significant divergence from the Component Approach, which may not prioritize optimization as strictly.

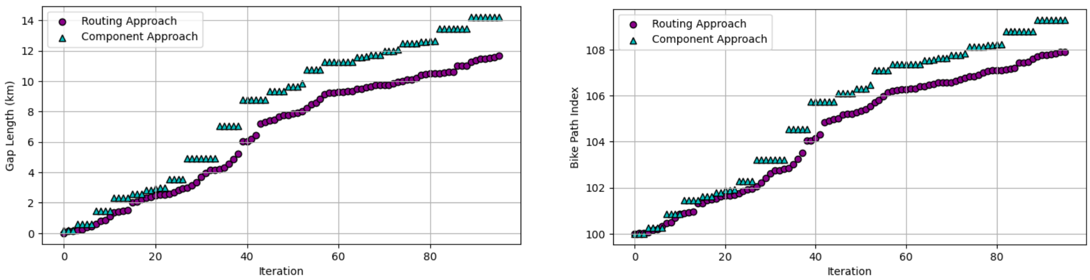
Figure 4.16 compares path coverage between the Component and Routing Approaches. The Component Approach generally achieves a higher path coverage ratio, indicating a broader integration of network paths as bike paths. The Routing Approach, however, optimizes the addition of routes within components, leading to a more selective increase in path coverage.



(a) Gap Length over Iterations

(b) Bike Path Length Index over Iterations

Figure 4.14: Bike Path Length and Gap Length for the Component and Routing Approach. Note: Comparison matches the routes of the Routing Approach to the corresponding components in the Component Approach.



(a) Gap Length for the First 30 Iterations

(b) Bike Path Length Index for the First 30 Iterations

Figure 4.15: Bike Path Length and Gap Length for the Component and Routing Approach for the First 30 Components. Note: See Figure 4.14.

This comparison highlights a trade-off between comprehensive coverage and optimized path selection.

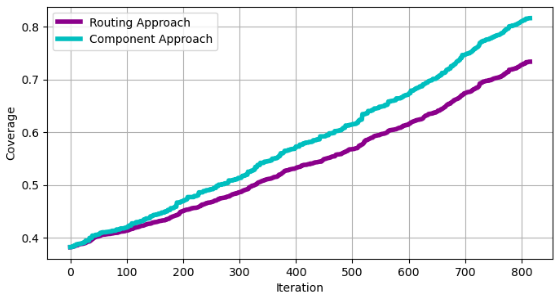
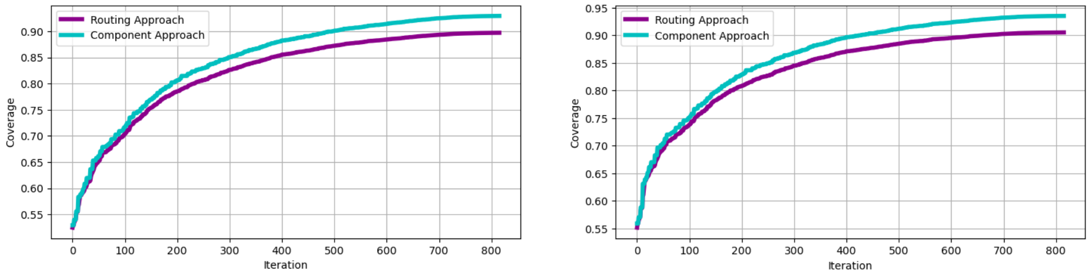


Figure 4.16: Path Coverage for the Component and Routing Approach. Note: See Figure 4.14

Despite the Routing Approach covering fewer routes, the difference in trip coverage between the approaches is minimal, as shown in Figure 4.17. The Component Approach adds more bike paths overall, not specifically targeting optimization but covering entire components. The

Routing Approach prioritizes routes with higher demand, resulting in efficient path selection and achieving comparable levels of trip and route coverage.

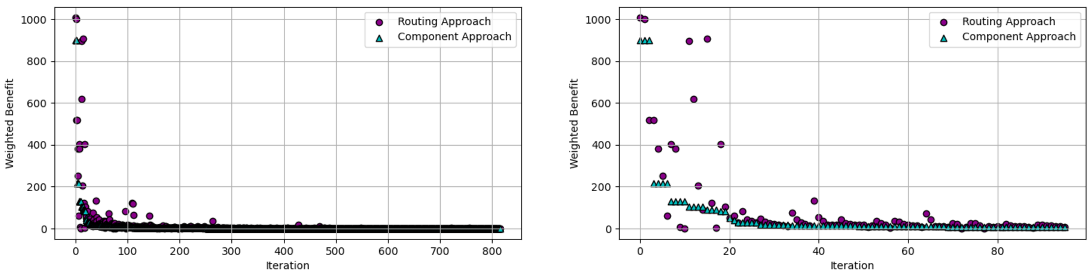


(a) Route Coverage over Iterations

(b) Trip Coverage over Iterations

Figure 4.17: Routes and Trips Coverage for the Component and Routing Approach. Note: See Figure 4.14

Figure 4.18 compares the weighted benefit between the Component and Routing Approaches. The Component Approach shows a straightforward progression, selecting components with the highest benefit sequentially. In contrast, the Routing Approach recalculates the benefit after each route selection, causing fluctuations in the benefit metric. This reflects the dynamic optimization process inherent to the Routing Approach.

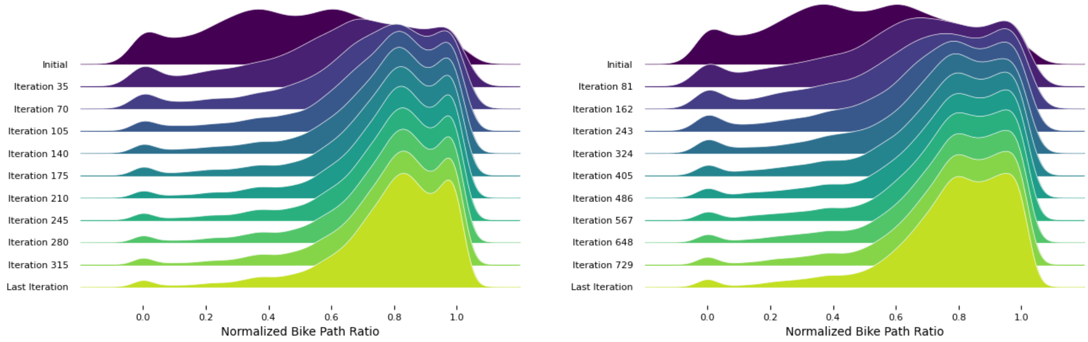


(a) Weighted Benefit over Iterations

(b) Weighted Benefit for the First 30 Iterations

Figure 4.18: Weighted Benefit for the Component and Routing Approach. Note: See Figure 4.14

To analyze how the addition of bike paths affects route coverage, Figures 4.19a and 4.19b illustrate the distribution of bike path coverage across routes over successive iterations. Initially, many routes have low bike path coverage, but as iterations progress, a higher proportion of routes become fully covered, as indicated by the rightward shift in the distribution. This progression demonstrates the Component Approach's effectiveness in gradually increasing bike path coverage. A similar trend is observed with the Routing Approach, which prioritizes routes with higher demand, further enhancing network efficiency.



(a) Component Approach

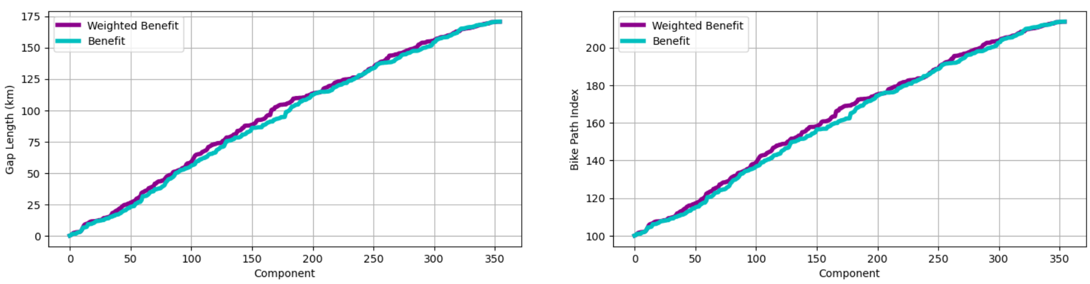
(b) Routing Approach

Figure 4.19: Comparison of Evolution of Bike Path Ratios for Different Approaches. Note: Both plots display the evolution of the Bike Path Ratio during the iterative addition of bike paths, adjusted using a normalized distribution.

4.3.2 Comparison of Metrics: Component-Based Approach

This section compares the two methods used for component prioritization: "Weighted Benefit" and "Benefit." Figure 4.20 shows that gap length decreases as more components are added, indicating improved network connectivity. Both metrics display similar trends in covering gaps, although the specific order of component addition may vary. This suggests that both methods are effective, with the choice between them depending on priorities such as emphasizing flow coverage (Weighted Benefit) or to emphasize network load (Benefit).

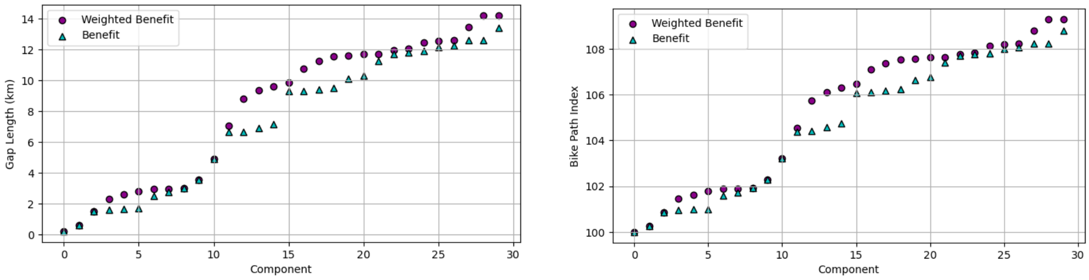
A closer look at the first 30 components, as shown in Figure 4.21, reveals slight variations in the early stages. Both metrics effectively reduce gap length initially, with differences arising due to the criteria used for component selection.



(a) Gap Length over Iterations

(b) Bike Path Length Index over Iteration

Figure 4.20: Bike Path Length and Gap Length Comparison: Benefit and Weighted Benefit for the Component Approach.



(a) Gap Length for the First 30 Iterations

(b) Bike Path Length Index for the First 30 Iterations

Figure 4.21: Bike Path Length and Gap Length Comparison: Benefit and Weighted Benefit for the Component Approach for the First 30 Components.

Figures 4.22 and 4.23 illustrate the evolution of path coverage over iterations for both metrics. While path, trip, and route coverage exhibit similar trends, the Weighted Benefit approach achieves network coverage marginally faster. This increased efficiency underscores the advantage of prioritizing components based on the weighted benefit.

These results imply that the Weighted Benefit metric offers a more efficient approach to covering routes and trips. This method enables a more strategic network expansion, focusing on components with the greatest impact.

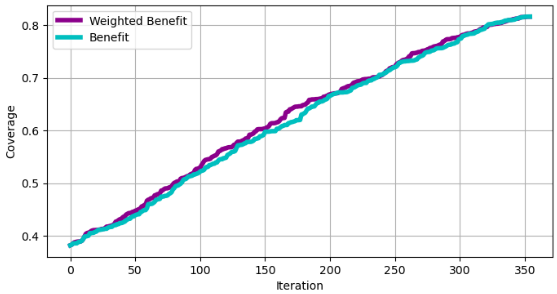
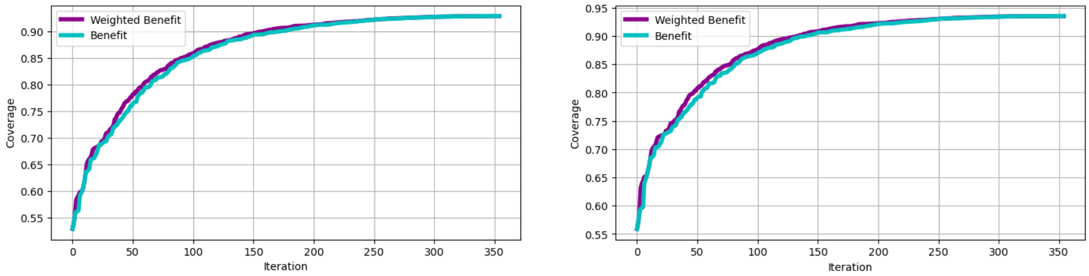


Figure 4.22: Path Coverage Comparison: Benefit and Weighted Benefit for the Component Approach.

4.4 Summary of Findings

In summary, the Component approach added 175 km of infrastructure, while the Routing approach required only 125 km, demonstrating greater efficiency by optimizing route selection and minimizing new infrastructure. Prioritizing components based on the Weighted Benefit metric achieved faster and more comprehensive network coverage. The analysis identified 355



(a) Route Coverage over Iterations

(b) Trip Coverage over Iterations

Figure 4.23: Routes and Trips Coverage: Benefit and Weighted Benefit for the Component Approach.

gaps, notably in critical areas like the city center and train station, underscoring the need for targeted improvements. Overall, the Routing Approach and Weighted Benefit metric proved more effective in enhancing bike network connectivity and efficiency.

5 Conclusion

This study evaluated two strategies for expanding Padova's bike-sharing infrastructure: the Component and Routing approaches. The Component approach provided a comprehensive view, identifying 355 gaps, particularly in high-demand areas like the city center and near the train station, requiring approximately 175 km of new infrastructure. The Routing approach, refining the Component method, proved more efficient by focusing on detailed routing within components, adding around 125 km of new paths while maintaining similar coverage.

The methodology involved a structured framework combining both approaches. It began with network creation using OSMnx, followed by flow analysis to assess demand, and simplified the network to focus on core paths. Disconnected components were prioritized using a modified centrality measure that included flow data, and routing optimization improved network connectivity within components.

Key metrics used included Weighted Link Closure and Weighted Benefit, adapting existing measures to balance infrastructure needs with cyclist demand. The study highlights the importance of integrating empirical traffic data, utilizing RideMovi data to prioritize projects effectively. The two-step process—macro identification of deficiencies and detailed routing optimization—ensures targeted improvements.

Prioritization should focus on:

- **Central Areas:** Significant gaps were identified in the city center due to complex routing constraints.

- **Train Station:** High priority due to its role as a major transit hub.
- **High-Demand Zones:** Areas with higher cyclist activity needing targeted improvements.
- **Gaps Distribution:** Concentrated in key transport hubs and high-demand areas.

The key findings for each methodology step are:

- **Component Identification:** Identified 619 disconnected components, reduced to 355 after filtering. Significant gaps were particularly found in central and high-demand areas such as the city center and train station.
- **Component Prioritization:** The Weighted Benefit metric prioritized components based on cyclist demand and strategic importance, providing a more efficient and targeted solution compared to the standard Benefit metric.
- **Routing Approach:** Demonstrated greater efficiency by reducing required infrastructure from 175 km (Component Approach) to 125 km. The Routing Approach proved more effective in connecting key zones with fewer additional kilometers.
- **Quality Metrics:** The Routing Approach and Weighted Benefit metric showed superior performance in optimizing network connectivity and minimizing new infrastructure. Path, trip, and route coverage were achieved faster with the Weighted Benefit approach.

The primary limitations of this study include the availability of up-to-date bike path data and the potential for incorporating alternative methods for determining origin and destination points. Key challenges involved accurately tracking existing bike paths and avoiding overlaps with established routes. Future work should focus on refining both the Component and Routing approaches by adjusting thresholds and hyperparameters to optimize component reduction and routing effectiveness. Overall, the Routing Approach proved to be a more efficient solution for improving bike network connectivity, while the Component Approach offered valuable insights into broader network coverage.

References

[1] W. F. Lamb, T. Wiedmann, J. Pongratz, R. Andrew, M. Crippa, J. G. Olivier, D. Wiedenhofer, G. Mattioli, A. Al Khourdajie, J. House *et al.*, "A review of trends and drivers of greenhouse gas emissions by sector from 1990 to 2018," *Environmental research letters*, vol. 16, no. 7, p. 073005, 2021.

[2] D. Banister, *Unsustainable transport: city transport in the new century*. Routledge, 2005.

[3] "European declaration on cycling," European Commission, 2023. [Online]. Available: https://transport.ec.europa.eu/system/files/2023-11/European_Declaration_on_Cycling_en_0.pdf

[4] M. J. Nieuwenhuijsen and H. Khreis, "Car free cities: Pathway to healthy urban living," *Environment international*, vol. 94, pp. 251–262, 2016.

[5] J. J. De Hartog, H. Boogaard, H. Nijland, and G. Hoek, "Do the health benefits of cycling outweigh the risks?" *Environmental health perspectives*, vol. 118, no. 8, pp. 1109–1116, 2010.

[6] "Planning suitable transport networks for e-scooters to foster micromobility spreading," *Sustainability*, mdpi.com. [Online]. Available: https://www.mdpi.com/2071-1050/12/20/8346

[7] "Historic milestone: 54 countries adopt the pan-european master plan for cycling promotion," eCF. [Online]. Available: https://ecf.com/news-and-events/news/historic-milestone-54-countries-adopt-pan-european-master-plan-cycling-promotion

[8] "Political priorities for micro-mobility 2024-2029," micro-mobility for Europe. [Online]. Available: https://micromobilityforeurope.eu/political-priorities-for-micro-mobility-2024-2029

[9] M. Fischer, M. Fleischer, M. Mellauner, K. Machata, and A. Soteropoulos, "Best practices in cycling infrastructure: Strategies, planning, implementation, maintenance

and assessment," EuroRAP Institute on behalf of SABRINA Project Partners, Dunajska cesta 128, 1000 Ljubljana, Sep. 2022, with contributions from all SABRINA Project Partners. Contact: Olivera Rozi, Project Director (olivera.rozi@eurorap.org). Graphic design by Identum Communications GmbH, Vienna. Image credits: iStock, SABRINA Project Partners. [Online]. Available: www.eira-si.eu

[10] R. de Groot, *Design Manual for Bicycle Traffic*, illustrated, revised ed., ser. C.R.O.W. record. CROW, 2016, vol. 28.

[11] F. H. Administration, *Bicycle and Pedestrian Data Collection Manual*. U.S. Department of Transportation, 2016.

[12] ——, *Separated Bike Lane Planning and Design Guide*. U.S. Department of Transportation, 2015.

[13] E. C. Federation, *Cycle Infrastructure Design*. ECF, 2019.

[14] ——, *Bike Barometer*. ECF, 2018.

[15] N. A. of City Transportation Officials, *Urban Bikeway Design Guide*. NACTO, 2014.

[16] ——, *Designing for All Ages and Abilities*. NACTO, 2017.

[17] Eco-Counter, "Tracking the bike boom: Interactive bike count dashboard," https://www.eco-counter.com/cycling-data-tracker/, 2024, accessed: 2024-06-13.

[18] F. Kon, É. C. Ferreira, H. A. de Souza, F. Duarte, P. Santi, and C. Ratti, "Abstracting mobility flows from bike-sharing systems," *Public Transport*, vol. 14, no. 3, pp. 545–581, 2022.

[19] L. G. Natera Orozco, F. Battiston, G. Iñiguez, and M. Szell, "Data-driven strategies for optimal bicycle network growth," *Royal society open science*, vol. 7, no. 12, p. 201130, 2020.

[20] M. Szell, S. Mimar, T. Perlman, G. Ghoshal, and R. Sinatra, "Growing urban bicycle networks," *Scientific reports*, vol. 12, no. 1, p. 6765, 2022.

[21] A. Vybornova, T. Cunha, A. Gühnemann, and M. Szell, "Automated detection of missing links in bicycle networks," *Geographical Analysis*, vol. 55, no. 2, pp. 239–267, 2023.

[22] C. Steinacker, D.-M. Storch, M. Timme, and M. Schröder, "Demand-driven design of bicycle infrastructure networks for improved urban bikeability," *Nature computational science*, vol. 2, no. 10, pp. 655–664, 2022.

[23] R. Gerike, S. Weikl, C. Koszowski, and K. Bogenberger, "Network level design for cycling," in *Advances in Transport Policy and Planning*. Elsevier, 2022, vol. 10, pp. 77–109.

[24] P. Folco, L. Gauvin, M. Tizzoni, and M. Szell, "Data-driven micromobility network planning for demand and safety," *Environment and planning B: Urban analytics and city science*, vol. 50, no. 8, pp. 2087–2102, 2023.

[25] P. Ye, B. Wu, and W. Fan, "Modified betweenness-based measure for prediction of traffic flow on urban roads," *Transportation Research Record*, vol. 2563, no. 1, pp. 144–150, 2016.

[26] B. Resch and M. Szell, "Human-centric data science for urban studies," p. 584, 2019.

[27] J. Bao, T. He, S. Ruan, Y. Li, and Y. Zheng, "Planning bike lanes based on sharing-bikes' trajectories," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 1377–1386.

[28] S. McDaniel, M. B. Lowry, and M. Dixon, "Using origin–destination centrality to estimate directional bicycle volumes," *Transportation Research Record*, vol. 2430, no. 1, pp. 12–19, 2014.

[29] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, pp. 35–41, 1977.

[30] D. Delling, P. Sanders, D. Schultes, and D. Wagner, "Engineering route planning algorithms," in *Algorithmics of large and complex networks: design, analysis, and simulation*. Springer, 2009, pp. 117–139.

[31] J. Broach, J. Dill, and J. Gliebe, "Where do cyclists ride? a route choice model developed with revealed preference gps data," *Transportation Research Part A: Policy and Practice*, vol. 46, no. 10, pp. 1730–1740, 2012.

[32] L. C. Commons. Comune di padova. [Online]. Available: http://padovanet.it

[33] G. Boeing, "Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks," *Computers, environment and urban systems*, vol. 65, pp. 126–139, 2017.

[34] F. Foti, P. Waddell, and D. Luxen, "A generalized computational framework for accessibility: from the pedestrian to the metropolitan scale," in *Proceedings of the 4th TRB conference on innovations in travel modeling. Transportation research board*, 2012, pp. 1–14.

[35] A. Hagberg, P. J. Swart, and D. A. Schult, "Exploring network structure, dynamics, and function using networkx," Los Alamos National Laboratory (LANL), Los Alamos, NM (United States), Tech. Rep., 2008.

[36] E. W. Dijkstra, "A note on two problems in connexion with graphs," in *Edsger Wybe Dijkstra: His Life, Work, and Legacy*, 2022, pp. 287–290.

[37] A. Shimbel, "Structural parameters of communication networks," *The bulletin of mathematical biophysics*, vol. 15, pp. 501–507, 1953.

[38] "Kepler.gl: A data-agnostic webgl-powered application for visual exploration of large-scale geospatial data sets," https://kepler.gl/.

Acknowledgments

After two years in the Data Science Master's program at the University of Padova, this thesis marks the culmination of an enriching academic journey filled with challenges, growth, and discovery—both professionally and personally.

I would like to express my deepest gratitude to Professor Silvestri for his invaluable guidance throughout the thesis process, playing a pivotal role in the successful completion of this research project. I am also thankful to the Mobility Research Group, where I had the opportunity to engage in insightful discussions during my internship, exploring various research papers and exchanging ideas. Special thanks go to Ridemovi for providing the essential data that formed the backbone of this study's methodology.

To my husband, Mario: Thank you for your constant support and encouragement. Your technical assistance and thoughtful discussions around my thesis have been invaluable, pushing me to think deeper and refine my work.

To my family—Mom, Dad, and my siblings Elvis, Eduardo, and Karen: Thank you for always supporting me in every academic decision and motivating me to continue growing, never considering this to be my final step in learning.

My master's studies, and consequently this research, were funded by the National Agency for Research and Development (ANID)/ Scholarship Program / MAGISTER BECAS CHILE EN EL EXTRANJERO/2022 – 73220397.