POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Bike-sharing: data-driven approach for the development of SSEI, a new station effectiveness indicator

TESI DI LAUREA MAGISTRALE IN
MOBILITY ENGINEERING

Author: **Nicolò Pozzati**

Student ID: 242262
Advisor: Fabio Borghetti
Academic Year: 2024-25

Abstract

Bike-sharing systems play an increasingly important role in urban mobility, supporting short-distance trips, last-mile connectivity, and the integration with public transport. Despite their diffusion, the planning and evaluation of station-based bike-sharing services are challenging, due to the difficulty of accurately predicting demand and the strong dependence of existing approaches on operational variables, such as bikes and docks availability at stations and fleet rebalancing strategies. Moreover, many studies focus on single case studies, limiting the interpretability and applicability of results in other contexts.

This thesis proposes a data-driven methodological approach aimed at identifying the main urban context factors influencing bike-sharing demand and at developing a new quantitative indicator, named *Station Static Effectiveness Index* (SSEI), which is designed to evaluate the effectiveness of bike-sharing stations on a scale ranging from 0 (critical) to 100 (excellent), while remaining independent of service operational characteristics. The analysis is based on a multi-city dataset comprising 140 bike-sharing stations located in several Italian cities, characterised by a set of objective variables that describe land use, Points of Interest (POIs) density and integration with transportation infrastructure. Machine learning techniques, such as linear regression models and ensemble methods, are implemented in Python programming language and then applied to investigate the relationship between these contextual variables and observed demand.

Outcomes show that bike-sharing demand is primarily influenced by factors related to the functional density of the station's surroundings, such as the presence of points of interest for personal needs, tourism, and leisure activities, as well as by distance from the city centre and, moderately, by the integration with individual and public transportation networks. Conversely, physical characteristics of the stations, such as lighting or weather protection, are found to have a limited impact on demand. Building on these findings, the *Station Static Effectiveness Index* is calibrated using a hybrid approach that combines statistical methods – such as correlation analysis and Principal Component Analysis – with survey-based techniques. Specifically, the calibration process relies on a survey conducted on 127 respondents from all over the world, including both bike-sharing users and non-users, aimed at identifying the main drivers of demand. At the same time, applying the same machine learning algorithms, analyses are conducted on local demand in each examined city and on the elements that most affect user satisfaction.

The new index is then applied through two main case-study analyses, both set in Italian bike-sharing services. First, a stand-alone evaluation is performed to assess the current state of bike-sharing services, using the SSEI as a network auditing tool. Second, an integrated analysis combining the SSEI with a cycling infrastructure quality index is conducted to explore the consistency between station effectiveness and local bikeability conditions. Results highlight the ability of the SSEI to uncover spatial patterns and planning priorities that may not emerge from traditional performance indicators, in addition to its ability to support bike-sharing planning, evaluation, and decision-making processes.

Key-words: bike-sharing; mobility; sustainable mobility; bikeability; transportation planning; demand drivers; performance evaluation; urban integration; machine learning.

Abstract in italiano

I servizi di bike-sharing rivestono un ruolo sempre più importante nella mobilità urbana, incentivando gli spostamenti di breve distanza, facilitando la connettività del primo e dell'ultimo miglio e garantendo l'integrazione con la rete del trasporto pubblico. Nonostante la loro sempre più ampia diffusione, la pianificazione e la valutazione delle prestazioni dei sistemi di bike-sharing station-based presentano numerose criticità, dovute alla difficoltà di stimare accuratamente la domanda e alla forte dipendenza dalle variabili operative del servizio, quali la disponibilità di stalli e biciclette nelle stazioni e le strategie di bilanciamento della flotta. Inoltre, diverse ricerche si focalizzano su singoli casi studio, limitando l'interpretazione dei risultati e la loro applicazione in altri contesti.

La presente tesi propone una nuova metodologia data-driven, volta a identificare quali siano i principali fattori del contesto urbano che influenzano maggiormente la domanda di servizi di bike-sharing. Ciò consente lo sviluppo di un indice innovativo, il *Station Static Effectiveness Index* (SSEI). Esso si propone di quantificare l'efficacia dell'integrazione urbana di una stazione di bike-sharing in un intervallo compreso tra 0 (critica) e 100 (eccellente), indipendentemente dalle caratteristiche operative del servizio. A tal fine, si è costruito un dataset di 140 stazioni appartenenti a diverse città italiane, caratterizzato da un insieme di fattori oggettivi che descrivono il contesto di inserimento della stazione. Per comprendere meglio i legami tra queste variabili e la domanda, il dataset è stato analizzato attraverso appositi codici sviluppati in linguaggio Python che prevedono l'uso di algoritmi di machine learning, dalla regressione lineare ai metodi ensemble.

I risultati mostrano che la domanda è principalmente influenzata da fattori legati alla densità funzionale nei dintorni della stazione, quali la presenza di punti di interesse per necessità personali (ristoranti, negozi, ambulatori…), turismo o attività ricreative, nonché dalla distanza dal centro città e, in misura secondaria, dall'integrazione con la rete di trasporti. Al contrario, le caratteristiche fisiche delle stazioni, come l'illuminazione o la protezione dagli agenti atmosferici, hanno un impatto limitato sulla domanda. Sulla base di tali evidenze, il *Station Static Effectiveness Index* è stato calibrato mediante un approccio ibrido che integra metodi statistici – quali l'analisi di correlazione e l'Analisi delle Componenti Principali – con strumenti survey-based. In particolare, il processo di calibrazione si avvale di un'indagine condotta su un campione internazionale di 127 persone, comprendente sia utenti di bike-sharing sia non utenti, finalizzata a individuare i fattori che influenzano maggiormente la domanda. In parallelo, gli stessi metodi di machine learning sono stati utilizzati per analizzare le peculiarità della domanda nelle singole città esaminate e i fattori che maggiormente influenzano il gradimento da parte degli utenti.

Infine, allo scopo di verificarne la validità, si propongono due casi studio di applicazione del SSEI, entrambi ambientati in Italia. Nel primo, l'indice viene utilizzato singolarmente per valutare l'efficacia attuale di servizi bike-sharing. Nel secondo, invece, il SSEI è affiancato da un indicatore di ciclabilità per valutare la coerenza tra lo sviluppo del servizio e la qualità delle infrastrutture ciclabili presenti nei dintorni delle stazioni. I risultati mostrano l'abilità del SSEI di individuare pattern geografici e priorità d'intervento che potrebbero non emergere attraverso l'uso dei tradizionali indicatori di performance, nonché la sua utilità nel supportare la pianificazione, l'analisi e i processi decisionali relativi ai servizi di bike-sharing.

Parole chiave: bike-sharing; mobilità sostenibile; mobilità ciclabile; pianificazione dei trasporti; driver della domanda; valutazione delle performance; integrazione urbana; machine learning.

Contents

1 Introduction

Urban mobility systems are currently undergoing a phase of continuous transformation, driven by strict environmental constraints, increase in travel demand and by the need of effective first and last mile solutions. In this rapidly changing context, cities are required to adopt new solutions capable of improving accessibility while reducing congestion and environmental impacts. Shared mobility has emerged as a key paradigm shift, promoting a transition from private vehicle ownership to more flexible, usage-based mobility services that complement traditional public transport. Within this framework, bike-sharing systems stand out as one of the most widespread and effective shared mobility solutions, thanks to their suitability for short-distance trips, low spatial and environmental footprint, and role in enhancing first- and last-mile connectivity. This first section of the thesis aims to explore the aforementioned topics in depth, specifying the operative framework in which it was developed. Firstly, features of bike-sharing systems are analysed; then, the motivations and objectives of the study are outlined in detail.

1.1. Shared mobility

Shared mobility is defined as transportation services and resources that are shared among users, either concurrently or one after another [1] and it is one of the most promising trends that impacts the mobility market. Since the inception of the first sharing programs, the concept of shared mobility has profoundly reshaped the mobility market, challenging the traditional notion of vehicle ownership, which has long influenced people's mobility habits. Therefore, the primary impacts of shared mobility are expected to include a significant reduction in the number of circulating private vehicles, a consequent decrease in emissions, and the release of urban space previously allocated to vehicle parking.

Nowadays, it is possible to identify three typologies of shared mobility, which differ in vehicle typology and ownership:

- *Micromobility*: small lightweight vehicles such as bicycles, moped or electric scooters which are available for shared public use upon payment of a predefined fee. Usually, these services are supported by a digital platform through which users find and unlock vehicles as well as pay for trips;

- *Ride hailing*: the process of requesting a car with a driver for immediate use, typically facilitated through a mobile application [2]. The difference with traditional taxi service is that drivers can also be unlicensed, allowing common people to earn money by offering this service through their own personal vehicle;
- *Car sharing*: consumers can reserve and use company-provided cars, typically for a limited period. Also in this case, users are helped by digital platforms, which can geo-localise vehicles and parking areas.

All categories of shared mobility share the same important necessity: the support of an infrastructure, both physical and digital. It is fundamental to have a robust digital structure that can assist users at each stage of their journey, from planning to final payment. On the other hand, it is also extremely important to have a proper physical infrastructure that allows the full potential of shared mobility solutions to be exploited. Thus, many important elements such as road network, car parks, micromobility facilities (stations and dedicated lanes) and recharging columns must be organised and coordinated to provide shared mobility services the optimal environment to operate.

1.2. Bike-sharing

1.2.1. Definition and current framework

Among shared mobility opportunities, there is one that stands out significantly above the others in terms of collective benefits and potential to transform the world of public transport: bike-sharing.

Bike-sharing (BS) is defined as a type of transportation based on the use of public bikes to cover relatively short distances in urban areas. It is used both in conjunction with traditional public transport to complete the "last mile", or as an alternative for its flexibility [3]. Users can utilise the service by signing up for subscriptions of varying durations or using it as a single-use service, paying a predefined fee per minute. Often, for monthly or yearly subscriptions, bike-sharing schemes (BSS) also offer a certain number of free ride minutes at the beginning of each trip.

In 2022, BSS were present in almost 1600 cities around the world, and, in 2025, the shared bike fleet exceeded 10 million bikes for the first time in history [4]. Italy is one of the countries with the most developed bike-sharing schemes, as evidenced by the 2024 sharing mobility report [5] published by the sharing mobility national observatory. Rides taken on shared bicycles have increased in recent years, reaching almost 35 million km travelled in 2023, as presented in Figure 1. This means that Italian citizens greatly appreciate bike-sharing, even though the number of vehicles available is declining (Table 1) due to the parallel growth of scooter-sharing services.

Table 1. Bike-sharing schemes in Italy. Source: Osservatorio nazionale sharing mobility. [5]

Year	Number of services	Number of rentals	Number of vehicles
2015	24	5,854,144	8,410
2016	24	7,423,545	8,810
2017	31	10,389,553	31,200
2018	33	12,337,719	24,240
2019	39	12,753,037	33,370
2020	38	5,762,669	34,700
2021	40	8,059,976	27,630
2022	41	13,901,264	48,570
2023	47	15,561,130	38,840

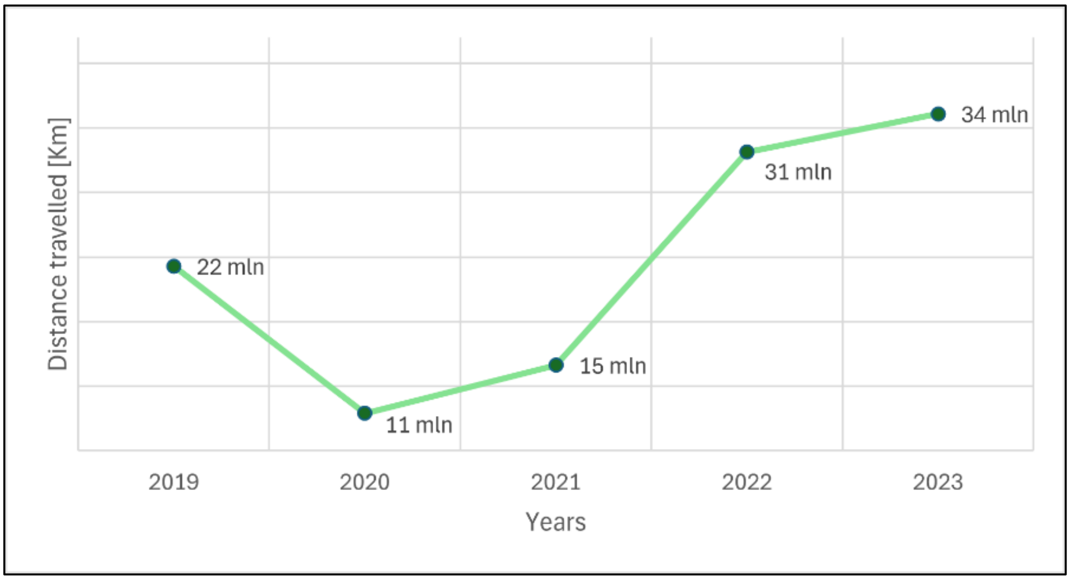


Figure 1. Overall travel distance by BS bikes in Italy. Source: Osservatorio nazionale sharing mobility. [5]

Bike-sharing is conquering the shared mobility world due to the many benefits it offers both to the individual citizen and to the community as a whole. The main ones are explained below, following the study conducted by Shaheen et al. [6]:

- *Modal shift and "last mile" connection*. Bike-sharing is a sustainable transport mode that can substitute other modes for short trips and seamlessly connect with public transit [7]. So, it could replace cars in the first or last part of the trip, the so-called "first and last mile", supporting commuters in planning trips and enhancing the completeness of the transportation network by adding new fictitious feeder lines to the main ones. For example, a worker could take the bike at a bike-sharing station near his home and ride towards the train station, where he will leave the bike and take the train. Thus, positioning bike-sharing stations near other public transport stops can help in creating a much more effective "hub and spoke" network;

- *Flexible mobility*. Users can take the shared bike whenever they want, by simply unlocking the bike using their smartphones. Thus, bike-sharing is not subject to timetables;

- *Emission reductions*. Bike-sharing is an environmentally friendly transportation since bicycles, both traditional and electric, do not emit any pollutants, unlike other public transportation options or private cars;

- *Congestion reduction*. This concept is strictly linked to the modal shift: if people use bicycles more, there would be fewer cars on the road, freeing up the roads and reducing daily traffic. At the same time, modal shift must be properly supported by the presence of bike lanes or roads suitable for cyclists to ensure their safety;

- *Individual financial savings*. Using bike-sharing is more cost-effective than travelling the same distances by other forms of public transport. However, this phenomenon is mainly limited to large cities and short journeys;

- *Bicycle ownership*. Bike-sharing programs solve some of the primary concerns of bicycle ownership, such as losses due to vandalism, lack of parking or storage, and maintenance requirements;

- *Health benefits*. Riding a bike everyday ensures people the right quantity of physical daily activity, improving their health and preventing cardiovascular problems.

1.2.2. History

The history of bike-sharing began in 1965, when the first BSS was introduced in Amsterdam, which is considered the "city of bikes" (Table 2). It consisted of a fleet of bikes, called "white bikes", openly available for free. However, the main drawback of this system was the absence of any kind of deposit station, so it collapsed in a few days. 30 years later, in 1995, the first large-scale scheme was introduced in Copenhagen. It was implemented using low-cost bikes and, for the first time, docking stations were introduced. Furthermore, it was not free of charge, but users had to make a coin deposit to use it. The first system with credit card access was opened in Paris in 1998. Instead, the fourth and latest generation of bike-sharing appeared on the market in 2005, the year in which the first large-scale system was introduced in Lyon, involving the use of electric bicycles as well as traditional ones, for a total of 1,500 bicycles. An important innovation of this new system was the ability to geo-localise bicycles in real time, reducing service disruptions and making balancing operations more effective.

Nowadays, the bike-sharing market is deeply consolidated, with three different typologies of services:

- *Station-based systems*, in which bicycles are picked up and dropped off at predefined points, called "docking stations". Stations are placed in cities' strategic locations, such as bus terminals, train stations or tourist points of interest. The use of these services is very sensitive to both bicycle and parking slot availability at stations. Thus, the main critical issues during the planning phase are the dimensioning of parking slots and the rebalancing of the bike fleet among the stations.

- *Dockless systems*. In this kind of system, there are no stations. Users can figure out where a bike is parked via a digital platform, use it, and then drop it off wherever they want inside a dedicated operative area, which is usually delimited by the city boundaries. Main issues of this typology are the software implementation for geo-localising and unlocking the bicycle, and the recovery of sparse bicycles.

- *Hybrid systems*. It could be considered as a station-based system with fictitious stations, since users can pick up and drop off bikes only in predefined and precisely limited areas, spread around the city. If this does not happen, the user will have to pay an additional charge for the service. Another form of hybrid system consists of those in which physical stations are present, but users can also park their bicycles outside of them. However, this typology is less widespread than the previous ones.

Table 2. Bike-sharing generations. Source: elaboration from Midgley, 2011. [8]

Generation	Year	Components	Features
First	1965	Bicycles	Distinct bicycles Unlocked bikes Free of charge No stations
Second	1995	Bicycles and docking stations	Distinct bicycles Locked bikes Coin access Specific stations
Third	1998	Bicycles and docking stations	Distinct bicycles Locked bikes Smart card access Specific stations Access kiosk
Third+	2005	Bicycles (also electric) and docking stations	Distinct bicycles Locked bikes Smart card access Specific stations Real-time availability Access kiosk GPS tracking

1.2.3. Literature themes

The bike-sharing literature covers the various stages of service planning, focusing on the critical issues relevant to each stage. Academic articles can be grouped into six clusters:

- *Demand drivers* [9] [10]. Planning a bike-sharing service is challenging due to the difficulty of predicting demand, which depends on modal shift, weather conditions, land use, and travel patterns. Demand varies significantly across space and time and cannot be reliably estimated using aggregate assumptions alone. For this reason, identifying the main drivers influencing bike-sharing demand is essential to reduce uncertainty and support effective system design and planning decisions.

- *Station allocation* [3] [11]. This topic is the most studied in the field of bike-sharing, as it originates from the application of operational research algorithms already used in other fields, such as the study of the best allocation of warehouses in distribution chains. In bike-sharing, many algorithms are applied to define the best configuration of the service network, while respecting constraints such as the total investment, demand coverage, or the need to ensure a geographically and socially equitable service.

- *Service equity* [12]. Strictly related to the previous one, papers about this theme try to verify whether the current network configuration is efficient from the point of view of guaranteeing the same opportunity to different levels of the social pyramid (vertical equity) and to different areas of the city (horizontal equity). Many studies are applied to existing services, but equity constraints could also be present in station allocation algorithms.
- *Bicycle allocation and station capacity* [13]. It concerns the definition of the number of parking slots and bikes to allocate in each docking station. These values should not be equal, due to rebalancing needs. In this phase, the main constraints are represented by elements studied in the previous steps, such as demand distribution or equity needs and the city's geographical features.
- *Stations rebalancing* [14]. It faces the challenge of moving bicycles between stations in order to constantly satisfy demand and users' needs. Rebalancing deals with constraints such as the number of available workers and repositioning vans.
- *Performance evaluation* [15] [16]. This is usually considered as an ex-post phase in which planners use or define dedicated indices which will be applied to evaluate the effectiveness of the bike-sharing system, based on different constraints already mentioned: demand coverage, vertical and horizontal equity, bicycle and parking availability, safety, and users' satisfaction.

1.3. Motivation

A significant portion of the existing research primarily focuses on either demand prediction or the assessment of operational efficiency. These studies often rely on factors such as origin-destination matrices, fleet availability, and rebalancing strategies. While these elements are crucial for managing services, they fall short in addressing early-stage planning decisions or providing a reliable comparative framework for evaluating stations in various contexts. Moreover, models developed for specific cities tend to be customised to local conditions, which diminishes their applicability in other locations and limits their efficacy as general decision-support tools. This highlights a distinct need for evaluation methods that are less constrained by short-term operational changes and more aligned with the structural characteristics of urban environments.

Additionally, the increasing availability of urban open data is fostering innovative approaches to examining bike-sharing systems from a context-oriented perspective. Data regarding land use, accessibility, functional density, and integration with transportation infrastructure is now more accessible than ever, allowing for a comprehensive assessment of potential station locations. However, synthesising this diverse array of information into a clear and coherent evaluation remains a challenge, particularly at the station level. In the absence of suitable analytical frameworks, there is a risk of oversimplifying complex urban dynamics or generating findings that are difficult to interpret and implement.

Considering these challenges, there is a compelling need to develop analytical methods that can accurately identify the influence of urban contextual factors on bike-sharing static performance, regardless of service operational features (strictly related to demand). Such methodologies could assist planners in determining whether service usage issues stem from inadequately located stations, insufficient infrastructure, or broader urban constraints. Ultimately, adopting a context-based evaluative perspective can yield a more cohesive understanding of bike-sharing systems, wherein the service's effectiveness is viewed as a product of both network design choices and the quality of the surrounding urban environment. This conceptual framework underpins our current research and justifies the creation of a framework aimed at facilitating more informed, clear, and adaptable planning decisions for bike-sharing services.

1.4. Objectives

This thesis lies at the intersection between demand drivers analysis, station allocation and performance evaluation. The primary goal is to deeply investigate the main relevant factors influencing bike-sharing demand through Machine Learning (ML) methods, adopting a new methodology which is completely independent of operational characteristics of the service, such as bikes and docks availability. Only factors referred to as ‘objective elements’ are taken into consideration. Such factors are *those aspects that are based on facts and therefore can be unequivocally measured after assuming some premises and uncertainties* [17]. Examples of objective factors include the presence and number of points of interest near the station, the distance from transport infrastructures, and the physical characteristics of the station itself. The selected approach involves creating a dataset of bike-sharing stations belonging to various bike-sharing services. It will be analysed using machine learning algorithms (linear regression and ensemble methods) to identify the urban context factors that have the greatest impact on BS demand. After that, the findings from this investigation are used to achieve the following objectives, listed in order of importance:

- Creation of a new composite index able to quantify the effectiveness of bike-sharing stations integration in urban areas in relation to demand;
- Examination of the most influential demand drivers at local level by breaking down the main dataset into single cities;
- Implementation of a new model that targets user satisfaction through the reviews they leave. This will make it possible to see whether the factors that influence satisfaction are different from those that influence demand.

The first-mentioned objective includes the most important innovation proposed by this research: the creation of a new effectiveness index, which will be called the *Station Static Effectiveness Index* (SSEI). Its first benefit is that it is completely independent of the Origin-Destination (OD) matrix in the study area. This makes it possible to overcome potential issues arising from the calculation of this matrix, which could be challenging when calculating for a cycling mode. Since the OD matrix is a proxy for demand, it can be stated that the SSEI is created on the basis of demand attractors but is independent of demand itself, intended as an indication of a user's movement from point A to point B. This greatly expands the possibilities for utilising the SSEI, making it easy to calculate in various application contexts. The main relevant potential applications are the following:

- *Stand-alone evaluation of service effectiveness*. Computation of the *Station Static Effectiveness Index* for each station of the service to assess the current state of the bike-sharing service. Thus, the SSEI is used as a network auditing tool, allowing planners to evaluate how effectively stations are integrated into their urban context independently of operational performance. A case study related to this application is discussed in Chapter 5.1.

- *Integrated SSEI-bikeability analysis*, through analysis of the trade-off between the proposed index and a bikeability indicator that reflects the quality of cycling infrastructure in the surrounding area. A bike-sharing service is more effective when supported by an adequate cycling network. Thus, this is an opportunity to assess the consistency of the bike-sharing service with urban development, evaluating priorities for action from both a service and infrastructure perspective. A case study related to this application is discussed in Chapter 5.2.

- *Checking of station allocation phase*. Calculating SSEI for each station of the service makes it possible to verify the previous design phase before proceeding to define the number of parking docks and bicycles for each station.

- *Multi-objective research of optimal station positioning*. The SSEI formula could be used as an objective function to be maximised during the station allocation phase, either on its own or in conjunction with other techniques already in use, such as maximal demand coverage or minimisation of impedance between possible trips in the study area. Furthermore, even at this stage, it is possible to consider a bikeability index in parallel, using a common formulation to define a new objective function.

To summarise the above considerations, the SSEI is proposed as a new quantitative indicator that measures the effectiveness of a bike-sharing station's integration into its surrounding context and in relation to bike-sharing demand, while being independent of demand in its formulation. In this way, it allows for the evaluation of the "static" performance of a station without explicitly considering its operational characteristics (intended as "dynamic" elements). Therefore, SSEI can be applied as an evaluation parameter of the current state of the service, as an evaluation parameter of the station allocation phase, or as an element integrated into the algorithms used in the same phase.

1.5. Research structure

The thesis is structured as follows. The next section analyses the state of the art and the main sources from which the study draws inspiration, particularly with reference to the index definition already mentioned. Section 3 is dedicated to the methodological framework: without recourse to numerical values, the data-driven approach will be presented from a purely theoretical point of view. First, we will focus on the creation of the SSEI, then shift our attention to the analysis of local demand and user satisfaction. Section 4 puts what has been proposed into practice, using the method to perform the data analysis and models necessary to calibrate the index and to perform the other two complementary examinations. Finally, Section 5 will present two case studies, suggesting two applications of the newly created SSEI. The first is aimed at calculating the index for the stations that make up the initial dataset, simulating a static performance evaluation on an already operative service. The second studies the trade-off between the SSEI and a cycling quality index for bike-sharing stations in the proximity of the Brescia underground stations. Conclusions, further developments and limitations of the study will be presented in the final pages of the document.

2 State of the art

This section focuses on the presentation of academic studies which have been useful to identify the literature gap and how to fill it. The chapter follows the flow suggested by Figure 2: the first part will be devoted to Bike-sharing demand drivers, while the other two examine the literature about BS station allocation methods and performance evaluation to understand where the SSEI could fit in and how it could solve various problems. Thus, the topics are presented in the order of the design phase. Although this thesis will not present applications of the SSEI in the station allocation phase, the literature on this topic is nevertheless analysed to understand the factors that converge in this planning phase and the innovations introduced by the new index. Lastly, please note that machine learning methods for data analysis will be briefly presented in Chapters 3.3.2 and 3.3.3, starting with linear regression and progressing to ensemble methods.

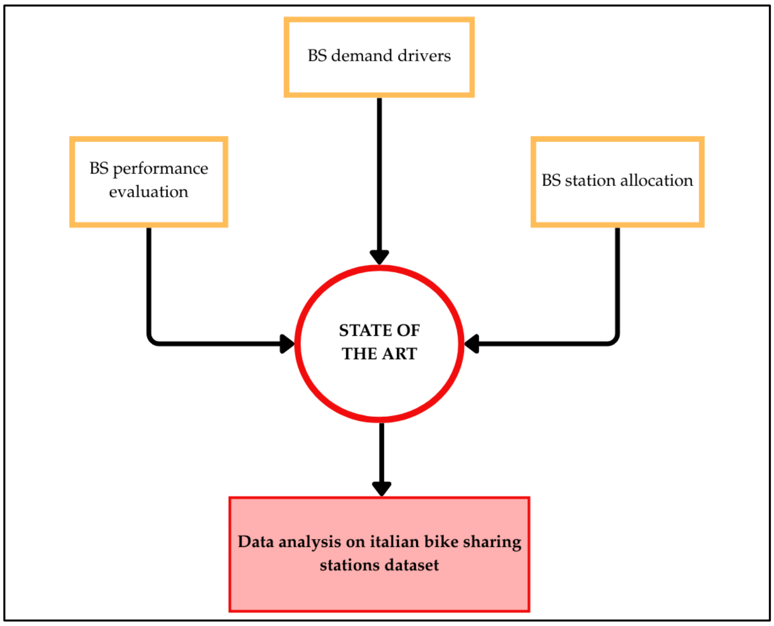


Figure 2. State of the art workflow.

During the literature review phase, 42 academic articles were reviewed. They are listed in Table 3 and 4 and were mainly consulted through portals such as ResearchGate [18] and Elsevier [19]. For each of the topics presented above, the following sub-chapters outline the articles that were most relevant to the development of this study.

Table 3. State of the art sources.

Ref.	Author(s)	Year	Title	Topic
[20]	Harkey et al.	1998	The Bicycle Compatibility Index: a level of service concept, implementation manual	Bikeability
[21]	Paul DeMaio	2009	Bike-sharing: History, Impacts, Models of Provision, and Future	Bike-sharing Overview
[22]	Larsen et al.	2011	Build it. But where? The use of geographic information systems in identifying locations for new cycling infrastructure	Station allocation
[8]	Peter Midgley	2011	Bicycle-sharing schemes: enhancing sustainable mobility in urban areas	Bike-sharing Overview
[23]	Enrica Papa, Pierluigi Coppola	2012	Gravity-Based Accessibility Measures for Integrated Transport-Land Use Planning (GraBAM)	Accessibility
[24]	Alexander Rixey	2012	Station-Level Forecasting of Bike-sharing Ridership: Station Network Effects in Three U.S. Systems	Demand drivers
[11]	Garcia-Palomares et al.	2012	Optimizing the location of stations in bike-sharing programs: A GIS approach	Station allocation
[25]	Pizzolorusso Giancarlo	2013	Sviluppo di un indice di qualità ciclistica di strade e piste ciclabili urbane	Bikeability
[26]	Herbie Huff, Robin Liggett	2014	The Highway Capacity Manuals Method for Calculating Bicycle and Pedestrian Levels of Service: the Ultimate White Paper	Bikeability
[3]	Edoardo Croci, Davide Rossi	2014	Optimizing the position of bike-sharing stations. The Milan case.	Station allocation
[27]	Chen et al.	2015	Bike-sharing Station Placement Leveraging Heterogeneous Urban Open Data	Station allocation
[28]	Ines Frade, Anabela Ribeiro	2015	Bike-sharing stations: A maximal covering location approach	Station allocation
[29]	Mateo-babiano et al.	2016	How Does Our Natural and Built Environment Affect the Use of Bicycle Sharing?	Demand drivers
[30]	Munkàcsy et al	2017	Potential User Profiles of Innovative Bike-Sharing Systems: The Case of BiciMAD (Madrid, Spain)	Demand drivers
[31]	Cristian Kloimullner, Gunther Raidl	2017	Hierarchical Clustering and Multilevel Refinement for the Bike-Sharing Station Planning Problem	Station allocation
[16]	Albinski et al.	2018	Performance analysis of a hybrid bike-sharing system: A service level-based approach under censored demand observations	Performance evaluation
[32]	Soriguera et al.	2018	A simulation model for public bike-sharing systems	Station allocation
[33]	Rito Brata Nath, Tarun Rambha	2018	Modelling Methods for Planning and Operation of Bike-Sharing Systems	BS management
[34]	Liu et al.	2019	Utilizing Bicycle Compatibility Index and Bicycle Level of Service for Cycleway networks	Bikeability
[35]	Chen et al.	2019	Locating stations in bike-sharing service: a special maximal covering location problem	Station allocation
[36]	Fang et al.	2019	A location-routing problem for the public bike-sharing system with service level	BS management
[15]	Bhuyan et al.	2019	GIS-Based Equity Gap Analysis: Case Study of Baltimore Bike Share Program	Performance evaluation
[37]	Song et al.	2020	Impact Evaluation of Bike-Sharing on Bicycling Accessibility	Accessibility
[38]	Kazemzadeh et al.	2020	Expanding the Scope of the Bicycle Level-of-Service Concept: A Review of the Literature	Performance evaluation
[17]	Maria Nogal, Pilar Jimenez	2020	Attractiveness of Bike-Sharing Stations from a Multi-Modal Perspective: The Role of Objective and Subjective Features	Performance evaluation

Table 4. Cont. state of the art sources.

Ref.	Author(s)	Year	Title	Topic
[12]	Duran-Rodas et al.	2020	Howfair is the allocation of bike-sharing infrastructure? Framework for a qualitative and quantitative spatial fairness assessment	Station allocation
[39]	Caggiani et al.	2020	An equality-based model for bike-sharing stations location in bicycle-public transport multimodal mobility	Station allocation
[40]	Francesc Soriguera, Enrique Jimenez-Merono	2020	A continuous approximation model for the optimal design of public bike-sharing systems	Station allocation
[41]	Kong et al	2020	Deciphering the relationship between bikesharing and public transit: Modal substitution, integration, and complementation	BS integration with public transit
[42]	Nikiforiadis et al.	2020	Determining the optimal locations for bike-sharing stations: methodological approach and application in the city of Thessaloniki, Greece	Station allocation
[43]	Kamel et al.	2020	A composite zonal index for biking attractiveness and safety	Performance evaluation
[44]	Leonardo Caggiani, Rosaria Camporeale	2021	Accessibility indicators for fair bike-sharing systems based on level of service	Performance evaluation
[9]	Maas et al.	2021	Spatial and temporal analysis of shared bicycle use in Limassol, Cyprus	Demand drivers
[39]	Torrisi et al.	2021	Exploring the factors affecting bike-sharing demand: evidence from student perceptions, usage patterns and adoption barriers	Demand drivers
[45]	Beairsto et al.	2022	Identifying locations for new bike-sharing stations in Glasgow: an analysis of spatial equity and demand factors	Station allocation
[46]	Kyounguk Kim	2023	Investigation of modal integration of bike-sharing and public transit in Seoul for the holders of 365-day passes	BS integration with public transit
[47]	Hsu et al.	2023	A Hybrid Model for Evaluating the Bikeability of Urban Bicycle Systems	Bikeability
[48]	Ceccarelli et al.	2023	Learning from Imbalanced Datasets: The Bike-Sharing Inventory Problem Using Sparse Information	BS management
[49]	Song et al.	2024	A station location design problem in a bike-sharing system with both conventional and electric shared bikes considering bike users' roaming delay costs	Station allocation
[10]	Dziecielski et al.	2024	Understanding the determinants of bike-sharing demand in the context of a medium-sized car-oriented city: The case study of Milton Keynes, UK	Demand drivers
[50]	Sebastian L. Gruner, Matthias Kowald	2025	Bike-sharing, why not? A framework of utility perceptions of BSSs' non-users based on qualitative data	Demand drivers
[51]	Roig-Costa et al.	2025	Unpacking the docked bike-sharing experience. A bike-along study on the infrastructural constraints and determinants of everyday bike-sharing use	Demand drivers

2.1. Bike-sharing demand drivers

2.1.1. Spatial and temporal analysis of shared bicycle use in Limassol, Cyprus [9]

Taking as a reference the Mediterranean city of Limassol, Cyprus, the authors investigate the main drivers of bike-sharing demand, analysing historical data about trip data in the city. They use GIS data to study the presence of points of interest in a 300 m buffer around stations. Then, regression models have been applied to identify the influence of such factors upon users' choices in terms of where, when, and why to interact with bike-sharing schemes.

Explanatory variables included in the regression model are clustered into four groups, presented in the next Table 5.

Table 5. Explanatory variables. Source: [9]

Category	Variables
Land use variables	Residential land use; commercial land use; number of hotels, restaurants, beaches, parks, university buildings, clothes shops; length of cycling paths; distance from the nearest cycling path; distance from nearest coastline; distance from the nearest bus stations; distance from the nearest university building; count of nodes in transport network.
Socio-economic variables	Population density; percentage of residents with tertiary education; percentage of Men in M/F ratio; percentage of population over 65 years; percentage of unemployed and foreign people.
Network variables	Station distance from the centre of the BSS; number of stations in 600m buffer around each station; number of stations in 1200m buffer around each station.
Temporal variables	Average monthly maximum temperature; Total monthly precipitation; total monthly tourist arrivals in Cyprus.

Random forest regressor allows for the discovery of interesting insights, such as:

- Cycling path availability and bike lanes network quality have a much more positive effect on BSS use;
- Areas with higher population density and attractive activities encourage demand;
- Both maximum temperature and quantity of rain have a negative impact on demand;
- Other variables with a negative influence on BSS use are: number of hotels in the buffer, presence of bus station, station distance from BSS centre and the percentage of people with tertiary education.

2.1.2. Attractiveness of Bike-Sharing Stations from a Multi-Modal Perspective: The Role of Objective and Subjective Features [17]

This paper still tries to study demand influencing factors, proposing a new perspective, categorising factors into objective and subjective. The former are those aspects which can be unequivocally measured, and so they don't depend on the user's experience; on the other hand, the latter are factors which are subject to the user's perceptions and preferences. Furthermore, variables objectivity is explained using the concept of safety and security (Table 6). Safety, the grade in which people are protected from a hazard, is a proxy of objective factors, while security is linked to subjective ones, since it depends on user's perception.

Table 6. Explanatory variables. Source: [17]

Variable	Safety/security
Land topography	Safety
Traffic interaction	Safety
Bike lane layout	Safety
Environmental conditions	Safety and security
Presence of facilities to ride comfortably (resting areas, benches, panel information…)	Security
Weather protection	Safety and security
Presence of point of interest nearby (shopping areas, parks, landmarks…)	Safety
Walkway layout (referred to people who walk to the station to pick up the bike)	Safety
Presence of BSS station within a 600m buffer	-
Presence of BSS-related facilities (information panels, screens and contact telephone)	Security

Presented variables are studied in relation to pedestrians, cyclists and BSS stations, and a case study in Dublin is analysed. Authors conclude that planning policies should be based on a combination of objective and subjective metrics to guide decisions that are more sensitive to users' needs and more representative of reality. The proposed tools will also allow for reducing re-balancing costs and maximising user satisfaction, which the authors intend as an indirect proxy of urban sustainability.

2.1.3. Understanding the determinants of bike-sharing demand in the context of a medium-sized car-oriented city: The case study of Milton Keynes, UK [10]

Milton Keynes is the biggest of the planned cities built under the UK government's "New towns" programme, but it is characterised by being a car-centric city. For this reason, authors conduct an analysis of bike-sharing rentals for over one year in Milton Keynes, identifying the most relevant factors influencing BSS demand and users' behaviours in different months and seasons. This will help planners in boosting bike-sharing in the city, efficiently meeting population needs.

In this case, the considered variables are:

- Number of public transport stops in the catchment area;
- Number of universities in the catchment area;
- Number of touristic attractions in the catchment area;
- Presence of Park in the catchment area;
- Number of office blocks in the catchment area;
- Number of schools (primary and secondary) in the catchment area;
- Number of malls in the catchment area;
- Number of cinemas in the catchment area;
- Number of residents aged 15 or more per hectare divided by 1000;
- Length of cycleway (in km) in catchment area.

Data exploratory analysis on bike rentals demonstrated that for every season, the distribution of rentals during the day is bimodal, with two peaks at 8.00 and 17.00, respectively. This suggests that a not negligible part of bike-sharing rentals is for work commuting. The weather has a high impact on rentals, too. In fact, the average number of rentals during summer months is over three times as high as for the winter months. Moreover, regression analysis shows that the number of public transport stops, offices and universities are positively associated with BSS use, while the presence of malls, schools and population density are the variables which have the most negative impact.

2.1.4. Discussion

Among academic studies on bike-sharing, the topic of factors influencing demand has not been explored in detail. Among the three papers presented and others in the literature [51] [39] [46], two important common features can be identified:

- The variables considered are very similar, such as integration with public transport and the presence of tourist or educational attractions. However, the number of variables included in each model is often small, which could lead to other important elements being neglected (e.g., the number of transport lines in the catchment area, proximity to healthcare facilities, car parking, the presence of a restricted traffic zone in the area, morphological characteristics, etc.). Furthermore, only in some cases, such as [17], intrinsic characteristics of the station are considered, such as the type of positioning, the presence of information panels, or protection from weather conditions.
- In all cases, the model is trained using data from a single bike-sharing service. This significantly limits the possible applications of the model, since in other cities, with different service characteristics and urban layout, the considerations made may no longer be valid. In other words, it can be asserted that the model is overly adapted (overfitted) to the service on which it is trained and when applied to other cities, the model may result in distorted results.

2.2. Bike-Sharing station allocation

2.2.1. Optimizing the location of stations in bike-sharing programs: A GIS approach [11]

This article proposes a GIS (geographical information system) based method to determine the optimal location for bike-sharing stations, identifying their most relevant characteristics and assessing the accessibility of each of them. The main starting hypothesis is that the distance between the origin or destination point and the BS station must be considerably small to gain user acceptance, as well as the distance between stations should be appropriate to allow trips by bicycle. The authors proposed a four-stage model for optimal station location in a bike-sharing scheme. The steps are reported below:

1. Analysis of the distribution of potential demand;
2. Application of location-allocation models, using data from candidate locations. Solutions are computed for two algorithms in parallel: p median, which minimises impedance (weighted travel costs) between demand points, and maximise coverage model, which tries to maximise the portion of demand covered by the service;
3. Station capacity and characterisation. After having analysed simulation scenarios (which differ in the overall number of stations), bicycles and docks for each station are allocated. Moreover, stations are clustered based on the ratio between attracted trips and the overall number of trips, creating four groups: Generators, attractors, mixed, and high attractors;
4. Lastly, an accessibility analysis for each station is performed using an indicator which relates the accessibility of a location directly with the number of opportunities available and inversely with the distance needed for those opportunities to be taken.

In conclusion, the article demonstrates that GIS-based location–allocation models can effectively optimise bike-sharing station placement, improving demand coverage and accessibility, but with diminishing returns as stations increase. Its innovation lies in integrating spatial demand analysis with optimisation models, enabling both the identification of optimal locations and the functional characterisation of stations for more efficient planning.

2.2.2. Bike-sharing stations: A maximal covering location approach [28]

This research, staged in Coimbra, a Portuguese university city, develops an optimisation method to design bike-sharing services based on the demand coverage maximisation, taking the public investment as the major constraint. Furthermore, it also proposes a method to define the number of docks and allocated bicycles for each station and to rebalance them during the day.

The model receives as inputs demand in the study area, maximum and minimum capacity of the stations, the price of the stations and bicycles, the relocations and maintenance costs, the total investment budget and the annual supplementary budget, as well as the discount and growth rate and the project's horizon years. In this way, authors pay great attention to financial factors which influence the overall service life cycle. Maximising the scalar product between the proportion of demand covered in a zone and demand in the same zone, the model simultaneously computes the number of stations in each zone, the capacity of the stations, the number of bicycles in each zone, and, to relocate in each time step, the fleet size, the annual revenue and the annual expenses.

However, unlike other models found in the literature, this one does not compute the exact station location, but it only identifies areas in which it is recommended to place a station. This could be a remarkable drawback of this model. In fact, authors suggest using the model in coordination with others which can minimise the distance between people and stations and to compute the exact position of the stations.

2.2.3. Determining the optimal locations for bike-sharing stations: methodological approach and application in the city of Thessaloniki, Greece [9]

The article investigates the problem of station allocation from the operator's perspective. It elaborates a methodological approach for the maximisation not only of the demand, but also of the area (i.e. built environment) coverage, as well as minimising the needs for bike redistribution.

Adopting this methodology, the optimal selection of location for bike-sharing stations is set as a multi-objective problem. Even in this case, the proposed model is structured in four stages:

1. Determination of candidate stations;
2. Selection of criteria that will be taken into account in the objective functions. As mentioned before, the three selected criteria are: demand coverage, area (built environment) coverage, and bikes re-distribution needs;

3. Definition of constraints. This phase is crucial because it must reflect the operator's strategy. Thus, operators must carefully select constraints based on their available sources. In the presented application, the defined constraints are: maximum monthly budget for renting public or private space, maximum number of new stations, minimum distance between stations (taking into account the existing stations), type of available space (public or private);
4. The last step is the mathematical formulation of the multi-objective optimisation problem. Three dedicated functions are formulated for each criterion selected before. Then, they are optimised in according to the chosen constraints.

One of the key features of multi-objective problems is the need to choose the weight given to each objective function. Authors recommend carefully calibrating these parameters because results could slightly differ changing their values. Lastly, authors anticipate future developments, which will also consider cycling infrastructure in cities.

2.2.4. Optimizing the position of bike-sharing stations. The Milan case [3]

This research differs significantly from the three previously presented. In fact, it does not present an algorithm that outputs the recommended location for a station. Rather, it presents a method that studies the most influential elements in the use of bike-sharing stations. For this reason, this article has significant methodological similarities with the idea presented in this thesis. However, the elements considered by the two authors are very few, unlike what will be explained in the following sections. In addition, the article is rather dated, having been published in 2014. Over the years, factors influencing bike-sharing demand may have changed.

Edoardo Croci and Davide Rossi, the two authors, develop a simple linear regression model to estimate the effect of attractors on the daily use of bike-sharing stations. Potential attractors considered are the following:

- Metro and train stations;
- Bus and tram stops;
- University;
- Museum;
- Cinemas and theatres.

Data about proximity and visibility from these attractors are processed, considering also weather data, computing the relative weight of each attractor on BS daily use.

In addition to the limited number of potential factors considered, another weakness of this model is the possibility that it may be subject to overfitting. Thus, it may be too closely tailored to the service for which it was calibrated (the city of Milan) and may not be applicable in other contexts. This could happen for two main reasons:

- The linear regression model does not include a penalisation term to avoid overfitting, such as what happens in Lasso or Ridge regression;
- The dataset is built using data from a single bike-sharing service.

2.2.5. Discussion

The common feature among most of the studies presented above and others analysed [45] is the strong dependence of the station allocation phase on travel demand in the study area. Many models need as input demand to perform algorithms such as maximal demand coverage or impedance minimisation. No model explicitly includes, either as an objective function or as constraints, the numerous characteristics of the context – such as integration with public transport and the cycle network or the presence of demand generators or attractors – which could play a significant role in the decision-making process for station location.

No model explicitly includes, either as an objective function or as constraints, the numerous characteristics of the context—such as integration with public transport and the cycle network or the presence of demand generators or attractors—which could play a significant role in the decision-making process for station location. Many might argue that these considerations are unnecessary, given that demand data is available. However, it is not always easy to obtain accurate data on demand, as it is the result of a time-consuming process and demand detectors, such as sensors from cellular networks, Wi-Fi, or credit cards, are not always reliable. Furthermore, the considered travel demand is generic and not targeted at individual bicycle trips. Thus, there is no exact knowledge of the portion of trips that are undertaken by bicycle.

2.3. Bike-sharing performance evaluation

2.3.1. Performance analysis of a hybrid bike-sharing system: A service level-based approach under censored demand observations [16]

One of the most widespread indicators used to evaluate bike-sharing stations' performance is the station level of service. The authors of this paper suggest a new methodology for calculating the level of service, highlighting the role of censored demand, i.e. demand not satisfied because of bicycle unavailability, in the computation of such an indicator.

First of all, a definition of service level is provided, indicating how many requests are successfully served with the available bikes. Then, they compute two indices which allow for an adequate measure of performance of the bike-sharing scheme:

- *α-service level*, which states the availability of the system, indicating the probability that all bike requests are fulfilled within a certain time threshold;
- *β-service level*, or fill rate, which indicates the number of requests that could be satisfied with the current available fleet.

Lastly, starting from the assumption that real demand is higher than the observed one, the authors propose a data-driven approach based on sales patterns about historical demand observations in order to boost the level of service computation and have a better understanding of system's performance, which is presented as much linked to the availability of both bikes and parking docks.

2.3.2. Accessibility indicators for fair bike-sharing systems based on level of service [44]

In this case, system performance is evaluated from a new perspective, that is, by analysing the relationship between the accessibility of the study area and the presence of the bike-sharing service. Thus, authors compute two new equity indicators to evaluate the accessibility of station-based services.

In this work, the authors aim to go beyond traditional accessibility measures based exclusively on distance and travel time. In fact, in their view, the mere presence of stations does not guarantee equitable access. Therefore, they have developed two indicators named *Bike-Sharing Equity*, one for active accessibility and the other for passive accessibility, considering multiple factors such as:

- Spatial accessibility: travel times have been adjusted considering walking distance thresholds (simulating people's willingness to walk to the station or to the destination);
- Station's level of service in terms of bicycles and docks availability;
- Social factors such as the presence of disadvantaged individuals (low income, no car owners, young and old people…) in the study area.

A high BSE indicates that an area enjoys reliable, equitable and continuous access to bike-sharing, while a low BSE indicates service deficiencies. Thus, BSE allows for the identification of peripheral or weak areas where bike-sharing is less accessible despite spatial coverage, showing that disadvantaged groups are at risk of having poorer access.

2.3.3. GIS-Based Equity Gap Analysis: Case Study of Baltimore Bike Share Program [15]

Since the expansion of bike-sharing schemes is now global, many questions arise regarding accessibility, equity, expansion and bikeability. Thus, the authors of this paper developed a new equity-based methodology for service planning. Using GIS tools, they compute a bike equity index and then analyse the trade-off between it and the level of the traffic stress index.

Bike Equity Index (BEI) is a composite measure calculated at the census block group level. Formulation is reported below:

$$BEI_i = E_i + Y_i + C_i + M_i + L_i$$

Where:

- E_i is the elderly population (over 65) index;
- Y_i is referred to the young population;
- C_i is related to car households;
- M_i indicates the amount of people minorities;
- L_i is linked to the low-income population.

A high BEI indicates a greater concentration of vulnerable populations and thus a stronger priority area for bike-sharing investment. Afterwards, the authors examine the trade-off between BEI and the *Level of Traffic Stress* (LTS), a bikeability indicator that measures the comfort and safety of cycling infrastructure, at the census block level. This approach makes it possible to pinpoint critical areas where cycling infrastructure fails to adequately support bike-sharing or where the service itself is underdeveloped. In this sense, the application of BEI is comparable to the index developed in this thesis, although the model variables differ substantially.

2.3.4. Discussion

Bike-sharing performance evaluation is a deeply investigated theme in the literature. Despite different possible approaches, the final goal is the same: identifying critical areas or stations where it is necessary to modify the current service. However, it is possible to recognise two important features of performance evaluation methods:

- Performance is often evaluated using the concept of level of service, which is strictly related to stations' capacity, which is modelled upon travel demand.
- If the level of service is not considered, the focus of the performance analysis is no longer the single station but a sub-portion of the entire study area.

Thus, there arises the need to set up a new method for performance analysis, which is referred to as the single station and is independent of travel demand. In this way, immediately after having decided the right exact station positioning, it may be possible to have an initial indication of its "static performance", using a new indicator based exclusively on the intrinsic station's characteristics and on the features of the urban context in which the station is inserted. This is exactly what we will be seeking to do in the following sections of this document.

2.4. Evidence of literature gap

The state of the art analysis presented in the previous paragraphs highlighted the most significant issues that arise in the study of bike-sharing systems. In particular, three major drawbacks can be highlighted for each area explored:

- Demand drivers: each article examined focuses on only one city or bike-sharing service. This can lead to an excessive adaptation of the obtained outcomes to the case study under consideration. It therefore becomes very difficult to apply the same methodology in other contexts.
- Station allocation: the algorithms in this phase aim to maximise demand coverage or minimise expenses. Therefore, the characteristics of the urban context in which the station will be located are not taken into consideration.
- Performance evaluation: currently used indicators all relate to the functioning and operational characteristics of the service, taking into account dynamic elements such as the availability of parking spaces and bicycles at each station. In contrast, the station's performance evaluation does not take into account the context in which it operates.

Consequently, the current state of the literature reiterates the need for and importance of a new data-driven method for studying the demand and the static performances of a bike-sharing service. In particular, this research aims to resolve the issues highlighted above by proposing a new methodological framework (Figure 3 and Figure 4), which includes:

- Collection of data from different bike-sharing services in a single dataset. This will enable the development of insights by simultaneously analysing different contexts with distinct social, cultural and geographical features. As a result, data analysis will be more robust, and it will be possible to replicate the proposed method in multiple case studies.
- Consideration of the various elements of the urban context in which a station is located, such as the number of attractions in the area surrounding the station, accessible transport lines or the physical characteristics of the station itself (the objective elements mentioned in [17]). In this way, previously ignored elements can be studied, investigating which ones have the greatest influence on demand and user satisfaction. Furthermore, by developing what has been elaborated, it will help to create new methodologies for the station allocation phase.

- The development of a new synthetic indicator that can quantify the quality level of the station's urban integration, regardless of the operational characteristics of the service.

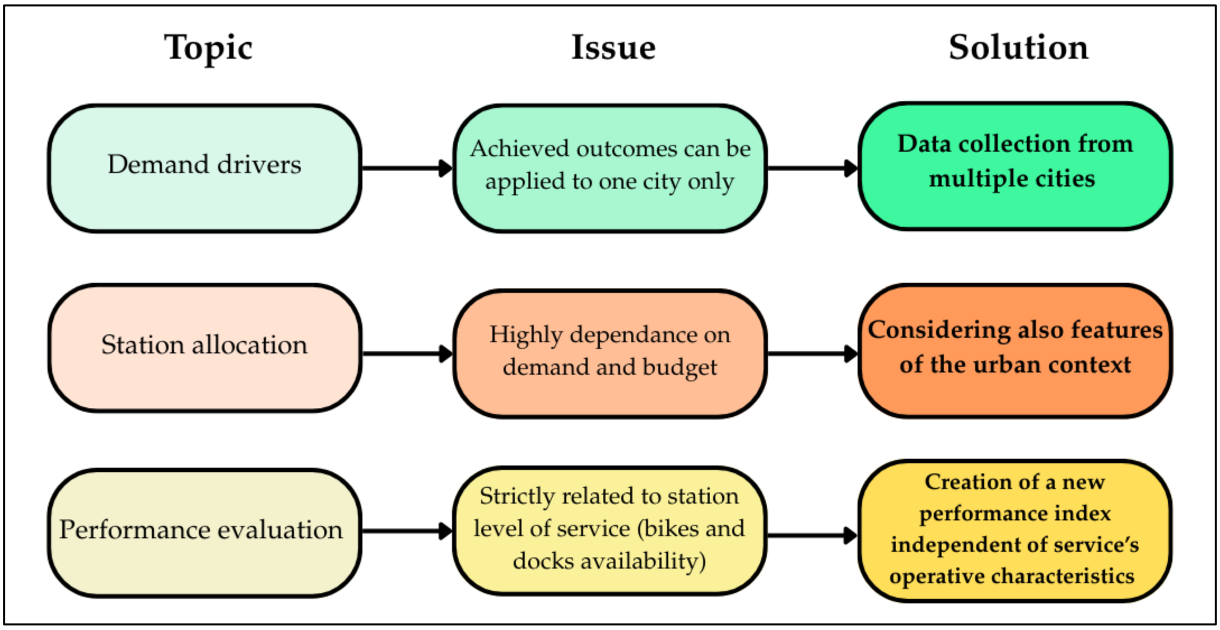


Figure 3. State of the art: topics, issues and solutions.

Thus, the thesis proposes a new data-driven methodology for studying the demand for bike-sharing services through machine learning methods. This analysis, which is more robust than those found in the literature, will offer a wide range of possible further developments. Among these, the most relevant innovation is the formulation of a new performance index: the *Station Static Effectiveness Index* (SSEI). This index, developed in Chapter 4.4, would allow service planners to quantify the quality of the station's positioning. At the same time, it allows the identification of intervention priorities for improvement in the areas where the service is active. In addition, the equation of SSEI could also be used during the station allocation phase as an objective function of the algorithm, ensuring the introduction of new elements considered in this first phase of service planning.

In addition to the index calibration, the proposed method allows two other important analyses to be carried out: the study of demand at the local level (Chapter 4.5), separating the initial dataset into individual cities, and the analysis of the factors that influence user satisfaction, presented in the Chapter 4.6, which will make it possible to verify whether these differ from those that influence demand.

3 Methodological framework

The steps taken to achieve the previously identified objectives and to fill the literature gap are presented in this section. The order of chapters reflects the workflow used to develop the entire thesis, following the various stages of the research, outlined in Figure 4.

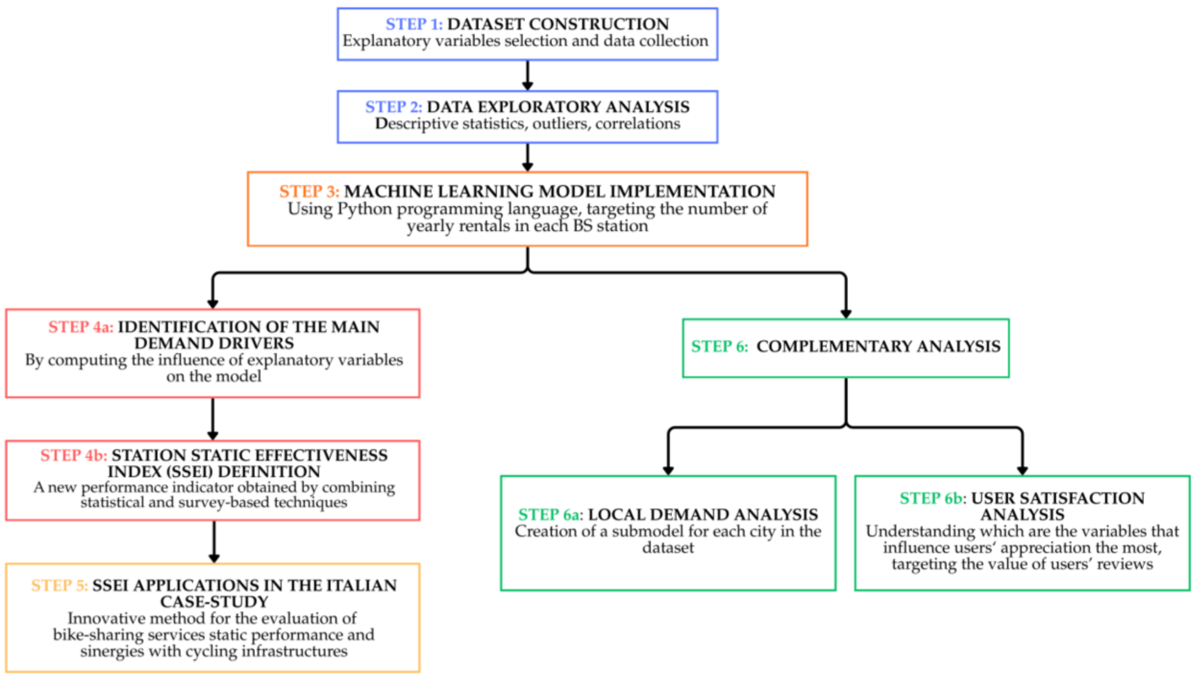


Figure 4. Study workflow.

Since this section is intended as the first approach to the technical part of the thesis, no numerical values will be included. The chapter aims to help readers understand the contents developed in the following sections at a purely theoretical level, so that the study can be easily replicated in other contexts.

3.1. Dataset construction

3.1.1. Variable selection

The primary purpose of the study is to analyse the main drivers of demand and determine which are most significant. To this end, the model will use the total number of rentals per station as a target, intended here as the most relevant proxy for demand, consistent with what has also been presented in the literature [9]. The only exception will be the model created on users' reviews (see chapter 3.5), which instead uses the value of the reviews themselves as the target.

In accordance with what was presented in the introduction, explanatory variables must all be related to objective factors characterising the urban context surrounding the station. Examples of urban context variables might include those related to integration with public transit, the presence of Points of Interest (POIs) or station physical features, such as the presence of information panels on how to use the service or the type of separation from the roadway.

Re-organising insights from literature and proposals from the author, variables which will be considered are presented in Table 7, indicating also the relative source and the reference to articles from literature which suggest using that variable (the letter P means that the variable has been considered after an author's assumption, since it was not explicitly considered in other articles). Let's note that quantitative variables – i.e., ones that indicate the number of something in the surroundings of the station – are referred to a buffer around the points where the station is located (Buffers will be defined later; see Section 4.1 for more details).

Table 7. Study explanatory variables. P = personal assumption.

Variable	Unit of measure	Data source	Literature reference
Presence of a bus stop	(0,1)	Google Maps data	[3] [10] [9]
Presence of train station	(0,1)	Google Maps data	[3]
Presence of metro station	(0,1)	Google Maps data	[3]
Number of overall bus lines	-	Google Maps data	P
Number of overall accessible public transit lines	-	Google Maps data	P
Distance from the nearest bike lane	m	Google Earth, Openstreetmap	[9] [46] [50]
Distance from the nearest car park	m	Google Earth, Google Maps	[10]
Presence of Low Traffic Zone	(0,1)	GIS, public administrations websites	P
Distance from the nearest bike-sharing station	m	Google Earth, GIS	[9]
Rentals in the nearest BS station	-	GIS, data from BS companies	P
Distance from the city centre	m	Google Earth	[9]
University buildings density	/km²	Google Maps data	[3] [10] [17] [9]
Schools density	/km²	Google Maps data	[10] [17]
Restaurants/bar/café density	/km²	Google Maps data	[9]
Hotels density	/km²	Google Maps data	[9]
Parks and green areas density	/km²	Google Maps data	P
Boutiques/shops/malls density	/km²	Google Maps data	[10] [17] [9]
Theatres and cinemas density	/km²	Google Maps data	[3] [10] [17]
Museums density	/km²	Google Maps data	[3] [17]
Touristic Point of Interest density	/km²	Google Maps data	P
Sport centres density	/km²	Google Maps data	P
Healthcare facilities (hospitals, dentists, doctors, pharmacies) density	/km²	Google Maps data	P
Other working activities (banks, post offices, insurance, hairdressers, wellness centres, police stations) density	/km²	Google Maps data	P
Population density	Peop./km²	ISTAT report	[9]
Housing density[1]	Apart./km²	ISTAT report	[9]
Presence of information panels at the station	(0,1)	Google Street View	[17]
Type of separation from the roadway	(0,1,2)[2]	Google Street View	[51]
Elevation of the station	m	Google Earth	[9] [46]
Presence of coverage from weather phenomena	(0,1)	Google Street View	[17]
Presence of station's adequate lighting	(0,1)	Google Street View	[3] [51]

[1] It indicates the number of apartments per square km. Population and housing density have been retrieved from ISTAT report on census areas [52].

[2] 0 = station alongside the road with no physical separation, 1 = station alongside the road, 2 = station far and separated by the road. See Chapter 4.2.3 for more details.

For ease of reading and interpretation, the variables in Table 7 have been grouped into different categories, as shown in Figure 5.

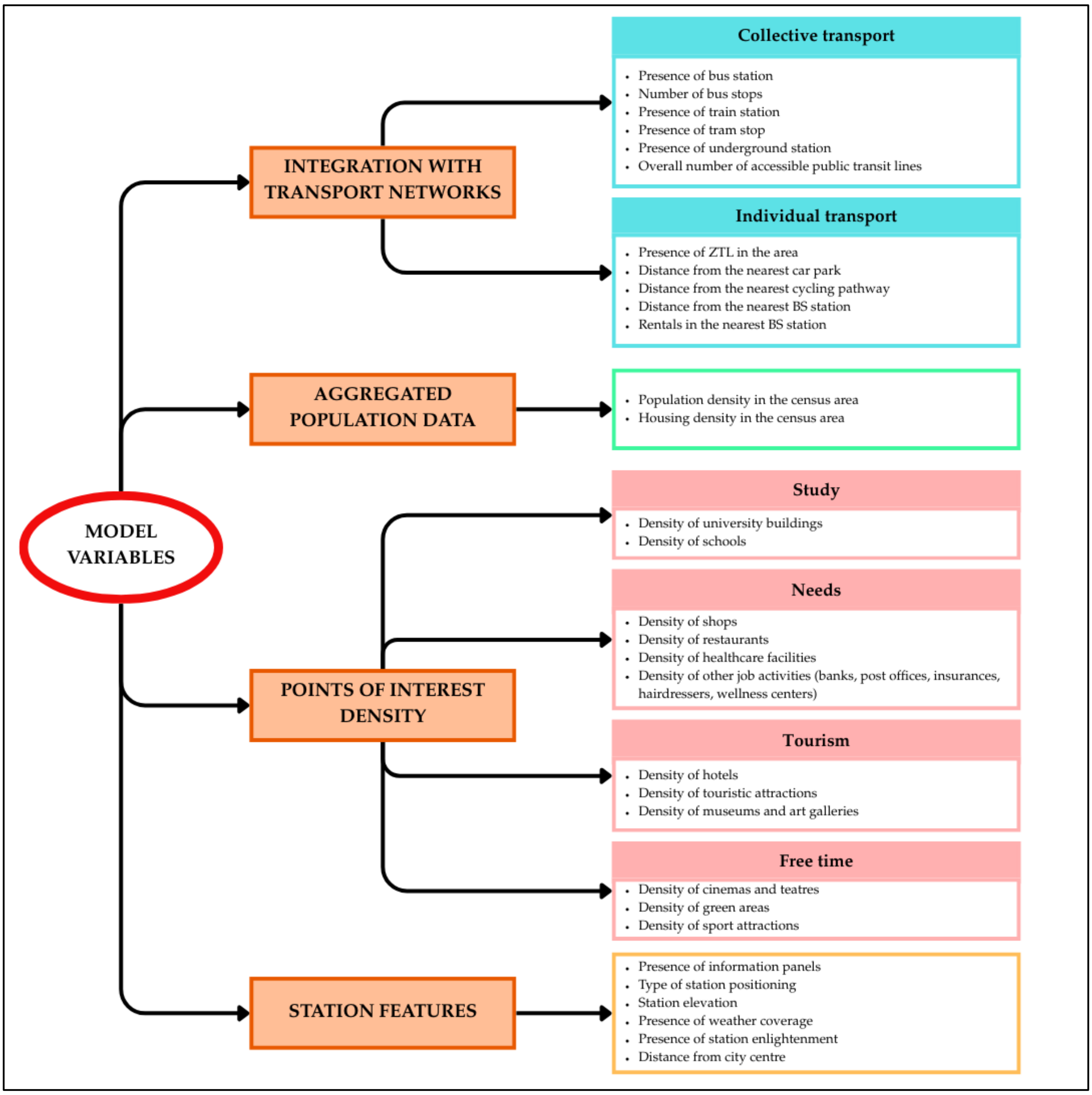


Figure 5. Study explanatory variables.

3.1.2. Data collection

The data collection phase consists of computing the value of each variable for each of the stations in our dataset. This is the most time-intensive part of this study, even if it strongly depends on the number of stations we want to analyse. Data collection is composed of three subsequent phases:

1. Definition of study stations (scenario identification);
2. Computation of buffer independent variables;
3. Computation of the buffer dependent variables.

Scenario identification is the first phase; it's very tricky, since it has a great impact on the final results. It is of critical importance to choose a set of stations that is as heterogeneous as possible. Moreover, it is necessary to verify that data regarding rentals are easily accessible, since they constitute the target of the models developed later. In our study, it was decided to analyse a dataset starring several Italian university cities with fewer than 600.000 inhabitants, located in different geographical contexts to boost heterogeneity in the dataset. The details will be presented in the dedicated chapter.

Once stations are identified, variables about their physical characteristics (from Google Street View) and distance variables (using Google Earth) can be computed.

In the last phase, dedicated to the computation of quantitative variables, it is necessary to identify a buffer which will be helpful to define a study area for each station. Following a brief literature analysis on this theme, it was decided to adopt a "dynamic" buffer, which depends on station distance from the city centre. This method allows for avoiding interferences due to the fact that, far from the city, centre stations are less dense, and people are willing to walk more to reach their own destination of the station itself. Thus, chosen buffers are the following:

- Stations within 1 km from the city centre, 250 m radius buffer (0.200 km^2 area);
- Stations between 1 km and 2 km from the city centre, 350 m radius buffer (0.385 km^2 area);
- Stations more than 2 km from the city centre, 450 m radius buffer (0.686 km^2 area).

At this point, using the Google Maps database, variables are computed. They are considered not in absolute terms but as density (over the reference area) to overcome the differences in size between the buffers. Furthermore, in the case of highly relevant POIs (e.g. supermarkets or large hospitals), a multiplier factor is added to enhance their presence.

An exception to the buffer method is the calculation of population density and housing density. They are calculated at the census area level, since they are presented in this way in the ISTAT report [52], the source of these data. If the station is located on the border between two census areas or in an area with null values, population and housing densities will be calculated as the average of the surrounding census areas.

3.2. Data exploratory analysis

Before getting into data modelling, it is fundamental to visualise and explore collected data. This will help to have a first critical look at phenomena we aim to analyse, highlighting statistics, trends and anomalies. Except for population data, exploration will be conducted on each variable category. It will investigate the following aspects:

- Descriptive statistics: mean, median, spread, percentiles;
- Distribution of unique values (for binary and categorical features);
- Density distribution for numerical variables;
- Outlier analysis;
- Correlation analysis through *Pearson Coefficient*[3];
- Multicollinearity analysis through the *Variance Inflation Factor* (VIF).

Because of data exploration outcomes, planners could decide to modify the set of variables. Examples could be:

- If a binary variable results in more than a predefined percentage of null values, it is suggested not to consider it in the model.
- If a group of variables results in being highly correlated to each other – i.e. the correlation matrix highlights a high values region – it is possible to aggregate them. After aggregation, it will be useful to make a new correlation analysis with aggregated variables.
- If there is evidence of one or more pairs of variables which are highly correlated with each other (Pearson coefficient, $r > 0.9$), it is necessary to choose only one of the two, opting for the one that is most significant for the model to be created.

[3] Pearson Correlation Coefficient, indicated in this thesis with the letter r, measures linear correlation between two sets of data X and Y. It is the ratio between the covariance of the two variables and the product of their standard deviations, as follows: $r_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

3.3. Data-driven approach for the identification of main demand drivers

This step aims to implement models that will help analyse data to understand the main relevant features influencing bike-sharing demand. Thus, the expected output is a list of the most significant variables that explain the variability of the target, according to different evaluation methods. After the necessary data preparation, two types of algorithms will be used in model implementation:

- Linear regression, both simple and penalised (Ridge and Elastic net);
- Ensemble methods (Random Forest and Gradient Boosting).

They will all be implemented using Python programming language.

3.3.1. Data preparation

Collected data must be properly prepared for use in the machine learning models, since raw data could introduce bias into the model, altering outcomes. The following paragraphs outline the three preparation steps adopted in this study.

3.3.1.1. Outliers management

Based on the previous outlier analysis, it is necessary to define an appropriate strategy for their treatment. Values which deviate substantially from the rest of the distribution can significantly distort model outcomes, especially when using simple algorithms such as Linear Regression, which is particularly sensitive to the presence of outliers. In this study, all data are real and manually collected; therefore, no outliers are expected to originate from sensor malfunctions or measurement errors. For this reason, one could argue for retaining all values in the analysis. Nevertheless, some stations exhibit observations that are markedly distant from the overall distribution. To avoid undue influence on model estimation, a partial winsorization approach has been adopted:

- For all variables except population and housing density, outliers are identified only in the right tail of the distribution. Values above the 90th percentile are capped at the 90th percentile, but only if such extreme observations constitute more than 5% of the dataset;

- For population and housing density, outliers are considered in both tails of the distribution, as the mode of these variables is far from zero. In addition to capping right-tail values above the 90th percentile, values below the 5th percentile are capped at the 5th percentile, only if they exceed 5% of the total observations.

3.3.1.2. Data normalisation

Normalisation is needed when data belong to different scales or have different origins. The chosen dataset collects stations from different cities; therefore, local normalisation could be necessary. Two variables, distance from the city centre and station's elevation, must be corrected, since they are much influenced by the characteristics of the single service and by the orography of the single city. Thus, they're normalised as follows, using min-max normalisation:

$$z_i = \frac{x_i - min_{city}(x)}{max_{city}(x) - min_{city}(x)} \qquad 3.1$$

In this way, each city will have a station which will have a value equal to zero and one equal to 1 for each of the normalised variables. Lastly, please note that, if data refer to a unique city, normalisation is not required but still recommended.

Furthermore, some regression models, such as simple and penalised linear regression, require data normalisation to perform better. Thus, before their implementation, variables with data not already comprised between 0 and 1 will be normalised. In this case, normalisation is not local but global, since it is not limited to a single city.

3.3.1.3. Target preparation

The target variable is the overall number of rentals in each BS station, since it has been identified as the best proxy of bike-sharing demand. It is calculated by summing the number of pick-ups and drop-offs at the station, thus counting the frequency with which a station is involved in a trip. However, the dataset includes stations from several cities; therefore, rentals must be adjusted to ensure comparability of the target variable across different urban contexts. A simple min–max normalisation is not the best solution in this case, as it might lead to loss of information. For this reason, correction of the target variable must explicitly account for the scale of each bike-sharing system, expressed by both the total number of stations and the overall number of rentals recorded in each city. Thus, rentals are rescaled city by city by incorporating these structural characteristics, as shown in the following equation 3.2.

$$Target_i = \frac{rentals_i}{\sum_i^k rentals_i} \cdot k_{city} \quad 3.2$$

Where:

- i: considered station;
- $rentals_i$: number of overall rentals in the station i;
- k: number of overall stations in the dataset from the considered city.

Once these operations have been performed, both the explanatory variables and the target are ready. Therefore, we can focus on the data analysis models.

3.3.2. Linear regression

Linear regression algorithms study the linear relationship between a target variable and one or more explanatory variables, optimised using Ordinary Least Squares and fitting the following equation:

$$Target = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \quad 3.3$$

Trying to improve performance, *K-fold cross validation* with $K = 10$ is used. Then, many performance indicators are computed:

- R^2(*coefficient of determination*): measures the proportion of variance in the dependent variable that is explained by the model, providing an overall indication of goodness of fit [53];
- R^2_{adj}: corrects the R^2 value by accounting for the number of explanatory variables included in the model, allowing for a fairer comparison between models with different levels of complexity [53];
- $RMSE$: quantifies the average magnitude of prediction errors by measuring the square root of the mean squared difference between observed and predicted values, with greater weight given to larger errors [53];
- $SMAPE$: expresses the average relative prediction error as a percentage, symmetrically penalising over- and under-estimation and allowing for comparability across different scales [53];
- *Test t-student* on coefficients: is used to assess the statistical significance of individual model coefficients by testing whether their estimated values are significantly different from zero under the assumption of normally distributed errors [53].
- Visual representation of residual values and the relationship between observed and predicted values.

Following an initial attempt with simple linear regression, Ridge and Elastic Net regression will be performed. These algorithms introduce a penalisation term to regression coefficients, which is helpful to reduce overfitting. Even if we are using algorithms to analyse data and not to create a predictive model on the target – so overfitting is not a real issue – it is better to prevent it to perform a more accurate analysis on demand drivers. Penalisation terms are of two typologies, with different functions:

- Penalisation *L1*, applied to absolute values of coefficients, which selects the most significant variables, setting to 0 the least significant ones. A minimum of ten explanatory variables will be applied.
- Penalisation *L2 or alpha*, applied to square values of coefficients, which aims at reducing multicollinearity between explanatory variables.

Ridge regression uses only the L2 penalisation, while Elastic Net includes both penalisation terms. Optimal values of these two penalisation terms will be computed using an automated function in Python (Appendix A). Then, the same evaluation metric of simple linear regression will be computed and compared to the ones obtained before.

Both for simple and Elastic Net regression, the most significant features are identified based on the value of coefficients: the higher the regression coefficient, the more significant that variable is. Moreover, if the computed p-value is higher than 0.05, it means that the coefficient is not statistically significant (5% has been chosen as the predefined significance level). Thus, there is insufficient evidence to conclude that the corresponding predictor has a meaningful linear relationship with the target variable**.** In this case, the coefficient computed with this method is not valid, and it is suggested to use another machine learning algorithm.

3.3.3. Ensemble methods

Ensemble methods are the natural development of simple regression or classification algorithms. They make predictions combining the effects of more than one individual model, aiming at creating a more robust and accurate model, less inclined to overfitting compared to one.

In the presented case study, the basis of ensemble methods is a single regression tree, a type of decision tree used to predict continuous numerical values. It works by recursively partitioning the dataset into smaller, more homogeneous subgroups based on features. To increase the performance of regression trees, two ensemble algorithms have been selected: Random Forest Regressor and Gradient Boosting Regressor. They combine the effects of a set of decision trees with different approaches, briefly explained in the next paragraphs. Nevertheless, both models share the same underlying principle for computing Feature Importance (equation 3.4). The most influential variables—representing the key output of this analysis step—are those that yield the greatest reduction in *Mean Squared Error* (MSE) at the tree nodes where they are used for splitting. In other words, the more a variable decreases prediction error when creating a split, the more important it is considered. When multiple trees are involved, feature importance is obtained by averaging $FI_j^{(t)}$, the MSE reductions associated with each variable across all nodes and all trees in the ensemble (previously computed through equation 3.4).

$$FI_j^{(t)} = \frac{\sum_{i:node\ i\ splits\ on\ feature\ j} RI_i^{(t)}}{\sum_{i \in all\ nodes} RI_i^{(t)}} \tag{3.4}$$

Where RI indicates impurity reduction for the node i in tree t, computed as ΔMSE across the node. The ten variables characterised by the highest Feature Importance will be selected as the most significant and will constitute the foundation of the SSEI formulation and calibration.

Then, the ensemble models will be evaluated using the same metrics used for linear models, except for the t-student test, which is no longer applicable since there is no numerical coefficient for each variable considered.

3.3.3.1. Random forest

Random forest is an ensemble method used both for classification and regression problems. It mitigates the correlation between decision trees by employing a random selection of samples and features. In fact, predictions are made by training in parallel a predefined number of trees. Each tree is built with just a sample of the total amount of explanatory variables. The final prediction is given by the average value among trees.

Parameters used in creating the Random Forest model are the following

- Number of trees in parallel (`n_estimators`): 300;
- Maximum number of considered features at each split (`max_features`): "sqrt", which means that it is equal to the square root of the overall number of features;
- No limitations on the depth of the tree and minimum number of samples for creating a new split or leaf.

3.3.3.2. Gradient boosting

Gradient boosting is another ensemble algorithm that combines the results of regression trees to improve the accuracy and robustness of the model. However, the combination method is different compared to Random Forest. In fact, Gradient Boosting uses a technique called boosting, which iteratively trains trees one after another to correct previous mistakes. As a result, the last tree will be the most accurate.

Parameters used in the Gradient Boosting implementation will be the following:

- Number of trees created iteratively (`n_estimators`): 200;
- Learning rate, which indicates how fast the model "learn" from each tree. In other words, it is a proxy of the contribution of each single tree (`learning_rate`): 0.05;
- `max_features`: "sqrt";
- No limitation in the maximum depth of trees and minimum samples for each slit or leaf.

3.4. Definition of the Station Static Effectiveness Index (SSEI)

The creation of a new quantitative index for evaluating station effectiveness in urban contexts is one of the most promising expansions of the previous data analysis. The index will help planners both during the design of the service and during the evaluation and management phase, providing them with a quantitative indication of how effectively a station is integrated into the city context. The index will be called SSEI: *Station Static Effectiveness Index*. It refers to a single station, and it does not have a dedicated unit of measure. The word "static" has been chosen to indicate that only static variables are considered, neglecting others, such as the availability of docks and bicycles, intended as "dynamic". For other details and SSEI fields of application, see Chapter 1.4.

The next sub-chapters will explain in detail, step by step, the methodology adopted in SSEI creation, starting from insights obtained from the foregoing data analysis and concluding indicating possible applications of the new index.

3.4.1. Variable weights definition

The output of the previous machine learning data analysis is the pinpointing of the ten most significant variables influencing bike-sharing demands. Therefore, it is necessary to understand how quantitatively they influence such demand. In other words, we need to find a "weight" for each of the variables formerly selected. To do this, it was decided to adopt a hybrid approach, exploiting the strengths of both statistical and survey-based methods.

In detail, weights for each variable are computed including different contributions:

- 50% statistical methods, of which:
    - 25% for the Pearson correlation coefficient between the considered variable and the number of rentals;
    - 25% for the quote of variance explained by the considered variable j, in the linear combination, obtained by applying *Principal Component Analysis* (PCA) on the selected variables.
- 50% survey-based method. A survey will be conducted among people, both bike-sharing users and non-users. It will ask them to give a weight (from 1 to 5) to each of the selected variables. The 20% truncated mean of grades will constitute the variable's weight suggested by the survey.

Then, each term has been normalised on the variable subset using a softmax normalisation[4]. Thus, definitive weights for the variable j are computed with the following formula:

$$W_j = 10\left[0.25 \cdot Pearson_j + 0.25 \cdot |w_{PCA}| + 0.5 \cdot w_{survey}\right] \quad 3.5$$

Multiplying by 10 is suggested for ease of interpretation, and the sign of the result is given only by the correlation term, since it is considered just the absolute value of the PCA term. Lastly, since the variables are characterised by different scales and units of measure, it will be necessary to normalise the calculated weight. To do this, we use the standard deviation of the variables calculated from the initial dataset, as shown in equation 3.6.

$$W_{dj} = \frac{W_j}{\sigma_j} \quad 3.6$$

3.4.2. SSEI equation

To create an index which can be easily used in different contexts, allowing users to make comparisons between stations, it is fundamental that it is computed not with

[4] Softmax normalisation is a type of normalisation that tends to reduce the influence of extreme data (or outliers) without explicitly eliminating them. The normalisation equation is the following: $\sigma(z) = \frac{e^{\beta z_i}}{\sum_{j=1,\dots,K}^{K} e^{\beta z_j}}$ in which β is chosen as equal to $\frac{1}{0.25}$ and $K = 10$, in this case.

reference to a predefined scale. Furthermore, it is relevant that the index is as easy to interpret as possible, both technically and graphically. Following the research carried out, it was decided to calibrate the SSEI with reference to the logistic function, characterised by the following equation and shape (Figure 6).

$$f(x) = \frac{L}{1 + e^{-k(X-c)}} \tag{3.7}$$

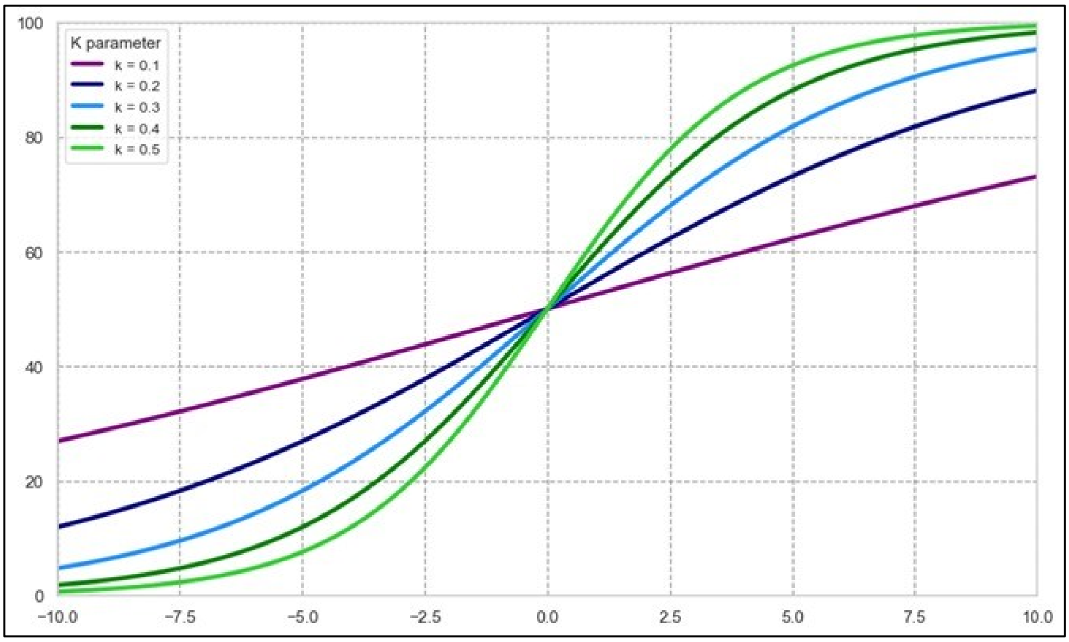


Figure 6. Logistic function for different k values.

This choice was due to three key reasons:

- The logistic function allows for setting the upper limit through the term L, while the lower limit is always equal to zero. In the case of SSEI, an upper limit equal to 100 is established. Moreover, it is possible to set also the central value of the curve, so the inflexion point, indicated by the term c, which will be calibrated using training data.
- The parameter k, which regulates the slope of the curve, helps to manage the distribution of outputs. The steeper the curve, the less the distribution of output values. Conversely, as the slope decreases, the values will be more widely distributed. By calibrating the value of k on the training data, it will be possible to find the value that most closely approximates the ideal configuration (see next chapter).
- The logistic function allows the calculation of the index even if not all the features' data are available. Thus, in cases where it is not possible to collect

certain data, the index can still be calculated, and it will still have a value between 0 and L. However, this value will certainly not be as accurate as that calculated using all the variables. This characteristic of the logistic function also helps in cases where, by choice, it is decided not to consider certain types of variables. For example, if we want to analyse stations that are all far from the city centre, we may choose not to consider the distance from the city centre in the calculation of the SSEI.

With reference to a single station, the SSEI equation appears as follows:

$$SSEI_i = \frac{100}{1 + e^{-k(\sum_j^V x_{ij} W_{dj} - c)}} \tag{3.8}$$

In which:

- i denotes the considered station;
- V is the number of main relevant variables obtained after data analysis;
- x_{ij} is the value of the variable j for the station i;
- W_{dj} is the weight computed through equation 3.5 for the variable j.

If one or more of the variables are binary, it is recommended not to include them in the exponential term but to consider them separately using specific dummy β variables and relative multiplicative coefficients (equation 3.9). It will be necessary to modify the numerator term appropriately so as not to exceed the upper limit of 100.

$$SSEI_i = \frac{(100 - q)}{1 + e^{-k(\sum_j^V x_{ij} W_{dj} - c)}} + q \cdot \beta_i \tag{3.9}$$

Where:

- q is the weight given to the considered binary variable;
- β_i is the dummy variable which indicates the value of the binary variable ($1 = True, 0 = False$).

3.4.3. SSEI calibration

After choosing the most suitable equation for SSEI, calibration is necessary for its two main parameters: growth rate k and c. c, the x-coordinate of the inflexion point, is computed as the average of $x_{ij} W_{dj}$ over stations from the original dataset, as reported in equation 3.10.

$$c = mean_i \left[\sum_{j=1}^{V} x_{ij} W_{dj} \right] \tag{3.10}$$

Where V is the overall number of features in equation 3.9 and i denotes the single station.

Calibration of k is conducted by exploiting an iterative automatic function developed in Python. In detail, the function works following the steps reported below:

1. From the initial dataset, extract random samples with a dimensionality z equal to multiples of 5;
2. Assuming c as computed before and different possible values of k between 0.1 and 2 (with a span of 0.1), compute SSEI for each combination of z and k, based on data from the dataset;
3. For each value of k, assign each station to one of these categories based on its SSEI (Figure 7);
   i. $80 \leq SSEI \leq 100$: $excellent$;
   ii. $60 \leq SSEI < 80$: $good$;
   iii. $40 \leq SSEI < 60$: $medium$;
   iv. $20 \leq SSEI < 40$: low;
   v. $0 \leq SSEI < 20$: $critical$;
4. For each combination of z and k, compute Δ as the difference between the number of stations in the most represented class and the number of stations in the least represented class (equation 3.11 or 3.12);

$$\Delta = n_{stations_{most\ represented\ class}} - n_{stations_{least\ represented\ class}} \tag{3.11}$$

$$\Delta\% = \frac{n_{stations_{most\ represented\ class}} - n_{stations_{least\ represented\ class}}}{n_{stations_{most\ represented\ class}}} \tag{3.12}$$

5. For each sample, select the value of k which minimises Δ or Δ%;
6. Count the frequency of the selected value of k. The optimum value of k is the most frequent among the selected ones. It will be used in the SSEI equation.

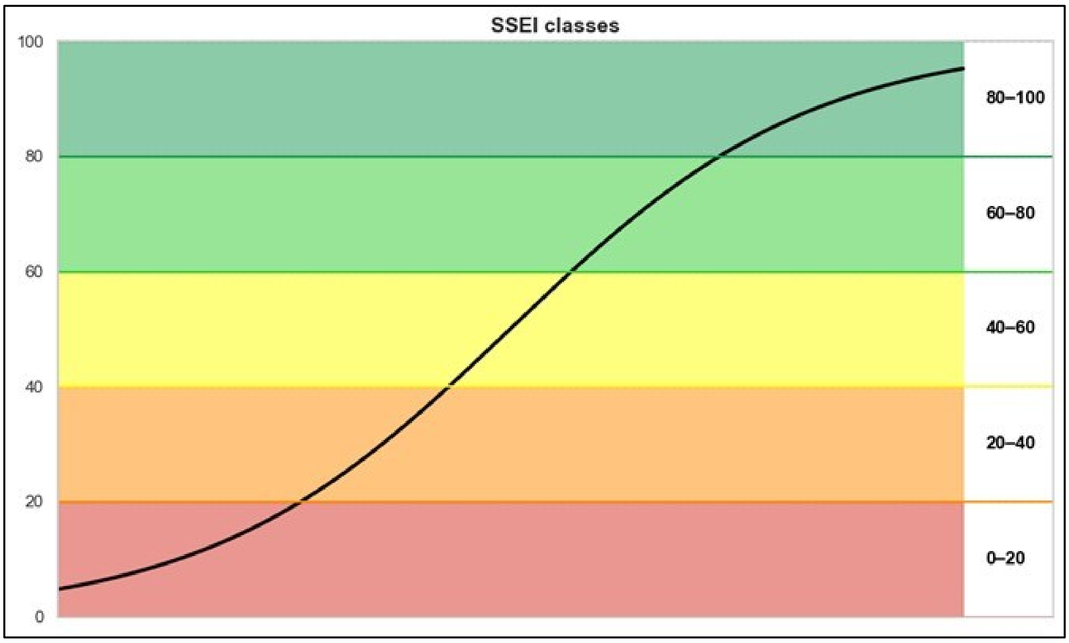


Figure 7. Logistic function with SSEI classes.

Lastly, it should be noted that the calibration of these two parameters is strongly influenced by the characteristics of the initial dataset. Therefore, even though the chosen methodology aims to make the SSEI index more universal and interpretable, it may not be suitable for explaining the effectiveness of the urban integration of bike-sharing stations in contexts much different from the one analysed in Chapter 4.1.

Once the equation of the SSEI is defined, it will be tested by computing it for each station in the original dataset (Chapter 5.1). This will allow us to understand if the index specifications are adequate for our case study. If the response is positive, the calculated values are expected to fit the set logistic curve perfectly (except for any binary variables), ensuring that no class stands out from the others in terms of number, with SSEI values evenly distributed among the classes. Then, the applications of the SSEI will be studied in depth in Chapter 5, presenting a simulation of the performance evaluation of a service (Chapter 5.1) and an analysis of synergies with the cycling infrastructure (Chapter 5.2).

3.5. Local demand analysis framework

One of the advantages offered by the initial dataset is the possibility to carry out further analyses to understand the characteristics of local demand. The term 'local demand' refers to demand in the cities that are part of the original dataset.

The first necessary step in local analysis is the decomposition of the initial dataset. From it, K smaller dataset will be created, where K is the number of cities present in it. For each of the new datasets, after a brief data exploration, a data analysis will be conducted based on the framework presented in Chapter 3.2. The goal is always the same: identifying the variables that most influence demand in the individual cities analysed, using machine learning methodologies. Considering the reduced dimensionality of the new datasets, linear regression will not be taken into account in this case. Thus, demand drivers will be investigated by adopting only ensemble methods, such as random forest and gradient boosting. Local data analysis will be facilitated by faster variable preparation. In fact, variable normalisation will no longer be necessary, since data come from the same origin (i.e. the same city).

Outcomes will help planners in the identification of common demand patterns and areas to focus on to improve the quality of service. It will be intriguing to compare the results obtained with those of other cities to detect similarities or differences in users' behaviour. A further development of the analysis could be the creation, through the use of a chosen algorithm, of a new predictive model that could help to obtain an initial estimate of the number of rentals for a certain bike-sharing station in the city in question. This could be very useful during the service planning phase.

3.6. Users' satisfaction modelling framework

Bike-sharing schemes offer users the opportunity to review the service, giving a grade (usually between 1 and 5) for each station. The grades are a synthetic indication of the user's satisfaction level. Thus, service managers can have a measure of the quality of their own service by simply collecting users' reviews. Moreover, bike-sharing stations are often present in Google Maps, so everyone can write a review and give them a grade. These elements offer the opportunity to adapt the previously used Machine Learning model targeting bike-sharing station grade, aiming at studying which are the main influencing factors on users' satisfaction. However, before proceeding with the assessment, it should be noted that this analysis differs significantly from the others presented in this document, as it is the only one that also considers the operational characteristics of the service, even if implicitly. In fact, although the explanatory variables are still the same, the target variable necessarily also depends on operational characteristics such as the availability of bicycles at each station, which undoubtedly influences user satisfaction.

The Machine Learning model will use as target a new variable `Grade`, combining the two grade families previously presented:

$$Grade = 0.7 \cdot G_{serv} + 0.3 \cdot G_{Google} \qquad 3.13$$

In which:

- G_{serv} is the station's grade – from 1 to 5 – released by service users on the service application/website;
- G_{Google} is the stations's grade – from 1 to 5 – reported on Google maps.

It has been decided to give much more importance to reviews on the service website, as they can only be posted by people who have actually used the bike-sharing service. Google, on the other hand, allows anyone to leave a rating on the stations.

Due to various reasons, such as low usage, reviews are not always available for every station considered or are available in very limited amounts. Consequently, not all the stations present in the original dataset will be considered in the new analysis. Moreover, in some cities, the review system might not be working. To overcome these issues, the following measures have been taken, in line with the proposed order:

- For each city in the initial dataset, if more than 40% of the overall stations have no reviews, stations from that city will not be considered in the analysis.
- For each of the two categories, in case the review is not present, that station assumes a grade equal to the average value of the other ones from the same city for the same category.

Once the target variable is defined, an exploration of it and its components will be performed, analysing average values, distribution and confidence intervals, both in overall terms and for single cities. Then, Machine Learning analysis will be conducted, attempting to train linear regression and ensemble models as was done in the main data analysis. The final output will be a rank which includes all the model variables, sorted in order of importance in determining user satisfaction.

4 Models implementation and SSEI calibration

This section is dedicated to applying the proposed methodology to demonstrate its robustness and reliability. The goal is to propose an approach to data analysis of bike-sharing stations, totally independent of the OD demand matrix and service operative features, which will lead to the creation of the SSEI, an innovative index that can help analyse the performance of bike-sharing services in relation to the urban context in which the stations are located. The structure is the same as section 3: from dataset construction to complementary analysis, each sub-chapter will explain in detail what was performed, no longer limiting itself to a theoretical exposition, but including data, numbers and qualitative considerations about outcomes.

4.1. The Italian multi-city dataset

4.1.1. Variable selection

As presented in paragraph 3.1.1, variables must be chosen among the ones reported in Table 7. To ensure completeness and robustness, it was decided to retain all the variables listed in the table, reserving the right to perform additional checks after the initial data exploration phase. Thus, selected dataset explanatory variables are reported in Table 8, alongside the identification name given to each of them.

Table 8. Dataset features.

Variable	ID	Variable	ID
Presence of a bus stop	bus	Presence of train station	train
Presence of metro station	metro	Distance from the nearest bike lane	dist_cic
Number of overall accessible public transit lines	n_transport	Presence of Low Traffic Zone	ZTL
Distance from the nearest car park	dist_carpark	Rentals in the nearest BS station	rent_near_bs
Distance from the nearest bike-sharing station	dist_bs	University buildings density	uni_dens
Distance from the city centre	dist_centre	Restaurants/bar/cafes density	rest_dens
Schools density	school_dens	Parks and green areas density	park_dens
Hotels density	hot_dens	Theatres and cinemas density	cin_tea_dens
Boutiques/shops/malls density	shop_dens	Touristic Point of Interest density	tour_dens
Museums density	mus_dens	Healthcare facilities (hospitals, dentists, doctors, pharmacies) density	health_dens
Sport centres density	sport_dens	Population density	pop_dens
Other working activities (banks, post offices, insurance, hairdressers, wellness centres, police stations) density	other_act_dens	Presence of information panels at the station	info
Housing density	ab_dens	Elevation of the station	elev
Type of separation from the roadway	pos	Presence of adequate lighting nearby station	lit
Presence of coverage from weather phenomena	cover		

4.1.2. Data collection

In the present case, some bike-sharing stations in Italian mid-sized university cities are taken into account. This choice was due to two main reasons: firstly, the ease of retrieving the necessary data. Secondly, the fact that bike-sharing is a transport solution widely popular among students, as stated by many articles in the literature [39]. This increases interest in a case study such as this one.

Table 9 shows a brief presentation of the selected cities, which are: Brescia, Genoa, Forlì, Parma, Pisa and Trieste. They have been selected to obtain a heterogeneous dataset, due to the different orographic and cultural contexts that characterise each city.

Table 9. Cities in the dataset.

City	Population	Area	Elevation (Avg.)
Brescia	200,857	90.34 km^2	149 m
Genoa	565,301	240.29 km^2	20 m
Forlì	116,700	228.2 km^2	34 m
Parma	202,111	260.60 km^2	55 m
Pisa	98,778	185 km^2	4 m
Trieste	198,668	85.11 km^2	2 m

As of 2024, bike-sharing services in these cities are all entirely station-based. They differ greatly in terms of size and management methods. The main characteristics are shown in Table 10, which also highlights the source of the data about rentals per station, then reported in the dataset. Except for Parma, the other services are all provided by local companies that report to *Weelo*, an italian company that manages micro-mobility services nationwide.

Table 10. Bike-sharing services in the dataset.

City	Service name	Provider	Number of stations	Average yearly number of rentals per station	Data source
Brescia	*BiciMia*	Brescia Mobilità	98	17,061	Brescia Mobilità
Genoa	*ZenaByBike*	Genoa Parcheggi	14	473	Weelo
Forlì	*MiMuovoInBici*	FMI	14	1,080	Weelo
Parma	*MiMuovoInBici*	Infomobility	46	2,780	Infomobility
Pisa	*CicloPI*	PisaMo	39	2,200	Weelo
Trieste	*BiTS*	Trieste Trasporti	23	2,810	Weelo

To avoid introducing excessive imbalance into the dataset, it is necessary to find a configuration in which each city is adequately represented while not dominating the others. This results in stations from a single city comprising a range between 10% and 25% of the overall number of stations in the dataset. Figure 8 shows the best dataset partition identified: 140 total stations from Brescia (22%), Genoa (10%), Forlì (10%), Parma (23%), Pisa (19%) and Trieste (16%).

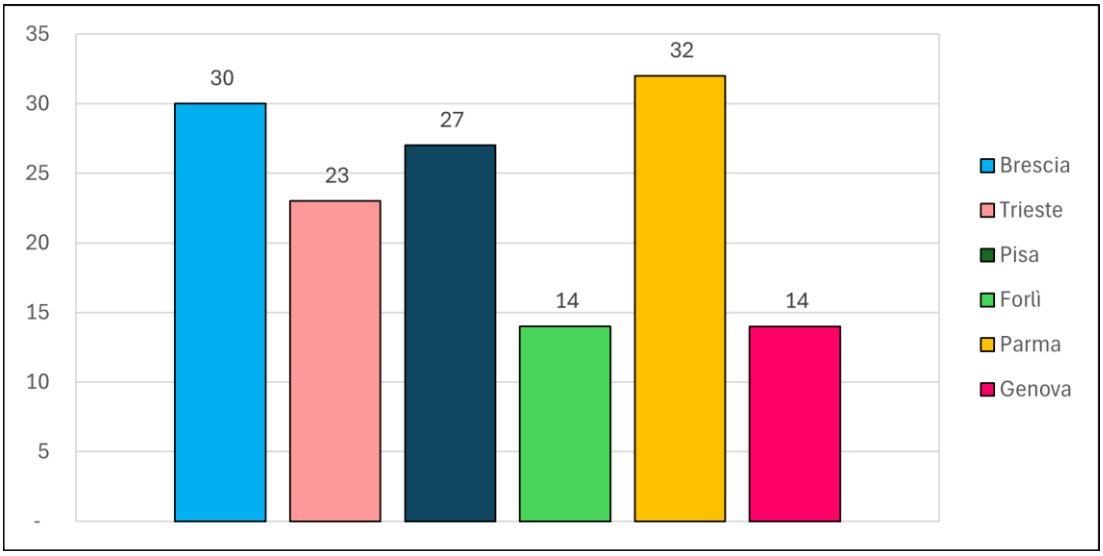
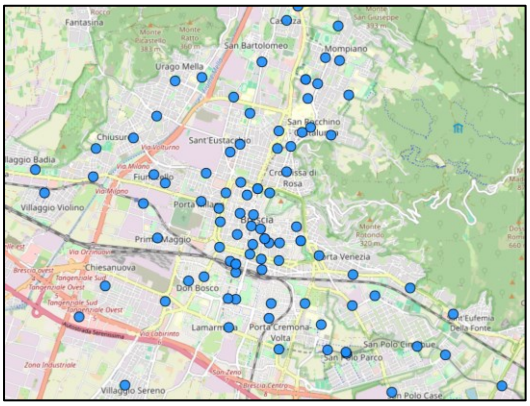


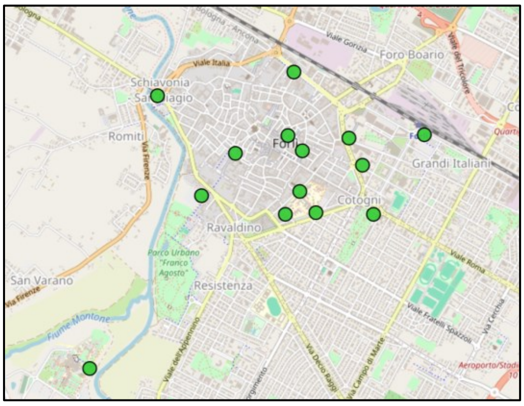
Figure 8. Dataset composition by number of stations.

After that, it is possible to fill the dataset by computing values for each variable and station. To do this, buffers around stations have been defined adopting "dynamic buffer" technique exposed in paragraph 3.1.2. Figure 10 displays the situation in each city, after drawing buffers on Google Earth application, while Table 11 points out some statistics about buffers. Then, the dataset fields have been filled using methods reported in Table 8. All data have been collected or computed manually, and no relevant issues have been identified at this stage of the work. Therefore, all fields in the dataset are complete for each of the 140 stations present.

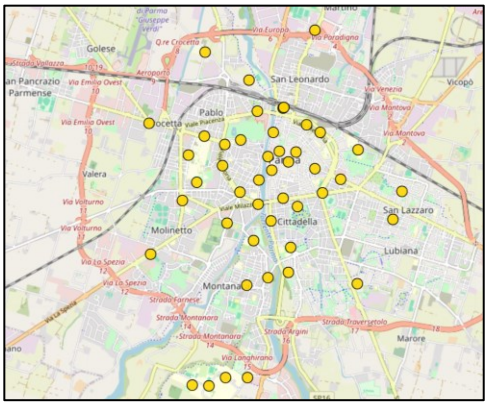


a) Brescia

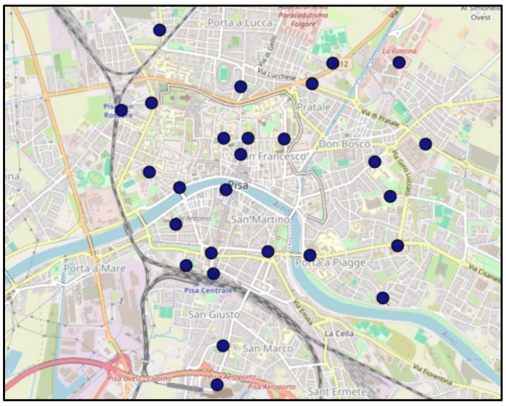
b) Genoa



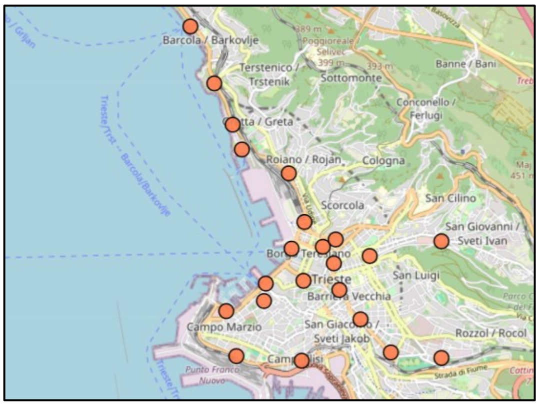
c) Forlì



d) Parma

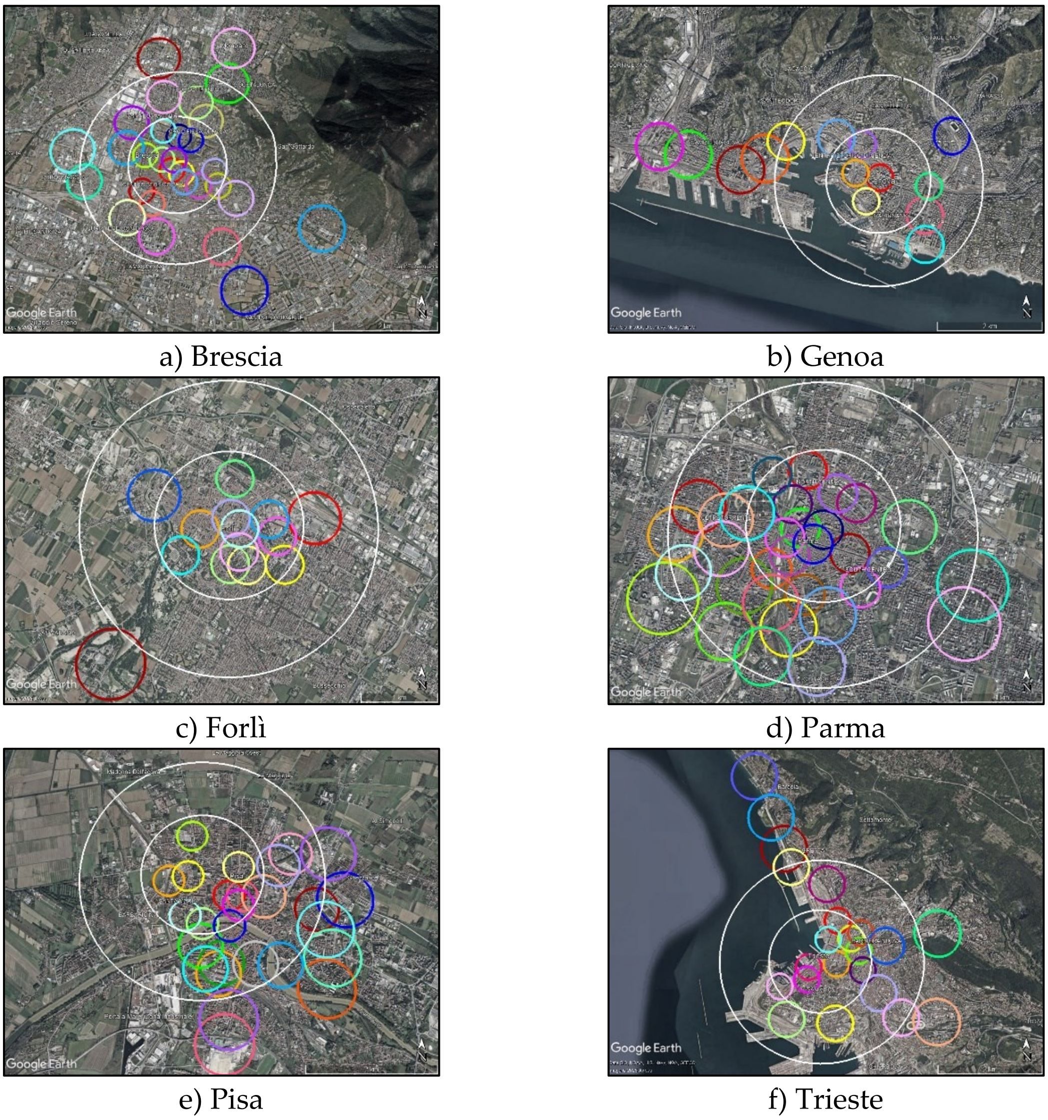


e) Pisa



f) Trieste

Figure 9. Maps of bike-sharing services in the dataset.



a) Brescia b) Genoa c) Forlì d) Parma e) Pisa f) Trieste

Figure 10. Station buffers.

Table 11. Stations' buffers statistics.

City	City centre	Distance	Number of stations	Area coverage [%]
Brescia	*Piazza della Loggia*	$< 1km$	13	81.3 %
		$1km < d < 2km$	11	54.0 %
		$> 2km$	6	-
Forlì	*Piazza Saffi*	$< 1km$	11	68.8 %
		$1km < d < 2km$	2	23.3 %
		$> 2km$	1	-
Genoa	*Piazza De Ferrari*	$< 1km$	5	34.5 %
		$1km < d < 2km$	5	27.7 %
		$> 2km$	4	-
Parma	*Piazza del Duomo*	$< 1km$	14	87.5 %
		$1km < d < 2km$	15	67.8 %
		$> 2km$	3	-
Pisa	*Campo dei Miracoli*	$< 1km$	10	62.5 %
		$1km < d < 2km$	10	46.3 %
		$> 2km$	7	-
Trieste	*Piazza Unità d'Italia*	$< 1km$	10	62.5 %
		$1km < d < 2km$	7	37.1 %
		$> 2km$	6	-

4.2. Exploratory data analysis results

For the sake of clarity, data exploration is now presented separately for the different variable categories reported in Figure 5. An exploratory analysis was carried out for each of these, in line with the methodology proposed in the previous section. The most interesting insights for each category are reported in the next paragraphs. After that, correlation analysis will be presented in a dedicated chapter.

4.2.1. Points of interest

Descriptive statistics for each type of point of interest variable have been computed. They're presented in Table 12.

Table 12. Points of interest: descriptive statistics.

	uni_dens	mus_dens	cin_tea_dens	shop_dens	rest_dens	hot_dens	park_dens	school_dens	health_dens	sport_dens	tour_dens	other_act_dens
Mean	10.7	6.8	2.2	73.2	64.0	30.5	7.4	13.5	59.9	12.6	23.6	61.6
Std	14.9	11.7	3.5	68.3	61.7	40.0	5.6	11.8	44.2	9.0	27.5	50.7
Min	0	0	0	1.6	0	0	0	0	0	0	0	0
Med	3.1	2.6	0	49.4	45.0	13.6	5.2	10.2	52.0	10.4	10.4	50.9
Max	61.1	61.1	20.4	356.5	259.7	188.4	25.5	61.1	213.9	40.7	137.5	275.0

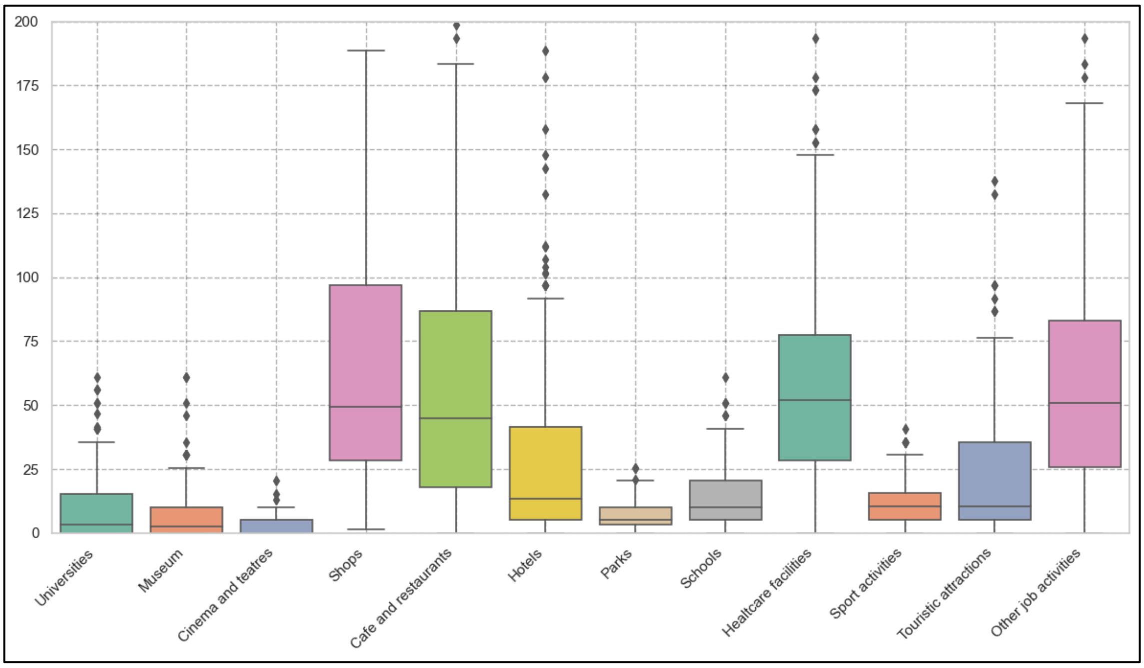


Figure 11. Points of interest boxplots.

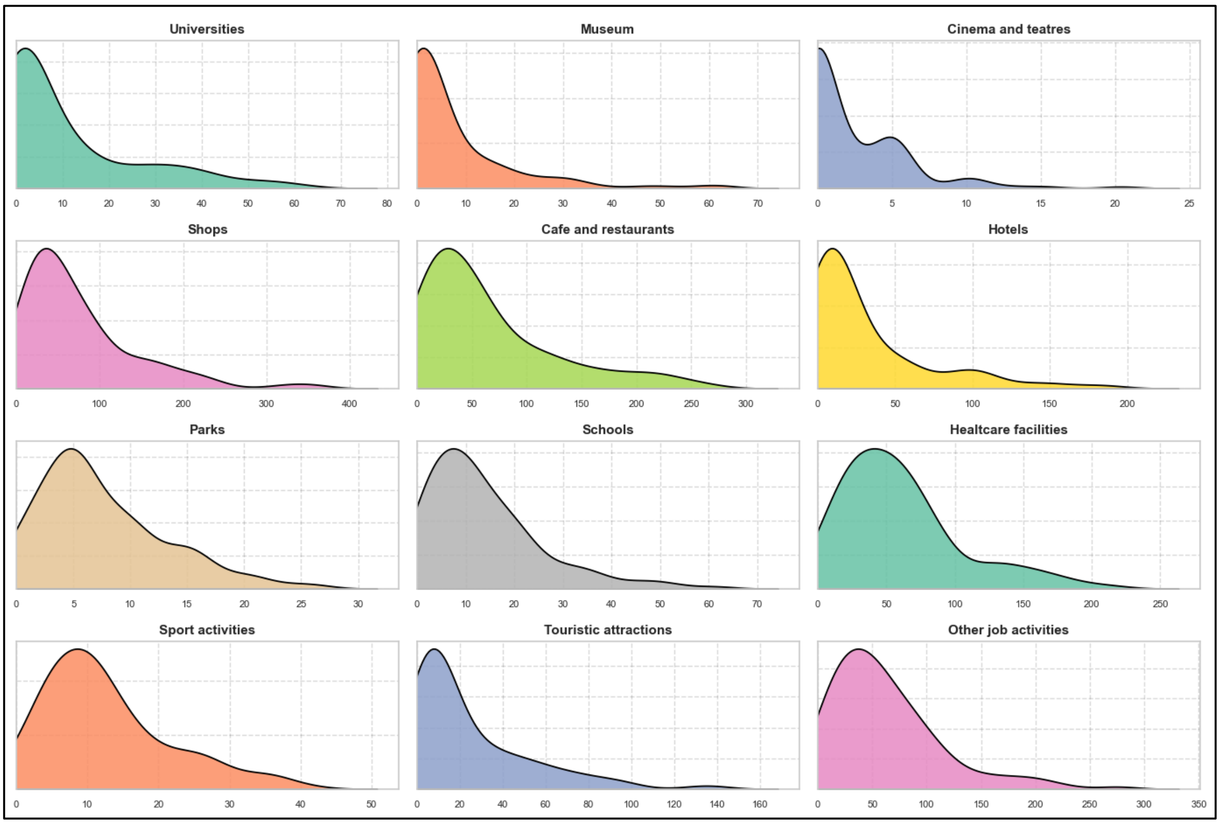
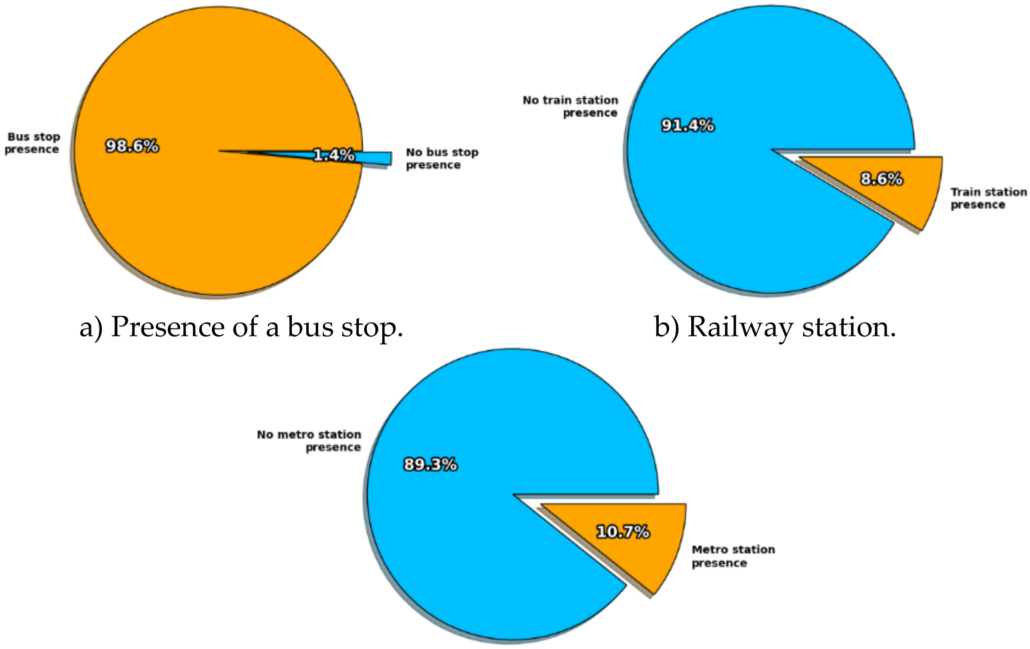


Figure 12. Points of interest Kernel Density Plots.

Boxplots (Figure 11) highlight a quite significant number of outliers for each POI variable. They will be deeply investigated in chapter 4.3.1, in which it will be identified the most suitable tactics to manage them. On the other hand, density plots (Figure 12) point out that variables share a common unimodal distribution with the sole exception of cinemas and theatres density, which appears to be bimodal.

4.2.2. Integration with transport network

Binary variables about the presence of bus stop, train station or metro station have been investigated by analysing the percentage of stations in which these variables are equal to 1 or not (Figure 13).



a) Presence of a bus stop.

b) Railway station.

c) Underground station.

Figure 13. Transportation infrastructure presence pie plots.

Outcomes show that there are just two stations (1.4%) without bus stops in the buffer. For this reason, `bus` variable won't be considered in the next working steps. Instead, there is a quite relevant percentage of stations characterised by the proximity to a train or underground station. The latter comes from the cities of Brescia and Genoa, the only ones in the dataset with an underground service.

As regards the total number of accessible transport lines, the average is 13.6 per station. While the minimum is zero, the maximum number is 82, relating to the bike-sharing station at Parma railway station. In general, bike-sharing stations of this type are those with the most accessible transport lines. This is because railway stations are seen as multimodal centres, where several modes of transport, such as buses, trams, and taxis, converge in addition to rail transport itself.

Low Traffic Zone (ZTL) is present in the buffer of a quarter of all stations in the dataset (Figure 14). In Brescia, there are areas both with 24h ZTL and less than 24h ZTL, while in other cities, like Parma and Genoa, the restrictions are dependent on the considered area.

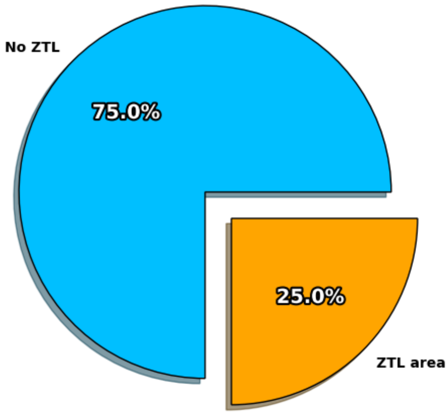


Figure 14. ZTL presence pie plot.

The distance variables (from cycle paths, other bike-sharing stations and car parks) were the last to be analysed. For each of the three, values far from the average are highlighted, and the distribution appears to be unimodal in each case, with minimal dispersion in terms of distance from cycle paths and other bike-sharing stations, while it is more marked in terms of distance from car parks.

Table 13. Distance from individual transport infrastructures descriptive statistics.

	Dist_cic [m]	Dist_bs [m]	Dist_carpark [m]
Mean	108.9	437.0	166.8
Std	162.5	316.2	155.9
Min	5.0	120.0	10.0
Median	30.0	400.0	135.0
Max	900.0	3240.0	700.0

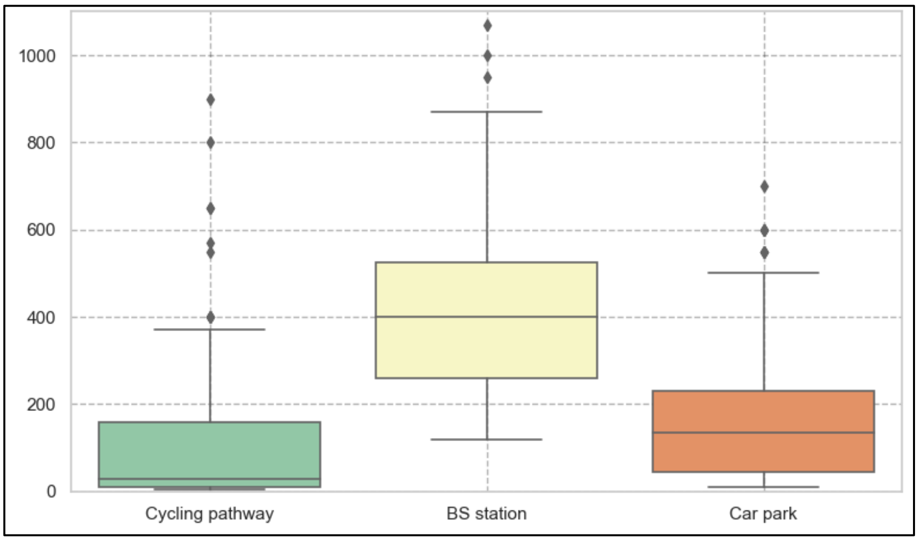


Figure 15. Distance from individual transport infrastructures boxplots.

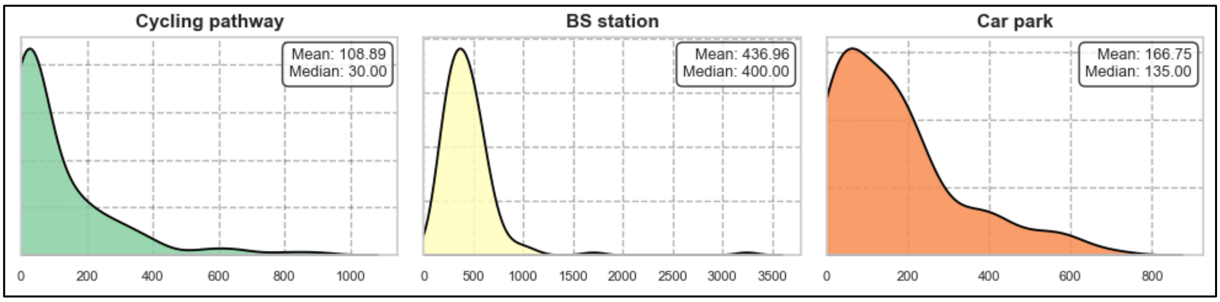
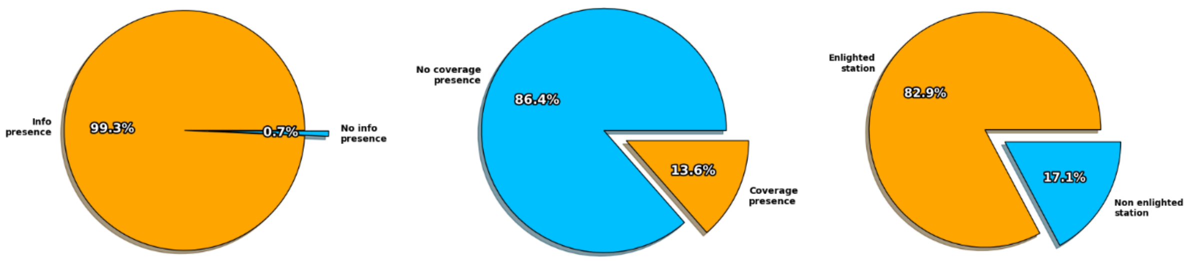


Figure 16. Distance from individual transport infrastructures Kernel Density plots.

4.2.3. Station physical features

Data exploration concludes with a discussion of the station's physical characteristics. Variables from this category are three binary (`info`, `cover`, `lit`), one numerical (`elev`) and one categorical (`pos`). Elevation of stations has low relevance in this phase. Please refer to the chapter about data exploration broken down by city (Chapter 4.5.1) for a better understanding of this feature.

Just one station over 140 (0.7%) results in not being equipped with information panels on how to use the service. In line with what was done previously with the variable `bus`, it is decided not to include feature `info` in the subsequent analysis. Many stations lack of coverage against weather phenomena (Figure 17); this may be due to high construction costs and the risk of vandalism. Most of the covered stations are located at key points in cities, such as railway stations or hospitals. Lastly, adequate lighting is present in the proximity of more than 80% of stations. This may encourage the use of the station during the hours of darkness, as well as speeding up pre- and post-journey operations.



a) info
b) cover
c) lit

Figure 17. Station physical features pie plots.

The feature `pos` deserves special attention; it is not a binary variable, but rather a categorical one. Each category represents a different possible type of separation between the station and the road. Thus, the aim is to represent a proxy of the safety perceived by users during the choice of the station: if the station is adequately separated from the road, it is expected that the user will feel safer. Three categories are:

- *0*: no separation between the station and the road (Figure 19a, color violet in Figure 18),
- *1*: station alongside the road with separation like curbs, shoulders or small barriers (Figure 19b, color fuchsia in Figure 18),
- *2*: station far from the road and separated by it. An example could be a station located in pedestrian areas or in squares where vehicular traffic is prohibited. (Figure 19c, color pink in Figure 18)

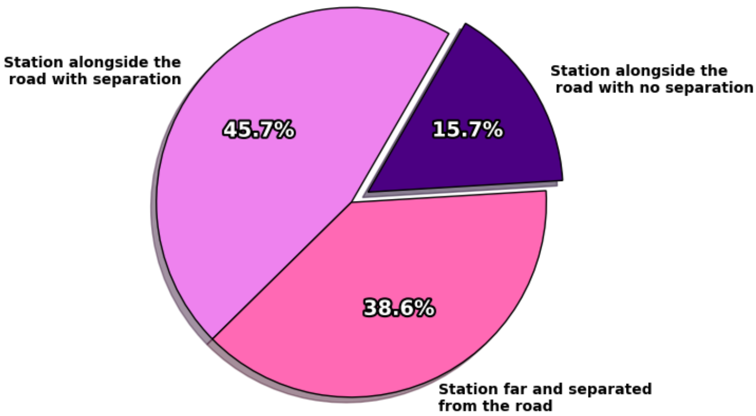


Figure 18. Station position pie plot.

a) 0 b) 1 c) 2

Figure 19. Types of station positioning.

4.2.4. Correlation analysis

Correlation was investigated among all variables selected after data exploration. So, variable `bus` and `info` are not included. Pearson correlation coefficients were computed, creating a correlation matrix reported in Figure 20. Two variables, `dist_centr` and `elev` have been normalised as suggested in paragraph 3.3.1.2. Thus, they have been renamed `dist_centr_norm` and `elev_norm`.

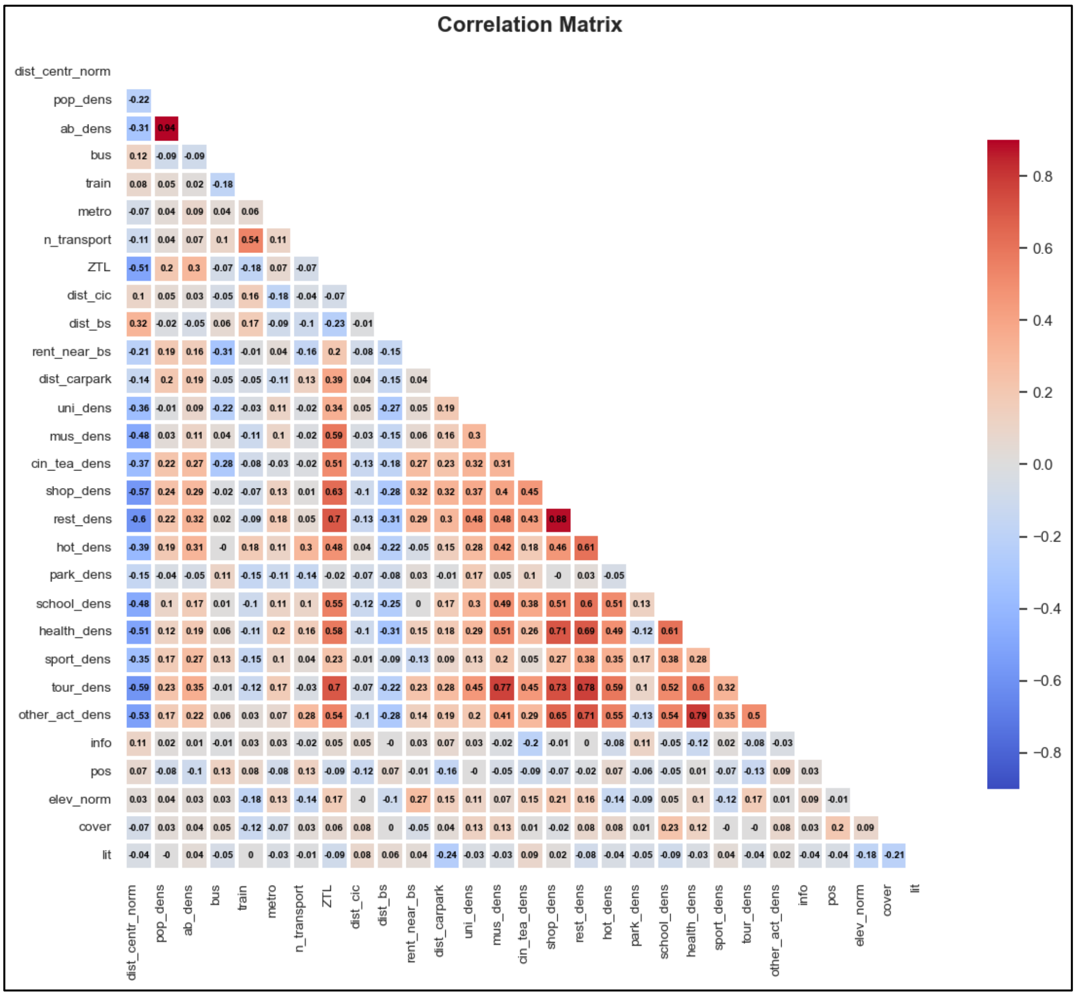


Figure 20. Correlation matrix.

Results suggest important insights:

- As expected, by moving away from the city centre, the density of points of interest decreases (i.e. the correlation is negative). This, in some ways, confirms that it is correct to use a dynamic buffer approach, as performed.
- Population and housing density are highly correlated (r = 0.94) with each other. Thus, it is necessary to delete one of them. To not distort models, it has been decided to remove `ab_dens`, since the focus of the study is on urban context features and not on people.
- Presence of a ZTL is moderately correlated with the presence of points of interest. This is because ZTLs are often present in the city centre, where points of interest are denser.
- It is possible to identify a cluster of red boxes, which is related to the correlation among different types of points of interest. This could introduce heavy

interference in our model. To overcome this issue, these variables will be aggregated in families, as follows:

- `Study_dens`: `uni_dens, school_dens;`
- `Needs_dens`: `shop_dens, rest_dens, health_dens, other_act_dens;`
- `Tourism_dens`: `hot_dens, tour_dens, mus_dens;`
- `Free_time_dens`: `cin_tea_dens, park_dens, sport_dens.`

They will all be computed by summing the number of single points of interest and then dividing by the buffer area.

Following these adjustments, the new correlation matrix is shown in Figure 21, where red boxes have been slightly reduced.

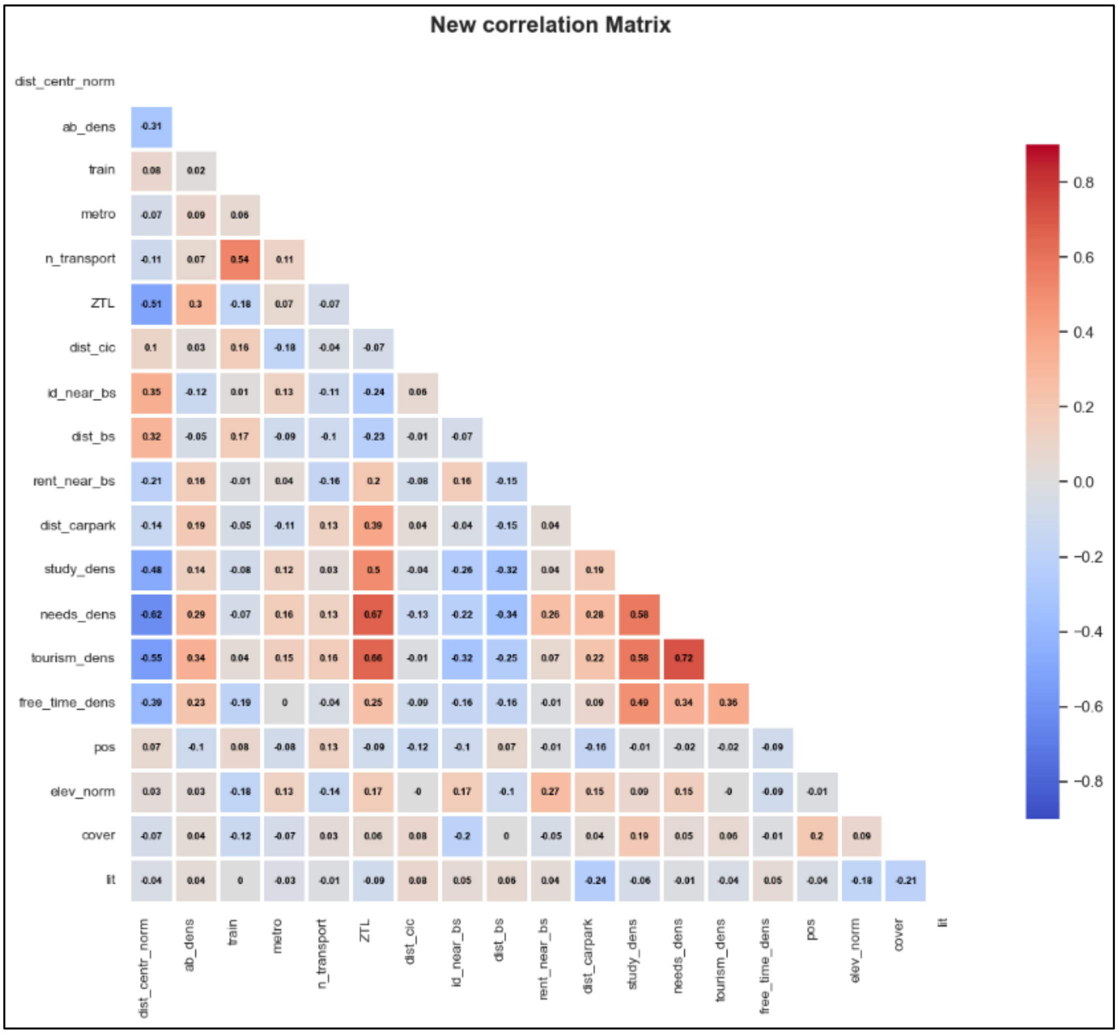


Figure 21. Correlation matrix (after points of interest aggregation).

4.2.5. Multicollinearity analysis

A multicollinearity analysis was conducted to assess the degree of linear dependence among the variables selected as predictors in the regression model. Since highly correlated predictors may distort coefficient estimates and reduce model interpretability, this step ensures the robustness of the analytical framework. The assessment was performed exclusively on variables retained after correlation analysis, including those aggregated into families. This allowed evaluation of multicollinearity only within the final predictor set, ensuring coherence with the modelling strategy. Outcomes, reported in Table 14, show some features with $VIF > 5$; `need_dens` is the most critical one, since its VIF is equal to 8.15. In general, aggregated variables are the ones with the highest values of VIF. This behaviour was predictable, considering the high correlation between them. Without the aggregation carried out previously, the multicollinearity coefficient would have been even higher, with the risk of significant distortions in the analysis model.

Table 14. Multicollinearity analysis outcomes.

Variable	VIF	Variable	VIF
Dist_centr_norm	4.92	Dist_carpark	2.87
Ab_dens	2.84	Study_dens	4.85
Train	2.11	Needs_dens	8.15
Metro	1.37	Tourism_dens	5.38
N_transport	3.78	Free_time_dens	5.41
ZTL	3.53	Pos	4.29
Dist_cic	1.70	Elev_norm	3.60
Dist_bs	2.66	Cover	1.46
Rent_near_bs	1.78	lit	5.59

4.3. Demand model implementation and results

Linear regression, Random Forest and Gradient Boosting are the three machine learning methods selected to analyse our dataset and to identify which are the most influential variables on bike-sharing demand. This chapter is the heart of the entire study, since the models here implemented will provide the basis for all the developments shown in the next sections. Following the previous data preparation, which is essential to avoid encountering unpleasant difficulties during the implementation of the model, models will be created and explained in detail. Consistently with the methodological framework presented in the previous section, the analysis will first introduce and evaluate the linear regression model, before moving on to the two ensemble techniques. This sequential approach allows establishing a baseline and then assessing the added value provided by more advanced machine learning methods.

4.3.1. Data preparation

Data exploration has highlighted the presence of outlier values for many variables (see Figure 11 and Figure 15). Consequently, they must be treated to avoid introducing bias in the models. Partial winsorization has been chosen as the most suitable outlier management technique: data will be capped at the 90th percentile only if the outlier percentage exceeds 5% of the total amount of data. Outlier analysis is applied only to numerical variables. As suggested in paragraph 3.2.1.1, outliers on the right tail of the distribution are investigated for all numerical features in the dataset, while only housing density has been investigated for both tails of the density distribution. Outlier detection has identified outliers in many features. Outcomes are reported in Table 15.

Table 15. Outlier analysis.

Variable	Number of outliers	% of outliers	Winsorization (yes/no)
`Ab_dens`	28 (14 upper, 14 lower)	10.00% upper, 10.00% lower	Yes
`Need_dens`	14	10.00%	Yes
`Tourism_dens`	14	10.00%	Yes
`Dist_centre`	13	9.29%	Yes
`Dist_bs`	13	9.29%	yes
`Study_dens`	13	9.29%	Yes
`N_transport`	11	7.86%	Yes
`Dist_carpark`	11	7.86%	Yes
`Free_time_dens`	9	6.43%	Yes

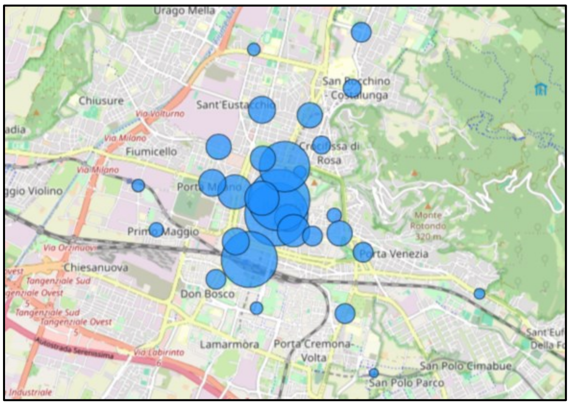
For other features, not reported in Table 15, no outlier has been detected.

The final step before creating the model is selecting and preparing the target variable. The target variable chosen is the number of rentals registered for each station of the dataset during the year 2024. The time interval is wide enough to overcome possible issues due to seasonal peaks in rentals during the year. Thanks to the companies listed in Table 10, it was possible to retrieve the necessary data. For each city, Table 16 resumes the three stations characterised by the highest and lowest number of rentals, among the ones included in the dataset. A common trend is that high numbers of rentals occur at bike-sharing stations located near intermodal transport hubs, such as railway or underground stations. In addition, the most frequently used stations are often associated with the city's main tourist attractions, as in the case of Brescia and Forlì (*P.zza della Loggia* and *P.zza Saffi*).

Table 16. Most and least used stations.

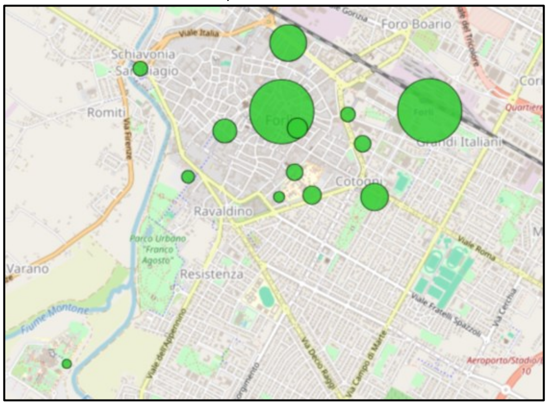
City	Most used BS stations	Least used BS stations
Brescia	1) *Vittoria* (91,270 rentals) 2) *Rovetta* (88,890) 3) *Stazione FS* (78,820)	1) *Poliambulanza* (4,320) 2) *Wuhrer* (7,555) 3) *Villa Glori* (9,498)
Genoa	1) *Caricamento* (957) 2) *De Ferrari* (940) 3) *Rossetti* (649)	1) *Fiumara* (115) 2) *Marina* (141) 3) *Stadio* (165)
Forlì	1) *Piazza Saffi* (3,719) 2) *Stazione FS* (3,700) 3) *Vittorio Veneto* (1,812)	1) *Ospedale* (21) 2) *Lombardini* (182) 3) *Park dell'argine* (213)
Parma	1) *Stazione FS* (16,530) 2) *Santa Croce* (7,851) 3) *Garibaldi* (6,879)	1) *Passo della Cisa* (275) 2) *Rustici* (675) 3) *Lubiana* (784)
Pisa	1) *Stazione FS* (8,450) 2) *Largo Pontecorvo* (6,662) 3) *San Rossore* (5,461)	1) *San Giusto* (383) 2) *Sesta Porta* (420) 3) *La Fontina* (646)
Trieste	1) *Stazione – Piazza Libertà* (8,590) 2) *Stazione marittima* (8,122) 3) *Barcola* (5,194)	1) *Campi Elisi* (349) 2) *Cumano – Musei* (426) 3) *Foro Ulpiano* (966)

The next images (Figure 22) show a portrait of bike-sharing services in the different cities, in which each station is represented by a circle with a radius proportional to its number of rentals.

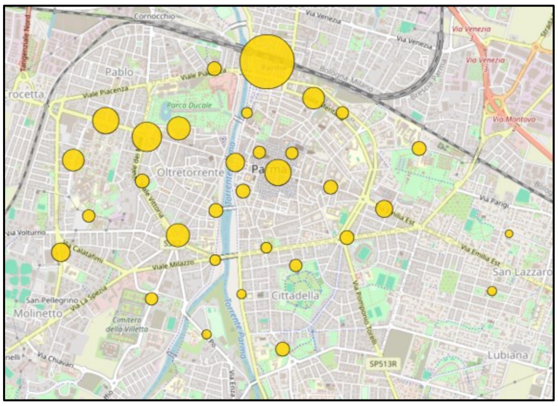


a) Brescia

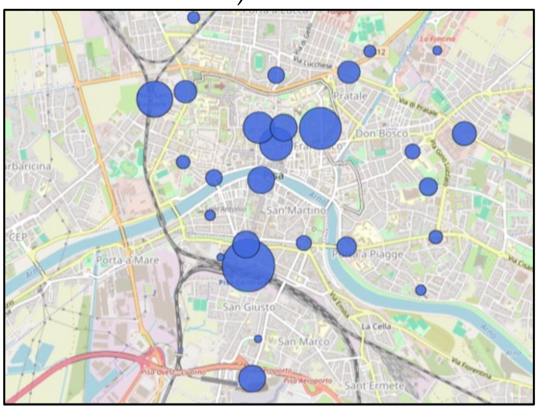
b) Genoa



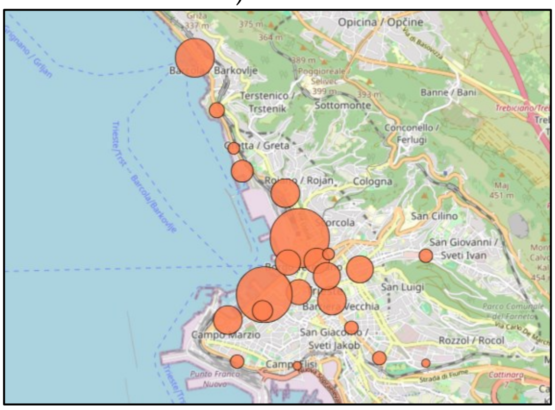
c) Forlì



d) Parma



e) Parma



f) Trieste

Figure 22. Maps of bike-sharing services in the dataset. Each station is represented by a circle with a radius proportional to the number of rentals in year 2024.

Lastly, rentals must be normalised, adapting equation 3.2 to the case of every single city, as shown in Table 17. Normalisation in this case is a very powerful tool, since it allows us to overcome differences in scale among services, both in terms of the number of rentals and stations. Once this operation is performed, the dataset is ready for model creation and implementation.

Table 17. Target variable preparation.

City	Stations in the dataset (k_{city})	$\sum_{i}^{k} rentals_i$	Equation
Brescia	30	943,311	$Target_i(BS) = \frac{rentals_i}{943{,}311} \cdot 30$
Genoa	14	6,619	$Target_i(GE) = \frac{rentals_i}{6{,}619} \cdot 14$
Forlì	14	15,118	$Target_i(FO) = \frac{rentals_i}{15{,}118} \cdot 14$
Parma	32	109,211	$Target_i(PR) = \frac{rentals_i}{109{,}211} \cdot 32$
Pisa	27	76,256	$Target_i(PI) = \frac{rentals_i}{76{,}256} \cdot 27$
Trieste	23	64,628	$Target_i(TS) = \frac{rentals_i}{64{,}628} \cdot 23$

4.3.2. Linear regression

The first modelling step consists of applying an Ordinary Least Squares linear regression. Then, Ridge and Elastic Net regression are introduced to mitigate potential overfitting issues, given that the multicollinearity analysis revealed groups of predictors that are highly correlated with one another. As discussed in the methodological chapter, the models developed here are not intended for predictive purposes, but rather to enhance the understanding of the relationships within the dataset. In this perspective, overfitting is not the primary concern; however, limiting it remains advisable, as excessive overfitting may distort coefficient estimates and introduce unwanted bias into the interpretation of the model.

Linear regression algorithms require data normalisation to perform better. Therefore, all explanatory variables not already comprised between 0 and 1 have been rescaled.

The next three pages point out the application of the three selected linear regression algorithms, indicating their coefficients, t-stat and p-values (Table 18, Table 20, Table 22), computing evaluation metrics (Table 19, Table 21, Table 23) and, lastly, displaying a visual representation of residual and predicted values (Figure 23, Figure 24, Figure 25). Penalisation terms have been computed through a dedicated optimisation function developed in Python programming language (reported in Appendix A).

Application of linear regression demonstrates that these algorithms are not the best solution to study our Italian multi-city dataset. In fact, despite being the most widely used method in literature, in this case, it proved ineffective, as can be seen from the p-values computed for the derived coefficients. In simple regression, only 4 cases out of 17 explanatory variables meet the significance threshold of 0.05. In the penalised cases, however, no term meets this threshold. Furthermore, for each of the cases analysed, the evaluation metrics (especially R^2) are insufficient to adequately explain the variability of the data. Such low values of the evaluation metrics do not even allow us to get an idea of the relative importance of the single explanatory variables, since some of these take on different coefficients and importance depending on the model. For instance, variable `n_transport` has a negative coefficient in the classic model but a positive one in penalised models. Thus, linear regression is not the most suitable method to explain the relationship between the number of accessible transport lines and bike-sharing demand, as confirmed by its p-value, close to 1. This phenomenon also occurs for many other explanatory variables.

In conclusion, outcomes demonstrate that linear regression is not suitable for studying this dataset. Ensemble methods, which are more robust and powerful, may be a better choice.

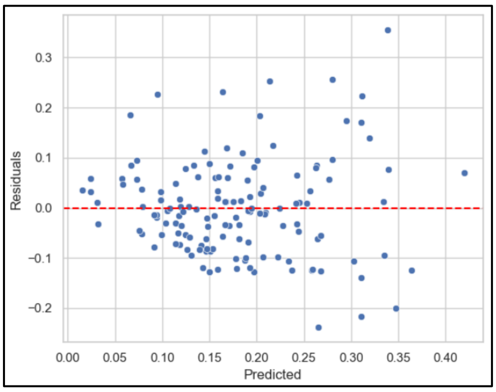
4.3.2.1. Simple linear regression

Table 18. Simple linear regression outcomes.

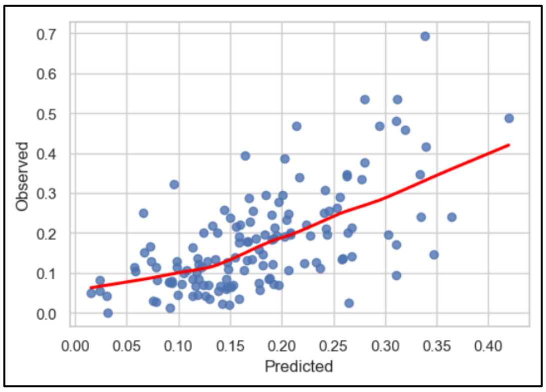
Simple OLS linear regression (target: rent_raw_norm)				
Feature	Coefficient	Standard deviation	t-stat	p-value
Train	0.045	0.011	4.176	0.0001
Needs_dens	0.035	0.017	2.104	0.038
Dist_carpark	0.032	0.011	2.884	0.005
Lit	0.012	0.009	1.295	0.198
Tourism_dens	0.010	0.016	0.673	0.503
Pos	0.009	0.009	0.975	0.331
Ab_dens	0.004	0.010	0.404	0.687
Metro	0.003	0.009	0.352	0.726
Elev_norm	0.001	0.010	0.133	0.894
Rent_bear_bs	-0.002	0.010	-0.160	0.873
Study_dens	-0.003	0.013	-0.168	0.867
Cover	-0.003	0.010	-0.328	0.744
N_transport	-0.008	0.011	-0.700	0.486
Dist_bs	-0.012	0.011	-1.100	0.274
ZTL	-0.013	0.015	-0.905	0.367
Free_time_dens	-0.015	0.011	-1.372	0.173
Dist_centr_norm	-0.015	0.013	-1.188	0.237
Dist_cic	-0.027	0.010	-2.766	0.007

Table 19. Simple linear regression evaluation metrics.

Evaluation metrics: simple OLS linear regression			
$R^2 = 0.402$	$R^2_{adj} = 0.313$	$RMSE = 0.096$	$SMAPE = 47.36\%$



a) Residuals.



b) Observed vs predicted.

Figure 23. Simple linear regression residuals and predicted analysis.

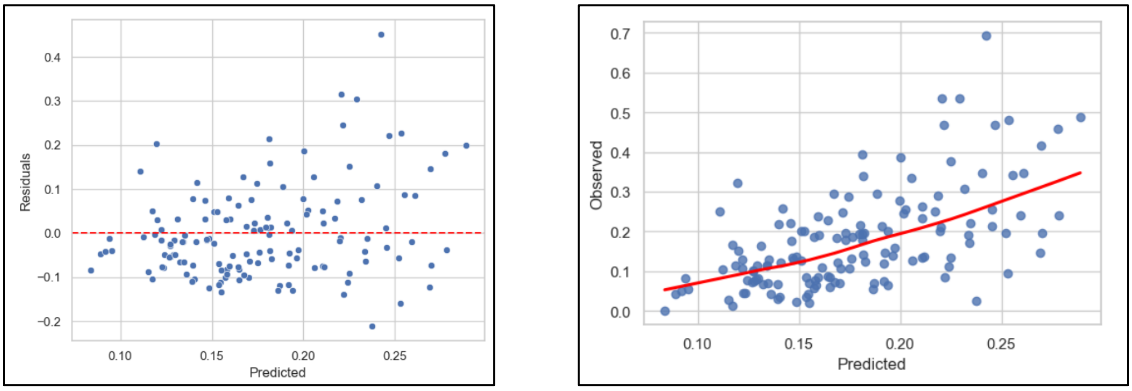
4.3.2.2. Ridge regression

Table 20. Ridge regression outcomes.

Ridge regression (alpha: 187.38, target: rent_raw_norm)				
Feature	**Coefficient**	**Standard deviation**	**t-stat**	**p-value**
Train	0.017	0.012	1.452	0.149
Needs_dens	0.012	0.018	0.656	0.513
Dist_carpark	0.010	0.012	0.830	0.408
Tourism_dens	0.008	0.017	0.475	0.636
N_transport	0.007	0.012	0.579	0.564
Rent_near_bs	0.005	0.011	0.434	0.665
Ab_dens	0.004	0.011	0.397	0.692
Pos	0.004	0.010	0.435	0.664
Metro	0.003	0.010	0.326	0.745
Study_dens	0.003	0.014	0.216	0.829
ZTL	0.003	0.016	0.182	0.855
Lit	0.003	0.010	0.279	0.781
Elev_norm	-0.0002	0.011	-0.017	0.986
Cover	-0.004	0.011	-0.333	0.740
Free_time_dens	-0.004	0.012	-0.360	0.719
Dist_bs	-0.007	0.012	-0.592	0.555
Dist_centr_norm	-0.007	0.014	-0.546	0.586
Dist_cic	-0.009	0.011	-0.847	0.399

Table 21. Ridge regression evaluation metrics.

Evaluation metrics: Ridge linear regression			
$R^2 = 0.305$	$R^2_{adj} = 0.201$	$RMSE = 0.104$	$SMAPE = 48.51\%$



a) Residuals. b) Observed vs predicted.

Figure 24. Ridge regression residuals and predicted analysis.

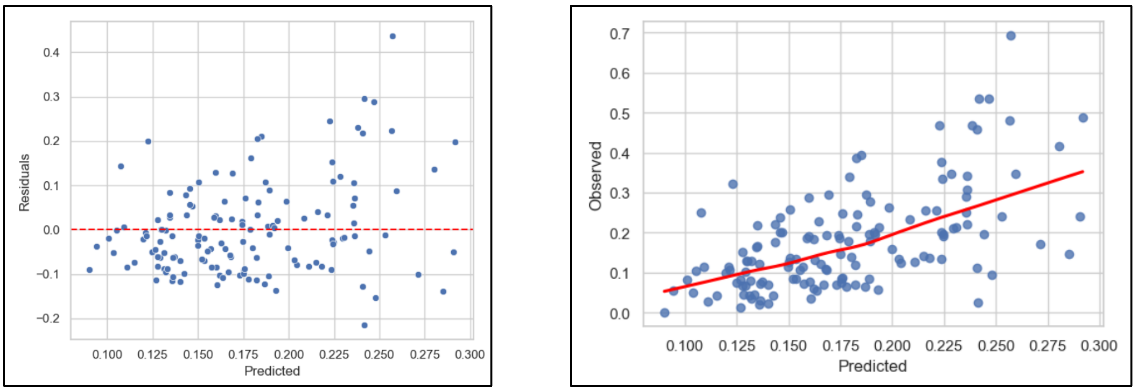
4.3.2.3. Elastic Net regression

Table 22. Elastic net regression outcomes.

Elastic net regression (alpha: 0.25, L1 ratio: 0.05, target: rent_raw_norm)				
Feature	**Coefficient**	**Standard deviation**	**t-stat**	**p-value**
`Train`	0.027	0.012	2.318	0.022
`Needs_dens`	0.023	0.018	1.252	0.213
`Dist_carpark`	0.011	0.012	0.886	0.377
`Tourism_dens`	0.006	0.017	0.357	0.722
`N_transport`	0.0001	0.012	0.007	0.994
`Dist_bs`	-0.005	0.012	-0.398	0.692
`Dist_centr_norm`	-0.005	0.014	-0.365	0.716
`Dist_cic`	-0.010	0.011	-0.931	0.353

Table 23. Elastic net regression evaluation metrics.

Evaluation metrics: Elastic Net regression			
$R^2 = 0.307$	$R^2_{adj} = 0.204$	$RMSE = 0.104$	$SMAPE = 48.72\%$



a) Residuals. b) Observed vs predicted.

Figure 25. Elastic net regression residuals and predicted analysis.

4.3.3. Ensemble methods

The previous Chapter demonstrates that Linear regression models are sensitive to multicollinearity, which can inflate coefficient variance and reduce interpretability. In addition, they assume that the relationship between the target and predictors is linear; this strongly limits their ability to capture interactions or non-linear patterns. Ensemble methods could be an effective solution to these issues. They use multiple regressors at the same time, aggregating them based on different methods. Therefore, variance and model instability are considerably reduced. Moreover, their structure enables them to naturally capture complex and non-linear relationships in the data. The next pages present in detail outcomes from the application of two ensemble regressors: Random Forest, which trains in parallel multiple regressor trees, and Gradient Boosting, which trains iteratively many decision trees, ensuring that each tree corrects the errors introduced by the previous one.

4.3.3.1. Random Forest

Table 24. Random forest: features importance.

Feature	Feature importance	Feature	Feature importance
Needs_dens	0.119	Tourism_dens	0.108
Dist_centr_norm	(-)0.085	Dist_bs	(-)0.084
Study_dens	0.084	Free_time_dens	0.081
Dist_carpark	0.072	Dist_cic	(-)0.066
N_transport	0.064	Rent_near_bs	0.056
Ab_dens	0.052	Elev_norm	0.045
train	0.039	pos	0.016
ZTL	0.009	metro	0.009
Cover	(-)0.006	lit	0.005

Table 25. Random Forest performance metrics.

Random Forest regressor (n_estimators: 300, max_features: "sqrt")			
$R^2 = 0.889$	$R^2_{adj} = 0.873$	$RMSE = 0.041$	$SMAPE = 23.69\%$

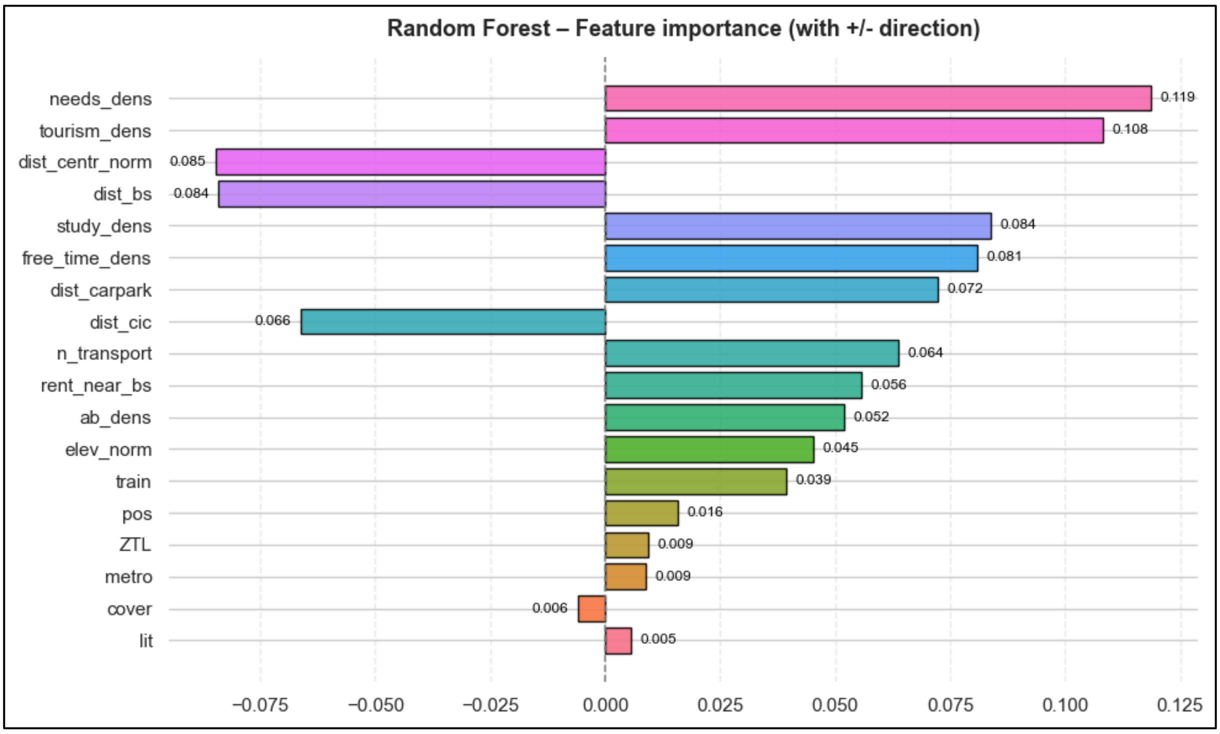


Figure 26. Random forest: graphical representation of feature importance.

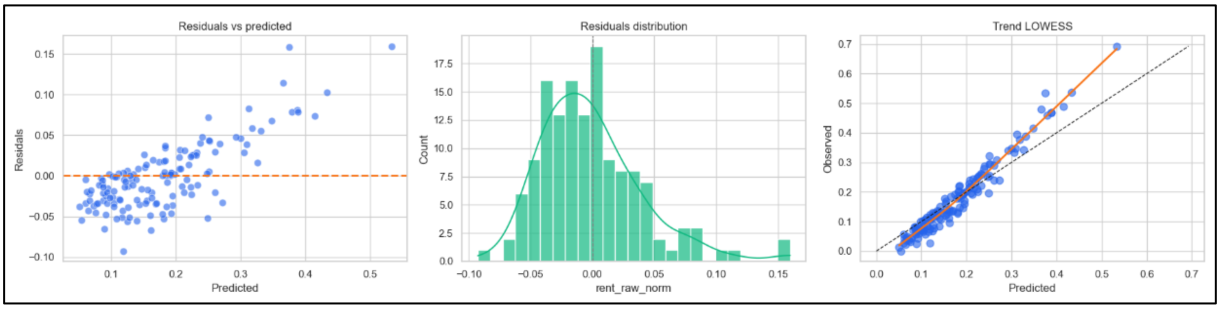


Figure 27. Random forest diagnostics.

4.3.3.2. Gradient Boosting

Table 26. Gradient Boosting: features importance.

Feature	Feature importance	Feature	Feature importance
tourism_dens	0.128	needs_dens	0.121
Free_time_dens	0.097	Dist_centr_norm	(-)0.090
Study_dens	0.079	N_transport	0.074
Dist_cic	(-)0.073	Dist_carpark	0.070
Dist_bs	(-)0.069	Ab_dens	0.066
train	0.041	Elev_norm	0.040
Rent_near_bs	0.038	pos	0.007
lit	0.003	metro	0.002
cover	0.002	ZTL	0.001

Table 27. Gradient boosting: evaluation metrics.

Gradient Boosting regressor (`n_estimators`: 200, `learning_rate`: 0.05, `max_features`: "sqrt")			
$R^2 = 0.954$	$R^2_{adj} = 0.992$	$RMSE = 0.027$	$SMAPE = 18.5\%$

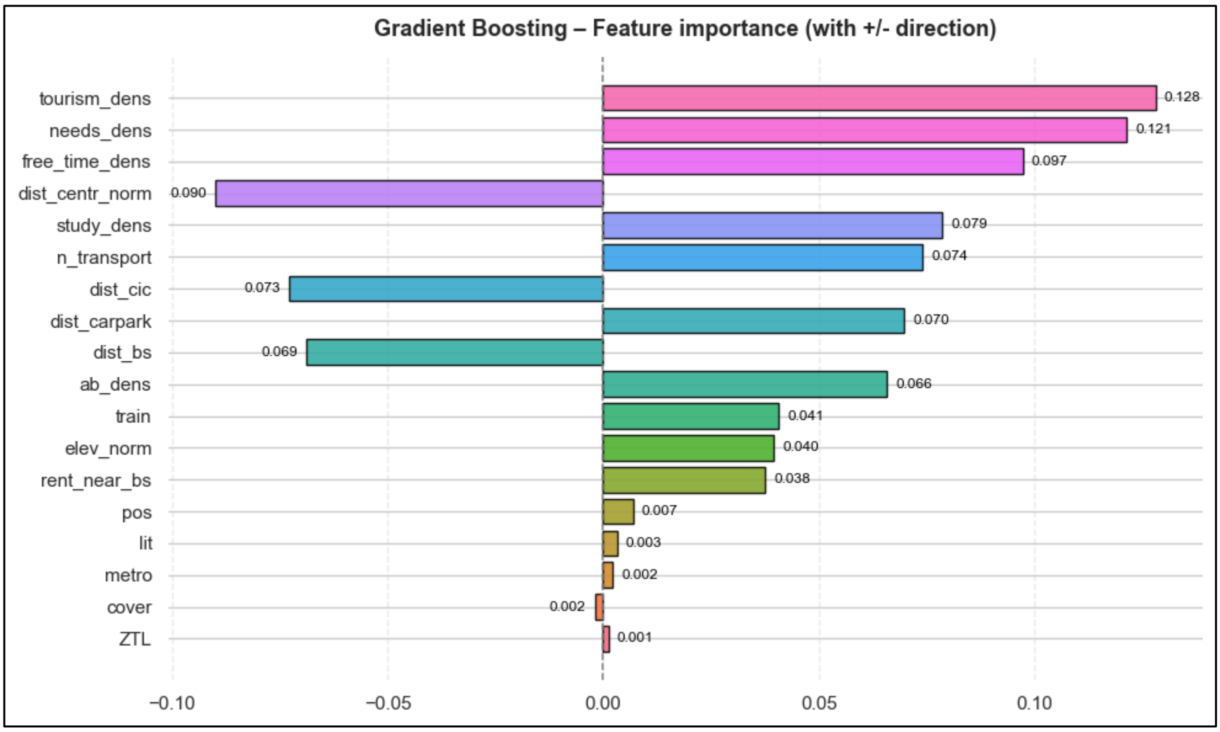


Figure 28. Gradient Boosting. Graphical representation of feature importance.

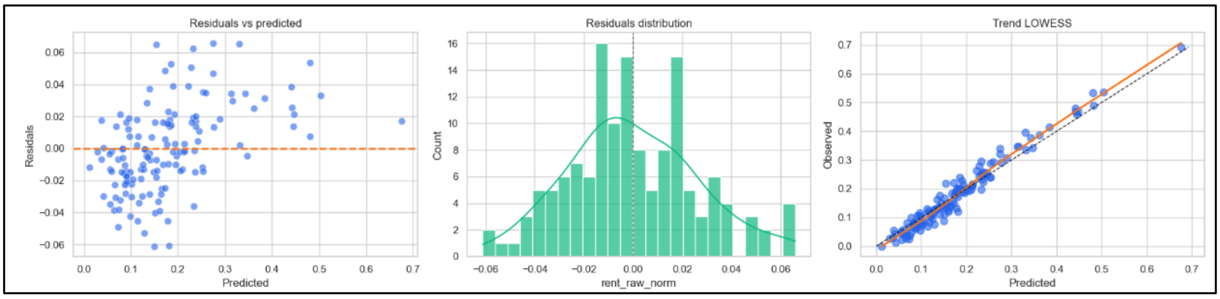


Figure 29. Gradient Boosting diagnostics.

Outcomes of the two ensemble algorithms are consistent, since the variables with the highest feature importances are almost the same in both cases. Despite the number of estimators being lower, Gradient boosting performance is better than Random Forest, as witnessed by R^2 and R^2_{adj} metrics. This might be due to the algorithm's working method, which is designed to correct errors at each iteration. The residuals in the random forest are distributed more to the left, with respect to the zero level. This means that the model tends to underestimate the values. On the contrary, the distribution of gradient boosting residuals is distributed around the zero value, showing no particular tendency to underestimate or overestimate the output. In both cases, the lowess graph shows how the model works well for low to medium values, while uncertainty tends to increase as the number of rentals at a single station increases.

Lastly, it is relevant to underline that the feature importance for each variable cannot be negative. However, we opted to represent all variables that are negatively correlated with the target on the left side of the graphs in Figure 26 and Figure 28 (Correlations have been measured through Pearson coefficients; results are presented in Table 28). This means that as their value increases, the number of expected rentals decreases. The negatively correlated variables, except for weather protection, are all distance variables, indicating that the closer the station is to the element under consideration, the greater the number of rentals will be. This is consistent with real-world experience.

Table 28. Correlation coefficients between rentals and explanatory variables.

Feature	Pearson coefficient	Feature	Pearson coefficient
Need_dens	0.415	Tourism_dens	0.362
Train	0.329	Dist_carpark	0.261
N_transport	0.258	ZTL	0.237
Study_dens	0.234	Ab_dens	0.204
Rent_near_bs	0.166	Metro	0.122
Pos	0.082	Lit	0.029
Free_time_dens	0.028	Dist_cic	-0.164
Cover	-0.053	Dist_centr_norm	-0.304
Dist_bs	-0.265		

At the end of the analysis, the Gradient Boosting model appears as the most appropriate for describing the relationship between demand and the variables of the urban context in which a bike-sharing station is located.

In conclusion, the implemented model provides a clear indication of the urban context factors that most influence bike-sharing demand. It shows that proximity to points of interest is the most important category of variables, especially concerning proximity to touristic attractions, restaurants, shops and leisure POIs (while the presence of school/educational POIs is less relevant). Distance variables also have a significant influence on demand, led by distance from the city centre. Demand results in being higher when the station is close to cycle paths or other bike-sharing stations (which can encourage an alternative to the origin or destination of the journey), while it decreases when the station is close to car parking areas. This could mean that users perceive cars and bike sharing as alternatives to one another, rather than complementary modes of transport, preferring the integrated use of bike sharing and public transport. In fact, the number of accessible transport lines in the surrounding area of the station is of considerable interest, confirming the role of bike sharing as a tool for multimodal integration with the local transport network. Lastly, the residential nature of the neighbourhood where the station is located, the presence of a railway station, the altitude of the station and the number of rentals at the nearest station (indicating whether or not it is located in an area of high demand) also appear to be of moderate importance. In contrast, the influence of the physical characteristics of the station (position in relation to the road, presence of weather protection, lighting) is almost irrelevant, as is the presence of a restricted traffic zone in the area.

Drawing up a ranking, the top 10 variables are as follows, in order of relevance:

1. Proximity to touristic points of interest, such as hotels, monuments, churches, museums;
2. Proximity to personal need points of interest, such as shops, restaurants, and healthcare facilities;
3. Proximity to free time points of interest, such as cinemas and theatres, green areas and fitness attractions;
4. Distance from city centre;
5. Proximity to educational points of interest, such as university buildings, libraries and schools;
6. Number of overall accessible public transport lines;
7. Distance from the nearest bike lane;
8. Distance from the nearest car park;
9. Distance from the nearest bike-sharing station (bike-sharing stations' density);
10. House density near station (thus, an indication of the residential character of the neighbourhood)

The 11th variable also deserves attention, as it relates to the presence of a railway station. During the data exploration phase, it was observed that the bike-sharing station corresponding to the railway station is often the one with the highest number of rentals (Table 16). However, there are a few items in the dataset with a positive value for this binary variable, masking the real importance of the variable `train`. Consequently, it has been decided to include it in SSEI equation.

The eleven selected variables are collectively responsible for 90.7% of the total explanatory importance (obtained by summing the single feature importance values). This reinforces the hypothesis of creating a synthetic index based solely on these explanatory variables, as they are capable of explaining most of the variance in demand. However, this does not mean that the variables not included are irrelevant. Rather, they are only of lesser importance than those selected, as demonstrated by the feature importance calculation.

4.4. SSEI calibration: a combined survey-based and statistical approach

The opportunities offered by having a new performance indicator are manifold. For this reason, it was decided to use data analysis outcomes to create a new quantitative index that can measure the effectiveness of stations' urban integration. For the specifications of this index, called *Station Static Effectiveness Index* (SSEI), please refer to Chapter 3.4. SSEI is independent of the number of available docks and bikes for each station. Therefore, it can be used to perform a "static" performance evaluation without having already established the operative features of the service. In other words, it allows adjustments to decisions made during the station allocation phase. Moreover, another potential application of SSEI could be its use during the design phase, exploiting its equation as one of the functions to be optimised when choosing the best location for stations.

The next paragraphs will incorporate the theoretical concepts from chapter 3.4, explaining in detail the definition and calibration of SSEI using outcomes from the performed data analysis. Then, in section 5, two case studies about the SSEI application will be examined. These involve the computation of the index for the stations in the initial dataset and its integration with a cycling indicator that can measure the quality of cycling infrastructure in the vicinity of a bike-sharing station.

4.4.1. Variable weights definition

The list of main relevant dataset features drawn up as the final output of chapter 4.3 is now resumed to compute the relative numerical influence of each feature on bike-sharing demand. Weights are defined considering simultaneously Pearson correlation coefficients between variables and rentals, variables' linear combination through PCA and an online survey conducted in November 2025 on 127 people – 53.1% males and 45.3% females (1.6% not specified) – both users and non-users of bike-sharing services. Survey's results are reported in Appendix B; it was conducted considering the population of bike-sharing Italian users of 10,800 people, a confidence level of 5% and a margin of error equal to 10%. This results in the minimum sample dimension equal to 96. Weights are then computed, conveying these three elements in equation 3.5 and using soft-max normalisation. Outcomes are reported in Table 29.

Table 29. W_j computation.

Feature	$Pearson_j$	$\lvert w_{PCA} \rvert$	w_{survey}	w_j	W_j
Tourism_dens	0.36	0.43	3.31	0.095	0.95
Needs_dens	0.42	0.45	3.60	0.152	1.52
Free_time_dens	0.03	0.29	3.71	0.131	1.31
Dist_centre	-0.30	0.40	3.40	-0.092	-0.92
Study_dens	0.23	0.40	3.43	0.090	0.90
N_transport	0.26	0.15	3.83	0.201	2.01
Dist_cic	-0.16	0.01	3.58	-0.083	-0.83
Dist_carpark	0.26	0.21	2.29	0.041	0.41
Dist_bs	-0.27	0.28	3.37	-0.072	-0.72
Ab_dens	0.20	0.28	2.80	0.043	0.43

W_j must be now normalised, considering different scales of variables. For example, consider housing density and distance from the nearest cycle path: the former generates values in the order of thousands, while the latter is often equivalent to a few hundred. Final weights W_{dj} are listed in Table 30, in which the standard deviation (obtained from the initial dataset) of each variable has been used to compute the final weights.

Table 30. Weights computation.

Feature	W_j	Standard deviation σ_j	W_{dj}
Tourism_dens	0.95	68.23	13.92
Needs_dens	1.52	201.89	7.53
Free_time_dens	1.31	13.35	98.12
Dist_centre	-0.92	954.54	-0.96
Study_dens	0.90	22.69	39.67
N_transport	2.01	13.41	149.86
Dist_cic	-0.83	161.90	-5.13
Dist_carpark	0.41	155.38	2.64
Dist_bs	-0.72	315.08	-2.29
Ab_dens	0.43	4529.92	0.09

4.4.2. SSEI calibration

SSEI's mathematical formulation takes inspiration from the logistic function, as presented in chapter 3.4.2. Equation 3.9 is chosen as the best formulation of SSEI. In this way, there are 11 variables considered in the equation: 10 numerical variables, already presented in Table 30, and one binary variable, which refers to the presence of a railway station in the buffer around the individual bike-sharing station. The reasons for including this variable have already been explained in paragraph 3.4.2. Furthermore, it is divided into two sub-variables: one, more relevant, relating to the presence of the city's main station, and one, less relevant, relating to the presence of a secondary station. The final equation of the SSEI index is therefore broken down as follows:

- A logistic part, which generates values between 0 and 90, considering the ten variables previously selected;
- A linear part, which generates values between 0 and 10, considering only the variable related to the presence of a train station, ten points in the case of a main station and five in the case of a secondary station.

In this way, SSEI will always have a value between 0 and 100, with the maximum only achievable if there is a main railway station in the buffer. The final SSEI equation for the station i, takes the following form:

$$SSEI_i = \frac{90}{1 + e^{-k(\sum_j^V \frac{x_{ij}W_{dj}}{1000} - c)}} + 10\beta_{main} + 5\beta_{sec} \quad 4.1$$

Where $\beta_{main}, \beta_{sec}$ are dummy variables about the presence of a train station, k and c are parameters that will be calibrated later, V is the number of considered variables j, and $x_{ij}W_{dj}$ is equal to:

$$\begin{aligned} x_{ij}W_{dj} = 149.86 \cdot n_{transport_i} + 98.12 \cdot freetime_{dens_i} + 39.67 \cdot study_{dens_i} \\ + 13.92 \cdot tourism_{dens_i} + 7.53 \cdot needs_{dens_i} - 5.13 \cdot dist_{cic_i} + 2.64 \\ \cdot dist_{carpark_i} - 2.29 \cdot dist_{bs_i} - 0.96 \cdot dist_{centre_i} + 0.09 \cdot ab_{dens_i} \end{aligned} \quad 4.2$$

The division by 1000 of $x_{ij}W_{dj}$ in equation 4.1 was introduced to improve the interpretability of the calculation, avoiding the introduction of numbers of excessive orders of magnitude. k e c are the two parameters of the logistic function, and it is necessary to find their optimum values. k indicates the function's growth rate, while c is the x-coordinate of the inflexion point of the curve.

As regards c, it is calculated as the average of $x_{ij}W_{dj}$ for all stations in the original dataset.

$$c = mean_i \left[\sum_{j=1}^{V} \frac{x_{ij}W_{dj}}{1000} \right] = 6.2 \quad 4.3$$

As proposed in the methodology chapter, k will be calibrated using the data from the initial dataset. Random data groups will be extracted in multiples of 5, for which the value of k will be calculated to minimise the difference in number (expressed in %) between the most represented and the least represented class reported in Chapter 3.4.3. Outcomes of calibration are presented in Table 31. Calibration of parameter k.

Table 31. Calibration of parameter k.

Z (subgroup dimension)	Δ	Δ%	k_{opt}
5	2	40%	0.1
10	3	30%	0.3
15	3	20%	0.3
20	5	25%	0.3
25	6	24%	0.2
30	4	13%	0.4
35	5	14%	0.3
40	7	17%	0.3
45	6	13%	0.3
50	7	14%	0.4
55	7	13%	0.4
60	11	18%	0.3
65	13	20%	0.3
70	10	14%	0.3
75	13	17%	0.3
80	12	15%	0.3
85	11	13%	0.3
90	12	13%	0.3
95	10	11%	0.4
100	17	17%	0.3
105	19	18%	0.3
110	18	16%	0.3
115	20	17%	0.3
120	21	17%	0.3
125	15	12%	0.3
130	24	18%	0.3
135	22	16%	0.3
140	26	19%	0.3

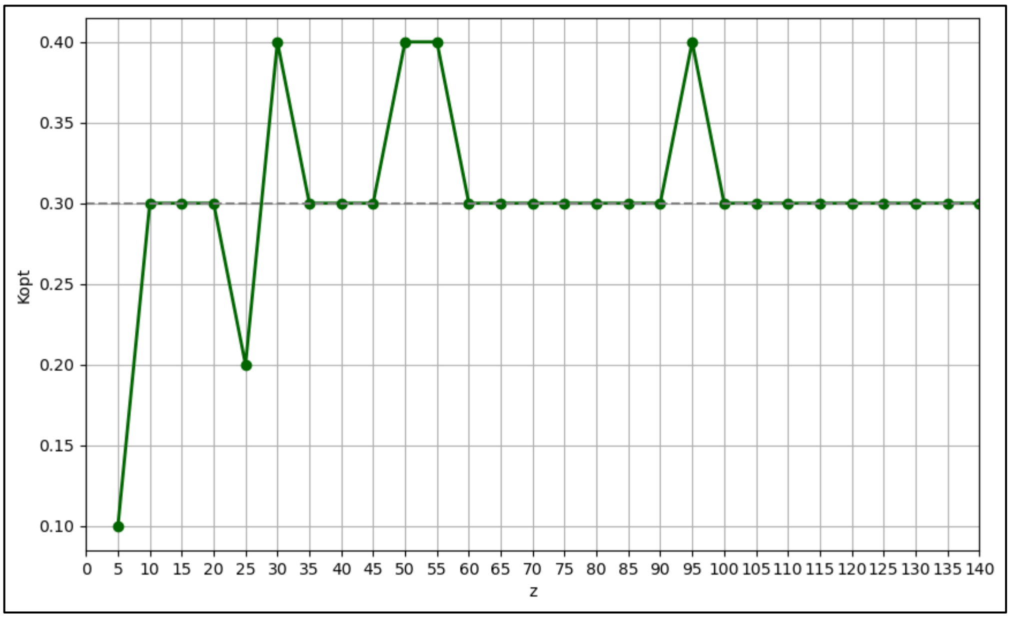


Figure 30. K_{opt} for different z.

The study on the parameter k demonstrates that in 22 cases over 28 (78.6%), the optimum value of k is equal to 0.3. In addition, as the sample size increases, the value of k stabilises at 0.3. Conversely, for smaller samples, variability increases. Figure 31 confirms the choice, highlighting the optimal values of k as the sample size and delta vary. For these reasons, it is concluded that the optimal value of k for the SSEI equation is 0.3. The equation takes the following expression:

$$SSEI_i = \frac{90}{1 + e^{-0.3\left(\sum_j^V \frac{x_{ij} W_{dj}}{1000} - 6.2\right)}} + 10\beta_{main} + 5\beta_{sec} \qquad 4.4$$

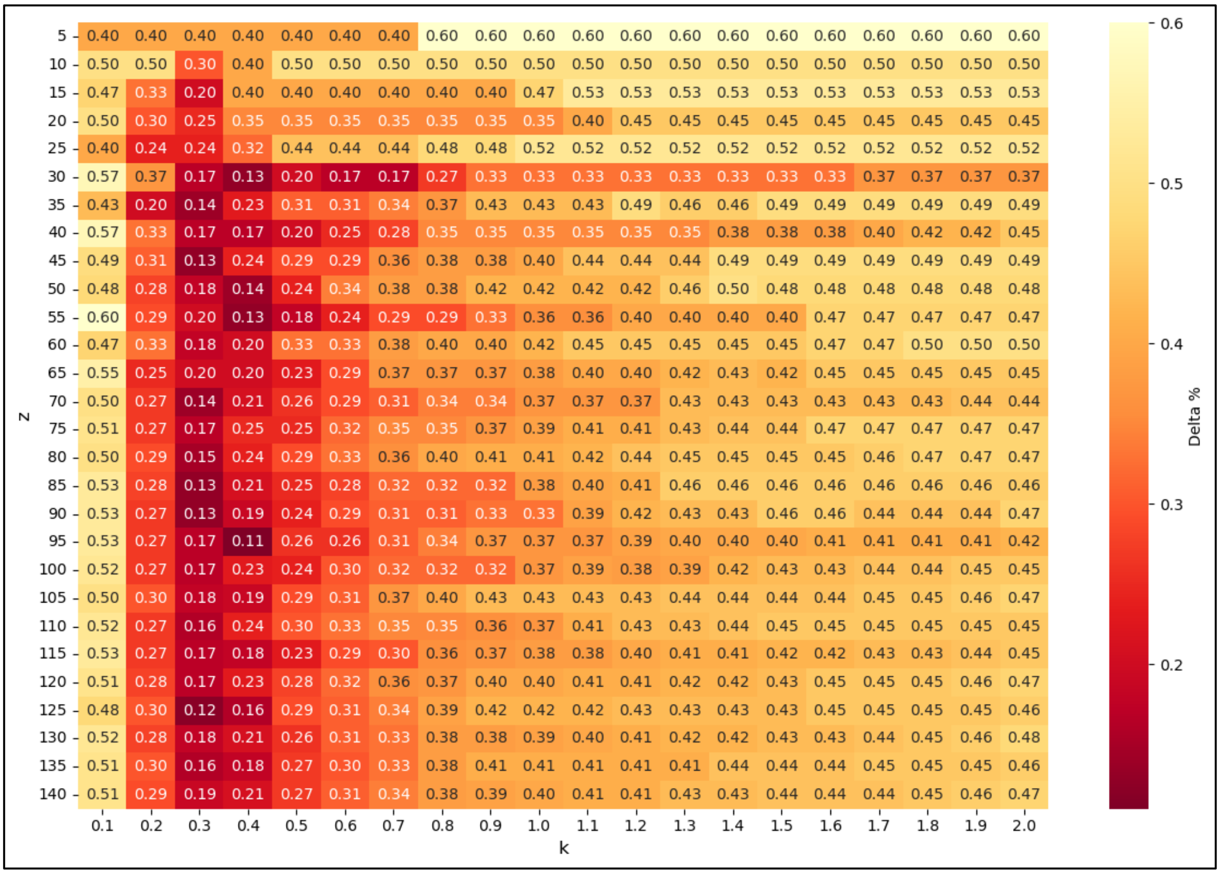


Figure 31. Heatmap of parameter k calibration.

4.5. Local demand analysis: city-level results

The new data analysis methodology is now applied in every single city of the dataset with the aim of studying demand locally, analysing differences with previously studied general trends and among different cities. Local demand study will help planners in different cities to better understand which are the main demand drivers, suggesting which stations to focus on for possible future infrastructure improvements, or in identifying areas where a new bike-sharing station needs to be built.

After an introductory data exploration, the original dataset will be disaggregated into six smaller datasets, one for each city. An adapted Gradient Boosting algorithm will be applied to each dataset, creating a model able to explain the characteristics of demand in each city. Lastly, the obtained outcomes will be discussed, with particular attention to deviations from general trends, analysed in Chapter 4.3.

Models implemented in this section will not compare data from different scales and origins. Thus, normalisation is not required. Features such as elevation and distance from the city centre are therefore used in their original units, without transformation. Similarly, the target variable, representing bike rentals, is also left unnormalised.

4.5.1. Disaggregated data exploration

A brief data exploration, city by city, is now performed. It will not cover all the features in the dataset, but only those deemed most relevant. In particular, graphical representations of points of interest (disaggregated, in terms of counts and not in terms of densities), distance variables from individual transport infrastructures and the number of accessible transport lines per station will be presented and discussed in the next paragraphs, highlighting dissimilarities among cities.

4.5.1.1. Points of interest

POIs exploration (Figure 32) points out that Genoa emerges as the most service-rich city for almost all POI categories, while Forlì is consistently the least service-dense. This phenomenon is mainly due to the dimensions and area coverage of these cities, which are respectively the most and the least populated among those in the study. At the same time, Brescia and Parma show mixed patterns, with some categories well represented and others less. Lastly, Pisa and Trieste usually occupy middle positions, with moderate POI numbers in some categories, such as hotels and tourist attractions, which reflect the touristic profile of these cities.

Pisa, Parma and Genoa are on the podium of the cities with the highest average of university buildings around bike-sharing stations. These suggest that the service is designed with a particular focus on young people and, in particular, on university students who, as seen above, make up a large proportion of the regular users of bike-sharing services.

Genoa is the city with the highest museum density, mainly due to its famous nautical and maritime museums (*Acquario di Genoa, Biosfera, Museo Del Mare…).* Cinemas and theatres values are consistently low across all cities, reflecting a generally low presence of entertainment venues.

Regarding shops, cafes and restaurants, Forlì again appears as the city with the lowest counts while Genoa leads the ranking. Parma, Brescia and Pisa show high variability, indicating heterogeneous commercial conditions across station locations.

Trieste, Pisa and Genoa stand out from other cities in terms of the number of hotels. This is likely due to tourist- or business-related demand. At the same time, Brescia is characterised by high dispersion, indicating big differences between central and peripheral stations.

Tourist attractions show a marked variation across cities. Genoa records by far the highest concentration of attractions near bike-sharing stations, reflecting its strong touristic character and dense historical core. Pisa presents only moderate values, suggesting that major attractions are not uniformly covered by the system. The remaining cities show relatively low levels, indicating that bike-sharing stations are generally located in areas with limited touristic functions.

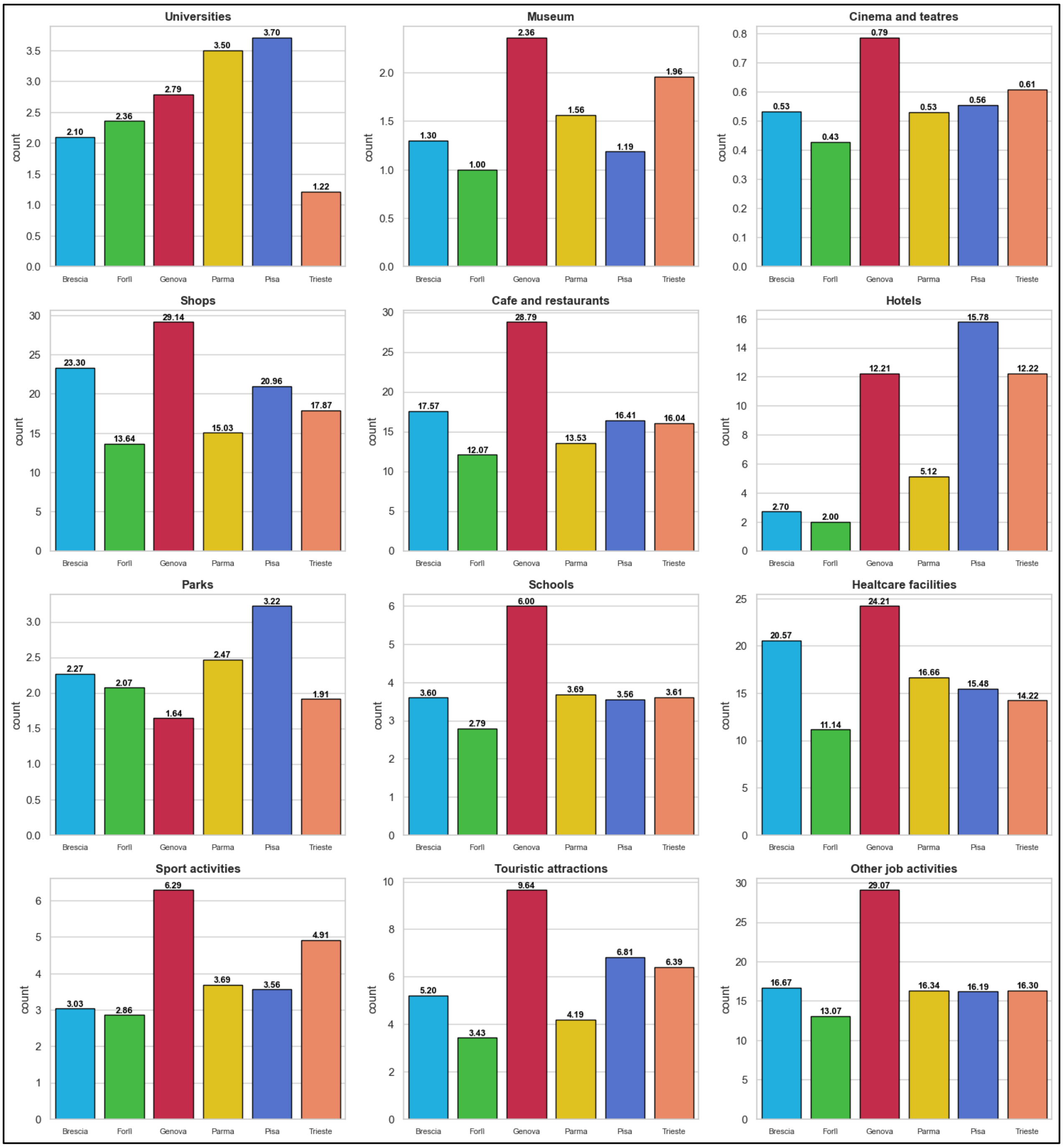


Figure 32. Disaggregated POIs histograms by city.

4.5.1.2. Distances from individual transport infrastructures

Pisa appears as the city with the longest distances between a bike-sharing station and the nearest bike lane. This may indicate a lack of cycling infrastructure in the city. Forlì and Trieste rank second and third, while Brescia and Genoa are characterised by minimal distances between cycle paths and bike-sharing stations. In these cities, therefore, the cycling infrastructure is adequately positioned to support the use of bicycles through bike-sharing, encouraging the start or end of the trip by a dedicated lane.

Distances between bike-sharing stations record values around 500 metres. This suggests a good density of stations, which emphasises bike usage for short to medium trips. In particular, Genoa emerges as the city with the lowest distances, because of the geographical disposition of stations (Figure 9), equally spaced from each other.

Lastly, distance from a car park is the variable that presents more variability, especially for cities like Pisa and Parma. The highest distances are registered in Genoa, Pisa and Parma, while Brescia, Trieste and Forlì have similar average values, creating a cluster below the other cities. The location of a bike-sharing station near a car park can be a way to encourage modal shift, especially concerning first and last-mile mobility. Therefore, in the three cities mentioned above, the use of bike-sharing services could be encouraged precisely because of the relatively short distances from car parks.

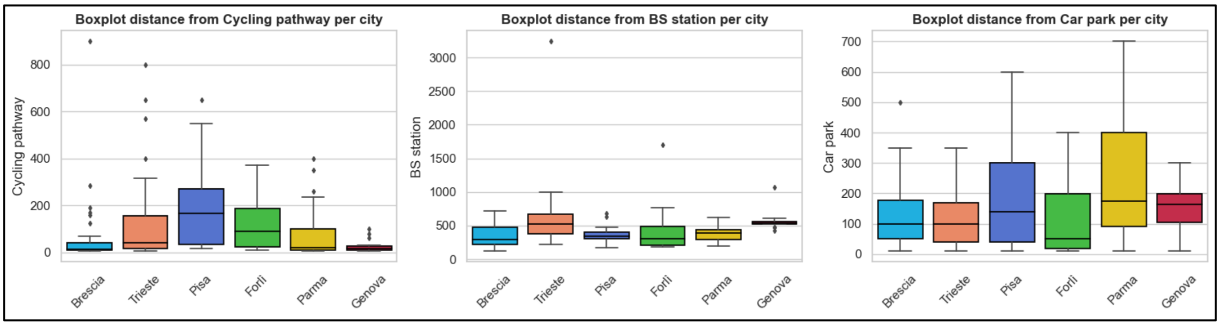


Figure 33. Distances from individual transport infrastructures by city boxplots.

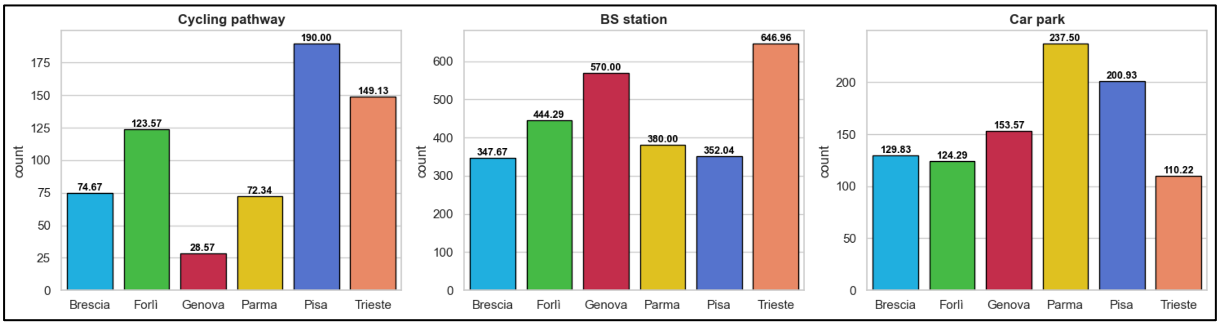


Figure 34. Distances from individual transport infrastructures by city histograms.

4.5.1.3. Number of overall accessible transport lines

The number of accessible public transport lines around each bike-sharing station varies significantly across the analysed cities. Parma and Genoa present the highest average values, indicating strong multimodal integration and a dense transport supply that can facilitate intermodal trips. Pisa, Trieste, and Forlì show intermediate levels of accessibility, with distributions (Figure 35 and Figure 36) characterised by moderate variability across stations. Brescia, on the other hand, records the lowest values, suggesting limited public transport coverage in the immediate surroundings of its bike-sharing network. Overall, these differences reflect distinct urban contexts and planning strategies, with some cities offering well-connected multimodal hubs and others characterised by more locally oriented station placement.

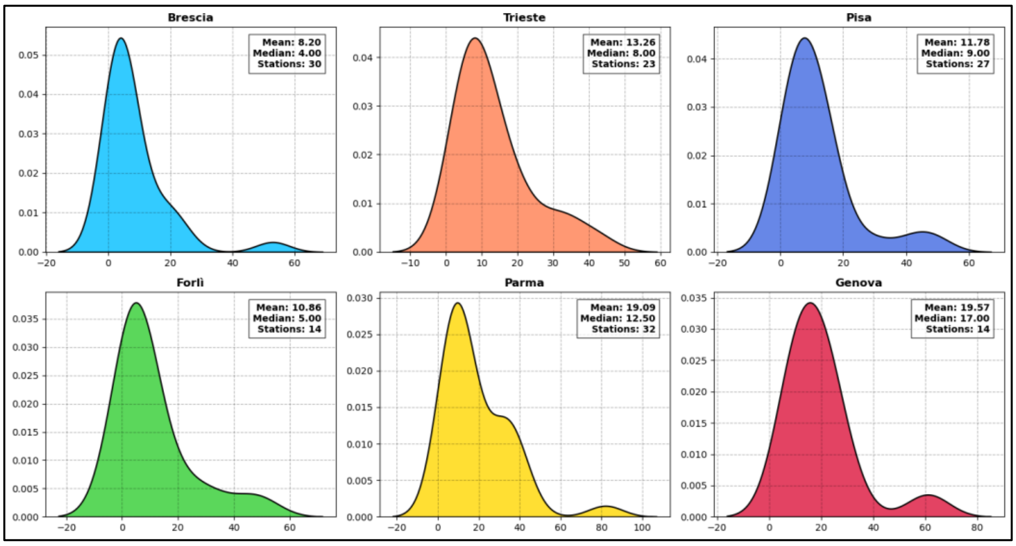


Figure 35. Number of overall accessible transport lines by city KDE plots.

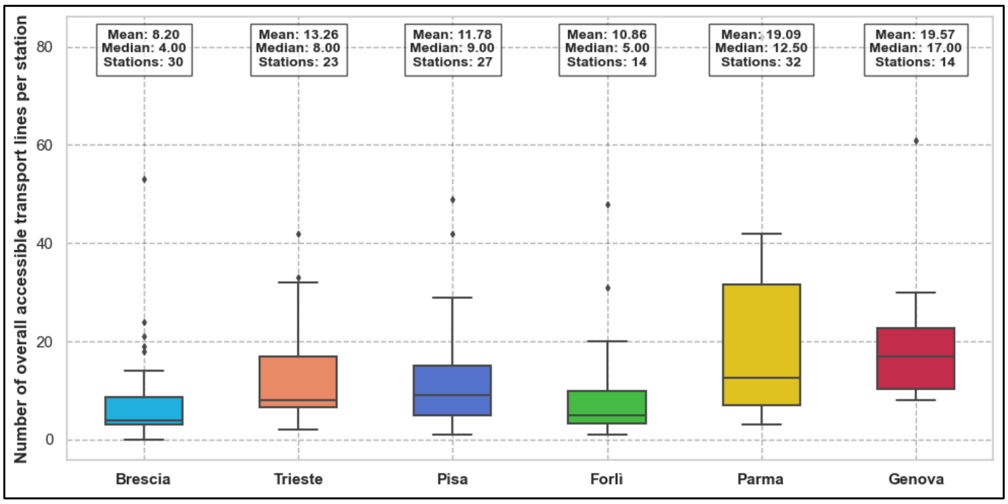


Figure 36. Number of overall accessible transport lines by city boxplots.

4.5.2. Brescia

The bike-sharing service in Brescia, called *Bicimia*, is the largest in the dataset. It is characterised by significantly higher rental numbers than other cities, often exceeding them by two orders of magnitude. The high usage by citizens, mostly young people, is mainly because the service is currently free for the first 45 minutes of the ride. Brescia's municipal administration is particularly keen on the city's bike-sharing service. So recently, new stations have been introduced to expand the service and reach an increasing number of citizens.

The analysis of outliers (Table 32), conducted on the 30 selected stations using the same methodology as before, showed a constant percentage (10%) of values exceeding the imposed limits for the categories listed in Table 32. Consequently, these data are processed, modified and set to the appropriate percentile value so as not to interfere with the analysis model that will be implemented later. The consistent presence of outliers across density and distance variables indicates heterogeneous spatial conditions among Brescia's stations, with some situated in uniquely dense or peripheral contexts.

Table 32. Brescia: outliers analysis.

Feature	Number of outliers	% of outliers	Winsorization (yes/no)
`Ab_dens`	6 (3 upper, 3 lower)	10.0% upper, 10.0% lower	Yes
`Dist_centre`	3	10.0%	Yes
`N_transport`	3	10.0%	Yes
`Dist_cic`	3	10.0%	Yes
`Dist_bs`	3	10.0%	yes
`Dist_carpark`	3	10.0%	Yes
`Study_dens`	3	10.0%	Yes
`Needs_dens`	3	10.0%	Yes
`Tourism_dens`	3	10.0%	Yes
`Free_time_dens`	3	10.0%	Yes

The model application shows good values for performance metrics (Table 33), achieving an excellent fit (R^2 and R^2_{adj}). This indicates that features in the Brescia dataset collectively explain almost all the variability of rentals. Similarly, error metrics (RMSE and SMAPE) indicate high predictive accuracy and stable model behaviour. Let's note that RMSE now has a greater value than the previous applications, because the target variable is no longer normalised.

Table 33. Brescia: model evaluation metrics

Gradient Boosting regressor (`n_estimators`: 100, `learning_rate`: 0.05, `max_features`: "sqrt")			
$R^2 = 0.995$	$R^2_{adj} = 0.989$	$RMSE = 1638.4$	$SMAPE = 6.80\%$

When calculating the feature importance (Figure 37), three variables stand out above the others, making them the most influential on demand:

- Distance from city centre, which is negatively correlated with rentals, meaning that the closer a station is to the city centre (*Piazza Della Loggia*), the higher the number of rentals associated with it will be.
- Needs (Restaurants, shops, healthcare…) density is the second most important feature, showing a strong positive relationship with rental volumes.
- Distance from the nearest bike-sharing station is also a feature with a substantial positive influence. The greater the density of stations (the shorter the distance between them), the higher the demand for journeys will be. This could mean that Brescia citizens use the service for short journeys, covering distances shorter than the average distance between stations (about 350m).

Comparing outcomes to the previously studied general trends, it can be observed that the presence of tourist attractions decreases, while the importance of distance from the cycle path and the presence of educational attractions declines drastically.

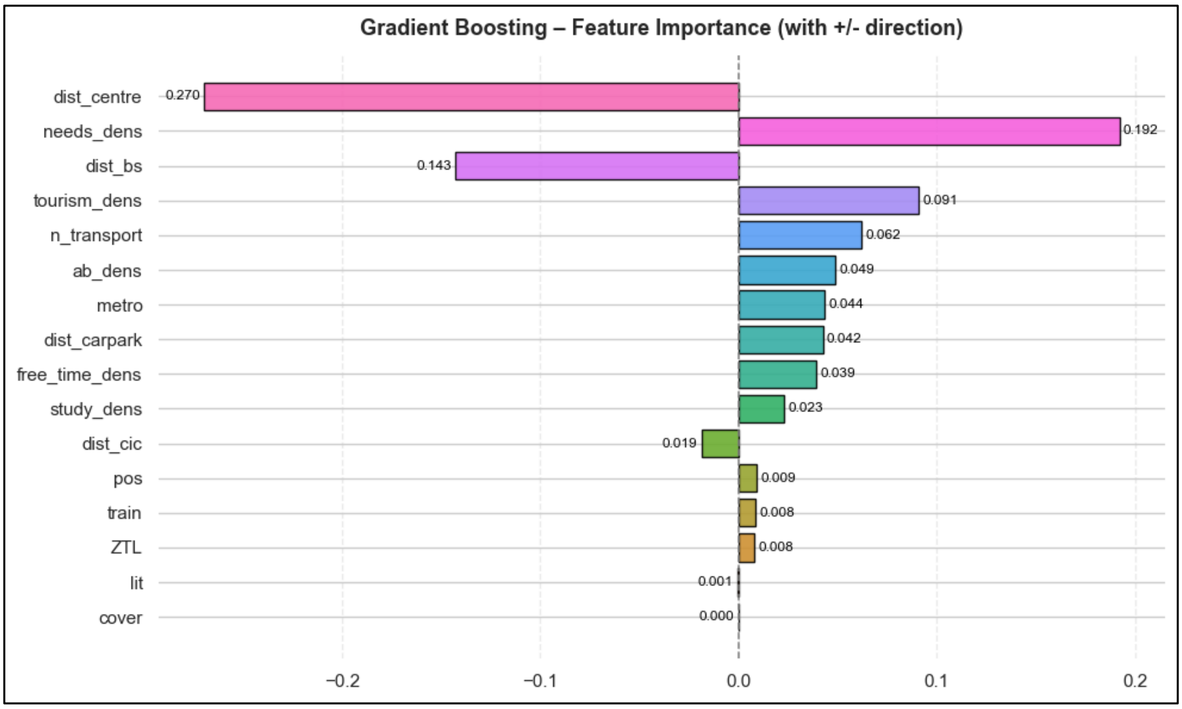


Figure 37. Brescia: graphical representation of Feature Importance.

4.5.3. Genoa

The bike-sharing service in Genoa runs along the coast of the Ligurian capital and in the city centre (around *P.zza De Ferrari*), where the stations with the highest number of rentals are located. Despite being the largest city in terms of size and population, the bike-sharing service is not widely used, with no station exceeding 1,000 total rentals in 2024. Overall, there are 14 active stations in the service. The service requires an annual subscription of €20, which allows users to travel for free for the first 20 minutes of each trip.

Outlier analysis (Table 34) shows a slightly high proportion of outliers (14.3%) across most features, indicating great variability and heterogeneity in the spatial and functional characteristics of its bike-sharing. For almost all variables, the number of outliers is equal to 2, meaning that there are usually two stations that exceed limit values. The fact that there is only one outlier for variable `dist_bs` means that stations in Genoa are equally spaced from each other, as mentioned during data exploration.

Table 34. Genoa: outliers analysis.

Feature	Number of outliers	% of outliers	Winsorization (yes/no)
`Ab_dens`	4 (2 upper, 2 lower)	14.3% upper, 14.3% lower	Yes
`Dist_centre`	2	14.3%	Yes
`N_transport`	2	14.3%	Yes
`Dist_cic`	2	14.3%	Yes
`Dist_carpark`	2	14.3%	Yes
`Study_dens`	2	14.3%	Yes
`Needs_dens`	2	14.3%	Yes
`Tourism_dens`	2	14.3%	Yes
`Free_time_dens`	2	14.3%	Yes
`Dist_bs`	1	7.1%	Yes

The Gradient Boosting model achieves perfect fit metrics (Table 35). This is because the number of stations is very low (14). Thus, a powerful algorithm like Gradient Boosting can discover the inner relationships between the analysed data. So, the model is good for data analysis but not suitable for the number of rentals prediction, since the risk of overfitting is very high. At the same time, the model shows very good error metrics.

Table 35. Genoa: model evaluation metrics.

Gradient Boosting regressor (`n_estimators`: 100, `learning_rate`: 0.05, `max_features`: “sqrt”)			
$R^2 = 1$	$R^2_{adj} = 1$	$RMSE = 3.8$	$SMAPE = 1.23\%$

Among the feature importances (Figure 38), the `pos` variable is particularly noteworthy, climbing many positions compared to the previous classifications. This means that the Genoa residents tend to choose stations that give them a greater sense of safety (recall that as the value of this variable increases, the level of physical separation between the bike-sharing station and the road increases). In second place is `tourism_dens`, which was already high in the general ranking. The housing density variable, which reflects the type of neighbourhood in which the station is located, is also particularly important. Therefore, there is high demand in neighbourhoods with a strong residential character (high housing density). Despite the presence of the underground, proximity to stations does not have a particularly significant impact on demand for bike-sharing. The same behaviour is also observed for proximity to railway stations, although the bike-sharing station near *Brignole* railway station has the highest number of rentals.

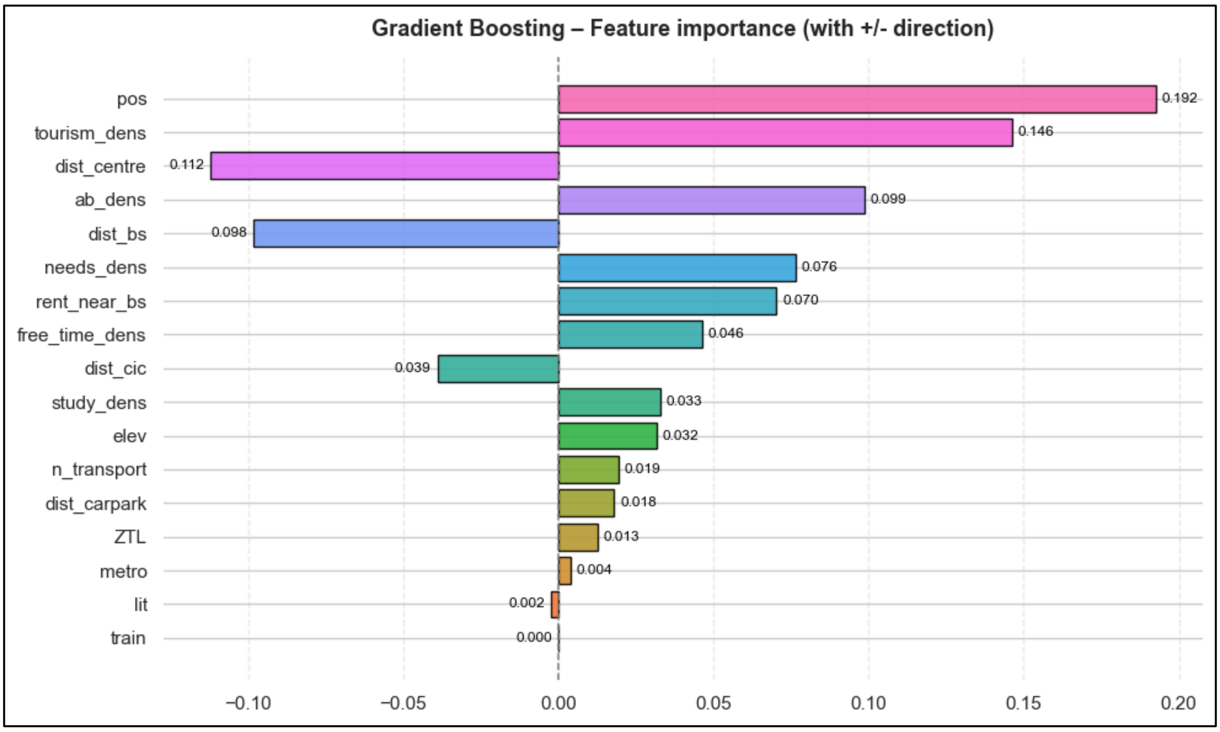


Figure 38. Genoa: graphical representation of Feature Importance.

4.5.4. Forlì

The bike-sharing service in Forlì, part of the *"Mi Muovo in bici"* project, includes 14 bike-sharing stations. Specifically, 13 of these are located near the city centre, while one is located near the hospital on the outskirts. The service, managed by *Weelo*, involves more than 18,000 subscribers, corresponding to almost 20% of Forlì's citizens. People can subscribe to the service through an annual subscription (for €25) or a daily subscription (for €5). Subscribers can enjoy half an hour free for each trip. The busiest stations are those near the railway station and the city's central square (*P.zza Saffi*).

Analysis of outliers (Table 36) reveals trends similar to those in the city of Genoa, with a constant percentage of values exceeding the imposed limits, which will therefore be treated as explained above.

Table 36. Forlì: outliers analysis.

Feature	Number of outliers	% of outliers	Winsorization (yes/no)
`Ab_dens`	4 (2 upper, 2 lower)	14.3% upper, 14.3% lower	Yes
`Dist_centre`	2	14.3%	Yes
`N_transport`	2	14.3%	Yes
`Dist_carpark`	2	14.3%	Yes
`Needs_dens`	2	14.3%	Yes
`Tourism_dens`	2	14.3%	Yes
`Free_time_dens`	2	14.3%	Yes
`Dist_bs`	2	14.3%	Yes
`Dist_cic`	1	7.1%	Yes
`Study_dens`	1	7.1%	Yes

Since the size of the service is similar to that of Genoa, the evaluation metrics of the Gradient Boosting model will have the same characteristics (Table 37). In fact, the model variables can fully explain the variability of rentals at each station, while the percentage error is just 7.83%.

Table 37. Forlì: model evaluation metrics.

Gradient Boosting regressor (`n_estimators`: 100, `learning_rate`: 0.05, `max_features`: "sqrt")			
$R^2 = 1$	$R^2_{adj} = 1$	$RMSE = 22.3$	$SMAPE = 7.83\%$

For the first time, the variable related to distance from car parks ranks first among the most relevant drivers of demand (Figure 39). The greater the distance from a car park, the higher the number of rentals expected. This could suggest that Forlì citizens see bicycles as a substitute for cars: in areas where it is more difficult to park and therefore use a car, people prefer to use bike-sharing. In second place is `dist_centre`, meaning that stations in the centre are more attractive. The number of accessible transport lines is also relevant; this reflects the fact that stations are interpreted as intermodal hubs. For the first time, we see that `study_dens` is negatively correlated with the number of rentals. This means that stations near universities (in Forlì, there are two bike-sharing stations near the main university campus) are not very attractive to students. Finally, the physical characteristics of the station once again play an almost negligible role.

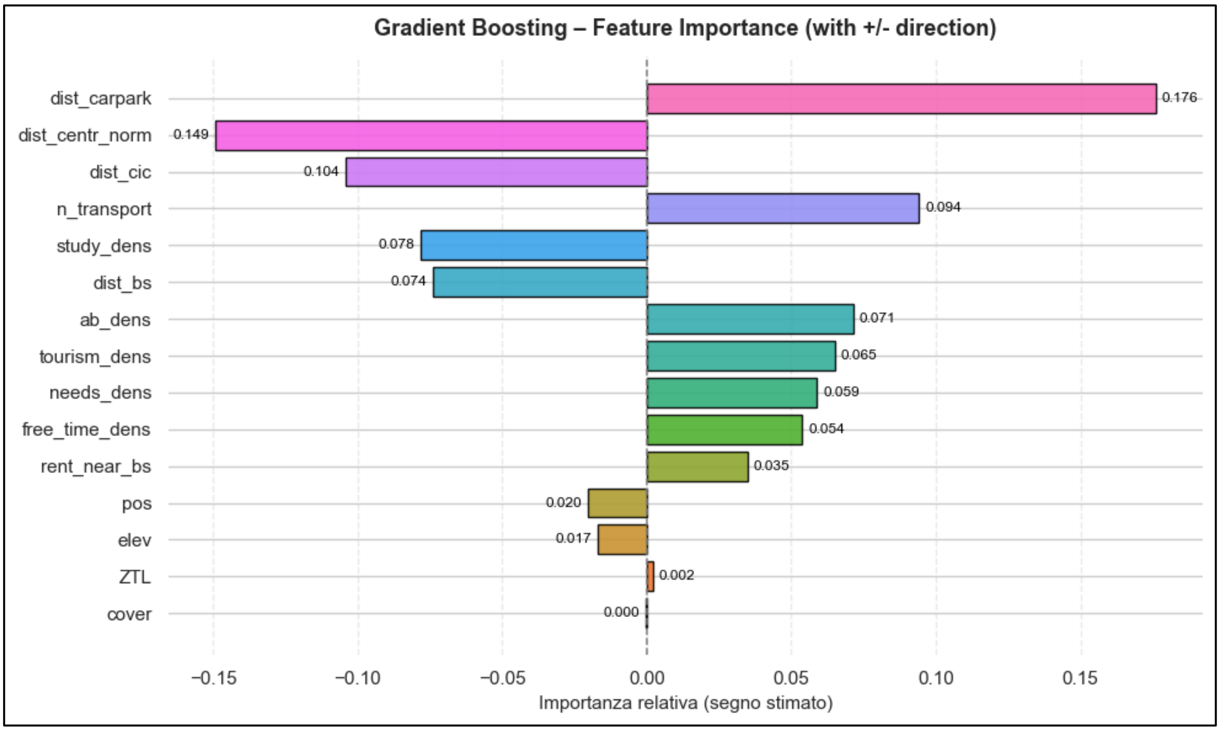


Figure 39. Forlì: graphical representation of Feature Importance.

4.5.5. Parma

The station-based Parma bike-sharing service analysed in the dataset is no longer operational since December 2024. However, it was decided to include it in the analysis since it has been replaced by a hybrid service, in which users can leave bikes only in dedicated areas. Thus, dedicated areas have replaced the previous bike-sharing stations, ensuring that the study of the urban context surrounding them can still be relevant. Currently, the service can rely on 550 bicycles, both traditional and electric. There are different typologies of subscription: yearly-pass, monthly-pass or two-day-pass. Subscribers can benefit from 30 minutes of free travel before starting to pay a fixed fee every half hour. In addition, there is also the possibility of an occasional subscription, which costs 1,20€ each 30 minutes of travel.

Outlier analysis (Table 38) shows that there is an almost constant percentage of anomalous data for each explanatory variable, suggesting that stations belong to heterogeneous contexts. In particular, `ab_dens` stands out as the variable with many outliers (six, considering both lower and upper limits), pointing out big differences in terms of housing density in the surroundings of stations.

Table 38. Parma: outliers analysis.

Feature	**Number of outliers**	**% of outliers**	**Winsorization (yes/no)**
`Ab_dens`	6 (3 upper, 3 lower)	9.7% upper, 9.7% lower	Yes
`Dist_centre`	3	9.7%	Yes
`N_transport`	3	9.7%	Yes
`Dist_cic`	3	9.7%	Yes
`Dist_carpark`	3	9.7%	Yes
`Study_dens`	3	9.7%	Yes
`Needs_dens`	3	9.7%	Yes
`Tourism_dens`	3	9.7%	Yes
`Free_time_dens`	3	9.7%	Yes
`Dist_bs`	2	6.5%	Yes

Performance of the model (Table 39) is still very high, even if not perfect, since the number of stations included in the model is higher than cities like Genoa and Forlì. R^2 and R^2_{adj} are very high, with values near one. Moreover, prediction errors also show good performance indications.

Table 39. Parma: model evaluation metrics.

Gradient Boosting regressor (`n_estimators`: 100, `learning_rate`: 0.05, `max_features`: "sqrt")			
$R^2 = 0.990$	$R^2_{adj} = 0.977$	$RMSE = 196.4$	$SMAPE = 7.22\%$

Feature importances (Figure 40) show some deviations from the general trends. `needs_dens` stands out as the most relevant explanatory variable, followed by the proximity to bike lanes. The presence of educational points of interest (`study_dens`) assumes a moderate importance, higher than in other cities. This reflects the university character of the city of Parma, whose university is renowned and well attended. Students, who often do not have their own motor vehicles, use bike-sharing to travel to and from university buildings. Also, distance from car park, proximity to popular bike-sharing stations and presence of tourist attractions play an important role in defining target variability. Compared to the general case, distance from the city centre has less relevance, suggesting that Parma stations are still attractive even if far from the cultural and economic central point of the city. Presence of free time attractions and housing density have slight importance, while `n_transport` has low feature importance. This might suggest that Parma residents do not see bike-sharing as a tool for promoting intermodal exchange, but rather as a method for making the entire journey, from origin to destination. Lastly, station positioning, cover and lighting have the lowest feature importance. The fact that the variable related to the presence of a railway station has zero feature importance should not be misleading. In fact, there is only one bike-sharing station with a positive value for this variable, but it may not influence the split of the model decision trees.

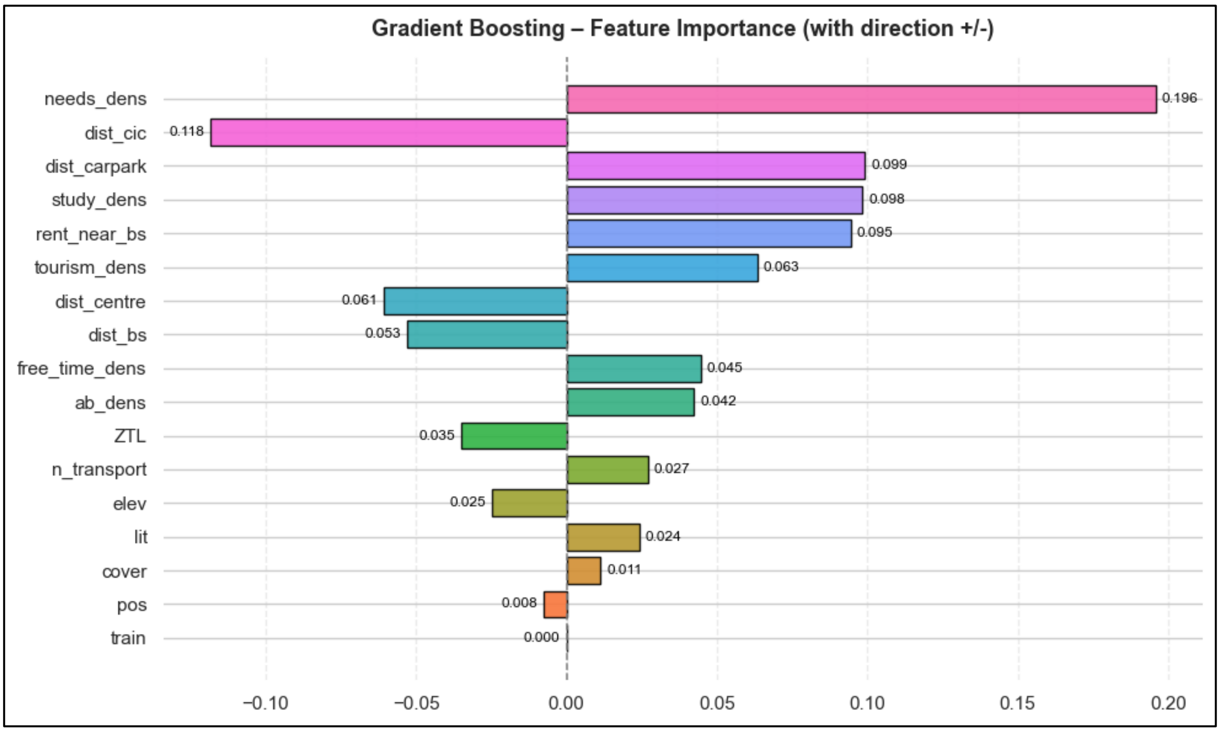


Figure 40. Parma: graphical representation of Feature Importance.

4.5.6. Pisa

City of Pisa bike-sharing, called *CicloPi*, provides citizens with 200 bicycles distributed in 39 different stations in Pisa and in the surrounding suburbs. Pisa is the only city, among the analysed ones, with a bike-sharing station in the proximity of the city airport. Currently, there are almost 14,000 subscribers. People can subscribe to an annual pass for 35€. There is a special offer for students who can subscribe to the service for just 25€ per year. Subscription includes the first 30 minutes of free trip. After that, the fees will increase for each half-hour of travel.

All variables for which anomalous values were recorded were winsorised (Table 40). The percentage of outliers reported is slightly higher than in other cities with more than 15 stations previously analysed. This indicates greater heterogeneity in Pisa's urban conditions, given that the city has stations in different contexts, from the city centre areas dense with tourist and educational attractions to the more peripheral areas with poor infrastructure.

Table 40. Pisa: outliers analysis.

Feature	Number of outliers	% of outliers	Winsorization (yes/no)
`Ab_dens`	6 (3 upper, 3 lower)	11.1% upper, 11.1% lower	Yes
`Dist_centre`	3	11.1%	Yes
`N_transport`	3	11.1%	Yes
`Dist_cic`	3	11.1%	Yes
`Dist_bs`	3	11.1%	Yes
`Study_dens`	3	11.1%	Yes
`Tourism_dens`	3	11.1%	Yes
`Free_time_dens`	3	11.1%	Yes
`Dist_carpark`	2	7.4%	Yes
`Needs_dens`	2	7.4%	Yes

Model performance (Table 41) is slightly better than the Parma case. Explanatory features can explain almost the whole target variability ($R^2 = 99.8\%$) while accuracy is increased with respect to the previous cities.

Table 41. Pisa: model evaluation metrics.

Gradient Boosting regressor (`n_estimators`: 100, `learning_rate`: 0.05, `max_features`: "sqrt")			
$R^2 = 0.998$	$R^2_{adj} = 0.994$	$RMSE = 87.0$	$SMAPE = 4.5\%$

The podium of the most relevant features is composed of three variables related to points of interest density, with `study_dens` in first place (Figure 41). In this case, too, the same considerations made before can be drawn. Students particularly appreciate the bike-sharing service, as many stations are located near major university campuses, providing students with a sustainable and economical mode of transport to reach their places of study, especially in cases where they do not own a vehicle. The second and third place are occupied by `needs_dens` (bars, cafes and restaurants near the stations) and `tourism_dens`, much relevant since the strong touristic character of the city, known mainly for *Campo dei Miracoli*, home of the famous Pisa Tower. Other similarities with Parma can be highlighted: distance from the city centre and `n_transport` are significantly less important than in other cities. Furthermore, `n_transport` in Pisa has the lowest feature importance of all the cities analysed. This may indicate a lack of transport infrastructure or, as mentioned for the city of Parma, that users prefer to use bike-sharing as a means of completing the entire journey rather than as a means of covering the first or last mile of the journey. Lastly, station lighting and the presence of coverage are of hardly any relevance in the study of demand variability.

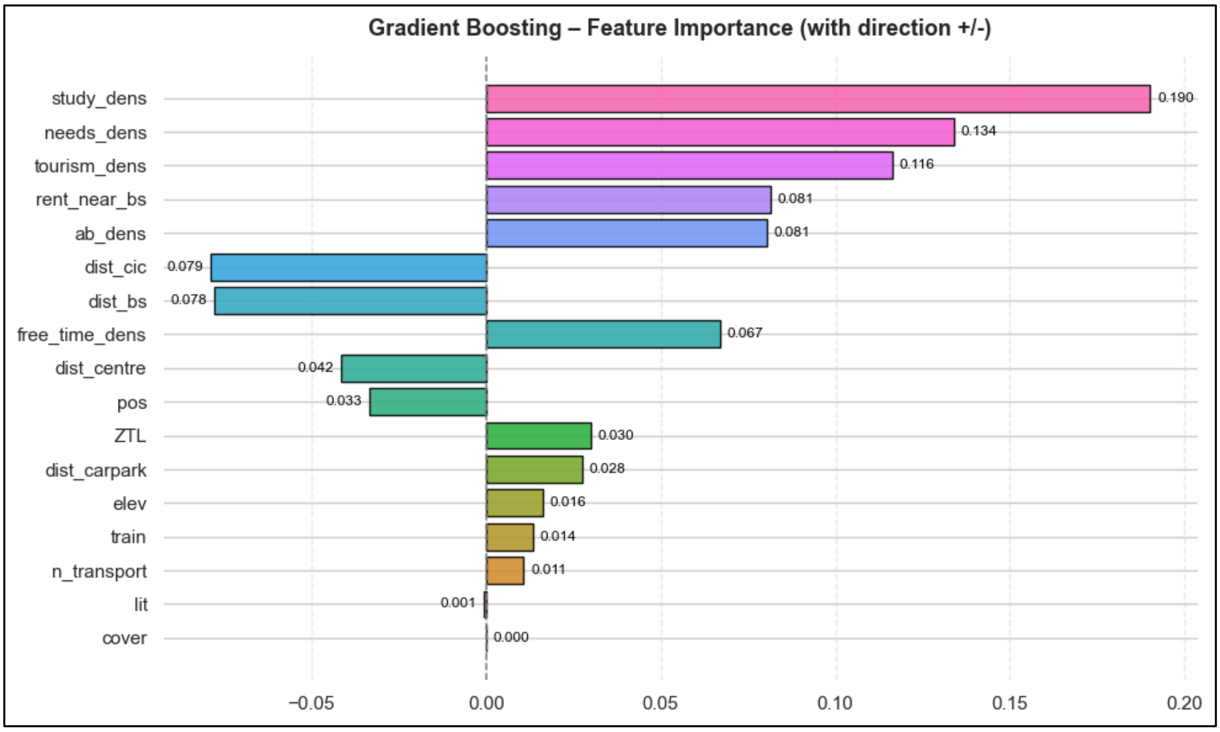


Figure 41. Pisa: graphical representation of Feature Importance.

4.5.7. Trieste

Trieste Bike-sharing service is managed by *Trieste Trasporti* and available via the *Weelo* app. It relies on more than 100 bicycles strategically distributed in 23 stations. All of them are included in the dataset. Stations can be categorised into two big clusters: one in the city centre and one on the seafront to the north of the city, where the stations are equally spaced so that users can cycle along the entire coastline from the city to Miramare, a popular tourist destination because of its famous castle. Users have two subscription options: they can subscribe to a daily pass (8€ for a maximum of six hours of daily trips or 17€ without any limit) or an annual pass at 9€ + 3€ for the annual. With the yearly pass, subscribers have 30 minutes of free travel for each rental.

Trieste bike-sharing stations are characterised by a constant percentage of outliers, the highest among the analysed cities (Table 42). Trieste has always been known as a multifaceted city, with diverse cultural influences converging there. Similarly, the urban context is remarkably heterogeneous, as reflected in the high percentage of anomalous data detected during the study of outliers.

Table 42. Trieste: outliers analysis.

Feature	Number of outliers	% of outliers	Winsorization (yes/no)
Ab_dens	6 (3 upper, 3 lower)	13.0% upper, 13.0% lower	Yes
Dist_centre	3	13.0%	Yes
N_transport	3	13.0%	Yes
Dist_cic	3	13.0%	Yes
Dist_bs	3	13.0%	Yes
Dist_carpark	3	13.0%	Yes
Study_dens	3	13.0%	Yes
Need_dens	3	13.0%	Yes
Tourism_dens	3	13.0%	Yes
Free_time_dens	3	13.0%	Yes

The machine learning model demonstrates a highly reliable capacity to capture the demand dynamics of Trieste's bike-sharing system. It shows excellent predictive performance, with an R^2 of 0.997 and an adjusted R^2 of 0.988, indicating that nearly all variance in station rentals is explained by the selected features. The RMSE value of 111.9 confirms a low prediction error relative to the scale of the target variable, while the SMAPE of 5.27% reflects strong accuracy and stability across observations.

Table 43. Trieste: model evaluation metrics.

Gradient Boosting regressor (`n_estimators`: 100, `learning_rate`: 0.05, `max_features`: "sqrt")			
$R^2 = 0.997$	$R^2_{adj} = 0.988$	$RMSE = 111.9$	$SMAPE = 5.27\%$

Distance from the city centre appears to be the most relevant explanatory variable (Figure 42). This is consistent with the dispositions of the station. In fact, the mentioned on the northern seafront are very far from the city centre (the furthest, *Miramare*, is 7 km from *P.zza Unità d'Italia*) and also characterised by poor densities of points of interest. As for general trends, other relevant features are densities of tourist, leisure and educational points of interest, while the presence of restaurants and cafes is of reduced significance compared to other cities. Variables such as distance from bike lanes, number of accessible transport lines, presence of railway station and station positioning have a moderate influence on demand variability. The altitude of individual stations also plays an important role and is negatively correlated with the number of rentals. This means that stations on the coast (practically at sea level) are more attractive than others (especially in the east, where the altitude increases). Housing density and bike-sharing station density have little influence in Trieste. Similarly, the presence of weather protection is still of almost no relevance.

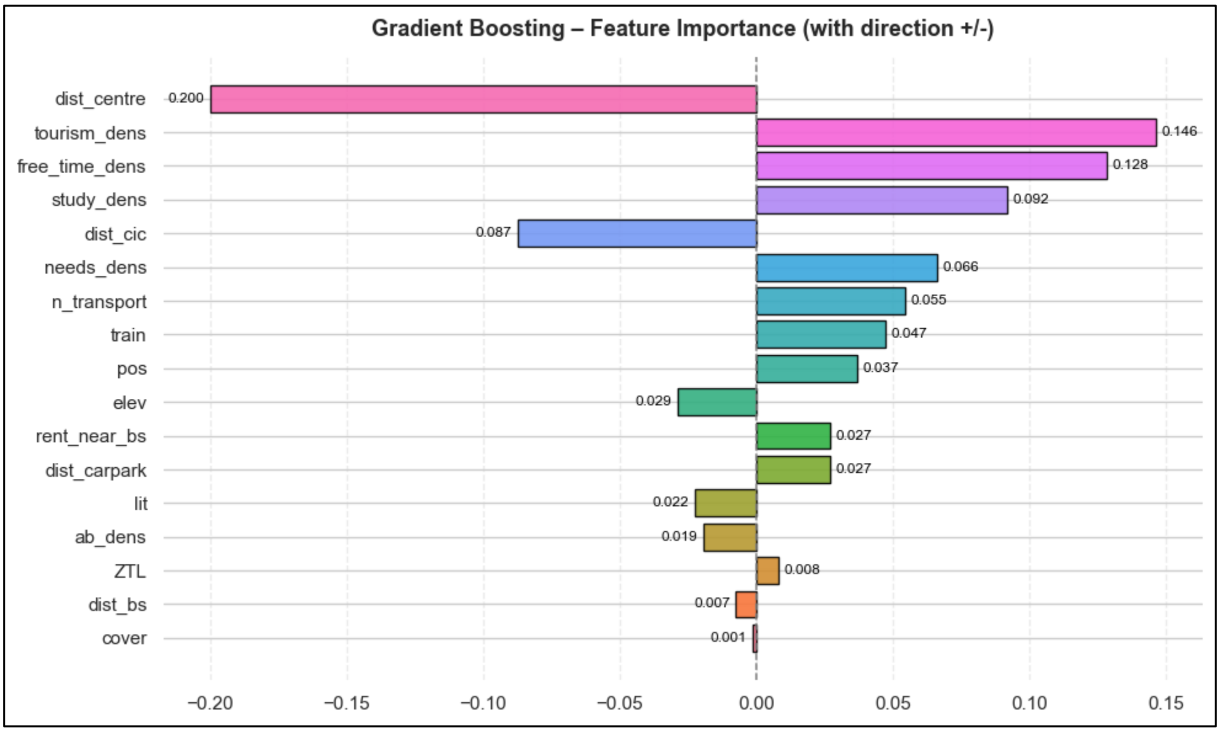


Figure 42. Trieste: graphical representation of Feature Importance.

4.6. Users' satisfaction analysis: empirical results

Applying the methodology proposed in Chapter 3.6, this section aims to analyse user satisfaction towards bike-sharing services in the dataset, studying which urban context characteristics have the greatest influence on users' feedback. Satisfaction grades for bike-sharing stations belong to two main families:

- Using the service's own website or application. This is only allowed for bike-sharing service subscribers. They can evaluate each station, giving it a grade between 1 and 5. Grades of this typology are collected in the feature `bic`, from the name of the service's website;
- Using Google Maps reviews. However, this is allowed for everyone, even those who have never used the service. The voting method is the same: voters can assign a score from 1 to 5 for each BS station. Grades from Google Maps are collected in the feature `Google`.

As suggested by equation 3.13, grades will be aggregated, giving more importance to variable `bic`, since reviews from those who have actually used the service are considered more reliable. A new variable `grade` for each station i will be computed as follows:

$$grade_i = 0.7 \cdot bic_i + 0.3 \cdot Google_i \qquad 4.5$$

Unfortunately, only a limited number of cities could be selected for the analysis of user reviews. In fact, the cities of Genoa and Trieste recorded insufficient reviews (less than 40% of stations had an adequate number of reviews), both from the service's website and from Google, so it was decided not to include them in the study. Therefore, the analysis presented in the following pages focuses on the cities of Brescia, Forlì, Parma and Pisa. This results in a dataset which includes 104 bike-sharing stations.

The in-depth analysis of user reviews begins with the exploratory study of the variables mentioned above, investigating descriptive statistics, distribution, and confidence intervals for the mean. Subsequently, a Gradient Boosting model will be applied. Then, Feature importances are computed, identifying which variables have the greatest influence on user satisfaction. Finally, the results obtained from this analysis will be compared with those obtained from the general analysis about demand drivers.

4.6.1. Reviews exploratory analysis

As expected, data mining shows that Google reviews are, on average, higher than those posted directly on the service's app. This is because Google reviews are mostly posted by users who do not regularly use the bike-sharing service. On the contrary, subscribers, who can post reviews on the company's app, tend to be more accurate in their ratings, having used the service several times and knowing its strengths and weaknesses.

Box plots and KDE plots (Figure 43 and Figure 44) confirm this trend:

- Google reviews are mostly concentrated around the average, with reduced dispersion. Their distribution is strongly bimodal, suggesting the presence of two significant clusters, the first relating to satisfied users (around a value equal to 3) and the second relating to users who are very satisfied with the service (around a value equal to 4.5).
- Bic reviews are characterised by more spread values, ranging from a minimum of 0.2 to a maximum of 4.40, highlighting great variability for this variable. The average value (2.56) is sensibly lower than Google reviews, while the distribution is basically unimodal, with a greater accumulation of values at the left tail of the curve.

The aggregation of the two variables creates a new feature with a greatly reduced dispersion of values compared to the two previous cases. Consistent with the calculation equation, the average (equal to 2.96) appears closer to that of `bic` than to that of `google`. The dispersion is highly unimodal, with almost perfect symmetry between the left and right tails. There is only one outlier in the left tail of the curve. This number is not high enough to justify the elimination or winsorisation of the data.

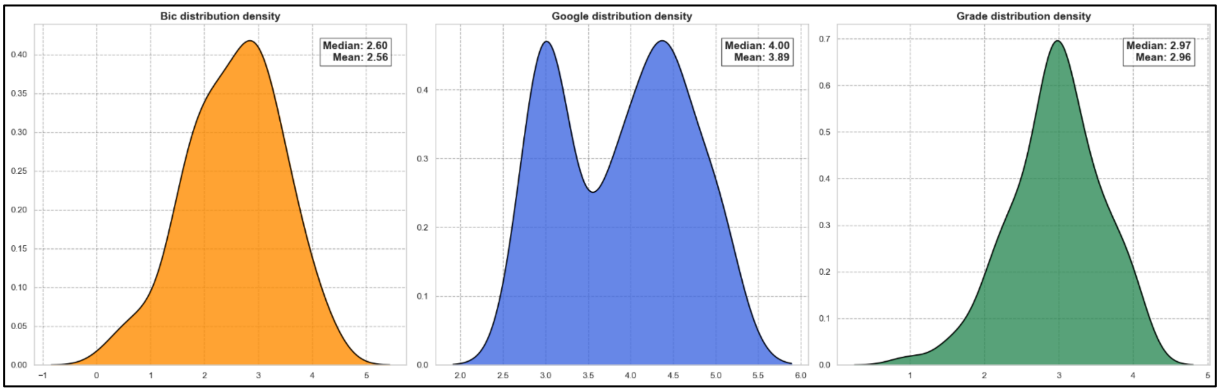


Figure 43. Users' reviews KDE plots.

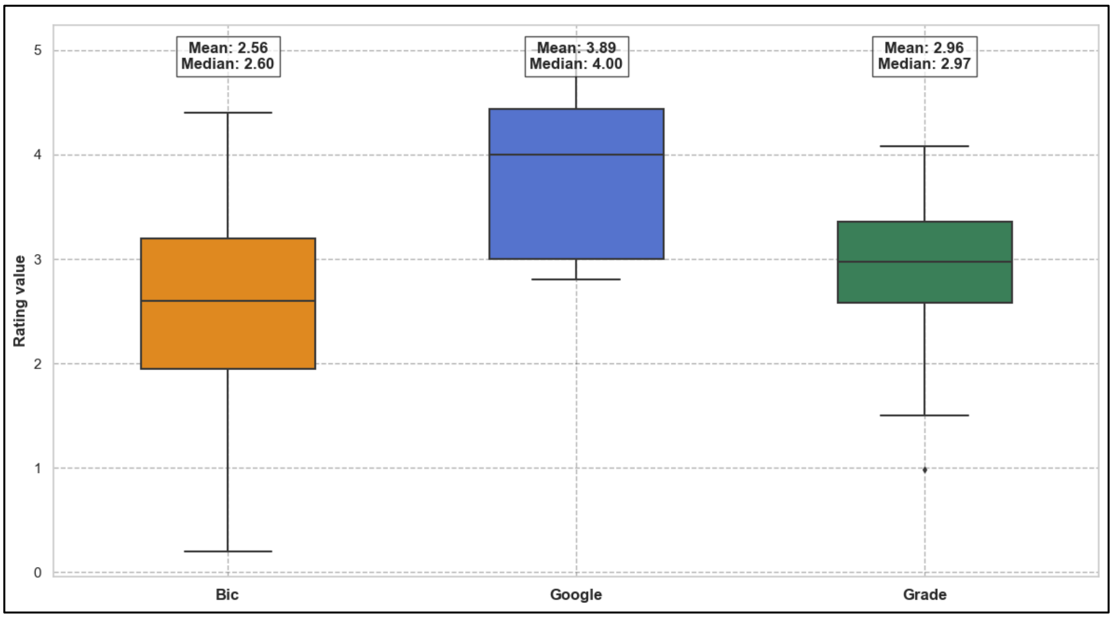


Figure 44. Users' reviews boxplots.

Confidence intervals (Figure 45) further reinforce the differences observed between the two rating sources. In all cities, Google reviews systematically show higher mean values than `bic` reviews, highlighting a generally more positive perception among occasional or external users. Brescia is the city with the largest gap between the two means, suggesting a particularly strong divergence between internal users—typically more critical—and the broader public. The only partial exception is Parma, where the two confidence intervals lie closer together, indicating a more aligned perception between the two user groups. This suggests that Parma's bike-sharing service may be comparatively more appreciated by regular users, who tend to rate the service on Bic, than by occasional users who post reviews on Google. Overall, the analysis confirms that differences in user type and usage frequency strongly shape the evaluation patterns across cities.

Regarding the aggregated variable `grade`, Brescia is the city where the bike-sharing service is most appreciated (Figure 46). So, the service appears to be very effective, explaining the high utilisation rate mentioned in previous sections. Forlì and Pisa are characterised by the lowest scores (2.87 and 2.88) and, at the same time, by the greatest dispersions. This could be due to two causes: either the presence of significant disparity among user reviews, as not all users share the same views, or the presence of two clusters of stations, the former with high ratings and the latter less appreciated by users. Parma, on the other hand, ranks second in terms of average values, while it is characterised by the lowest data dispersion.

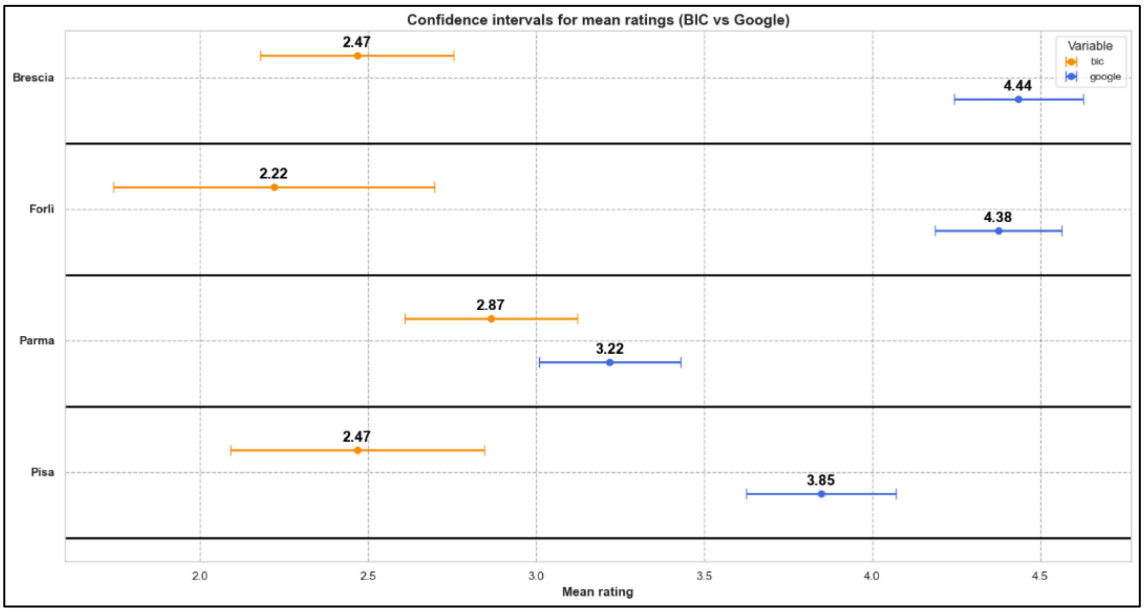


Figure 45. Confidence intervals for the average review score by city.

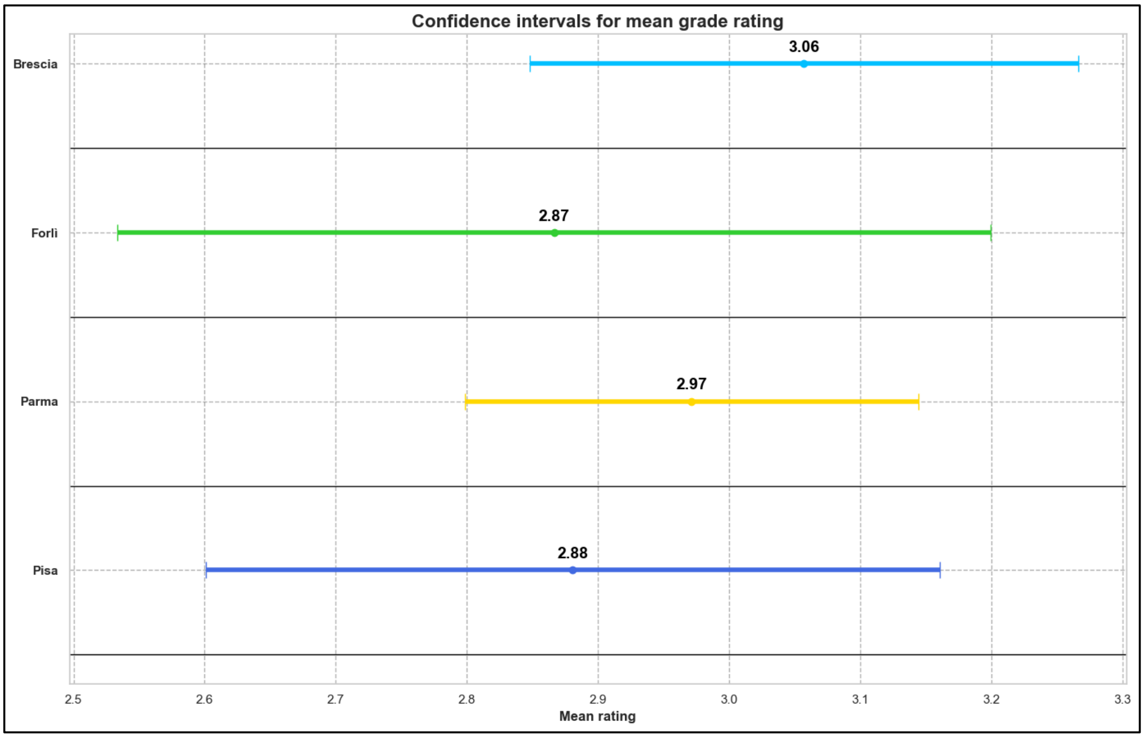


Figure 46. Confidence intervals for aggregated users' reviews.

4.6.2. Users’ reviews model

The Gradient Boosting algorithm is now applied to study users’ reviews and their variability. This method has been chosen because it has already proven its robustness and effectiveness in data analysis applications. The specifications are the same as the general case, with `n_estimators` equal to 200 and `learning_rate` equal to 0.05. In this case, the target variable changes: the variable `grade` is chosen as the model target. Unlike the demand model, it is not necessary to normalise the target variable because the scale (1-5) is the same for all the considered cities. On the other hand, variables belonging to different scales, such as distance from the city centre and station elevation, must be transformed to make them comparable. The normalisation method used here is the same as that explained in paragraph 3.3.1.2: variables will be rescaled with output values between 0 and 1.

Model application reveals good performance (Table 44) in terms of explaining target variability, resulting in high values for R^2 and R^2_{adj}. At the same time, error metrics are good, with low values in prediction errors, both in absolute value and percentage. This once again confirms the robustness of the iterative method of gradient boosting.

Table 44. Users' reviews model: evaluation metrics.

Gradient Boosting regressor (`n_estimators`: 200, `learning_rate`: 0.05, `max_features`: "sqrt")			
$R^2 = 0.963$	$R^2_{adj} = 0.957$	$RMSE = 0.109$	$SMAPE = 3.23\%$

Computation of Feature Importances (Figure 47) shows deep deviations from the model on rentals and many features, which are now negatively correlated with the target variables.

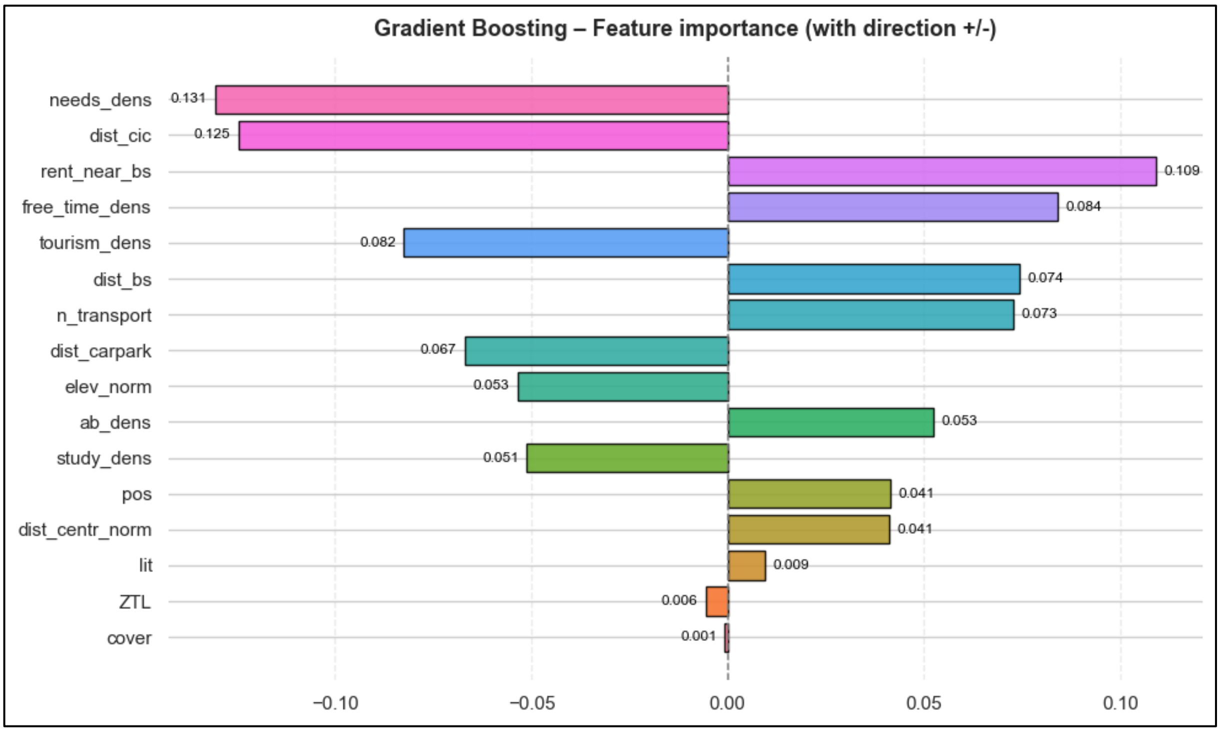


Figure 47. Users' reviews model: graphical representation of Feature Importance.

Unlike the previous models, `needs_dens`, `tourism_dens`, `study_dens` have now a different interpretation: the presence of these kinds of points of interest has a negative influence on user reviews, meaning that they're not relevant for users' satisfaction. On the other hand, the most important explanatory variables with traditional correlation are:

1. Distance from the nearest bike lane;
2. Proximity to other popular bike-sharing stations;
3. Presence of free time attractions;
4. Density of bike-sharing stations in the surroundings;
5. Number of overall accessible public transport lines.

Out of a total of five variables, four are related to integration with the city's transport network. The conclusion of the research indicates that users appreciate stations that offer a greater level of integration with the transport network, both individual (other bike-sharing stations or cycle paths) and collective (accessible transport lines). Similarly, proximity to points of interest for leisure activities, such as fitness centres or sports facilities, cinemas and theatres, is also important. Users particularly appreciate stations that offer easy access to these types of attractions. Once again, the explanatory variables relating to the physical characteristics of the station appear to have the least influence on the target variable.

Table 45 presents a comparative analysis of the key factors identified by Gradient Boosting models in relation to demand and user satisfaction. It emphasizes that these two dimensions are influenced by distinct mechanisms. For bike rentals, the primary determinants are closely associated with the presence of nearby points of interest—such as those related to tourism, personal necessities, leisure activities, and study environments— indicating that demand is significantly driven by the accessible activities from the station. Proximity to the city centre and multiple public transport options also play crucial roles in augmenting rental activity. On the other hand, examining user reviews, the hierarchy of influential factors alters. Elements such as the density of amenities and the distance from cycling paths seem to negatively affect satisfaction. Conversely, the presence of leisure facilities nearby and proximity to popular other BS stations tend to improve perceived quality. This likely suggests that enhanced usage conditions and a more inviting urban environment contribute positively to user satisfaction. Notably, certain factors that promote demand—such as tourism density—may exhibit a weaker or even negative correlation with satisfaction.

This reinforces the idea that high usage does not necessarily equate to positive user experiences.

As outlined in the methodology chapter, the discrepancy between the two models might be partially attributed to the fact that user satisfaction is influenced not only by the fixed characteristics of the station's urban setting but also by the operational features of the service, such as the number of available bicycles and docks. Therefore, while demand models primarily capture the main factors influencing bike-sharing usage, satisfaction models offer additional insights into user perceptions of service quality at each station. This comparison underscores the necessity of evaluating bike-sharing systems from multiple perspectives rather than presuming that only one of these elements suffices to define overall service success.

Table 45. Users' reviews model: main features comparison with the rentals model.

Top ranked features by Feature importance	
Target: rentals	**Target: users' reviews**
1. Tourism_dens (+)	1. Needs_dens (-)
2. Needs_dens (+)	2. Dist_cic (-)
3. Free_time_dens (+)	3. Rent_near_bs (+)
4. Dist_centr_norm (-)	4. Free_time_dens (+)
5. Study_dens (+)	5. Tourism_dens (-)
6. N_transport (+)	6. Dist_bs (+)
7. Dist_cic (-)	7. N_transport (+)
8. Dist_carpark (+)	8. Dist_carpark (-)
9. Dist_bs (-)	9. Elev_norm (-)
10. Ab_dens (+)	10. Ab_dens (+)

5 SSEI: the Italian case-study

Once the mathematical formulation of the SSEI has been defined, it is necessary to propose applications for the brand-new index in order to verify its accuracy, robustness and applicability in different contexts. This section proposes two possible applications of the *Station Static Effectiveness Index*, both related to Italian bike-sharing services. The first involves calculating the index for the dataset used for the previous data analysis. In this case, the SSEI will be used on its own to evaluate the performance of the service. In contrast, in the second application, the SSEI will be used in parallel with the *Bicycle Compatibility Index* (BCI), which indicates the level of service of the cycling infrastructure. This will make it possible to analyse the synergies between service and infrastructure, as well as identify priorities for improvement in these areas.

5.1. SSEI stand-alone evaluation of service effectiveness

The first application concerns the calculation of the index for the 140 stations that compose the dataset, simulating a performance evaluation of existing bike-sharing services. In the following pages, the SSEI will be calculated for each of these stations; first, an aggregate analysis will be performed, followed by a disaggregated analysis for each of the cities present.

5.1.1. Aggregate analysis

SSEI computation on the original dataset demonstrates once again that stations characterised by the highest quality level are the ones near railway stations or in the historical city centre. In fact, the five stations with the highest SSEI values (Table 46) are all located in the city centre (*De Ferrari* in Genoa, *Borgo Stretto* in Pisa) or in the proximity of a railway station (*Brignole* in Genoa, *Vittorio Emanuele* in Pisa, *Stazione FS* in Parma). In contrast, stations with the lowest SSEI values are mostly located far from the city centre, in scarcely populated areas, which are more likely to be the destination of the journey rather than its origin (such as *Miramare* in Trieste, *Ospedale* in Forlì or *Aeroporto* in Pisa in Table 46).

Table 46. Stations with the highest and lowest SSEI.

Highest			Lowest		
City	**Station**	**SSEI**	**City**	**Station**	**SSEI**
Genoa	*Brignole*	97.0	Forlì	*Ospedale*	2.2
Pisa	*Vittorio Emanuele*	89.6	Pisa	*SMS Biblioteca*	3.6
Parma	*Stazione FS*	89.5	Trieste	*Cumano-Musei*	3.6
Genoa	*De Ferrari*	88.6	Trieste	*Miramare*	5.4
Pisa	*Borgo Stretto*	87.9	Pisa	*Aeroporto*	6.0

By categorising the stations into classes based on the SSEI value (see Chapter 4.4.2), it is observed that they are distributed homogeneously across the classes, with no one class standing out above the others (Table 47 and Figure 48). This confirms the correctness of the mathematical formulation of the index and, in particular, of the calibration of the k parameter, which has been designed to optimise the distribution of output values and equalise the probability of belonging to the various classes. This trend is once again confirmed by Figures 50 and 51, in which each circle represents a station in the dataset. The first represents stations divided by SSEI value, the second by city. The division by city also shows no city that is significantly better than the others in terms of SSEI values. In line with expectations, stations effectively follow the calibrated logistic curve, never exceeding the lower and upper limits. Stations that deviate from the logistic function trend are those characterised by a railway station in the buffer. In fact, the presence of the related dummy variables causes the circles to deviate from the expected trend of the logistic function with $k = 0.3$.

Table 47. Stations classification based on SSEI.

Class	Number of stations
Excellent (80-100)	18
Good (60-80)	36
Medium (40-60)	20
Low (20-40)	31
Critical (0-20)	35

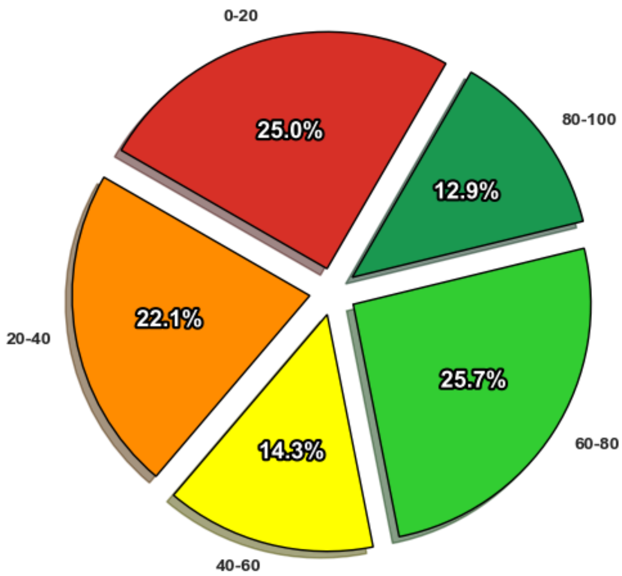


Figure 48. Stations classification based on SSEI pie plot.

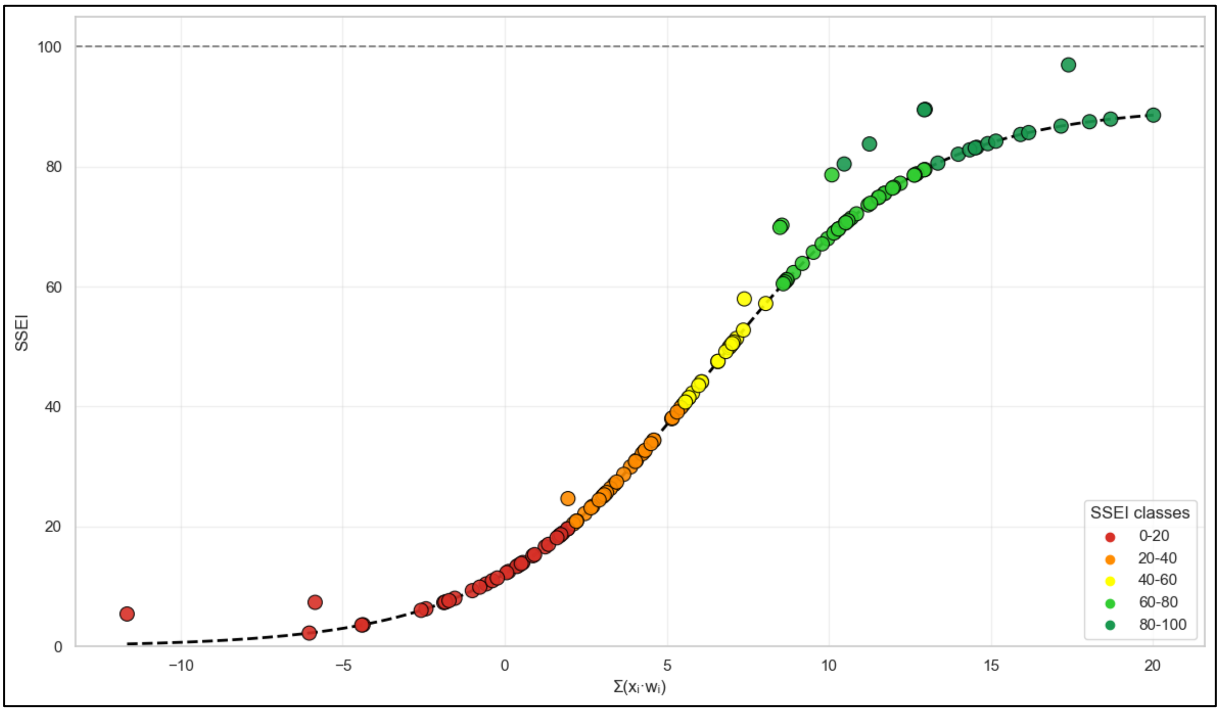


Figure 49. SSEI logistic function. Classification by classes.

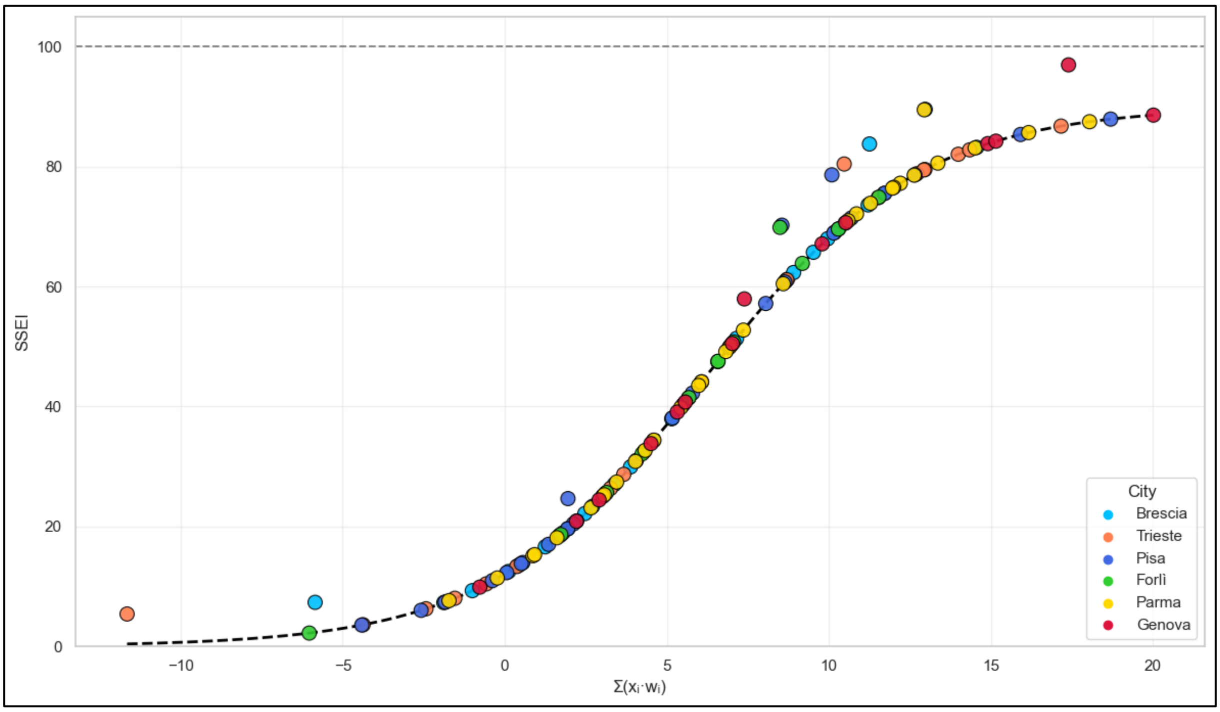


Figure 50. SSEI logistic function. Classification by cities.

Moreover, SSEI values show a moderate correlation with the number of rentals per station ($r = 0.50$, Figure 51). As expected, the stations with the highest number of rentals belong to the best SSEI classes. However, the inverse relationship is not always true, as there are some stations with high SSEI values but low rental numbers. At the same time, there are also stations with a moderate number of rentals belonging to lowest SSEI classes. This behaviour confirms the validity of the mathematical formulation of the SSEI, which not only explains the effectiveness of a bike-sharing station in relation to the total number of rentals, but also rewards good integration with the context in which it is located, giving significant importance to the citizens' perceptions (through the survey reported in Appendix B).

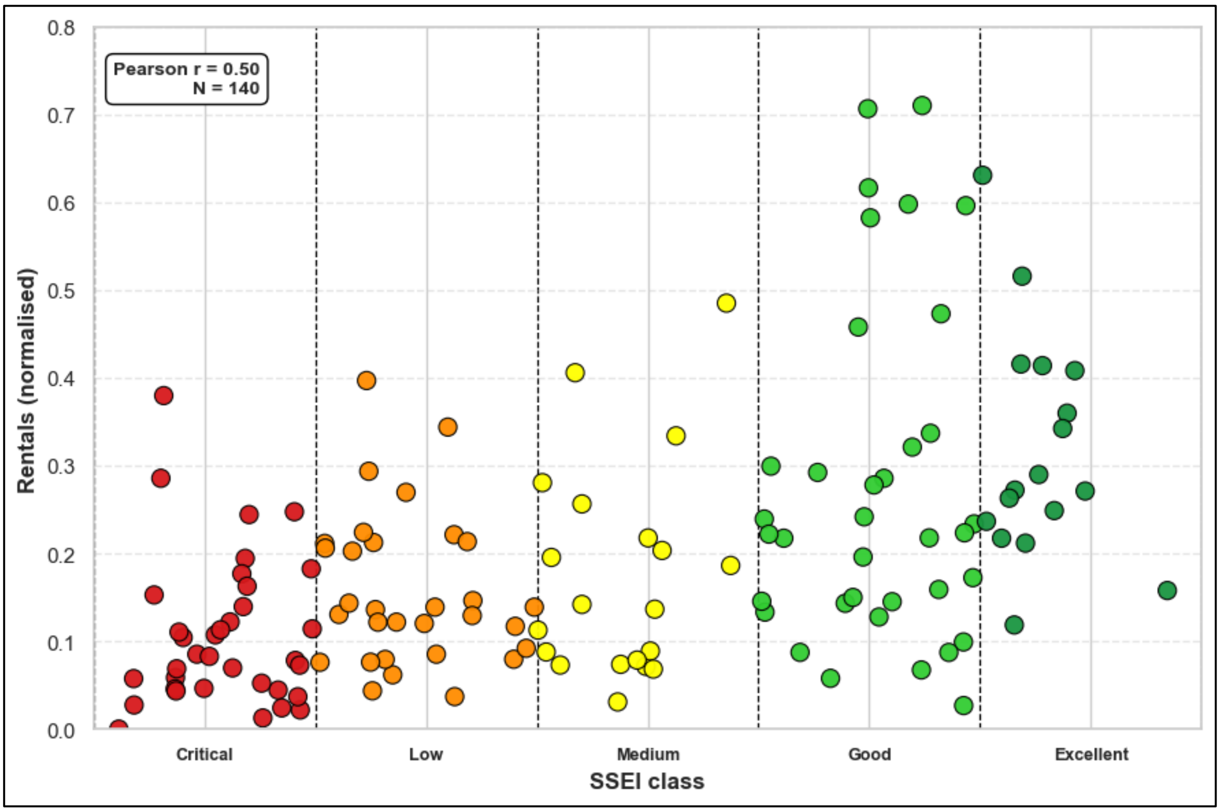


Figure 51. SSEI - rentals scatterplot.

Lastly, the correlation between SSEI and its constitutive variables is analysed (Table 48). Results appear consistent with the mathematical formulation adopted and with the considerations presented above.

Table 48. Correlation coefficients between SSEI and its explanatory variables.

Feature	Pearson coefficient
`Needs_dens`	0.78
`Tourism_dens`	0.73
`Study_dens`	0.64
`N_transport`	0.54
`Free_time_dens`	0.51
`Dist_carpark`	0.34
`Ab_dens`	0.34
`Train`	0.20
`Dist_cic`	-0.18
`Dist_bs`	-0.38
`Dist_centre`	-0.67

In particular, there is a strong negative correlation with the variable about the distance from the city centre, which in turn is moderately correlated with the presence of different types of points of interest. However, a quick calculation of the SSEI without including the variable relating to distance from the centre shows that the index values vary by an average of 5.45%, a percentage that is not significantly high enough to justify removing this variable from the equation. In any case, the chosen formulation allows the user to consider only variables of their choice or for which data can be easily retrieved. Therefore, especially when calculating the SSEI for groups of stations where the `dist_centre` variable has little influence (e.g. stations that are all very far from the city centre), it may be decided not to include it in the model. As a final note, it should be pointed out that this method can be applied to any of the other variables included in the SSEI.

5.1.2. Disaggregate analysis

Computing the SSEI separately for each city makes it possible to uncover previously unseen patterns and to obtain a comparative assessment of bike-sharing system performance across the selected urban areas. Confidence intervals for the average values of SSEI have been computed (Figure 52). The results show that Genoa is the city with the highest SSEI values, with an average score of 54.9. This outcome can be attributed to two main factors:

- Firstly, Genoa is the biggest city in the dataset. Thus, it covers a larger area. The natural consequence of this is that the city centre is larger than in other cities, increasing the density of points of interest, even for bike-sharing stations located far from the point chosen as the city centre (*Piazza De Ferrari*).

- Secondly, the service is less capillary than in other cities. Planners decided to build a few stations in key locations around the city rather than many to cover a larger area.

The second-best city in terms of average SSEI is Parma, while the remaining four cities exhibit very similar average values, ranging from 41.0 in Brescia to 43.8 in Forlì.

Regarding the variability of station performance, Trieste shows the highest dispersion, with a spread of 31.9 SSEI points; this indicates a strong heterogeneity in station quality across the city, likely reflecting its complex topography and uneven urban structure. Conversely, Forlì displays the lowest dispersion (22.0 points), suggesting a more homogeneous network where stations tend to perform similarly, possibly due to a more uniform urban layout and a consistent distribution of facilities.

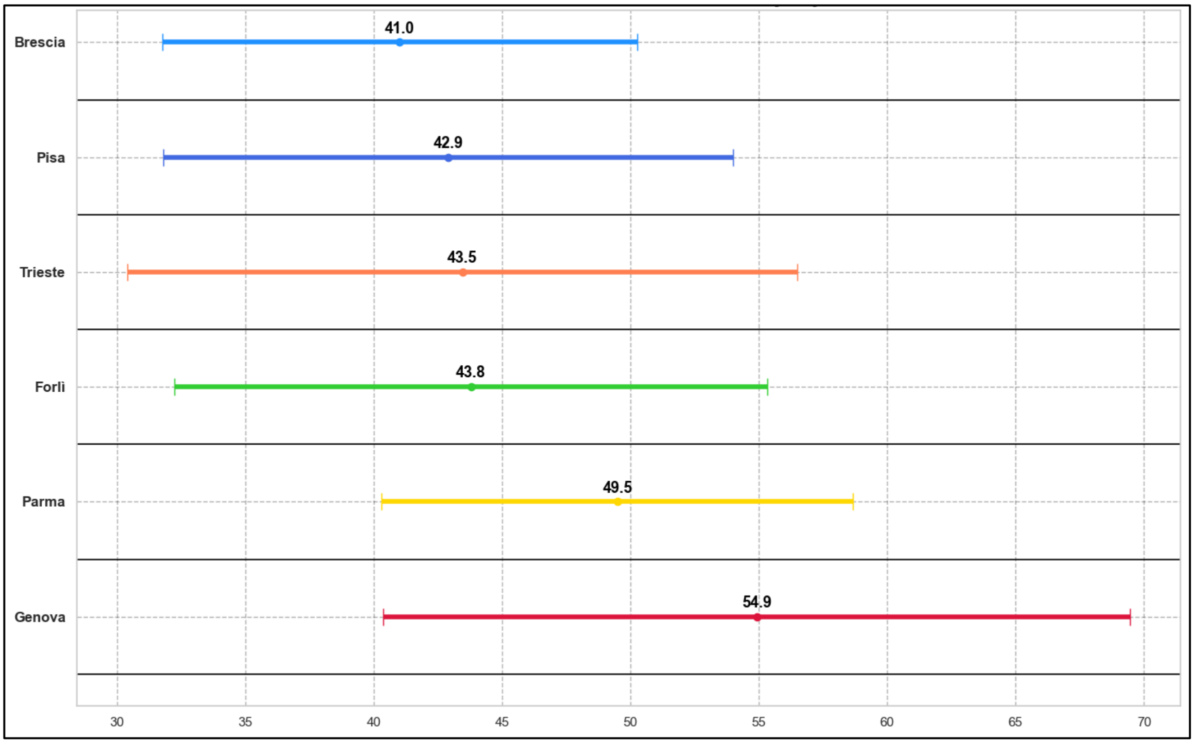


Figure 52. Confidence intervals for mean SSEI by city.

Brescia

Table 49. Brescia: stations with the highest and lowest SSEI.

Highest		Lowest	
Station	**SSEI**	**Station**	**SSEI**
Stazione FS	83.8	*Wuhrer*	7.3
Zanardelli	83.2	*Villa Glori*	7.4
Paolo VI	75.5	*1° Maggio*	9.3
Vittoria	73.6	*Via scuole*	12.5
Rovetta	70.6	*Europa*	13.4

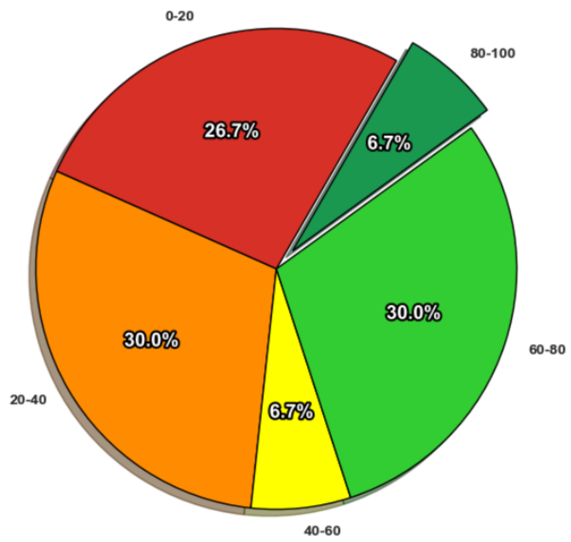


Figure 53. Brescia: Stations SSEI classification by classes.

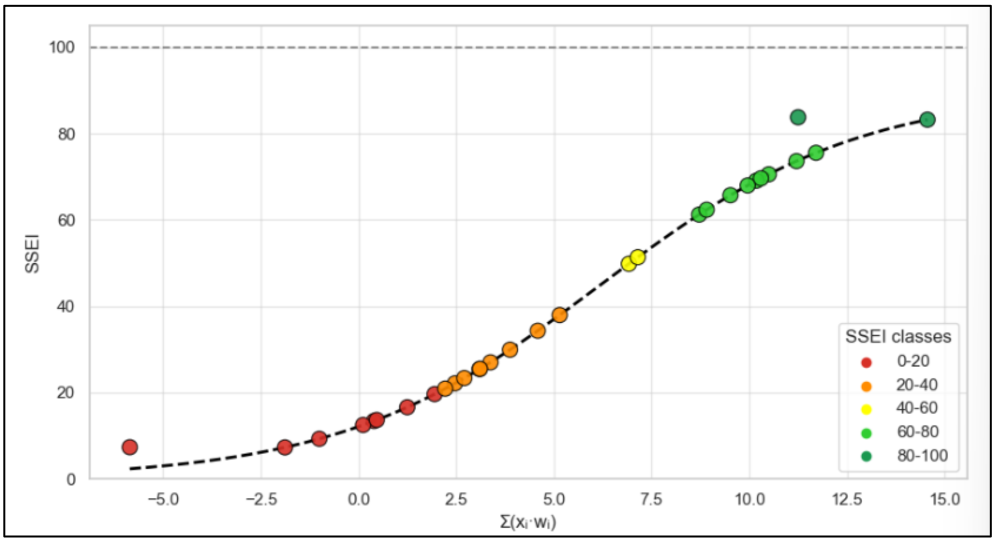


Figure 54. Brescia: SSEI logistic function.

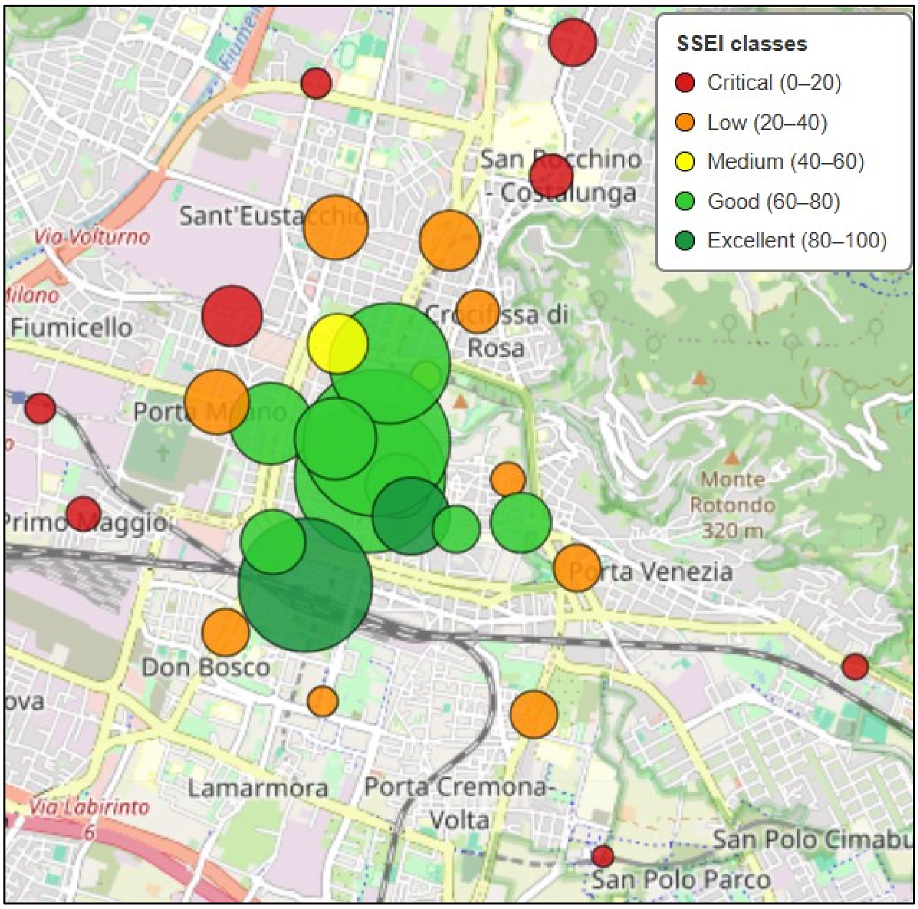


Figure 55. Brescia SSEI map. Radius of the circles is proportional to number of rentals.

The stations with the highest SSEI are the ones with the most rentals. Except for the one near the main railway station, the others are all in the historic city centre. Among these is *Zanardelli* station, located in a central area with lots of shopping and leisure spots, as well as restaurants and cafes. Low SSEI values are recorded far from the city centre, even if the stations serve stand-alone neighbourhoods or are close to underground stations (e.g. *Europa* station). There are stations which, although located near the city centre and characterised by moderate demand, record poor SSEI values. These stations should be taken into consideration when evaluating the current service. The Brescia bike sharing service will be the focus of the second SSEI application, presented in Chapter 5.2, which features the stations of this service located near underground stations.

Genoa

Table 50. Genoa: stations with the highest and lowest SSEI.

Highest		Lowest	
Station	**SSEI**	**Station**	**SSEI**
Brignole	96.8	*Fiumara*	9.9
De Ferrari	88.6	*San Benigno*	20.9
Zecca	84.3	*Matitone*	24.4
Caricamento	83.8	*Di Negro*	33.8
Darsena	70.7	*Questura*	39.1

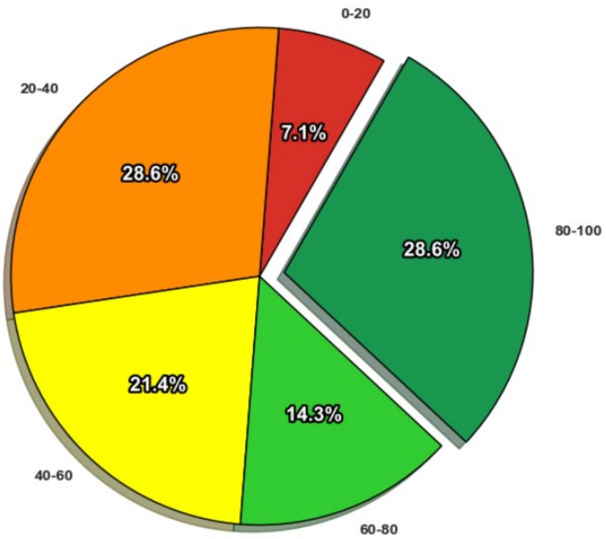


Figure 56. Genoa: SSEI classification by classes.

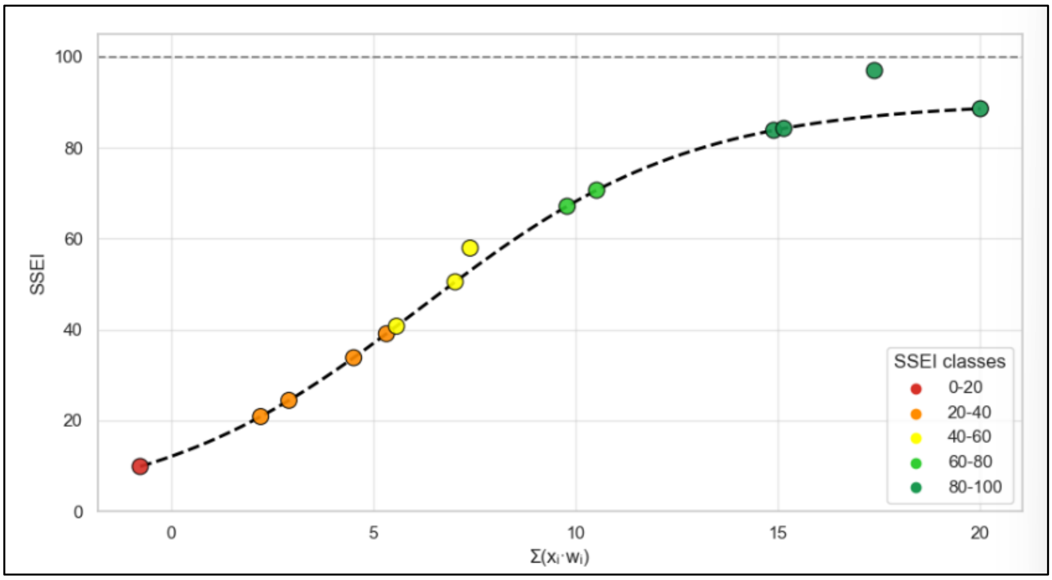


Figure 57. Genoa: SSEI logistic function.

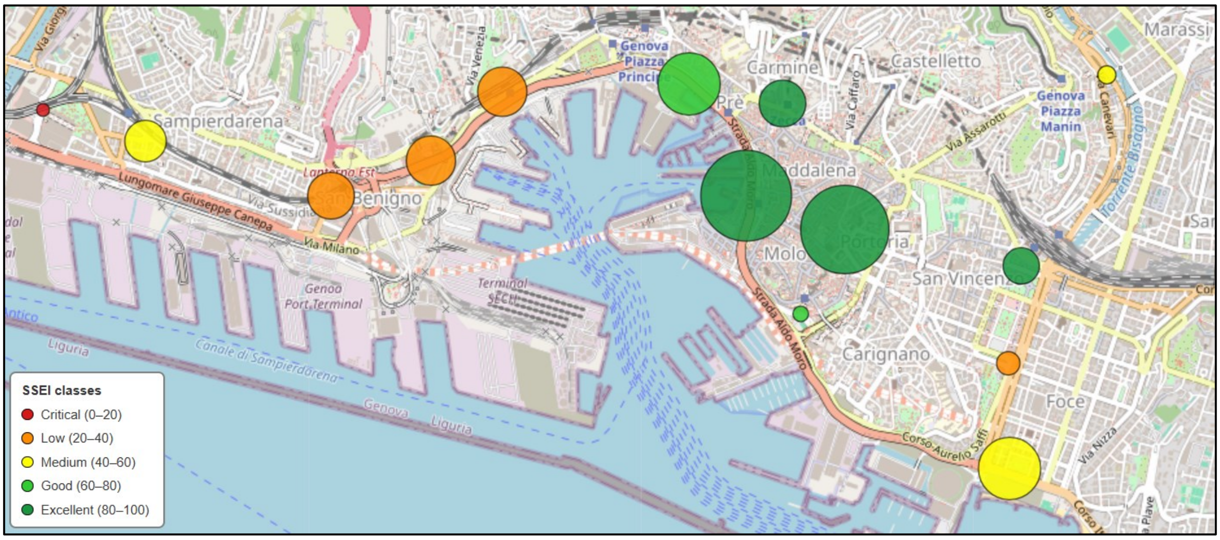


Figure 58. Genoa SSEI map. Radius of the circles is proportional to the number of rentals.

Brignole station, which has the highest SSEI, is not the one with the highest number of rentals. In fact, the most demanded stations are *Caricamento* and *De Ferrari*, located in the historic and tourist centre of the city, and both belong to the best SSEI category. *Fiumara*, in the western part of the city, is the bike-sharing station with both the lowest SSEI value and the lowest number of rentals. Despite its proximity to a major shopping centre, this is probably due to low housing density and its distance from the primary needs of the city centre. The situation is similar for the station serving the stadium in the north-east of the city, which is characterised by low demand and belongs to SSEI category 20-40. This is because, apart from the aforementioned stadium, there are no other significant points of interest in the surrounding area.

Forlì

Table 51. Forlì: stations with the highest and lowest SSEI.

Highest		Lowest	
Station	SSEI	Station	SSEI
Piazza Saffi	74.9	*Ospedale*	2.2
Stazione FS	69.9	*Park dell'argine*	18.4
Vittoria	69.6	*Park Schiavonia*	18.7
Park Oriani	63.9	*Campus Universitario II*	25.7
XX Settembre	50.8	*Vittorio Veneto*	32.1

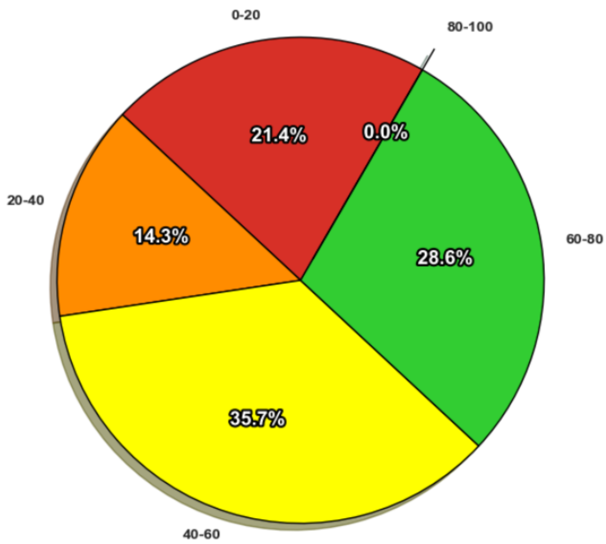


Figure 59. Forlì. SSEI classification by classes.

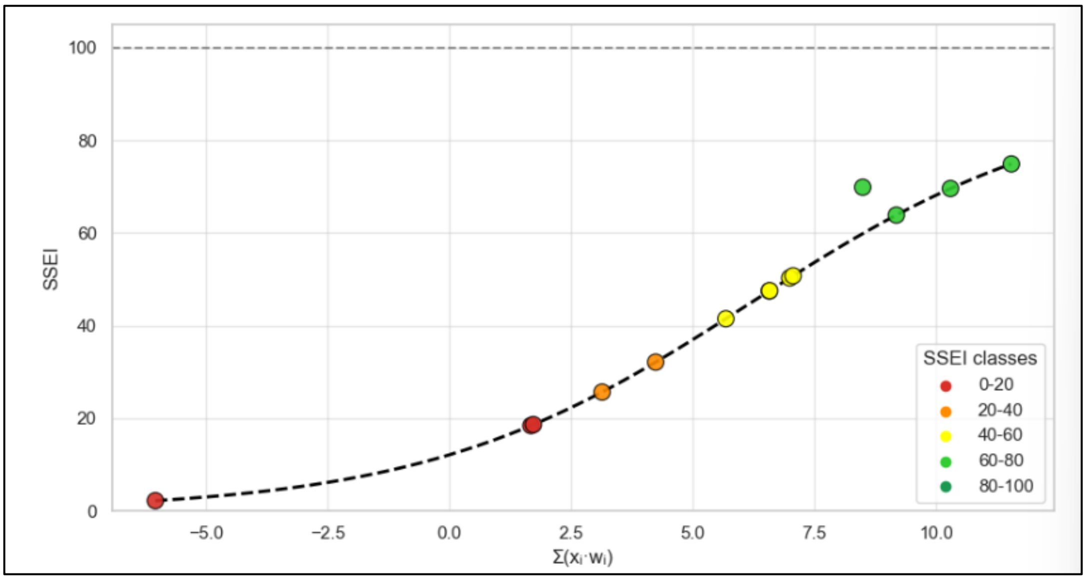


Figure 60. Forlì. SSEI logistic function.

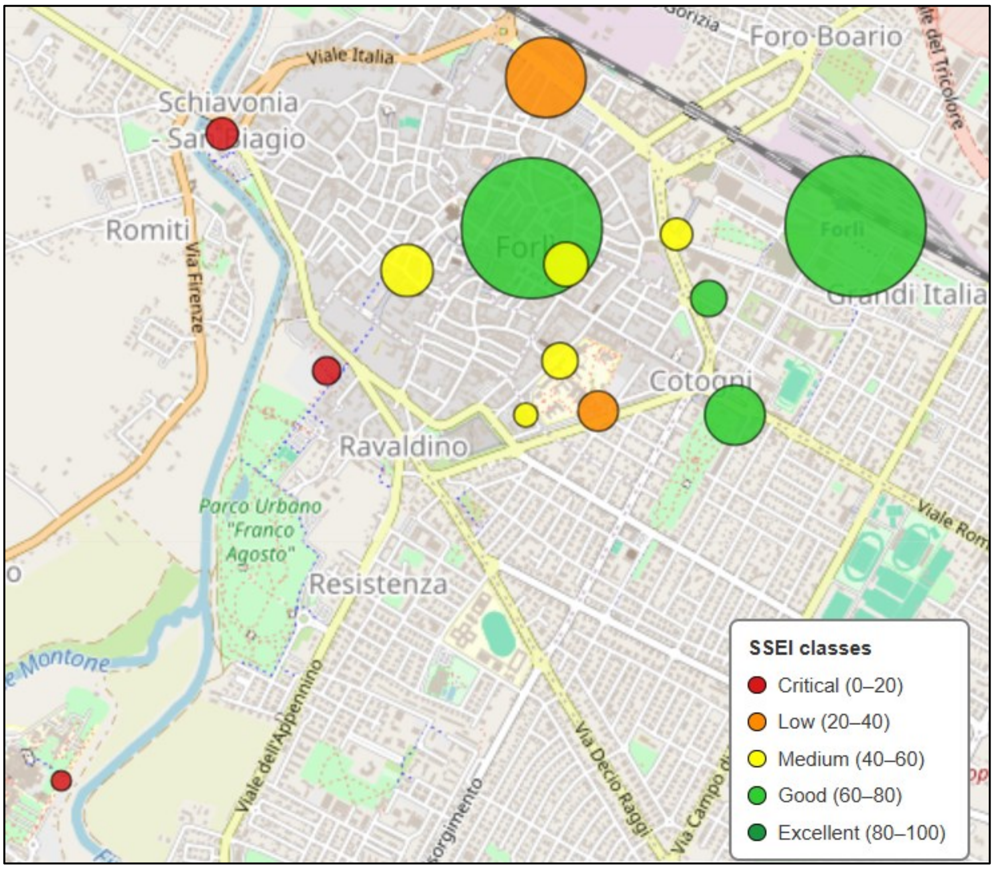


Figure 61. Forlì SSEI map. Radius of the circles is proportional to the number of rentals.

Forlì is the only city that does not have a station in the highest SSEI class. The stations with the highest SSEI are those related to the main railway station and the central square. Forlì is also the city with the worst station in terms of SSEI. This is the *Ospedale* station, which has only 21 rentals per year and a SSEI value of 2.2, probably because, in this location, it can only serve the hospital itself, being completely isolated from the city and the rest of the bike-sharing stations. The average SSEI value in the city of Forlì is lower than in the other cities, as it is the city with the lowest density of points of interest, as well as the smallest in surface area. However, the map shows the robustness of the proposed index, as there are stations with good SSEI values even far from the centre and stations with poor SSEI values even close to the city centre.

Parma

Table 52. Parma: stations with the highest and lowest SSEI.

Highest		Lowest	
Station	**SSEI**	**Station**	**SSEI**
Stazione FS	89.5	*Ospedale Volturno*	7.7
Kennedy	87.5	*Villetta*	11.4
Garibaldi	85.7	*Lauro Grossi*	15.1
Barezzi	83.1	*Passo della Cisa*	15.3
Ponte di mezzo	80.6	*Palasport*	18.1

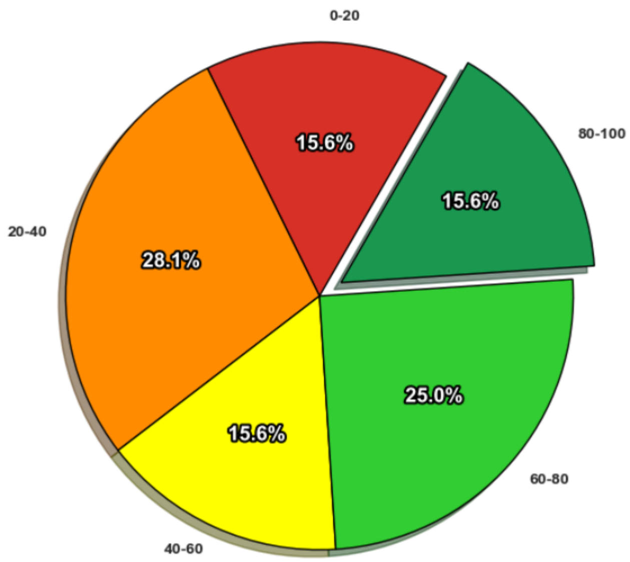


Figure 62. Parma: SSEI classification by classes.

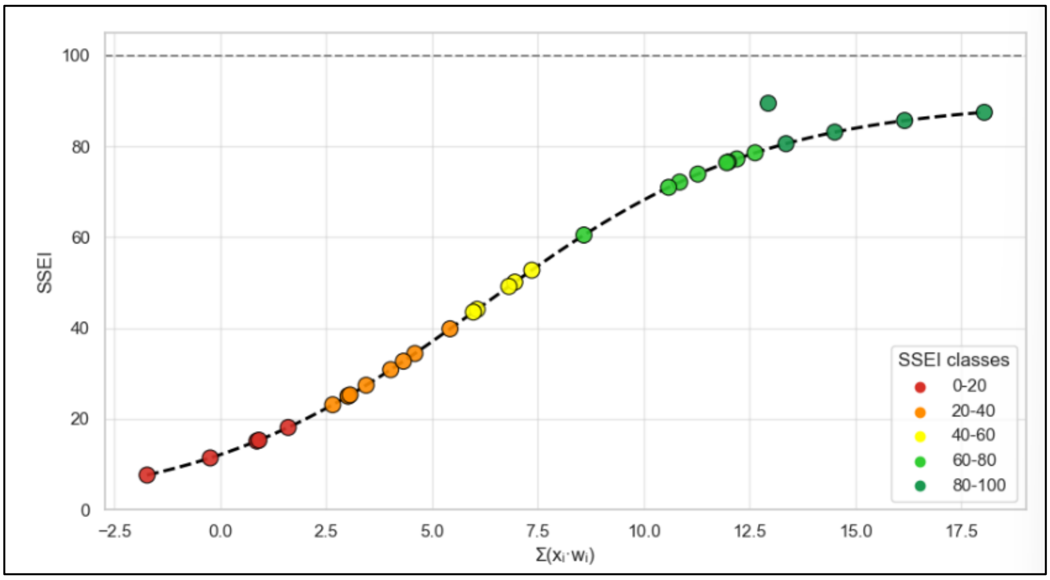


Figure 63. Parma: SSEI logistic function.

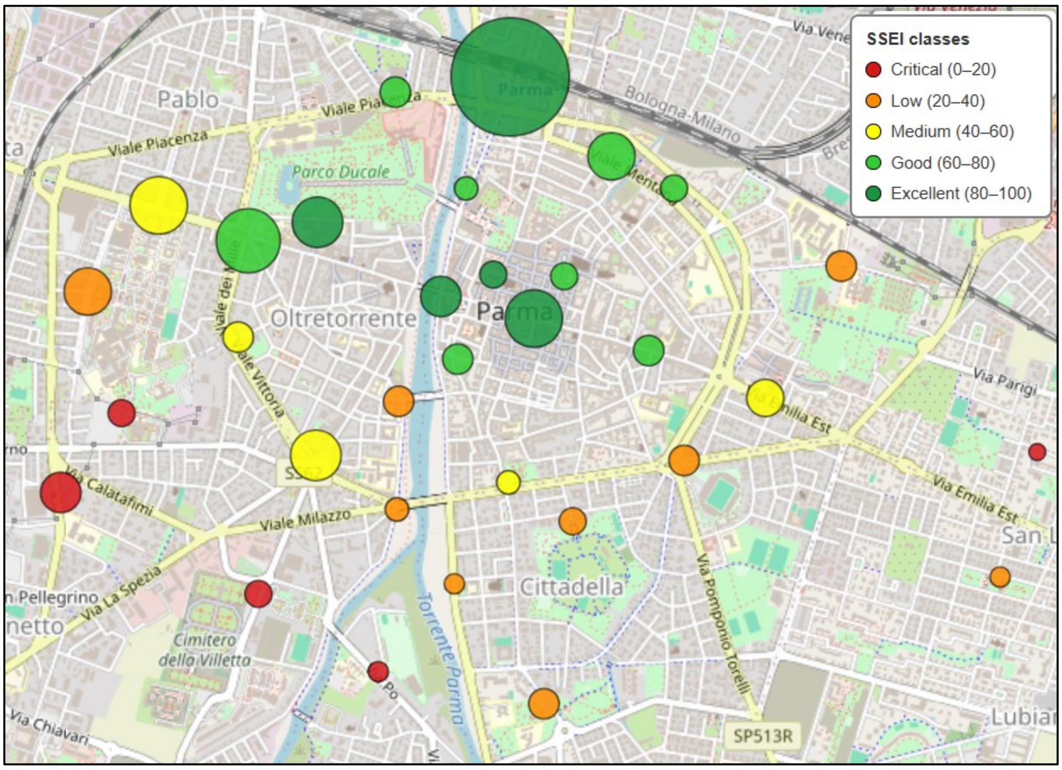


Figure 64. Parma SSEI map. Radius of the circles is proportional to the number of rentals.

The best SSEI stations are all concentrated in the northern part of the city, an area with high demand for bike-sharing and the site of the railway station and the historic, cultural and tourist centre. It is possible to identify a triangle that encompasses most of the best stations, whose vertices are the city centre, the railway station and the university campus located near *Parco Ducale*. Conversely, there are also five stations which fall in the SSEI critical class (values between 0 and 20). Four of them are located in the south-west, while *Passo della Cisa* station is the only one located in the eastern part of the map. There are no particular differences between the SSEI values in the base case and when the distance from the city centre is not taken into account, as in the general case.

Pisa

Table 53. Pisa: stations with the highest and lowest SSEI.

Highest		Lowest	
Station	**SSEI**	**Station**	**SSEI**
Vittorio Emanuele	89.6	*SMS biblioteca*	3.6
Borgo Stretto	87.9	*Aeroporto*	6.0
Martiri della libertà	85.4	*La fontina*	7.4
Duomo	78.6	*Galleria Gerace*	11.0
Sesta Porta	78.6	*USL Garibaldi*	12.3

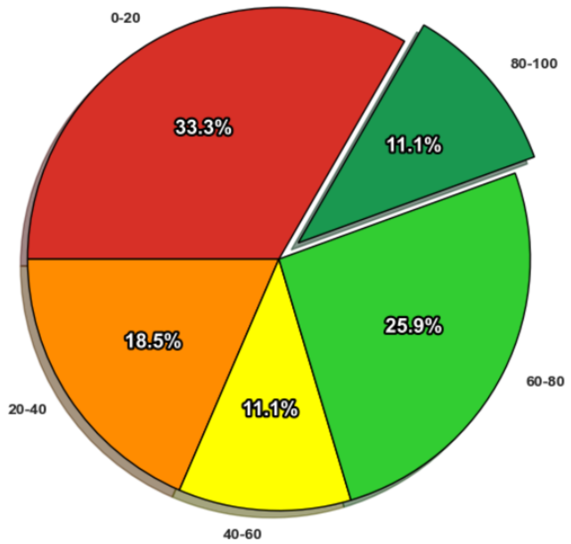


Figure 65. Pisa: SSEI classification by classes.

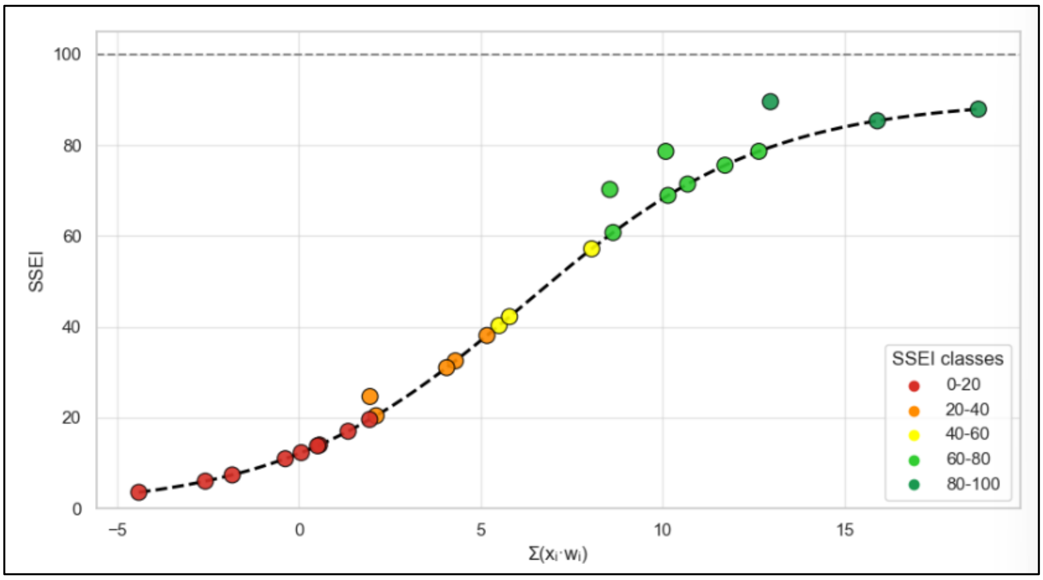


Figure 66. Pisa: SSEI logistic function.

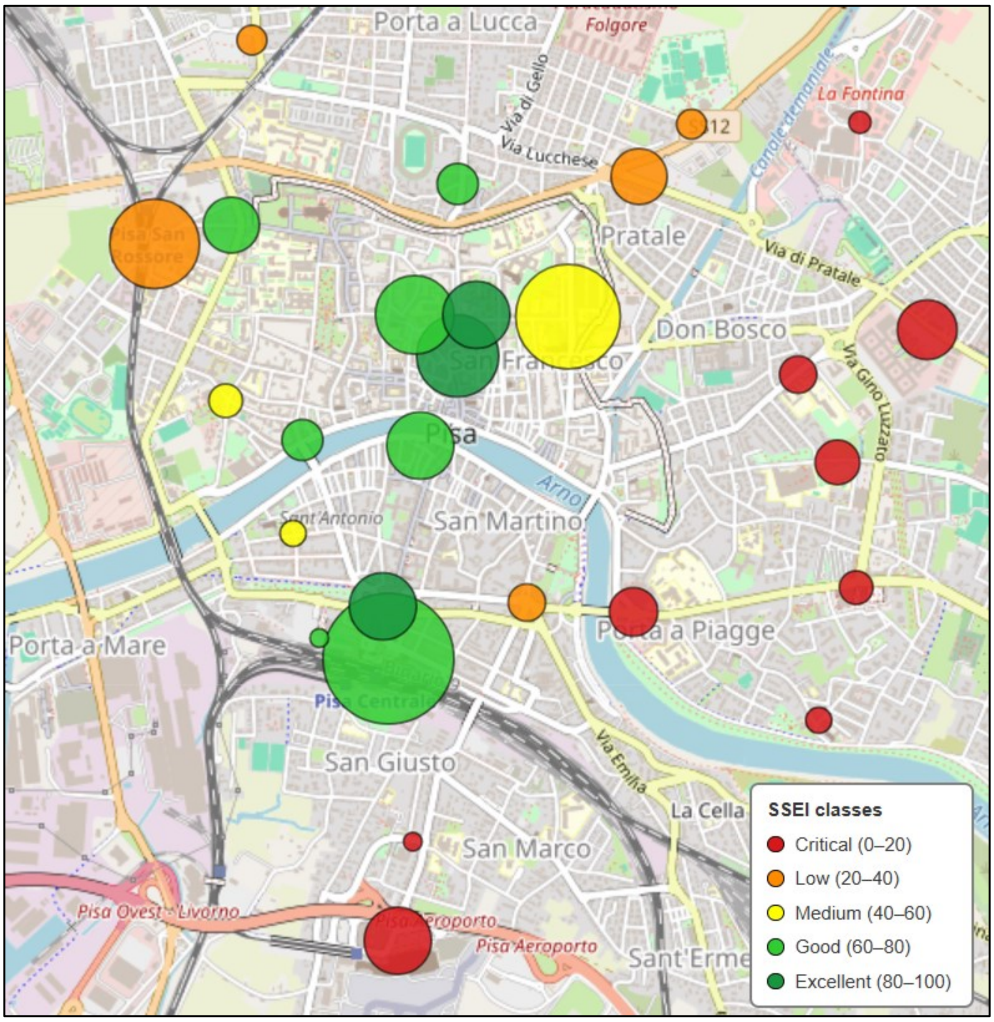


Figure 67. Pisa SSEI map. Radius of the circles is proportional to the number of rentals.

There is a cluster of stations in the highest classes extending from the historic centre to the railway station area, where there are three bike-sharing stations: *Vittorio Emanuele, Sesta Porta* and *FS station*. On the contrary, low SSEI values are recorded for stations in the eastern part. The most striking case is that of the station near the airport. It is characterised by a moderate number of rentals but a low SSEI value, presumably for the same reasons mentioned for the Forlì hospital station. In addition, Pisa is the only city in the dataset that is served by two different railway stations, both located not far from the city centre. Both are characterised by a high number of rentals. However, while the BS station near the main railway station belongs to the "good" SSEI class, the BS station near *San Rossore* station does not stand out for its high SSEI value, due to its low functional density, especially in the area on the west of the railway.

Trieste

Table 54. Trieste: stations with the highest and lowest SSEI.

Highest		Lowest	
Station	SSEI	Station	SSEI
Battisti	86.8	*Cumano-musei*	3.6
Largo Barriera	82.8	*Miramare*	5.4
Oberdan	82.1	*Barcola*	6.3
Piazza libertà-Stazione	80.4	*Porto Vecchio - Miramare*	7.4
Teatro Romano	79.5	*Park Bovedo*	8.0

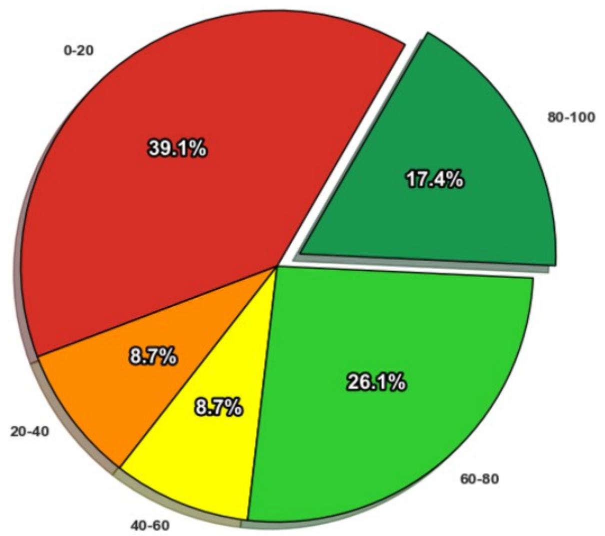


Figure 68. Trieste. SSEI classification by classes.

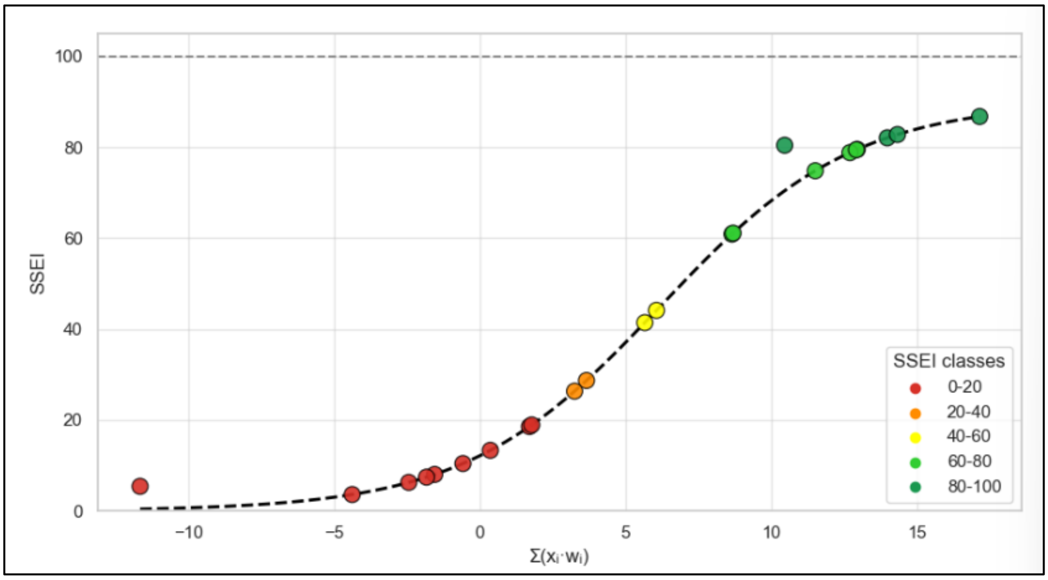


Figure 69. Trieste: SSEI logistic function.

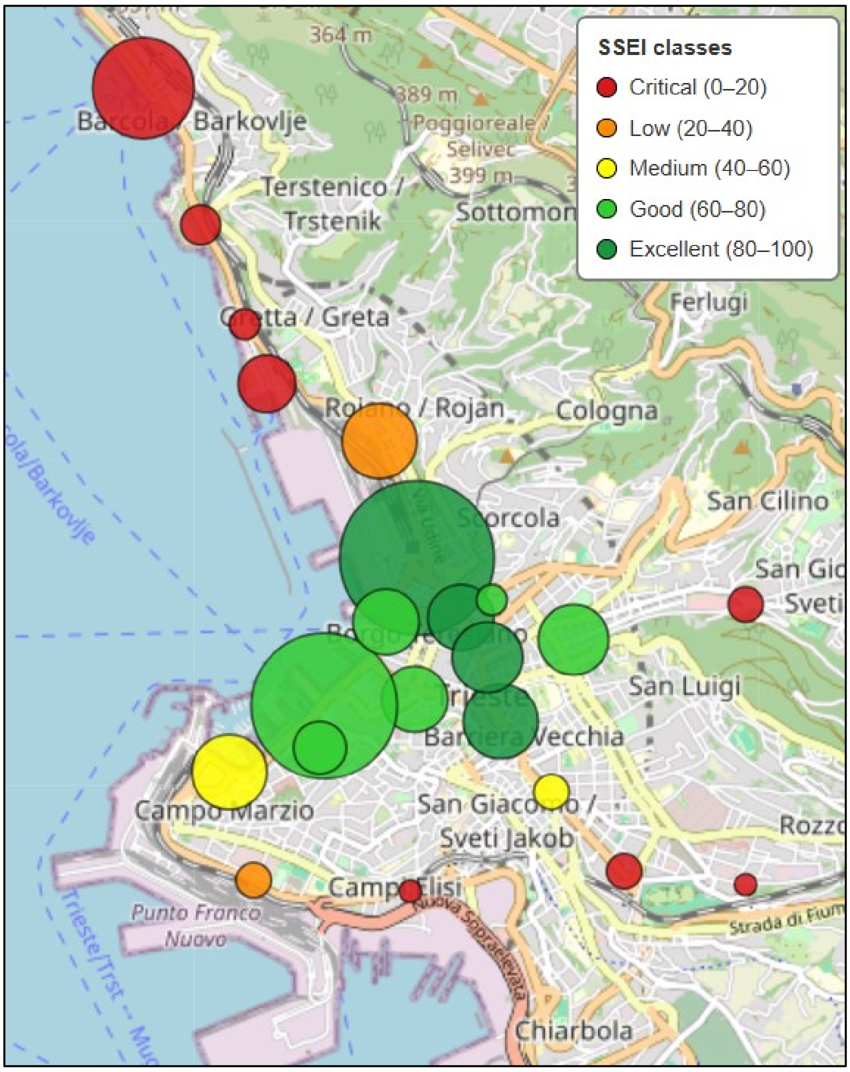


Figure 70. Trieste SSEI map. Radius of the circles is proportional to the number of rentals.

In Trieste, the best stations are all located along the city's main road, which runs from the railway station to the old barrier. The tourist area of *Piazza Unità d'Italia*, on the coast, also has stations with good SSEI scores. On the other hand, the worst stations are located in the suburbs and, in particular, on the northern seafront, a tourist area with several stations and moderate bike-sharing demand but characterised by a low density of cafes, restaurants and shops. However, this is a notable case in which the bike-sharing service has adapted to tourist demand. In fact, thanks to the stations located along the coast, tourists and citizens can cycle along the coastal cycle lane, which connects the city centre with *Miramare Castle,* located 7 km from the city centre. The bike-sharing station there, which is not shown on the map for reasons of limited space, is characterised by a significant number of rentals but a critical SSEI value.

5.2. Integrated SSEI-bikeability analysis

The newly created *Station Static Effectiveness Index* is ideal for an application that analyses the effectiveness of bike-sharing service integration with the cycle infrastructure network. The need for such an application arises from the recognition that a bike-sharing service is even more effective when it is adequately supported by a high-performance cycling infrastructure network (cycle paths, intersections, etc.).

Bikeability indices generally classify roads into different quality levels by measuring their level of service from the cyclist's perspective. The factors considered by such indicators are numerous, embracing various elements, such as traffic characteristics, road geometry and traffic regulations (e.g. speed limits). For these reasons, it was decided not to include these factors in the formulation of the SSEI: they are numerous and already included in other types of indicators. Therefore, it is more appropriate to consider the two types of indices separately, analysing their trade-off for a single station, calculating the SSEI alongside a bikeability index that reflects the level of service of the surrounding roads. By taking this approach, service planners can analyse two complementary aspects of urban cycling quality simultaneously, capturing both the node dimension (stations) and the network dimension (cycling infrastructure). At the end of the study, it will be possible to identify eventual intervention priorities on the service, identifying areas in which to invest in cycling infrastructure or those in which to suggest to the municipality to invest in urban development.

The assessment does not merely analyse the synergies between the cycle network and the bike-sharing service, but goes further, guaranteeing the following benefits:

- Evaluation of the consistency between bike-sharing planning and cycling infrastructure planning, highlighting whether the expansion of the bike-sharing system has been consistent with the development of the cycling network or whether there are misalignments between the two strategies.
- Supports an operational classification of stations, grouping them into homogeneous groups based on intervention needs, facilitating differentiated intervention strategies.
- Explicitly separates infrastructure interventions (road and cycle network) from service interventions (changes to the location of current or future bike-sharing stations).

For this application, a modified version of the *Bicycle Compatibility Index* (BCI) has been chosen among the ones suggested by the literature ([38], [54], [25]). BCI is an indicator developed in the USA during the late 1990s by a team of researchers guided by David L. Harkey [20]. It considers many parameters listed in Table 55, assigning the road to a class (from the best A to the worst F, reported in Table 56) after computing its level of service using a linear combination method. BCI was developed two decades ago, so it has been decided to modify the original formulation of BCI by adding a new parameter, PBL (*Protected Bike Lane*), related to the type of separation between the eventual bike lane and the road. PBL could assume three possible values:

- 0, if there is no separation;
- 1, if separation is intermittent or partial;
- 2, if there is a constant separation (guard rail or traffic separators).

The weight chosen for PBL is 0.485, equal to half of BL, related to the presence of a bike lane.

Table 55. BCI variables.

Variable Identification	Variable definition	Weight
-	Intercept	3.67
BL	Presence of a bicycle lane or paved shoulder ≥ 0.9m (1 = yes, 0 = no)	-0.966
BLW	Bicycle lane (or paved shoulder) width [m]	-0.410
CLW	Curb lane width [m]	-0.498
CLV	Curb lane volume in one direction in [veich/h/lane]	0.002
OLV	Other lane volume [veich/h/lane]	0.0004
SPD	85th percentile speed of traffic [km/h]	0.022
PKG	Presence of a parking lane alongside the cyclist lane (1 = yes, 0 = no)	0.506
AREA	Type of roadside development (1 = residential, 0 = other type)	-0.264
f_t	Adjustment factor for truck volumes (based on large truck volumes, always < 1)	1
f_p	Adjustment factor for parking turnover (based on parking time limit, always < 1)	1
f_{rt}	Adjustment factor for right-turn volume (based on road typology, always < 1)	1

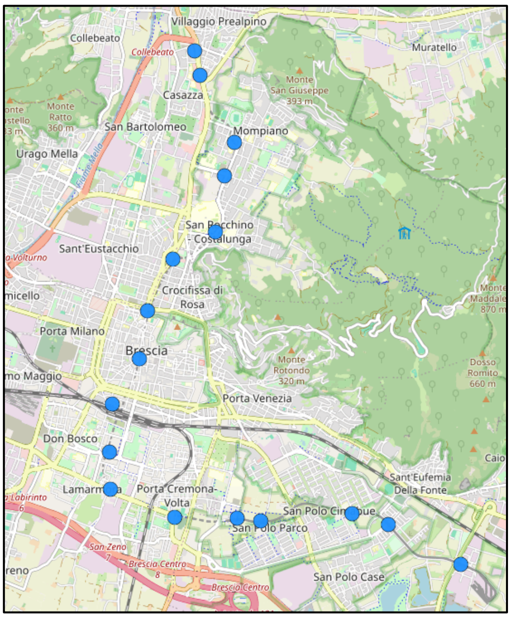
$$\begin{gathered} BCI = 3.67 - 0.966BL - 0.485PBL - 0.410BLW - 0.498CLW + 0.002CLV \\ + 0.0004OLV + 0.022SPD + 0.506PKG - 0.264AREA \\ + f_t + f_p + f_{rt} \end{gathered} \tag{5.1}$$

Since no actual traffic volume data were available, values were estimated based on the functional classification of roads defined by Italian regulations (*A*: motorway, *B*: main suburban road, *C*: secondary suburban road, *D*: urban expressway, *E*: urban neighbourhood road, *F*: local urban road). Similarly, as it was not possible to calculate the 85th percentile of circulating vehicles, a value equal to 105% of the speed limit on the road in question was assumed.

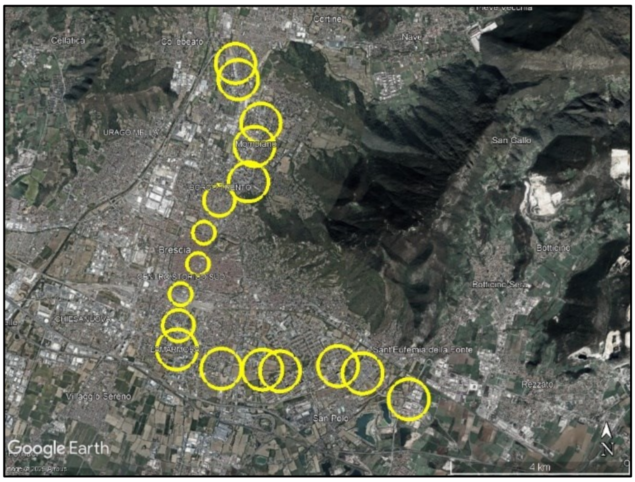
Table 56. BCI classes.

Class	BCI range	Compatibility level
A	≤ 1.50	Extremely high
B	1.51 ÷ 2.30	Very high
C	2.31 ÷ 3.40	Moderately high
D	3.41 ÷ 4.40	Moderately low
E	4.41 ÷ 5.30	Very low
F	> 5.30	Extremely low

The case study analyses the bike-sharing stations in Brescia, located close to underground stations (Figure 71a). The Brescia underground was inaugurated in 2013 and consists of a single driverless line connecting the *Buffalora* (south) and *Prealpino* (north) terminus stations, running through the city centre, thanks to the *Vittoria* and *San Faustino* stations. For each of the BS stations considered, the SSEI will be calculated using equation 4.4, while BCI refers to the main roads within the buffer around them (Figure 71b), as presented above. The average BCI for the roads considered will constitute the BCI for the bike-sharing station, which will be used in the trade-off analysis.



a) OpenStreetMap

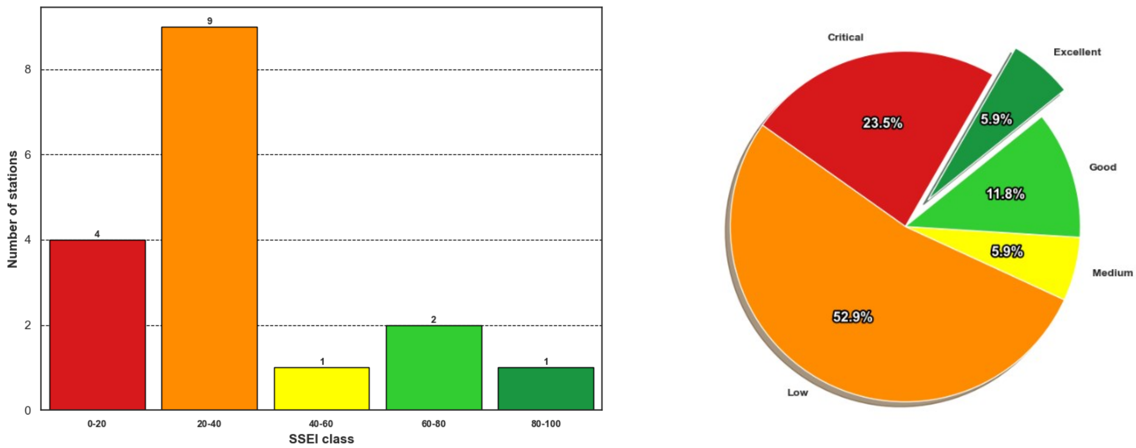


b) Buffers

Figure 71. BS stations near underground stops.

5.2.1. SSEI computation

SSEI has been computed for each station using the mathematical formulation proposed in Chapter 4.4.2, equation 4.4. The distance from the city centre has not been taken into account, since the case study examines bike-sharing stations with high variability for this variable. In fact, the minimum `dist_centre` recorded is equal to 150m (*Vittoria* station) while the maximum is 5,600m (*Buffalora* station). This is a further opportunity to reiterate the robustness of the SSEI index, as it demonstrates that its calculation is possible even when certain explanatory variables are not taken into account, yielding comparable results across different bike-sharing stations. SSEI values are reported in Table 57, while Figure 75 shows a portrait of stations painted with the corresponding SSEI class. As expected, *Stazione FS* is the station with the highest SSEI and the only one in the "excellent" class because of its high level of integration with the transport network and, at the same time, moderate density of points of interest. On the other hand, distribution graphs (Figure 72) suggest the presence of a big portion of stations which belong to the lowest SSEI classes. These could signify a partial isolation of some station from the main city's spots. All stations belonging to the 'low' class are located in the south-eastern suburbs of the city, as shown on the map in Figure 75a. This area of the city is mainly used for industrial development, unlike the northern suburbs, which are more residential and therefore have a higher housing density and more points of interest (thus raising the SSEI of the stations located there).



a) Histogram b) Pie plot

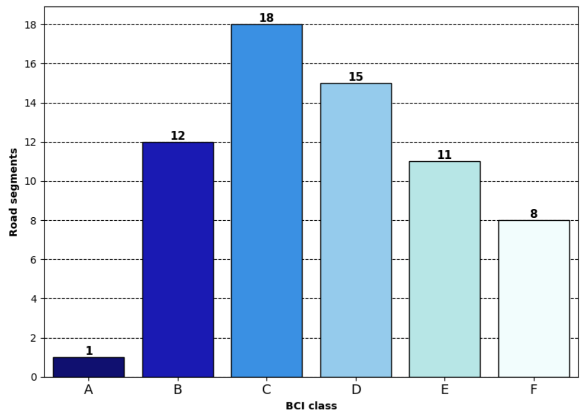
Figure 72. SSEI distribution by class.

5.2.2. BCI computation

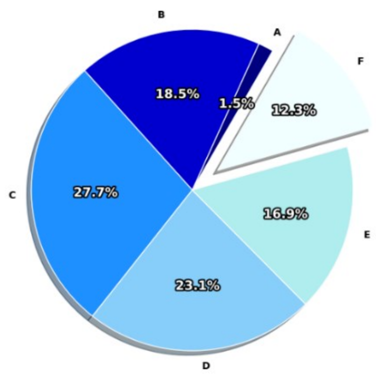
BCI has been computed using equation 5.1, averaging the main roads located in the buffer of each considered bike-sharing station. The number of studied roads within the buffers is comprised between 3 and 5, since it may vary based on various factors. In computation, the following rules have been respected:

- If the characteristics of the road change based on the considered direction (e.g. presence of a parking lane in just one direction of the road), BCI was computed two times, one for each direction;
- If there is a major arterial road in the buffer, it is analysed regardless of other factors, even if it does not directly connect the bike-sharing station to the rest of the network;
- All road connections relating to a single bike-sharing station have been studied (e.g. if a station is located near a junction of four roads, the BCI is calculated for each arm of the junction).

The findings demonstrate that the analysed roads are concentrated mainly in the four central classes of the BCI (Figure 73), with class C having the highest frequency (18 roads), followed by D and B. The only road belonging to class A is *Via Cimabue* in the eastbound direction, near the *San Polo-Cimabue* bike-sharing station. This road has a dual-lane cycle lane, completely separated from the main road by a parking lane (which does not interfere with cyclists) and a pavement. On the other hand, class F is represented by eight different roads, all characterised by the absence of a dedicated cycle lane.



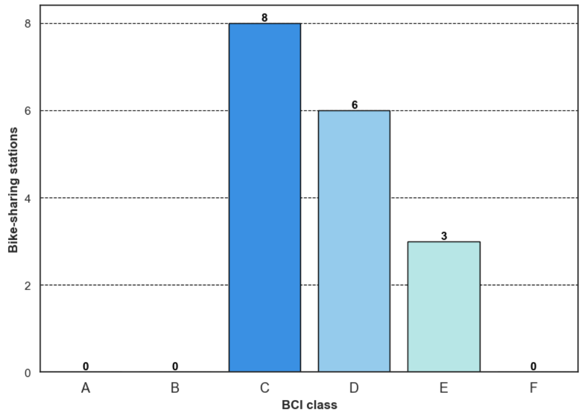
a) Histogram



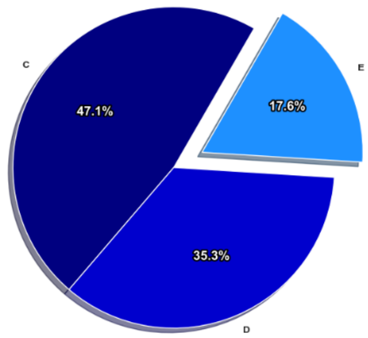
b) Pie plot

Figure 73. Roads BCI distribution by class.

Aggregating roads per station (Figure 74) and computing BCI for each of them causes a further densification in the central classes. This is consistent with the chosen method, which involves calculating the average BCI of the roads around each station. Classes A, B and F have no representatives, while almost half of the stations lie in class C.



a) Histogram



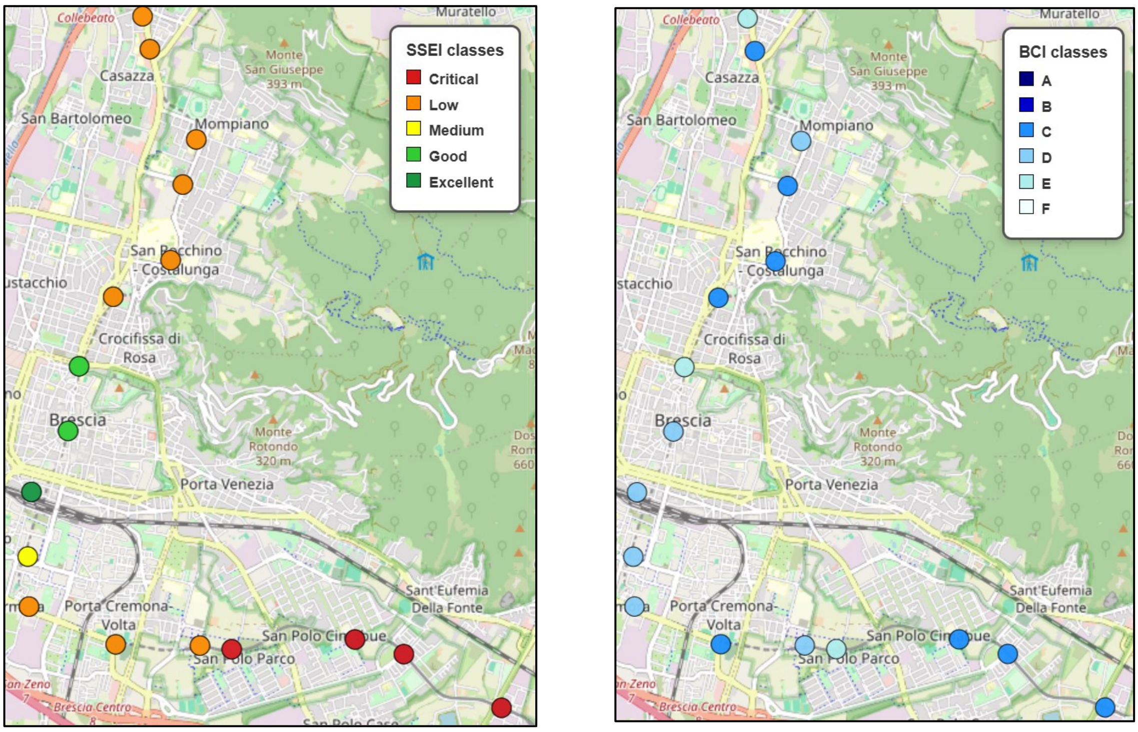
b) Pie plot

Figure 74. Stations BCI distribution by class.

As reported in Table 57, *Sanpolino* bike-sharing station is the best from the point of view of BCI (2.48), followed by *Ospedali Civili* (2.53) and *Europa* (2.54). These three stations are characterised by the fact that they are not close to the city centre and are therefore located in areas where there is undoubtedly more space to build cycle lanes, often adequately protected from vehicular traffic. Conversely, *Prealpino* results in being the worst bike-sharing station based on BCI because of its proximity to high traffic roads with no dedicated lanes for cyclists. Maps in Figure 75 propose a graphical representation of stations painted based on their BCI class. No relevant correlations with other variables such as the number of rentals ($r = 0.35$) or distance from the city centre ($r = -0.3$) are evident.

Table 57. SSEI and BCI for case study BS stations.

Station	SSEI	SSEI class	BCI	BCI class
Brescia Due	46.4	Medium	3.42	D
Buffalora	2.1	Critical	2.99	C
Casazza	25.1	Low	3.31	C
Europa	26.3	Low	2.54	C
Lamarmora	22.1	Low	3.79	D
Marconi	31.6	Low	3.31	C
Mompiano	27.9	Low	3.62	D
Ospedali Civili	30.8	Low	2.53	C
Poliambulanza	31.2	Low	4.14	D
Prealpino	23.9	Low	4.77	E
San Faustino	72.4	Good	5.02	E
San Polo Cimabue	16.8	Critical	2.77	C
San Polo Parco	19.5	Critical	4.62	E
Sanpolino	3.8	Critical	2.48	C
Stazione FS	87.1	Excellent	4.17	D
Vittoria	74.2	Good	3.83	D
Volta	23.5	Low	2.67	C



a) Grouped by SSEI

b) Grouped by BCI

Figure 75. Maps of case study BS stations.

5.2.3. Trade-off analysis

The measurements presented in the two previous paragraphs have demonstrated that there is no concrete correlation between SSEI and the selected bikeability index. This confirms that a trade-off analysis is necessary, as the two indices measure complementary aspects of urban cycling quality. Consequently, a good station location does not guarantee good local cycling conditions, and, similarly, good cycling conditions do not automatically imply that a station is well integrated into its surroundings. This is clearly illustrated by the map in Figure 76, which shows, through a simple sight of colour distribution, that, especially for the BCI, they are well distributed without the presence of homogeneous groups.

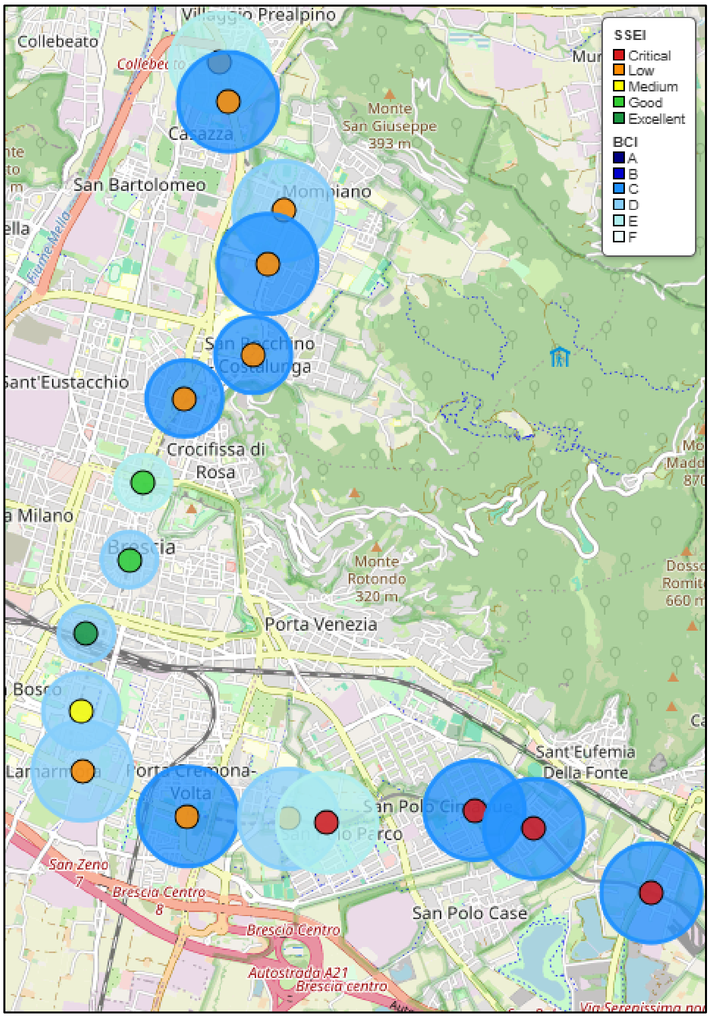


Figure 76. SSEI-BCI trade-off maps.

The first conclusions suggested by the scatter plot in Figure 77 confirm what was stated previously. The relationship between SSEI and BCI is neither linear nor monotonic, with stations appearing to be very scattered.

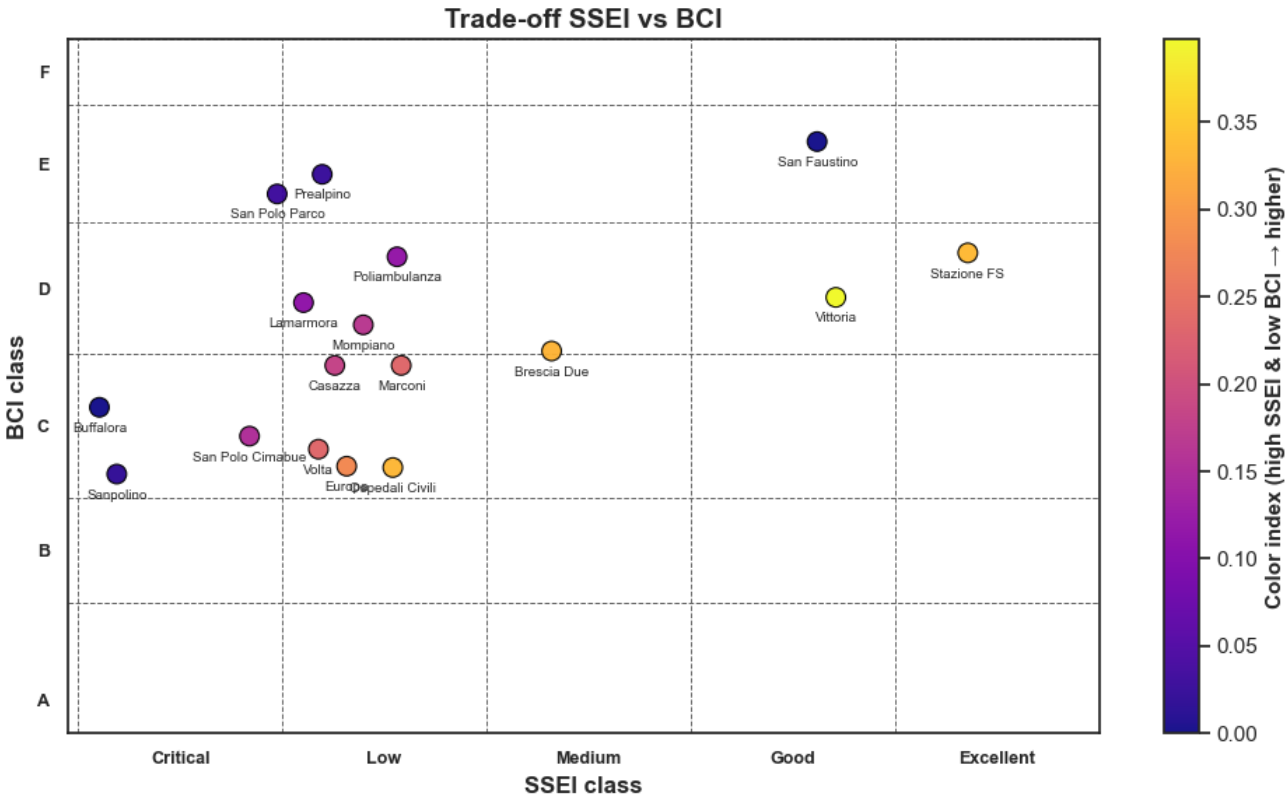


Figure 77. SSEI-BCI trade-off scatterplot.

Combining the two indices, the best bike-sharing station is *Vittoria*, characterised by a very high SSEI and an average BCI (class D). Moreover, most stations are concentrated between low/medium SSEI and medium-high BCI classes. This means that many improvements can be performed in terms of both infrastructure and service. However, it should be noted that *Bicimia* is an extensive bike-sharing service (97 stations currently, in 2025), so it is inevitable that some stations will not be optimally integrated into the urban context (low SSEI).

The bike-sharing station near the railway station is the best in terms of SSEI. However, the cycling quality in the surrounding area is not optimal, as it is located near multi-lane urban roads with heavy traffic. *San Faustino* station, on the northern edge of the city centre, performs even worse. It is located at the intersection of three roads with more than one lane in each direction. Despite the presence of a dedicated cycle lane, the quality of cycling is affected by high traffic volumes and narrow lanes. For these reasons, the *San Faustino* BS station is the only one that falls under the BCI F class.

Some stations are skewed towards positive BCI but with critical urban integration. These are the bike-sharing stations in *Buffalora* (an industrial area in the south-eastern suburbs) and *San Polo Cimabue* and *Sanpolino*, located in suburban neighbourhoods in the southern part of the city. Other stations that stand out for their good BCI are *Volta*, *Ospedali Civili* and *Europa*; their SSEI is slightly higher (class "low"), probably because they are closer to the city centre. The stations in this area of the graph are therefore already satisfactory in terms of infrastructure. On the contrary, they are lacking in terms of service. One action that could be taken is to advise the municipality to introduce incentives for those who invest in the construction of new points of interest (restaurants, shops, sports activities, etc.) in these areas, as well as introducing new road transport lines.

The worst stations are located in the upper left part of the graph. *San Polo Parco* is by far the worst in the trade-off analysis, as it is characterised by a critical SSEI and a very low BCI. This is because it is isolated from the rest of the urban development to the north, while there are few points of interest to the south. Alongside it is *Prealpino* station, the northern terminus of the underground, which has a slightly better SSEI but is by far the station with the highest BCI, as specified in the previous chapter. Stations such as these require significant improvements in terms of cycling infrastructure, as well as investment to add more points of interest in their surroundings.

Summarising the considerations above, it is possible to create an evaluation matrix that helps to immediately understand the status of each station. This matrix is shown in Table 58 and mapped in Figure 78. Please note that it has been created for index values related to this case study, as there is a considerable density of stations in the central classes. For example, the BCI classes are divided unevenly ("low" = classes E, F; "medium" = class D; "high" = classes A, B, C). Consequently, it cannot be used with the same classes in other contexts. The colours specify both the status of the stations and the priority of intervention in the surrounding area:

- *Dark green*: excellent quality level, no action required;
- *Light green*: good quality level, low priority for intervention;
- *Yellow*: average quality level, medium priority for intervention;
- *Orange*: low quality level, high priority for intervention;
- *Red*: critical quality level, maximum priority for intervention.

Lastly, it should be noted that a more in-depth analysis should also include a study of the number of rentals for each station. For example, *Stazione FS* bike-sharing station is the third station in terms of number of rentals, but the cycling quality could be improved. Therefore, it is advisable to intervene at an infrastructural level because of the high usage by users, even though it is already located in the green area of the matrix.

Table 58. SSEI-BCI trade-off evaluation matrix.

	Critical SSEI	Low-medium SSEI	Good-excellent SSEI
High BCI (E-F)	*San Polo Parco*	*Prealpino*	*San Faustino*
Medium BCI (D)	-	*Lamarmora* *Mompiano* *Poliambulanza* *Brescia Due*	*Vittoria* *Stazione FS*
Low BCI (A-B-C)	*Buffalora* *Sanpolino* *San Polo Cimabue*	*Volta* *Europa* *Ospedali Civili* *Casazza* *Marconi*	-

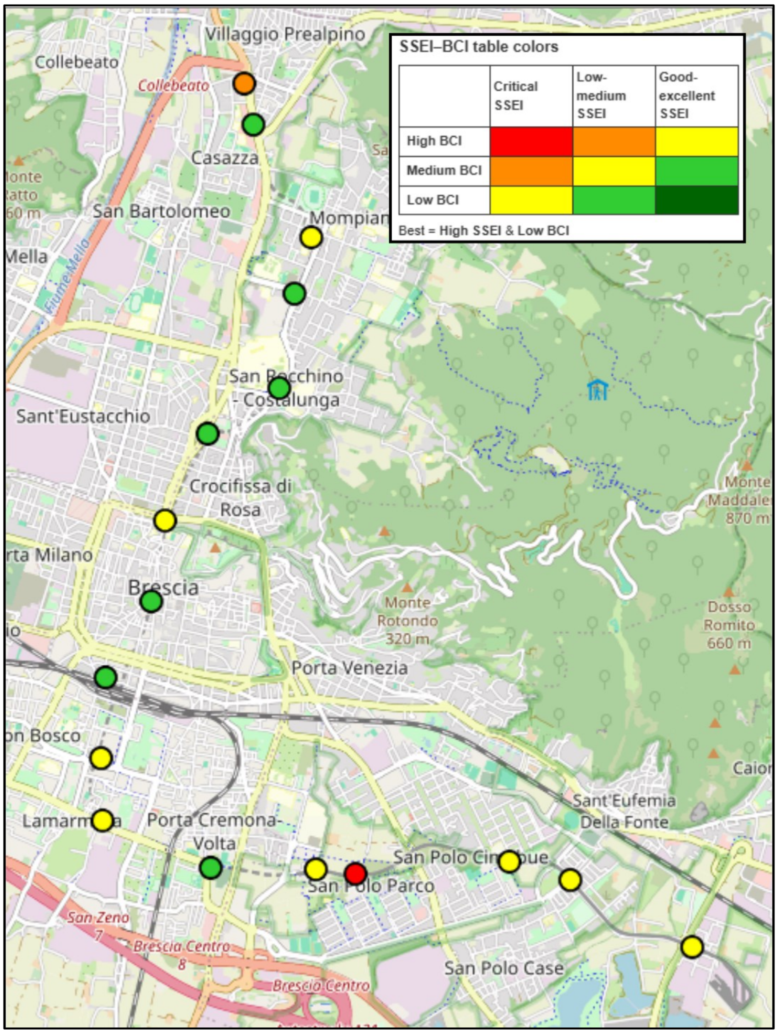


Figure 78. SSEI-BCI trade-off map based on evaluation matrix.

To wrap up the case study, an unsupervised clustering analysis in the SSEI–BCI space has been conducted (Figure 79). This was done to independently validate the previous classification matrix. First off, we normalised both indices to make sure they were comparable, and then we combined them in a weighted K-Means framework, giving more weight to the BCI dimension to better reflect local cycling conditions. The best number of clusters (equal to five clusters) has been figured out by using silhouette analysis, which helped to find a solution that strikes a good balance between how tightly the groups hold together and how distinct they are from each other. The clusters we got show a clear organisation based on the main factors of station integration quality (SSEI) and cycling compatibility (BCI). Most stations ended up in clusters that are mainly within one or two neighbouring cells of the SSEI–BCI matrix, which confirms that our matrix captures the data's underlying structure well. Any leftover differences mostly arise from the different methods used: clustering depends on distance-based decision boundaries, while the matrix sets fixed threshold values for easier interpretation. Lastly, the optimal number of clusters is lower than the number of cells in Table 58, meaning that our previous assessment is robust, since we analysed more "classes" than what is suggested by the Machine Learning method. These results reinforce that our matrix is a solid, practical tool for evaluating the outcomes of this case study.

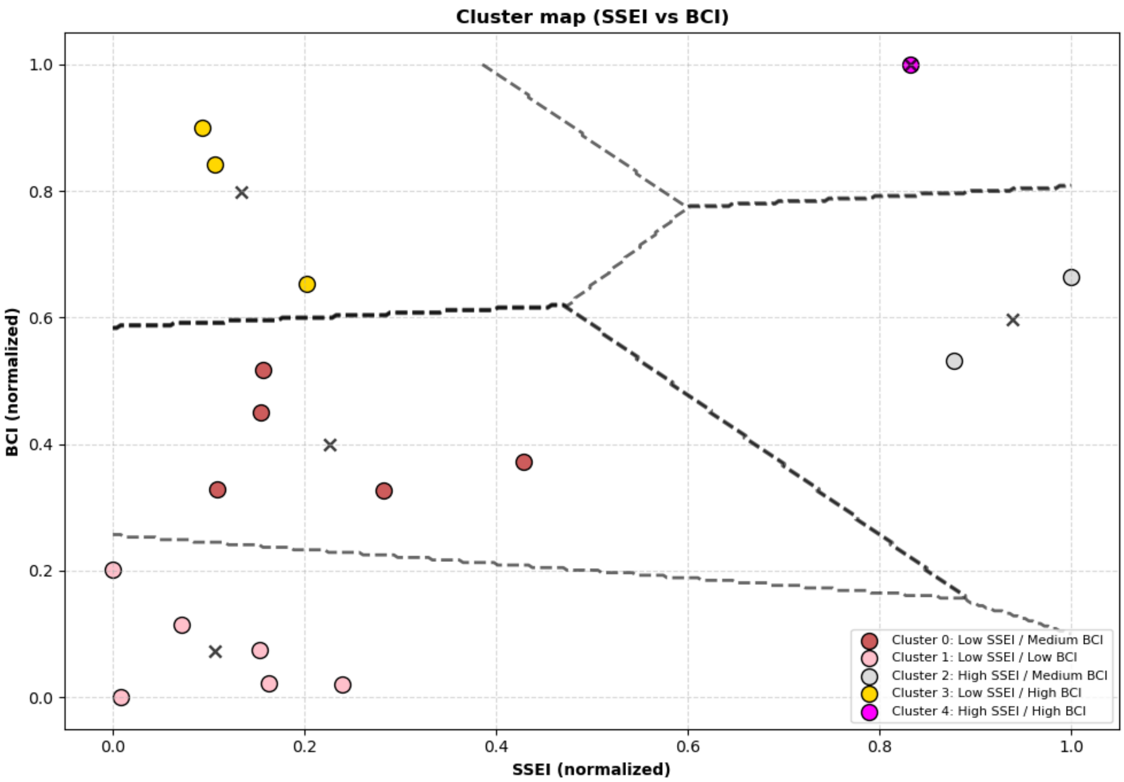


Figure 79. Case study BS stations clustering.

6 Conclusions

To conclude the research, a critical analysis of the obtained outcomes is now provided, examining the main conclusions reached and reiterating the proposed innovations. In addition, the potential future developments of the proposed method and its main limitations are presented and explored in depth. Therefore, this last section of the document offers a further overview of the most important achievements obtained through the data-driven approach used and the subsequent development of the *Station Static Effectiveness Index*.

6.1. Critical analysis of results

The thesis aimed to propose a new method for analysing data related to bike-sharing stations to determine which urban context factors most influence service demand. The literature review highlighted recurring issues, such as excessive adaptation to the analysed city, a heavy reliance on demand and budget constraints when choosing station locations, and an overemphasis on the operational characteristics of the service when evaluating performance. Therefore, the proposed method is innovative, as it considers bike-sharing stations from different contexts (various services and cities), and takes into account only objective factors that depend on the context of the station's location and are not linked to the operational characteristics of the service, such as the availability of bicycles and parking slots. This enables significant in-depth assessments to be carried out. Among these, the most relevant is the creation of a quantitative index that measures the effectiveness of the urban integration of a station. It suggests an innovative method for the "static" evaluation of the service offered, for the creation of a hierarchy of interventions and also for determining the optimal location of stations, alongside existing approaches.

The developed Gradient Boosting model established that the main drivers of demand are those related to functional density in the proximity of the station, such as the presence of points of interest for personal, tourist, or leisure activities, the distance from the city centre, and integration with individual transport, including the distance from cycle lanes or car parks and proximity to other bike-sharing stations. A high number of accessible public transport lines, the presence of a railway station, and location within a residential neighbourhood are also of considerable relevance. On the other hand, the physical characteristics of the stations, such as their positioning in relation to the road, the presence of weather protection, or lighting conditions, have little influence. The brand-new index, named the *Station Static Effectiveness Index* (SSEI), condenses these elements to provide a quantitative measure of how effectively a station is integrated into its surroundings. It is based on a logistic function, whose slope was calibrated on the dataset under analysis, and includes a binary variable related to the presence of a railway station. This choice proved effective, as the application on stations from the study dataset results in having an adequately heterogeneous distribution of computed index values, with no SSEI class standing out over the others.

The case study discussed extensively in the final section highlighted two possible applications of SSEI: internal auditing to assess the status of the service and parallel analysis with the quality of cycling infrastructure in the surroundings of the station, which allows for the identification of intervention priorities in the analysed areas. The findings showed that the index is capable of uncovering previously unnoticed insights, pinpointing areas where bike-sharing is most effective, while also highlighting situations in which stations are well integrated into the urban environment but insufficiently supported by cycling infrastructure, or vice versa. The case study analysis confirmed the wide range of potential applications of the SSEI, which are not limited to the two previously presented examples but also extend to the support of station allocation techniques.

In addition, two complementary analyses were conducted in parallel: the analysis of local demand and the investigation of factors influencing user satisfaction. The former highlighted that, in the individual contexts analysed, the main drivers of demand may deviate from general trends, particularly with respect to the relative importance of points of interest or specific local conditions, such as in the case of the city of Genoa, where the factor with the greatest influence on users' choices is the type of separation between the station and the roadway. The latter, on the other hand, demonstrated that proximity to points of interest has a weaker influence on user satisfaction than on demand, except for points of interest related to personal needs. Conversely, proximity to cycle paths and other bike-sharing stations was found to be more relevant in determining user satisfaction.

Overall, this study demonstrates that bike-sharing performance can be effectively interpreted through a static, context-based perspective, complementing traditional demand- and operation-oriented approaches. By focusing exclusively on objective urban features, the proposed framework provides a robust and transferable tool for understanding how stations interact with their surrounding environment. The SSEI offers a clear and synthetic representation of this interaction, enabling meaningful comparisons across different cities and services. In this sense, the index represents not only an evaluation metric but also a planning support instrument, capable of bridging demand analysis, urban context assessment, and infrastructure quality. The results confirm that bike-sharing effectiveness is deeply embedded in the spatial and functional structure of cities, reinforcing the need for integrated planning strategies that jointly address service provision and urban design.

6.2. Study limitations

Despite the robustness of the proposed framework and the consistency of the results obtained, some limitations of this study should be acknowledged. These limitations do not undermine the validity of the proposed framework but rather better define its scope of applicability.

The main limitation is related to the operative framework of the research. The entire study presented in this document is based on a dataset of 140 bike-sharing stations located in various Italian cities. Although the dataset was designed to capture even subtle peculiarities of the urban context across different services, it is not possible to guarantee that the results obtained would remain valid when applied to contexts that differ substantially from those used for model calibration. Consequently, the application of SSEI and the interpretation of the results may be less accurate when analysing stations located in cities with more than 600,000 inhabitants, and potentially significantly distorted in metropolitan areas with populations exceeding one million. Similar considerations may also apply to much smaller urban contexts. However, it should be noted that bike-sharing services are relatively uncommon in cities of this size, particularly when they are located far from larger urban centres.

Another limitation is inherent to the methodological choices adopted in this study. Specifically, the framework was intentionally designed to remain independent from the operational characteristics of the service. While this allows for a clear assessment of the role of the urban context, it also implies that certain aspects of service performance are not captured by the analysis. Nevertheless, it should be emphasised that the proposed method is not intended to replace operational performance evaluations. Instead, the SSEI is conceived as a complementary tool to be applied in parallel with other evaluation indices already in use.

6.3. Further developments

Building on the results achieved and the methodological framework proposed, several directions for future research could be identified. Future developments may involve both methodological extensions of the index itself and the exploration of new application domains, ranging from urban planning to operational decision support.

At the methodological level, as already widely discussed in the previous sections, a new algorithm could be developed for the station allocation phase using the *Station Static Effectiveness Index* as an objective function to be maximised. This formulation could be adopted either as a stand-alone objective function or in combination with others, leading to the definition of a multi-objective optimisation model. In addition, using the same machine learning methods and the same set of variables, it would be possible to implement a model capable of estimating the expected number of rentals for a given station. This would greatly support service planners in the station sizing process, helping to bridge the gap between the station allocation and service start-up phases.

One other extension of the SSEI is its integration with a selected bikeability index. By applying appropriate statistical techniques, such as Principal Component Analysis, the formulations of the two indices could be combined in a unique equation, obtaining a more comprehensive indication of the station's static performance. However, Chapter 5.2 has shown that the two indices are not strongly correlated, as they capture distinct aspects of cycling mobility. Therefore, particular care would be required in defining the relative weights assigned to each of the two components (SSEI and bikeability index) when constructing a composite indicator.

Lastly, at the applicative level, the SSEI could be used in the future as a tool for monitoring the service over time, by periodically computing the index for each station, especially following the construction of new infrastructure or other urban interventions. Nevertheless, even in this case, caution would be necessary, as demand drivers may evolve over time and the SSEI formulation calibrated in the present study may no longer be fully representative after several decades.

Overall, these potential developments highlight the flexibility and scalability of the proposed framework. The SSEI can evolve from a static evaluation indicator into a broader planning and monitoring tool, supporting both strategic and operational decisions. Its integration with predictive models and complementary indices would further strengthen its applicability, reinforcing the role of context-based, data-driven approaches in the planning and evaluation of bike-sharing systems.

7 Appendix A – Python codes

This appendix contains all the functions written in Python programming language that were used to create the model. The codes are presented for the main model. Still, the same ones, with appropriately replaced input variables, were also used for the other models proposed in the thesis and are easily exportable, allowing them to be used in other contexts.

Outliers analysis

```
def cap_outliers(df, exclude_cols=None, both_sides_cols=None, percentile=0.9,
threshold_perc=5):
    """
    Analizza e gestisce gli outlier nelle colonne del DataFrame, applicando un capping
selettivo.
    Parametri:
        df (pd.DataFrame): DataFrame di input.
        exclude_cols (list): Colonne da escludere completamente dall'analisi.
        both_sides_cols (list): Colonne per cui analizzare outlier sia sopra che sotto il
percentile.
        percentile (float): Percentile di riferimento per identificare i valori limite
(default = 0.9).
        threshold_perc (float): Percentuale soglia di outlier oltre la quale viene
applicato il capping.
    Ritorna:
        df_mod (pd.DataFrame): DataFrame aggiornato con eventuali valori capped.
        report_upper (pd.DataFrame): Report per le colonne con capping solo superiore.
        report_both (pd.DataFrame): Report per le colonne con capping su entrambi i lati.
    """
    df_mod = df.copy()
    if exclude_cols is None:
        exclude_cols = []
    if both_sides_cols is None:
        both_sides_cols = []
    report_upper = []
    report_both = []
    cols = [c for c in df.columns if c not in exclude_cols]
    n_upper, n_both = 0, 0
    for col in cols:
        series = df_mod[col].dropna()
        q_upper = series.quantile(percentile)
        q_lower = series.quantile(1 - percentile)

        # Analisi outlier superiore
        n_out_upper = (series > q_upper).sum()
        perc_out_upper = n_out_upper / len(series) * 100
        # Analisi outlier inferiore (solo se richiesto)
        n_out_lower, perc_out_lower = 0, 0
        if col in both_sides_cols:
            n_out_lower = (series < q_lower).sum()
            perc_out_lower = n_out_lower / len(series) * 100
```

```
        # Capping superiore
        cap_upper = False
        if perc_out_upper >= threshold_perc:
            df_mod[col] = np.where(df_mod[col] > q_upper, q_upper, df_mod[col])
            cap_upper = True
        # Capping inferiore (solo per entrambe le direzioni)
        cap_lower = False
        if col in both_sides_cols and perc_out_lower >= threshold_perc:
            df_mod[col] = np.where(df_mod[col] < q_lower, q_lower, df_mod[col])
            cap_lower = True
        # Report
        if col in both_sides_cols:
            report_both.append({
                'Variabile': col,
                f'Limite inferiore ({int((1 - percentile) * 100)}°p)': round(q_lower, 2),
                f'Limite superiore ({int(percentile * 100)}°p)': round(q_upper, 2),
                'Outlier inferiore': n_out_lower,
                'Outlier superiore': n_out_upper,
                'Percentuale inf (%)': round(perc_out_lower, 2),
                'Percentuale sup (%)': round(perc_out_upper, 2),
                'Capping inferiore': cap_lower,
                'Capping superiore': cap_upper
            })
            n_both += int(cap_upper or cap_lower)
        else:
            report_upper.append({
                'Variabile': col,
                f'Limite superiore ({int(percentile * 100)}°p)': round(q_upper, 2),
                'Outlier superiori': n_out_upper,
                'Percentuale outlier (%)': round(perc_out_upper, 2),
                'Capping superiore': cap_upper
            })
            n_upper += int(cap_upper)
    # Conversione in DataFrame
    report_upper = pd.DataFrame(report_upper).sort_values(by='Percentuale outlier (%)',
ascending=False)
    report_both = pd.DataFrame(report_both).sort_values(by='Percentuale sup (%)',
ascending=False)
    print(f" DataFrame aggiornato: {n_upper} colonne modificate per limite superiore,
{n_both} per limiti superiore/inferiore.")
    return df_mod, report_upper, report_both
```

Multicollinearity analysis

```
from statsmodels.stats.outliers_influence import variance_inflation_factor

def calcola_vif(df, colonne=None):
    """
    Calcola il Variance Inflation Factor (VIF) per ogni colonna numerica.
    Gestisce casi di collinearità perfetta evitando warning e valori infiniti non gestiti.
    """
    # Se non specificate, considera solo colonne numeriche
    if colonne is None:
        colonne = df.select_dtypes(include=['float64', 'int64']).columns.tolist()

    # Rimuove righe con NaN
    X = df[colonne].dropna()

    vif_data = []
    for i, col in enumerate(colonne):
        try:
            vif_val = variance_inflation_factor(X.values, i)
            if np.isinf(vif_val):
                print(f"Collinearità perfetta rilevata per '{col}' (VIF → ∞).")
        except Exception:
            vif_val = np.inf
            print(f"Errore nel calcolo VIF per '{col}' — impostato a ∞.")
        vif_data.append({'Variabile': col, 'VIF': round(vif_val, 2)})

    vif_df = pd.DataFrame(vif_data)
    return vif_df
```

Linear regression

```
from sklearn.model_selection import KFold, cross_val_score, GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression, Lasso, Ridge, ElasticNet
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error,
median_absolute_error
from scipy import stats

def regression_model(df, target, model_type='simple', k=10, normalize=True,
random_state=42, min_nonzero_vars=10):
    """
    Applica un modello di regressione lineare (OLS, Lasso, Ridge, Elastic Net) con
validazione K-Fold,
    tuning dei parametri, diagnostica e controllo sul numero minimo di variabili con
coefficiente != 0.
    Ora calcola anche MAPE e sMAPE.
    """

    # --- Prepara i dati
    X = df.drop(columns=[target]).copy()
    y = df[target].copy()
    n, p = X.shape

    # --- Normalizzazione (necessaria per modelli penalizzati)
    scaler = None
    penalized_models = ['lasso', 'ridge', 'elasticnet']
    if normalize or model_type in penalized_models:
        scaler = StandardScaler()
        X_scaled = pd.DataFrame(scaler.fit_transform(X), columns=X.columns, index=X.index)
    else:
        X_scaled = X.copy()

    # --- Selezione modello e griglia parametri
    if model_type == 'simple':
        model = LinearRegression()
        param_grid = None
    elif model_type == 'lasso':
        model = Lasso(max_iter=10000, random_state=random_state)
        param_grid = {'alpha': np.logspace(-6, 3, 100)}
    elif model_type == 'ridge':
        model = Ridge(max_iter=10000, random_state=random_state)
        param_grid = {'alpha': np.logspace(-6, 3, 100)}
    elif model_type == 'elasticnet':
        model = ElasticNet(max_iter=10000, random_state=random_state)
        param_grid = {
            'alpha': np.logspace(-6, 3, 80),
            'l1_ratio': np.linspace(0.01, 0.99, 25)
        }
    else:
        raise ValueError("Usa: 'simple', 'lasso', 'ridge', 'elasticnet'.")

    # --- Tuning (se applicabile)
    if param_grid:
        grid = GridSearchCV(model, param_grid, cv=k, scoring='r2', n_jobs=-1)
        grid.fit(X_scaled, y)
        model = grid.best_estimator_
        best_params = grid.best_params_
        print(f" Parametri ottimali trovati per {model_type.upper()}: {best_params}")
    else:
        best_params = None
        model.fit(X_scaled, y)

    # --- Fit finale
```

```
    model.fit(X_scaled, y)
    coeffs = model.coef_
    intercept = model.intercept_

    # --- Controllo numero variabili attive (solo per Lasso / ElasticNet)
    nonzero_vars = np.sum(np.abs(coeffs) > 1e-6)
    if model_type in ['lasso', 'elasticnet']:
        if nonzero_vars < min_nonzero_vars:
            print(f"ATTENZIONE: solo {nonzero_vars} variabili attive (non nulle) "
                  f"nel modello {model_type.upper()} – min richieste: {min_nonzero_vars}")
        else:
            print(f"Numero variabili attive nel modello {model_type.upper()}:
{nonzero_vars}")

    # --- Statistiche dei coefficienti
    y_pred = model.predict(X_scaled)
    residuals = y - y_pred
    mse = np.sum(residuals ** 2) / (n - p - 1)
    try:
        var_b = mse * np.linalg.inv(X_scaled.T @ X_scaled).diagonal()
    except np.linalg.LinAlgError:
        var_b = np.repeat(np.nan, len(coeffs))
    sd_b = np.sqrt(var_b)
    with np.errstate(divide='ignore', invalid='ignore'):
        t_stats = coeffs / sd_b
        p_values = [2 * (1 - stats.t.cdf(abs(t), n - p - 1)) if not np.isnan(t) else np.nan
for t in t_stats]

    # --- Coefficienti standardizzati
    X_std = X_scaled.std()
    y_std = y.std()
    coeffs_std = coeffs * (X_std / y_std)

    coeff_df = pd.DataFrame({
        "Variabile": X.columns,
        "Coeff_regressione": np.round(coeffs, 5),
        "Coeff_standardizzato": np.round(coeffs_std, 5),
        "Deviazione_std": np.round(sd_b, 5),
        "t_stat": np.round(t_stats, 3),
        "p_value": np.round(p_values, 4)
    }).sort_values(by="Coeff_standardizzato", ascending=False).reset_index(drop=True)

    # --- Metriche di performance
    r2 = r2_score(y, y_pred)
    r2_adj = 1 - (1 - r2) * (n - 1) / (n - p - 1)
    kf = KFold(n_splits=k, shuffle=True, random_state=random_state)
    cv_scores = cross_val_score(model, X_scaled, y, cv=kf, scoring='r2')
    r2_cv_mean = cv_scores.mean()
    r2_cv_std = cv_scores.std()

    rss = np.sum(residuals ** 2)
    tss = np.sum((y - np.mean(y)) ** 2)
    rmse = np.sqrt(mean_squared_error(y, y_pred))
    mae = mean_absolute_error(y, y_pred)
    mdae = median_absolute_error(y, y_pred)

    # --- MAPE e sMAPE
    # MAPE (con esclusione di y=0)
    y_nonzero = np.where(y == 0, np.nan, y)
    mape = np.nanmean(np.abs((y - y_pred) / y_nonzero)) * 100

    # sMAPE: 2*|y - y_pred| / (|y| + |y_pred|)
    denom = np.abs(y) + np.abs(y_pred)
    smape = np.mean(
        np.where(denom == 0, 0, 2.0 * np.abs(y - y_pred) / denom)
```

```
    ) * 100

    # --- Warning diagnostici
    warnings = []
    if r2 < 0.5: warnings.append("R² basso (<0.5)")
    if r2_adj < 0.5: warnings.append("R² adj basso (<0.5)")
    if mape > 20: warnings.append("MAPE alto (>20%)")
    if smape > 20: warnings.append("sMAPE alto (>20%)")
    if mae > np.std(y) * 0.5: warnings.append("MAE elevato")
    if mse > np.var(y) * 0.5: warnings.append("MSE elevato")
    if model_type in ['lasso', 'elasticnet'] and nonzero_vars < min_nonzero_vars:
        warnings.append(f"Poche variabili attive ({nonzero_vars} < {min_nonzero_vars})")

    metrics_df = pd.DataFrame([{
        "Modello": model_type.upper(),
        "R²": round(r2, 3),
        "R²_adj": round(r2_adj, 3),
        "R²_CV_mean": round(r2_cv_mean, 3),
        "R²_CV_std": round(r2_cv_std, 3),
        "RSS": round(rss, 3),
        "TSS": round(tss, 3),
        "MSE": round(mse, 3),
        "RMSE": round(rmse, 3),
        "MAE": round(mae, 3),
        "MdAE": round(mdae, 3),
        "MAPE (%)": round(mape, 2),
        "SMAPE (%)": round(smape, 2),
        "Variabili_attive": nonzero_vars if model_type in ['lasso', 'elasticnet'] else p,
        "Warning": ", ".join(warnings) if warnings else "✅ Tutto regolare"
    }])

    metrics = {
        "R²": round(r2, 3),
        "R²_adj": round(r2_adj, 3),
        "R²_CV_mean": round(r2_cv_mean, 3),
        "R²_CV_std": round(r2_cv_std, 3),
        "MAPE (%)": round(mape, 2),
        "SMAPE (%)": round(smape, 2)
    }

    results = {
        "coeff_df": coeff_df,
        "metrics_df": metrics_df,
        "metrics": metrics,
        "best_params": best_params,
        "best_model": model,
        "scaler": scaler,
        "intercept": intercept,
        "nonzero_vars": nonzero_vars
    }

    # --- Output sintetico
        print(f"R² = {r2:.3f} | R² adj = {r2_adj:.3f} | CV = {r2_cv_mean:.3f} ±
{r2_cv_std:.3f}")
    print(f"RMSE = {rmse:.3f} | MAE = {mae:.3f} | MAPE = {mape:.2f}% | sMAPE =
{smape:.2f}%")
    print(f"Variabili attive: {nonzero_vars} / {p}")
    if warnings:
        print("warning", ", ".join(warnings))
    else:
        print("Nessun problema rilevato")

    return coeff_df, metrics_df, results
```

Linear regression diagnostics

```
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import r2_score, mean_squared_error

def check_linearity_report(df, target, model_result, rf_estimators=500, plot=True,
export_excel=False):
    """
    Verifica se la relazione tra predittori e target può essere spiegata da un modello
lineare.

    Compatibile con l'output della funzione regression_model().
    """

    # =========================
    # Recupera elementi dal dizionario dei risultati
    # =========================
    if not isinstance(model_result, dict) or "best_model" not in model_result:
        raise ValueError("L'input deve essere il dizionario restituito da
regression_model() (contiene 'best_model').")

    model = model_result["best_model"]
    scaler = model_result.get("scaler", None)

    X = df.drop(columns=[target])
    y = df[target]

    # Se è stato usato lo scaler, lo applichiamo
    if scaler is not None:
        X_scaled = pd.DataFrame(scaler.transform(X), columns=X.columns, index=X.index)
    else:
        X_scaled = X.copy()

    # =========================
    # Predizioni e residui
    # =========================
    y_pred = model.predict(X_scaled)
    residuals = y - y_pred

    # =========================
    # Analisi grafica dei residui
    # =========================
    if plot:
        fig, axes = plt.subplots(1, 2, figsize=(12, 5))
        sns.scatterplot(x=y_pred, y=residuals, ax=axes[0])
        axes[0].axhline(0, color='red', linestyle='--')
        #axes[0].set_title("Residui vs Valori Predetti")
        axes[0].set_xlabel("Predicted")
        axes[0].set_ylabel("Residuals")

        # Q-Q plot empirico
        from scipy import stats
        stats.probplot(residuals, dist="norm", plot=axes[1])
        #axes[1].set_title("Q-Q Plot dei residui")
        plt.tight_layout()
        plt.show()

    # Pattern dei residui
    pattern_flag = np.abs(pd.Series(residuals).rolling(5).mean().std()) > 0.5 *
np.std(residuals)

    # =========================
```

```
    # Confronto con Random Forest
    # =========================
    rf = RandomForestRegressor(n_estimators=rf_estimators, random_state=42)
    rf.fit(X_scaled, y)
    y_pred_rf = rf.predict(X_scaled)

    r2_lin = r2_score(y, y_pred)
    r2_rf = r2_score(y, y_pred_rf)
    rmse_lin = np.sqrt(mean_squared_error(y, y_pred))
    rmse_rf = np.sqrt(mean_squared_error(y, y_pred_rf))

    # =========================
    # LOWESS Plot
    # =========================
    if plot:
        plt.figure(figsize=(6, 4))
        sns.regplot(x=y_pred, y=y, lowess=True, line_kws={'color': 'red'})
        #plt.title("LOWESS smoothing tra valori predetti e osservati")
        plt.xlabel("Predicted")
        plt.ylabel("Observed")
        plt.show()

    # =========================
    # Decisione automatica
    # =========================
    flags = {
        "pattern_residui": pattern_flag,
        "RF_molto_migliore": (r2_rf - r2_lin) > 0.1
    }

    non_linear_flags = sum(flags.values())

    if non_linear_flags >= 1:
        decision = "La relazione non può essere spiegata adeguatamente da una regressione 
lineare."
        conclusion_text = (
            "L’analisi dei residui e il confronto con un modello non lineare (Random 
Forest) "
            "mostrano che l’andamento tra i predittori e la variabile target non è lineare. 
"
            "Il modello lineare non rappresenta in modo adeguato la relazione sottostante."
        )
    else:
        decision = "Una relazione lineare appare appropriata."

    # =========================
    # Report finale
    # =========================
    report = pd.DataFrame([{
        "R² lineare": round(r2_lin, 3),
        "R² random forest": round(r2_rf, 3),
        "Δ R² (RF - lineare)": round(r2_rf - r2_lin, 3),
        "RMSE lineare": round(rmse_lin, 3),
        "RMSE random forest": round(rmse_rf, 3),
        "Residui con pattern": bool(pattern_flag),
        "Conclusione": decision
    }])

    display(report)
    print("\nValutazione automatica dei test:\n", flags)
    print(decision)

    # =========================
    # Esportazione
    # =========================
```

```
    if export_excel:
        report.to_excel("linearity_report.xlsx", index=False)
        print("Report salvato come 'linearity_report.xlsx'")

    return report
```

Ensemble methods

```
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import KFold, cross_val_score
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error

def ensemble_methods(
    df,
    target,
    model_type='random_forest',
    n_estimators=300,
    max_depth=None,
    min_samples_split=2,
    min_samples_leaf=1,
    max_features='sqrt',
    learning_rate=0.05,
    subsample=1.0,
    k=10,
    plot=True,
    log_target=False,
    bootstrap_smoothing=False,
    random_state=42
):
    """
    Versione BASE del modello ensemble (Random Forest o Gradient Boosting).
    Modello diretto e senza regolarizzazioni aggiuntive.

    Parametri principali modificabili (con default coerenti):
    - Random Forest: max_depth=None, min_samples_split=2, min_samples_leaf=1,
max_features='sqrt'
    - Gradient Boosting: learning_rate=0.05, max_depth=3, subsample=1.0

    Restituisce:
        imp_df, metrics_df, model
    """

    X = df.drop(columns=[target])
    y_raw = df[target].copy()
    y = np.log1p(y_raw) if log_target else y_raw.copy()
    n, p = X.shape

    # --- Selezione modello
    if model_type == 'random_forest':
        model = RandomForestRegressor(
            n_estimators=n_estimators,
            max_depth=max_depth,
            min_samples_split=min_samples_split,
            min_samples_leaf=min_samples_leaf,
            max_features=max_features,
            bootstrap=True,
            random_state=random_state,
            n_jobs=-1
        )
    elif model_type == 'gboost':
        model = GradientBoostingRegressor(
```

```
            n_estimators=n_estimators,
            learning_rate=learning_rate,
            max_depth=3,
            min_samples_leaf=min_samples_leaf,
            subsample=subsample,
            max_features=max_features if isinstance(max_features, (int, float)) else
'sqrt',
            random_state=random_state
        )
    else:
        raise ValueError("Usa 'random_forest' o 'gboost' come model_type.")

    # --- Addestramento
    model.fit(X, y)
    preds_log = model.predict(X)
    preds = np.expm1(preds_log) if log_target else preds_log.copy()

    # --- Cross-validation
    kf = KFold(n_splits=k, shuffle=True, random_state=random_state)
    cv_r2 = cross_val_score(model, X, y, cv=kf, scoring="r2")
    mean_r2, std_r2 = cv_r2.mean(), cv_r2.std()

    # --- Metriche
    r2 = r2_score(y, preds_log if log_target else preds)
    r2_adj = 1 - (1 - r2) * (n - 1) / (n - p - 1)
    rmse = np.sqrt(mean_squared_error(y_raw, preds))
    mae = mean_absolute_error(y_raw, preds)

    def smape(y_true, y_pred):
        denom = (np.abs(y_true) + np.abs(y_pred))
        denom[denom == 0] = np.nan
        return 100 * np.nanmean(2 * np.abs(y_true - y_pred) / denom)

    def mase(y_true, y_pred):
        naive_forecast = np.roll(y_true, 1)[1:]
        mae_naive = np.mean(np.abs(y_true[1:] - naive_forecast))
        return np.mean(np.abs(y_true - y_pred)) / mae_naive

    smape_val = smape(y_raw.values, preds)
    mase_val = mase(y_raw.values, preds)

    # --- Importanza e segno
    importances = model.feature_importances_
    signs = [np.sign(np.corrcoef(X[col], preds)[0, 1]) if not np.isnan(np.corrcoef(X[col],
preds)[0, 1]) else 0 for col in X.columns]

    imp_df = pd.DataFrame({
        "Variabile": X.columns,
        "Importanza": importances,
        "Segno": signs
    })
    imp_df["Direzione"] = imp_df["Segno"].map({1: "Positiva", -1: "Negativa", 0: "N/A"})
    imp_df["Importanza firmata"] = imp_df["Importanza"] * imp_df["Segno"]
    imp_df = imp_df.sort_values(by="Importanza", ascending=False).reset_index(drop=True)

    # --- Grafico
    if plot:
        sns.set_theme(style="whitegrid")
        plt.figure(figsize=(12, 5))
        top10 = imp_df.head(10)
        colors = ["#5CB85C" if s > 0 else "#D9534F" if s < 0 else "#AAAAAA" for s in
top10["Segno"]]
        bars = plt.barh(top10["Variabile"], top10["Importanza firmata"], color=colors,
alpha=0.85, edgecolor="black")
        plt.axvline(x=0, color="gray", linestyle="--")
```

```
        plt.gca().invert_yaxis()
        plt.title(f"{model_type.upper()} - Top 10 variabili più influenti", fontsize=13,
weight="bold")
        for bar, val in zip(bars, top10["Importanza firmata"]):
            xpos = val + (0.002 if val > 0 else -0.002)
            plt.text(xpos, bar.get_y() + bar.get_height()/2, f"{abs(val):.3f}",
va="center", ha="left" if val > 0 else "right", fontsize=7)
        plt.tight_layout()
        plt.show()

    metrics_df = pd.DataFrame([{
        "Modello": model_type.upper(),
        "R²_train": round(r2, 3),
        "R²_adj": round(r2_adj, 3),
        "R²_CV_mean": round(mean_r2, 3),
        "R²_CV_std": round(std_r2, 3),
        "RMSE": round(rmse, 3),
        "MAE": round(mae, 3),
        "SMAPE (%)": round(smape_val, 2),
        "MASE": round(mase_val, 3)
    }])

    print(f"\n {model_type.upper()} — BASE")
    print(f"R² = {r2:.3f} | CV = {mean_r2:.3f} ± {std_r2:.3f}")
    print(f"RMSE = {rmse:.3f} | MAE = {mae:.3f} | SMAPE = {smape_val:.2f}% | MASE =
{mase_val:.3f}")

    return imp_df, metrics_df, model
```

Ensemble methods diagnostics

```
def check_ensemble_report(
    df,
    target,
    model,
    model_name="Ensemble (Auto)",
    k=10,
    plot=True,
    export_excel=False,
    random_state=42,
    log_target=False
):
    """
    Diagnostica coerente con ensemble_methods_auto/v2:
    - Calcola R², RMSE, MAE, SMAPE, MASE
    - Se log_target=True, riconverte le predizioni
    - Grafici affiancati: residui, distribuzione, trend LOWESS
    """

    X = df.drop(columns=[target])
    y = df[target]
    y_pred = model.predict(X)

    if log_target:
        y_pred = np.expm1(y_pred)

    residuals = y - y_pred

    # --- Metriche
    r2 = r2_score(y, y_pred)
    rmse = np.sqrt(mean_squared_error(y, y_pred))
    mae = mean_absolute_error(y, y_pred)

    def smape(y_true, y_pred):
```

```
        denom = (np.abs(y_true) + np.abs(y_pred))
        denom[denom == 0] = np.nan
        return 100 * np.nanmean(2 * np.abs(y_true - y_pred) / denom)

    def mase(y_true, y_pred):
        naive_forecast = np.roll(y_true, 1)[1:]
        mae_naive = np.mean(np.abs(y_true[1:] - naive_forecast))
        return np.mean(np.abs(y_true - y_pred)) / mae_naive

    smape_val = smape(y.values, y_pred)
    mase_val = mase(y.values, y_pred)

    # --- Cross-validation
    kf = KFold(n_splits=k, shuffle=True, random_state=random_state)
    cv_r2 = cross_val_score(model, X, y, cv=kf, scoring="r2")
    mean_r2, std_r2 = cv_r2.mean(), cv_r2.std()

    # --- Grafici
    if plot:
        sns.set_theme(style="whitegrid")
        fig, axes = plt.subplots(1, 3, figsize=(18, 5))
        plt.suptitle(f"Diagnostica modello: {model_name}", fontsize=15, weight="bold")

        sns.scatterplot(x=y_pred, y=residuals, color="#2563EB", s=60, alpha=0.6,
ax=axes[0])
        axes[0].axhline(0, color="#F97316", linestyle="--", lw=2)
        axes[0].set_title("Residuals vs predicted")
        axes[0].set_xlabel("Predicted")
        axes[0].set_ylabel("Residals")

        sns.histplot(residuals, kde=True, color="#10B981", bins=25, alpha=0.7, ax=axes[1])
        axes[1].axvline(0, color="gray", linestyle="--", lw=1)
        axes[1].set_title("Residuals distribution")

        sns.regplot(x=y_pred, y=y, lowess=True,
                    scatter_kws={'color': "#2563EB", 'alpha': 0.6, 's': 60},
                    line_kws={'color': "#F97316", 'lw': 2}, ax=axes[2])
        axes[2].plot([y.min(), y.max()], [y.min(), y.max()], 'k--', lw=1)
        axes[2].set_title("Trend LOWESS")
        axes[2].set_xlabel("Predicted")
        axes[2].set_ylabel("Observed")

        plt.tight_layout(rect=[0, 0, 1, 0.95])
        plt.show()

    # --- Diagnosi overfitting
    overfit_flag = (r2 - mean_r2) > 0.1
    pattern_flag = np.abs(pd.Series(residuals).rolling(5).mean().std()) > 0.5 *
np.std(residuals)
    if overfit_flag:
        decision = "Possibile overfitting: R² train >> R² CV."
    elif pattern_flag:
        decision = "Residui con struttura → mancano effetti o interazioni."
    else:
        decision = "Il modello mostra buona generalizzazione."

    report = pd.DataFrame([{
        "Modello": model_name,
        "R²_train": round(r2, 3),
        "R²_CV_mean": round(mean_r2, 3),
        "R²_CV_std": round(std_r2, 3),
        "RMSE": round(rmse, 3),
        "MAE": round(mae, 3),
        "SMAPE (%)": round(smape_val, 2),
        "MASE": round(mase_val, 3),
```

```
        "Overfitting": overfit_flag,
        "Residui con pattern": pattern_flag,
        "Conclusione": decision
    }])

    display(report)
    print(f"\n{decision}")

    if export_excel:
        fname = f"{model_name.lower().replace(' ', '_')}_report.xlsx"
        report.to_excel(fname, index=False)
        print(f"Report salvato come '{fname}'")

    return report
```

Feature importance graphical representation

```
def plot_ensemble(imp_df, model_name="Modello Ensemble", save_path=None, figsize=(10,6)):
    """
    Crea un grafico a barre orizzontali con linea centrale (0)
    per rappresentare tutte le variabili di un modello ensemble,
    distinguendole per colore.

    Parametri
    ----------
    imp_df : pd.DataFrame
        DataFrame restituito dalla funzione `ensemble_feature_analysis()`
        contenente almeno le colonne:
        ['Variabile', 'Importanza firmata', 'Segno'].
    model_name : str, default='Modello Ensemble'
        Titolo del grafico (es. 'Random Forest', 'Gradient Boosting').
    save_path : str, opzionale
        Se specificato, salva il grafico in formato PNG nel percorso indicato.
    figsize : tuple, default=(10,6)
        Dimensione del grafico.

    Output
    -------
    Mostra il grafico e lo salva (opzionale).
    """

    # --- Controllo colonne
    required_cols = {"Variabile", "Importanza firmata", "Segno"}
    if not required_cols.issubset(imp_df.columns):
        raise ValueError(f"Il DataFrame deve contenere le colonne: {required_cols}")

    df_plot = imp_df.copy().sort_values(by="Importanza", ascending=True)

    # --- Colori differenti per ogni variabile
    palette = sns.color_palette("husl", n_colors=len(df_plot))

    # --- Setup grafico
    sns.set_theme(style="whitegrid")
    plt.figure(figsize=figsize)

    bars = plt.barh(
        df_plot["Variabile"],
        df_plot["Importanza firmata"],
        color=palette,
        edgecolor="black",
        alpha=0.9
    )

    # Linea centrale
```

```
plt.axvline(x=0, color="gray", linestyle="--", linewidth=1.1)

# --- Etichette ed estetica
plt.title(f"{model_name} – Feature importance (with +/- direction)",
          fontsize=13, weight="bold", pad=12)
#plt.xlabel("Feature importance (with +/- direction)", fontsize=11)
plt.ylabel("")
plt.grid(axis='x', linestyle='--', alpha=0.4)
sns.despine(left=True, bottom=True)
plt.tight_layout()

# --- Etichette numeriche ai margini
for bar, val in zip(bars, df_plot["Importanza firmata"]):
    xpos = val + (0.002 if val > 0 else -0.002)
    ha = "left" if val > 0 else "right"
    plt.text(
        xpos,
        bar.get_y() + bar.get_height() / 2,
        f"{abs(val):.3f}",
        va="center",
        ha=ha,
        fontsize=8.5,
        color="black"
    )

plt.show()
```

8 Appendix B – Survey results

Appendix B presents the results of the survey used in Chapter 4.4.1 *Variable weights definition* to define the weights of the variables included in the *Station Static Effectiveness Index*. It should be noted that this survey was proposed to bike-sharing users and non-users in November 2025, collecting a total of 127 responses.

Sample description and respondents' profile

This part of the survey aims to understand the characteristics of the sample studied, through socio-demographic attributes. The results are presented in Table 59.

Table 59. Sample socio-demographic attributes.

Question	Results
Respondents' age	1. 39.8% 25-34 years; 2. 31.3% 18-24 years; 3. 15.6% 35-49 years; 4. 9.4% 50-64 years; 5. 2.3% < 18 years; 6. 1.6% > 65 years.
Respondents' gender	1. 53.1% male; 2. 45.3% female; 3. 1.6% not specified.
Respondents' occupation	1. 41.4% workers over 26 years; 2. 41.4% university students; 3. 11.7% workers under 26 years; 4. 3.1% other; 5. 2.6% High-school students.
Respondents' country	1. 85.7% Italy; 2. 7.1% Asia; 3. 7.1% other European countries.
City size where respondents usually use bike-sharing	1. 54.8% > 1,000,000 citizens; 2. 21.4% currently don't use BS; 3. 9.5% > 500,000 citizens; 4. 7.1% > 10,000,000 citizens; 5. 4.8% < 50,000 citizens; 6. 2,4% > 100.000 citizens.

Bike-sharing drivers demand

This section is the core of the survey. For each of the previously selected variables, respondents were asked to assign a score from 1 to 5 based on their degree of attraction. Table 60 and Table 61 show the results obtained, specifying the percentage of respondents for each of the grades.

Table 60. BS demand drivers survey results.

Variable	Results
Distance from city centre	1. 8.6% 2. 13.3% 3. 31.3% 4. 29.7% 5. 17.2%
Number of accessible transport lines	1. 6.3% 2. 9.4% 3. 21.9% 4. 32.0% 5. 30.5%
Distance from the nearest bike lane	1. 8.6% 2. 14.1% 3. 23.4% 4. 28.1% 5. 25.8%
Distance from other BS stations	1. 12.5% 2. 13.3% 3. 29.7% 4. 25.0% 5. 19.5%
Distance from the nearest car park	1. 35.2% 2. 24.2% 3. 16.4% 4. 14.8% 5. 9.4%
Housing density in the neighbourhood	1. 20.3% 2. 18.0% 3. 31.3% 4. 17.2% 5. 13.3%

Table 61. Cont. BS demand drivers survey results.

Variable	Results
Proximity to educational Points of Interest (university buildings, schools, libraries)	1. 10.9% 2. 16.4% 3. 21.9% 4. 32.0% 5. 18.8%
Proximity to touristic Points of Interest (Hotels, museums, squares, churches…)	1. 12.5% 2. 15.6% 3. 28.1% 4. 25.0% 5. 18.8%
Proximity to personal needs Points of Interest (Restaurants, cafes, shops, healthcare facilities…)	1. 3.9% 2. 13.3% 3. 28.9% 4. 34.4% 5. 19.5%
Proximity to free time Points of Interest (Parks, gyms, sports centres…)	1. 8.6% 2. 7.0% 3. 24.2% 4. 39.8% 5. 20.3%

Bibliography

[1] Shared-use mobility center, “What is shared mobility?,” 2025. [Online]. Available: https://sharedusemobilitycenter.org/what-is-shared-mobility/. [Accessed 24 9 2025].

[2] “Ride-Hailing,” ScienceDirect, 2020. [Online]. Available: https://www.sciencedirect.com/topics/social-sciences/ride-hailing. [Accessed 24 9 2025].

[3] E. Croci and D. Rossi, “Optimizing the position of bike sharing stations. The Milan case.,” IEFE - The Center for Research on Energy and Environmental Economics and POlicy at Bocconi University, Milan, 2014.

[4] R. Meddin, “The Meddin bike-sharing world map,” 30 9 2025. [Online]. Available: https://bikesharingworldmap.com/index.php#/all/2.3/0/51.5/. [Accessed 1 10 2025].

[5] Osservatorio nazionale sharing mobility, “8° Rapporto nazionale sulla sharing mobility,” Osservatorio nazionale sharing mobility, Roma, 2024.

[6] S. Shaheen, S. Guzman and H. Zhang, “Bikesharing in Europe, the Americas, and Asia: Past, Present, and Future,” *Institute of Transportation Studies, UC Davis, Institute of Transportation Studies, Working Paper Series,* pp. 159-167, 2010.

[7] H. Zhenhua, L. Xin and h. Pan, “The Impact of Bike Sharing on Urban Transportation.,” *The Frontiers of Society, Science and Technology,* vol. 1, no. 4, pp. 288-298, 2019.

[8] P. Midgley, “Bicycle-sharing schemes: enhancing sustainable mobility in urban areas,” in *United nations department of economic and social affairs, commission on sustainable development*, New York, 2011.

[9] S. Maas, N. Paraskevas, M. Attard and L. Dimitriou, “Spatial and temporal analysis of shared bicycle use in Limassol, Cyprus,,” *Transportation Research procedia,* vol. 93, 2021.

[10] M. Dziecielski, A. Nikitas, A. Radzimski and B. Caulfield, “Understanding the determinants of bike-sharing demand in the context of a medium-sized car-oriented city: The case study of Milton Keynes, UK,” *Sustainable cities and society,* vol. 114, 2024.

[11] J. C. García-Palomares, J. Gutiérrez and M. Latorre, “Optimizing the location of stations in bike-sharing programs: A GIS approach,” *Applied geography,* no. 35, pp. 235-246, 2012.

[12] D. Duran, D. Villeneuve, F. Pereira and G. Wulfhorst, “How fair is the allocation of bike-sharing infrastructure? Framework for a qualitative and quantitative spatial fairness assessment,” *Transportation Research Part A General,* no. 140, pp. 299-319, 2020.

[13] T. Raviv, M. Tzur and I. A. Forma, “Static repositioning in a bike-sharing system: models and solution approaches.,” *Journal on transportation and logistics,* vol. 2, no. 3, pp. 187-229, 2013.

[14] D. Chemla, F. Meunier and R. Calvo, "Bike sharing systems: Solving the static rebalancing problem.," *Discrete optmization,* no. 10, pp. 120-146, 2013.

[15] I. Bhuyan, C. Chavis, A. Nickkar and P. Barnes, "GIS-Based Equity Gap Analysis: Case Study of Baltimore Bike Share Program," *Urban science,* vol. 3, no. 42, 2019.

[16] S. Albiński, P. Fontaine and S. Minner, "Performance analysis of a hybrid bike sharing system: A service-level-based approach under censored demand observations," *Transport research,* no. 116, pp. 59-69, 2018.

[17] M. Nogal and P. Jiménez, "Attractiveness of Bike-Sharing Stations from a Multi-Modal Perspective: The Role of Objective and Subjective Features," *Sustainability,* vol. 12, no. 9062, 2020.

[18] "ResearchGate," [Online]. Available: https://www.researchgate.net/. [Accessed 14 12 2025].

[19] "Elsevier," [Online]. Available: https://www.eu.elsevierhealth.com/. [Accessed 14 12 2025].

[20] D. Harkey and D. K. M. Reinfurt, "Development of the Bicycle Compatibility Index," *Transportation research,* no. 1636, pp. 13-20, 1998.

[21] P. DeMaio, "Bike-sharing: History, Impacts, Model of Provision, and Future," *Journal of Public Transportation,* vol. 12, no. 4, pp. 41-56, 2009.

[22] J. Larsen and A. M. El-Geneidy, "Build It. But Where? The Use of Geographic Information Systems in Identifying Locations for New Cycling Infrastructure," *International Journal of Sustainable Transportation,* vol. 7, no. 10, 2013.

[23] E. Papa and P. Coppola, "Gravity-Based Accessibility measures for Integrated Transport-land Use Planning (GRABAM)," in *Accessibility Instruments for Planning Practice*, COST office, 2012, pp. 117-124.

[24] A. Rixey, "Station-Level forecasting of Bike Sharing Ridership: Station Network Effects in Three U.S. Systems," in *TRB 2013 Annual Meeting*, 2012.

[25] G. Pizzolorusso, "Sviluppo di un indice di qualità ciclistica di strade e piste ciclabili urbane," Politecnico di Milano, Milan, 2013.

[26] H. Hff and R. Ligget, "The Highway Capacity Manuals Method for Calculating Bicycle and Pedestrian Levls of Service: the Ultimate White Paper," Lewis Center for Regional Policy Studies and Institute of Transportation Studies, University of California, Los Angeles, 2014.

[27] L. Chen, D. Zhang, G. Pan, X. Ma, D. Yang, K. Kushlev, W. Zhang and S. Li, "Bike Sharing Station Placement Leveraging Heterogeneous Urban Open Data," in *UBICOMP '15*, Osaka, Japan, 2015.

[28] I. Frade and A. Riberiro, "Bike-sharing stations: A maximal covering location approach," *Transportation research part A,* no. 82, pp. 216-227, 2015.

[29] I. Mateo-Babiano, R. Bean, J. Corcoran and D. Pojani, "How Does Our Natural and Built Environment Affect The Use of Bicycle Sharing?," *Transportation Research Part A: Policy and Practice,* vol. 94, pp. 295-307, 2016.

[30] A. Munkacsy and A. Monzòn, "Potential User Profiles of Innovative Bike-Sharing Systems: The Case of BiciMAD," *Asian Transport Studies,* vol. 4, no. 3, pp. 621-638, 2017.

[31] C. Kloimüllner and G. Raidl, "Hierarchical Clustering and Multilevel Refinement for the Bike-Sharing Station Planning Problem," in *International Conference on Learning and Intelligent Optimization*, 2017.

[32] F. Soriguera, V. Casado and E. Jiménez Meroño, “A simulation model for public bike-sharing systems,” *Transportation Research Procedia,* vol. 33, pp. 139-146, 2018.

[33] R. Nath and T. Rambha, “Modelling Methods for Planning and Operation of Bike-Sharing Systems,” *Journal of the Indian Institute of Science,* vol. 99, pp. 1-26, 2019.

[34] Q. Liu, R. Homma and K. Iki, “Utilizing Bicycle Compatibility Index and Bicycle Level of Service for Cycleway networks,” in *MATEC Web of Conferences*, 2019.

[35] H. Chen, T. Cheng and Y. Zhang, “Locating stations in bike-sharing service: a special maximal covering location problem.,” in *Geographical Information Science Research- UK*, 2019.

[36] Y. Fang, Y. Song, D. Chen, P. Wu and F. Chu, “A location-routing problem for the public bike-sharing system with service level,” in *International Conference on Industrial Engineering and Systems Management (IESM)*, 2019.

[37] S. Mingzhu, K. Wang, Y. Zhang, M. Li and H. Qi, “Impact Evaluation of Bike-Sharing on Bicycling Accessibility,” *Sustainability,* vol. 12, 2020.

[38] K. Kazemzadeh, A. Laureshyn, L. Winslott Hiselius and E. Ronchi, “Expanding the Scope of the Bicycle Level-of-Service concept: a review of the literature,” *Sustainability,* vol. 12, 2020.

[39] V. Torrisi, M. Ignaccolo, G. Inturri, G. Tesoriere and T. Campisi, “Exploring the factors affecting bike-sharing demand: evidence from student perceptions, usage patterns and adoption barriers,” *Transportation research procedia,* no. 52, pp. 573-580, 2021.

[40] F. Soriguera and E. Jiménez-merono, “A continuous approximation model for the optimal design of public bike-sharing systems,” *Sustainable Cities and Society,* vol. 52, 2020.

[41] H. Kong, S. Jin and D. Sui, “Deciphering the relationship between bikesharing and public transit: Modal substitution, integration, and complementation,” *Transportation Research Part D: Transport and Environment,* vol. 85, 2020.

[42] A. Nikiforiadis, G. Aifadopoulou, J. M. Salanova Grau and N. Boufidis, “Determining the optimal locations for bike-sharing stations: methodological approach and application in the city of Thessaloniki, Greece,” *Transportation Research Procedia,* vol. 52, pp. 557-564, 2021.

[43] M. Kamel, T. Sayed and A. Bigazzi, “A composite zonal index for biking attractiveness and safety,” *Accident; analysis and prevention,* vol. 137, 2020.

[44] L. Caggiani and R. Camporeale, “Accessibility indicators for fair bike-sharing systems based on level of service,” in *IEEE International Conference on Environment and Electrical Engineering*, Bari, 2021.

[45] J. Beairsto, Y. Tian, L. Zheng, Q. Zhao and J. Hong, “Identifying locations for new bike-sharing stations in Glasgow: an analysis of spatial equity and demand factors,” *Annals of GIS,* vol. 28, no. 2, pp. 111-126, 2022.

[46] K. Kim, “Investigation of modal integration of bike-sharing and public transit in Seoul for the holders of 365-day passes,” *Journal of transport geography,* no. 106, 2023.

[47] C.-C. Hsu, Y.-W. Kuo and J. Liou, “A Hybrid Model for Evaluating the Bikeability of Urban Bicycle Systems,” *Axioms,* vol. 12, no. 155, 2023.

[48] G. Ceccarelli, G. Cantelmo, M. Nigro and C. Antoniou, “Learning from Imbalanced Datasets: The Bike-Sharing Inventory Problem Using Sparse Information,” *Algorithms,* vol. 16, no. 351, 2023.

[49] J. Song, B. Li, W. Szeto and X. Zhan, "A station location design problem in a bike-sharing system with both conventional and electric shared bikes considering bike users' roaming delay costs," *Transportation Research Part E: Logistics and Transportation Review,* vol. 181, 2024.

[50] S. L. Grüner and M. Kowald, "Bike-sharing, why not? A framework of utility perceptions of BSSs' non-users based on qualitative data," *Transport Policy,* vol. 168, pp. 220-243, 2025.

[51] O. Roig-Costa, C. Miralles-Guasch and O. Marquet, "Unpacking the docked bike-sharing experience. A bike-along study on the infrastructural constraints and determinants of everyday bike-sharing use," *Journal of transport geography,* no. 125, 2025.

[52] ISTAT, "Basi territoriali e variabili censoriali," October 2025. [Online]. Available: https://www.istat.it/notizia/basi-territoriali-e-variabili-censuarie/. [Accessed October 2025].

[53] D. Witten, G. James, R. Tibshirani and T. Hastie, An Introduction to Statistical Learning: with Applications in R, Springer, 2013.

[54] Transportation Research board, "Highway Capacity Manual 2010," National Research Council, Washington DC, 2010.

[55] Technische Universität Dresden, "Verkehrsökologische Schriftenreihe," 12 4 2016. [Online]. Available: https://tud.qucosa.de/landing-page/?tx_dlf[id]=https%3A%2F%2Ftud.qucosa.de%2Fapi%2Fqucosa%253A29431%2Fmets. [Accessed 2 10 2025].

[56] M. J. Zaki and W. Meira Jr., Data Mining and Machine Learning: Fundamental Concepts and Algorithms, 2nd edition, Cambridge university press, 2020.

[57] IBM, "What is overfitting?," IBM, 2021. [Online]. Available: https://www.ibm.com/think/topics/overfitting. [Accessed 3 October 2025].

[58] B. Beura, L. Veera, C. Haritha and B. Prasanta, "Defining Bicycle Levels of Service Criteria Using Levenberg–Marquardt and Self-organizing Map Algorithms," *Transportation in developing economies,* no. 4, 2018.

[59] L. Bai, P. Liu, C.-Y. Chan and Z. Li, "Estimating level of service of mid-block bicycle lanes considering mixed traffic flow," *Transportation Research part A: policy and practice,* vol. 101, pp. 203-217, 2017.

List of figures

List of tables