University of Padova

Department of Mathematics

Master Thesis in Data Science

The 15-Minute City for Tourists

Supervisor
Francesco Silvestri
University of Padova

Master Candidate
Tapia Montero, Mario Alejandro

Academic Year

2023-2024

Dedication.
To my incredibly sweet, beautiful, and beloved wife, Alejandra, for this incredible journey in life.

Abstract

This Thesis focuses on identifying the most walkable areas of a city for tourists, using a proposed Tourism Walkability Index within the framework of the 15-minute city concept. It examines Naples and Venice to assess how well hotels and BnBs are connected to well-reviewed nearby attractions and restaurants, using data from TripAdvisor, Airbnb, and OpenStreetMap. In both cities, hotels, which are more concentrated, tend to have better walkability than BnBs, which are more spread out. At the city level, the results show that Venice has more consistently distributed walkability, while Naples exhibits greater variability, with a higher proportion of accommodations having high walkability scores. The relationship between walkability and accommodation pricing is also analyzed, revealing that accommodations with higher walkability tend to have higher prices. This analysis was conducted using multiple linear regression and regression trees. The study highlights the economic and experiential benefits of walkable environments for both cities and tourists, respectively.

Contents

Listing of figures

Listing of tables

Listing of acronyms

GDP .... Gross Domestic Product

OLS .... Ordinary Least Squares

OSM .... OpenStreetMap

POI .... Point(s) Of Interest

TWI .... Tourism Walkability Index

VIF .... Variance Inflation Factor

1
Introduction

Tourism is a significant component of the economic activity in many cities. It is defined as a social, cultural, and economic phenomenon that involves the movement of people to destinations outside their usual environment. Concurrently, there is a growing trend toward the implementation of the concept of 15-minute cities, which aims to ensure that residents can access essential amenities such as grocery stores, banks, and pharmacies within a short walking distance from their homes. Developing neighborhoods with this approach offers benefits in terms of reduced pollution and improved public health, making cities more livable and sustainable.

This Thesis explores the intersection of these two concepts—tourism and the 15-minute city. The central question is how to identify the most walkable areas within cities, particularly considering the location of accommodations in relation to high-quality attractions and restaurants, while also examining the relationship between walkability and accommodation prices. Walkability is a key component of the 15-minute city concept and is crucial not only for residents but also for enhancing the experience of tourists, making it a significant factor in the desirability of a destination.

To approach this problem, the Thesis begins with a literature review in Chapter 2, which establishes the context and summarizes existing research on tourism and the concept of 15-minute cities. This review provides the theoretical foundation for the study, highlighting the importance of walkability in urban planning and tourism. Chapter 3 follows with a discussion of the essential concepts and tools that underpin the methodology.

Chapter 4 presents the methodology used to evaluate the walkability around hotels and BnBs and its influence on accommodation pricing. For this, data on points of interest and accommodations is gathered through web scraping. A ranking of the most attractive points of interest is then created, and accessibility from each accommodation within a walkable distance is assessed using shortest-path calculations. By calculating a walkability index and analyzing its relationship to accommodation prices, the study provides insights into how hotels and BnBs are distributed across the city according to their walkability level, and how this impacts accommodation pricing.

Chapter 5 presents the experiments conducted and the results obtained for the cities of Naples and Venice in Italy, applying the methodology previously mentioned. Finally, the conclusions and future work are discussed in Chapter 6, where the findings provide a framework for diagnosing and comparing cities based on their walkability. This framework allows for assessing the attractiveness of different urban areas, as well as the city as a whole, from a tourism perspective, aligning with the principles of the 15-minute city to promote more sustainable and enjoyable urban environments.

1.1 Objectives

The overall objective of this Thesis is to determine which areas of a city are attractive to tourists from a walkability perspective. To achieve this, the Thesis will investigate the relationship between the location of tourist accommodations and their proximity to points of interest and assess how these factors relate to accommodation pricing.

The particular objectives are the following:

- To determine whether well-connected, central accommodations with high accessibility (t-minutes) to well-reviewed points of interest like restaurants, attractions, and metro stations, have higher accommodation prices compared to less connected ones.
- To develop a metric or index that captures these variables and assigns scores to different accommodations.
- To identify the attractiveness of different city areas for tourists based on the previous concepts.

2
Literature Review

In this chapter, the concepts of Tourism, Walking, and 15-minute cities are presented and discussed. These concepts represent the foundational elements of the research conducted in this Thesis. Furthermore, some prior studies that integrate these concepts are also briefly reviewed.

2.1 The importance of Tourism

Tourism can be defined as a social, cultural, and economic phenomenon related to the movement of people to places outside their usual place of residence, pleasure being the usual motivation [1]. It is a form of productive consumption that connects people, organizations, material objects, environments, and technologies [2].

Its impact in economic growth has been examined in various regions, including India [3], Spain [4], OIC countries [5], and Italy [6].

In this context, Figure 2.1 illustrates the proportion of travel and tourism's overall contribution to the global GDP in 2019 and 2022, along with projections for 2023 and 2033.

It is evident from the data that this sector plays a significant role in the economies of various countries. Its relevance is further emphasized by the rising trend in international tourist arrivals over the years, as depicted in Figure 2.2. Since 2005, Europe has emerged as the most popular destination for tourists, as shown in Figure 2.3.

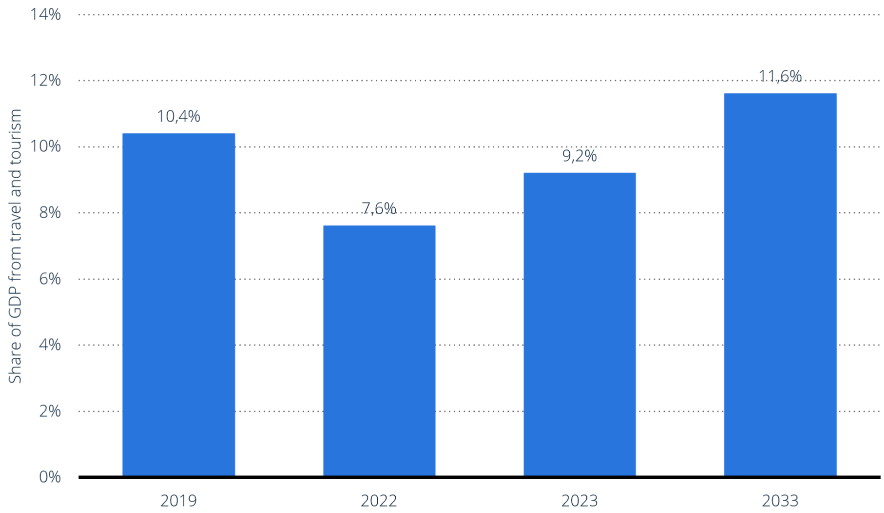


Figure 2.1: Share of travel and tourism's total contribution to GDP worldwide in 2019 and 2022, with a forecast for 2023 and 2033 [7].

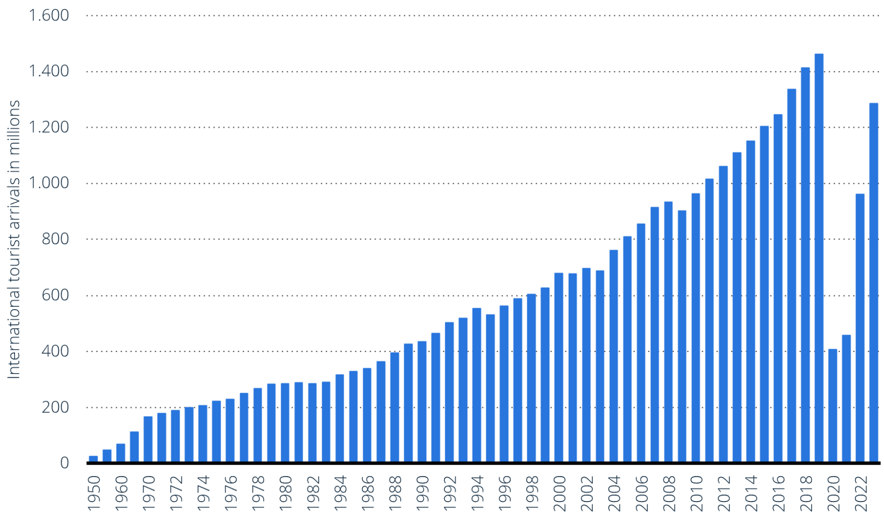


Figure 2.2: Number of international tourist arrivals worldwide from 1950 to 2023 (in millions) [8].

In this work, tourism is viewed as an activity that comes with a unique set of needs that are distinct from those of residents. This distinction is crucial for accurately defining the methodology, which will be elaborated on later.

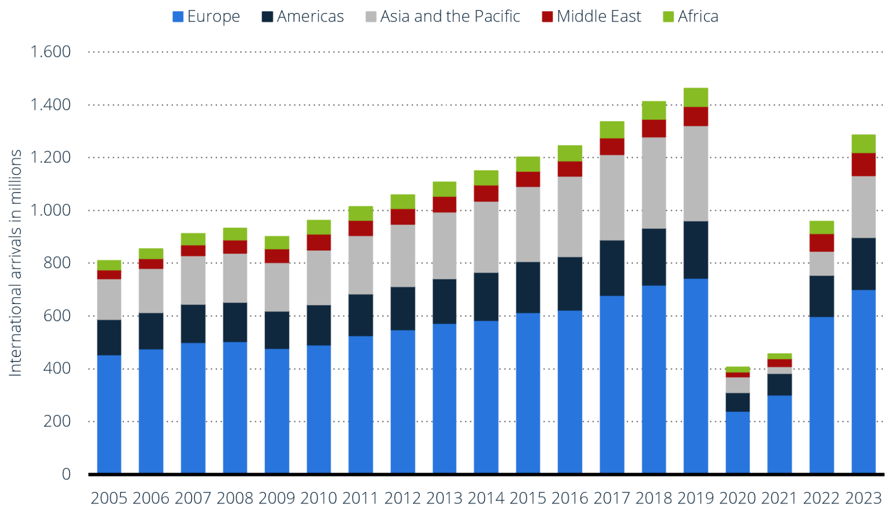


Figure 2.3: Number of international tourist arrivals worldwide from 2005 to 2023, by region (in millions) [9].

2.2 Walking as a Sustainable and Inclusive Mode of Transport

The historical growth of cities and advancements in transport technologies have given rise to new forms of mobility. These can cover larger distances, thereby maintaining the coherence of the expanding city. The modes of transport, ranging from horsecars, bicycles, streetcars, to rails, buses, and automobiles, provide to the diverse needs of urban trips. However, this approach to transport planning aimed at minimizing travel time for individual motorized transport is unsustainable, as it contributes to environmental and health degradation, and is heavily reliant on non-renewable energy sources [10].

The economic stress of auto mobility alone amounts to €500 billion annually in the EU. This includes accidents, air pollution, climate change, noise, and congestion, as well as other external costs linked to up and downstream processes, i.e. energy, vehicle, and infrastructure production [11]. The EU Urban Mobility Framework is actively combating these issues by striving to enhance the quality of life for urban populations within the EU [12]. This is achieved by addressing various urban mobility challenges such as air pollution, congestion, accessibility, urban road safety, and the growth of e-commerce. The framework also aims to increase the prevalence of sustainable transport modes, particularly public transport and active mobility. In this contemporary perspective, all modes of transport coexist in a hierarchy that gives precedence to walking, cycling, and public transport over other forms of transport [10, 13]. As per [11], walking and cycling yield an annual societal benefit worth €90 billion due to their positive health impacts. Moreover, walking is considered the most inclusive form of mobility because it does not require any means of transport [14, 15], it enhances safety [15], and is also a profoundly social activity [16] enabling, in comparison to other modes of transport, a greater diversity of social interactions [14].

According to [17, 18, 19, 20], the distribution of human mobility trips is heavy tailed, meaning that most of the time people travel only over short distances, occasionally taking longer trips. These short distances are dominated by walking or bus travel modes (see Figure 2.4). Moreover, approximately 83% of all trips are short, for non-work purposes, and occur relatively close to home [21]. Medium-distance trips are optimally covered by bicycle or car. Long-distance trips are preferred to be taken by airplane [22].

Walking, as a mode of transportation, aligns perfectly with the global shift towards sustainable practices. It is a zero-emission activity that contributes to reducing our carbon footprint, thus playing a crucial role in mitigating the effects of climate change. Moreover, it promotes health and well-being, offering physical benefits such as reduced all-cause mortality and better mental health [15].

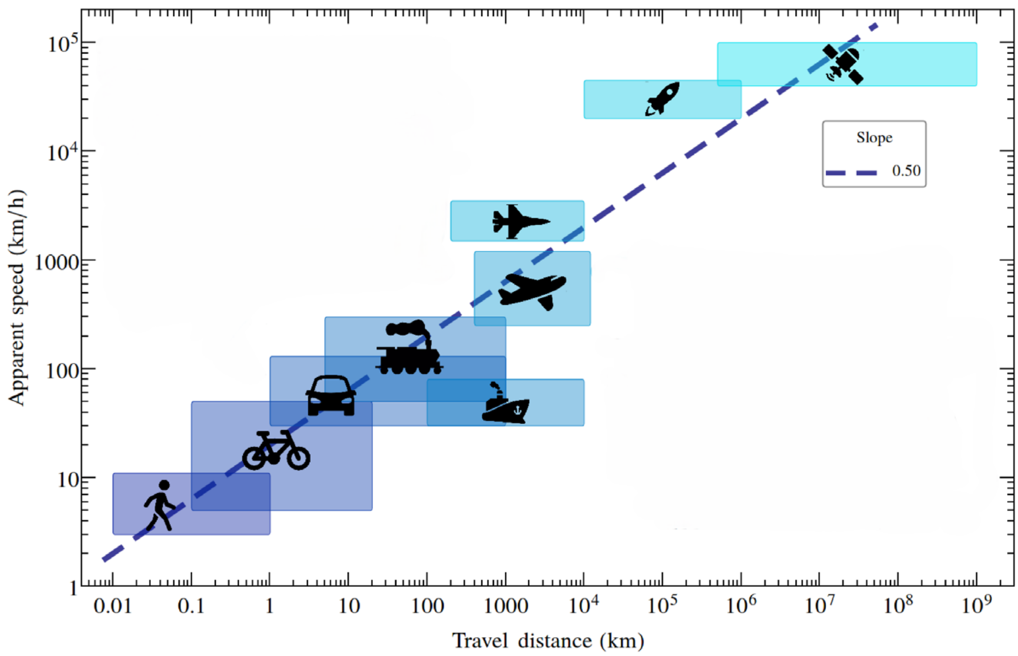


Figure 2.4: Velocity and distance scales of human travel (adapted from [22]).

It is worth noting that various authors use different walking speeds for their computations. In references [23, 24], the authors used a speed of 1 m/s, while in [14], a speed of 1.39 m/s was used. Reference [25] reported an average walking speed of 1.22 m/s for all pedestrians. Additionally, they cited another author who found a speed of 1.51 m/s for pedestrians under 65 years old and 1.25 m/s for those aged 65 and above. For a more comprehensive analysis of walking speed distribution, considering factors such as age, gender, and traffic control conditions, refer to [26].

2.3 Global Implementations of Proximity-Based Urban Planning

There is a heavy relation between the objective of improving sustainability and proximity [24]. Instead of seeking fast ways to travel long distances to reach the services they need, citizens should find those services close to them. In such a perspective, fast mobility would be replaced by proximity [14]. This vision takes concrete form in the so-called 15-minute city, which focuses on the pedestrian walkable distance from their home to nearby urban services and spaces [24] in no more than fifteen minutes. The range of services, as determined by the researchers' criteria, can be classified as belonging to the following categories: living, working, commerce, healthcare, education, and entertainment [27].

In recent years, the concept of the 15-minute city has been gaining popularity and has been implemented in major urban areas. This idea has particularly grabbed attention due to its successful implementation as a policy in Paris, France, which has significantly contributed to its popularity [28].

Another example is the city of Melbourne, Australia. In 2017, a new urban plan (Plan Melbourne 2017–2050 [29]) was presented pursuing the aim of shaping the city into "20-minute neighborhoods", where residents can meet their daily (non-work) needs within a short, non-motorized, trip from home [30].

The 20-minute neighborhood has been promoted also by the city of Portland, USA. The City of Portland Bureau of Planning and Sustainability defines this as a place with convenient, safe, and pedestrian-oriented access to the places people need to go to and the services people use nearly every day: transit, shopping, healthy food, school, parks, among others (City of Portland Bureau of Planning and Sustainability, 2012 [31]).

Finally, the Greater Sydney Commission, the planning agency for the Sydney region, came up with the concept "30-Minute City", referring to a new plan that should make Sydney region more interconnected and should allow its citizens to reach one of the three main metropolitan centers in less than 30 minutes walking, bike riding or by public transport [14].

More specifically, in a 15-minute analysis, the time factor is frequently assessed using a buffer analysis, isochrone maps, or closest facility analysis [24]. A measure that indicates a city's coverage is determined by calculating the percentage of accessible areas within each considered time range, or by considering an index (e.g. Frank's walkability index, Walk Score, NEXI).

Several studies, such as [14] and [24], illustrate the examination and prospective utilization of this concept in various cities.

In the case of [14], the authors developed an index for measuring the level of local proximity to services by walking in the city of Ferrara, Italy, according to the principles of 15-minute cities. Based on this, they were able to identify which parts of the city were able to access services within a certain time range. The service categories considered were commerce and retail, education, entertainment, grocery stores, healthcare, postal and banking services, offices, public parks, and restaurants. Figure 2.5 shows the index distribution for some of the categories. Based on this, the authors were able to answer the question "how many people in the city of Ferrara live in a 15-minute area?", using the pie chart from Figure 2.6.

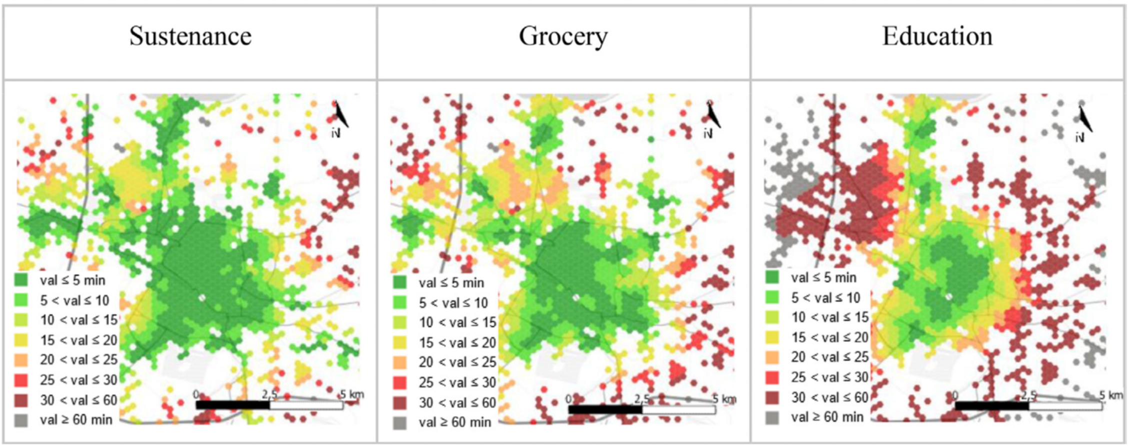


Figure 2.5: NEXI-Minutes index, categories comparison for the city of Ferrara [14].

In [24], authors explore the 15-minute city concept in a neighborhood in Parma, using a GIS-based method. The service categories considered were kindergarten and neighborhood

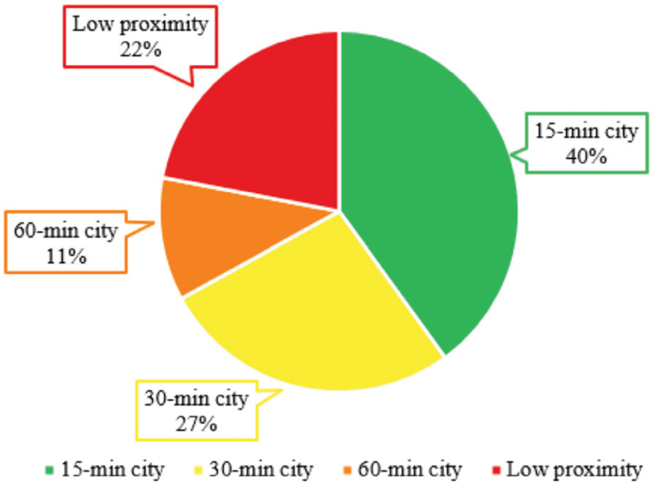


Figure 2.6: Percentage of population living in the different proximity levels in the city of Ferrara [14].

core (urban nodes well served by necessities shops and services, such as supermarkets, grocery stores, bars, drugstores, and banks). The results, in form of isochrone maps, can be seen in Figures 2.7 and 2.8. Specifically, authors measure the walkability of the neighborhood using the percentage of inhabitants that can access the services within the 15-minute time gap (e.g. more than 70% of the neighborhood have access to kindergartens, and 91% to neighborhood cores located in the same district within walking distance of 15-minute from the services).

The strengths of this study include the authors' consideration of a slower walking speed. They argue that this lower speed allows for a more inclusive approach to all road users. This is particularly relevant in certain urban areas, such as Italian cities, where there is a high population of older pedestrians. The authors also considered a delay factor that accounts for the impact on walking time of both signalized (with traffic lights) and unsignalized crossing paths, providing a more realistic approach. However, there are some concerns regarding privacy due to the way the authors present the population distribution, including the number of residents and residents of kindergarten age, across the neighborhood. This data was obtained from the official demographic registry of the Municipality of Parma. Additionally, the results could potentially be biased as they do not consider the proximity of services from adjacent neighborhoods that might be preferred by the residents.

As per the authors of the study on the 30-minute city [23], they estimate that pedestrians spend approximately 8 minutes waiting at traffic lights during a 30-minute walk. This results in a loss of 46% of their accessibility, as they can traverse significantly less area within that time

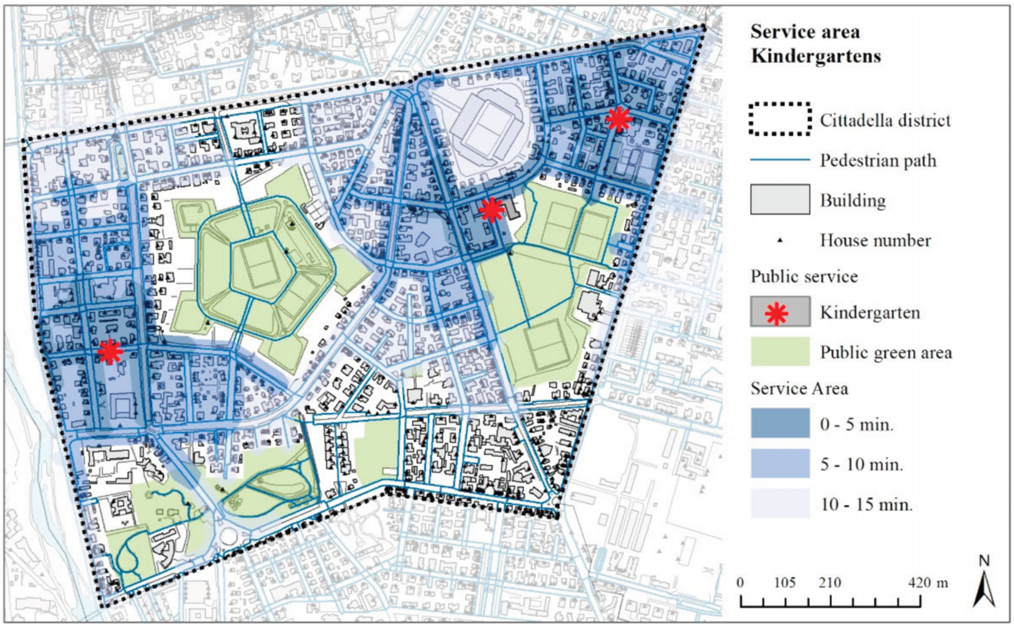


Figure 2.7: Service Area isochrone map concerning preschool facilities (kindergartens) [24].

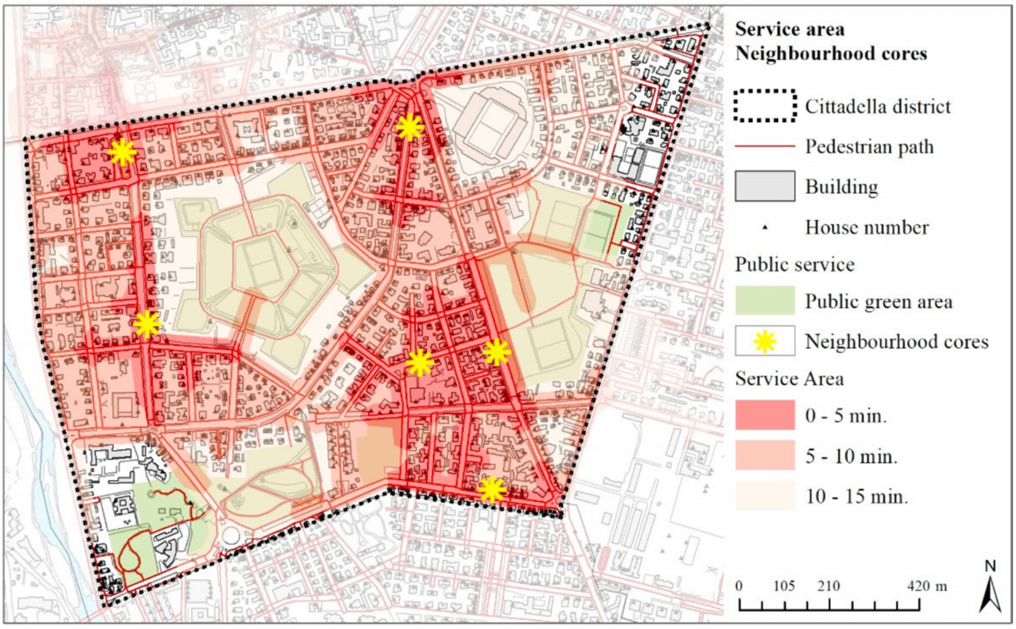


Figure 2.8: Service Area isochrone map concerning neighbourhood cores [24].

frame, making it an important variable to consider.

These two examples clearly demonstrate that the facilities used in the computation of the

15-minute city concept can differ based on the needs of the population and/or the nature of the analysis.

2.4 Walkable Urban Tourism: A 15-Minute Perspective

As presented in the preceding sections, there is a growing interest in making places more sustainable by enhancing their walkability and in encouraging walking as a means of active transport. Walking has also long been regarded as an important dimension of leisure and travel; a fundamental, universal, and significant tourist activity [32]. However, despite growth in research on walking in a residential context relatively little has been written on the contribution that walkability might make to tourism [28]. This is surprising given that walkable places are often considered as attractive for locals, visitors, and tourists alike, and as a measure of urban quality, with some studies suggesting that many tourists believe that the best way to experience a city is to walk it [28, 33], and others stating that tourists walk primarily to explore the urban spaces [34] as it increases the "imageability" of a place; the quality that makes it recognizable and memorable [15].

Walking encompasses time and rhythm, which, in the specific case of walking tourism, awakens the senses, facilitating a deeply situated spatio-temporal engagement. This allows travelers to engage more, not only with the host community, the surrounding environment, and the local culture but also with themselves and each other. Greater self-esteem, relaxation, freedom, and absence of stress are some of the feelings associated with the experience of walking tourism [35]. This positive experience might contribute to the results presented in [36], where it was found that nearly 60% of visitors in Munich primarily used public transportation, either independently or combined with walking. Furthermore, in [37] authors reported that international visitors in London were more likely to use walking as a means of transport than domestic tourists.

The key factors of tourism attractiveness have been identified in several studies:

- In [38], the movements of 557 tourists' daytrips tracked by GPS in Hong Kong revealed that the location of tourist accommodation is a major determinant of the spatial patterns of movements of tourists in urban destinations, with a large share of the total tourist time budget spent in the immediate vicinity of the hotel.

- In [39], a multiple criteria study and correlation with post purchase evaluation were used to identify the importance of several factors in Istanbul's hotels. Among these, authors found that the most important criterion of tourists' hotel location choice is the number of tourist attractions within walking distance of the hotel.

- In [40], authors carried out a study in Barcelona to compare spatial patterns of hotels and Airbnb listings, using census data and geolocated photographs from tourists. They found that Airbnb listings benefit more from a greater proximity to the city's main tourist attractions than does the hotel sector.

- In [41], authors analyze spatiotemporal patterns of tourism in Europe using conventional and big data sources (Bed-places, nights-spend, tourism demand, and accommodation location and capacity). The study reveals significant differences in the spatial distribution of tourism (see Figure 2.9), including sprawled patterns (London and Paris), clustered (Santorini), concentrated (Venice), and linear (Rimini), owing to local geography and typology of tourism.

- In [42], authors state that major visitor attractions stand out as the strongest pull-factors of a destination and are considered as key destination resources for development and marketing.

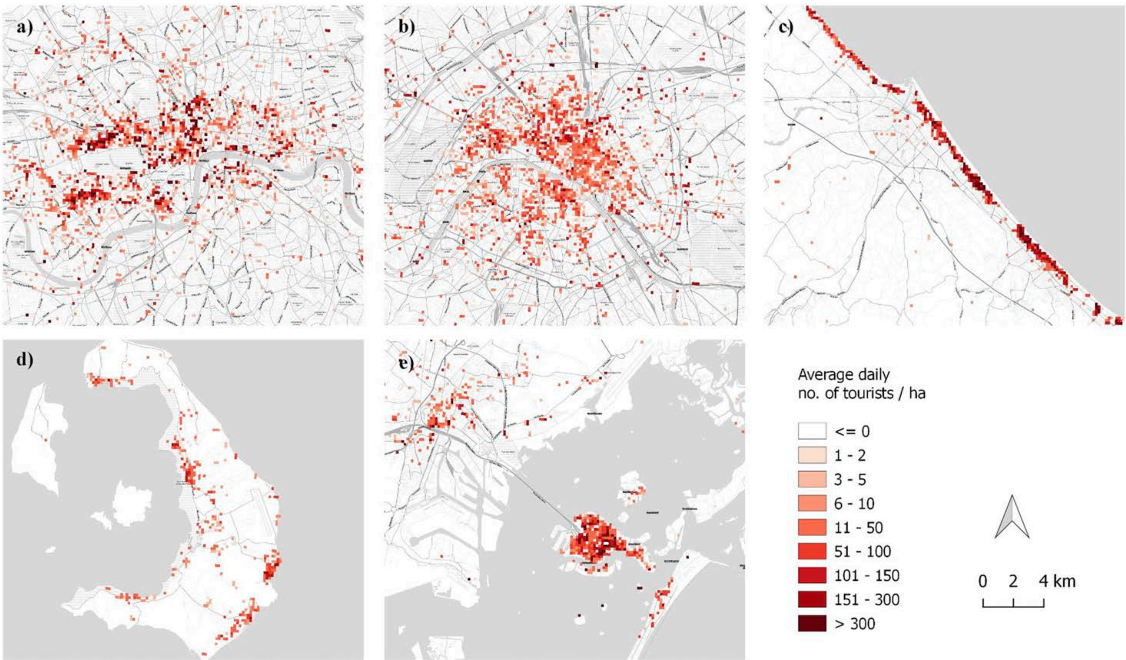


Figure 2.9: Tourist density in August 2016 in selected locations: a) London, b) Paris, c) Rimini, d) Santorini, e) Venice [41].

In summary, the key factors are the quantity and quality of attractions near lodging, given that tourists are inclined to spend more time around their accommodation. This pattern can be directly linked to the 15-minute cities concept, by substituting residents with tourists, and focusing on city attractions, restaurants, and metro stations rather than grocery stores, healthcare facilities, and banking services in a t-minute framework.

Despite the interesting results obtained in the previously mentioned studies, none of them focuses on walkability as it has been explored in the context of t-minute cities. Examples of the latter are the studies in [43] and [28].

In [43], authors analyzed 81 hotels and 97 Airbnb listings in the city of Tel Aviv, using Walk Score metric. The results showed that most accommodations were found in the walkable places of the city, even though the effect of Walk Score on prices and number of reviews is not linear or strong, as seen in Figures 2.10 and 2.11.

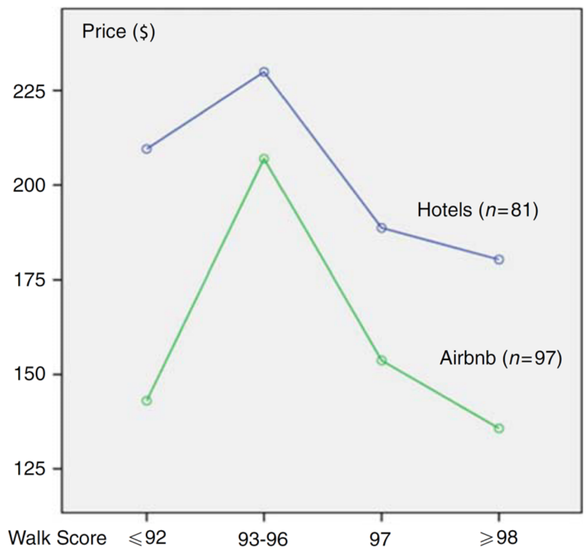


Figure 2.10: The effect of Walk Score on accommodation price [43].

In [28], authors focus on Walk Score and the top 330 tourist attractions from TripAdvisor ratings in the city of England. They observed only weak and nonlinear correlations between the Walk Score values and both the number of visitors and the number of reviews on TripAdvisor for the top attractions.

Even though these studies offer interesting insights into the relationship between walkability and tourism, they rely on the Walk Score metric. This metric assesses the walkability of any ad-

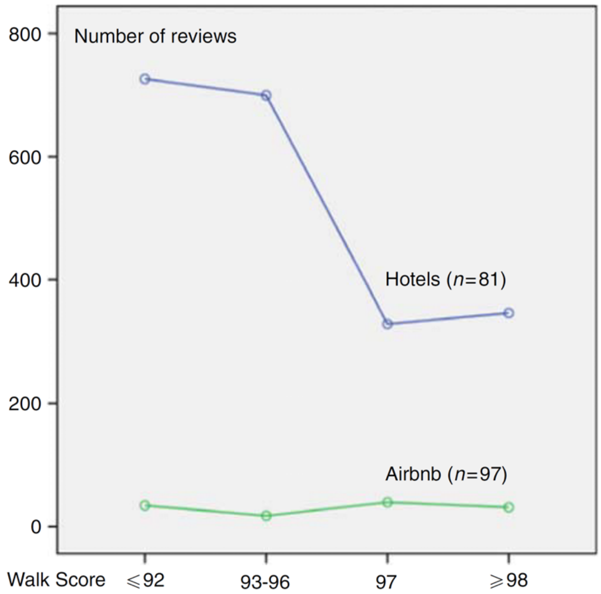


Figure 2.11: The effect of Walk Score on the number of reviews [43].

dress through a patented system [44], indicating that it is not feasible to obtain details regarding the specific variables employed in the calculation, nor the respective weights assigned to each. What is known is that the index assesses the walking potential of a place through a combination of three elements: the intersection density around the origin point, the block length, and the shortest distance to a group of preselected destinations. In [45], the authors test the validity and reliability of this metric for estimating access to objectively measured walkable amenities in Rhode Island. Significant correlations were identified between Walk Score and all categories of aggregated walkable amenities within a 1-mile buffer of the 379 random addresses, as can be seen in Figure 2.12. These categories correspond to:

- Food vendors (restaurants, coffee shops, bars)
- Grocery stores
- Parks
- Movie theaters
- Schools
- Libraries
- Fitness facilities

- Drug stores
- Retail (clothing, music, book, hardware stores)

Although these findings support Walk Score as a reliable and valid tool for estimating areas with a high density of walkable amenities, it is essentially designed as a metric for residents, not tourists. This is because the amenities included in its calculation are typically those needed for daily living. This focus might be the reason for the weak connections between variables identified in [43] and [28], which opens up the possibility to explore the concept of proximity between attractions and accommodations in a tourism context.

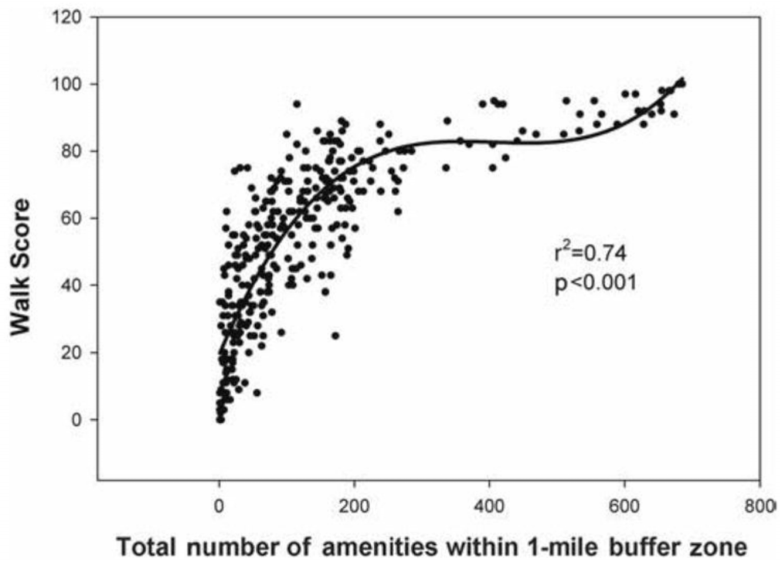


Figure 2.12: Association between Walk Scores and the total number of walkable amenities within a 1-mile buffer [45].

Walkability alone is not attractive to all tourists. But given the interest in encouraging visitors to engage in more sustainable forms of mobility it is something that needs to be investigated further [28].

3
Preliminaries

This Chapter covers the essential concepts and background information needed to understand the methodology discussed in Chapter 4.

3.1 Tools

In this section, the main tools used in the work are introduced. All coding is performed in Python, hence the libraries mentioned are specific to this programming language.

3.1.1 OpenStreetMap

OpenStreetMap is an open-source collaborative project that focuses on collecting and sharing high-quality spatial data worldwide. It operates as a user-generated geographic database of the world, following a peer production model similar to Wikipedia. The data collected through OSM is distributed as open data under the ODbL license and is widely reused across various sectors due to its accuracy, high update rate, global coverage, and uniformity of collected attributes. The platform is particularly beneficial for building and analyzing transportation networks, including street networks, subways, and railways, and can be combined to create multi-modal transport networks.

The core of a wiki-style process is that all users share an interest in maintaining accurate data. If inaccurate data is introduced, either maliciously or accidentally, the remaining majority of users have the ability to verify, correct, or eliminate it. The vast majority of well-intentioned participants can automatically compensate for the occasional outliers. In this regard, the accuracy of the outcomes derived in this Thesis is closely tied to the accuracy and reliability of the OSM data, particularly concerning the nodes and edges comprising the city's graph network [46, 47, 13, 14].

3.1.2 OSMnx and Pandana

Multiple tools have been developed to obtain and analyze data on transportation. One of the best known is OSMnx, a Python package that facilitates the downloading, modeling, analysis, and visualization of street networks and other geospatial features from OSM. It is built on top of NetworkX, another Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks [48]. OSMnx allows for the downloading and modeling of walking, driving, and biking networks. Particularly, it enables working with urban amenities/points of interest, building footprints, transit stops, elevation data, street orientations, speed/travel time, and routing with equal ease [49].

On the other hand, Pandana is a Python library for network analysis that uses contraction hierarchies to calculate super-fast travel accessibility metrics and shortest paths [50]. Figure 3.1 presents an overview of the main open-source analysis tools related to networks and transport.

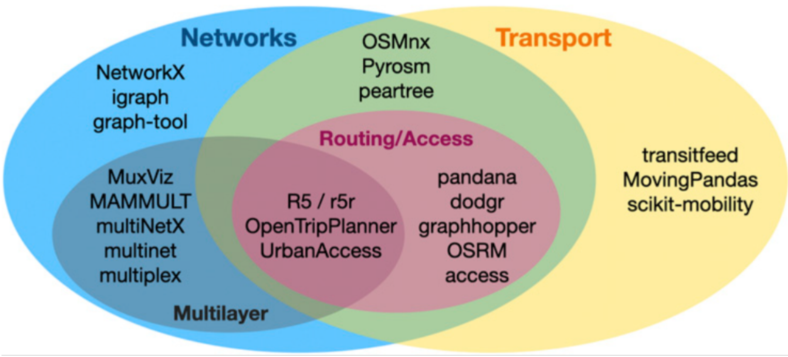


Figure 3.1: Overview of the main open-source analysis tools related to networks and transport [13].

3.1.3 Selenium

Selenium is a versatile tool for automating web browsers, commonly used for testing web applications to ensure functionality across different browsers and platforms. It supports multiple programming languages and includes components like Web-Driver for browser automation, Selenium IDE for script recording, and Selenium Grid for parallel test execution [51].

3.2 Graph Networks

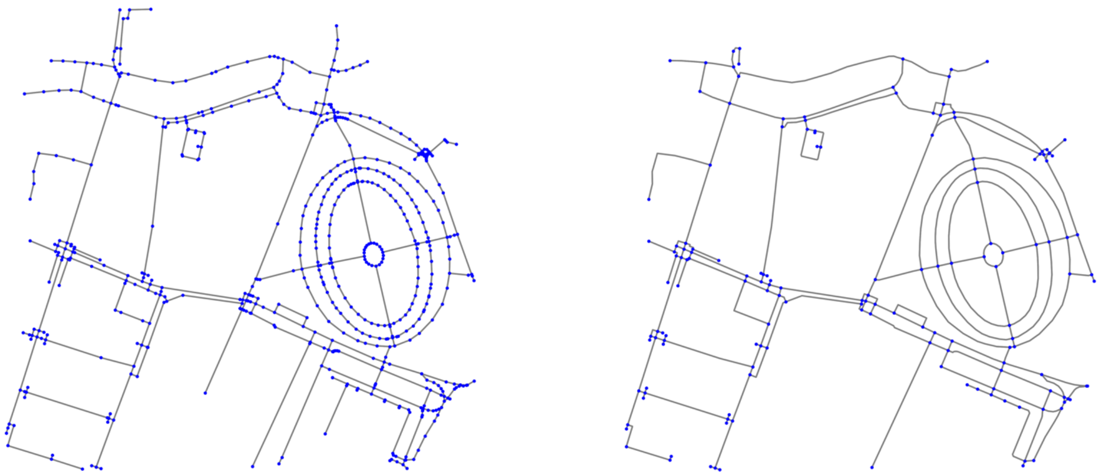
OSMnx models spatial networks as primal, non-planar, weighted, directed multigraphs with possible self-loops. The library allows to retrieve any spatial network data from the Overpass API (a read-only service that provides selected parts of the OSM data based on custom queries) such as streets, paths, rail, canals, among others. OSMnx models a one-way street as a single directed edge from node u to node v, but a bidirectional street is modeled with two reciprocal directed edges (with identical geometries)—one from u to v and another from v to u—to represent both possible directions of flow. Because these graphs are non-planar, they correctly model the topology of interchanges, overpasses, and underpasses [49].

After simplifying raw data using OSMnx, nodes represent intersections or dead-ends in a street network, and edges represent the street segments that connect these nodes. Figure 3.2 illustrates the process before and after simplification, where non-intersection and non-dead-end nodes are deleted, and the edges between them are merged into new simplified edges, retaining the complete true edge geometry as an edge attribute. Subsequently, nearby nodes are merged to create a more accurate model. This is crucial because many real-world street networks feature complex intersections and traffic circles, which result in clusters of graph nodes where there is actually just one true intersection, as it would be considered in transport planning or urban design.

3.3 Bayesian Score Calculation

Not all POI in a city are of interest to tourists. To create a ranked list based on the attractiveness of these POI, both the review score and the number of reviews are considered. By combining these factors, a single value that captures not only the user-given score for each place but also the confidence level of that score (which is based on the number of reviews), can be derived. Once the values are sorted, it is possible to filter the top-*k* attractions/restaurants.

The Bayesian score is a method that adjusts the average rating of products based on the number of ratings they receive. It is used to ensure that products with fewer ratings have less influence on the overall ranking compared to those with more ratings. This technique is crucial



(a) Before simplification

(b) After simplification

Figure 3.2: Graph network of Prato della Valle, Padova, Italy.

because it creates a more reliable comparison between products by using a prior belief (the overall arithmetic mean). It prevents items with a small number of high ratings from outranking popular items with many slightly lower ratings, thus reflecting a better balance of rating quality and quantity [52, 53, 54, 55].

This method is applied to the current work by considering the mentioned products as the POI of the city. Specifically, the Bayesian score for each POI is described by the following equation,

$$s_{ij} = \frac{n_{ij}}{n_{ij} + n_{j,avg}} \cdot A_{ij} + \left(1 - \frac{n_{ij}}{n_{ij} + n_{j,avg}}\right) \cdot S_j \tag{3.1}$$

where i represents the i-th POI of category j, s_{ij} the Bayesian value, n the number of reviews, $n_{j,avg}$ the average number of reviews, A_{ij} the review score as reported by TripAdvisor/AirBnB, and S_j the average review score of the category.

For each POI category, this scoring function considers that:

- For a POI with a fewer than average number of ratings, the score should be around the overall arithmetic mean
- For a POI with a substantial number of ratings, the score should be the POI's review score as reported by TripAdvisor/AirBnB

- As the number of ratings that a POI receives increases, the score should gradually move from the overall mean to the POI's mean

To illustrate the behaviour of the weights in Equation (3.1), let's consider the following toy example: a POI with review score of 4.5, another POI with review score of 3.5, an average review score of 3.0 for the whole category, and an average number of reviews of 500 for the same category. Under this scenario, Figure 3.3 depicts the change in the weights value as a function of the number of reviews, where it is possible to see that the intersection of the curves occurs at the average number of reviews. The score moves gradually from the overall mean to the POI's mean, as depicted in Figure 3.4.

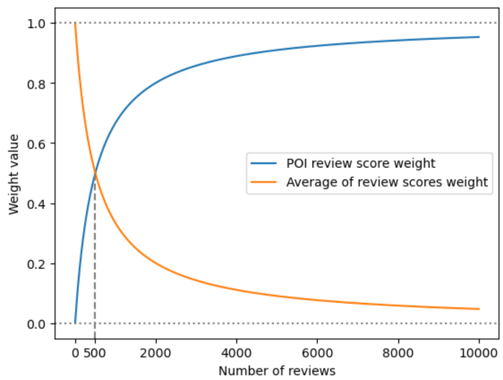


Figure 3.3: Weights of Bayesian score as a function of the number of reviews.

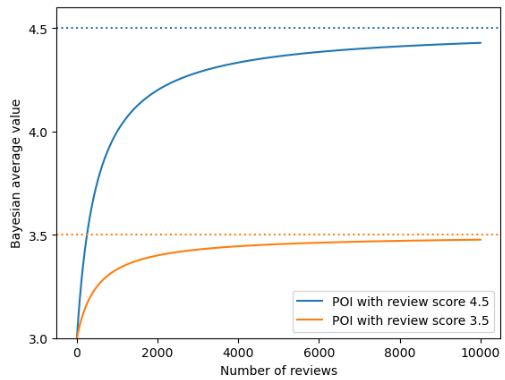


Figure 3.4: Bayesian score value for two POI, as a function of the number of reviews.

3.4 Shortest-path distance calculation

The shortest path between nodes is calculated using Pandana, a library that implements Contraction Hierarchies—a speed-up technique utilized in route planning and navigation systems to efficiently compute shortest paths in graphs [50]. The calculation involves three steps, divided into preprocessing and query phases: node ordering and node contraction (preprocessing phase), and bidirectional query (query phase). A summarized description of each step is provided below, based on [56, 57].

3.4.1 Preprocessing phase: Node ordering

Node ordering is a crucial step in the preprocessing phase, where the sequence of node contractions is determined based on their importance. In the Pandana library, nodes are ranked using a heuristic that evaluates the impact of each node's contraction on the overall graph structure.

The importance of a node is gauged by calculating a contraction cost, which considers the edge difference (the number of shortcut edges that would need to be added minus the number of original edges removed if the node were contracted). Nodes are then ordered from least to most costly to contract. The smaller the contraction cost, the less important the node, and the earlier it is contracted.

This node ordering is designed to optimize the graph for the subsequent query phase by ensuring that the remaining, more important nodes maintain the network's efficiency and connectivity, thereby helping the graph retain its ability to efficiently find the shortest paths.

3.4.2 Preprocessing phase: Node contraction

After determining the node order, nodes are contracted sequentially, starting from the least important. Node contraction involves removing a node from the graph and adding shortcut edges between its neighbors to ensure that the shortest paths are preserved.

Consider a graph $G = (V, E)$, where V is the set of vertices (nodes) and E is the set of edges. For a vertex $v \in V$, let U represent the set of vertices with edges incoming to v, and W represent the set of vertices with edges outgoing from v. For each pair of vertices (u, w), where $u \in U$ and $w \in W$, if the path $\langle u, v, w \rangle$ is the unique shortest path between them, a shortcut edge uw is added to replace the path through v. The weight of this shortcut edge is $\omega(u, v) + \omega(v, w)$, where ω denotes the edge weight.

The need for a shortcut edge is determined by running a localized Dijkstra search from each node $u \in U$. The algorithm checks whether the direct path from u to w (via v) is shorter than any alternative path that bypasses v. If the direct path is shorter, a shortcut is added. If not, no shortcut is necessary.

This process ensures that even after node v is removed, the graph still contains all necessary connections to compute the shortest paths between any pair of nodes. The result of this step is an overlay graph G^*, which consists of the original graph plus the added shortcuts.

3.4.3 Query phase: Bidirectional query

In the query phase, a modified bidirectional Dijkstra's algorithm is employed to efficiently compute the shortest path between two nodes in the preprocessed graph G^*. This approach reduces the search space by conducting searches from both the source and the target simultaneously.

The bidirectional search operates on two subgraphs that are defined by the node ordering from the preprocessing phase:

- The upward graph G_U^* includes edges where the start node was contracted before the end node (each node connects to nodes of higher importance).
- The downward graph G_D^* includes edges where the start node was contracted after the end node (each node connects to nodes of lower importance).

Forward (from the source) and backward (from the target) searches are performed using Dijkstra's algorithm on the upward graph G_U^* and the downward graph G_D^*, respectively (see Figure 3.5 for a schematic comparison between the Dijkstra and bidirectional search spaces).

Figure 3.5: Schematic search spaces of Dijkstra's algorithm (left) and bidirectional search (right) (edited from [58]).

These searches continue until all reachable nodes are settled in both the forward and backward graphs. After both searches are complete, the algorithm calculates the shortest-path by examining all nodes that have been settled by both searches.

For each such node v, the algorithm calculates the sum of distances $\text{dist}(s, v) + \text{dist}(t, v)$, where $\text{dist}(s, v)$ is the distance from the source node s to v found by the forward search, and $\text{dist}(t, v)$ is the distance from the target node t to v found by the backward search. The shortest path corresponds to the minimum of these sums given by the expression,

$$\text{dist}(s, t) = \min\{\text{dist}(s, v) + \text{dist}(t, v)\}, \quad v \in L \tag{3.2}$$

Algorithm	Scanned vertices	Time [μs]
Dijkstra	9,326,696	2,195,080
Bidirectional Dijkstra	4,914,804	1,205,660
Contraction Hierarchies	280	110

Table 3.1: Performance comparison of some speedup techniques on Western Europe [58].

where L is the set of all nodes settled in both searches.

The use of shortcut edges, added during the preprocessing phase, significantly reduces the number of nodes and edges that need to be explored, making the shortest-path computation much faster.

3.4.4 About the method's speed

In [58], various algorithms were used to perform a series of shortest path queries, recording the average number of vertices scanned and the average time taken per query. These tests were conducted on a road network graph of Western Europe. The results, presented in Table 3.1, illustrate the significant speedup achievable with the Contraction Hierarchies algorithm.

3.5 Isochrone maps

An isochrone map is a type of map that visually represents areas accessible within a certain amount of time from a specific point. It shows contours or boundaries around a location, indicating how far one can travel within a given time period, typically by walking, driving, or using public transportation. For example, on an isochrone map centered on a particular location, a 10-minute isochrone might show all the areas that can be reached within 10 minutes of travel time from that point. These maps are useful for urban planning and transportation analysis, as they help visualize accessibility and understand the spatial distribution of services or amenities relative to travel time.

3.6 Price analysis

In this study, Multiple Linear Regression is employed to analyze prices. The primary objective is to determine whether the price of an accommodation is influenced by its walkability, which, in this context, is quantified by the Tourism Walkability Index defined in the following section. The effectiveness of the model is assessed based on the goodness of fit, model complexity, and the significance of the explanatory variables. Additionally, the Regression Tree method is used to further confirm and explain the results. These methods allow for high explainability, enabling an understanding of the impact of walkability.

3.6.1 Definition of OLS for Multiple Linear Regression

Ordinary Least Squares is a statistical method used to estimate the parameters in a regression model. The goal of OLS is to minimize the sum of the squared residuals, which are the differences between the observed and predicted values.

3.6.2 Multiple Linear Regression

In this method, the dependent variable is predicted as a linear combination of multiple independent variables. The OLS method is used to find the coefficients of the independent variables. The model can be represented as,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon \tag{3.3}$$

where Y is the dependent variable, X_n represents the n independent variables, β_n denotes the coefficients of the model, and ϵ is the error term. The numerical variables are standardized, and the categorical variables are converted into dummy variables.

3.6.3 Correlation

Correlation measures the strength and direction of the linear relationship between two individual variables. Since the variables are numerical and categorical, three approaches are used to detect correlation:

- **Pearson Correlation**: The Pearson correlation coefficient measures the linear relationship between two continuous variables, ranging from -1 (indicating a perfect negative linear relationship) through 0 (indicating no linear relationship) to $+1$ (indicating a perfect positive linear relationship). The threshold is set to ± 0.5
- **Cramer's V**: Cramer's V is a measure of association between two categorical (nominal) variables, ranging from 0 (indicating no association) to 1 (indicating a perfect association), without indicating the direction of the relationship. The threshold is set to 0.5

- **Point Biserial**: The point biserial correlation coefficient measures the relationship between a binary variable and a continuous variable, ranging from -1 (indicating a perfect negative relationship) through 0 (indicating no relationship) to $+1$ (indicating a perfect positive relationship). The threshold is set to ± 0.5

Multicollinearity refers to the presence of high correlations among the predictor variables, which can lead to unreliable and unstable estimates of regression coefficients. To detect multicollinearity, the Variance Inflation Factor (VIF) is calculated for each predictor. The VIF for a predictor X_j is defined as,

$$\text{VIF}(X_j) = \frac{1}{1 - R_j^2} \tag{3.4}$$

where R_j^2 is the coefficient of determination obtained by regressing X_j on all other $n - 1$ predictors in the model. A VIF value of 1 indicates complete absence of multicollinearity. A value that exceeds 5 or 10 indicates a problematic amount multicollinearity, suggesting that the predictor X_j is highly correlated with other predictors.

3.6.4 Backward elimination

To identify the most significant variables, backward elimination is utilized. The following steps are followed:

1. **Fit the Full Model**: Begin with all potential predictor variables included in the model.
2. **Evaluate the Least Significant Predictor**: Calculate the statistical significance (using p-values) for each predictor variable.
3. **Remove the Least Significant Predictor**: Identify the predictor with the highest p-value (i.e., the least significant one) and remove it from the model if its p-value is above a certain threshold (0.05 in this case).
4. **Refit the Model**: Refit the model without the removed predictor and repeat the process of evaluating and removing the least significant predictor.
5. **Stop When All Remaining Predictors Are Significant**: Continue the process until all remaining predictor variables in the model are statistically significant (p-values below the threshold).

As there is risk of excluding variables that might be significant in combination with others, a manual check is performed afterwards.

3.6.5 Assumptions on Residuals

The assumptions related to residuals in OLS include:

- **Linearity**: Linear relationship between the dependent and independent variables in terms of the coefficients. This assumption is checked by a visual inspection of the Residuals vs. Fitted Values plot.
- **Homoscedasticity**: The variance of the residuals is constant across all levels of the independent variables. It is checked by a visual inspection of the scale-location plot.
- **Normality**: The residuals are normally distributed. It is checked by a visual inspection of the QQ-plot.

These assumptions are necessary for the OLS estimates to be unbiased, efficient, and consistent, which are desirable properties of a good estimator.

3.6.6 Regression trees

A regression tree is a type of decision tree used for analyzing continuous numerical outcomes. It works by recursively splitting the data into subsets based on the values of input features, creating a tree structure where each node represents a decision based on a feature, and each leaf represents a calculated value.

4 Methodology

This Chapter details the steps undertaken for the analysis, based on the concepts discussed in Chapter 3.

4.1 Procedure description

The methodology employed in this study follows a structured approach to assess urban walkability and its impact on accommodation pricing. The process begins with web scraping to collect relevant data on POI and accommodations. Once the data is gathered, POI are mapped to the city's street network to accurately reflect their proximity and accessibility.

A ranking system for attractions and restaurants is developed using a Bayesian Score, focusing on the top-rated POI. A t-minute walking framework is then established, converting shortest-path distances from accommodations to POI into walking times, and identifying those within the t-minute threshold. Then isochrone maps are created to visualize accessible areas.

After a process of feature engineering, the Tourism Walkability Index is calculated for each accommodation, and a histogram of the index values is plotted to evaluate city-wide walkability.

Finally, a price analysis using multiple linear regression and regression tree methods is conducted, exploring the relationship between walkability and accommodation pricing. This structured approach provides insights into the economic value of walkable urban environments.

In the following sections, a detailed explanation of each step is presented.

4.2 Assumptions

Some important assumptions and simplifications of this study are presented:

- **Tourist Type**: The tourist considered in this study is one who visits a city seeking natural, cultural, entertainment activities, and dining experiences of good quality within the city limits. This tourist prefers walking and public transportation, specifically metro (or equivalent) services, as means of active transport.

- **Walking Speed**: For the purpose of calculations, the tourist's walking speed is assumed to be constant. This simplification does not account for time spent at traffic lights or variations in street steepness, factors that could directly influence the walkability index.

- **Shortest-Path**: It is assumed that the tourist follows the shortest path when moving from the accommodation to a POI. This assumption is based on the typical behavior of tourists who are unfamiliar with the city's streets and rely on navigation services that provide the shortest routes.

- **Isochrone Concept**: The study is based on the isochrone concept, which analyzes accessibility within a given time frame. As such, it does not consider behaviors dependent on specific times, such as a tourist deciding to visit a nearby location after reaching an initial destination.

4.3 Web Scraping of Data and features

Web scraping is a method that is used to extract large volumes of data from websites. This data is then stored in a local file on a computer or in a database in a table format. Typically, websites do not provide an option to save their data for personal use. The only alternative is to manually copy and paste the data, which can be a time-consuming task. Web scraping automates this process, allowing the same task to be completed much more quickly [59].

In the context of this Thesis, Selenium is used to scrape and update existing hotel data from TripAdvisor. TripAdvisor is the world's largest travel site, reaching 350 million unique monthly visitors and 385 million reviews and opinions for more than 6.6 million accommodations, restaurants, and attractions [28]. Based on this, the focus was on collecting information about these three categories.

Particularly, the extracted information includes the following variables:

- **gcode**: Unique identifier used by TripAdvisor for the city where the place is located
- **dcode**: Unique identifier used by TripAdvisor for the place
- **name**: Name of the place as displayed in TripAdvisor
- **review_score**: The overall score reported by TripAdvisor, which considers the quality, quantity, and recency of evaluations made by users
- **number_of_reviews**: total count of individual ratings that users have left based on their experience
- **Location coordinates**: GPS coordinates of the place
- **municipality**: Location of the place at municipality level
- **price_range**: Minimum and maximum prices based on the average rates for a standard room (in USD)
- **location_score**: Optional rating given by users based on their experience with the hotel's location. This could take into account factors such as proximity to tourist attractions, accessibility of transportation, safety of the neighborhood, and more
- **rooms**: Total count of rooms of the hotel
- **hotel_class**: Star-based score as given by GIATA company. It ranges from 1 to 5 stars

The hotel class is determined by the amenities offered by the accommodation and is a standardized measure given by the GIATA company, a provider of high-quality hotel content and mapping services. This company assigns official star categories, making it possible to compare hotels across the platform [60]. A summary of some amenities for each hotel class is presented in Table 4.1.

In the case of AirBnB, the data was obtained from InsideAirBnB [62]. InsideAirBnB is a mission-driven project that provides data and advocacy regarding AirBnB's impact on residential communities. The project works towards a vision where communities are empowered with data and information to understand, decide, and control the role of renting residential homes to tourists. Through data collection, analysis, and visualization, Inside AirBnB aims to promote transparency and support informed decision-making about short-term rentals and their effects on local housing markets and neighborhoods.

Particularly, the extracted information includes the following variables:

Star Rating	Amenities
⋆	Daily room cleaning, shower/WC or bath tub/WC in all rooms, TV and remote function in all rooms, WIFI in public areas and rooms, soap or body wash, bath towels
⋆⋆	All 1-star amenities, breakfast buffet, reading light next to bed, bath and hand towels, bilingual staff, linen shelves
⋆ ⋆ ⋆	All 2-star amenities, 10 hours staffed reception, lobby with seats and beverage service, luggage service on demand, laundry and ironing service
⋆ ⋆ ⋆⋆	All 3-star amenities, 16 hours staffed reception, breakfast buffet with service or equivalent breakfast menu card, bath robe and slippers on demand, international TV channels
⋆ ⋆ ⋆ ⋆ ⋆	All 4-star amenities, 24 hours staffed reception, valet parking service, shuttle or limousine service, Personalized greeting with flowers or present, minibar and 24 hours room service

Table 4.1: Hotel stars criteria [61].

- **id**: Unique identifier for each listing
- **host_is_superhost**: Indicator of whether the host is a Superhost (a status given to top-rated and experienced hosts)
- **neighbourhood_cleansed**: The specific neighborhood where the listing is located
- **Location coordinates**: GPS coordinates of the place
- **property_type**: The type of property (e.g., apartment, house)
- **room_type**: The type of room being offered (e.g., entire home/apartment, private room)
- **accommodates**: The number of guests the listing can accommodate
- **bathrooms**: The number of bathrooms available in the listing
- **bedrooms**: The number of bedrooms available in the listing
- **beds**: The number of beds available in the listing.
- **amenities**: The amenities provided with the listing (e.g., Wi-Fi, kitchen)
- **price**: The price per night for the listing (in USD)
- **minimum_nights**: The minimum number of nights required to book the listing

- **maximum_nights**: The maximum number of nights a guest can stay in the listing
- **availability_365**: The number of days the listing is available for booking over the next 365 days
- **number_of_reviews**: The total number of reviews the listing has received
- **review_scores_rating**: The overall rating score of the listing based on guest reviews
- **review_scores_accuracy**: The rating score for the accuracy of the listing description
- **review_scores_cleanliness**: The rating score for the cleanliness of the listing
- **review_scores_checkin**: The rating score for the check-in process
- **review_scores_communication**: The rating score for the communication with the host
- **review_scores_location**: The rating score for the location of the listing
- **review_scores_value**: The rating score for the value of the listing

In both cases, non-consistent data are removed from the datasets prior to any analysis.

4.4 Mapping of POI and accommodations

Using GPS coordinates, POI and accommodations are assigned to the nearest existing node in the network. This approach is employed to avoid the necessity of modifying the network structure, which could complicate analysis and increase computational demands. By assigning these locations to the closest node, the integrity of the existing network is maintained, while still accurately representing the proximity of these points. Figure 4.1 illustrates this process, where the Campanile di San Marco in Venice (green star) is assigned to the closest node in the network (blue point).

4.5 Creation of POI rankings

Rankings for attractions and restaurants in the city are created using the Bayesian Score criterion. This method allows for the identification of high-quality attractions and restaurants. A score of 4 or higher is arbitrarily defined as indicating good quality, and only those places that meet this threshold are retained for further analysis.

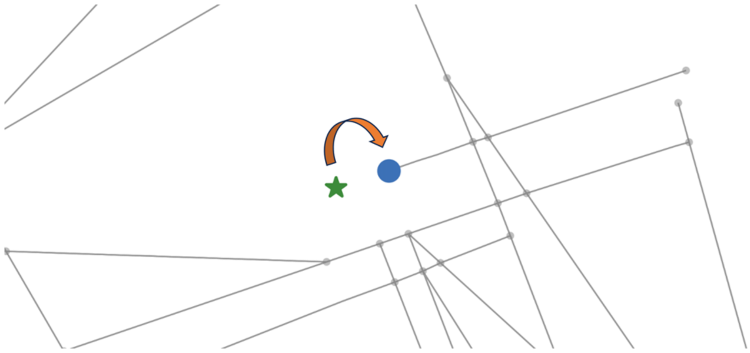


Figure 4.1: Example of node mapping for Campanile di San Marco, Venice.

4.6 Shortest-Path calculation

The *t*-minute walking framework is set to 15 minutes, with a walking speed of 1 m/s, based on an inclusion criterion that considers people of all ages. This speed is selected as it is the lowest found in the analyzed literature.

Afterwards, the shortest path from each accommodation to each POI is calculated. The weights of these paths are then converted from distance to time using the defined walking speed. Subsequently, for each accommodation, only the POIs within a 15-minute walking distance are retained.

An example of this process is shown in Figure 4.2, where the star marks the Concordia Hotel in Venice, the blue points denote the POI (from left to right: Chiesa di Santa Sofia, Basilica di San Marco, and Campanile di San Marco), and the red paths indicate the shortest routes from the hotel to each POI. Only the POI circled in green are kept, given that they are reachable within 15 minutes from the hotel.

4.7 Isochrone maps

An isochrone map is generated by evaluating the accessibility to POI from each accommodation, identifying the nodes of the network that are reachable within a *t*-minute walk from each hotel or BnB. The map incorporates varying travel times, ranging from 15 to 45 minutes, with nodes colored accordingly. Subsequently, the POI are overlaid on the map to visually assess their reachability from at least one of the accommodations (accommodations are not displayed).

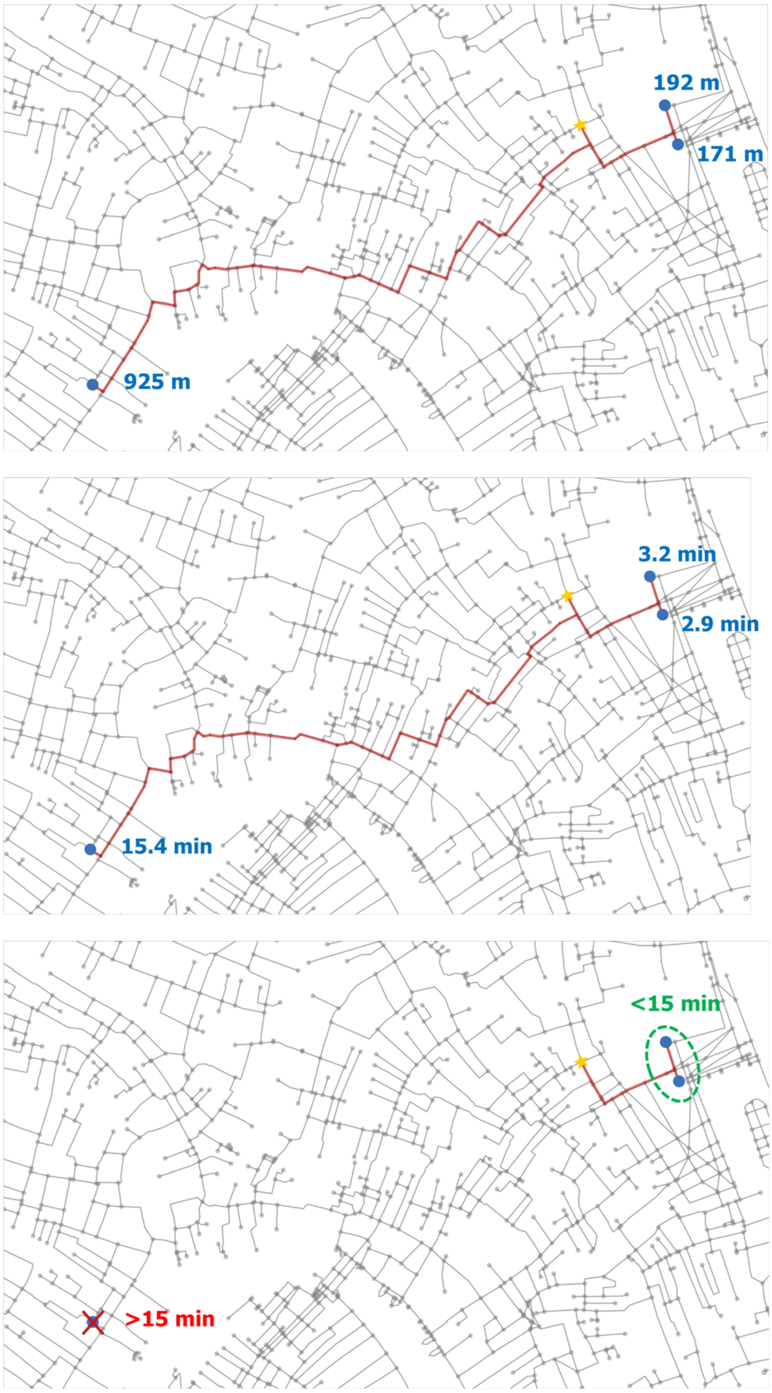


Figure 4.2: Example of shortest-path calculation and POI selection for a hotel in Venice.

4.8 Feature engineering

There are several variables created from the original data to support the analysis of the network: Count, Tourism Walkability Index, Metro, Closeness centrality, Bayesian score of accommodation, Bayesian score mean for attractions, and Bayesian score mean for restaurants. In the case of the AirBnB data, two additional variables (Count of amenities and Amenities type) are engineered. A description of each one is given next.

4.8.1 Count

The Count feature considers the number of attractions and number of restaurants (separately) within walking distance from each accommodation (t-minute criteria).

4.8.2 Tourism Walkability Index

This subsection explains a proposed Tourism Walkability Index (TWI) to measure the walkability of each accommodation. The methodology is based on Frank's Walkability Index [63], which is applied to the neighborhood (residential needs) and is supported by similar approaches from other authors (see [64], [65], [66]). A step-by-step procedure to develop the index is given next.

Scaling of variables

The first step in constructing the composite index is to scale each variable that could potentially be part of the index. This is achieved through standardization using the Z-score, which is calculated with the following equation,

$$x'_{ij} = \frac{x_{ij} - \bar{x_j}}{\sigma_{x_j}} \tag{4.1}$$

where i is the i-th observation of the j-th variable, x' is the Z-score, x is the original value, $\bar{x}$ is the mean, and σ is the standard deviation.

Index calculation

Once the variables are preprocessed to a common scale, they are aggregated to create a single value. In this case, the additive method is used. This combination method is relatively simple to interpret and is commonly used by a variety of indices. It allows high values in one variable to compensate for low values in another variable (see Figure 4.3). Hence, the proposed TWI is calculated as follows,

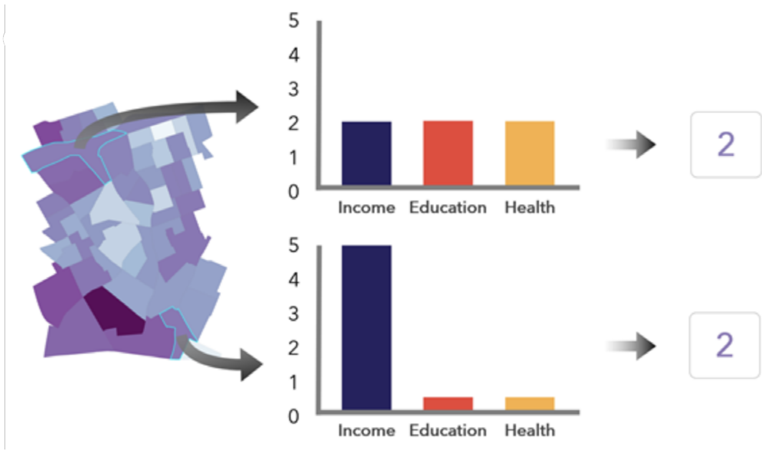


Figure 4.3: Example of compensation in an aggregation method (mean method) [66].

$$\text{TWI} = 2Z_{Count_A} + Z_{Count_R} \tag{4.2}$$

where Z_{Count_A} is the Z-score of the variable "Count of attractions", and Z_{Count_R} is the Z-score of the variable "Count of restaurants". The index aims to capture the attractiveness of an accommodation in terms of its walkability, based on the two main POI categories for tourists according to the available data: attractions and restaurants within walkable distance.

Weighting

It is important to highlight that the number of attractions is weighted by a factor of 2, making it relatively more important compared to restaurants. This proposition is based on the information presented in Chapter 2 regarding the significance tourists place on surrounding attractions. Additionally, attractions encompass, in general, a wide variety of natural, cultural, and recreational activities, justifying the increased weight. The validation of the TWI is addressed in the final step of the current subsection.

Index range	Walkability level
$0 \leq \text{TWI} \leq 20$	Poor
$20 < \text{TWI} \leq 40$	Basic
$40 < \text{TWI} \leq 60$	Average
$60 < \text{TWI} \leq 80$	High
$80 < \text{TWI} \leq 100$	Excellent

Table 4.2: Classification of the TWI using equal intervals.

Scaling and classification of the index

In this step, the index is scaled between 0 and 100 for easier interpretation, where 0 represents the minimum possible walkability and 100 represents the maximum possible walkability. The following equation is used,

$$\text{TWI}'_i = a + \frac{(\text{TWI}_i - \min(\text{TWI}))(b - a)}{\max(\text{TWI}) - \min(\text{TWI})} \tag{4.3}$$

where TWI_i is the i-th original value, $\min(\text{TWI})$ is the minimum value found in the index, $\max(\text{TWI})$ is the maximum value found in the index, a is the specified minimum value (0 in this case), b is the specified maximum value (100 in this case), and TWI'_i is the i-th scaled value.

Five qualitative levels are proposed for accommodations based on their TWI levels, using equal interval classification, as displayed in Table 4.2. Grouping accommodations according to their TWI class makes it easier to interpret a city's attractiveness for tourists in terms of walkability from the accommodation. A histogram based on this classification is used for this purpose.

Validation of the index

Validating the index requires external information, such as expert opinions or surveys of tourist preferences. In this case, to understand if the index is capturing some of the preferences of users regarding walkability, the location scores are used.

For the hotels, the Location Score from TripAdvisor is used. As it is previously mentioned, the Location Score is an optional field provided to users after they rate their overall experience at an accommodation (see Figure 4.4). It allows users to rate the place based on the perceived attractiveness of the accommodation's location. For AirBnB accommodations, the location ranking (see Figure 4.5) is derived from guest reviews and ratings. Guests provide feedback through AirBnB's review system, where they rate various aspects of their stay, including location.

On the other hand, Spearman's correlation measures the strength and direction of the association between two ranked variables. It ranges from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation. It assesses the relationship without assuming linearity or normality, making it suitable for non-linear associations and non-normally distributed data.

Figure 4.4: Location Score of a hotel [67].

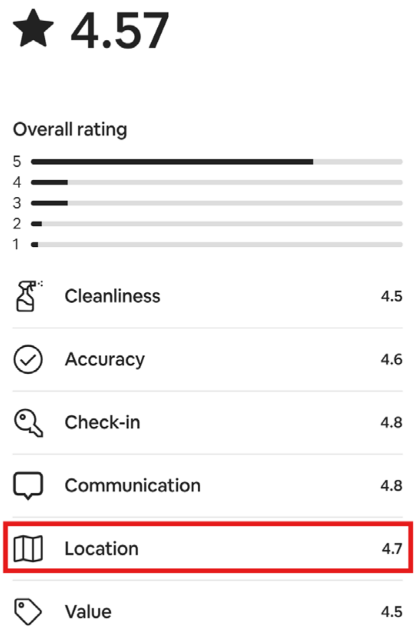


Figure 4.5: Location Score of an AirBnB [68].

Spearman's correlation is calculated between the location scores and the TWI. The hypothesis is that a correlation different from zero would indicate that the TWI captures some of the user preferences regarding the availability of POI within walking distance from the accommodation. In other words, it would suggest that the selection and weighting of the count of attractions and restaurants within a t-minute distance are appropriate and a good approximation. This fact is further emphasized by analyzing the coefficient and significance levels in the price analysis section.

4.8.3 Metro

The Metro feature assesses whether there is at least one metro (or its equivalent) within walking distance from each accommodation. This is determined using a t-minute walking criterion and the shortest-path calculation from the accommodation to all the stations/stops and selecting the closest one. The feature is represented as a binary variable, where a value of zero indicates the absence of any metro station within the specified walking time.

4.8.4 Closeness centrality

Normalized closeness centrality is calculated for each accommodation, taking into account the attractions and restaurants separately, regardless of whether they are within the t-minute distance or not. The idea is to capture how central an accommodation is with respect to all the POI of a given category in the city. This metric is calculated as follows,

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} dist(u,v)} \tag{4.4}$$

where $n-1$ is the number of reachable nodes (POI belonging to a certain category) from node u, v is the destination node, and $dist(u,v)$ is the shortest-path distance from u to v.

4.8.5 Bayesian score of accommodation

This feature considers the previously described Bayesian score for the accommodation, which is interpreted as the quality score for the place.

4.8.6 Bayesian score mean for attractions

This feature considers the quality of the surrounding amenities by calculating the mean Bayesian score of attractions within 15 minutes of the accommodation.

4.8.7 Bayesian score mean for restaurants

This feature considers the quality of the surrounding amenities by calculating the mean Bayesian score of restaurants within 15 minutes of the accommodation.

4.8.8 Count of amenities

For AirBnB's data, this feature counts the number of amenities of the accommodation.

4.8.9 Amenities type

For Airbnb's data, this feature classifies each of the accommodation's amenities into one of the following categories:

- **TVs**: Various types of televisions available in the property
- **Wifi**: Internet connectivity options
- **Parking**: Vehicle parking facilities
- **Kitchen_Appliances**: Appliances in the kitchen
- **Bathroom_Amenities**: Items provided in the bathroom
- **Child_Amenities**: Facilities for children
- **Entertainment_Systems**: Entertainment options
- **Miscellaneous**: Various amenities not classified in the previous ones

4.9 Price analysis

For the price analysis, multiple linear regression and regression trees are utilized (the latter only for BnBs, based on the high number of observations available).

For the multiple linear regression, the process begins with the log transformation of accommodation prices, followed by the removal of outliers. The explanatory variables are then grouped based on their distribution. Highly correlated variables are identified, and one variable from each correlated pair is removed. Multicollinearity is assessed using VIF, and variables with high values exceeding 5 are eliminated. Finally, a backward elimination process is conducted, followed by a manual inspection of the remaining variables. The assumptions of linearity, homoscedasticity, and normality are also examined.

Regression trees are utilized to confirm previous findings. The tree is constructed by performing a grid search to identify the minimum number of samples per leaf that maximizes R^2, using 5-fold cross-validation. As in the case of linear regression, highly correlated variables are identified and one of them removed prior to model construction.

SHAP analysis is then performed to validate the results on price from previous calculations using linear regression, ensuring the reliability and consistency of the model's findings.

5
Experiments and Results

This Chapter presents the experiments carried out, the results, and the discussion on them.

5.1 Napoli

5.1.1 Description

The first city under study corresponds to Naples, Italy. The network consists of $21,161$ nodes and $58,218$ edges. The considered walking network is displayed in Figure 5.1.

The city has 592 attractions whose review count and score are different from zero. Among these, 541 are retained because they have a Bayesian Score of 4 or higher, which is arbitrarily considered indicative of good quality for attractions. Figure 5.2 shows the location of each one of these attractions. The list of the top-10 attractions is presented in Table 5.1.

The city has $2,103$ restaurants whose review count and score are different from zero. Among these, 111 are retained because they have a Bayesian Score of 4 or higher, which is arbitrarily considered indicative of good quality for restaurants. Figure 5.3 shows the location of each one of these restaurants. The list of the top-10 restaurants is presented in Table 5.2.

For the metro, Table 5.3 presents the list of metro stations considered, which are displayed in Figure 5.4.

Figure 5.1: Considered walking network of Naples, Italy.

Ranking	Name	Score	Review count	Bayesian score
1	Galleria Borbonica	5.0	10,365	4.975244
2	Catacombe di San Gennaro	5.0	5,680	4.955772
3	Catacombe di San Gaudioso	5.0	1,659	4.863877
4	Museo delle Arti Sanitarie	5.0	1,002	4.793913
5	Parco Archeologico Pausilypon	5.0	981	4.790470
6	Napulitanata	5.0	933	4.782153
7	SantAnna dei Lombardi	5.0	873	4.770780
8	Chiesa Museo di Santa Luciella ai Librai	5.0	860	4.768158
9	MUSA Museo Universitario	5.0	714	4.733976
10	La Chiesa Di San Giovanni A Carbonara	5.0	574	4.690174

Table 5.1: Top-10 attractions of Naples, Italy, according to the Bayesian score criteria.

Ranking	Name	Score	Review count	Bayesian score
1	Hachi Ristorante Giapponese	5.0	7,757	4.958912
2	Januarius	5.0	986	4.739152
3	Re Pazzo Pizza	5.0	967	4.735178
4	Pizzeria Speranzella	5.0	932	4.727530
5	Pizzeria Ciccio	5.0	681	4.656362
6	Opera restaurant	5.0	384	4.502650
7	Pizzeria Starita a Materdei	4.5	10,111	4.481692
8	Osteria Il Gobbetto	4.5	3,982	4.455363
9	Tandem	4.5	3,270	4.446411
10	Antica Capri	4.5	3,118	4.444013

Table 5.2: Top-10 restaurants of Naples, Italy, according to the Bayesian score criteria.

Chiaiano	Vanvitelli	Quattro Giornate
Materdei	Museo	Dante
Montedonzelli	Piscinola-Scampia	Lala
Mergellina	Augusto	Università
Frullone	Toledo	Duomo
Policlinico	Colli Aminei	Rione Alto
Medaglie d'Oro	Garibaldi	Mostra
Salvator Rosa	Municipio	

Table 5.3: Considered metro stations for Naples, Italy.

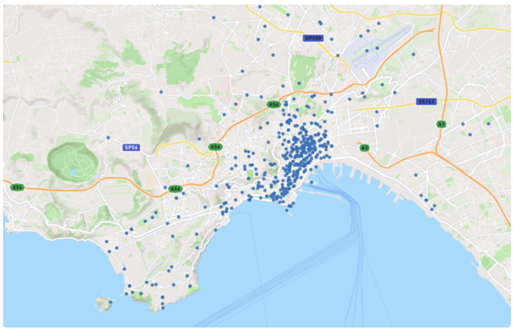


Figure 5.2: Considered attractions for Naples, Italy.

Figure 5.3: Considered restaurants for Naples, Italy.

Figure 5.4: Considered metro stations for Naples, Italy.

5.1.2 Hotels

The city of Naples has 243 hotels, which are displayed in Figure 5.5. Most of these hotels are concentrated in Municipalities 1 and 2. The distribution based on hotel class is shown in Figure 5.6, illustrating that the majority of the hotels are 3-star, followed by 4-star establishments. Figure 5.7 provides a visual representation of the hotels on the map, with points colored according to their respective hotel class.

Figure 5.5: Hotels location in Naples, Italy.

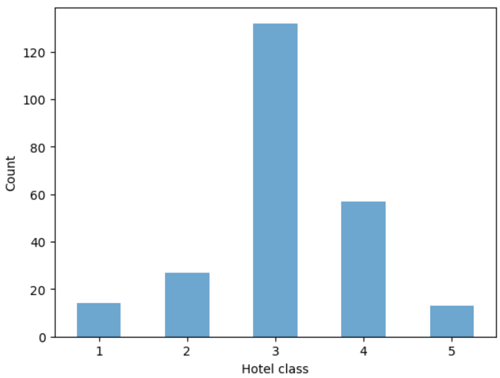


Figure 5.6: Hotels distribution by class in Naples, Italy.

Figure 5.7: Hotels distribution by class on the map in Naples, Italy. Classes represented are 1 (red), 2 (orange), 3 (yellow), 4 (light green), and 5-stars (green).

5.1.3 BnBs

The city of Naples has 6,887 AirBnBs, which are displayed in Figure 5.8. As the hotels, most of these accommodations are concentrated in Municipalities 1 and 2. The distribution based on accommodation type shows that around 74% of the BnBs are entire homes or apartments, and the rest private rooms.

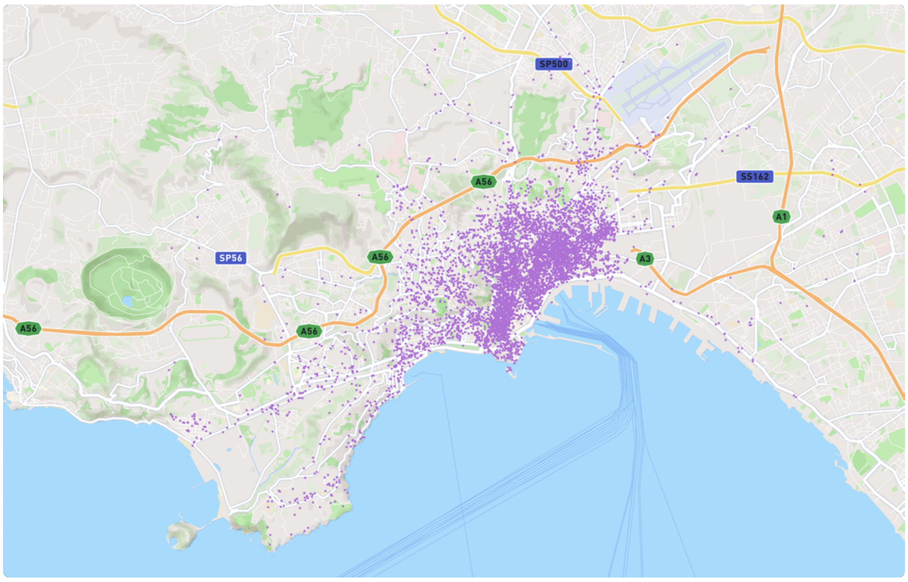


Figure 5.8: AirBnB locations in Naples, Italy.

5.1.4 Isochrone Maps

The isochrone maps from each accommodation are displayed in Figures 5.9 and 5.10. A range of 15 to 45 minutes is used to assess the reachable nodes within that time frame. In both cases, most of the attractions and restaurants in the city are within 15 minutes from at least one accommodation, which is a good indicator of the accessibility and convenience for visitors staying at these locations. For the BnBs, most of the city is reachable in less than 15 minutes, as these accommodations are more sparsely distributed around the city.

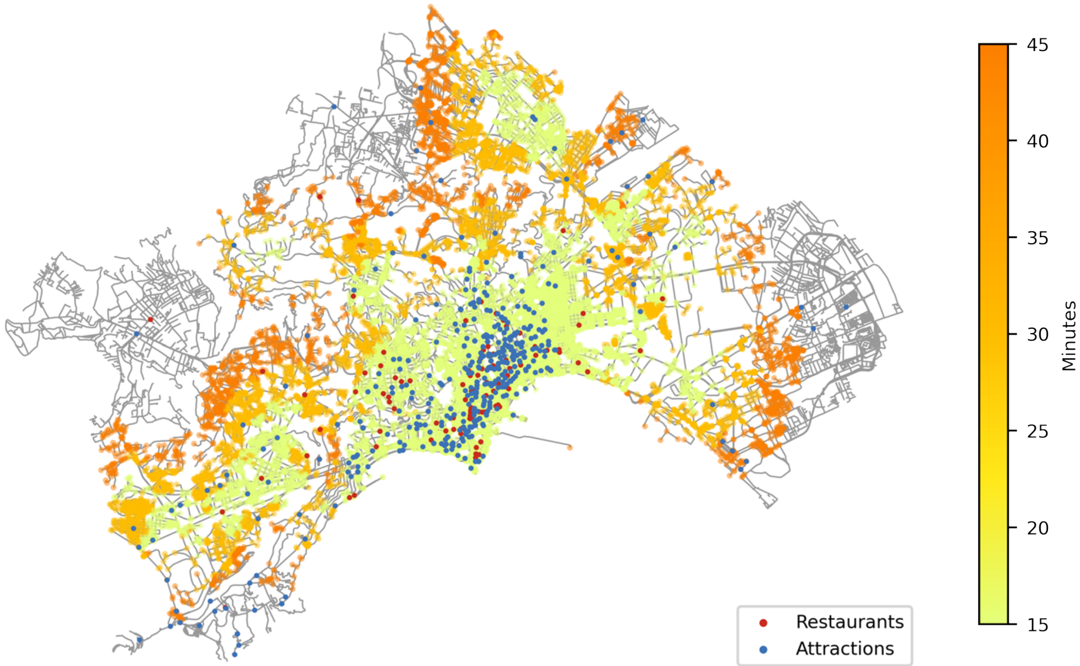


Figure 5.9: Isochrone map for hotels in Naples, Italy.

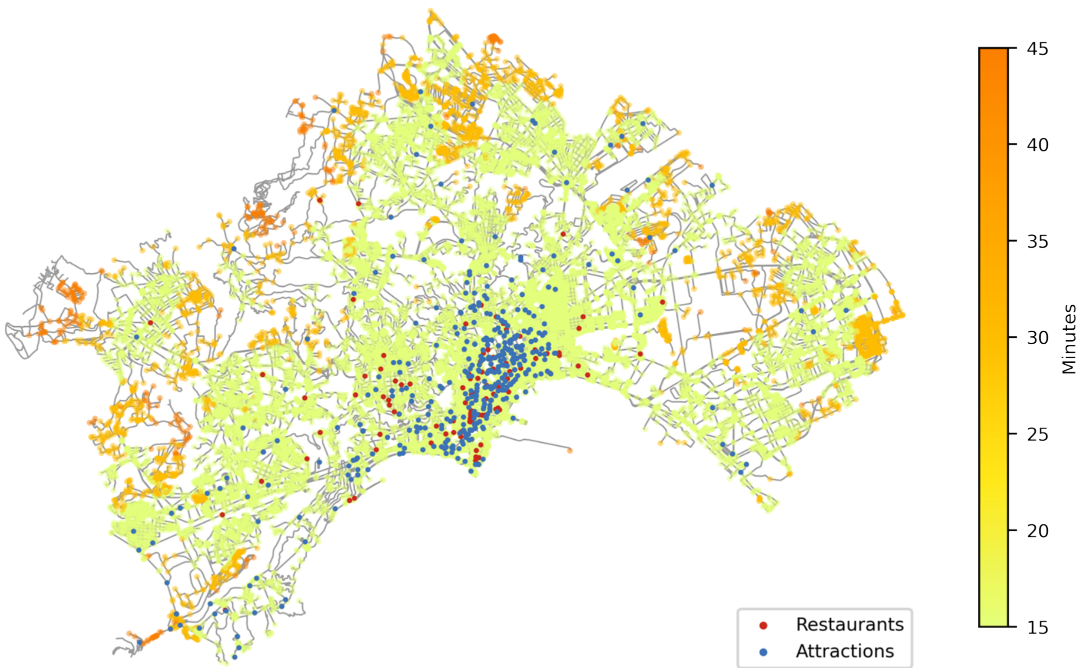


Figure 5.10: Isochrone map for BnBs in Naples, Italy.

5.1.5 TWI Distribution

For each accommodation, the TWI is calculated. While isochrone maps provide information about the areas of the city that are reachable within specific time frames, the TWI offers a metric that captures the attractiveness of an accommodation by considering the number and quality of nearby attractions and restaurants.

Figure 5.11 shows the normalized distribution of the TWI for both hotels and BnBs. The effect of the sparsity of the BnBs is evident in the comparison of the bars, where BnBs show a higher percentage of Poor, Basic, and Average walkability compared to hotels. In contrast, hotels, which are primarily located in the city center near most attractions and restaurants, exhibit a higher percentage of High and Excellent walkability.

Figures 5.12 and 5.13 provide a visual representation of the hotels and BnBs on the map, with points colored according to their respective TWI levels (two colors per level for better visualization). It is important to highlight that the TWI distribution observed is consistent with the distribution of attractions and restaurants in the city from Figures 5.2 and 5.3, which are mainly located in Municipalities 1 and 2.

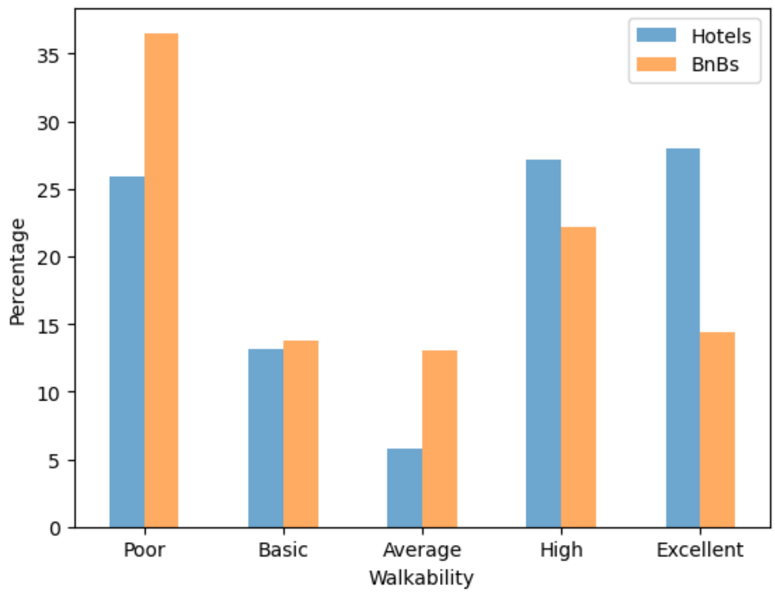


Figure 5.11: Distribution of accommodations according to its TWI of hotels and BnBs in Naples, Italy.

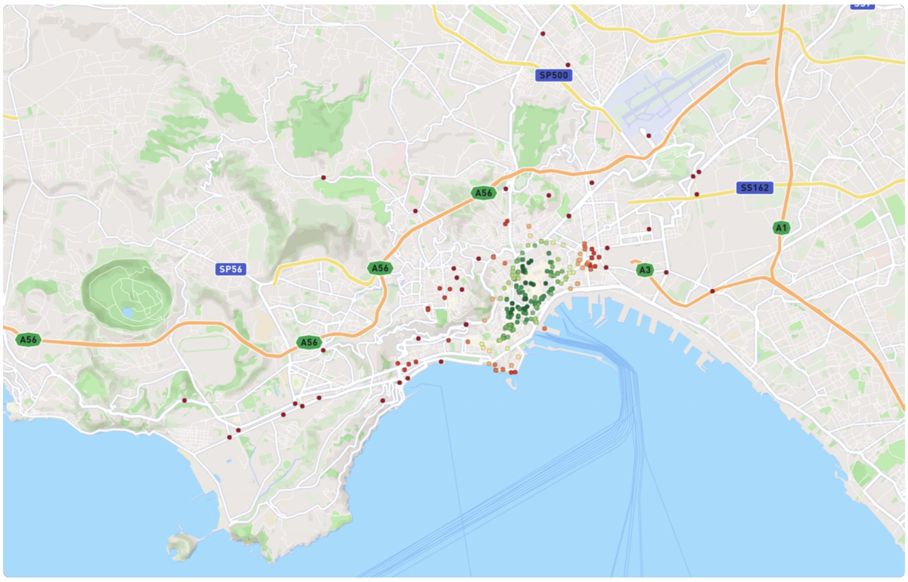


Figure 5.12: Hotels distribution by TWI level on the map in Naples, Italy. Levels represented are Poor (dark-red/red), Basic (orange-red/orange), Average (light-orange/yellow), High (light yellow-green/light green), and Excellent (green/dark-green).

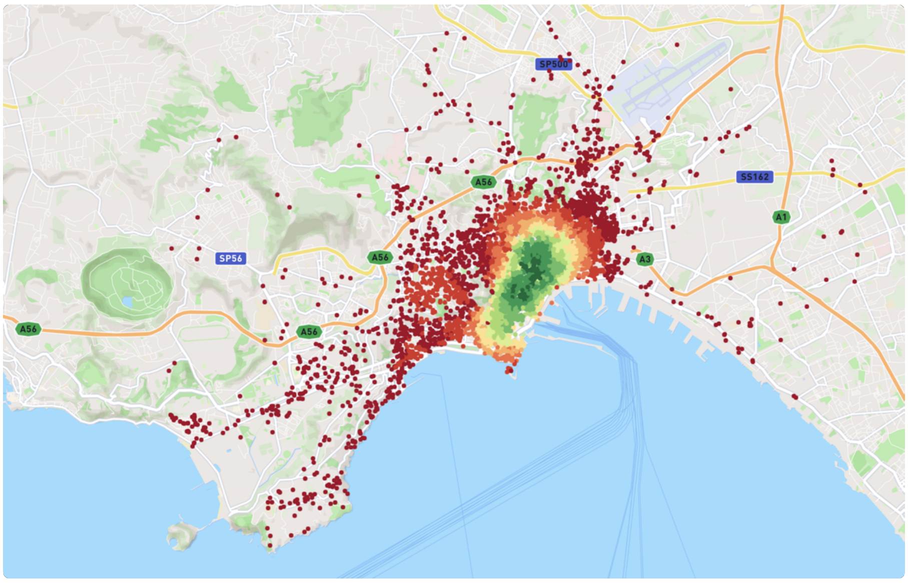


Figure 5.13: BnBs distribution by TWI level on the map in Naples, Italy. Levels represented are Poor (dark-red/red), Basic (orange-red/orange), Average (light-orange/yellow), High (light yellow-green/light green), and Excellent (green/dark-green).

5.1.6 Price Analysis for Hotels

The process starts by calculating the log (natural) of the price to stabilize the variance of the residuals and improve linearity. After this, outliers are removed by keeping the data that falls

Metro_station	Closeness_Centrality_Attractions	location_score
Closeness_Centrality_Restaurants	Review_count_H	Score_H
Bayesian_avg_H	Bayes_mean_A	Bayes_mean_R
price_range_min	price_range_max	n_rooms_quant_0
n_rooms_quant_1	n_rooms_quant_2	n_rooms_quant_3
Walkability_num_1	Walkability_num_2	Walkability_num_3
Walkability_num_4	Walkability_num_5	Hotel_class_12
Hotel_class_3	Hotel_class_45	

Table 5.4: Considered variables for hotel analysis in Naples and Venice, Italy.

between $Q_1 - 1.5 * \text{IQR}$ and $Q_3 + 1.5 * \text{IQR}$, where Q_1 and Q_3 are the first and third quartiles, and IQR is the interquartile range. This process leaves 233 accommodations for processing.

The Spearman correlation between TWI and the Location Score is 0.57, so it is concluded that some of the preferences of the users regarding the walkability of the place are captured by the index.

The variables considered are displayed in Table 5.4.

For the hotel class variable, due to its distribution as shown in Figure 5.6, classes 1 and 2 are combined, as well as classes 4 and 5, to address the imbalance with class 3.

For the variable number of rooms, it is not considered important to treat it as a numerical variable, so it is transformed into a categorical variable using quartiles.

The TWI is converted into the categorical variable Walkability_num to analyze the impact of the different walkability levels separately. Afterwards, numerical variables are standardized and the procedure from Section 4.9 is followed. This leads to the results reported in Table 5.5.

In Model 1, the results suggest that higher walkability is associated with higher prices, indicating that walkability has an impact on pricing. However, the Average walkability level does not significantly differ from the base level (Poor). This lack of significant difference could be due to two factors: either there is no substantial price difference between Poor and Average walkability, or the sample size for Average walkability is affecting the results because of its smaller number of examples compared to other levels (as shown in Figure 5.11). The small sample size can result in a high standard error, which reduces the likelihood of achieving statistical significance due to insufficient data to accurately estimate the effect of this category.

Hotel quality also appears to be an important factor. Both the Bayesian Score and the number of reviews are positively related to price. Although the number of reviews contributes to

	Model 1	Model 2
Constant	4.7450***	4.8434***
	(0.049)	(0.043)
n_rooms_quant_0	−0.1441***	−0.1361***
	(0.040)	(0.041)
Hotel_class_3	0.1210**	0.1186**
	(0.044)	(0.046)
Hotel_class_45	0.2854***	0.2738***
	(0.054)	(0.055)
Walkability_num_2	0.1247*	-
	(0.044)	
Walkability_num_4	0.1319**	-
	(0.054)	
Walkability_num_5	0.1503***	-
	(0.043)	
Bayesian_avg_H	0.0655***	0.0745***
	(0.018)	(0.018)
Review_count_H	0.0561***	0.0565***
	(0.012)	(0.012)
AIC	18.23	27.95
R^2	0.353	0.308
Adj.R^2	0.330	0.293
N	233	233

Table 5.5: Multiple linear regression results for hotels with and without TWI for Naples, Italy.

the calculation of the Bayesian Score, the formula combines these variables in a non-linear way. This allows the two variables to be included in the model while maintaining a small correlation between them.

Hotel class is positively related to price, with higher star ratings indicating higher prices. This is expected, as hotels with more stars offer more amenities. Another interesting result is that hotels with small number of rooms seem to be associated with lower prices.

As it was mentioned in Section 4.9, linearity, homoscedasticity, and normality assumptions are assessed. Figure 5.14 shows the residuals vs. fitted plot. Given that the residuals appear to be randomly scattered around the red zero line, it suggests that the linearity assumption holds reasonably well. Figure 5.15 shows the scale-location plot, where the slight curvature in the LOESS line (in violet) might indicate minor deviations, but these do not appear significant enough to suggest a serious violation of the homoscedasticity assumption. For the normality, Figure 5.16 shows the QQ-plot, where most of the residuals closely follow the red reference line.

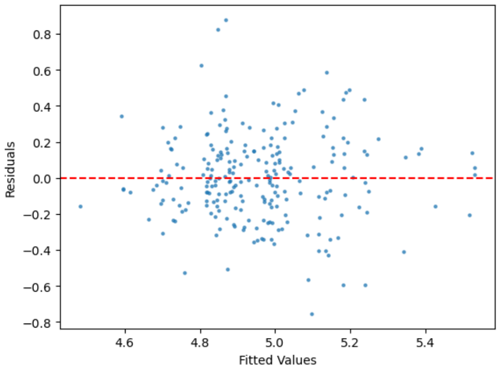


Figure 5.14: Residuals vs. fitted plot for hotels in Naples, Italy.

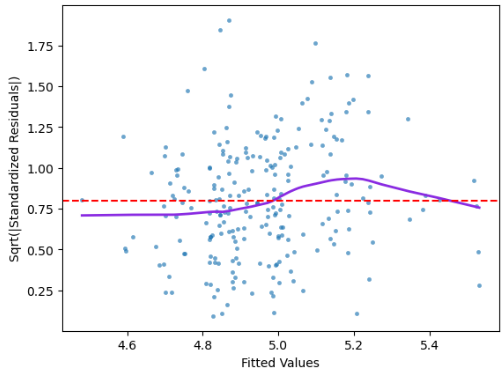


Figure 5.15: Scale-location plot for hotels in Naples, Italy.

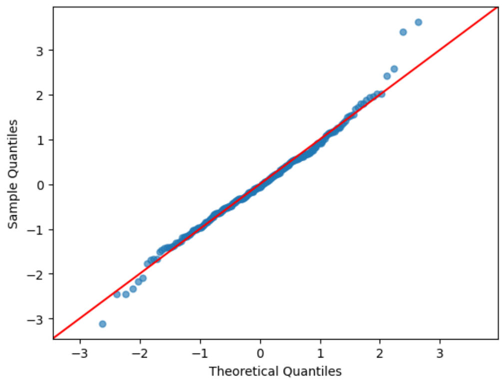


Figure 5.16: QQ-plot for the hotels in Naples, Italy.

In Model 2, the impact of the TWI is assessed by removing the associated variables from Model 1. This results in the AIC increasing from 18.23 to 27.95, confirming the importance of the TWI in explaining the variation in price. The significant increase in AIC indicates that the model without TWI variables has a poorer fit, highlighting the explanatory power of TWI-related factors in determining the price. The effect can also be observed in the R^2 coefficient which decreases by 0.045 percentual points. It is important to highlight that the small value of R^2 arises from the lack of variables that could better explain the phenomenon, such as seasonality, criminality, tourism trends, among others. This fact is further confirmed with the residuals vs. fitted plot (linear model is appropriate and it explains a significant portion of the phenomenon).

To further analyze the impact of the TWI on prices, Figure 5.17 presents a comparison of prices as a function of the TWI for hotel classes 1/2, 3 and 4/5. In this bar plot, the following variables are kept constant, which correspond to the median case: number of rooms in the second, third, and fourth quartiles (ranging from 5 to 397), a Bayesian score of 4.42, and number of reviews of 217. The graph clearly shows a tendency for prices to increase as walkability improves, especially when comparing Poor level to the others. For Basic, High, and Excellent levels the difference in price is not substantial between them.

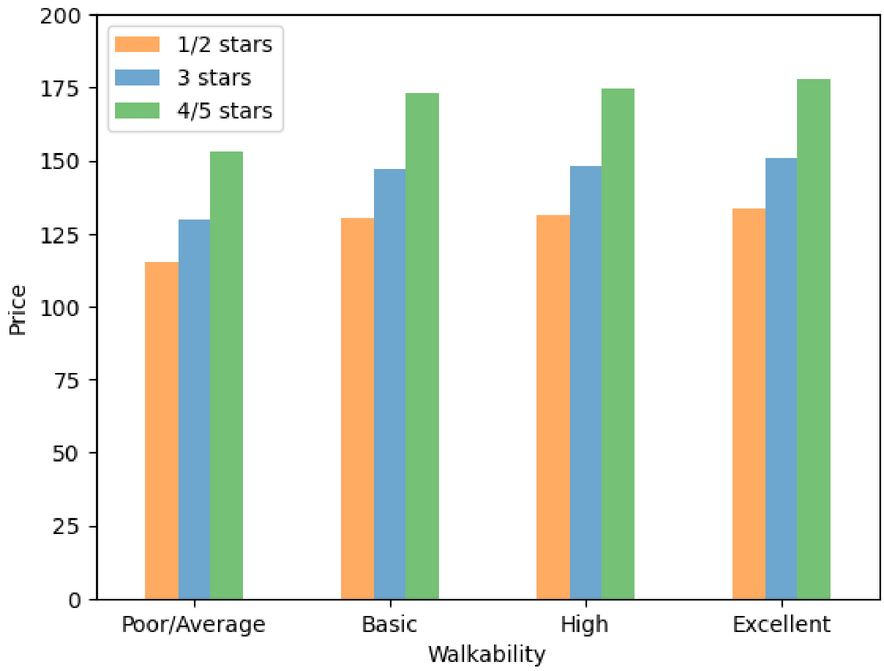


Figure 5.17: Comparison of prices as a function of the TWI and class for hotels in Naples, Italy.

5.1.7 Price Analysis for BnBs

Multiple Linear Regression

The process starts by calculating the log (natural) of the price to stabilize the variance of the residuals and improve linearity. After this, outliers are removed by keeping the data that falls between $Q_1 - 1.5 * \text{IQR}$ and $Q_3 + 1.5 * \text{IQR}$, where Q_1 and Q_3 are the first and third quartiles, and IQR is the interquartile range.

The considered categories are "Entire home/apt" and "Private room" for room type. The property type categories considered are "Entire rental unit" ($2,741$ samples), "Entire condo" ($1,093$), "Private room in bed and breakfast" (717), "Entire home" (565), "Private room in rental unit" (558), "Entire vacation home" (287), "Private room in condo" (205), "Entire loft" (104), and "Others" (412). The neighbourhood categories considered are San Lorenzo ($1,217$ samples), Pendino (671), San Ferdinando (625), Montecalvario (520), Chiaia (509), Avvocata (449), San Giuseppe (401), Porto (356), Stella (333), San Carlo all'Arena (319), Vomero (292), Vicaria (176), Arenella (153), Mercato (124), Fuorigrotta (122), Posillipo (116), and Others (299). Other variables considered are bathrooms (1, 2, and 3 or more), bedrooms (1, 2, and 3 or more), and beds (1, 2, 3, and 4 or more). This process leaves $6,528$ accommodations for processing.

The Spearman correlation between TWI and the Location Score is 0.34, so it is concluded that some of the preferences of the users regarding the walkability of the place are captured by the index.

The variables considered are displayed in Table 5.6. The TWI is converted into the categorical variable Walkability_num to analyze the impact of the different walkability levels separately. Afterwards, numerical variables are standardized and the procedure from Section 4.9 is followed. This leads to the results reported in Table 5.7.

In Model 1, the results suggest that higher walkability is associated with higher prices, indicating that walkability has an impact on pricing. Neighborhood and property type appear to influence accommodation prices in varying ways. Additionally, the number of accommodates, bathrooms, and bedrooms are positively correlated with price, likely serving as proxies for property size. As expected, the coefficients for the number of bathrooms and bedrooms increase as their counts rise. Furthermore, the Bayesian Score is positively associated with accommodation prices.

As it was mentioned in Section 4.9, linearity, homoscedasticity, and normality assumptions are assessed. Figure 5.18 shows the residuals vs. fitted plot. Given that the residuals appear to be randomly scattered around the red zero line and no distinguishable pattern, it suggests that the linearity assumption holds well. Figure 5.19 shows the scale-location plot, where the the LOESS line (in violet) follows closely the reference line, suggesting that the assumption of

Metro_station	Closeness_Centrality_Attractions	Closeness_Centrality_Restaurants
amenities_count	Review_count_H	Score_H
Bayesian_avg_H	location_score	accommodates
minimum_nights	maximum_nights	availability_365
review_scores_accuracy	review_scores_cleanliness	review_scores_checkin
review_scores_communication	review_scores_value	TVs
Wifi	Kitchen_Appliances	Parking
Child_Amenities	Bathroom_Amenities	Entertainment_Systems
Miscellaneous	Bayes_mean_A	Bayes_mean_R
TWI	neighbourhood_cleansed_0	neighbourhood_cleansed_1
neighbourhood_cleansed_2	neighbourhood_cleansed_3	neighbourhood_cleansed_4
neighbourhood_cleansed_5	neighbourhood_cleansed_6	neighbourhood_cleansed_7
neighbourhood_cleansed_8	neighbourhood_cleansed_9	neighbourhood_cleansed_10
neighbourhood_cleansed_11	neighbourhood_cleansed_12	neighbourhood_cleansed_13
neighbourhood_cleansed_14	neighbourhood_cleansed_15	neighbourhood_cleansed_16
property_type_0	property_type_1	property_type_2
property_type_3	property_type_4	property_type_5
property_type_6	property_type_7	property_type_8
room_type_0	room_type_1	host_is_superhost_0
host_is_superhost_1	bathrooms_1	bathrooms_2
bathrooms_3	bedrooms_1	bedrooms_2
bedrooms_3	beds_1	beds_2
beds_3	beds_4	Walkability_num_1
Walkability_num_2	Walkability_num_3	Walkability_num_4
Walkability_num_5		

Table 5.6: Considered variables for BnB analysis in Naples, Italy.

homoscedasticity is reasonable. For the normality, Figure 5.20 shows the QQ-plot, where most of the residuals closely follow the red reference line, suggesting that linearity holds well.

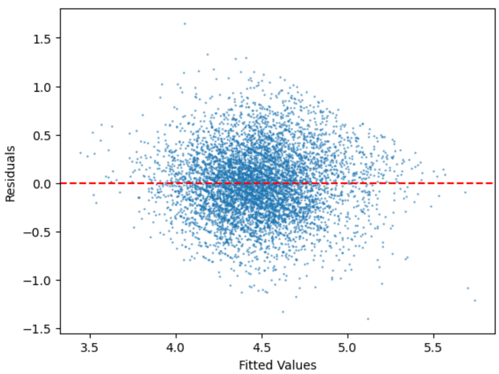


Figure 5.18: Residuals vs. fitted plot for BnBs in Naples, Italy.

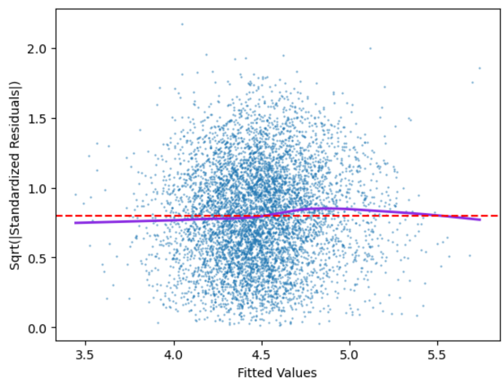


Figure 5.19: Scale-location plot for BnBs in Naples, Italy.

In Model 2, the impact of the TWI is assessed by removing the associated variables from Model 1. This results in the AIC increasing from $4,881$ to $5,023$, confirming the importance of the TWI in explaining the variation in price. The increase in AIC indicates that the model

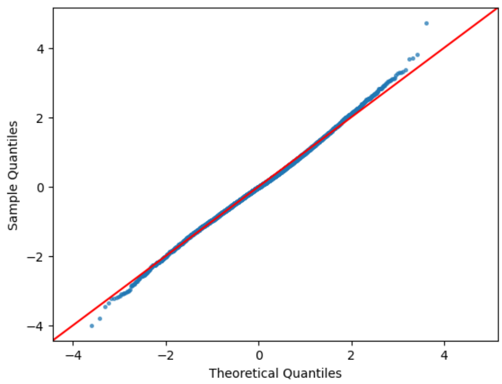


Figure 5.20: QQ-plot for the BnBs in Naples, Italy.

without TWI variables has a slightly poorer fit, highlighting the explanatory power of TWI-related factors in determining the price. The effect can also be observed in the R^2 coefficient which decreases by 0.014 percentual points. It is important to highlight that the small value of R^2 arises from the lack of variables that could better explain the phenomenon, such as seasonality, criminality, tourism trends, property size, among others. This fact is further confirmed with the residuals vs. fitted plot (linear model is appropriate and it explains a significant portion of the phenomenon), and by fitting more complex models that do not assume specific data distributions, such as regression trees and GAMs, which report similar metrics (all this in the case of BnBs that present significantly more number of samples).

To further analyze the impact of the TWI on prices, Figure 5.21 presents a comparison of prices as a function of the TWI. In this bar plot, the following variables are kept constant, which corresponds to the median case: Private room in rental unit located in Pendino neighbourhood (second municipality), single bathroom single bedroom, parking spot available, no child amenities, 16 reviews, a cleanliness review score of 4.88, Bayesian score of 4.74, up to four accommodates, availability of 237 days in the following 365, and a minimum number of nights of 2. The graph clearly shows a tendency for prices to increase as walkability improves.

Regression Tree

To confirm the findings regarding the significance and impact of TWI on price, a regression tree model is used. A grid search is performed to identify the optimal minimum samples per leaf that maximizes the R^2 metric, utilizing 5-fold cross-validation. Within the range of 20 to 50, the optimal minimum samples per leaf is determined to be 32. The mean and standard deviation of

	Model 1	Model 2
Intercept	4.3190*** (0.014)	4.4202*** (0.012)
Walkability_num_2	0.0891*** (0.015)	-
Walkability_num_3	0.1239*** (0.015)	-
Walkability_num_4	0.1459*** (0.014)	-
Walkability_num_5	0.1808*** (0.018)	-
Bayesian_avg_H	0.0543*** (0.004)	0.0542*** (0.004)
neighbourhood_cleansed_0	−0.0787*** (0.023)	−0.1622*** (0.022)
neighbourhood_cleansed_1	−0.0925*** (0.022)	−0.1484*** (0.021)
neighbourhood_cleansed_7	0.1423*** (0.016)	0.1792*** (0.016)
neighbourhood_cleansed_9	0.0647** (0.021)	0.1327*** (0.020)
neighbourhood_cleansed_10	0.2692*** (0.035)	0.1803*** (0.034)
neighbourhood_cleansed_11	−0.0900*** (0.021)	−0.1140*** (0.021)
neighbourhood_cleansed_12	0.2165*** (0.018)	0.1586*** (0.017)
neighbourhood_cleansed_13	0.1875*** (0.023)	0.0966*** (0.022)
neighbourhood_cleansed_16	0.1102*** (0.022)	0.1890*** (0.019)
property_type_1	−0.0664** (0.022)	−0.0536* (0.022)
property_type_2	−0.1144*** (0.017)	−0.1186*** (0.018)
property_type_3	0.1097*** (0.016)	0.1036*** (0.016)
property_type_4	0.0864*** (0.019)	0.0866*** (0.020)
property_type_5	−0.0774*** (0.017)	−0.0757*** (0.017)
property_type_6	−0.1311*** (0.026)	−0.1378*** (0.026)
property_type_8	0.0271* (0.013)	0.0310* (0.013)
bathrooms_2	0.1171*** (0.013)	0.1218*** (0.014)
bathrooms_3	0.1913*** (0.031)	0.1999*** (0.032)
bedrooms_2	0.0945*** (0.013)	0.0851*** (0.013)
bedrooms_3	0.1883*** (0.024)	0.1740*** (0.024)
Parking	0.0326*** (0.010)	0.0174 (0.010)
Child_Amenities	0.0672*** (0.010)	0.0709*** (0.010)
accommodates	0.1282*** (0.007)	0.1307*** (0.007)
availability_365	0.0816*** (0.007)	0.0785*** (0.007)
minimum_nights	−0.0111*** (0.003)	−0.0101** (0.003)
Review_count_H	−0.0605*** (0.003)	−0.0573*** (0.003)
review_scores_cleanliness	0.0264*** (0.004)	0.0255*** (0.004)
AIC	4,881	5,023
R^2	0.376	0.362
Adj.R^2	0.373	0.359
N	6,528	6,528

Table 5.7: Multiple linear regression results for BnBs with and without TWI for Naples, Italy.

the R^2 score per fold are assessed to prevent overfitting. This hyperparameter selection results in an R^2 of 0.423 (Adj. R^2 of 0.419) when the model is fitted to the entire dataset.

To specifically evaluate the impact of TWI on price, the SHAP summary plot from Figure 5.22 is used. The plot indicates that TWI has a substantial impact on price, with SHAP values ranging approximately from -15 to $+15$ price units. This range highlights TWI as an important feature in the model.

The plot also reveals that higher TWI values are generally associated with higher SHAP values, indicating a positive correlation between walkability and price. Conversely, lower TWI values tend to correspond with lower SHAP values, suggesting that improvements in walkabil-

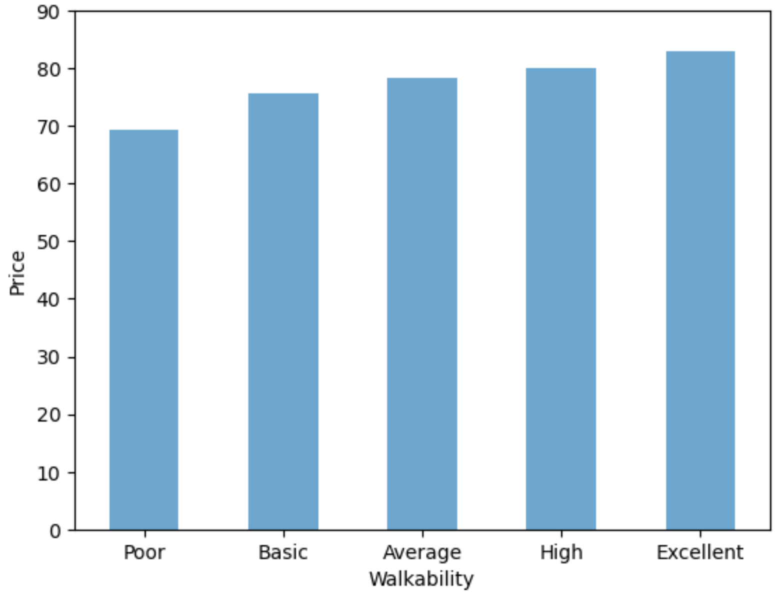


Figure 5.21: Comparison of prices as a function of the TWI for BnBs in Naples, Italy.

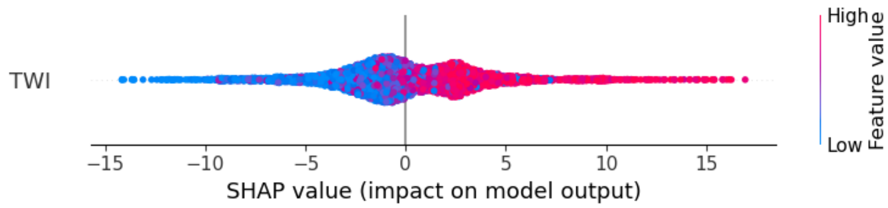


Figure 5.22: SHAP summary plot of the variable TWI for BnBs in Naples, Italy.

ity positively affect the price.

Additionally, the clusters in the plot show that, in general, lower TWI values have less impact on price, as they are closer to the center. In contrast, higher TWI values, which are farther from the center, have a more significant impact.

Finally, after fitting this model, which does not rely on any assumptions about the data distribution, and obtaining a similar R^2 value, it can be inferred that the current set of variables is insufficient to fully explain the observed variability in the data.

5.2 Venezia

5.2.1 Description

The second city under study corresponds to Venice (Santa Lucia), Italy. The network consists of 5, 520 nodes and 14, 220 edges. The considered walking network is displayed in Figure 5.23.

Figure 5.23: Considered walking network of Venice, Italy.

The city has 579 attractions whose review count and score are different from zero. Among these, 564 are retained because they have a Bayesian Score of 4 or higher, which is arbitrarily considered indicative of good quality for attractions. Figure 5.24 shows the location of each one of these attractions. The list of the top-10 attractions is presented in Table 5.8.

The city has 942 restaurants whose review count and score are different from zero. Among these, 176 are retained because they have a Bayesian Score of 4 or higher, which is arbitrarily considered indicative of good quality for restaurants. Figure 5.25 shows the location of each one of these restaurants. The list of the top-10 restaurants is presented in Table 5.9.

For the vaporetto, Table 5.10 presents the list of vaporetto stops considered, which are displayed in Figure 5.26. These stops belong to Line 1, and 20 out of 21 are considered (Lido

S.M.E. lies outside the area of analysis).

Ranking	**Name**	**Score**	**Review count**	**Bayesian score**
1	Centro Storico di Venezia	5.0	$2,218$	4.859084
2	Gioielleria Eredi Jovon	5.0	476	4.573338
3	Canal Grande	4.5	$41,167$	4.495788
4	Piazza San Marco	4.5	$36,935$	4.495311
5	Basilica di San Marco	4.5	$28,570$	4.493956
6	Campanile di San Marco	4.5	$9,064$	4.481476
7	Collezione Peggy Guggenheim	4.5	$8,433$	4.480150
8	Cannaregio	4.5	$4,481$	4.464023
9	Teatro La Fenice	4.5	$4,409$	4.463482
10	Basilica Santa Maria Gloriosa dei Frari	4.5	$4,170$	4.461565

Table 5.8: Top-10 attractions of Venice, Italy, according to the Bayesian score criteria.

Ranking	**Name**	**Score**	**Review count**	**Bayesian score**
1	Trattoria Al Gazzettino	4.5	$12,295$	4.477414
2	Da Mamo	4.5	$3,696$	4.430527
3	Ristorante Florida	4.5	$3,640$	4.429575
4	Bistrot de Venise	4.5	$3,600$	4.428879
5	Trattoria Bar Pontini	4.5	$3,397$	4.425122
6	Osteria La Zucca	4.5	$3,060$	4.417925
7	Osteria Al Squero	4.5	$2,804$	4.411461
8	Trattoria Alla Ferrata	4.5	$2,750$	4.409965
9	Riviera	4.5	$2,749$	4.409937
10	Spaghetteria 6342 A Le Tole Pizzeria	4.5	$2,569$	4.404560

Table 5.9: Top-10 restaurants of Venice, Italy, according to the Bayesian score criteria.

P.le Roma F	Ferrovia b	Riva de Biasio
S. Marcuola-Casino' B	San Stae	Ca' D'Oro
Rialto Mercato	Rialto	S. Silvestro
S. Angelo	S. Toma' A	Ca' Rezzonico
Accademia	S. Maria del Giglio	Salute
S. Marco	S. Marco-San Zaccaria	Arsenale
Giardini A	S. Elena-Stadio Penzo C	

Table 5.10: Considered vaporetto stops for Venice, Italy.

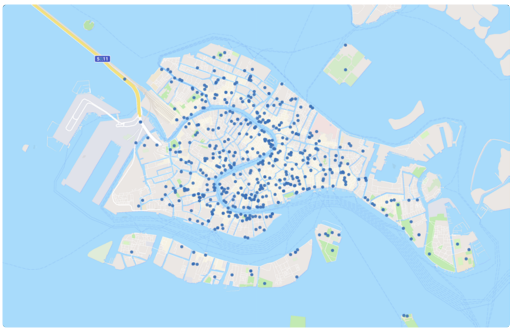


Figure 5.24: Considered attractions for Venice, Italy.

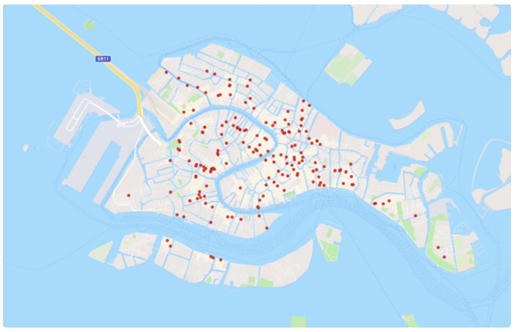


Figure 5.25: Considered restaurants for Venice, Italy.

Figure 5.26: Considered vaporetto stops for Venice, Italy.

5.2.2 Hotels

The city of Venice has 312 hotels, which are displayed in Figure 5.27. Most of these hotels are concentrated in districts San Marco and Castello. The distribution based on hotel class is shown in Figure 5.28, illustrating that the majority of the hotels are 3-star, followed by 4-star establishments. As a whole, the pattern distribution is similar to the one from Naples shown in Figure 5.6. Figure 5.29 provides a visual representation of the hotels on the map, with points colored according to their respective hotel class.

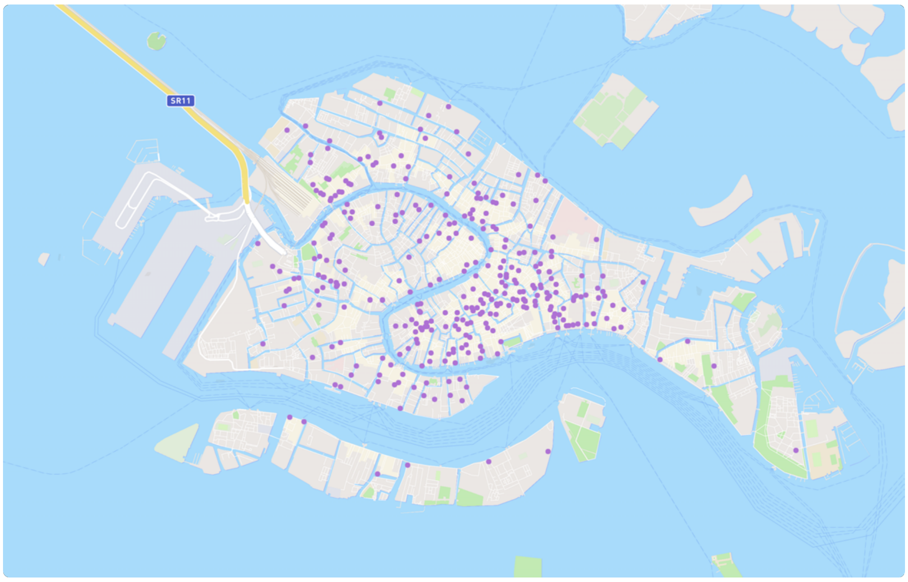


Figure 5.27: Hotels location in Venice, Italy.

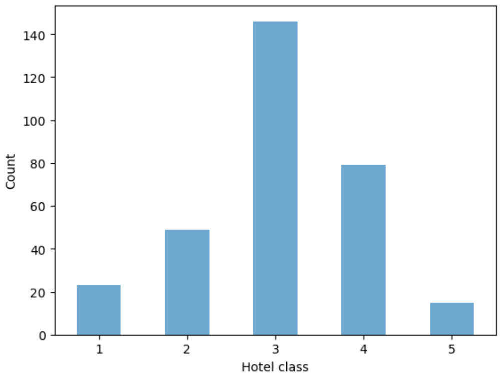


Figure 5.28: Hotels distribution by class in Venice, Italy.

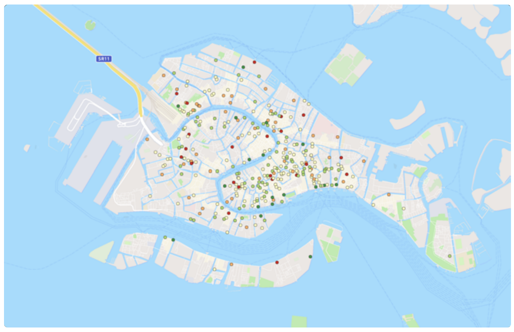


Figure 5.29: Hotels distribution by class on the map in Venice, Italy. Classes represented are 1 (red), 2 (orange), 3 (yellow), 4 (light green), and 5-stars (green).

5.2.3 BnBs

The city of Venice has 4, 851 AirBnBs, which are displayed in Figure 5.30. Unlike hotels, these accommodations are scattered around the city with more presence in the Castello, Cannaregio, and San Marco districts. The distribution based on accommodation type shows that around 87% of the BnBs are entire homes or apartments, and the rest private rooms.

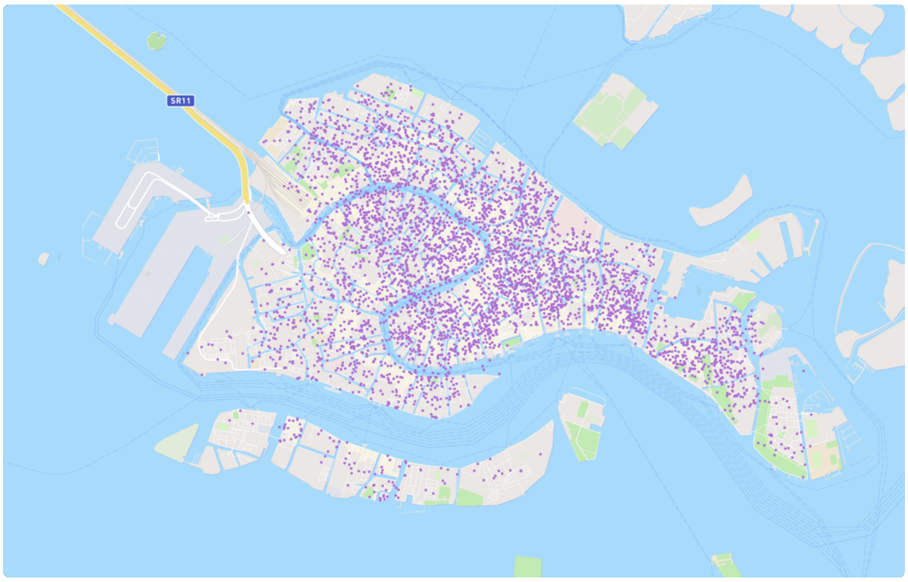


Figure 5.30: AirBnB locations in Venice, Italy.

5.2.4 Isochrone Maps

The isochrone maps from each accommodation are displayed in Figures 5.31 and 5.32. A range of 15 to 45 minutes is used to assess the reachable nodes within that time frame. In both cases, most of the attractions and restaurants in the city are within 15 minutes from at least one accommodation, which is a good indicator of the accessibility and convenience for visitors staying at these locations.

5.2.5 TWI Distribution

For each accommodation, the TWI is calculated. Figure 5.33 shows the normalized distribution of the TWI for both hotels and BnBs. The effect of the sparsity of the BnBs is evident in the comparison of the bars, where BnBs show a higher percentage of Poor and Basic walkability compared to hotels. In contrast, hotels, which are primarily located in the city center near most attractions and restaurants, exhibit a higher percentage of Average, High, and Excellent walkability.

Figures 5.34 and 5.35 provide a visual representation of the hotels and BnBs on the map, with points colored according to their respective TWI levels (two colors per level for better visualization). It is important to highlight that the TWI distribution observed is consistent with the distribution of attractions and restaurants in the city from Figures 5.24 and 5.25.

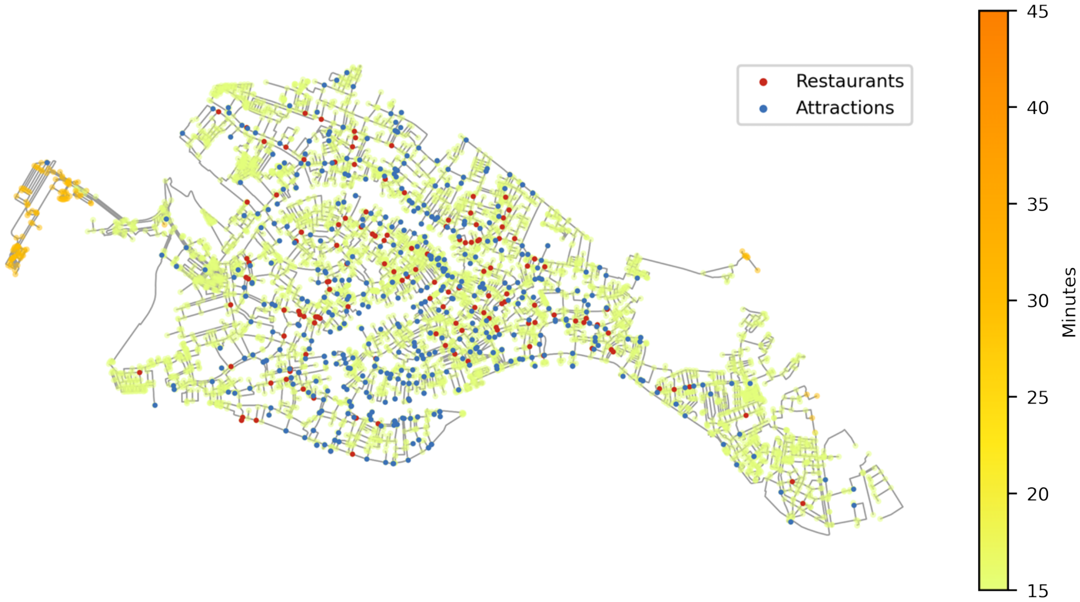


Figure 5.31: Isochrone map for hotels in Venice, Italy.

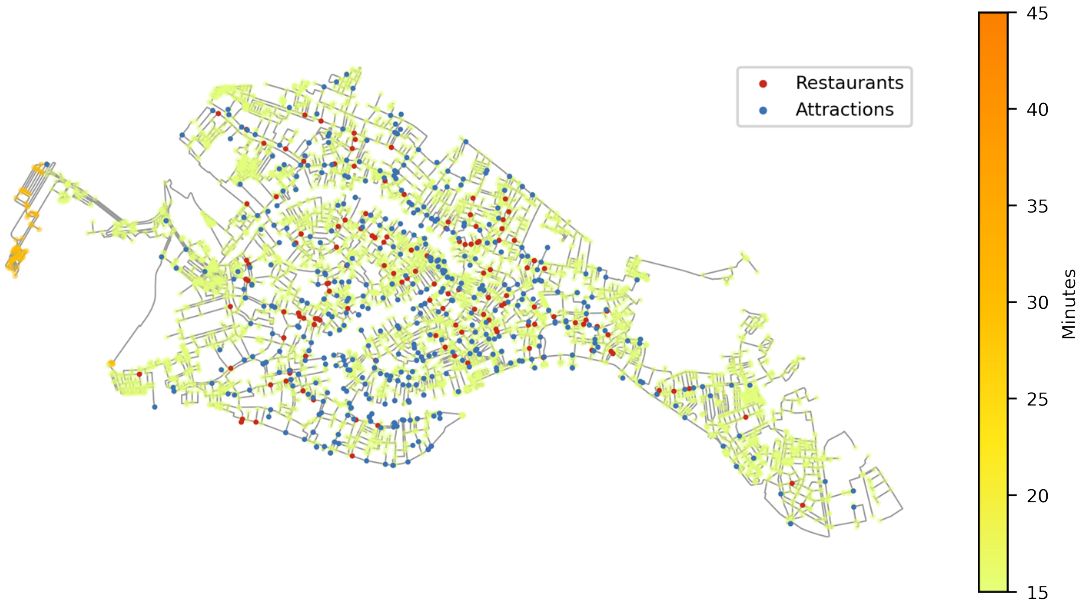


Figure 5.32: Isochrone map for BnBs in Venice, Italy.

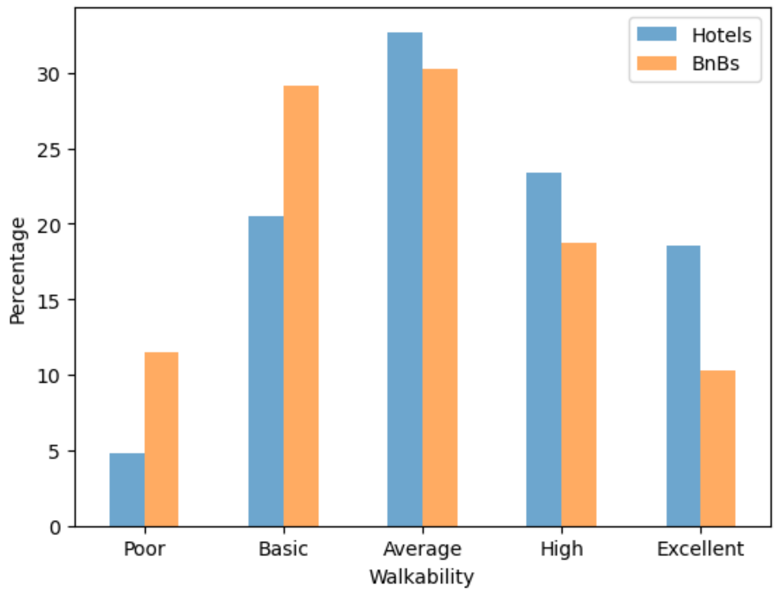


Figure 5.33: Distribution of accommodations according to its TWI of hotels and BnBs in Venice, Italy.

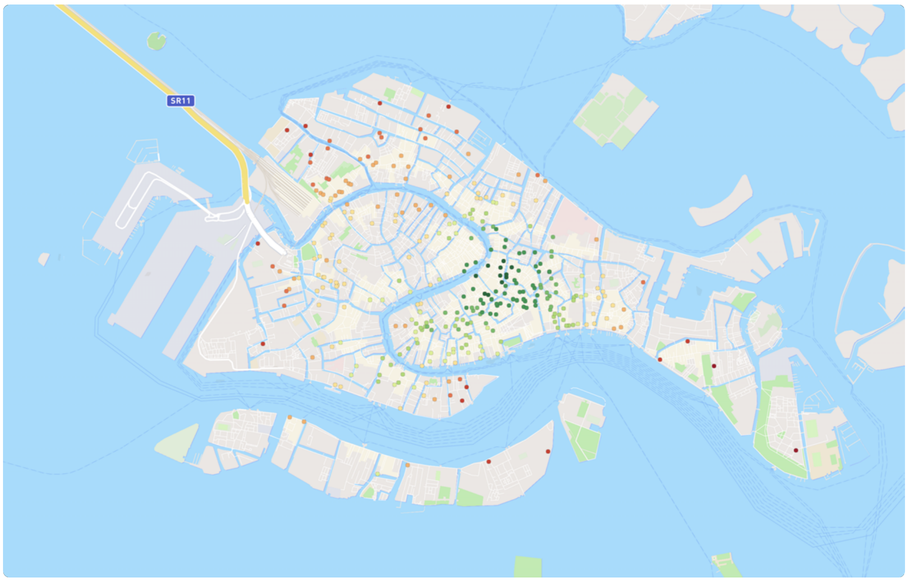


Figure 5.34: Hotels distribution by TWI level on the map in Venice, Italy. Levels represented are Poor (dark-red/red), Basic (orange-red/orange), Average (light-orange/yellow), High (light yellow-green/light green), and Excellent (green/dark-green).

5.2.6 Price Analysis for Hotels

The followed process similar to the one presented in the previous Section. In this case, 301 accommodations are left for processing.

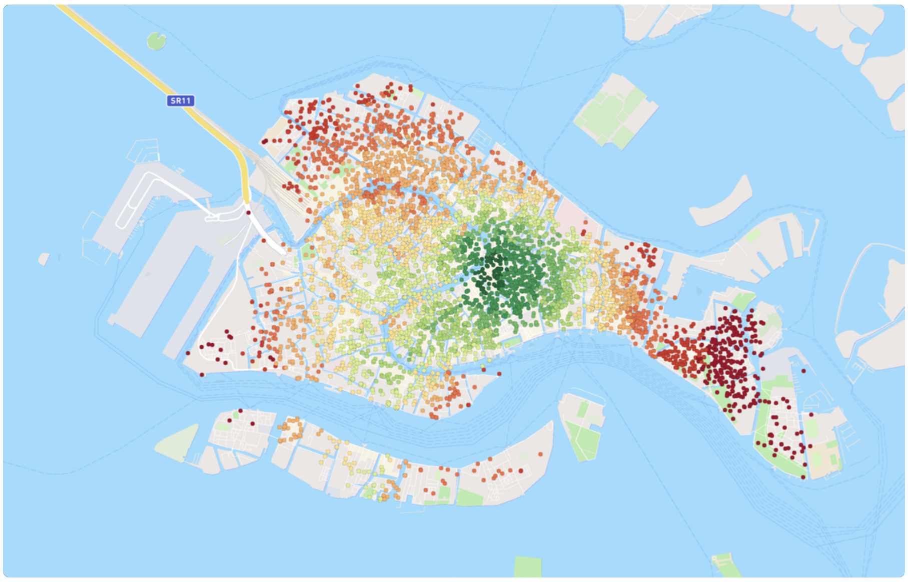


Figure 5.35: BnBs distribution by TWI level on the map in Venice, Italy. Levels represented are Poor (dark-red/red), Basic (orange-red/orange), Average (light-orange/yellow), High (light yellow-green/light green), and Excellent (green/dark-green).

The Spearman correlation between TWI and the Location Score is 0.39, so it is concluded that some of the preferences of the users regarding the walkability of the place are captured by the index.

The variables considered are the same as in the previous case. These are displayed in Table 5.4, leading to the results reported in Table 5.11. In this model, the TWI is eliminated during the process of backward elimination, even when considering shorter radii of 10, 7, and 5 minutes, which makes sense considering the size of the city. One possible explanation for this is that hotels are mainly located close to the main attractions and restaurants, resulting in no significant difference in price among them.

Hotel class is positively related to price, with higher star ratings indicating higher prices. This is expected, as hotels with more stars offer more amenities.

Hotel quality also appears to be an important factor, though it influences the price to a lesser extent. Additionally, unlike in Naples, the average restaurant quality does appear as a significant factor.

As it was mentioned in Section 4.9, linearity, homoscedasticity, and normality assumptions are assessed. Figure 5.36 shows the residuals vs. fitted plot. Given that the residuals show no clear pattern, it suggests that the linearity assumption holds well. Figure 5.37 shows the scale-location plot, where the fitted line shows a relatively flat behaviour, suggesting that there is no strong pattern of increasing or decreasing variance. For the normality, Figure 5.38 shows the QQ-plot, where most of the residuals closely follow the red reference line, confirming that the data is mostly linear.

	Model
Constant	5.3465***
	(0.030)
Hotel_class_3	0.1357***
	(0.036)
Hotel_class_45	0.4711***
	(0.041)
Bayesian_avg_H	0.0936***
	(0.016)
Bayes_mean_R	0.0361*
	(0.017)
AIC	18.40
R^2	0.441
Adj.R^2	0.433
N	301

Table 5.11: Multiple linear regression results for hotels in Venice, Italy.

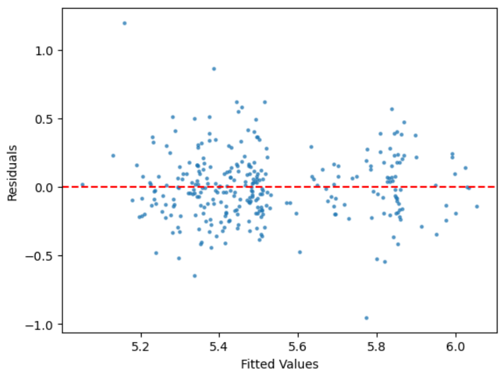


Figure 5.36: Residuals vs. fitted plot for hotels in Venice, Italy.

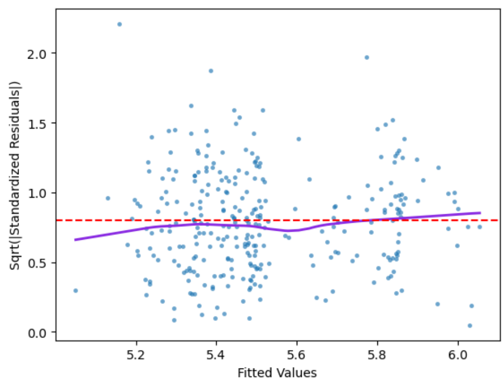


Figure 5.37: Scale-location plot for hotels in Venice, Italy.

Just as in the case of the previous city, it is important to highlight that the small value of R^2 arises from the lack of variables that could better explain the phenomenon, such as seasonality,

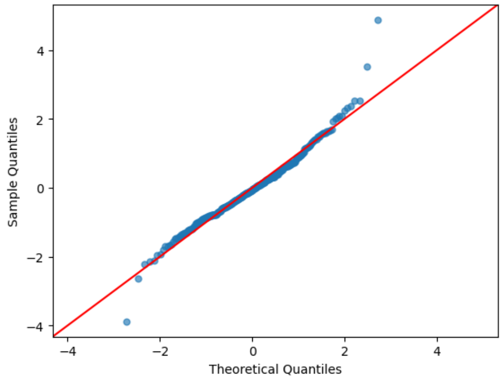


Figure 5.38: QQ-plot for the hotels in Venice, Italy.

criminality, tourism trends, among others. This fact is further confirmed with the residuals vs. fitted plot (linear model is appropriate and it explains a significant portion of the phenomenon).

5.2.7 Price Analysis for BnBs

Multiple Linear Regression

The same procedure from Section 5.1 is followed. The considered categories are "Entire home/apt" and "Private room" for room type. The property type categories considered are "Entire rental unit" (2, 924 samples), "Entire condo" (742), "Private room in bed and breakfast" (717), "Private room in rental unit" (210), "Entire home" (157), "Private room in bed and breakfast" (105), and "Others" (452). The neighbourhood categories considered are Castello (1, 211 samples), Cannaregio (1, 180), San Marco (705), San Polo (479), Santa Croce (475), and Others (594). This process leaves 4, 535 accommodations for processing.

The Spearman correlation between TWI and the Location Score is 0.27, so it is concluded that some of the preferences of the users regarding the walkability of the place are captured by the index.

The variables considered are displayed in Table 5.12. The TWI is converted into the categorical variable Walkability_num to analyze the impact of the different walkability levels separately. Afterwards, numerical variables are standardized and the procedure from Section 4.9 is followed. This leads to the results reported in Table 5.13.

Vaporetto_stop	Closeness_Centrality_Attractions	Closeness_Centrality_Restaurants
amenities_count	Review_count_H	Score_H
Bayesian_avg_H	location_score	accommodates
minimum_nights	maximum_nights	availability_365
review_scores_accuracy	review_scores_cleanliness	review_scores_checkin
review_scores_communication	review_scores_value	TVs
Wifi	Kitchen_Appliances	Parking
Child_Amenities	Bathroom_Amenities	Entertainment_Systems
Miscellaneous	Bayes_mean_A	Bayes_mean_R
TWI	neighbourhood_cleansed_0	neighbourhood_cleansed_1
neighbourhood_cleansed_2	neighbourhood_cleansed_3	neighbourhood_cleansed_4
neighbourhood_cleansed_5	property_type_0	property_type_1
property_type_2	property_type_3	property_type_4
property_type_5	room_type_0	room_type_1
host_is_superhost_0	host_is_superhost_1	bathrooms_1
bathrooms_2	bathrooms_3	bedrooms_1
bedrooms_2	bedrooms_3	beds_1
beds_2	beds_3	beds_4
Walkability_num_1	Walkability_num_2	Walkability_num_3
Walkability_num_4	Walkability_num_5	

Table 5.12: Considered variables for BnB analysis in Venice, Italy.

	Model 1	Model 2
Intercept	4.8539*** (0.018)	4.9482*** (0.013)
Walkability_num_2	0.0759*** (0.021)	-
Walkability_num_3	0.1354*** (0.016)	-
Walkability_num_4	0.1405*** (0.018)	-
Walkability_num_5	0.1505*** (0.021)	-
Bayesian_avg_H	0.0870*** (0.005)	0.0855*** (0.005)
TVs	0.0532*** (0.010)	0.0573*** (0.010)
Bathroom_Amenities	0.0246* (0.011)	0.0249* (0.011)
neighbourhood_cleansed_4	0.1128*** (0.015)	0.1530*** (0.013)
property_type_1	−0.1771*** (0.023)	−0.1764*** (0.023)
property_type_4	0.1090*** (0.031)	0.1148*** (0.032)
bathrooms_2	0.1653*** (0.012)	0.1733*** (0.012)
bathrooms_3	0.3535*** (0.028)	0.3654*** (0.028)
bedrooms_2	0.0611*** (0.008)	0.0631*** (0.011)
beds_2	−0.0236* (0.011)	−0.0262* (0.011)
beds_3	−0.0263* (0.013)	−0.0279* (0.013)
amenities_count	0.0434*** (0.008)	0.0401*** (0.008)
Review_count_H	−0.0972*** (0.005)	−0.0931*** (0.005)
location_score	0.0163*** (0.003)	0.0210*** (0.003)
accommodates	0.1628*** (0.007)	0.1619*** (0.007)
maximum_nights	0.0595*** (0.010)	0.0609*** (0.011)
availability_365	0.0708*** (0.008)	0.0747*** (0.008)
AIC	2,276	2,358
R^2	0.458	0.447
Adj.R^2	0.456	0.445
N	4,535	4,535

Table 5.13: Multiple linear regression results for BnBs with and without TWI for Venice, Italy.

In Model 1, just as in the case of Naples, the results suggest that higher walkability is associated with higher prices, indicating that it has an impact on pricing. Neighborhood San Polo appears to be significantly different from all the others. Additionally, the number of accommodates, bathrooms, and bedrooms are positively correlated with price, likely serving as proxies for property size. As expected, the coefficients for the number of bathrooms and bedrooms increase as their counts rise. Furthermore, the Bayesian Score is positively associated with accommodation prices.

As it was mentioned in Section 4.9, linearity, homoscedasticity, and normality assumptions are assessed. Figure 5.39 shows the residuals vs. fitted plot. Given that the residuals appear to be randomly scattered around the red zero line and no distinguishable pattern, it suggests that the linearity assumption holds well. Figure 5.40 shows the scale-location plot, where the the LOESS line (in violet) follows closely the reference line, suggesting that the assumption of homoscedasticity is reasonable. For the normality, Figure 5.41 shows the QQ-plot. This plot shows a slight tendency towards a right skewed distribution, yet most of the residuals closely follow the red reference line, so it is concluded that the linearity assumption holds relatively well.

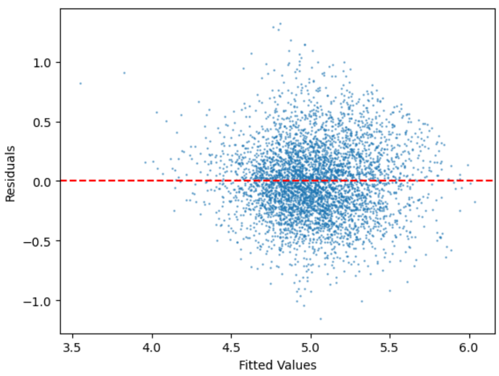


Figure 5.39: Residuals vs. fitted plot for BnBs in Venice, Italy.

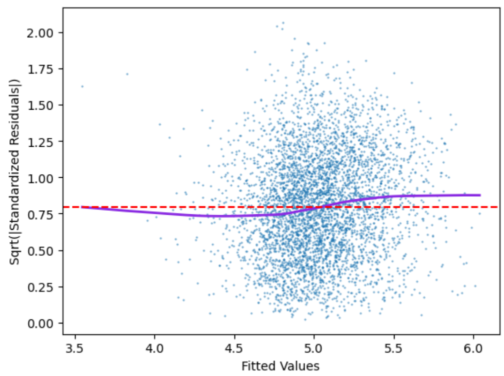


Figure 5.40: Scale-location plot for BnBs in Venice, Italy.

In Model 2, the impact of the TWI is assessed by removing the associated variables from Model 1. This results in the AIC increasing from $2,276$ to $2,358$, confirming the importance of the TWI in explaining the variation in price. The increase in AIC indicates that the model without TWI variables has a slightly poorer fit, highlighting the explanatory power of TWI-related factors in determining the price. The effect can also be observed in the R^2 coefficient which decreases by 0.011 percentual points. It is important to highlight that the small value of R^2 arises from the lack of variables that could better explain the phenomenon, such as seasonality, criminality, tourism trends, property size, among others. This fact is further confirmed

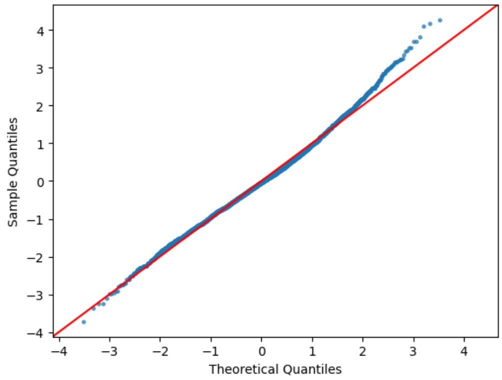


Figure 5.41: QQ-plot for the BnBs in Venice, Italy.

with the residuals vs. fitted plot (linear model is appropriate and it explains a significant portion of the phenomenon), and by fitting more complex models that do not assume specific data distributions, such as regression trees and GAMs, which report similar metrics (all this in the case of BnBs that present significantly more number of samples).

To further analyze the impact of the TWI on prices, Figure 5.42 presents a comparison of prices as a function of the TWI. In this bar plot, the following variables are kept constant, which corresponds to the median case: Property type different from Entire Home or Condo located in a neighbourhood different from San Polo, single bathroom (with amenities), one, three, or more bedrooms, one, four, or more beds, TV available, 28 amenities in total, 59 reviews, a location score of 4.92, Bayesian score of 4.75, up to four accommodates, availability of 185 days in the following 365, and a maximum number of nights of 365. The graph clearly shows a tendency for prices to increase as walkability improves, but it does not seem to be significantly different between walkability levels Average, High, and Excellent.

Regression Tree

To confirm the findings regarding the significance and impact of TWI on price, a regression tree model is used. A grid search is performed to identify the optimal minimum samples per leaf that maximizes the R^2 metric, utilizing 5-fold cross-validation. Within the range of 70 to 100, the optimal minimum samples per leaf is determined to be 71. The mean and standard deviation of the R^2 score per fold are assessed to prevent overfitting (although in this case there could be some overfitting, given that the proportion of the standard deviation to the mean is 12%). This hyperparameter selection results in an R^2 of 0.434 (Adj. R^2 of 0.429) when the

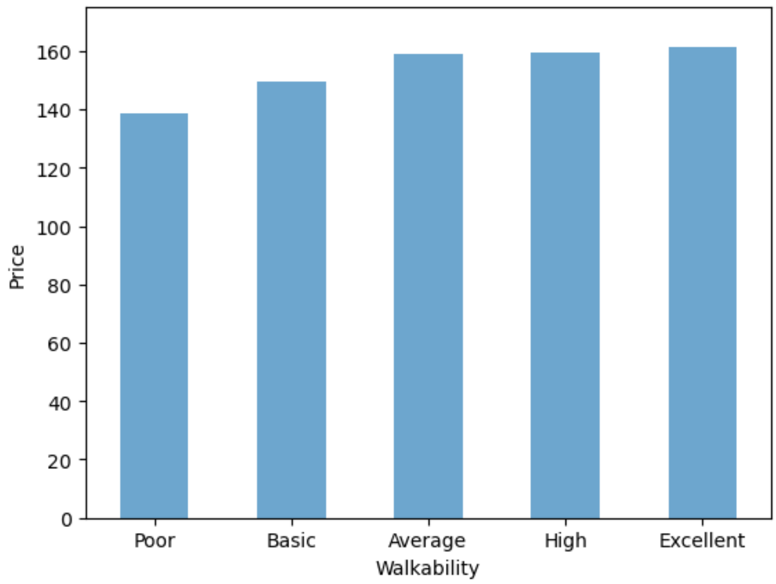


Figure 5.42: Comparison of prices as a function of the TWI for BnBs in Venice, Italy.

model is fitted to the entire dataset.

To specifically evaluate the impact of TWI on price, the SHAP summary plot from Figure 5.43 is used. The plot indicates that TWI has a substantial impact on price, with SHAP values ranging approximately from -20 to $+20$ price units. This range highlights TWI as an important feature in the model.

The plot also reveals that higher TWI values are generally associated with higher SHAP values, indicating a positive correlation between walkability and price. Conversely, lower TWI values tend to correspond with lower SHAP values, suggesting that improvements in walkability positively affect the price.

Additionally, the clusters in the plot indicate that both high and low walkability values impact the price similarly (positively and negatively, respectively), as they are at similar distances from the center.

Finally, after fitting this model, which does not rely on any assumptions about the data distribution, and obtaining a similar R^2 value, it can be inferred that the current set of variables is insufficient to fully explain the observed variability in the data.

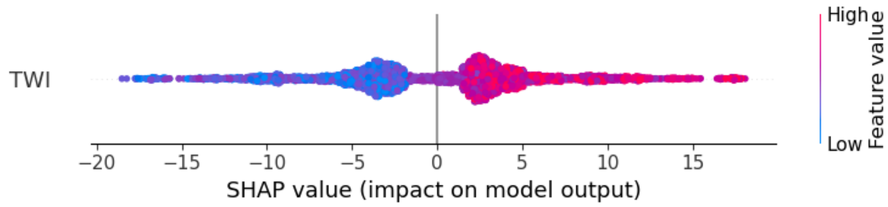


Figure 5.43: SHAP summary plot of the variable TWI for BnBs in Venice, Italy.

5.3 Comparison of Cities

5.3.1 Naples and Venice

To compare the cities and identify the most attractive one in terms of walkability, and to highlight the utility of the developed index, the approach used involves comparing the TWI distributions between cities at the hotel and BnB level. This method allows for a detailed analysis of how walkability varies between different types of accommodations in each city, providing valuable insights into which city offers better walkability for tourists.

Figure 5.44 illustrates the distribution of the categorical TWI at the hotel level. The results indicate that Naples has a significantly higher proportion of hotels with a Poor walkability level compared to Venice. This disparity may be attributed to the larger size of Naples, where hotels are more dispersed. Conversely, Venice exhibits a greater percentage of hotels with Average walkability. When examining High and Excellent walkability levels, Naples emerges as more attractive than Venice.

Figure 5.45 shows the distribution of the categorical TWI at the BnB level. The results are similar to the ones for hotels, as they indicate that Naples has a significantly higher proportion of BnBs with a Poor walkability level compared to Venice. Conversely, Venice exhibits a greater percentage of BnBs with Basic and Average walkability. When examining High and Excellent walkability levels, Naples emerges as more attractive than Venice.

To compare the cities from another perspective, Figures 5.46 and 5.47 present the box plots of the continuous TWI for each city at the hotel and BnB levels, respectively.

In these plots, a higher median walkability index implies that most accommodations are situated in areas with better access to attractions and restaurants. This can lead to a more convenient and enjoyable experience for tourists.

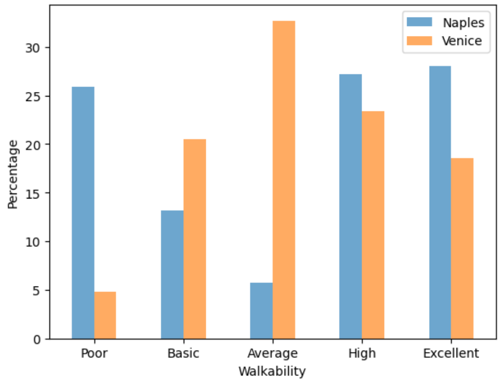


Figure 5.44: Comparison of TWI distribution (categorical) between cities at hotel level.

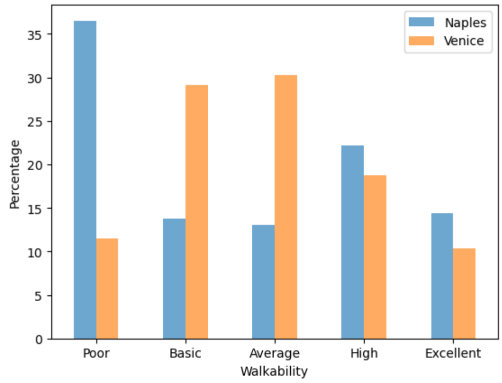


Figure 5.45: Comparison of TWI distribution (categorical) between cities at BnB level.

At the hotel level, 50% of the accommodations in Naples have a walkability score above 65. In contrast, in Venice, the same percentage of accommodations have a walkability score slightly above 50, making Naples more attractive. On the other hand, Naples exhibits greater variability in walkability scores compared to Venice, making Venice's walkability experience more consistent for tourists.

At the BnB level, 50% of the accommodations in Naples have a walkability score above 40. In contrast, in Venice, the same percentage of accommodations have a walkability score above 45 approximately, making Venice more attractive. In this case, Naples also exhibits greater variability in walkability scores compared to Venice, making Venice's walkability experience more consistent for tourists.

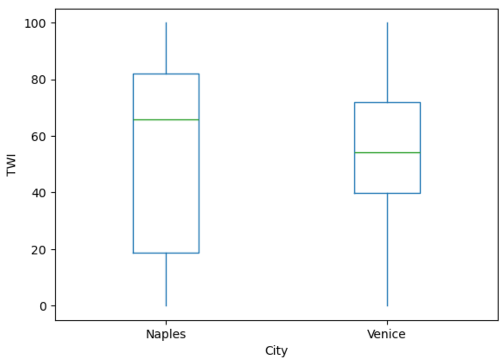


Figure 5.46: Comparison of TWI distribution (numerical) between cities at hotel level.

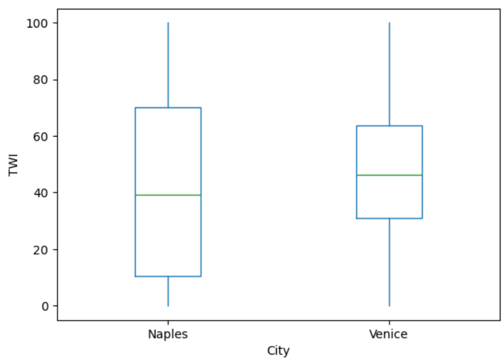


Figure 5.47: Comparison of TWI distribution (numerical) between cities at BnB level.

5.3.2 Insight from Larger Cities: Santiago, Chile

As an additional insight, a map of Santiago, Chile, is presented in Figure 5.48 to demonstrate how larger cities may exhibit multiple clusters of high walkability. Unlike the single walkability clusters observed in Venice and Naples, Santiago shows two distinct areas with high TWI values, indicating multiple attractive zones for tourists in terms of walkability.

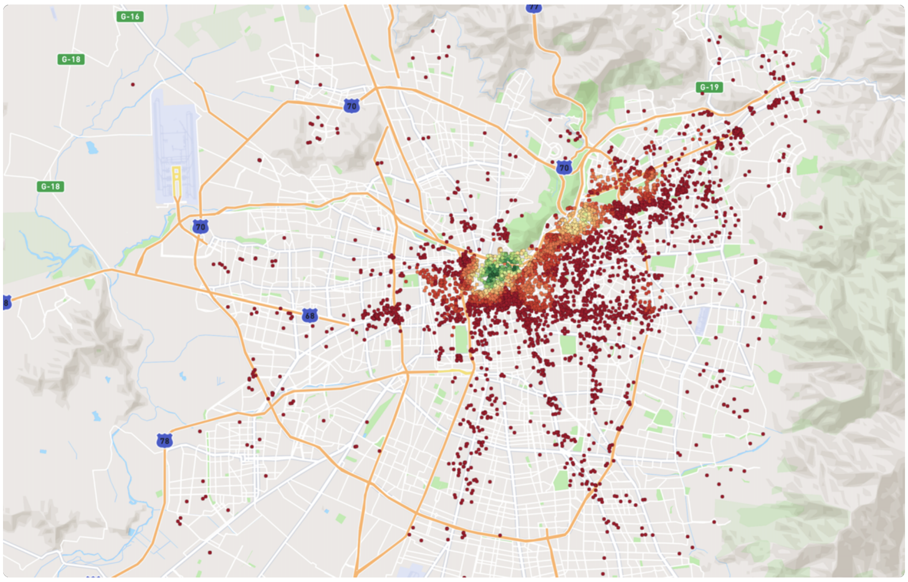


Figure 5.48: BnBs distribution by TWI level on the map in Santiago, Chile. Levels represented are Poor (dark-red/red), Basic (orange-red/orange), Average (light-orange/yellow), High (light yellow-green/light green), and Excellent (green/dark-green).

This suggests that in larger cities, tourist-friendly walkable areas may not be limited to a single concentrated zone but could be spread across different parts of the city.

6

Conclusions and Future Work

This chapter presents the conclusions of the study and outlines potential directions for future research.

6.1 Conclusions

The primary objective of this Thesis was to assess the walkability of urban areas surrounding tourist accommodations, such as hotels and BnBs, and to analyze how walkability impacts accommodation pricing. This objective was achieved by collecting and analyzing data on points of interest and accommodations, developing a Tourism Walkability Index, and examining its relationship with accommodation prices.

In particular, the conclusions are the following:

- Both Naples and Venice offer good accessibility, with most attractions and restaurants within a 15-minute walk from at least one accommodation.
- Hotels in both cities have higher walkability scores compared to BnBs, a difference attributed to the more centralized locations of hotels versus the more spread locations of BnBs.
- Venice's compact layout offers more consistent walkability, while Naples, being larger and more dispersed, exhibits greater variability. Specifically, Naples is more attractive at the hotel level, whereas Venice stands out at the BnB level. However, when focusing on

areas with High and Excellent walkability, Naples proves to be more appealing in both cases.

- Greater walkability is linked to higher accommodation prices in both cities, with more walkable accommodations having higher rates. The multiple linear regression models for both cities confirm that walkability, as measured by the TWI, is a statistically significant factor influencing prices. However, the models also suggest that other unconsidered factors, such as seasonality and crime rates, may further explain the variability in accommodation prices.

- Larger cities may naturally exhibit multiple clusters of high walkability accommodations due to their greater size and spatial complexity, offering tourists several distinct areas with easy access to key POI.

The findings emphasize the importance of walkable environments for tourists. Cities that prioritize accessibility to amenities can benefit economically, as walkable neighborhoods are associated with higher accommodation prices. For tourists, staying in walkable areas enhances convenience and the overall travel experience by providing easy access to attractions and restaurants, while also promoting better health and reducing pollution.

6.2 Future Work

This study offers insights into walkability and accommodation pricing, but there are areas for further exploration. Future research could enhance understanding in the following ways:

- **TWI Weights** : The weights used in defining the TWI could be refined by incorporating expert opinions or survey results. This would help the index better reflect tourists' preferences.

- **TWI Factors**: While the walkability index was designed to capture the accessibility to attractions and restaurants, it does not account for other walkability factors such as street safety, sidewalk slope, or presence of traffic lights, which could influence the actual walkability experience.

- ***t*-minute definition**: As demonstrated, various policies have been implemented globally to enhance urban walkability. It is logical that the radius for walkability changes depending on the size of the city, with larger cities often adopting policies like the 30-minute walkability framework. It would be valuable to explore how such policies affect both the walkability experience and accommodation prices within the city.

- **POI Quality**: In this work, POI with a Bayesian score of 4 or higher were defined as indicative of good quality for tourists. This could directly affect the number of locations included in the TWI calculation, making it worthwhile to analyze the impact of adjusting this threshold.

- **Walkability at country level**: The TWI could be used to analyze and compare the walkability experience of different accommodations at a country level. However, this would likely require acquiring and processing a large amount of data, which could pose a limitation.

References

[1] U. N. D. of Economic and S. Affairs, "International recommendations for tourism statistics 2008," United Nations, New York, Studies in Methods, 2010.

[2] M. Cisani and C. Rabbiosi, "Exploring tourism 'slow' mobilities," in *Reimagining Mobilities across the Humanities*, 1st ed. Routledge, 2023.

[3] R. Ohlan, "The relationship between tourism, financial development and economic growth in india," *Future Business Journal*, vol. 3, no. 1, pp. 9–22, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2314721017300063

[4] J. F. Perles-Ribes, A. B. Ramón-Rodríguez, A. Rubia, and L. Moreno-Izquierdo, "Is the tourism-led growth hypothesis valid after the global economic and financial crisis? the case of spain 1957–2014," *Tourism Management*, vol. 61, pp. 96–109, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0261517717300031

[5] S. Behzad, P. Hossein, and S. Razzaghi, "Assessing the dynamic economic impact of tourism for oic members," vol. 32, pp. 1098–1105, 01 2014.

[6] R. Bronzini, E. Ciani, and F. Montaruli, "Tourism and local growth in italy," *Regional Studies*, vol. 56, no. 1, pp. 140–154, 2022. [Online]. Available: https://doi.org/10.1080/00343404.2021.1910649

[7] World Travel Tourism Council. (2023) Share of travel and tourism's total contribution to gdp worldwide in 2019 and 2022, with a forecast for 2023 and 2033. [Online]. Available: https://www.statista.com/statistics/1099933/travel-and-tourism-share-of-gdp/

[8] United Nations World Tourism Organization. (2024) Number of international tourist arrivals worldwide from 1950 to 2023 (in millions). [Online]. Available: https://www.statista.com/statistics/209334/total-number-of-international-tourist-arrivals/

[9] ——. (2024) Number of international tourist arrivals worldwide from 2005 to 2023, by region (in millions). [Online]. Available: https://www.statista.com/statistics/186743/international-tourist-arrivals-worldwide-by-region-since-2010/

[10] *Unsustainable Transport: City Transport in the New Century*.

[11] S. Gössling, A. Choi, K. Dekker, and D. Metzler, "The social cost of automobility, cycling and walking in the european union," *Ecological Economics*, vol. 158, pp. 65–74, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921800918308097

[12] European Commission. Sustainable urban mobility. [Online]. Available: https://transport.ec.europa.eu/transport-themes/urban-transport/sustainable-urban-mobility_en

[13] L. Alessandretti, L. Natera Orozco, M. Saberi, M. Szell, and F. Battiston, "Multimodal urban mobility and multilayer transport networks," *Environment and Planning B: Urban Analytics and City Science*, vol. 50, no. 8, pp. 2038–2070, 2023. [Online]. Available: https://doi.org/10.1177/23998083221108190

[14] B. Olivari, P. Cipriano, M. Napolitano, and L. Giovannini, "Are italian cities already 15-minute? presenting the next proximity index: A novel and scalable way to measure it, based on open data," *Journal of Urban Mobility*, vol. 4, p. 100057, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667091723000134

[15] S. Claris, D. Scopelliti, C. Luebkeman, and J. Hargrave, "Cities alive: Towards a walking world," ARUP, London, Report, 2016.

[16] *Ways of Walking: Ethnography and Practice on Foot*.

[17] L. Alessandretti, P. Sapiezynski, S. Lehmann, and A. Baronchelli, "Multi-scale spatio-temporal analysis of human mobility," *PLOS ONE*, vol. 1, no. 2, pp. 3–5, 2021.

[18] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, pp. 462–5, 02 2006.

[19] M. C. Gonzalez, C. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–82, 07 2008.

[20] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1177170

[21] B. Saelens, J. Sallis, and L. Frank, "Environmental correlates of walking and cycling: Findings from the transportation, urban design, and planning literatures," *Annals of behavioral medicine : a publication of the Society of Behavioral Medicine*, vol. 25, pp. 80–91, 02 2003.

[22] L. Varga, A. Kovács, G. Tóth, I. Papp, and Z. Néda, "Further we travel the faster we go," *PLOS ONE*, vol. 11, pp. 1–9, 02 2016. [Online]. Available: https://doi.org/10.1371/journal.pone.0148913

[23] D. M. Levinson, *The 30-Minute City: Designing for Access*. Network Design Lab, 2020. [Online]. Available: https://hdl.handle.net/2123/21630

[24] B. Caselli, M. Carra, S. Rossetti, and M. Zazzi, "Exploring the 15-minute neighbourhoods. an evaluation based on the walkability performance to public facilities," *Transportation Research Procedia*, vol. 60, pp. 346–353, 2022, new scenarios for safe mobility in urban areasProceedings of the XXV International Conference Living and Walking in Cities (LWC 2021), September 9-10, 2021, Brescia, Italy. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352146521009455

[25] A. Forde and J. Daniel, "Pedestrian walking speed at un-signalized midblock crosswalk and its impact on urban street segment performance," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 8, no. 1, pp. 57–69, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S209575641830415X

[26] T. J. Gates, D. A. Noyce, A. R. Bill, and N. Van Ee, "Recommended walking speeds for timing of pedestrian clearance intervals based on characteristics of the pedestrian population," *Transportation Research Record*, vol. 1982, no. 1, pp. 38–47, 2006. [Online]. Available: https://doi.org/10.1177/0361198106198200106

[27] C. Moreno, Z. Allam, D. Chabaud, C. Gall, and F. Pratlong, "Introducing the "15-minute city": Sustainability, resilience and place identity in future post-pandemic cities," *Smart Cities*, vol. 4, no. 1, pp. 93–111, 2021. [Online]. Available: https://www.mdpi.com/2624-6511/4/1/6

[28] C. Hall and Y. Ram, "Measuring the relationship between tourism and walkability?: Walk score and english tourist attractions," *Journal of Sustainable Tourism*, vol. 27, 01 2018.

[29] Victoria State Department of Transport and Planning. Plan melbourne 2017-2050. [Online]. Available: https://www.planning.vic.gov.au/guides-and-resources/strategies-and-initiatives/plan-melbourne

[30] L. E. Thornton, R. Schroers, K. Lamb *et al.*, "Operationalising the 20-minute neighbourhood," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 19, no. 15, 2022.

[31] City of Portland, Oregon. (2012) The portland plan. [Online]. Available: https://www.portland.gov/bps/planning/about-bps/portland-plan#toc--about-the-plan

[32] S. Dihingia, M. Gjerde, and B. Vale, "Walking tourist: Review of research to date," *Journal of Urban Planning and Development*, vol. 148, no. 2, p. 04022017, 2022. [Online]. Available: https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29UP.1943-5444.0000829

[33] K. Thompson, "Urban transport networks and overseas visitors: analysis of the factors affecting usage and the implications for destination management," 2003. [Online]. Available: https://api.semanticscholar.org/CorpusID:260623127

[34] J. Farkic, D. Perić, M. Lesjak, and M. Petelin, "Urban walking: Perspectives of locals and tourists," *Geographica Pannonica*, vol. 19, pp. 212–222, 12 2015.

[35] C. Rabbiosi and S. Meneghello, "Questioning walking tourism from a phenomenological perspective: Epistemological and methodological innovations," *Humanities*, vol. 12, no. 4, 2023. [Online]. Available: https://www.mdpi.com/2076-0787/12/4/65

[36] D.-T. Le-Klähn, J. Roosen, R. Gerike, and C. M. Hall, "Factors affecting tourists' public transport use and areas visited at destinations," *Tourism Geographies*, vol. 17, no. 5, pp. 738–757, 2015. [Online]. Available: https://doi.org/10.1080/14616688.2015.1084527

[37] Transport for London. (2013) Visitor segmentation presentation. [Online]. Available: https://content.tfl.gov.uk/visitor-segmentation-research-report.pdf

[38] N. Shoval, B. McKercher, E. Ng, and A. Birenboim, "Hotel location and tourist activity in cities," *Annals of Tourism Research*, vol. 38, no. 4, pp. 1594–1612, 2011.

[39] S. Aksoy and M. Ozbuk, "Multiple criteria decision making in hotel location: does it relate to postpurchase consumer evaluations?" *Tourism Management Perspectives*, vol. 22, pp. 73–81, 2017.

[40] J. Gutiérrez, J. García-Palomares, G. Romanillos, and M. H. Salas-Olmedo, "The eruption of airbnb in tourist cities: Comparing spatial patterns of hotels and peer-to-peer accommodation in barcelona," *Tourism Management*, vol. 62, p. 278–291, 05 2017.

[41] F. Batista e Silva, M. A. Marín Herrera, K. Rosina, R. Ribeiro Barranco, S. Freire, and M. Schiavina, "Analysing spatiotemporal patterns of tourism in europe at high-resolution with conventional and big data sources," *Tourism Management*, vol. 68, pp. 101–115, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S026151771830044X

[42] Y. Ram, P. Bjork, and A. Weidenfeld, "Authenticity and place attachment of major visitor attractions," *Tourism Management*, vol. 52, pp. 110–122, 02 2016.

[43] C. Hall and Y. Ram, "Walk score tourist accommodation," *International Journal of Tourism Cities*, vol. 4, 04 2018.

[44] Redfin Corporation. Walk score methodology. [Online]. Available: https://www.walkscore.com/methodology

[45] L. Carr, S. Dunsiger, and B. Marcus, "Validation of walk scores for estimating access to walkable amenities," *British journal of sports medicine*, vol. 45, pp. 1144–8, 04 2010.

[46] OpenStreetMap, "Faq - openstreetmap wiki," 2023, accessed: 2024-05-27. [Online]. Available: https://wiki.openstreetmap.org/wiki/FAQ

[47] ——, "Open database license relicensing faq," 2023, accessed: 2024-05-27. [Online]. Available: https://wiki.openstreetmap.org/wiki/Open_Database_License_Relicensing_FAQ

[48] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using networkx," in *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Pasadena, CA USA, 2008, pp. 11–15.

[49] G. Boeing, "Modeling and analyzing urban networks and amenities with osmnx," 2024, working paper. [Online]. Available: https://geoffboeing.com/publications/osmnx-paper/

[50] F. Foti, P. Waddell, and D. Luxen, "A generalized computational framework for accessibility: From the pedestrian to the metropolitan scale," *Department of City and Regional Planning, University of California Berkeley*, February 2012.

[51] Selenium Contributors, "Selenium documentation," https://www.selenium.dev/documentation/, 2024, accessed: 2024-05-28.

[52] A. Bhayani, "Bayesian average," https://arpitbhayani.me/blogs/bayesian-average/, accessed: May 30, 2024.

[53] Fulmicoton, "Bayesian rating," https://fulmicoton.com/posts/bayesian_rating/, March 2013, accessed: May 30, 2024.

[54] Algolia, "Bayesian average," https://www.algolia.com/doc/guides/managing-results/must-do/custom-ranking/how-to/bayesian-average/#calculating-the-bayesian-constants, October 2023, accessed: May 30, 2024.

[55] X. Yang and Z. Zhang, "Combining prestige and relevance ranking for personalized recommendation," in *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, ser. CIKM '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 1877–1880. [Online]. Available: https://doi.org/10.1145/2505515.2507885

[56] R. Geisberger, P. Sanders, D. Schultes, and D. Delling, "Contraction hierarchies: Faster and simpler hierarchical routing in road networks," in *Experimental Algorithms*, C. C. McGeoch, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 319–333.

[57] J. Lazarsfeld, "Contraction hierarchies: An illustrative guide," https://jlazarsfeld.github.io/ch.150.project/, 2018, accessed: 06-2024.

[58] H. Bast, D. Delling, A. V. Goldberg, M. Müller-Hannemann, T. Pajor, P. Sanders, D. Wagner, and R. F. Werneck, "Route planning in transportation networks," in *Algorithm Engineering*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:14384915

[59] A. Jain and A. Kasbe, "Fake news detection," in *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, 2018, pp. 1–5.

[60] GIATA GmbH, "Giata - your provider for high quality hotel content and mapping services," 2024, accessed: 2024-05-28. [Online]. Available: https://www.giata.com/en/

[61] Hotelstars Union, "Criteria - hotelstars union," 2024, accessed: 2024-05-28. [Online]. Available: https://www.hotelstars.eu/criteria/

[62] InsideAirbnb, "Insideairbnb," https://insideairbnb.com/.

[63] L. D. Frank, J. F. Sallis, B. E. Saelens, L. Leary, K. Cain, T. L. Conway, and P. M. Hess, "The development of a walkability index: application to the neighborhood quality of life study," *British Journal of Sports Medicine*, vol. 44, no. 13, pp. 924–933, 2010. [Online]. Available: https://bjsm.bmj.com/content/44/13/924

[64] A. Bassiri Abyaneh, A. Allan, J. Pieters, and G. Davison, *Developing a GIS-Based Tourist Walkability Index Based on the AURIN Walkability Toolkit—Case Study: Sydney CBD*. Cham: Springer International Publishing, 2021, pp. 233–256. [Online]. Available: https://doi.org/10.1007/978-3-030-76059-5_13

[65] I. Jeong, M. Choi, J. Kwak, D. Ku, and S. Lee, "A comprehensive walkability evaluation system for promoting environmental benefits," *Scientific Reports*, vol. 13, no. 16183, 2023. [Online]. Available: https://www.nature.com/articles/s41598-023-43261-0

[66] ArcGIS Documentation, "How calculate composite index works," 2023, accessed: 2024-06-06. [Online]. Available: https://doc.arcgis.com/en/allsource/latest/analysis/geoprocessing-tools/spatial-statistics/how-calculate-composite-index-works.htm

[67] TripAdvisor, "Tripadvisor," https://www.tripadvisor.com.

[68] AirBnB, "Airbnb," https://www.airbnb.com/.

Acknowledgments

I want to thank my wife, Alejandra, for her constant support and helpful suggestions.

I am also deeply thankful to my parents, Elcira and Mario, and my sister, Rocío, for always being there, even from a distance, and extend the same thanks to my grandparents.

I would also like to thank Professor Francesco Silvestri for his guidance and assistance throughout this study.

This master's degree program was funded by the National Agency for Research and Development (ANID) through the Scholarship Program, Magíster en el Extranjero Becas Chile/2022 - 73220391.