UNIVERSITY OF TURIN

Ph.D. in Modeling and Data Science

XXXVI CYCLE

Final dissertation

UNIVERSITÀ
DI TORINO

Enhancing urban liveability and sustainability through geospatial large-scale data analytics

Supervisor: Prof. Rossano SCHIFANELLA

Candidate: Alice BATTISTON

ACADEMIC YEAR 2023/2024

Declaration of Authorship and Publications

I, Alice BATTISTON, declare that this thesis titled, "Enhancing urban liveability and sustainability through geospatial large-scale data analytics" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

The content of the dissertation is based on the technical report produced for the EU Horizon 2020 GoGreenRoutes project (Grant ID:869764) and the following publications:

Battiston, A., L. Napoli, P. Bajardi, A. Panisson, A. Perotti, M. Szell, and R. Schifanella. *Revealing the determinants of gender inequality in urban cycling with large-scale data.* EPJ Data Science 12, no. 1 (2023): 9.

Battiston, A., and R, Schifanella. *On the need for a multi-dimensional framework to measure accessibility to urban green.* npj Urban Sustainability 4, no. 1 (2024): 1-11.

Other publications and working papers not covered by the dissertation and produced during the Ph.D program are:

Battiston, A., E. Massaro, C. R. Binder, and R. Schifanella. *High spatial resolution dataset of La Mobilière insurance customers.* Scientific Data 9, no. 1 (2022): 81.

Montana, F., Pereira Barobza, E., Khomenko, S., Iungman, T., Mueller, N., Cirach, M., Daher, C., Chakraborty, TC, de Hoog, K., Battiston, A., Schifanella, R., Nieuwenhuijsen, M. *Healthy Urban Design Index (HUDI): How to Promote Health and Environmental Quality in European Cities.* [Submitted to The Lancet Planetary Health]

Sćepanović, S. Joglekar, S., Stephen, L., Zhou, K., Battiston, A., and Quercia, D. *Not All Forms of Vitamin N are Equal: Benefits of Different Forms of Urban Greenery for Health.* [Working paper]

Abstract

The rapid increase in urban populations and the growing availability of detailed, geo-referenced, large-scale data have led to the emergence of a new field known as the science of cities. This multidisciplinary field explores the complex phenomena arising from the structure and dynamics of urban environments to uncover the driving forces behind their evolution, growth, and broader dynamics. While current research typically offers a normative perspective, focusing on understanding these urban dynamics, the inclusion of cities and human settlements as key components of the 2015 Sustainable Development Goals (SDG 11) emphasises the need to integrate normative analyses with policy-oriented research. This integration is crucial for providing actionable insights and tools to enhance policymaking. In line with this approach, this dissertation focuses on developing tools to promote sustainable and healthy individual behaviours and provides insights for designing urban interventions that encourage these behaviours. The dissertation presents three distinct research contributions.

The first contribution investigates factors contributing to gender-specific barriers to cycling in Western cities using large-scale data from the sport-tracking application Strava. Previous academic research on gender differences in cycling has primarily consisted in stated-preferences studies relying on direct data collection, such as surveys. While these methods provide detailed insights into cycling preferences, their generalisability is limited by small sample sizes and geographical scope. By leveraging automatically collected data from a widely adopted sport-tracking application, this study aims to validate and expand upon the findings of earlier studies with an unprecedentedly large dataset. Through both macroscopic (cross-city) and microscopic (within-city) analyses, this research identifies a significant correlation between female cycling rates and the availability of safe cycling environments. These findings shed light on potential strategies to improve urban cycling conditions for women, thereby promoting more sustainable urban transport for all.

Aligning with the recommendations of the UN SDG 11.7, which calls for cities to provide access to safe, inclusive, and accessible greenspaces for all their residents, the second contribution is a computational framework to measure the accessibility of public greenspaces in urban areas. Due to computational constraints, evaluations of green exposure and accessibility are typically limited to a single metric. By contrast, our framework evaluates three families of green accessibility indicators encompassing metrics proposed in the academic literature as well as in the public policy domain. Through an analysis of population and area rankings generated by different indicators across more than 1,000 cities worldwide, I question the reliability of single-metric assessments in capturing the complexity of green accessibility within urban systems. The findings suggest that a single indicator may inadequately differentiate across areas or subgroups of the population, even when focusing on one form of green accessibility at a time. From a policy viewpoint, this indicates the need to switch to a multidimensional framework capable of organically evaluating a range of indicators at once. To enhance the usability of the computational framework, the associated interactive web interface, ATGreen, provides a range of functionalities designed for both the general public and policymakers specifically.

The third contribution continues the broader discussion on enhancing green accessibility but introduces a shift in focus from the structural aspects of green accessibility to strategies aimed at influencing individual behaviour to enhance exposure

to nature. To this scope, I introduce a novel routing system called ATGreenGO, designed to recommend nature-enriching walking routes with minimal detours compared to the shortest path, thus compatible with integration within daily routines. The validation of the system demonstrates that it balances the two dimensions —increased exposure to natural environments and overall duration of the suggested route—providing a significant increase in exposure while minimising the additional duration of the route.

Overall, the contributions presented in this dissertation combine advanced statistical and data analysis with digital tool development to enable a more comprehensive approach to the understanding and managing of urban phenomena and to inform the design of urban interventions to achieve the objectives outlined in the SDG 11.

Acknowledgements

Coming from an economic background, my journey towards a PhD in data science was particularly challenging at the beginning. I vividly remember struggling to connect to the remote server via SSH. Therefore, how could I not say my first *thank you* to my PhD advisor, Prof. Rossano Schifanella. He was always supportive and believed in me even when I struggled to believe in myself. His incredible patience and understanding were invaluable, especially when I lacked even the basic terminology that might have been straightforward for someone with his background. He also introduced me to an incredible group of collaborators, including Michael, Ludo, Paolo, and Linus. Here, I also want to express my gratitude to Prof. Laura Sacerdote. It was only during my third year, while serving as a student representative, that I truly realised the huge amount of work that goes into running a Ph.D. program. A sincere *thank you* also goes to Prof. Riccardo Di Clemente and Prof. Francisco Rowe, whose feedback greatly contributed to enhancing the final outcome of my PhD journey.

Who goes next? To ensure I don't overlook anyone, I'll proceed chronologically. My next *thank you* goes to London and my East London friends. The city itself deserves recognition; it's a place where you never feel too old to start a new adventure. Had I been in Italy in May 2020, I would have felt too old to start a Ph.D. at 28. But a city is mostly defined by the people you meet there. My East London friends were not only my best friends but also the ones that made me realise I was missing something. I remember one night, when everyone at the table except me had completed their Ph.D.s. Someone said "It was a challenge, but worth it!" and others even "really enjoyed the experience!". I distinctly remember Iac saying, "It was the most amazing period of my life". Maybe it was FOMO, but that conversation sparked my interest in pursuing a Ph.D. to gain the time to learn something new and transition from economics to urban data science. So, *thank you* Andre, Luisa, Fabfab, Jime, Vale, Lilli, Iac, Elif, and Jacopo. You were all there at that table and have continued to support me (remotely!) throughout this journey.

Since we are talking about the "good old times", how could I not say *thank you* to my CCA friends – Sarah, i Panorzi, Luca, MarcoIlBello (and more recently Nu!), Fede, Detective Marcus, Albi, and Ceci. You have been in my life for many years now, some in the same city and some on different continents, always showing your support, listening to my nonsense, and sharing our lives. Fede and Marcus, reconnecting with you in Torino is one of the things I treasure the most of these last four years.

I hope that one day I can settle with you all under that apricot tree.

It was October 2020 when I finally started the Ph.D. adventure. Despite the challenges of the ongoing pandemic, I made an effort to meet new friends and inspiring colleagues. And... it happened! *Thank you*, Dani, Gio, and Peter. Our dinners and tennis sessions (dalla Vecchia!) were invaluable and I am glad that our friendship continues despite the distance. To Duilio (whom I finally met in person only during my third year!), Narges, Valentina, and Chiara, we spent less time together, but it was no less meaningful. A *thank you* also goes to the guys of the *Bell Palace* office. Having in-person Ph.D. colleagues during my second year made my journey more real and much less isolating.

The new "stability" I had built in Torino during my second year made me hesitant about moving to Barcelona for my research visiting at BSC. I feared I would feel

lonely and isolated again. Barcelona was an exciting city, one I had loved during my Erasmus, but I was no longer 22 and felt very reluctant. I couldn't have been more wrong. The team at BSC was made of inspiring researchers and incredibly welcoming people. I am eternally grateful to Pato and Fernando for inviting me to spend three, then nine –'cause, I mean, I loved it–, months with the group. And Pato, you also introduced me to bouldering! On my last night at *Alice's garden* (named after Alice in Wonderland, not me!), I realised I had found a new family, and I knew I would miss everyone of you (Pato, Fer, Rens, Giovanni, Sol, Guille, Jero, Marta, Laura, Pause, Pau, David, and Carlos) very much. But Barcelona wouldn't have been the same without one of the most amazing flatmates anyone could have (yes, that's you Isa!), with whom I had some of the most sincere conversations of my life. And of course, Cali and Ra were also wonderful flatmates – except for THAT day, Ra, when you made it very hard for me to love you.

Some people come into your life and stay, while others eventually leave, my family has always been there and I am sure it will always be. *Thank you*, Ciottina, Papi, and Fede, for accepting my "all fine" even though you are used to have much longer conversations. That was all I needed.

And finally, Iac, I am not going to thank you. Not yet, at least. First, I owe you an apology. You are the one who had to survive my long conversations, my lack of self-confidence, my over-complicated thinking, my anxiety, my instability... what a nightmare I am! You have been there for all of this, even before there was an *us*. I will do better in the future. *Thank you* (now it is time for it) for the unconditional support of the last 8 years but also for encouraging me to grow into a better person. I couldn't be more excited about our –I mean, with Fred and Rillo too!– life together.

List of Abbreviations

EU	European Union
GHS-POP	Global Human Settlement - Population layer 2015 - revised version 2019A [1]
GHS-UCDB	Global Human Settlement - Urban Centre Database 2015 - revised version 2019A [2]
JRC	Joint Research Centre
NDVI	Normalized Difference Vegetation Index
OD	Origin-Destination
OSM	OpenStreetMap [3]
OSRM	Open Source Routing Machine [4]
OTP	OpenTripPlanner [5]
PGA	Public Green Area
PPGIS	Public Participation Geographic Information System
QS	Quality Score
SDG(s)	Sustainable Development Goal(s)
UC	Urban Centre
UN	United Nations
WC-ESA	World Cover data 2020 issued by the European Space Agency [6]
WHO	World Health Organization

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivations and background

The 21st century is witnessing an unprecedented wave of urbanisation. Currently, over half of the global population resides in urban areas, and projections of the United Nations (UN) indicate that this figure could rise to 70% within the next 25 years [7]. This rapid urban expansion has given rise to a new field of study known as the *science of cities*. This field encompasses a multidisciplinary approach to understanding the complex systems that define urban environments [8]. It integrates insights from urban planning, sociology, economics, environmental science, and data analytics to explore how cities grow, evolve, and function. The science of cities not only aims to understand historical transformations that have contributed to the definition of current cities but also seeks to employ advanced data analytics and modelling tools to optimise policy interventions to address the challenges associated with increasing urbanisation.

Contrary to the traditional view that urbanisation is inherently detrimental, there is growing recognition of its potential benefits [9]. Urban areas, when designed and managed effectively, can induce more sustainable models of living. Dense urban environments can reduce the per capita use of resources, lower energy consumption, and minimise the overall environmental footprint. For example, urbanisation can help limit soil erosion and preserve natural habitats by concentrating human activity within defined areas, thus sparing more land for nature [10]. Dense urban environments are also essential for the economic feasibility of public transportation systems [11]. Moreover, serving as hubs for innovation and efficiency, cities can drive the development and adoption of sustainable technologies and practices [12].

Despite the potential advantages of urbanisation, many cities struggle to combine economic productivity, social inclusion, and environmental sustainability. Research associated with the UN Sustainable Development Goals (SDGs) [13] indicates that while cities host just over half of the global population, they account for 60% to 80% of energy consumption and at least 70% of carbon emissions [14]. Additionally, cities often fail to provide social equality and sufficient living standards, with approximately 828 million urban residents (almost 21% of the 4.2 billion urban residents) estimated to live in slums, as of 2018 [7]. The acceleration of urbanisation in recent years has also been accompanied by a change in its form. While only 10 cities worldwide had more than 10 million inhabitants (defining a mega-city) in 1990, the number increased to 28 in 2014 and peaked at 34 in 2023. As most future urban expansion is projected to take place in developing countries, estimates indicate that 9 out of 10 future mega-cities will be in these regions [7].

Recognising the importance of shaping urbanisation to better address global challenges, the UN dedicated SDG 11 to the definition of objectives for the sustainable development of cities and urban communities, emphasising the need for cities to be inclusive, safe, resilient, and sustainable. SDG 11 outlines specific objectives in several key areas: housing provision, provision of basic services, access to transportation systems, environmental quality (including air quality and access to green spaces), and improving the resilience of cities to natural disasters. The integration of advanced data analytics and multidisciplinary approaches is crucial for monitoring progress towards the objectives set by SDG 11 and to inform future policy design. By leveraging the insights and tools provided by the science of cities, urban planners and policymakers can develop strategies that promote sustainable and liveable urban environments.

1.2 Aims and contributions

This dissertation presents three distinct research contributions within the domain of sustainable and liveable cities. The unifying element among these contributions is their alignment with the objectives and goals of the UN SDG 11, which focuses on making cities inclusive, safe, resilient, and sustainable. Collectively, this research aims to enhance our understanding of urban phenomena linked to this SDG while providing practical tools to advance progress towards its objectives.

The first contribution, presented in Chapter 2, is an analytical study that deepens our understanding of the determinants of gender inequalities in cycling. Cycling is widely recognised as a crucial component of sustainable urban mobility systems. However, cycling uptake globally remains below desired levels, with women exhibiting particularly low cycling rates, despite comprising approximately half of the urban population (see Section 2.2). Despite a growing body of research on gender differences in urban mobility [15, 16], the determinants of the so-called *gender cycling gap* remain insufficiently understood. Most of the existing research consists of stated preferences studies, relying on direct data collection, which are often limited in sample size and geographical scope (see Section 2.3). This study addresses this gap by leveraging large-scale data from the sport-tracking application Strava. By studying the behaviour emerging by aggregated online logs of activities by millions of cyclists, this research aims to complement, expand, and validate existing studies on the determinants of the gender cycling gap, thereby contributing to the design of more inclusive transportation systems. The chapter is based on the research article *Revealing the determinants of gender inequality in urban cycling with large-scale data*, published in EPJ Data Science [17].

In line with the goals of UN SDG 11.7, which recommends the provision of *universal access to safe, inclusive, and accessible green and public spaces, particularly for women, children, older persons, and persons with disabilities* [13], the second and third contributions focus on the development of tools to enhance green accessibility and exposure to nature in urban environments. Both contributions consist of a theoretical component and the development of a web-based interactive resource, and are discussed in more details in what follows.

The second contribution, presented in Chapter 3, directly engages with the ongoing debate on approaches and methods to measure green accessibility in urban environments. Recent research has concentrated on developing high-resolution spatial

indicators of green accessibility (see Section 3.3). Due to the established links between exposure to green spaces and public health, these indicators are typically employed for health impact assessments but more recently they have also been used to inform and optimise policy interventions aimed at enhancing green space accessibility. More broadly, this research effort is part of a larger shift in the urban planning paradigm towards data-driven policy design processes. In this context, developing spatially resolved indicators, supported by advancements in high-resolution geographical data and software, is seen as crucial for defining benchmarks to track progress towards specific targets and informing the design of new *structural* interventions, in line with the principle *what gets measured, gets done* [18]. Despite growing interest from both academic and policy communities, there is still no established standard for measuring green accessibility from a *structural perspective*, that is by evaluating the spatial relationship between the distribution of urban residents and the availability of urban green infrastructure. The classes of spatial indicators used, as well as the data sources employed, vary significantly based on the application (see again Section 3.3). Here, I contribute to this debate by proposing a computational approach that accommodates the measurement of several classes of spatial indicators of green accessibility within a unified framework. This framework is then utilised to examine the degree of interchangeability among these indicators and assess the impact of using one indicator over another within a policy design process. The proposed computational approach is made available to policymakers and the public through the development of a novel web-based interactive resource, currently deployed for over 1,000 cities worldwide. This tool illustrates the necessity of a multidimensional approach to green accessibility and includes functionalities designed to facilitate its adoption in policy design processes. The chapter is based on the research article *On the need for a multi-dimensional framework to measure accessibility to urban green*, published in NPJ Urban Sustainability [19]. The associated web-based interactive tool is available at ATGreen.

The third contribution, presented in Chapter 4, shifts the focus from the structural element of green accessibility to the development of tools that enhance the exposure of urban residents to nature by influencing their behaviour. To this end, I propose a novel routing engine (ATGreenGO) that recommends nature-enriching routes with minimal detours compared to the shortest path, thus compatible with the daily routines of potential users. The use of routing engines has increased dramatically in recent decades, driven by the widespread availability of internet access on mobile devices. While routing engines can offer a variety of functionalities (see Section 4.3), their primary purpose is to compute the optimal routes between two points. This computation can be customised based on several parameters, including the mode of transport (walking, cycling, driving, public transport, or multimodal options), arrival and departure times, and more. While the most common engine defines the optimality of the route based on the time required to complete the transit, there has been growing interest in designing algorithms that recommend urban routes based on alternative definitions of optimality. Recent examples include the perceived pleasantness of the route [20], its popularity [21], or the air quality [22]. The theoretical contributions of ATGreenGO are twofold: first, the design of a nature-enriching algorithm capable of characterising the street network of a city with features related to the presence of nearby natural elements; second, the definition of a routing optimisation problem, whose solution maximises green exposure while balancing the detour from the shortest route. This tool is integrated into an interactive application currently deployed for a specific set of cities, offering a practical tool for

planning nature-rich routes to users in these areas. The chapter is based on research conducted as part of the European Horizon 2020 project GoGreenRoutes (Grant ID: 869764). The associated web interface is available at ATGreenGO.

1.3 Research strategy and tools

The research strategy of this dissertation is grounded in a systematic approach aimed at investigating urban phenomena related to the concept of liveable and sustainable cities. This approach leverages large-scale datasets, software development, and quantitative analytical techniques to enhance our understanding of various aspects of urban dynamics, particularly those linked to the objectives of the SDG 11. Additionally, it introduces new tools to optimise urban planning processes. While each chapter constitutes an independent research study, collectively, the three contributions provide a cohesive quantitative examination of urban dynamics in line with SDG 11, advancing knowledge on how urban planning and policy can improve the sustainability and liveability of cities.

Systematic literature review For each contribution, the research activity commenced with a systematic literature review, which was essential for the identification, the evaluation, and the synthesising of relevant knowledge gaps. The initial phase of the literature review involved key-word searches in two major academic databases, Google Scholar and JSTOR, to gather a broad spectrum of relevant literature. This initial search was further expanded by exploring cross-references from the selected research articles to ensure a comprehensive coverage of the topic.

In Chapter 2, which examines the gender cycling gap, evidence was gathered not only from academic sources but also from reports by national and local authorities, technical reports on transportation survey data released by statistical agencies, and reports from other relevant stakeholders, such as private or public companies managing bike-sharing services. This multi-source approach provided a robust foundation for understanding the gender disparities in cycling adoption.

For Chapter 3, due to the policy relevance of the research topic on urban green space accessibility, the literature review was extended to include non-academic resources. These included policy briefs from UN and European Union (EU) agencies, as well as reports from national and local authorities, particularly from European countries. This broader scope ensured that the research was informed by both academic insights and practical policy considerations.

Quantitative analysis and software development The quantitative analyses presented in Chapters 2 through 4 employ a combination of advanced geospatial methods for data processing and advanced statistical and econometric techniques for data analysis.

All data processing and analysis were conducted using Python. The use of Python allowed for the integration of various libraries and tools, such as Pandas [23] for data manipulation, GeoPandas [24] for geospatial data processing, Osmium and OSMnx [25] for processing of georeferenced data from OpenStreetMaps (OSM) [3], and Statsmodels [26] and Scikit-Learn [27] for econometric and statistical analysis. Details on the exact data processing and analysis is provided in the Materials and Methods of each contribution. To ensure transparency and reproducibility, the

code developed for each contribution has been made publicly available in associated GitHub repositories. In Chapter 4, which introduces a novel routing engine designed to enhance exposure to nature, the system architecture is built using Java. The use of a Java-based architecture is common for routing engine due to its robustness and scalability.

Web interface development Two web interfaces were developed as part of this research, each designed to facilitate user interaction with the analytical tools and data. The front-end of these interfaces was developed using the ReactJS framework, a popular JavaScript library known for its efficiency and flexibility in building user interfaces. The back-end was managed using a PostGIS database, an extension of the PostgreSQL database system that supports geographic objects, enabling the efficient handling of spatial data.

1.4 Thesis outline

The dissertation is organised into five chapters. The first and last chapters serve as the Introduction and Conclusions, respectively, while Chapters 2 through 4 present the three analytical chapters detailing the research conducted during the doctoral programme.

The Introduction (this chapter) outlines the motivations and broader context of the research, along with the detailed aims and contributions of each study included in the dissertation. It also discusses the overall research strategy and tools and presents the structure of the dissertation.

The three analytical chapters follow a similar structure, beginning with an overview of the study in layman's terms, followed by the scope of the research, a discussion of the related literature, an overview of the main results, and a discussion of these findings. A detailed outline of each analytical chapter is presented below.

Chapter 2 focuses on the determinants of the so-called *gender cycling gap*. The chapter begins by outlining the scope of the research and discussing the existing evidence on the positive societal and individual impacts of cycling, as well as the presence of a persistent gender cycling gap. It continues with an overview of related work, specifically addressing the results and limitations of studies evaluating the determinants of gender differences in cycling behaviours, and highlighting the recent shift in research towards the use of automatically collected data for studies on active urban mobility. The chapter then presents the materials and methods of the study, along with a comprehensive overview of the characteristics of the data from the sport-tracking application Strava. The results of the study are presented in two separate sections, each addressing distinct research questions with data organised at different geographical levels. The chapter concludes with a discussion of the overall findings from these investigations.

Chapter 3 presents the research activity focused on *structural measures of green accessibility* in urban environments. The first section introduces the motivations behind the study by summarising previous theoretical and empirical contributions on the role of nature in improving the health outcomes of urban residents. The second section discusses related work, detailing in particular the existing approaches to the measurement of green accessibility and their adoption in the context of health impact assessments and urban planning. After presenting the materials and methods of the

study, the chapter provides a description of the proposed computational framework. The results of a statistical evaluation of the interchangeability of different indicators for policy design conducted using the framework are then provided, along with a description of the developed web interface. The chapter concludes with a discussion of the implications of the study and its limitations.

Chapter 4 introduces the *routing engine ATGreenGO*. The chapter begins with a discussion of the scope of the tool, as well as an overview of related research on the development of routing engines. A comprehensive description of the materials and methods of the study follows, including the approach for characterising the street networks of each city in terms of exposure to natural environments that can be experienced by walking on each street. The routing system is then introduced, and its performance is evaluated. The chapter concludes with a discussion of the characteristics of the suggested routes, the impact of the tool on residents of cities currently available in the system, and future research directions to test the system with experimental data on user preferences.

The concluding chapter (Chapter 5) summarises the main findings of the research conducted during the doctoral programme, their implications for theoretical analyses and policy design, and provides broader perspectives on the impact of these research activities. The limitations and the future outlook of research aimed at integrating large-scale data analytics in urban planning are the discussed. The dissertation concludes with some final thoughts on the importance of a multidisciplinary perspective to tackle urban challenges.

Chapter 2

Investigating the determinants of the gender cycling gap using large-scale automatically-collected data

2.1 Overview of the chapter

This chapter is based on the research article titled *Revealing the determinants of gender inequality in urban cycling with large-scale data*, published in EPJ Data Science by A. Battiston, L. Napoli, P. Bajardi, A. Panisson, A. Perotti, M. Szell, and R. Schifanella [17].

The study delves into gender disparities in urban cycling and explores how these inequalities relate to urban features, particularly focusing on safety-enhancing elements such as the presence of cycleways and low-speed-limit zones. Drawing on data from the sport-tracking application Strava, the analysis encompasses over 60 cities across four geographical regions: Italy, Benelux (Belgium, Luxembourg, and the Netherlands), the United Kingdom, and the United States. This approach, leveraging automatically collected data, represents a significant departure from previous studies that have typically relied on direct collection methods, like national transportation surveys or ad-hoc direct questionnaires. By utilising this novel data source, the study aims to complement, expand, and validate existing research, thus contributing to a deeper understanding of strategies for promoting non-motorised sustainable transportation, in line with the recommendations of the UN SDG 11.

The remainder of the chapter is structured into seven sections. Section 2.2 outlines the scope of the study, its motivations and its contributions to both academic research and policy-making. Section 2.3 reviews the related academic literature on both gender preferences in cycling and the use of large-scale automatically collected data for urban cycling research. Section 2.4 describes the materials and methods of the study, providing an overview of both the data sources employed for the analysis and the analytical and statistical techniques. Section 2.5 provide a thorough description of the data from the sport-tracking application Strava on cycling. The results of the study are then presented in Sections 2.6 and 2.7, each addressing one of the two primary research questions. Finally, Section 2.8 discusses the implications and limitations of the study.

2.2 Scope

Cycling offers a wide range of individual and societal benefits. On an individual level, it has been associated with improved physical and mental health outcomes, including enhanced cardio-respiratory fitness, reduced cardiovascular mortality risk, and decreased stress levels [28, 29, 30]. Furthermore, cycling serves as a cost-effective mobility option, contributing to the financial well-being of individuals while providing health benefits [31]. From a wider societal perspective, the promotion of cycling within integrated urban mobility systems is recommended as a solution to decoupling urban population growth from increased emissions levels and traffic congestion, bad air quality, and reduced road safety. Overall, increases in cycling uptake within an urban environment have been suggested to have a direct positive impact on 11 of the 17 UN SDGs [13, 32].

Despite the wide-ranging benefits associated with cycling, there is evidence of a large participatory gap between men and women. Cycling remains a male-dominated activity, particularly in cities where the overall level of cycling uptake is low. In major US cities such as New York, Boston, and Chicago, only one in four bicycle trips between 2014 and 2018 were made by women [33]. A similar disparity was observed in San Francisco, where only 29% of cyclists were women in 2018 [34]. Moreover, recent data from England reveals that not only do men take more bicycle trips per week, but they also cover longer distances compared to women [35].

However, some European countries, such as Denmark, Germany, and the Netherlands, where cycling uptake is in general higher, have managed to narrow this gender gap significantly, with women representing over 45% of all cyclists as early as 2005 [36]. This suggests that the gender gap in cycling is not intrinsic but rather arises from place- and culture-specific barriers. In this context, focusing only on improving mobility for the existing dominant (male) group, the cycling research and policy-making risks ignoring half of the population and hindering the development of sustainable mobility solutions for everyone [37, 38].

The analysis presented in this chapter adds to the debate on the gender gap in cycling by offering a novel large-scale investigation of its determinants. Unlike most previous analyses, which rely on small sample survey-based data collection (see Section 2.3), this study aligns with recent cycling research endeavours by using large-scale automatically collected data. Here, we used data from the sport-tracking application Strava to characterise the cycling behaviour of men and women. Despite the limitations discussed in section 2.8, with approximately 36 million users across 195 countries in 2018 [39], Strava provides an unparalleled source of information on cycling behaviour, both in terms of the number of cyclists involved and its geographical coverage.

For this study, we collected and analysed data from over 60 cities across the United States and Europe, organised into four geographical regions (Benelux, Italy, United Kingdom, and United States). The study seeks to address two research questions:

RQ1 What city-level characteristics are associated with a higher uptake of cycling by women?

RQ2 What is the association between the presence of dedicated cycling infrastructure and the volume of female cyclists on a street (relative to males)?

To answer the first question, we analyse how the gender gap in cycling varies across different cities in our dataset. We investigate the connection between this gender gap, which we measure at the city level, and various urban indicators. These indicators encompass different characteristics of cities, ranging from their physical features to road-safety indicators, such as the availability of cycleways and streets with low-speed limits. As these indicators were included to mirror the hypotheses from the psychological literature concerning the gender cycling gap, this analysis shed light on the extent to which macro-differences in actual cycling behaviour across cities align with micro-level findings on individual preferences from survey-based studies.

For the second question, we shift our focus from a macro comparison across cities to a micro-level analysis. We analyse the gender gap in cycling at the level of individual streets and investigate how specific street-level urban features influence this gap. Using logistic regression analysis, we examine specifically the role that the presence of dedicated cycling infrastructure and show how this type of infrastructure is crucial in supporting cycling uptake by women.

2.3 Related work

2.3.1 Research on the determinants of the gender cycling gap

Previous research on the determinants of the gender cycling gap has been mostly carried out within the context of cycling for commuting purposes and predominantly consists of stated preference studies, which are grounded in direct data collection. These studies have explored a range of determinants that can be broadly categorised into four key areas: built-environment and road safety determinants, physical, environmental, and morphological determinants, economic determinants, and wider cultural determinants.

Built-Environment and Road Safety Determinants Several stated preference studies have focused on the role of the built environment in shaping cycling behaviours, particularly among women. Key built-environment elements that have been evaluated include urban density and road network characteristics[1], the presence of cycling-supportive infrastructure (such as specialised traffic signals, secure bike parking, etc.), and safety-.enhancing, cycling-dedicated street-level infrastructure (hereafter referred to as cycling-dedicated infrastructure), such as protected or unprotected cycleways.

Features that reduce travel distances, such as high-density urban areas and well-integrated road networks, have been found to promote cycling participation across the general population [40, 41, 42]. This include women, despite this latter group typically displaying different travel patterns and cycling preferences compared to men [43]. In terms of cycling-supportive infrastructure, a positive association between the presence of secure bike parking, specialised traffic signals for cyclists, and public bicycle-sharing systems and women's cycling uptake has been identified through both stated and revealed preference approaches [44, 45].

Despite the recognised importance of the aforementioned classes of built-environment elements, much of the stated preference academic literature has concentrated on the

[1]Depending on interpretation, these elements may also be classified as *Physical, environmental, and morphological determinants*.

role of safety-enhancing cycling-dedicated infrastructure, such as dedicated cycleways. This focus stems from the traditional attribution of the gender cycling gap to psychological and risk perception factors, leading women to have a stronger preference for physically separated cycling infrastructure due to heightened perceived risks when cycling in mixed traffic. Several stated preference studies have confirmed this hypothesis, finding that dedicated cycling infrastructure positively influences women's cycling engagement [46, 47, 48]. The presence of such infrastructure not only increases overall cycling rates but also significantly impacts women's cycling choices. For instance, evidence from the United States indicates that women are more likely than men to choose longer routes if they are perceived to be safer [49]. Similarly, in Australia, women show a preference for safer off-road paths [50], with comparable findings observed in China [44]. However, evidence from a few studies investigating revealed preferences, which observe actual behaviour rather than stated intentions, is less consistent, with some showing weaker or mixed associations between dedicated cycling infrastructure and cycling uptake among women [51, 52].

Physical, Environmental, and Morphological Determinants Beyond road safety and built-environment factors, other physical, environmental, and morphological determinants can significantly influence the suitability of a cycling environment and have thus been investigated as potential contributors to the gender cycling gap. Among the morphological characteristics of the urban landscape, street inclination has been identified as a substantial determinant of this gap [53, 54]. A study on cycling adoption for commuting in San Francisco found that women are more likely to avoid steep slopes due to concerns related to physical exertion and safety [53]. Steep inclines not only require greater physical effort but also increase the perceived risk of accidents, particularly in adverse weather conditions.

Women are also less inclined to cycle in extreme weather conditions, such as high temperatures and adverse weather, and are generally less likely to cycle in environments considered unsuitable—defined by a combination of these factors—especially in contexts where cycling culture is limited [55, 54]. These physical, environmental, and morphological barriers further reduce the opportunities for women to cycle under optimal conditions, thereby contributing to lower cycling participation among women.

Economic Determinants Economic factors, though less extensively studied in the academic literature, may also play a role in the gender cycling gap. While cycling is often promoted as a cost-effective mode of transport, the initial costs associated with purchasing a bicycle, maintenance, and acquiring appropriate gear can be prohibitive, particularly for women in low-income settings. UK data from the 2019 analysed by the British charity Sustrans showed that 19% of individuals from low-income households found buying a cycle unaffordable and reported it as a barrier to cycling[2]. Moreover, economic constraints might intersect with other barriers, such as safety and cultural factors. For instance, women who cannot afford a high-quality bicycle or safety gear may feel less confident cycling in traffic, further deterring them from choosing this mode of transport.

[2]Resource available here.

Wider Cultural Determinants Finally, academic research has also focused on the role that wider cultural factors play in the gender cycling gap. In regions where cycling is deeply ingrained in the local travel culture, women are more likely to participate in cycling. This is particularly evident in areas with high overall cycling levels, where the number of female cyclists and the distances cycled by women are notably greater [56, 57]. The cultural acceptance of cycling as a legitimate mode of transport influences women's comfort and confidence in choosing cycling, thereby reducing the gender gap. Conversely, in regions where cycling is less common, women report additional barriers to cycling, including discomfort with long distances, adverse weather conditions, and lower confidence in their cycling abilities [58, 59].

2.3.2 Research on cycling behaviour using automatically collected data

The advent of new GPS-enabled technologies, such as sport-tracking applications and bike-sharing platforms, has significantly enhanced the availability and quality of automatically collected data on cycling behaviour. This increase in data availability has paved the way for a new line of research that uses automatically collected data to explore not only cycling-related questions but also broader patterns and preferences in so-called human mobility research [60] and in the smart and digital twins cities research [61]. These data have greatly facilitated the adoption of revealed preference approaches in urban mobility research, where users' preferences are inferred from observed behaviours at a fine-grained geographical level, rather than being directly stated. Additionally, they expand our ability to monitor urban mobility patterns and improve both individual and collective planning processes, through the development of routing engines or data-driven policy design.

In cycling-related studies, data from bike-sharing services have been utilised for various purposes, including monitoring service demand and analysing adoption and preferences across different demographic groups [62, 63]. For instance, in Oslo, Norway, gender-specific analyses of bike-sharing data have revealed a gap in the spatial distribution of facilities, which disadvantages women in accessing these services [64].

Automatically collected data on cycling behaviour are also available through partnerships with sport-tracking applications[3]. A commonly used source for such studies is Strava Metro, which aggregates logs from the sport-tracking application Strava and is accessible through partnerships between city councils and Strava itself[4]. Recently, these data have been employed to study the exposure of cyclists to air pollution in Glasgow [65] and to map cycling patterns in Johannesburg [66]. GPS-based data have also been utilised to investigate route choices among different demographic groups in cities such as San Francisco and Atlanta [47, 53].

These emerging research avenues highlight the significant potential of large-scale data to offer detailed insights into the spatial and temporal aspects of cycling, which are vital for effective urban planning and infrastructure development. However, while these extensive datasets provide valuable information, they predominantly reflect the behaviour of current cyclists, potentially overlooking the differences between cyclists and non-cyclists. If such differences are substantial, the conclusions

[3] Although the use of these tracking systems has increased for commuting purposes, the data are generally more suitable for studying recreational cycling, especially when the trip purpose cannot be inferred from the aggregated data.

[4] Strava Metro data differ from those used in this study, as they are aggregated from individual users' routes rather than being based on the statistics of Strava segments (see Section 2.5).

drawn from these data may lack broader applicability. Additionally, concerns have been raised about the representativeness of these datasets and the specific conditions required for their use in research contexts [67].

Therefore, revealed preferences studies based on large-scale data should be viewed as complementary to stated preference investigations, with the limitations of both approaches carefully considered within the context of the research question at hand. In this spirit, this study presents the first large-scale investigation into the determinants of the gender cycling gap, aiming to support, validate, and expand upon existing research based on stated preferences evidence.

2.4 Materials and Methods

This section outlines the materials and methods used in the study, with separate descriptions provided for the two components (RQ1 and RQ2) for clarity. Detailed information on the Strava data, including its description, processing, and validation, is excluded here and is comprehensively addressed in the following section of this chapter.

RQ1: definitions, data sources and processing and statistical analysis

Definition of city To address RQ1, we used data aggregated at the level of city[5], following the definition provided in the Urban Centre Database of the Global Human Settlement (GHS-UCDB), version R2019A [2]. The GHS-UCDB is a comprehensive collection of data on urban areas worldwide developed by the Joint Research Centre (JRC) of the European Commission. It provides detailed information on the location and characteristics of urban centres, including population size, density, infrastructure, and socioeconomic factors. The definition of an urban centre does not follow an administrative definition. Instead, these are defined by specific cut-off values on the resident population and built-up surface share in a 1x1 km uniform global grid. The final sample of cites for this strand of the study consisted of 61 cities. The city of New York was excluded due to the large discrepancy between the administrative area of this city and the bounding box of the GHS-UCDB.

Data sources To address RQ1, we compiled a dataset of city-level indicators covering four domains: Environment, Built-Environment and Industrialization, Socio-Economics and Demographics, Street-Network Morphology and Road Safety. A comprehensive list of the indicators included in the study, their definition and their sources is provided in Table 2.1. While some of the indicators were already available at the required geographical level, others were constructed from street-level metrics and then aggregated to define city-wide indicators. These latter cases are marked with a "*" next to the source in the table.

Construction of urban road safety indicators City-level indicators on urban road safety were constructed aggregating data collected at the level of the street network into city-wide metrics. For both indicators, street-network information was obtained from OSM [3] accessed via the Python library OSMnx [25]. The indicator on the proportion of streets with a max-speed limit of 20mph or 30km/h or below (referred to as *speed limit* throughout the manuscript) was constructed following this pipeline:

[5]Throughout the dissertation, the terms city, urban area, and urban centre are used interchangeably.

TABLE 2.1: City-level indicators: description and source.

Category	Variable name	Description	Data source
	σ_c	Proportion of kilometers rode by female cyclists to the overall kilometers rode by any cyclist within the urban area	Strava*
E	share green	Share of population living in the high green area in 2015 in the Urban Centre of 2015. Ranging between 0-1	[2]
	open space	Percentage of open-spaces within the spatial domain of the Urban Centre. Ranging between 0-100	[2]
	Global Gender Gap index	Global Gender Gap index (measured at country level), from the World Economic Forum. Used for the sensitivity analysis only.	[68]
BEI	built area	Amount of the built-up area per person in 2015 calculated within the spatial domain of the Urban Centre. Expressed in square meters per person	[2]
	light emissions	Average night time night-light emission calculated within the Urban Centre spatial domain. Expressed in nano-watt per steradian per square centimetre	[2]
	pm2.5	Total concentration of PM2.5 for reference epoch 2014, calculated over the Urban Centre. Expressed in $\mu g/m3$	[2]
SED	area	Area of the spatial domain of the Urban Centre. Expressed in square meters	[2]
	population	Population density within the spatial domain of the Urban Centre	[2] *
	GDP	GDP per capita for year 2015 within the Urban Centre. Expressed in US dollars	[2] *
M	degree	Average node degree of street network within the spatial domain of the Urban Centre	[69]
	grade	Average absolute inclination of streets within the spatial domain of the Urban Centre. Expressed in percentage	[69]
	orientation	Orientation order of street network bearings within the spatial domain of the Urban Centre.	[69]
	3-way crosses	Proportion of nodes that represent a 3-ways street intersection in the street network within the spatial domain of the Urban Area. Ranging between 0-1	[69]
	straightness	Ratio of straightline distances to street lengths for streets in the street network within the spatial domain of the Urban Area	[69]
RS	bike lanes	Proportion of streets with cycleways (either protected or unprotected) computed on streets within the spatial domain of Urban Centre	OSM*
	speed limit	Proportion of streets with a speed-limit equal or lower than 20 mi/h or $30km/h$ computed on streets within the spatial domain of Urban Centre	OSM*

Categories: E: Environment, BEI: Built-Environment and Industrialization, SED: Socio-Economics and Demographics, M: Street Network Morphology, RS: Road Safety.
*Indicates that the data from the original data sources required specific preprocessing described in the Methods.

TABLE 2.2: Definition of bike lanes for construction of city-level indicators of urban road safety. The table provides the OpenStreetMap $(key, value)$ pairs used for the identification of streets with some form of cycling dedicated infrastructure, simply indicated as *bike lane* in the main text. This information was then used to measure the size of the cycling-dedicated-infrastructure in the city and construct the corresponding city-level indicator.

OSM key	**OSM value**
highway	[*cycleway*]
cycleway	[*track*, $opposite_t rack$, *lane*, $opposite_l ane$, *opposite*, $share_b usway$, $shared_l ane$, *designated*, *yes*]
cycleway : $left$	[*track*, $opposite_t rack$, *lane*, $opposite_l ane$, *opposite*, $share_b usway$, $shared_l ane$, *designated*, *yes*]
cycleway : *right*	[*track*, $opposite_t rack$, *lane*, $opposite_l ane$, *opposite*, $share_b usway$, $shared_l ane$, *designated*, *yes*]
cycleway : *both*	[*track*, $opposite_t rack$, *lane*, $opposite_l ane$, *opposite*, $share_b usway$, $shared_l ane$, *designated*, *yes*]

1. For each city c, extract the bounding box of city c from the GHS-UCDB [2].
2. Extract the street network from the polygon defined in the bounding box via the OSMnx library [25]. Set: $network\_type = 'drive', retain\_all = True$.
3. Compute the proportion of streets satisfying the condition on the speed limit. Weight each street with its length.

The indicator on the proportion of streets with cycling-dedicating infrastructure (referred to as *bike lanes* throughout the manuscript) was constructed following this pipeline:

1. For each city c, extract the bounding box of city c from the GHS-UCDB [2].
2. Extract the street network from the polygon defined in the bounding box using the OSMnx library [25]. Set: $network\_type = 'bike', retain\_all = True$. Call this graph $G0$.
3. From OSM [3], extract the street network from the polygon defined in the bounding box using the OSMnx library. Set: $network\_type = 'drive', retain\_all = True$. Call this graph $G1$.
4. Define as cycleways all streets in $G0$ with the pairs of OSM attribute described in Table 2.2.
5. Sum over the length of all cycleways in $G0$.
6. Sum over the length of all streets in $G1$.
7. Define the index as the ratio between the metric computed at point 5 and the metric computed at point 6.

Regression analysis We estimated a linear regression model of the form:

$$\sigma_c = \sum_{j=1}^{N} \beta_j z_{j,c} + \epsilon_c \quad c = 1, .., 61 \tag{2.1}$$

via Ordinary Least Squares (OLS), where the list of regressors z_j in the preferred model includes: *speed limit*, *orientation*, *GDP*, *3-way crosses*, *night-light emissions*, *grade*, *pm2.5* plus three dummy variables for the macro area to which the city belong (US, UK, Benelux, *baseline*: Italy). All continuous regressors were normalised using a z-score transformation. Out of the initial 15 city-level indicators collected (provided in Table 2.1), the final subset of seven indicators (plus three country-level dummies) included in the regression was selected via an exhaustive search to minimise the Akaike Information Criterion (AIC) of the model. The model was estimated using the OLS function of the Python library *statsmodel* [26].

RQ2: data sources and statistical analysis

Data sources To address RQ2, a dataset with street-level data was compiled for each city in our sample. Street-level information consisted of data on the gender ratio of cyclists on each street and on the presence of protected or unprotected cycleways. In detail:

- the street-level gender ratio of cyclists was defined following the approach described in Section 2.5.
- the level of protection of streets was extracted using the mapping between OSM tags and protection-level outlined in Table 2.5.

For the city of New York only, for which additional street-level information can be obtained from administrative data sources, the dataset was expanded to include information on (Table 2.3):

- the location of all (any-vehicle) accidents and bike accidents only from the OpenData Portal of the city of New York [70]. These data were processed to compute the number of accidents per 10 meters for each street.
- the presence of public lighting, proximity to a park or the coastline, and whether the surface is paved.
- for streets in the largest component of the street network, we computed the edge-betweenness [71] via the Python library graph-tool [72]. Streets outside the largest component of the network (i.e., streets in the borough of Staten Island) were excluded from the sample.
- the borough in which the street is located. Shapefiles were downloaded from the OpenData Portal of the city of New York [70].

Multivariate logistic regression To assess the degree of association between σ_s and the presence of cycling-dedicated infrastructure, we estimated a multivariate logistic regression model for streets in the city of New York. We restricted the sample to streets belonging to the bottom and top 33% of the distribution of σ_s and classified streets in *Low* and *High* σ_s respectively. As a robustness check, the analysis was repeated for alternative values of this threshold (0.25 and 0.40, instead of 0.33). We used features described in Table 2.3 as predictors and the binarized σ_s as the target variable. Moreover, continuous predictors (*Any-vehicle crashes*, *Bike crashes* and *Edge-betweenness*) were scaled using a z-score-transformation to normalize the magnitude of the estimated coefficients. The model was estimated using the Logit function of the Python library *statsmodel* [26].

TABLE 2.3: List of street-level urban features included in the case study for the city of New York.

Variable name	Description	Data source	OSM key-values pairs
Unprotected cycleway	Dummy for presence of a shared or unprotected bike-lane	[3]	{*cycleway* : *right*: [*lane*, *opposite_lane*, *share_busway*, *shared_lane*, *designated*, *yes*], *cycleway* : *left*: [*lane*, *opposite_lane*, *share_busway*, *shared_lane*, *designated*, *yes*], *cycleway* : *both*: [*lane*, *opposite_lane*, *share_busway*, *shared_lane*, *designated*, *yes*], *cycleway*: [*lane*, *opposite_lane*, *opposite*, *share_busway*, *shared_lane*, *designated*, *yes*], *bicycle*: [*yes*, *permissive*, *destination*, *private*]}
Protected cycleway	Dummy for either the presence of a protected bike-lane or streets with no vehicles	[3]	{*highway*: [*cycleway*, *path*, *footway*, *bridleway*, *track*], *cycleway*: [*track*, *opposite_track*], *cycleway* : *left*: [*track*, *opposite_track*], *cycleway* : *right*: [*track*, *opposite_track*], *cycleway* : *both*: [*track*, *opposite_track*], *bicycle*: [*designated*]}
Public lighting	Dummy for the presence of public lighting	[3]	{*lit*:[*yes*, 24/7]} or missing tag
Unpaved surface	Dummy for unpaved surface	[3]	{*surface*:[*unpaved*, *compacted*, *fine_gravel*, *gravel*, *pebblestone*, *ground*, *earth*, *dirt*,*grass*, *grass_paver*, *sand*, *mud*]}
Park proximity	Dummy for streets next to a park (within 15 meters)	[3]	{*leisure*:*park*:} + 15m proximity from the geometry
Any-vehicle crashes	Number of crashes involving any type of vehicles per 10m of street length	[70]	Not applicable
Bike crashes	Number of bike crashes per 10m of street length	[70]	Not applicable
{borough_name}	Dummy for boroughs (baseline: Manhattan)	[70]	Not applicable
Coast proximity	Dummy for street next to the river coast	[3]	*natural*:*coastline* + 150m proximity from the geometry
Edge-betweenness	Edge betweenness of the streets computed for streets within the largest component of the street network	[3]	OSM street-network used to compute edge-betweenness on the largest component of the network. Network extracted using OSMNX, network_type=*bike*.

2.5 Strava data on recreational cycling

The sport-tracking application Strava Strava is a popular sport-tracking application and social network for athletes or individuals practicing outdoor activities. It allows users to track activities such as running, cycling, and swimming, using GPS-enabled devices, like smartphones, smartwatches or dedicated GPS devices. Users can record their routes, distance, speed, and other metrics during their workouts. The service supports up to 33 different activities, but it is mostly used for cycling and running. At the time of the data collection in 2018, Strava counted around 36 million users worldwide, corresponding to 0.6 billion recorded activities [39]. Of these, 284 million were cycling-activities (47%), and approximately one in five cycling-uploads were by women (50 million). Tracking of commuting is growing in popularity on Strava, however, the majority of uploads by 2018 referred to recreational and athletic cycling. One of the key features of Strava is the aggregation of data at the level of segments, i.e. specific sections of roads or trails where users can compare their performance with others. Users can compete for the fastest time in these segments, earning achievements and recognition for their efforts. Each user training on a segment automatically enters the associated *leaderboard*.

Data collection For this study, information on cycling behaviour was obtained from the segments of the sport-tracking application Strava. We collected Strava segments from cycling activities recorded in 62 cities across four geographical regions: the United States, the United Kingdom, Benelux (Belgium, Netherlands, and Luxembourg), and Italy. For sensitivity analysis, the dataset was expanded to include eight additional cities across other European countries. A Strava segment is a specific portion of a road or trail where Strava users compete by recording their times. The performance of each user training on a segment is automatically recorded in its leaderboard, which provides insights into the characteristics of users cycling on that particular trail.

The data collection process involved two phases, both conducted in November 2018. In the first phase, we obtained the entire set of segments (approximately 16.4 million) available at the time through the Strava API. Each data point consisted of the geographical information related to each segment (a line string) and the ID of the segment. In the second phase, we collected summary statistics from the female and male leaderboards (using separate queries for each) associated with each segment ID. As segments need not to be within the boundary of one city only, for each city in our sample, we made queries for the leaderboards of all segments whose geometry was contained for at least 75% of its length within the city boundary. The outcome of this data collection phase is a set of datasets (one for each city) where each data record corresponds to a Strava segment and includes: 1. geographic information about the segment in the form of a line string of latitude-longitude coordinates; 2. statistics extracted from the associated leaderboards on the total number of unique male and female cyclists training on the segment, calculated as the sum of the lengths of the female and male leaderboards. It should be noted that each cyclist is included only once in the corresponding leaderboard, based on their best performance on the segment. This characteristic is important, as it could introduce a measurement error in metric adopted to measure the gender cycling gap in the study if the tendency to train on "previously visited" segments differs between men and women differs across cities or within different areas of the same city. However,

FIGURE 2.1: **Visualisation of the raw Strava segments of nine cities across the four geographical areas** Graphical visualisation of the raw Strava segments of nine cities for the four main geographical areas covered by the study (United Kingdom: Nottingham, Bristol and Manchester; Benelux: Amsterdam and Liege, Italy: Rome and Turin, United States of America: Boston and Memphis). All segments are plotted with the same colour intensity.

since data on the exact number of activities by gender was not directly available, it was not possible to evaluate the impact of this limitation.

Characteristics of raw Strava segments Strava segments are not predefined by the app developers. Instead, they are directly created by Strava users, reflecting their training preferences and needs. The result of this user-driven generating process is a set of segments highly heterogeneous in length, both across cities and within the same city. Indeed, some segments may correspond to portions of a street, while others define long trails spanning multiple streets. Furthermore, segments can partially or completely overlap each other. A graphical illustration of the raw Strava segments for nine cities is provided in Figure 2.1. For this visualisation, all segments were plotted with the same colour intensity and darker areas on the map indicate a series of overlapping segments. For each city, Table 2.4 reports the total number of Strava segments, the length of the shortest and longest segment in the data collection

in kilometres (Min (km), Max (km)), the mean and standard deviation of the distribution of the lengths of segments in kilometres (Mean (km), Std (km)). In terms of length of segments, the information presented in Table 2.4 well depicts the discussed heterogeneity both across cities and within the same city. As an illustration, let's consider the city of London, for which the length of segments spanned from 0.00 km to 104.97 km, with a mean of 2.61 km. A smaller range of less than 7 km is observed for the cities of Apeldoorn and Luxembourg, both located in the macro area of Benelux. Raw information on the two cities of Rotterdam and The Hague is here presented separately, however, in the remainder of the study, these urban centres are analysed together[6].

Remapping of data from Strava segments to the street network In the first step of the data processing, the Strava segments were projected to the street network of the corresponding city. This step was designed to enrich each edge of the street network with data on the gender split of cyclists training on it. Hereafter, we use the term *street* to refer to each edge of the street network[7], rather than the toponymic definition of the street.

Street-network information of each city was extracted from OpenStreetMap (OSM) [3] via the open-source software OSMnx [25]. The data processing followed a 6-step pipeline, performed on each city in our sample separately:

1. Load the full set of Strava segments for the city.
2. Extract the bounding box of the city from the Global Human Settlement - Urban Centre Database 2015, version 2019A (GHS-UCDB) [2].
3. From OSM, extract the street network within the polygon defined in the bounding box using the OSMnx library [25]. Set: $network\_type = 'bike'$, $retain\_all = True$, $simplify = True$.
4. Classify streets in the street network based on OSM attributes in: street with `protected cycleway`, `unprotected cycleway` and `no cycleway`. The $(key, value)$ pairs for the classification are provided in the Table 2.4. All other bikeable streets are classified as `no cycleway`.
5. Proceed with the *preferential assignment* of Strava segments as follows. Buffer with a 10-meter radius the geometries of the street network. Select all streets categorised as `protected cycleway` and intersect each Strava segment with the network. Re-project each segment (or portion(s) of a segment) on all streets with an intersection of at least 30 meters. Finally, compute the geometries of Strava segments left unassigned - that could be either a full segment or portion(s) of a segment- and repeat the procedure selecting `unprotected cycleway` first and finally streets with `no cycleway`.

[6] This is done to match the definition of urban centre adopted in the Global Human Settlement - Urban Centre Database [2].

[7] This is a topological representation of a street, which does not necessarily adhere to the toponymic definition. Multiple edges typically fall under the same street name. The use of the term street to refer to the edge of the stret network only applies to this Chapter of the dissertation.

TABLE 2.4: Geographic characteristics of raw Strava segments by city

	City name	Area	N seg.	Min (km)	Max (km)	Mean (km)	Std (km)
0	Albuquerque	USA	1997	0.25	126.51	6.88	8.97
1	Almere	Benelux	818	0.08	24.03	2.31	3.13
2	Amsterdam	Benelux	2091	0.00	23.39	1.64	2.13
3	Antwerp	Benelux	1014	0.04	23.96	1.67	2.68
4	Apeldoorn	Benelux	357	0.03	7.12	1.14	1.04
5	Arnhem	Benelux	1608	0.00	19.75	1.24	1.79
6	Austin	USA	5310	0.07	143.23	7.16	12.00
7	Bari	Italy	129	0.03	14.50	2.25	2.63
8	Bologna	Italy	653	0.10	44.52	2.85	4.68
9	Boston	USA	1400	0.19	53.87	4.45	5.96
10	Breda	Benelux	457	0.07	18.42	1.64	2.24
11	Brescia	Italy	665	0.03	21.19	2.20	2.76
12	Bristol	Uk	3221	0.00	40.34	1.61	2.87
13	Catania	Italy	271	0.14	49.78	3.95	6.40
14	Charleroi	Benelux	355	0.07	13.51	1.12	1.53
15	Charlotte	USA	1664	0.05	136.31	7.50	13.18
16	Derby	UK	1381	0.04	18.27	1.34	2.04
17	Eindhoven	Benelux	779	0.03	18.11	1.27	1.57
18	Exeter	UK	1674	0.00	24.83	1.70	2.94
19	Florence	Italy	689	0.00	24.50	1.53	1.82
20	Genoa	Italy	2107	0.00	35.92	2.35	2.88
21	Ghent	Benelux	1170	0.04	20.71	1.26	1.97
22	Groningen	Benelux	1279	0.00	11.34	1.45	1.51
23	Haarlem	Benelux	220	0.01	10.78	1.11	1.29
24	Jacksonville	USA	1032	0.36	160.42	10.54	16.88
25	Las Vegas	USA	1948	0.01	102.45	7.63	9.40
26	Leeds	UK	9991	0.00	82.89	2.90	5.87
27	Liege	Benelux	770	0.08	14.62	1.19	1.20
28	London	UK	18232	0.00	104.97	2.61	4.69
29	Louisville	USA	1833	0.19	208.30	9.15	14.50
30	Luxembourg	Benelux	606	0.06	7.36	1.14	0.93
31	Manchester	UK	2812	0.02	23.17	1.65	2.47
32	Memphis	USA	709	0.06	102.86	9.68	14.41
33	Modena	Italy	191	0.13	17.24	3.27	3.52
34	Nashville	USA	2186	0.13	168.36	7.78	12.32
35	Newcastle upon Tyne	UK	2085	0.03	17.23	1.51	2.21
36	New York	USA	7122	0.00	256.20	11.13	19.44
37	Nijmegen	Benelux	551	0.10	5.75	0.93	0.86
38	Norwich	UK	1065	0.00	20.73	1.14	1.82
39	Nottingham	UK	2168	0.02	22.73	1.43	2.30
40	Oklahomacity	USA	1502	0.23	277.57	16.75	28.65
41	Padua	Italy	130	0.20	10.47	1.58	1.40
42	Palermo	Italy	929	0.10	57.93	3.15	4.23
43	Parma	Italy	155	0.18	17.93	3.77	3.23
44	Phoenix	USA	6564	0.05	219.64	11.22	17.39
45	Plymouth	UK	3503	0.03	47.29	1.63	2.70
46	Prato	Italy	288	0.17	15.19	1.94	2.02
47	Reading	UK	830	0.08	16.75	1.09	1.51
48	Reggio Emilia	Italy	178	0.06	28.69	3.55	3.91
49	Rome	Italy	3257	0.00	52.60	3.00	5.11
50	Rotterdam	Benelux	1871	0.00	20.35	1.76	2.42
51	San Antonio	USA	3277	0.11	167.88	8.68	14.84
52	Sheffield	UK	7733	0.00	88.06	2.50	5.11
53	Southampton	UK	1495	0.07	17.03	1.34	1.68
54	Taranto	Italy	152	0.07	11.87	1.95	2.25
55	The Hague	Benelux	1018	0.02	23.95	1.27	1.90
56	Tilburg	Benelux	525	0.05	24.37	1.62	2.41
57	Trieste	Italy	1092	0.00	19.19	1.92	2.22
58	Turin	Italy	1418	0.00	19.27	2.06	2.10
59	Utrecht	Benelux	1221	0.06	50.08	2.05	4.15
60	Venice	Italy	341	0.15	75.69	3.31	7.39
61	Verona	Italy	1154	0.07	29.3	2.19	2.43
62	York	UK	1711	0.03	52.88	2.12	3.45

TABLE 2.5: OpenStreetMap key-value pairs for the classification of cycleways in protected and unprotected.

Type	OSM key	OSM value
Protected cycleway	*highway*	[*cycleway*, *path*, *footway*, *bridleway*, *track*]
	cycleway	[*track*, $opposite_track$]
	cycleway : *left*	[*track*, $opposite_track$]
	cycleway : *right*	[*track*, $opposite_track$]
	cycleway : *both*	[*track*, $opposite_track$]
	bicycle	[*designated*]
Unprotected cycleway	*cycleway*	[*lane*, $opposite_lane$, $share_busway$, $shared_lane$, *designated*, *yes*]
	cycleway : *left*	[*lane*, $opposite_lane$, $share_busway$, $shared_lane$, *designated*, *yes*]
	cycleway : *right*	[*lane*, $opposite_lane$, $share_busway$, $shared_lane$, *designated*, *yes*]
	cycleway : *both*	[*lane*, $opposite_lane$, $share_busway$, $shared_lane$, *designated*, *yes*]
	bicycle	[*yes*, *permissive*, *destination*, *private*]

6. Compute the gender ratio of each street in the street network using statistics from the re-projected Strava segments. In particular, letting I be the set of segments re-projected to street s, F_i (M_i) the number of unique female (male) cyclists on segment i, the total number of female cyclists on streets s (and correspondingly for male cyclists) is defined as:

$$F_s = \sum_{i \in I} F_i \tag{2.2}$$

The gender ratio (σ_s) of cyclists on street s is then computed as:

$$\sigma_s = \frac{F_s}{M_s + F_s} = \frac{\sum_{i \in I} F_i}{\sum_{i \in I}(F_i + M_i)} \tag{2.3}$$

The rationale for the *preferential assignment* is that if a cycleway runs parallel to a street with no cycleway and the line string geometry for the Strava segment is compatible with both streets (i.e. it falls within the buffered geometry of both streets), we assume that the cyclists rode on the cycleway rather than on the street with no cycling-dedicated infrastructure. This approach prevents us from remapping the same portion of a Strava segment to multiple parallel streets with different characteristics.

Data filtering on street with a low number of cyclists The vast majority of Strava users are men. Because of this, the probability of observing $\sigma_s = 0$ is a decreasing function of the number of cyclists on the street. This means that observing no women on a street may be due to two factors. First, the street might not be sufficiently popular (among both genders): being women underrepresented among Strava users, the likelihood of observing a female cyclist is low. On the other hand, the segment might be overall sufficiently popular, but its attractiveness being low among women compared to men. These effects are hard to disentangle in the aggregated data, limiting our ability to interpret an extreme values of σ_s, on segments with low overall popularity. To mitigate this issue, we filter the data for the analysis of RQ2 to exclude those segments with a small number of cyclists. In particular, for a city c, the probability of observing $\sigma_s = 0$ on a segment s conditional to observing N_s cyclists

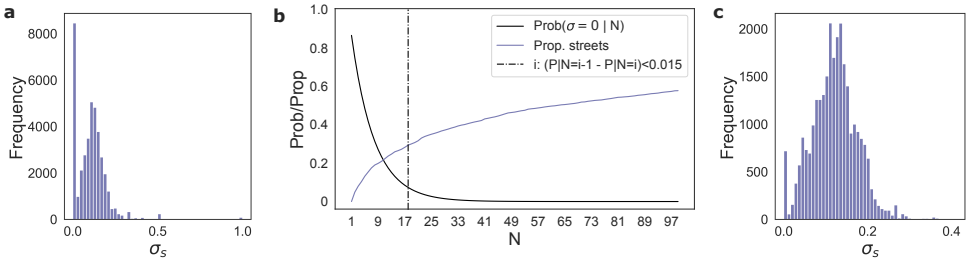

FIGURE 2.2: **Impact of streets filtering for the city of New York**. a) The distribution of σ_s for streets in the City of New York, before filtering. b) The black line depicts the probability of observing $\sigma_s = 0$ conditional on the number of cyclists. The light blue line depicts the proportion of streets in the street network as a function of the number of cyclists. The vertical dashed line depicts the selected threshold for the data filtering. c) The distribution of σ_s for streets in the City of New York, after filtering of segments with a number of cyclists below the selected threshold.

on s is given by (assuming replacement):

$$P_N = P(\sigma_s = 0|N_s) = \left(\frac{M_c}{M_c + F_c}\right)^{N_s} \tag{2.4}$$

where F_c (M_c) is the total number of female (male) cyclists in the city. P_N is a decreasing function of N_s. To limit the dependence of the observed gender ratio from the overall level of popularity of the segment, streets with an overall number of cyclists below a critical threshold were excluded from the sample. For each city, the filtering threshold N_s^* was set as:

$$N_s^* = min\{N_s|P_{N_s} - P_{N_{s-1}} < 0.015\} \tag{2.5}$$

The impact of the filtering on the distribution of gender ratio for the city of New York is depicted in Figure 2.2, where the filtering leads to the removal of streets with less than 17 cyclists. As expected, the observed zero inflation of the distribution appears to be less marked after the filtering, confirming the effectiveness of the adopted mitigation strategy. In addition, the new distribution is significantly less right-skewed.

Construction of the city-level index of the gender-cycling-gap For RQ1, we measure the gender cycling gap of city c through σ_c, defined as the ratio between the total kilometers travelled by female cyclists and the overall kilometers travelled by cyclists of both gender within the urban area. The rationale for the use of this metric is its ability to capture two forms of gender gaps described in the literature on cycling and gender: the propensity of women to make fewer trips than men and the propensity to cycle shorter distances. This measure is equivalent to the weighted sum of the gender ratio on streets (σ_s in the previous paragraph) within the urban area, with weights equal to the product of the length and the total popularity (total number of cyclists) of the street. I.e., letting S be the set of streets in the street network of the city c, F_s (M_s) the number of female (male) cyclists on s and l_s the length of street s expressed in kilometres, σ_c is defined as:

$$\sigma_c = \frac{\sum_{s \in S} F_s \cdot l_s}{\sum_{s \in S}(F_s + M_s) \cdot l_s} = \frac{\sum_{s \in S} \sigma_s \cdot l_s \cdot (F_s + M_s)}{\sum_{s \in S}(F_s + M_s) \cdot l_s} \tag{2.6}$$

TABLE 2.6: Ranking of cities by σ_c, by geographical area.

Ranking	City name	Geographical area	Country	σ_c
1	Groningen	Benelux	Netherlands	0.21
2	Utrecht	Benelux	Netherlands	0.17
3	Amsterdam	Benelux	Netherlands	0.16
4	Apeldoorn	Benelux	Netherlands	0.15
5	Nijmegen	Benelux	Netherlands	0.15
6	Haarlem	Benelux	Netherlands	0.14
7	Almere	Benelux	Netherlands	0.13
8	Eindhoven	Benelux	Netherlands	0.13
9	Arnhem	Benelux	Netherlands	0.13
10	Breda	Benelux	Netherlands	0.13
11	Tilburg	Benelux	Netherlands	0.12
12	Rotterdam/The Hague	Benelux	Netherlands	0.12
13	Ghent	Benelux	Belgium	0.10
14	Luxembourg city	Benelux	Luxembourg	0.09
15	Antwerp	Benelux	Belgium	0.09
16	Liege	Benelux	Belgium	0.06
17	Charleroi	Benelux	Belgium	0.06
1	Venice	Italy	Italy	0.09
2	Padua	Italy	Italy	0.07
3	Verona	Italy	Italy	0.07
4	Trieste	Italy	Italy	0.06
5	Florence	Italy	Italy	0.06
6	Parma	Italy	Italy	0.06
7	Turin	Italy	Italy	0.05
8	Palermo	Italy	Italy	0.05
9	Modena	Italy	Italy	0.05
10	Bologna	Italy	Italy	0.04
11	Bari	Italy	Italy	0.04
12	Brescia	Italy	Italy	0.04
13	Genoa	Italy	Italy	0.04
14	Reggio Emilia	Italy	Italy	0.04
15	Prato	Italy	Italy	0.04
16	Catania	Italy	Italy	0.03
17	Rome	Italy	Italy	0.03
18	Taranto	Italy	Italy	0.02
1	Exeter	UK	UK	0.17
2	York	UK	UK	0.15
3	London	UK	UK	0.14
4	Bristol	UK	UK	0.14
5	Nottingham	UK	UK	0.14
6	Norwich	UK	UK	0.14
7	Newcastle upon Tyne	UK	UK	0.13
8	Manchester	UK	UK	0.13
9	Derby	UK	UK	0.12
10	Leeds	UK	UK	0.12
11	Southampton	UK	UK	0.12
12	Plymouth	UK	UK	0.11
13	Reading	UK	UK	0.11
14	Sheffield	UK	UK	0.10
1	Albuquerque	US	US	0.19
2	Oklahomacity	US	US	0.19
3	Boston	US	US	0.18
4	Jacksonville	US	US	0.17
5	Las Vegas	US	US	0.17
6	San Antonio	US	US	0.16
7	Nashville	US	US	0.15
8	Austin	US	US	0.15
9	Memphis	US	US	0.14
10	Louisville	US	US	0.14
11	Charlotte	US	US	0.13
12	Phoenix	US	US	0.11

Table 2.6 provides the full ranking of cities in our sample (by geographical area) based on the gender ratio of cyclists in the urban area.

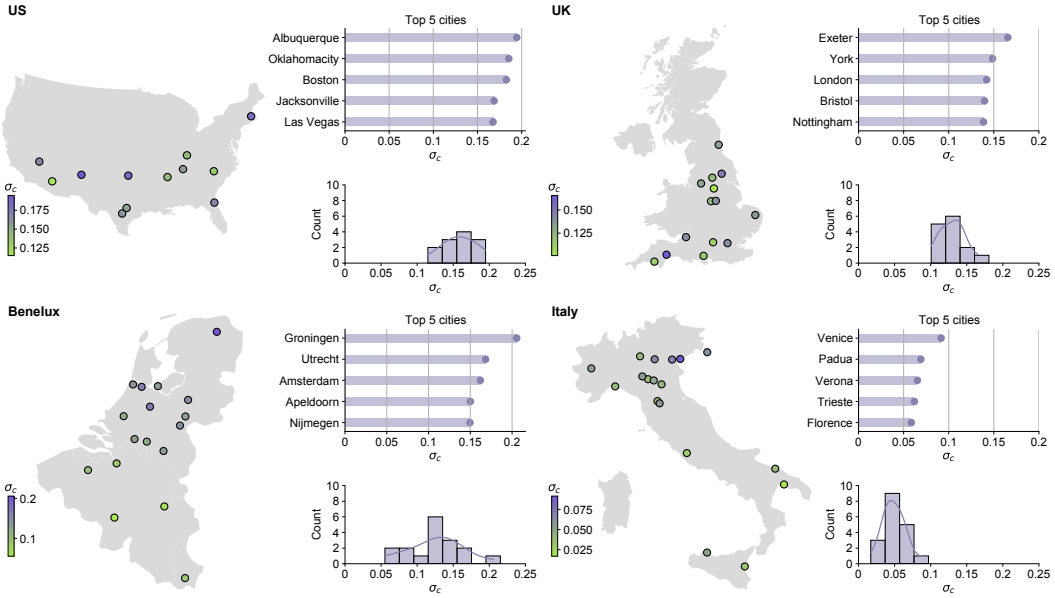

FIGURE 2.3: **Gender gap in recreational cycling in Strava: overview of cities included in the study** For each of the four geographical areas covered by the study, the figure depicts: the location of cities included in the analysis, the value of the female cycling rate σ_c for the five cities displaying the highest σ_c and the distribution of σ_c in the geographical area.

2.6 RQ1: What city-level characteristics are associated with a higher uptake of cycling by women?

2.6.1 Using Strava data to measure the gender gap in recreational cycling

In the first instance, we use Strava data to measure the gender gap in recreational cycling in 61 urban centres across four geographical areas: the United States, the United Kingdom, Italy, and Benelux. For each city c, we define the gender-cycling-gap as the ratio σ_c between the total kilometres travelled by female cyclists and the overall kilometres travelled by cyclists of both genders. This measure accounts both for gaps in trip shares among men and women and for differences in travelled distances. By construction, σ_c varies between 0 (no female cyclists) and 1 (no male cyclists): a value below 0.5 indicates the presence of a positive gender-cycling-gap (i.e. men cycling more than women). The closer the value to 0 the stronger the gap. For each geographical area covered by the study, Fig. 2.3 provides an overview of the cities included in the study, showing the five urban centres associated with the highest σ_c for each area (i.e. the lowest gender-cycling-gap), as well as the location and the distribution of σ_c of all covered cities (the full ranking is provided in Table 2.6). In our sample, the largest value for σ_c is 0.21 in the municipality of Groningen, Netherlands, indicating the presence of a substantial gender gap in recreational cycling for all cities under consideration.

Even within the same geographical area, we observe substantial heterogeneity in σ_c across cities. In the area of Benelux, in particular, σ_c ranges between 0.06 (Charleroi, Belgium) and 0.21 (Groningen, Netherlands). Dutch cities (particularly those in the northern regions) generally outperform cities in Belgium and Luxembourg. Among Italian cities, we observe a characteristic geographical pattern, with urban centres in the northeast displaying a lower gender-cycling-gap than cities in the south and

northwest. This north-south dichotomy is likely to be linked to the morphological characteristics of the country and the presence of a large flat land with a well-established cycling tradition. Differences in economic development might partially explain this structure as well. No geographical patterns are instead observable for cities in the United States and in the United Kingdom included in our sample. Interestingly, the is no evident link between the gender ratio and the size of a city. For instance, large cities such as Boston, Amsterdam and London perform high in the corresponding ranking, while top-ranking positions in Italy are dominated by relatively smaller urban areas.

It is important to note that the gender-cycling-gap measured using data from Strava may differ from official metrics on urban cycling provided by local and national administrations. For instance, in the Netherlands, national data indicate a similar cycling uptake among men and women [73]. This discrepancy likely arises because Strava is predominantly used for recreational purposes, while overall statistics do not distinguish between commuting and recreational purposes. Moreover, the level of adoption of Strava may vary across different geographic areas. For example, Strava usage data indicates varying usage patterns and adoption rates in the United States compared to other countries [74].

To limit the bias introduced by these elements, comparisons across cities are consistently performed within the same geographical area in this study. The fundamental assumption throughout the analysis is that the degree of penetration of Strava among men and women is similar across cities within the same geographical area [*homogeneity assumption*]. While we do not anticipate that the cycling behaviour recorded on Strava will be representative of the overall population's cycling behaviour, under the homogeneity assumption the observed variation in σ_c across cities within the same geographical area is not attributable to gender-specific patterns in the usage of the application.

2.6.2 Regression analysis

According to survey-based research on the gender gap in cycling, women tend to be more risk-averse than men resulting in a lower cycling rate by women (compared to men) in environments perceived as risky [46]. Based on this hypothesis, we examine the relationship between the gender ratio σ_c and two indicators of urban road safety. The first indicator (hereafter: *bike lanes*) measures the proportion of streets with cycleways (either protected or unprotected) in the street network. The second metric (hereafter: *speed limit*) provides the proportion of streets with a speed-limit of $20mph$ or 30 km/h or below. Both metrics are weighted using the length of each street. Figure 2.4 reports the scatter plots between the gender cycling gap σ_c and the two urban road safety metrics, for the four main geographical areas separately. Each marker corresponds to a city, the black line is the linear fit. Both measures of road safety display a positive correlation with the observed gender ratio for the area of Benelux. For cities in the United Kingdom, a positive (but weaker) correlation is only observable for the *speed limit* indicator. In contrast, for cities in Italy and the United States, both correlations are not statistically different from 0 (at neither a significance level of 0.05 nor 0.1). This lack of significant correlations could be attributed to the lower development of dedicated cycling infrastructure in these regions compared to cities in Benelux. For example, it is possible that a certain proportion of streets needs to be equipped with a cycleway for a city to be perceived as a safe cycling

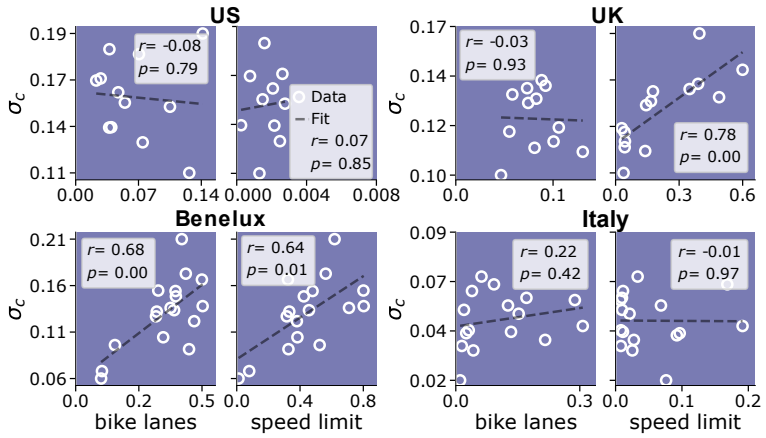

FIGURE 2.4: **Correlations between gender ratio and urban road safety indicators** The scatter plots show the correlations between two urban road safety indicators and the gender ratio σ_c, for cities in the four geographical areas separately. For each area, outliers to the three distributions of σ_c, *bike lanes* and *speed limit* were identified using the IQR Score method [75] and excluded. Each data point represents a city. The black line is the linear fit. The two urban road safety indicators capture the density of streets with a cycle lane in the street network (indicator: *bike lanes*) and the density of streets with a speed limit of up to 20 mi/h or 30 km/h in the street network (indicator: *speed limit*). A formal definition of the two indicators is provided in the Materials and Methods.

environment by women. In cities with limited dedicated cycling infrastructure development, the observed level of σ_c might be influenced by other factors related to the urban structure or the economic development of the area.

Although these findings are limited to specific geographic areas, the positive correlations suggest an association between road safety and σ_c, supporting the hypothesis that women's lower engagement with cycling (compared to men) may be due to a greater safety concern. To untangle the effect of confounding factors, we explore the relationship between σ_c and the two indicators of urban road safety controlling for a range of city-level indicators. To provide a thorough characterisation of each city, the indicators are chosen from four domains: 1) *E: Environment*, such as share of population in green areas, 2) *BEI: Built-Environment & Industrialization*, such as concentration of PM 2.5, 3) *SED: Socio-Economics & Demographics*, such as GDP per person, and 4) *M: Street Morphology*, such as average street grade. A full list of indicators is provided in Table 2.1 and the correlation matrix of the indicators across the entire sample is provided in Figure A.1 in Appendix A.

Coefficients (and 95% confidence intervals) of a linear regression model estimated via Ordinary Least Squares (OLS) are shown in Figure 2.5, with statistically significant coefficients at 0.05 level (two-tailed test) pictured in purple. Details on the model selection are provided in the Materials and Methods and the selected model presents an adjusted R^2 of 0.80. Overall, the regression analysis confirms the positive association between the gender ratio of cyclists and the *speed limit* indicator. This association indicates that urban centres with a relatively wider low-speed zone typically present a more balanced cycling uptake between men and women, after controlling for other confounding factors. Under the assumption that a wider low-speed zone indicates a less risky environment, this result also confirms that women

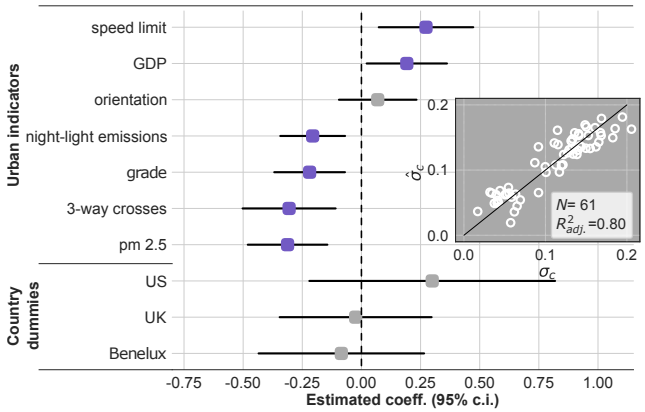

FIGURE 2.5: **Results of the regression analysis for RQ1** The main plot shows the estimated coefficients (square markers) with 95% confidence intervals (black lines) for the final set of regressors included in the model. Model estimated via Ordinary Least Square, selection performed via exhaustive search. Selection criterion: Akaike Information Criterion (AIC). Statistically significant coefficients at a 0.05 significance level are pictured in purple. The scatter plot displays the observed σ_c vs the fitted σ_c.

are more susceptible than men to the perceived level of risk of the cycling environment. Other insights emerge from the analysis of the control variables. First, we observe a negative association between σ_c and the proportion of 3-way crosses. From a topological viewpoint, cities with a high proportion of 3-way intersections deviate from grid-like street networks, that, by contrast, present a large prevalence of (mostly orthogonal) 4-way intersections [69]. This result can be interpreted again under the lens of the degree of safety of the urban environment for cycling. Indeed, the literature has shown that not only are crashes involving cyclists more likely to happen at non-orthogonal crosses than at right intersections, but the former are more likely to lead to severe injuries [76]. Another key urban feature relates to the morphology of the street network. The negative association between σ_c and the *grade* indicator shows that hillier cities display a larger gender gap in recreational cycling, controlling for all other factors. This result aligns with previous findings that women would have a preference for flatter routes [53] which may indicate a structural limit in the potential for cycling uptake by women in particular urban environments. Interestingly, the analysis also indicates a lower gender ratio in cities with worse air quality (higher concentration of PM 2.5). In the absence of a quasi-experimental setting, however, we are unable to determine whether the air quality is a relevant feature per se or if it acts as a proxy for other city-level characteristics such as motorized traffic. Finally, the results indicate a more balanced cycling uptake between men and women in relatively wealthier cities (with a larger GDP per person) and cities with a lower degree of night-light emissions (which can be a proxy for the size of the city). To test the robustness of this analysis, three additional models were estimated, using different mitigation strategies to account for the different levels of penetration of Strava worldwide. These strategies are described in Appendix A and differ in terms of: geographical coverage, specification of the geographical dummies and standardisation of the input and target variables (FigureA.2 in Appendix A). The results are largely consistent with the preferred specification presented here.

2.7 RQ2: What is the association between the presence of dedicated cycling infrastructure and the volume of female cyclists on a street (relative to males)

2.7.1 A case study: the city of New York

The findings for RQ1 show that aggregated urban features model well the heterogeneity of the gender gap in cycling observed across different cities. They also provide quantitative support for traditional hypotheses from the literature, which are often based on small-sample survey-based analyses. Although informative and affirmative, the previous analysis leaves two questions unanswered: Where do women prefer to cycle exactly? And what specific interventions could policymakers implement to enhance cycling for women?

To address these questions, we shift our focus from a macro-level comparison across cities to a micro-level setting, analysing streets within a city rather than the entire city itself. This change in perspective allows us to examine women's preferences for street-level characteristics in greater detail, identifying potential targets for interventions by policymakers. Among the available cities, we start by choosing New York City as a case study. Its extensive collection of administrative datasets provides an opportunity to enrich the analysis with data not otherwise available from OSM alone. A summary of all street-level features measured for this analysis is provided in Table 2.3.

Similar to the city-level analysis, we use Strava data on cycling to quantify female preferences for a street s. We measure the proportion of female cyclists out of all cyclists travelling via street s, denoted as σ_s. This indicator σ_s is a direct street-level extension of σ_c—indeed, σ_c can be constructed by averaging over σ_s with weights equal to the product of the length of each street and the total number of cyclists on it. The larger σ_s, the greater are women's preferences to cycle on street s.

Compared to a simple count of female cyclists, this relative measure has the advantage of quantifying female-specific preferences towards a street s, irrespective of the total level of "popularity" of the street. Therefore, the metrics will not be distorted towards streets that are very popular for cyclists in general (for instance, due to their position in the street network), but that may not present features that are particularly appreciated by our target group.

In addition, we adopt a data-driven approach to filter streets with a low number of cyclists (see Section 2.5). This filtering ensures that the observed σ_s is computed on a sufficiently large cyclist base. The distribution of σ_s is bell-shaped with a mean around 0.12 and a range between 0.00 and 0.41 (Figure 2.6). Stratifying the distribution by the protection level of the street (`No cycleway`, `Unprotected cycleway`, `Protected cycleway`), the analysis indicates that streets with no dedicated infrastructure typically have lower σ_s than streets with either protected or unprotected cycleways. The median value of σ_s for streets with `no cycleway` roughly corresponds to the 25th percentile of both the distributions of streets with a `protected cycleway` or an `unprotected cycleway`.

To further explore women's preferences for dedicated cycling infrastructure, we examine the degree of association between the presence of protected and unprotected cycleways and σ_s using multivariate logistic regression analysis. We classify streets into two classes, Low and High, corresponding to the bottom and top 33% of the

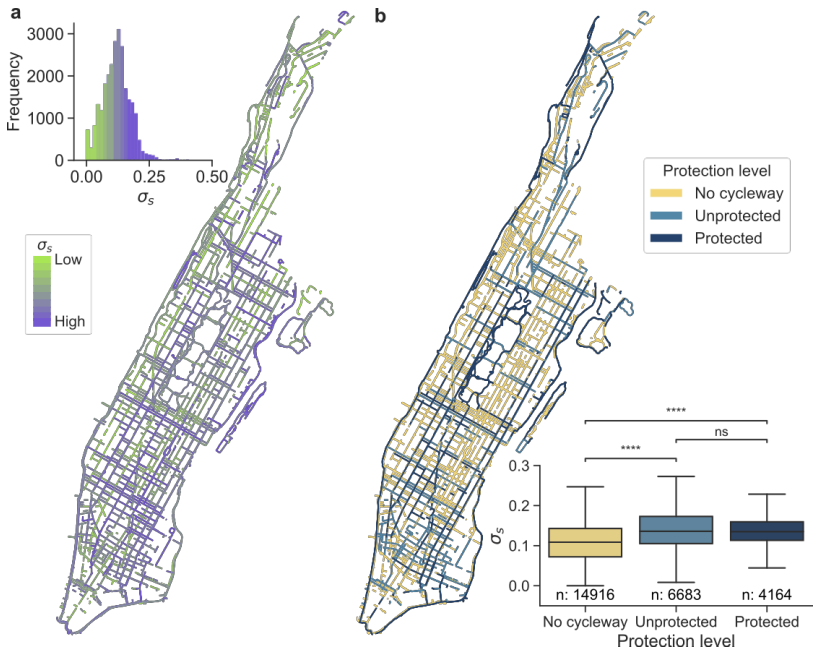

FIGURE 2.6: **The case of New York City: level of protection and σ_s** a) The map displays streets in the borough of Manhattan included in the final sample. A 10-quantile colour scheme has been used for the value of σ_s. The inset is the distribution of σ_s for streets included in the final sample (computed over the entire city of New York). b) The map displays the protection level of streets in the borough of Manhattan included in our sample. Yellow: no cycleway, light blue: unprotected cycleway, dark blue: protected cycleway. The inset displays the box plots of σ_s for streets with different levels of protection (computed over the entire city of New York). 'ns','*', '**', '***', '****' indicate the significance level of a Mann-Whitney-Wilcoxon test two-sided with Bonferroni correction, with the following p-values thresholds: 1e-4:'****', 1e-3: '***', 1e-2: '**', 0.05: '*', 1: 'ns'.

distribution of σ_s, and estimate the Odds Ratios (OR) via multivariate logistic regression. To check the robustness of the results, the analysis is repeated with different thresholds α (0.25 and 0.40, instead of 0.33) for the classification. The results (presented in Figure 2.7) are consistent across sample specifications, with generally slightly larger estimates on more extreme samples (lower values of the threshold α).

The main result of the regression analysis concerns the role of dedicated cycling infrastructure. With an estimated odds ratio (OR) of around 4.08 (95% confidence interval: [3.67, 4.54]), the analysis indicates that the odds of being classified as "High" are more than four times greater for protected cycleways than for streets with no cycleway (used as the baseline). This finding aligns well with the survey-based literature on the gender cycling gap, which suggests that women prefer physical separation more than men [46, 51, 47].

Although smaller in magnitude, we estimate a similarly positive association between the presence of an unprotected cycleway and σ_s. This analysis suggests that whenever protected cycleways are not feasible due to budget or physical constraints, the use of shared unprotected cycleways could still make the urban environment more accessible for female cyclists. In light of recent findings [77, 78], which suggest that unprotected cycleways do not necessarily enhance the safety of the road

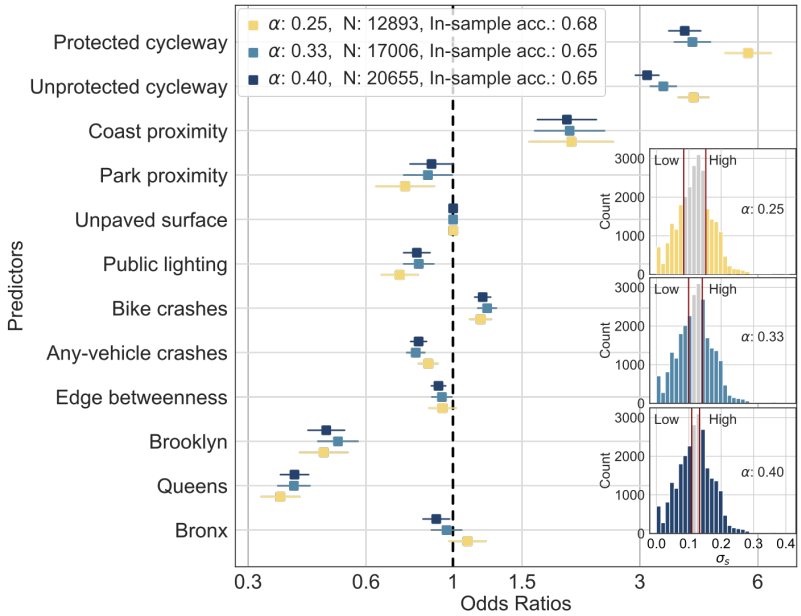

FIGURE 2.7: **Odds Ratios of multivariate logistic regressions, for several levels of the threshold α for the city of New York** a) the chart presents estimated ORs for a multivariate logistic regression where the target variable is the binarized σ_s and the predictors are listed in Table 2.3. The squared dots are the point estimates. The straight lines are the estimated 95% confidence interval for the corresponding OR. The model was estimated on three different sample selections, with the threshold α corresponding to 0.25 (yellow), 0.33 (light blue) and 0.40 (dark blue). The ORs are computed exponentiating the corresponding estimated coefficients. For each estimated model, the legend reports the value of the threshold α, the number of observations and the in-sample accuracy. b) The histograms show the mapping between the σ_s and the binarized σ_s for the three values of the threshold α: 0.2 5 (yellow), 0.33 (light blue) and 0.40 (dark blue).

network for cyclists, our results suggest that subjective safety may matter more than objective safety.

Regarding other control variables, consistent with the assumption that women prefer quieter streets, we estimate an OR below 1 for our proxy for traffic flow (edge-betweenness) and for the volume of accidents (by any type of vehicle). The positive association with the volume of bicycle crashes, on the other hand, is likely due to reverse causality: a more balanced gender ratio is typically associated with a larger volume of cyclists, increasing the likelihood of bicycle crashes.

The two dummies on coast and park proximity, inserted as proxies for the natural environment within which the street is located, appear to have opposite effects, with an estimated OR above 1 for coast proximity and below 1 for park proximity (with the latter statistically significant only at the 0.05 level for $\alpha = 0.25$). The high coast proximity value is due to the morphology of New York City and the presence of a long protected cycleway along the coastline of Manhattan, acting as an attractive infrastructure and impacting nearby streets too. The negative association with the park proximity dummy can be traced back to the location of the green areas under consideration, often in non-central locations (note that streets within Central Park largely fall into the excluded part of the distribution around the median value).

Information on the presence of public lighting is generally sparse in OSM, particularly for New York City (we assume public lighting to be absent only when explicitly stated, with less than 100 streets classified as without public lighting). Therefore, the negative estimated OR requires further analysis with more complete data.

Finally, although challenging to generalise to other urban contexts, we observe strong negative neighbourhood effects, particularly for the boroughs of Brooklyn and Queens (compared to the baseline borough of Manhattan).

2.7.2 Generalising the results to the other cities

In the previous section, we selected the city of New York as a case study due to the availability of additional street-level data from administrative sources. Furthermore, with its well-known city structure, New York serves as an ideal testing environment for urban studies.

In this section, we aim to assess the generalisability of the previous results on the role of protected and unprotected cycleways in other urban environments. For all cities in our sample, we conduct a similar analysis as the one undertaken for New York. We compute the odds ratios (ORs) of a minimal logistic regression, where the binarized street-level gender ratio is regressed against a categorical variable for the protection level of the street (three classes: No protection (baseline), Protected cycleway, and Unprotected cycleway). The processing and filtering of the data follow the same steps undertaken for New York. Results are reported in Fig. 2.8, where only ORs statistically different from 1 (at a significance level of 0.05) are depicted.

With the exception of eight cities (four cities in Italy, three in the US, and one in Benelux), the minimal models confirm a positive association between the presence of a protected cycleway and the probability that the street belongs to the high gender ratio class, as indicated by ORs above 1 for most cities in our sample. In contrast, no clear pattern is observed for unprotected cycleways across cities. This could be due to the fact that while protected cycleways ensure a certain degree of safety, the level of danger associated with unprotected ones likely depends on many additional features of the street, which are not necessarily encoded in this minimal model.

2.8 Discussion

Implications of the study

This chapter investigated the determinants of the gender cycling gap by examining patterns in cycling behaviours in over 60 cities in Europe and the United States. Unlike most previous analyses, which relied on survey-based data, we used large-scale data automatically collected from the online sport-tracking application Strava.

Initially, we examined the relationship between female cycling rates in various European and American cities and city-level characteristics [RQ1]. We found evidence supporting traditional hypotheses linking the observed gender gap in cycling to gender-specific preferences regarding road safety. Additionally, we observed higher female cycling rates in flatter cities compared to hillier ones, which aligns with findings in the literature [53]. This result suggests there may be structural, morphological, or cultural constraints in certain areas where increasing cycling uptake for women is more challenging [79]. For urban planning, this result implies that ad-hoc infrastructural interventions, such as providing cycleways or expanding low-speed

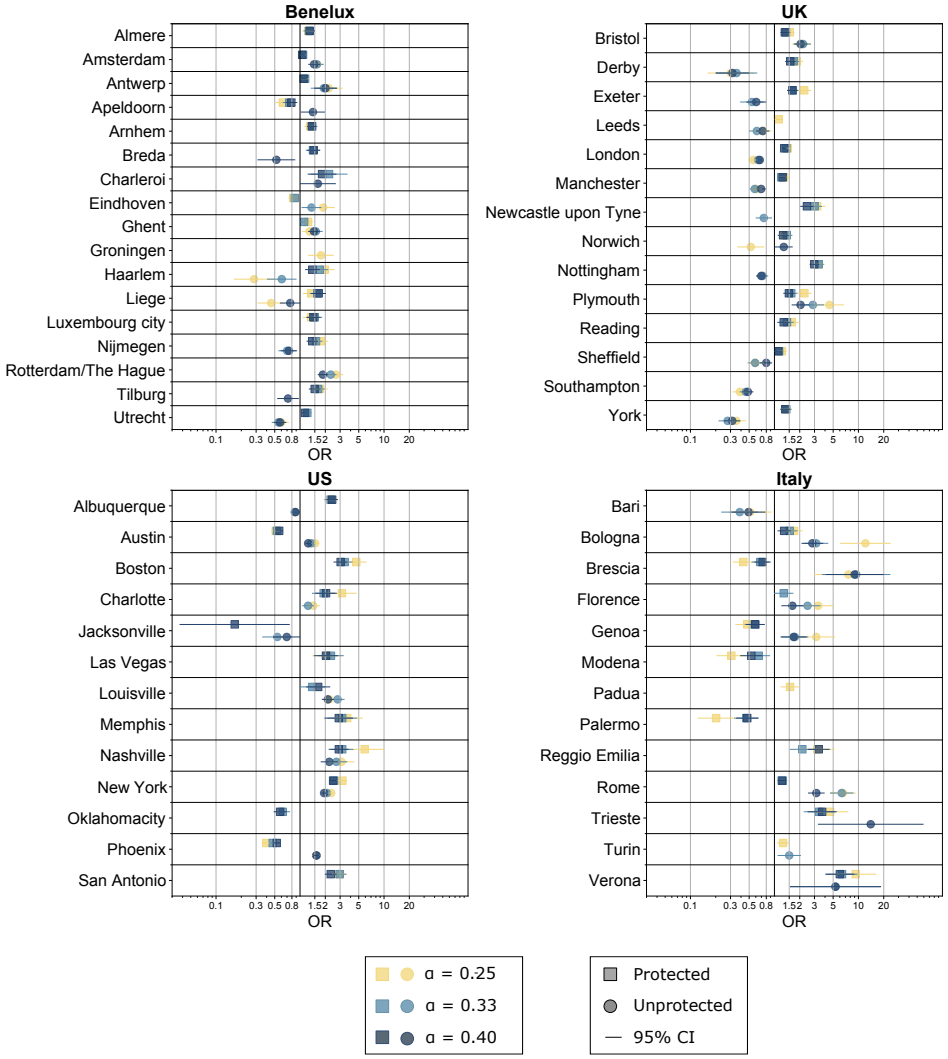

FIGURE 2.8: **Odd Ratios of minimal logistic models, all cities** The charts present estimated ORs of minimal logistic regressions, where the target variable is the binarized σ_s regressed against a categorical variable for the protection level of the street (three classes: *No protection* (baseline), *Protected cycleway* and *Unprotected cycleway*). The models are run for each city separately. The squared markers are the point estimates for *Protected cycleway* and the circular markers are the point estimates for *Unprotected cycleway*. The straight lines represent the estimated 95% confidence interval for the corresponding OR. Only statistically significant ORs (significance level 0.95) are pictured. For each city, the minimal model is estimated on three different sample selections, with the threshold α corresponding to 0.25 (yellow), 0.33 (light blue) and 0.40 (dark blue). The ORs are computed exponentiating the corresponding estimated coefficients. Only cities with at least one statistically significant ORs are pictured.

limit zones, may have limited efficacy in these contexts. Instead, they may require concurrent behavioural incentives, such as promoting the adoption of e-bikes.

In the second part of the study, we shifted the focus from a macro comparison across cities to a micro-level analysis, where edges of the street network are used as unit of analysis. Compared to the first research question, which provided evidence for and

expanded existing hypotheses (further validating our data as a reliable source on cycling behaviour), the second analysis aimed at capturing the role of urban features modelled at a higher resolution and delving deeper into the association between the gender-cycling-gap and the presence of dedicated cycling infrastructure. We selected the city of New York as a case study for this component of the study. Using multivariate logistic regression analysis, we have shown the existence of a positive association between the volume of female cyclists (relative to men) and the presence of dedicated cycling infrastructure. The positive association between σ_s and the presence of a protected cycleway was expected and well-documented in the literature, which highlights the strong preference of women for physical separation from motorized traffic [46, 51, 47]. More novel and interesting is the observed association with the presence of an unprotected cycleway. In light of recent studies showing that unprotected cycleways may not enhance the degree of objective road safety [78, 77], our result suggests that perceived safety may influence women to cycle more than actual safety in specific environments. However, this finding does not apply to all cities in our study, indicating that interventions in this regard may not be equally effective everywhere.

In contexts where physical separation is not possible (due to space or budget constraints), providing shared cycleways may still improve the perceived accessibility of urban environments for women in certain contexts. However, since the perceived increase in safety induced by this type of infrastructure may not always correspond to a decrease in actual risk, city planners should carefully evaluate the planning of such infrastructure. For example, they could prioritise specific solutions associated with higher safety levels.

Overall, our study quantitatively validated survey-based results using unprecedentedly large-scale automatically collected data. With approximately 36 million users worldwide in 2018, Strava was one of the major applications for sport tracking, providing reliable information on cycling behaviour for regular cyclists.

Limitations of the study and future research

The analysis is subject to three main limitations. The first limitation pertains to the representativeness of Strava users and the purposes of Strava trips. For example, having a considerable gender gap in the Netherlands (Figure 2.3), contrary to expectations [36], the Strava data are clearly not representative, and neither users nor purposes of use can be inferred. Nevertheless, we did our best to account for the representativeness challenge of this data, first by comparing only cities in the same geographical area, and second by comparing streets only in the same city, aiming to minimise user and trip purpose variation. Despite these mitigation strategies, it remains unclear to which extent our results can be generalised to cycling for purposes other than recreational, such as transport, and to less-skilled cyclists (occasional cyclists and not cyclists). Future analyses will require finding data sources that are able to reliably distinguish between such purposes and users since gender-based constraints can differ between these categories [80]. However, since the survey-based academic literature on gender-cycling-gap indicates that cycling preferences differ less among regular cyclists than among occasional ones [46, 81], the results of our analysis could be interpreted as a lower-bound and it is likely that the identified factors play an even larger role in explaining the gender-cycling-gap in the general population.

A second limitation of the Strava data set is the inability to extract exactly the potentially useful information of cyclist volumes [82], as the raw data are not individual cycling traces but Strava segments with only aggregated statistics. This aggregation also implies that the same cyclists may cycle on many segments in one or multiple sessions and we would not be able to identify them. The third limitation pertains to the cross-sectional nature of the available cycling data. The absence of a longitudinal dimension restricted the analysis of temporal variations in the data, thus hindering the use of statistical tools for policy evaluation to assess causal effects along with correlations.

Finally, there is a variety of gender-specific constraints apart from street safety that future studies should take into account, from cultural and psychological reasons [82, 79], to other environmental factors and harassment by motorists [80, 83]. Gender inequality and gendered transport habits may also play a large role, such as more frequent trip chaining by women due to childcare and other errands [84, 55]. Therefore, while street safety and urban design are undoubtedly important ingredients, there is no universal, simple fix for getting rid of the gender gap in cycling towards more sustainable mobility. It remains a complex societal issue that needs to be tackled from multiple angles [38].

Chapter 3

ATGreen: a multi-dimensional computational framework to evaluate accessibility to urban green

3.1 Overview of the chapter

This chapter is based on the publication titled *On the need for a multi-dimensional framework to measure accessibility to urban green*, published in NPJ Urban Sustainability by A. Battiston, and R. Schifanella [19].

This chapter focuses on understanding the degree of interchangeability of several structural metrics of green accessibility commonly adopted in the public health and urban planning communities. It proposes a computational approach to measure these frequently adopted spatial indicators of green accessibility within a unified framework. The framework is then used to evaluate the impact of using one metric compared to another in a policy design approach, highlighting the need to move towards multi-dimensional assessments. Along with the theoretical contribution, the chapter also introduces a novel web-based interactive resource (ATGreen) facilitating the use of the proposed framework and the shift towards the multi-dimensional perspective, for both policymakers and the general audience.

The remainder of the chapter is structured into seven sections. Section 3.2 outlines the motivations of the study, summarising the key findings on the relationship between green accessibility and public health. Section 3.3 summarises the main methodological approaches to the measurement of green accessibility. Section 3.4 details the materials and methods employed, providing an overview of data sources and analytical techniques. Sections 3.5 and 3.6 provide an overview of the computational framework and the statistical analysis evaluating the interchangeability of the several green accessibility metrics, respectively. The functionalities of the associated web interface are then presented in Section 3.7. Finally, Section 3.8 discusses the implications and limitations of the study.

3.2 Scope

As urban populations continue to expand, there is increasing interest among academics and policymakers in devising strategies to enhance the liveability and sustainability of cities. Among these strategies, urban greening interventions and nature-based solutions (NBSs) have emerged as critical approaches due to their capacity to mitigate the environmental impacts of urbanisation. A well-developed green infrastructure plays a vital role in biodiversity conservation, carbon sequestration, soil protection, and temperature regulation, providing multiple ecosystem services that contribute to the overall resilience and sustainability of urban areas [85, 86, 87, 88].

Simultaneously, several theories (including the biophilia hypothesis [89], attention restoration theory [90], and prospect-refuge theory [91]) have postulated a link between exposure to natural environments and improved health outcomes. These theories suggest that humans possess an intrinsic need for nature-rich environments, which can significantly enhance mental health, cognitive function, and overall well-being. Empirical research has substantiated these hypotheses, demonstrating that proximity to greenspaces is associated with a wide range of physical and mental health benefits, including lower mortality rates, reduced risk of cardiovascular disease, and positive effects on happiness and cognitive abilities [92, 93, 94, 95, 96, 97, 98, 99, 100, 101].

Building on this growing body of evidence, Milvoy et al. [102] introduced a comprehensive model in 2015 that links greenspace availability to the health and well-being of urban populations. This model identifies key health pathways—both direct and indirect mechanisms—through which the presence of nature leads to improved health outcomes. Direct pathways include enhanced physical activity, promotion of social interaction, reduction of stress levels, and exposure to improved air quality and cooler, shaded areas. Indirect pathways, on the other hand, relate to biodiversity support, which, according to the authors, offers ancillary benefits for human health. The insights provided by this model, along with empirical studies linking health outcomes to nature exposure, have contributed to the development of SDG 11.7[1] [13], which imposes a responsibility on local authorities to ensure equitable access to public green spaces for all, with particular attention to protected demographic groups, by 2030.

In alignment with the objectives outlined in SDG 11.7, various health organisations, local authorities, and institutional bodies have set a range of green-related targets for cities[2], which are utilised both for monitoring and evaluation, as well as for the purposes of optimising the design of greening interventions. The World Health Organization (WHO), for example, recommends that urban residents have access to at least 0.5–1 hectare of public green space within 300 metres of their residences [103]. The broad spectrum of benefits associated with exposure to nature has in some cases been incorporated into multi-level targets, setting different green space requirements for increasing distances from residential locations (e.g., [104, 105]). Similarly, the recently proposed 3-30-300 paradigm addresses the need for urban greenery to permeate the everyday lives of urban residents. According to this paradigm, *three trees* should be visible from every home, each neighbourhood should achieve

[1] SDG 11.7 states: *By 2030, provide universal access to safe, inclusive, and accessible green and public spaces, particularly for women and children, older persons, and persons with disabilities.*

[2] Although these targets align with the aims of SDG 11.7, some were established prior to the SDG and should not be regarded as directly derived from it.

a *30% tree canopy* cover, and every residence should have a greenspace within *300 metres* [106].

This chapter seeks to deepen our understanding of the potential to incorporate green accessibility indicators into urban planning policy decisions related to greenspace provision. Despite the growing interest in this area indeed, there is still no standard methodology for evaluating green accessibility, and the implications of using various metrics within a policy design context are not yet fully understood (see Section 3.3). To address this critical gap, this chapter introduces a novel computational framework for high-resolution greenspace accessibility measurement, currently implemented through a web interface (ATGreen) covering over 1,000 cities globally. This framework enables the comparison of multiple accessibility indicators and assesses the impact of introducing new greenspaces. Furthermore, I investigate the variability in accessibility patterns across different indicators, highlighting the necessity of adopting a multidimensional approach rather than relying on a single metric. This comprehensive analysis aims to inform and guide urban planning decisions, ensuring that greenspace provision is optimised to support both environmental sustainability and public health outcomes.

3.3 Related work

Over the past decade, advancements in high-resolution geographical data and computational capabilities have driven a shift in urban planning towards more data-driven policy design processes. An integral part of this transition is the creation of spatially resolved indicators. These indicators act as benchmarks for tracking progress towards specific goals and assist in the design of new urban interventions, aligning with the principle "what gets measured, gets done" [18, 107].

Despite significant interest from the research and policy communities in urban greening interventions, there remains no established, universally adopted framework to measure accessibility to urban greenspaces. Moreover, the various proposed approaches have not been evaluated within a unified framework. These approaches differ based on: (1) the type of data used to identify green urban features and/or (2) the family of indicators considered.

In terms of data on urban green features, three main types of sources can be distinguished: (1) **Land cover and land use data from (processed) satellite imaging, administrative data sources, or crowd-sourced initiatives**. Land cover data categorises the Earth's surface into different types, such as forests, urban areas, croplands, water bodies, and barren lands. Land use data reflects human activities and their impact on the landscape, including settlement patterns, agriculture, transportation, and infrastructure development. These data are frequently used for the evaluation of accessibility to greenspaces in single-city studies and often employ distance-based accessibility metrics [108, 109, 110, 111, 112, 113, 114]. Their main advantage, particularly for land use data, is the ability to recognise the functional use of greenspaces, for instance distinguishing between public and private areas. (2) **NDVI-based metrics from (processed) satellite imaging data**. Derived from the visible and near-infrared light reflected by vegetation, the Normalized Difference Vegetation Index (NDVI) assesses and quantifies the density of green vegetation in a specific area. NDVI quantifies each image pixel with values ranging from -1 to +1, with higher values indicating more greenness. Greenery can be calculated as a simple average of NDVI values in each pixel or by determining a suitable threshold on

the index to define green area pixels [115, 116]. NDVI-derived measures of greenery are prevalent in public health research [116, 117, 118, 119]. Unlike land cover data, NDVI-based metrics do not allow segmentation of green information based on function. (3) **View-based greenery**. Initially proposed by Yang et al. [120], this approach uses street view imagery to measure eye-level street greenery. Unlike previous approaches, street view imagery captures the landscape at eye level and has gained popularity due to its compatibility with human perception. However, due to its computational intensity, this approach is still limited to small-scale evaluations for single cities [121, 122, 123, 124, 125] or a limited number of urban areas [126].

Unlike metrics that look at the *average* characteristics of a city, the definition of spatial indicators poses specific computational and methodological challenges. Indeed, it requires considering the interplay between the spatial distribution of the population and greenspaces within a city, as well as the proximity and/or the walkable catchment of each sub-area resulting from the topology of its street network. As mentioned later, this latter element has been neglected in some specific applications due to its computational complexity. Among the several spatially-resolved green accessibility indicators proposed in the academic literature, it is possible to distinguish two broad categories [98, 115]: (1) **Cumulative opportunity (or exposure) measures**. These metrics measure the percentage or sum of green areas within a spatial unit or a circular buffer around the residential location. While the spatial unit could theoretically be based on the walkable catchment around a specific area, most large-scale studies use circular buffers for computational convenience. This type of metric is popular among public health studies aiming to compare the performance of many cities in terms of green exposure of their residents [127] (e.g. the proportion of residents that satisfies specific targets). However, fixed circular buffers do not account for discontinuities in measurement due to the thresholding approach. Recent studies, such as [128], acknowledge the importance of measuring short-range and long-range metrics by computing green exposure for buffered regions with progressively increasing radii around residential locations in both Global South and Global North cities. (2) **Distance-based green accessibility metrics**. Primarily used in urban planning, these metrics measure the distance (walking, cycling, or driving) to the closest area with specific characteristics and are typically instantiated with a threshold for the minimum size of a green area (e.g., 0.5, 1, 2 hectares) [115]. A binarised version of these metrics (e.g., whether an area with specified characteristics is available within a specific temporal or spatial threshold) forms the basis of institutional targets set by local authorities or other public bodies. These include the previously mentioned the World Health Organization (WHO) target, recommending access to at least 0.5-1 ha of public green within 300m of residential locations [103] and the multi-level targets proposed by specific local authorities and national agencies [104, 105]. Despite their wide adoption for urban planning, distance-based metrics have generally only been applied in single-city studies [109, 110, 111, 112] and have been disregarded for larger evaluations. In addition, studies based on this type of metric often lack a full discussion on the impact of selecting specific thresholds compared to alternative specifications. A first attempt towards the adoption of these metrics for large-scale evaluations is provided by [113], although the study is restricted to the evaluation of greenspaces within 500 meters of residential locations.

While the studies mentioned above examined green accessibility through the lens of the structural and topological features of urban areas, a different and more recent research line integrates behavioural information into evaluations of green exposure and accessibility. This behavioural information is derived from survey data

and, more recently, from automated user-generated geographical data, including social media, sports tracking apps, mobile phone traces, and public participation geographic information systems (PPGIS) [129, 130, 131, 132]. By using behavioural information, these studies quantify actual greenspaces' usage and green exposure levels (or *green demand*) in contrast to the *potential* green accessibility inferred solely from structural considerations. Although these metrics represent substantial progress in the characterisation of green accessibility and shed light on factors like the perceived quality or safety level of greenspaces, their large-scale implementation is often hindered by restricted data accessibility. Furthermore, concerns have been raised about the representativeness of some of these data sources, particularly regarding specific demographic groups [133, 134]. Due to these challenges, structural indicators are still predominantly adopted in evaluating green accessibility.

Despite the proliferation of green accessibility indicators, no research has yet evaluated how interchangeable these metrics are or the implications of using one over another for policy design. By defining a computational framework designed to operationalise the measurement of three families of green accessibility indicators and evaluating their stability under several parameterizations, this research aims to fill this gap. At the same time, it provides a tool for policymakers to evaluate multiple indicators, thus adopting a multidimensional perspective on green accessibility.

3.4 Materials and Methods

Definition of urban centre

For this study, urban centres[3] (UC) were defined according to the boundaries in the Urban Centre Database of the Global Human Settlement 2015, revised version R2019A (GHS-UCDB) [2]. Information on this dataset has already been provided in Section 2.4. Out of 13,000 urban centres recorded in the database, we retained the most populated 50 UCs per country (the internationally recognized three-letter ISO code identifies a country), provided that they had at least 100,000 inhabitants (corresponding to roughly 2,600 UCs). To ensure only cities with sufficient quality of the OpenStreetMap (OSM) data [3] –used extensively throughout this analysis– were included in the study, two additional filters were applied to the sample: 1. Exclusion of all UCs with less than 20 distinct public green areas recorded in OSM (about 600 cities were excluded, and 1,963 cities remained in the sample). To this scope, public green areas were defined according to the *key* : *value* pairs in Table 3.1. 2. Exclusion of 923 urban centres, with low data quality. The data validation was performed by comparing the green intensity appearing from OSM data (based on an extended definition of green) to the green intensity from the World Cover data 2020 from the European Space Agency (WC-ESA) [6] and defining ad-hoc acceptance intervals for urban centres with different size and average green intensity, based on the level of similarity observed for a set of reference cities. A detailed description of the validation approach is provided in section 3.4. The final sample for the study comprised 1,040 urban centres across 145 countries.

[3] As in the previous chapter, the terms urban centre, urban area, and city are here used interchangeably to indicate the units described in this paragraph.

Geographical units of analysis

The geographical space of each UC was divided into a regular grid with a spatial resolution of 9 arc-seconds (geographic projection: WGS-84), mimicking the grid of the population layer of the Global Human Settlement 2015 (revised version 2019A) (GHS-POP) [1]. Cells in the grid are the smallest geographical unit of analysis for this study, meaning that all metrics were measured at this geographical level and, whenever appropriate or needed, aggregated into higher geographical units.

Population data

Population data were extracted from the 9-arcsecond resolution grid of the GHS-POP [1]. The raw data consisted of residential population estimates for the year 2015, disaggregated from census or administrative units to grid cells and informed by the distribution and density of built-up as mapped in the corresponding Global Human Settlement Layer global layer. For each urban centre, the data were processed using the rioxarray package in Python, masking the global *.tiff* file with the boundary of the city enlarged with a three-km buffer. The clipped raster was then loaded into a PostGIS database.

Definition of PGAs and other green features

In this section, the terms *Public Green Areas* (PGAs) and *greenspaces* are used interchangeably to indicate accessible green areas of public use. The terms *urban green*, *green infrastructure*, and *green coverage* instead are used to refer to all green features in an urban centre, regardless of their use or their degree of accessibility. For each city, PGAs were extracted from OSM data following the pipeline described in the following three paragraphs and reclassified into three classes: *parks*, *grass*, and *forests*. For each city, the *green infrastructure* was extracted from the WC-ESA (codes: 10, 20, and 30) following the pipeline described in the subsequent paragraph.

Extraction of local osm.pbf dumps The OSM data dump for each continent was downloaded from the GeoFabrik Download Service in May 2022. From these, we generated the geographical extract of the OSM data for each UC. To ensure that locations at the boundary of the study area are not biased, the perimeter of the urban area defined in the GHS-UCDB was buffered with a three-km radius. The extraction was performed using the *smart strategy* of the method *extract* of the Osmium Library. The *smart strategy* runs in three passes and extracts: 1. all nodes inside the region and all ways referencing those nodes as well as all nodes referenced by those ways; 2. all relations referenced by nodes inside the region or ways already included and, recursively, their parent relations; 3. all nodes and ways (and the nodes they reference) referenced by relations tagged *type=multipolygon* directly referencing any nodes in the region or ways referencing nodes in the region. The default configuration was used for all other options of the method.

Extraction of public green areas from OSM For each UC, the second step consisted of the identification and extraction of public green areas from the local osm.pbf file. In this study, we defined as public green areas all those OSM spatial elements defining closed areas (i.e. closed ways and relations) with the *key* : *value* pairs in Table 3.1. The extraction was performed using the Python bindings of the osmium-tool. The geometries of relations were reconstructed from their members using the

Category	Key	Value
Parks	leisure	park
	leisure	garden
	landuse	recreation_ground
	landuse	village_green
Forests	landuse	forest
	natural	wood
Grass	landuse	grass
	landuse	meadow
	natural	grassland
	natural	shrubbery

TABLE 3.1: OSM *key* : *value* pairs used for the identification of public green areas

approach described in the official OSM guidelines. Elements with *key* : *value* pairs equal to *access:no* or *access:private* were excluded from the retrieved database. In the final step, we classified green OSM elements into three categories (`Parks`, `Forests`, `Grass`) according to the mapping in Table 3.1. The obtained green features were then loaded into the PostGIS database.

Remapping of public green area to the base grid The third step involved the remapping of geographical elements extracted from OSM to the base grid of each UC. For each cell in the grid – and for each combination of green elements (`Park`, `Forests`, `Grass`, `Parks & Forests`, etc.) intersecting the cell– we generated the following information on: 1. size of the intersection between the green element and the cell; 2. the overall size of the green element intersecting the cell. All values were expressed in hectares and rounded to the second decimal point so that the minimum resolution is a continuous green space of at least 100 square meters. It should be noted that green OSM elements might partially overlap (for instance, the same area could be mapped as *leisure* : *parks* and *landuse* : *grass*). To ensure the same area is not computed multiple times and continuous green (for instance, adjacent forests and grasslands) was appropriately identified, overlapping areas were dissolved together before undertaking the remapping exercise.

Green coverage from the World Cover data 2020 of the European Space Agency: extraction and processing The World Cover data 2020 [6] from the European Space Agency (WC-ESA) was used in this study for the computation of one class of green accessibility indicators (exposure indicator) as well as reference data for the sample validation, as described in the 3.4. The data provides global land cover maps for 2020 at 10 m resolution based on Copernicus Sentinel-1 and Sentinel-2 data. Out of all cover classes, classes *10 - Tree cover*, *20 - Shrubland*, and *30 - Grassland* were defined as *green coverage*. For each city, the data were extracted by clipping the worldwide file with the boundary of the UC buffered with a 3-km radius following the procedure described in the official guidelines and subsequently loaded into the PostGIS database. The data were then remapped to the base grid of each city following the approach described in Section 3.4 for the remapping of the OSM polygons to the base grid.

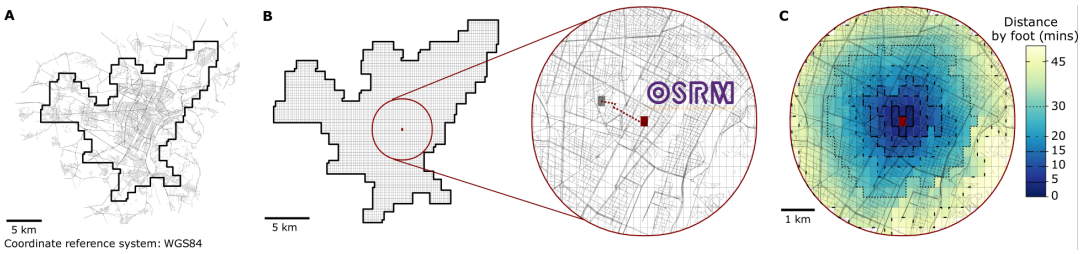

FIGURE 3.1: **Computation of walking distance matrices. The example of the city of Turin, Italy** A and B provide a schematic representation of the pipeline for the computation of the walking distances for the city of Turin, Italy. A depicts the street network for the buffered region (GHS-UCDB boundary buffered with a 3 km radius), extracted from the local OSM dump and the city boundary (solid black line). B depicts the base grid of the city of Turin. For each cell (depicted in red in the example), the walking distance between the cell's centroid and the centroids of all cells within a 3 km buffered region (red circle) was computed using the local routing engine Open Source Routing Machine. C provides an example of the walking distances computed using the described approach. The origin cell was depicted in red. Solid/Dashed/Dotted lines indicate several isochrones for the cell of interest.

Computation of walking distance matrices

In the proposed framework, the problem of computing the walking distances between residential areas and green spaces was simplified to the computation of walking distances between the centroids of cells composing the base grid. To this scope, we used the routing engine Open Source Routing Machine (OSRM) [4] and the street-network data from the local OSM dumps (see Section 3.4). A schematic representation of the adopted strategy for Turin, Italy, is provided in Figure 3.1.

In particular, for each UC, we adopted the following four-step pipeline.

1. From all cells in the extended grid of the city $\mathcal{M}$, define the set $\mathcal{Z}$ (with members: z) to be the subset of $\mathcal{M}$ with non-null population.
2. Characterise each cell z in $\mathcal{Z}$ and each cell m in $\mathcal{M}$ with the lat-long coordinates of their centroids.
3. For each z in Z: (a) Identify the subset of cells in $\mathcal{M}$ within a geodesic buffer of 3 km. Call this set $\mathcal{M}'_z$. For computational reasons, we did not compute the full distance matrix. Instead, for each starting cell, we restricted the computation of walking distances to cells within a radius of 3 km (geodesic distance). Given that this work focuses on walking accessibility, this simplification has little impact on our results, as cells outside the boundaries are unlikely to be within walking distance. (b) Use the matrix routing engine OSRM (with default configuration parameters for the *foot* profile) to compute the walking distance in minutes between the centroid of z and the centroids of all m' in $\mathcal{M}'_z$.

As the matrix computation of OSRM provides incorrect results if the starting/ending point is not sufficiently close to a walkable street, we used a heuristic approach to identify potentially problematic starting and ending cells by intersecting ex-ante the base grid with the walkable street network (defined mimicking the parameters of

the *foot* profile in OSRM). If a cell presents no intersecting street, it is assumed to be inaccessible, and all distances are set to NA.

Validation of the sample of cities

The quality of OSM data is constantly under scrutiny by researchers who rely on this data source for their studies [135]. On one side, this has led to the development of specific web services for visually comparing OSM data with other mapping systems, such as the OSM Map Compare tool provided by provided by BBBike and Geofabrik. On the other side, the scientific community has developed several methods to quantify the quality of OSM data. These range from intrinsic evaluation methods (among the first contributions: [136, 137, 138]), where metrics are defined using OSM data itself, to approaches based on comparing these with external data sources (reference data) (some of the first contributions include [139, 140, 141]).

As the suitability of OSM data for a particular application depends mainly on the specific problem being tackled, researchers cannot rely on a predefined list of areas. Instead, they should evaluate data quality in the context of their particular problem. Here, we propose an approach to validate the sample of cities to include in our study based on the comparison with the WC-ESA [6], here used as a reference dataset. The pipeline for the data validation comprised two phases:

1. Definition of a Quality Score of green elements for each city in the sample;
2. Identification of a set of cities with similarly high data quality using k-means clustering.

Each step is defined in detail below.

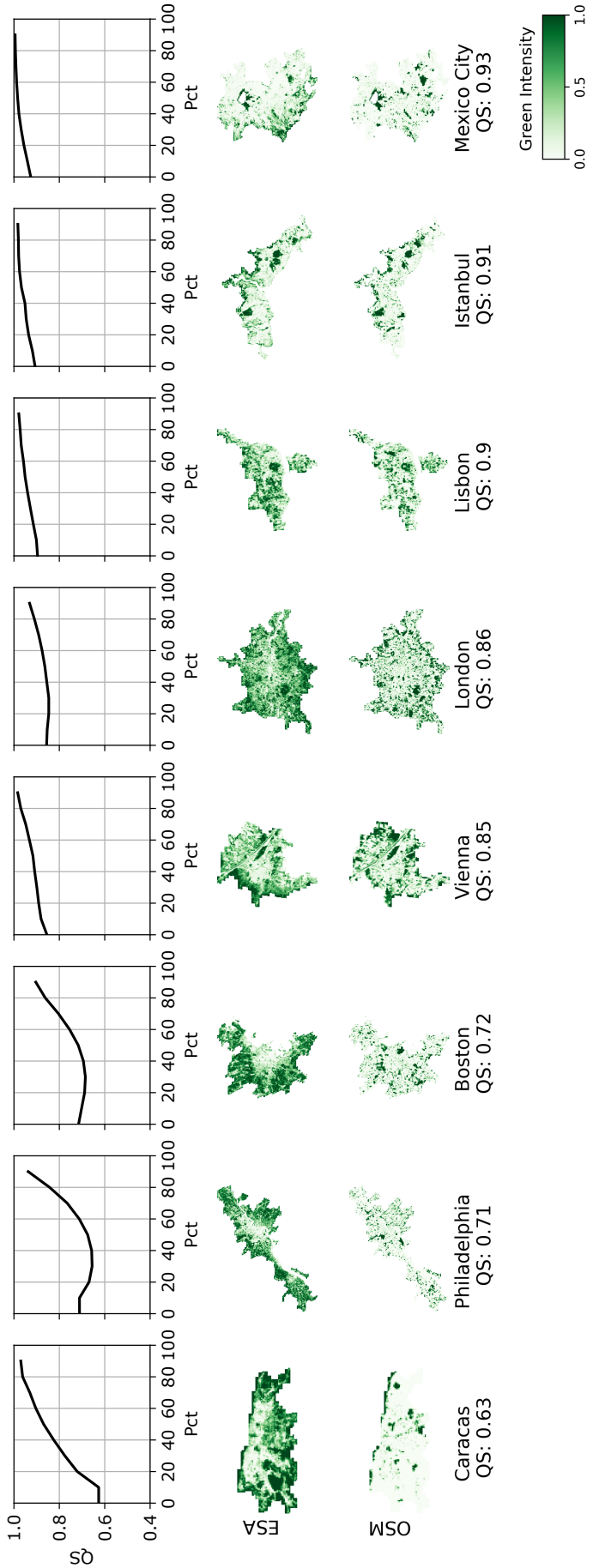

FIGURE 3.2: **Green intensity images and quality score: an example for eight cities** The chart displays the green intensity images generated using data from the European Spatial Agency World Cover 2020 data set (middle row) and data from OpenStreetMap (bottom row) for the cities of Caracas (Venezuela), Philadelphia (United States of America), Boston (United States of America), Vienna (Austria), London (United Kingdom), Lisbon (Portugal), Istanbul (Turkey) and Mexico City (Mexico). For each city, the top row displays the change in the Quality Score obtained by removing from the images cells/pixels of the progressively increasing population [i.e., the bottom 10% (20%, 30%, etc) of the cells (presented on the x-axis)].

Definition of a Quality Score of green elements, for each city in the sample For each city in our sample, we computed the OSM green data Quality Score (QS) based on the comparison of green features extracted from OSM and green features observable in the WC-ESA, here used as an external reference dataset. To this scope, we extended the list of green features in OSM to include features that -while not classifiable as public green areas- might appear as green cover in the satellite dataset. A list of additional OSM tags retained to this scope is provided in Table B.1 in Appendix B. Only OSM elements of the classes `closed ways` and `relations` were extracted, and the choice of tags was based on a manual assessment of randomly selected cities. The data were processed and remapped to the base grid following the approach described for the main OSM greenspaces, in Section 3.4.

Borrowing techniques from the digital image processing domain, the QS was defined to be proportional to the mean squared error between a green intensity image generated using OSM data and an equivalent picture generated using land cover data from the WC-ESA. In particular, for both data sources and each city, we first projected the green information on a 9-arcsec resolution grid and characterised each cell with a green intensity value corresponding to the proportion of the cell covered by a green feature (the values range between 0 and 1). By treating each cell in the grid as a pixel, we then generated two images of green intensity, one using data from OSM and the other using data from the reference dataset. For each city in the sample, the mean squared error among the two images was then computed as:

$$MSE_c = \frac{1}{N}\sum_{i \in I}(GI_{OSM,i} - GI_{ESA,i})^2 \tag{3.1}$$

where $GI_{data,i}$ is the green intensity value of pixel i from the image generated using data from *data*, I is the set of pixels (cells) composing the image, and N is the cardinality of I. By construction, the MSE is bounded between 0 (identical green intensity) and 1 (complementary green intensity). For each city c, the QS was then defined as

$$QS_c = 1 - MSE_c \tag{3.2}$$

Figure 3.2 provides a visual example for eight cities.

Identification of a set of cities with similarly high data quality using k-means clustering The second step involved the identification of a set of cities with sufficiently high and homogeneous quality. To this scope, we proceeded as outlined below:

1. First, we selected a sample of cities (*reference list*) from countries where OSM data are generally used for academic research. While we do not specifically evaluate the mapping quality for these cities, our scope was to identify all those cities in our dataset for which the data quality is comparable to cities in the reference list. We included in the *reference list* all those cities in our initial sample located in Italy, France, Netherlands, Germany, Austria, United Kingdom, Luxembourg, Belgium, the United States, and Canada.

2. For cities in the *reference list*, we then studied the association between the QS and two city-level metrics: the size (captured by the natural logarithm of the number of pixels in the image, or $log(N)$) and the average green intensity (measured taking the average of the green intensity of the image generated using the WC-ESA across all pixels, or $Mean\ GI = \frac{1}{N}\sum_{i \in I} GI_{ESA,i}$). As depicted

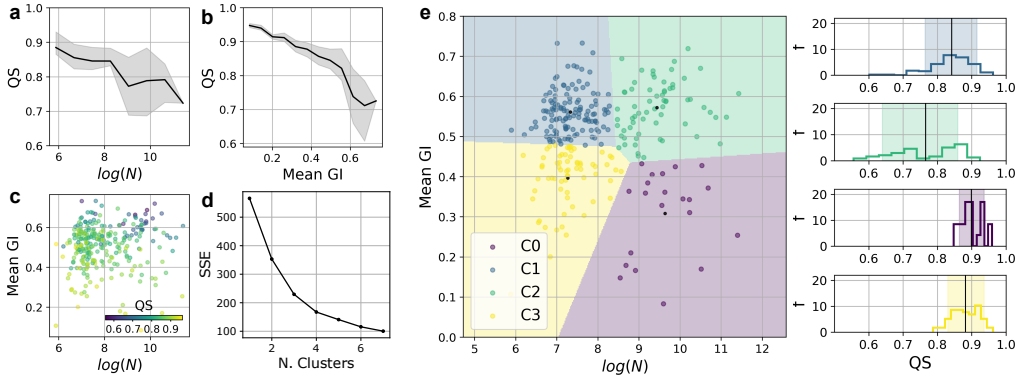

FIGURE 3.3: **Validation of the sample of cities** a) displays the Average (solid line) and IQR (shaded area) Quality Scores of cities in the *reference list* as a function of the logarithm of the number of pixels/cells ($log(N)$). b) displays the average (solid line) and IQR (shaded area) Quality Scores of cities in the *reference list* as a function of the Mean GI, i.e., the average value of green intensity for pixels within the boundary of the city. c) is a scatter plot of the $log(N)$ and *Mean GI*. Each point represents a city in the *reference list*. The color represents the Quality Score. d represents. d) represents the k-means inertia for different numbers of clusters, obtained when clustering cities in the *reference list* based on their tuple of values ($log(N)$, *Mean GI*). A z-scores transformation was performed before clustering the data. e) is a scatter plot of obtained clusters and corresponding regions identified by the corresponding decision boundaries. The insects on the right depict the distribution of the QS of cities in the *reference list* (histogram) and the admittable region (shaded area) for each cluster (C=: purple, C1: blue, C2: green, C3: yellow).

in panels a, b, and c of Figure 3.3, the QS decreases with the size and average greenness of the city – i.e., larger and/or greener cities have a lower QS than smaller and less green ones. It should be noted that this negative association depends on the definition of the QS rather than implying a difference in the quality of the mapping. Indeed:

- The QS of cities with low green coverage is artificially inflated by the absence of features to be mapped. This is the case because the lower the number of features to be mapped, the lower the information that the QS provides. A powerful example is the extreme case of cities with no green coverage. By construction, the QS would be equal to 1 in these cases. However, a QS equal to 1 does not guarantee a generally good quality of OSM data.

- For geometrical reasons, larger cities have more peripheral pixels with lower population densities. The positive association between the population density and the QS (see top row of Figure3.2) determines the lower QS observed for these large urban centres.

3. Due to the negative association between the size of the city and its average greenness and the QS, the use of a single cut-off value to identify cities to be included in our study might result in the inclusion of some urban centres with insufficient data quality (for instance if the city has poor green coverage) or the

exclusion of others with sufficient data quality (for example large cities). By k-means clustering cities in the *reference list* based on the size and the average greenness, we identified four regions of tuples of values ($log(N)$, *Mean GI*) (Panel e in Figure 3.3) and defined an admittable range of QS values as values between the $10th$ and the $90th$ percentile of the QS of cities in the *reference list* belonging to the specific cluster (shaded areas in the insects of panel e of Figure 3.3). Standard preprocessing was performed for the k-means clustering.

4. The final sample of cities was then defined as the union of cities in the *reference list* and cities outside the *reference list* whose QS fell within the admittable interval for their tuple of ($log(N)$, *Mean GI*) values.

The final sample for the study comprised 1,040 urban centres. Figure B.1 in Appendix B for information on the characteristics of cities included and excluded from the sample.

Stability metrics

In Section 3.5, we will introduce three families of green accessibility metrics, each defined by a set of underlying parameters. As an example, let's consider the family of minimum-distance indicators. For this family, a parameter is the minimum size of greenspaces, meaning that areas below this size are disregarded in the computation of the indicator. In section 3.6, we will then present an evaluation of the stability of each family of green accessibility indicators with respect to its underlying parameterization. Three metrics are used to evaluate the stability of the ranking induced by changes in the parameterization of each family of indicators. These metrics are discussed in greater detail when presenting the results, while here we provide a formal definition.

Throughout this section, the following notation is adopted: $\mathcal{N}_c$ is an ordered set of N elements representing the cells within the city boundary in the population grid of city c, $\mathcal{M}_c$ is an ordered set of M elements representing the cells in the extended (3km-buffered) grid for city c, D_c (D_c^+) is an $(N \times M)$-matrix ($(M \times M)$-matrix) where each element $d_{i,j}$ ($d_{i,j}^+$) represents the walking distance (as per street-network) in minutes between the i-th element of $\mathcal{N}_c$ ($\mathcal{M}_c$) and the j-th element of $\mathcal{M}_c$.

Kendall rank correlation coefficient Let $R_{Ind(x)}[N]$ be the ranking induced by the accessibility indicator Ind with parametrization (x) on the set of cell $\mathcal{N}_c$ of the urban centre c. The Kendall rank correlation coefficient [142] between two parametrizations (1) and (2) of indicator Ind is given by:

$$\tau = 1 - \frac{number\,of\,discordant\,pairs}{\binom{N}{2}} \tag{3.3}$$

where two pairs n_1 and n_2 in $\mathcal{N}_c$ are said to be *concordant* if either $R_{Ind(1)}[n_1] < R_{Ind(1)}[n_2]$ and $R_{Ind(2)}[n_1] < R_{Ind(2)}[n_2]$ or $R_{Ind(1)}[n_1] > R_{Ind(1)}[n_2]$ and $R_{Ind(2)}[n_1] > R_{Ind(2)}[n_2]$, otherwise they are *discordant*.

Naive targeting approach Let $Ind(x)[N]$ be the accessibility indicator Ind with parametrization (x) on the set of cell $\mathcal{N}_c$ of the urban centre c. For the minimum

distance indicator, let $t^{\star}_{x}(y)$ be

$$t^{\star}_{x}(y) = \min\left\{ t : \left[\frac{\sum_{n\in\mathcal{N}_c} P_n(1 - \mathbb{1}_{[0,t)}(Ind_x[n]))}{\sum_{n\in\mathcal{N}_c} P_n} \right] \cdot 100 \geq y \right\} \tag{3.4}$$

i.e. $t^{\star}_{x}(y)$ is the minimum value of the indicator associated with the performance of the bottom y% of the population (i.e. it is the minimum distance to a greenspace with determined characteristics of the individual performing better than y% of the total population of the urban centre). We call $t^{\star}_{x}(y)$ the cutoff value of the indicator associated with the y% targeting strategy.

Noticing that, unlike the minimum distance indicator, higher values of the exposure and per-person indicators are desirable, the cutoff value $t^{\star}_{x}(y)$ for these two families is defined as:

$$t^{\star}_{x}(y) = \min\left\{ t : \left[\frac{\sum_{n\in\mathcal{N}_c} P_n \mathbb{1}_{[0,t]}(Ind_x[n])}{\sum_{n\in\mathcal{N}_c} P_n} \right] \cdot 100 \geq y \right\} \tag{3.5}$$

Then, for any two parametrizations (1) and (2) of Ind, for the minimum distance indicator, we defined the proportion of the stable targeted population under the y% naive targeting approach as the following weighted Jaccard indicator:

$$S_{naive}(y)^{md} = \frac{\sum_{n\in\mathcal{N}_c} P_n(1 - \mathbb{1}_{1,t^{\star},y,n})(1 - \mathbb{1}_{2,t^{\star},y,n})}{\sum_{n\in\mathcal{N}_c} P_n(1 - \min[\mathbb{1}_{1,t^{\star},y,n}, \mathbb{1}_{2,t^{\star},y,n}])} \tag{3.6}$$

where $\mathbb{1}_{x,t^{\star},y,n} = \mathbb{1}_{[0,t^{\star}_{x}(y))}(Ind(x)[n])$. For the exposure and per-person indicators:

$$S_{naive}(y)^{exp,pp} = \frac{\sum_{n\in\mathcal{N}_c} P_n \mathbb{1}_{1,t^{\star},y,n} \mathbb{1}_{2,t^{\star},y,n}}{\sum_{n\in\mathcal{N}_c} P_n \max[\mathbb{1}_{1,t^{\star},y,n}, \mathbb{1}_{2,t^{\star},y,n}]} \tag{3.7}$$

where $\mathbb{1}_{x,t^{\star},y,n} = \mathbb{1}_{[0,t^{\star}_{x}(y)]}(Ind(x)[n])$. It is noteworthy that given the presence of potential ties (some cells may have the same accessibility value) induced by Ind(x), the number of people belonging to the bottom y% of the induced ranking population might differ under $Ind(1)$ and $Ind(2)$.

Most-disadvantaged targeting approach The approach to targeting the most disadvantaged is akin to the naive targeting method, with a key distinction. Rather than relying solely on a cutoff value determined by the ranking of the population, which remains agnostic to the actual value of the indicator, this approach determines the cutoff by incorporating information on the value of the indicator.

For the exposure and per-person indicators, we defined the target group as those with no exposure/per-person access under the parameterization (x). Thus, the proportion of the stable population in the target group under any two parameterizations (1) and (2) is defined as the resulting weighted Jaccard indicator:

$$S^{exp,pp}_{most-dis} = \frac{\sum_{n\in\mathcal{N}_c} P_n \mathbb{1}_{[0]}(Ind(1)[n]) \mathbb{1}_{[0]}(Ind(2)[n])}{\sum_{n\in\mathcal{N}_c} P_n \max[\mathbb{1}_{[0]}(Ind(1)[n]), \mathbb{1}_{[0]}(Ind(2)[n])]} \tag{3.8}$$

For the minimum distance indicator, we define the most-disadvantaged target population as that subgroup that performs y-times worse than the Thus, letting t^{m}_{x} be

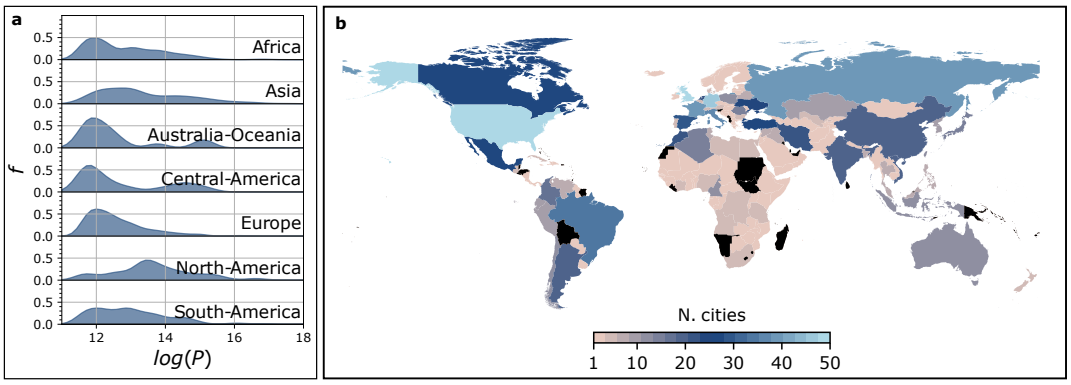

FIGURE 3.4: **Geographical coverage of the study** a) The density plot reports the size (measured as log-population) distribution of cities included in the sample for each macro-area. Cities in the Russian Federation have all been attributed to Europe. b) The map displays the number of cities included in the study for each country. Black indicates that no city was included for the corresponding country.

the average value of the indicator under the parametrization (x), then:

$$S^{md}_{most-dis}(y) = \frac{\sum_{n\in\mathcal{N}_c} P_n(1-\mathbb{1}_{1,t_1^m,y,n})(1-\mathbb{1}_{2,t_2^m,y,n})}{\sum_{n\in\mathcal{N}_c} P_n(1-\min[\mathbb{1}_{1,t_1^m,y,n},\mathbb{1}_{2,t_2^m,y,n}])} \tag{3.9}$$

where $\mathbb{1}_{x,t_x^m,y,n} = \mathbb{1}_{[0,yt_x^m)}(Ind(x)[n])$.

3.5 A computational framework to evaluate several families of green accessibility metrics

Building upon policy debate and recommendations of public-health authorities [103, 104, 105] on green accessibility and exposure, we defined a computational framework to measure three families of spatial indicators of green accessibility. The ability to measure multiple families of metrics simultaneously allows for a multidimensional characterisation of access and exposure to urban greenery. The three families are: Minimum distance, exposure (or cumulative opportunities), and per-person green availability. The framework enables a flexible parameterization of each family of indicators according to the minimum size of the PGAs/green features of interest, the type of greenery, and (when applicable) the time budget and is currently deployed in an interactive web interface for 1,040 cities worldwide – whose geographical distribution is depicted in Figure 3.4. Figure 3.5 provides a schematic representation of each family of indicators, while a formal definition is provided in what follows.

The following notation is used throughout the section: $\mathcal{N}_c$ is an ordered set of N elements representing the cells within the city boundary in the population grid of city c, $\mathcal{M}_c$ is an ordered set of M elements representing the cells in the extended (3km-buffered) grid for city c, D_c (D_c^+) is an $(N\times M)$-matrix ($(M\times M)$-matrix) where each element $d_{i,j}$ ($d^+_{i,j}$) represents the walking distance (as per street-network) in minutes between the i-th element of $\mathcal{N}_c$ ($\mathcal{M}_c$) and the j-th element of $\mathcal{M}_c$.

Minimum distance This family of indicators measures the walking distance (in minutes) from a residential location to the nearest public green area. The indicator can be parameterized according to the minimum size of the green area and the type of green (combinations of parks, forests, and grass).

For each cell i in the population grid $\mathcal{N}_c$, the indicator is defined as:

$$md_{i,c} = min(D_{i,c} \odot gd_c) \tag{3.10}$$

i.e., the minimum of the Hadamard product between the i-th row of D_c and the M-dimensional vector gd_c taking value 0 or 1 to indicate the absence/presence of a green feature (with given characteristics in terms of size/type of green) in the corresponding cell of the extended grid $\mathcal{M}_c$.

Exposure (or cumulative opportunities) This family of indicators measures the amount of urban green space available from a residential location within a walking time budget (t minutes), in hectares. In contrast to the minimum distance indicator, which focuses solely on green areas that are both *accessible* and *public*, the exposure indicator assesses the overall presence of green features in the vicinity of a residential area, regardless of their accessibility or intended use. This approach expands our definition of green elements to encompass additional urban greenery features like roadside tree lines. To implement this broader perspective, we use the WC-ESA [6] as our primary data source to identify green elements rather than relying on OSM. As for the previous class, the indicator can be parameterized according to the value of t and the minimum size of the green area.

For each cell i in the population grid $\mathcal{N}_c$, the indicator is defined as the sum of the green intensities measured on cells no more than t-minutes away from the cell of origin. I.e.:

$$exp_{i,t,c} = \mathbb{1}_{(0,t]}(D_{i,c}) \times gi_c \tag{3.11}$$

where $\mathbb{1}_{(0,t]}(D_{i,c})$ is an indicator function mapping each element of the M-dimensional vector $D_{i,c}$ equals to 0 if $d_{i,j}$ is greater than t and equals to 1 otherwise. gi_c is an M-dimensional vector representing the size of the green features (with given characteristics in terms of size) in the corresponding cell of the extended grid $\mathcal{M}_c$.

Per-person This family of indicators measures the per-person availability of public green areas within a time budget (t minutes) from a residential location in square meters. Unlike the previous two indicators, agnostic to population density, it incorporates the notion of competitiveness in using PGAs for specific activities. As a result, the level of public green available to a resident depends on the total green provision and the cumulative number of people living within the service area of each PGA. As for the previous classes, the indicator can be parameterized according to the value of t, the minimum size of the green area, and the type of green.

For each cell i in the population grid $\mathcal{N}_c$, the indicator is computed as:

$$pp_{i,t,c} = \mathbb{1}_{(0,t]}(D_{i,c}) \times gpp_{t,c} \tag{3.12}$$

where $gpp_{t,c}$ is an M-dimensional vector representing the squared meters of green available per-person in the corresponding cell of $\mathcal{M}_c$. More specifically, $gpp_{t,c}$ is

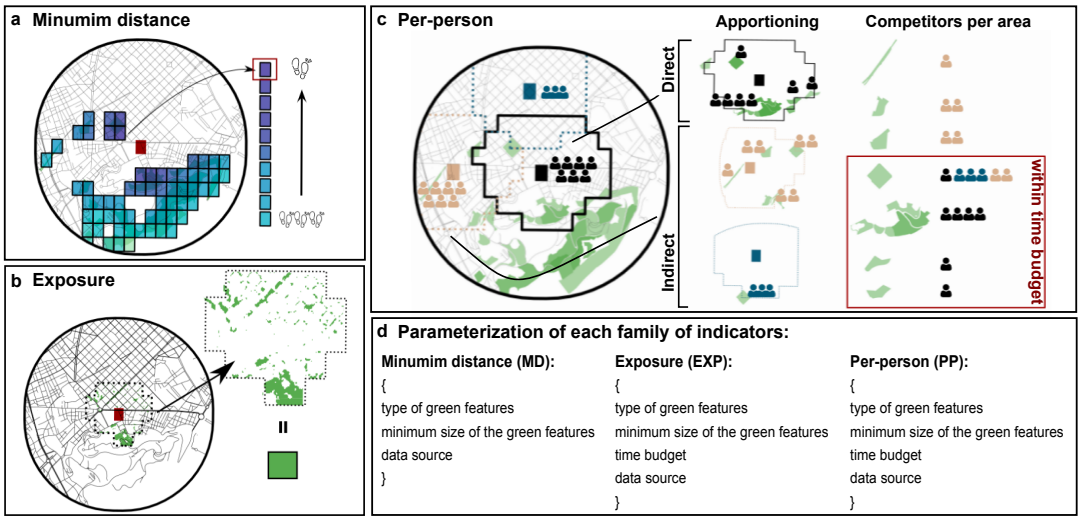

FIGURE 3.5: **Graphical representation of the three families of accessibility indicators** a) *Minimum distance*: The indicator measures the walking distance in minutes to the closest cell with a PGA with selected characteristics in terms of size and type of green for each residential cell. In the top-left panel, the red cell is the cell of interest. The remaining cells are nearby cells with a PGA, whose colour is proportional to the distance to the cell of interest. b) *Exposure*: The indicator measures the cumulative size of green features (in hectares) available within a walking time budget from a residential cell. In the top-right panel, the red cell is the cell of interest. The area within the walking time budget from the cell of interest is depicted with a dotted line. c) *Per-person*: This indicator is computed in two steps. In the first step, all residents in the urban centre are apportioned to PGAs – within the corresponding walking time budget – proportionally to the size of the PGA. In the second step, for each residential cell, the per-person indicator is computed as the ratio between the size of PGAs within the walking time budget from the cell of interest, and the total number of residents apportioned to these areas –irrespective of their residential location. In the bottom panel, the cell of interest, its residents, and the area within the corresponding walking time budget are depicted in black. The same information for competitor users from other residential cells is depicted in pink and blue. d) Summary of the parameterization of each family of indicators.

computed by dividing the green available in each cell by the total confluent population. I.e.:

$$gpp_{t,c} = gi_c \oslash AP_{(t,c)} \tag{3.13}$$

where $\oslash$ refers to element-wise division, AP is the M-dimensional vector whose element $ap_{j,t,c}$ equals the confluent population (for the time-threshold t) of the j-th element of $\mathcal{M}_c$. The affluent population of cell j is computed by assigning – for each residential cell i – shares of the population to cells in $\mathcal{M}$ within the time-budget t proportionally to the size of the available green in j. Formally, the confluent population of cell j in $\mathcal{M}$ for time-budget t is defined as:

$$ap_{j,t,c} = \sum_{i \in \mathcal{M}} P_i \frac{\mathbb{1}_{(0,t]}(d^+_{i,j,c}) \cdot gi_{j,c}}{\sum_{m \in \mathcal{M}} \mathbb{1}_{(0,t]}(d^+_{i,m,c}) \cdot gi_{m,c}} \tag{3.14}$$

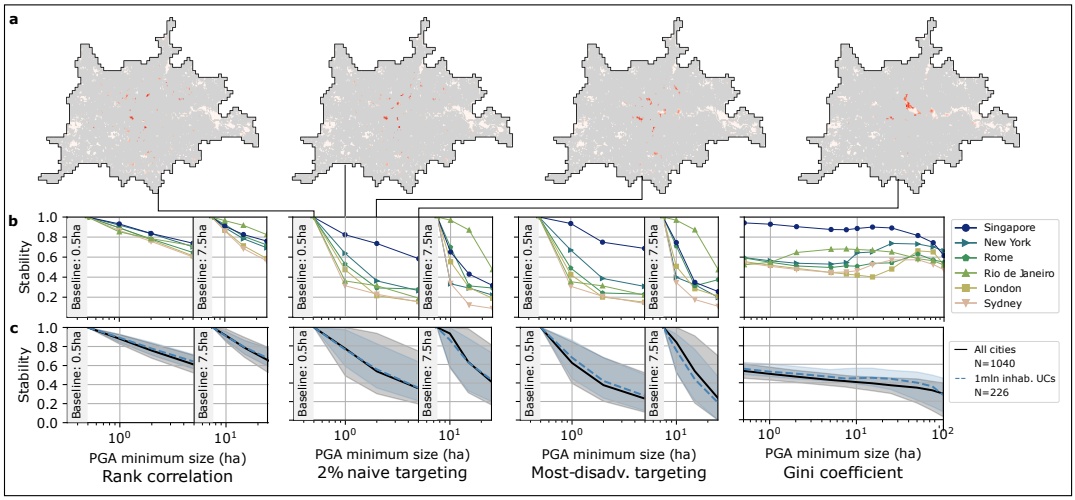

FIGURE 3.6: **Stability of minimum distance indicator to the minimum size of greenspaces** a) The maps depict in red target areas in London (UK) according to the *2% naive targeting approach* for the minimum distance indicator with increasing minimum greenspaces' size (from left to right: 0.5 ha, 1 ha, 2 ha, and 5 ha). The intensity of the red is proportional to the number of residents in the cell. b) The charts depict the level of stability of the minimum distance indicator to different parameterizations and according to different stability metrics for six cities across all continents. The *Most-disadvantaged targeting* targets residents performing worse than three times the mean citizen. For the *Rank correlation*, *2% naive targeting* and the *Most-disadvantaged targeting*, the comparison is provided with respect to the parametrization with minimum size equal to 0.5ha for minimum sizes up to 5ha and to 7.5ha for larger minimum sizes. For the *Gini indicator*, the chart reports the indicator's value under several parametrizations. c) The charts report the median value (solid line) and the IQR (shaded area) of the stability metrics for all cities in our sample (black) and for cities with more than 1 million inhabitants (blue).

3.6 Assessing the need for a multidimensional perspective in the evaluation of green accessibility in urban areas

Having defined a framework for the computation of several families of green accessibility indicators under a range of parameterizations, we then examined the importance of adopting a multidimensional perspective when using green accessibility metrics for policy design, as opposed to relying on a single indicator. In the next two sections, we present the results of two evaluations aimed at assessing the stability of the ranking of performance among different areas and the relative populations in each city induced by 1. different parametrizations of the same family of indicators, 2. green accessibility targets proposed by institutional bodies. A formal definition of the stability metrics was provided in section 3.4.

Stability of the green accessibility families under different parameterizations

Institutional targets and guiding principles frequently lack precision regarding the specific parametrization of an indicator, as illustrated, for instance, in [103, 104]. In

other cases, they recognise the importance of a multidimensional perspective, directly suggesting the use of multi-level green targets. Despite this, the vast majority of single-city or large-scale evaluations neglect this multidimensional perspective. Furthermore, there is still limited knowledge on how changes to the parametrization of green accessibility metrics affect the relative performance of different areas and sub-groups of the population within a city. To address this gap, we analysed the impact of minor changes in the parameterization of each family of indicators along *three dimensions*, introducing —where applicable— an appropriate stability metric. (i) The first dimension is the ranking, in terms of green accessibility, of the geographical units composing the city. We measured the stability of the rankings induced by two alternative parameterizations using the Kendall rank correlation coefficient [142]. This metric quantifies the degree of agreement between two ordered sets based on the ranking of their elements. In this first dimension, every geographical unit is equally important for assessing the ranking stability, and information on the number of residents in each cell is not factored in. (ii) To account for this, the second dimension extends the first one by additionally considering the population distribution in the city. Furthermore, rather than evaluating the entire ranking, for this dimension, we adopted a policy perspective by examining the stability of the population located at the lower end of the ranking, which is expected to be the focus of policy interventions. To this scope, we proposed two targeting strategies, i.e. strategies to identify under-served subgroups of the population of a city. With the first strategy, labelled *naive targeting*, we targeted the y% worst performing population, irrespective of the actual performance; with the second strategy, labelled *most-disadvantaged targeting*, we targeted subgroups of the population with low performance in either absolute term or relative to the rest of the population (see Section 3.4). We then defined the stability of an indicator to any two parametrizations (baseline vs. alternative) under a targeting strategy as the share of the overlapping targeted population. For a stability level s, the average proportion of the target population under one parameterization who would not be targeted under the alternative one (hereafter: `conflicting target population`) is given by $\frac{(1-s)}{(1+s)}$. (iii) As a third dimension, we computed the observed inequality level under the indicator's specific parameterization, measured through a weighted Gini indicator [143]. The two initial dimensions are grounded in the considerations of policy design. Although there is no universally agreed-upon planning strategy, there is a consensus that interventions should focus on *left-out* subgroups of the population to mitigate urban inequities [144]. The inclusion of the third dimension stemmed, instead, from the growing use of the Gini indicator in measuring urban inequalities in green accessibility studies [145, 146, 147, 148].

Figure 3.6 illustrates the stability of the minimum distance indicator in response to variations in the minimum size of the greenspaces across our entire sample of cities (mean and inter-quartile range (IQR)), cities exceeding 1 million inhabitants (mean and IQR), and selected major urban centres worldwide. For the evaluation, the ranking induced by parametrizations up to 5 hectares (ha) is compared to a baseline of 0.5 ha. Larger parametrizations are evaluated against a baseline of 7.5 ha. The use of two different baselines ensures that the comparison is provided across indicators measuring accessibility to greenspaces with a similar size, here used as a proxy for the potential uses of the area.

We observe a consistent decline in the median stability level as the minimum size

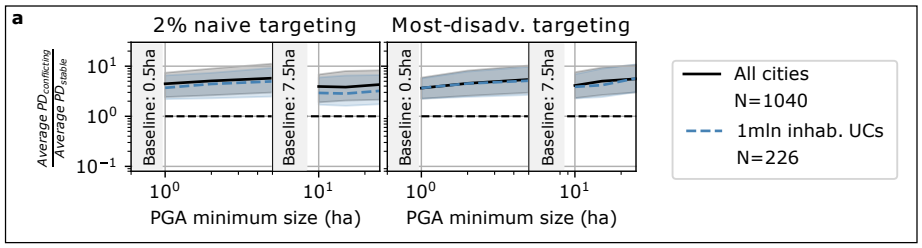

FIGURE 3.7: **Population density of conflicting and stable targeted areas** a) The panel depicts the ratio of the average population density (*PD*) of the areas associated with the conflicting targeted population to the average *PD* of areas linked to the stable targeted population under two targeting strategies and a range of alternative minimum PGA sizes for the minimum distance indicator. The line is the median across all cities (in black) and cities with more than 1 million inhabitants (blue); the IQR for both cases is depicted as a shaded area.

of greenspaces increases, for both the targeting approaches. This trend is consistent among all cities regardless of population size. Notably, the generally higher stability values at the area level (the Rank correlation displayed in Figure 3.6) mask significant reshuffling at the bottom of the population ranking. Under the *2% naive targeting strategy*, the median stability level across all cities shifts from 0.77 [IQR: 0.48-1] for a parameter change from 0.5 ha to 1 ha to 0.53 [IQR: 0.28-0.94] for a shift to 2 ha. This corresponds to a *conflicting target population* of 12% [IQR: 35%-0%] and 30% [IQR: 56%-3%], respectively. These results underscore substantial variability in stability depending on the specific green configuration of the urban centre, along with concerning levels of instability, particularly in selected large cities. For instance, using the same parameterizations (minimum PGA size of 0.5 ha and 1 ha), the stability levels for Sydney (AUS), Rio de Janeiro (BRA), and London (UK) are 0.32, 0.36, and 0.47, respectively, corresponding to *conflicting target population* levels of 51%, 47%, and 36%. Similar trends emerge with the more restrictive *most-disadvantaged targeting strategy*, which tightly focuses on populations with low accessibility levels as well as adopting different targeting levels from the naive targeting approach (presented in Figure B.4 in Appendix B). A visual assessment of the changes in the targeted population for the city of London (UK) is provided in the maps in the top row of Figure 3.6, which display – in a scale of reds –targeted areas using the *2% naive targeting strategy* under five parameterizations, with the intensity of the palette being proportional to the number of people living in the area. From the graphical comparison, we observe that:

1. For small changes in the parameter (e.g., from 0.5 ha to 1 ha), most of the stable population (i.e., populations targeted under both parameterizations) is concentrated in low-density areas. While some (but fewer) higher-density areas are targeted in both scenarios, they tend not to overlap. It is worth noticing that this is not only a peculiarity of London but holds for most cities. Indeed Figure 3.7 shows that the ratio between the population density of areas associated with a conflicting target population and areas with a stable target population is consistently above 1.

2. The degree of clustering of targeted areas increases along with the minimum size of the greenspaces. This is due to the physical and geographical constraints of larger greenspaces, which are typically fewer and less scattered around the city than smaller ones.

Name	Type	Size	Distance	Target	Data	Source
WHO	MD	0.5 ha	5 mins	5 mins	OSM	[103]
B1	MD	0.5 ha	10 mins	10 mins	OSM	[105]
N1	MD	2 ha	5 mins	5 mins	OSM	[104]
B2	MD	10 ha	15 mins	15 mins	OSM	[105]
N2	MD	20 ha	25 mins	25 mins	OSM	[104]
B3	PP	0.5 ha	15 mins	6 mq^2	OSM	[105]
ESA	EXP	100 mq^2	5 mins	0.5 ha	ESA	[103]

TABLE 3.2: **Operational definition of indicators and targets proposed by institutional bodies.** The proposed indicators are inspired by targets set by local authorities and public health bodies worldwide. An indicator is an underlying metric measuring whether an area satisfies the corresponding target. For instance, the WHO indicator measures the walking distance in minutes from the closest PGA of at least 0.5 hectares, and an area satisfies the WHO target if the walking distance to the closest PGA of at least 0.5 hectares is no more than 5 minutes. Column *Type* identifies the family of the indicator: MD - Minimum distance, PP - Per-person, and EXP- Exposure. Column *Size* refers to the minimum size of greenspaces extracted for the computation of the indicator. Column *Data* refers to the data source used to extract greenspaces. It should be noted that the ESA indicator is the only indicator using green data from the WC-ESA.

Finally, we explore the impact of parametrizations on the level of inequality within a city as measured through a weighted Gini indicator. Interestingly, we do not observe any typical pattern across cities. While some cities experience decreasing levels of inequality (e.g., Singapore (SGP)) as we increase the minimum size of the greenspaces, others show increasing levels (New York (USA)) or U-shaped patterns (Sydney (AUS) and London (UK)). While this is likely to result from the interplay between the size composition of greenspaces within a city and its spatial distribution, we do not assess the existence of specific regularities. Similar figures for the stability of the exposure and per-person indicators against the time budget (for both) and the minimum size of the greenspaces (for the latter only) are provided in Appendix B (Figures B.2-B.3), the sensitivity analysis to the y parameter of the targeting strategies for all indicators (Figures B.6-B.7). For the exposure indicator, we observe a median *Rank correlation* for cities in our sample of 0.69 [IQR: 0.63-0.74] for a shift in the time budget from 5 to 10 minutes and median stability levels of the *2% targeting strategy* of 0.38 [IQR: 0.23-0.63] for a similar change in the time budget. See Data availability statement for the release of detailed stability metrics across our entire sample.

Stability of institutional targets

In the previous section, we examined the stability of the green accessibility indicators across various parameterizations. This section focuses on the accessibility pictures emerging from institutional targets proposed by global public health organizations and local authorities. Each target can be traced back to a specific family of indicators, once a specific parametriation is set. The investigation consisted of two steps. First, we evaluated the performance of a city against the targets. As discussing the results for over 1,000 cities would not be feasible, here we present information on the median level of cities by continent. Detailed information for each city is available in

the web interface. Then, similar to the previous section, we analysed how the different institutional targets differ in terms of identifying under-served sub-populations. Once again, the web interface can be used to explore the results of each city. A comprehensive operational definition of each institutional target and the underlying indicator is provided in Table 3.2. While the indicator provides a detailed accessibility metric (such as the walking distance in minutes to the closest greenspace with specific characteristics), a target binarizes the indicator by distinguishing between values that meet a specified criterion and those that do not. For instance, to meet the WHO green accessibility target, an area must have access to a greenspace of at least 0.5 hectares within 300 meters, or a 5-minute walk.

The existence of well-defined green targets naturally induces a metric to measure the performance of cities in terms of green accessibility, i.e., measuring the proportion of the inhabitants of a city that satisfies the prescribed target. By coupling information on the green indicators and the population density, we estimated the proportion of the population meeting each target in each city. Figure 3.8a provides an overview of the performance of cities in our sample for each target, categorized by geographical area. Consistent with findings in other studies [128, 149], we observe a distinct geographical pattern across most indicators. Cities in Europe and Australia-Oceania generally outperform cities in the Global South and North America, particularly concerning minimum distance indexes and the per-person metric. While this pattern is less pronounced for the exposure metric, cities in Asia and Africa still exhibit more significant variation around the median than other geographical areas. Regardless of geographical location, a larger proportion of the urban population typically meets the exposure target than other targets. This disparity is more pronounced for cities in the Global South, where available green spaces are less likely to be organized in structured public areas, and for North American cities, which are often characterized by extensive suburbs with predominantly single-family homes featuring private gardens but fewer public spaces. The proliferation of green indicators reflects a recent surge in interest from local authorities and public health bodies in promoting greener urban environments. It also acknowledges the diverse array of benefits associated with exposure to nature. However, this proliferation reveals the existence of concurrent authorities, often operating at the same level, each establishing independent goals. In Figure 3.8b, we evaluated the interchangeability of the accessibility perspectives derived from these indicators. Similarly to the previous section, we quantified the degree of disagreement between any pair of two indicators for each city through the stability of the population not satisfying the target (henceforth: *targeted population*) proposed by the institutional body. We observe greater stability for indicators within the same class than across indicators belonging to different classes. The lowest stability is observed between the WHO indicator and the ESA, reflecting differences in the types of green features incorporated in these indicators, particularly the inclusion of elements beyond parks, grasslands, and forests in the latter. When comparing the three short-distance minimum-distance (MD) indicators, we note a smaller overlap in the *targeted population* between WHO and B1 compared to the overlap between WHO and N1. The distinction arises from variations in the time threshold for the former and the minimum size for the latter, suggesting that the time dimension has a more significant impact than the size of the greenspace in short-MD evaluation on the definition of the city performance and the targeted population.

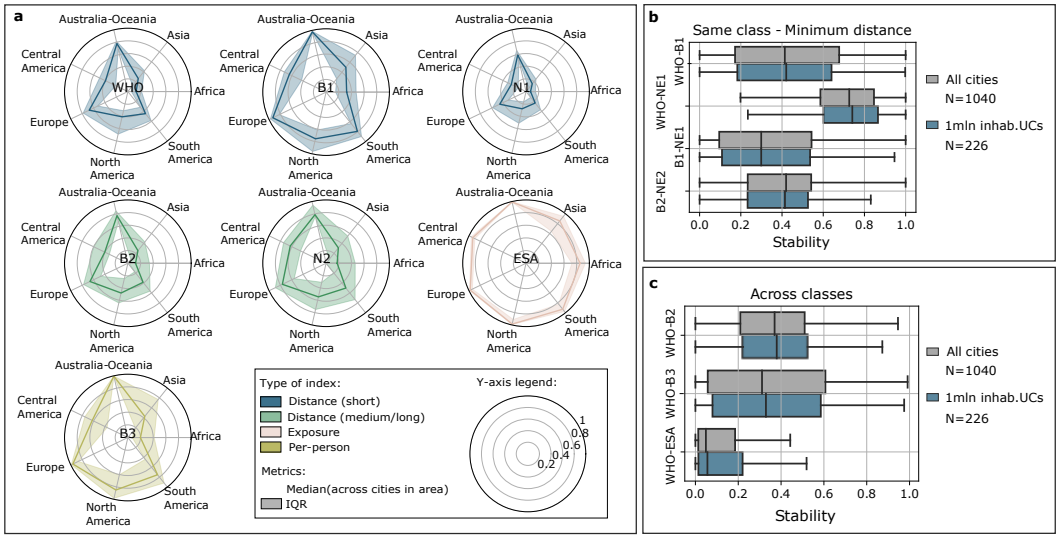

FIGURE 3.8: **Institutional green accessibility targets: performance and stability** a) For each institutional target (see Table 3.2), the radial plot depicts the median (solid line) and IQR (shaded area) across cities in each geographical macro-area of the proportion of the population satisfying the target. The colour of the plot reflects the class of the target. Cities in the Russian Federation have been attributed to Europe. b-c) The box plots depict the cross-indicators' level of stability in the population non-satisfying with the corresponding institutional targets for all cities in the sample (black) and for cities with more than 1 million inhabitants (red).

3.7 The interactive web interface

To facilitate the use of our multi-indicator framework by policymakers, we built an interactive web platform with five functionalities: *EXPLORE*, *MEASURE*, *COMPARE*, *CREATE* and *DRAW*. The back-end of the interface was developed using PostGIS, the front-end is developed using the ReactJS framework.

After selecting the urban centre of interest, the platform allows the user to:

1. *EXPLORE* public green areas identifiable in OSM. The user can filter the greenspaces based on detailed OSM tags, the size of the area (in hectares), and the name (as reported in OSM). This functionality is mostly meant to provide an overview of the green features included in the minimum-distance and per-person indicators. The interface does not provide the same functionality for the data from the WC-ESA since a similar tool is already available on the ESA website.

2. *MEASURE* several pre-computed green accessibility indicators at 9-arc sec geographical granularity. Policy recommendations by public health authorities and local governments inspire the indicators proposed. The map depicts the spatial variation of each indicator and the corresponding target. Summary metrics on the overall performance of the city in absolute terms and relative to the other cities are also provided.

3. *COMPARE* the performance of each cell across any two indicators, selected among the pre-computed metrics at the point before. The comparison is provided by splitting the distribution of each indicator into four groups based on

the distribution of the metric weighted by the population distribution and assigning each cell to the corresponding group. An example of this functionality for the city of Rome, Italy, is provided in Figure 3.9.

4. *CREATE* their indicators by setting each parameter to the desired level. The user can select the family of the indicator, the type of green (only green features from OSM can be selected for performance reasons), the minimum size of the green feature (in hectares), and, whenever applicable, the time budget (in minutes). By setting the desired level of the green accessibility (target), the user can then visualise areas satisfying the target and areas that are missing out.; and

5. *DRAW* a new green space and evaluate the impact on nearby areas in terms of enhanced green accessibility. After generating an initial indicator by selecting all the required parameters, the user can click on a specific location to simulate the impact of adding a greenspace in that area on the indicator itself. A black line in the resulting map identifies the area of the city affected by the intervention (i.e., areas where the metric value has improved), and information regarding the magnitude of the change in the indicator for the affected locations is displayed. It should be noted that the current version of the functionality assumes the additional greenspace does not extend beyond the selected cell; thus, the reported metrics are accurate only for relatively small interventions. An example of this functionality for the city of Rome, Italy, is provided in Figure 3.9

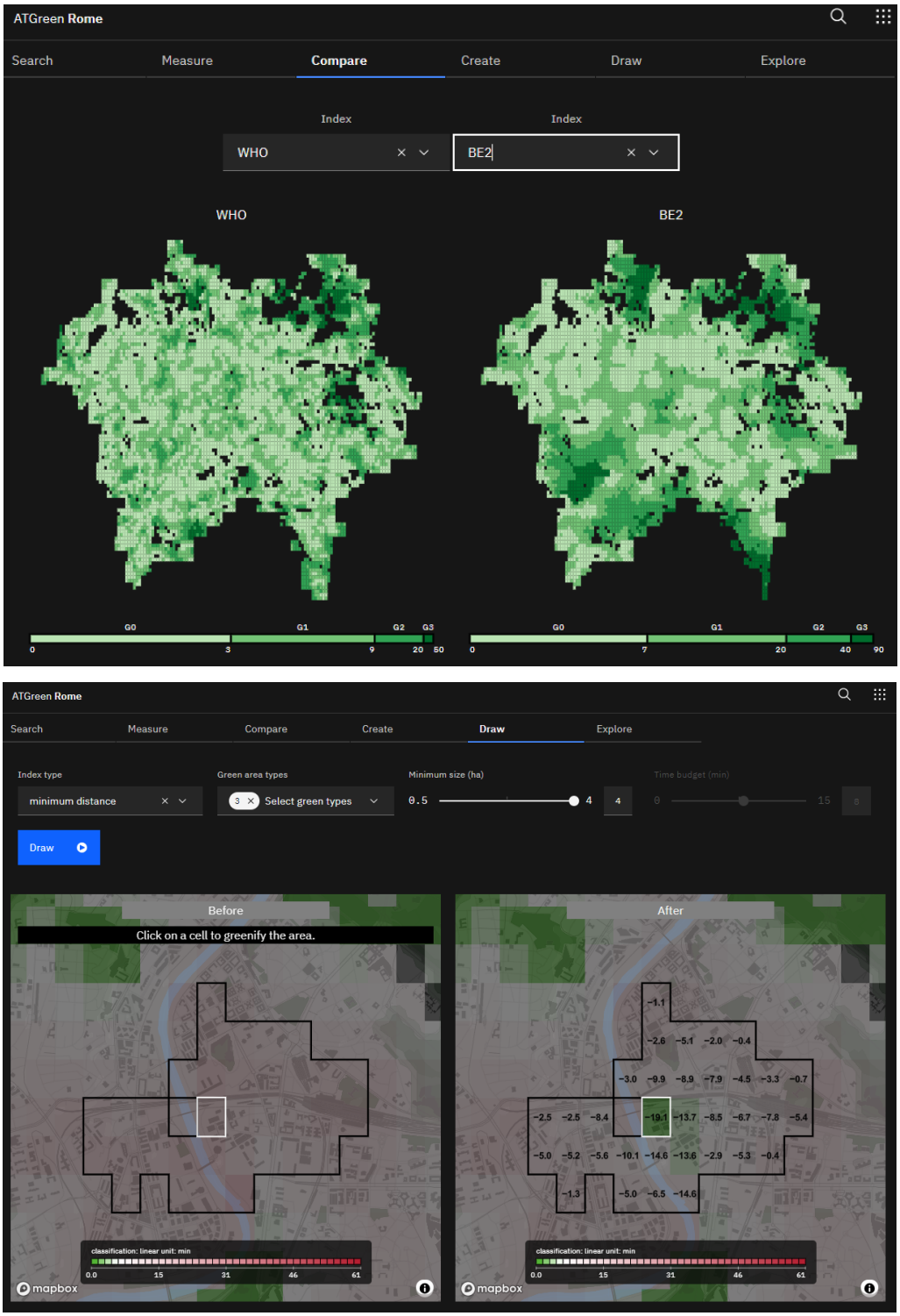

FIGURE 3.9: **ATGreen Front End: The web interface. An example.** The picture displays the front end of ATGreen. The top figure is a screenshot for the tab *COMPARE* (for the city of Rome, Italy). The bottom figure is a screenshot for the tab *DRAW* (for the city of Rome, Italy).

3.8 Discussion

Implications of the study

While recent studies on accessibility to urban green often compare cities or specific areas using a single indicator, our study emphasises the importance of recognizing the inherently multi-dimensional nature of green accessibility in urban environments. After introducing a framework to organically evaluate three families of structural green accessibility indicators, the framework was used to examine the similarity of accessibility outcomes arising from indicators within the same family but under different parameterizations and from institutional targets.

The findings in the first part of our study indicate significant instability in the ranking of both areas and populations, following perturbations in the parameterization of selected indicators. From a policy perspective, this suggests that relying on a single set of parameters may provide insufficient discrimination across areas and population subgroups with little green accessibility. This emphasizes the need to evaluate the impact of the adopted parametrization (or thresholding) when assessing the relative performance of areas/population subgroups in a city with respect to a specific structural green accessibility metric. In addition, the analysis showed that consistently under-performing areas are typically less densely populated than so-called *conflicting areas*, entailing an additional challenge from a policy design perspective, as these stable areas may not be sufficiently populated to be meaningful targets for intervention. It is important to stress that the simplistic ranking-based prioritization strategies assessed in the initial phase of this study do not seek to reproduce a realistic policy-design process. Numerous real-world factors, such as environmental or financial constraints, could impede the feasibility of greening interventions in severely under-performing areas. Additionally, there might be valid reasons to prioritise specific demographic groups, such as older adults or children, who may face more restricted mobility within the city. Moreover, realistic policy design processes in urban environments must consider other complex factors, including green gentrification phenomena, financial and built-environment constraints but also the implementation of co-creation processes to foster citizens' engagement [150, 151, 152, 153, 154, 155]. Given the broad geographical scope of the study, these elements could be factored in. Instead, we focused on the identification of subgroups of the population that are consistently (e.g., across several indicators or perturbation of the same indicator) under-performing. While this approach alone is insufficient to describe a realistic policy design approach, it is considered here instrumental for the design of interventions aimed at effectively promoting the reduction of urban inequalities [144].

The second part of the study focused on specific institutional targets, aiming to evaluate the interchangeability of the induced accessibility pictures. Similar to the preceding analysis, the interchangeability between any two targets was assessed by examining the size of the overlap in populations that do not satisfy these targets, e.g., the degree of overlap between populations that might be targeted for intervention using one or the alternative target. Unsurprisingly, the findings suggest a limited degree of interchangeability for indicators from different families (e.g., minimum distance vs. exposure) or those aiming to capture different forms of green accessibility (e.g., short distance to a small greenspace vs. longer distance to a larger greenspace). This confirms the need to evaluate each form of green accessibility independently to provide a comprehensive picture, in line with the multi-target recommendations

of specific institutional bodies [104, 105]. More interestingly, substantial discrepancies also emerge among targets seeking to capture similar forms of accessibility (e.g., targets WHO, B1, and N1, all short-distance indicators). This latter observation reinforces the results from the first part of the study on the impact of a fixed parameterization, here adopting more realistic targeting approaches designed by specific institutional bodies.

While single-indicator approaches have been shown to perform well in studying the impact of green accessibility and exposure on outcomes such as human health and wellbeing (as in [119]), this study suggests that transitioning toward a multi-indicator framework would provide a more nuanced picture of green accessibility. A multi-dimensional framework serves a dual purpose: 1. assessing the impact of a fixed-threshold approach induced by a specific parameterization of an indicator, 2. organically evaluating multiple forms of green accessibility.

Limitations of the study and future research

Despite the effort devoted to the cleaning and processing of the data to ensure the best possible standards, the main limitation of this analysis concerns the completeness of the mapping of green features in OSM. To limit the impact of this potential data bias, we undertook a series of filtering and checks to assess the quality of the OSM data in each urban centre, resulting in more than halving the initial sample of cities (from around 2,500 to 1,040). However, the accessibility metrics that we measured intrinsically depend on the quality of these data, so low green feature quality would necessarily result in biased indicators. To promote transparency and facilitate the identification of these biases, we made all our data (including the raw data) easily navigable to the public with our interactive platform. However, we deem the impact of this issue on the stability metrics performed in this study to be low as the analysis is designed to compare across areas within the same city (where the quality of the mapping is expected to be consistent) rather than across cities.

Another limitation of the study concerns the definition of greenspaces for the minimum distance and the per-person indicators. Our definition of urban green relies entirely on the tag of areas in OSM. In the presence of heterogeneous standards for the mapping of greenspaces among OSM mappers, the characteristics of areas classified as greenspaces (in terms of the level of green, type of services, and characteristics of the vegetation) might vary largely across countries or the climate zone of the urban centre. Once again, the interactive platform provides an initial attempt to control for this variability by allowing the user to customise the selection of green to incorporate in the indicator. To futher address this limitation, one potential avenue for future research involves expanding the framework by incorporating new data layers. These additional layers could automatically enrich the characterization of urban green areas with policy-relevant features, such as the availability of services and facilities, biodiversity levels, and other environmental quality or safety metrics. While collaboration with local authorities can inform this feature-augmentation process, extending it to large-scale evaluation poses notable challenges. An associated research direction evaluates the various data sources available for extracting green features, considering their completeness and suitability for constructing green accessibility indicators. Although our initial attempt in this direction is outlined in the Materials and Methods of this study, our evaluation is currently limited to two data sources. Future research efforts are necessary to assess the diverse sources available systematically.

Chapter 4

ATGreenGO: a routing engine recommending nature-enriching routes

4.1 Overview of the chapter

This chapter is based on on the outputs and technical report produced for the EU Horizon 2020 GoGreenRoutes project (Grant ID: 869764).

Continuing the broader discussion on enhancing green accessibility initiated in Chapter 3, this chapter introduces a shift in focus from the structural aspects of spatial relationships between urban residents' home locations and nearby greenspaces to strategies aimed at influencing individual behaviour to increase exposure to nature. To this end, I propose a novel routing engine for walking transits, named ATGreenGO, designed to maximise exposure to natural environments while minimising the level of detour compared to the shortest route, thereby providing routes that can be easily integrated into a daily routine. Although the routing engine is currently deployed for a small set of cities, its design relies solely on publicly available resources, making it replicable for any other urban area.

The remainder of the chapter is structured into seven sections. Section 4.2 outlines the motivations behind the development of the routing engine. Section 4.3 discusses academic research on the development of routing engines that adopt non-time or distance-based definitions of optimality. Section 4.4 provides an overall discussion of the materials and methods for the routing engine, including the characterisation of the street network in terms of the exposure to nature that each street guarantees. Section 4.5 and Section 4.6 describes the system in greater detail and provides an overview of the associated web interface. The performance of ATGreenGO is then evaluated in Section 4.7. Finally, Section 4.8 discusses the implications and potential impact of the developed routing engine.

4.2 Scope

In the previous chapter, I discussed the health implications of exposure and access to nature (see Section 3.2), which underpin many greening interventions implemented by local authorities. I also reviewed the methodologies used to evaluate green accessibility, with a specific focus on the widely used *structural metrics*. These metrics

assess the relationship between the green infrastructure of a city and the distribution of its residents, concentrating predominantly on their residential locations.

However, this home-centric approach has notable limitations, as it overlooks the potential for and actual nature exposure during various daily activities. People engage with green infrastructure not only around their residential areas but also in other locations throughout their day. Moreover, some exposure to nature may occur incidentally during activities such as commuting, running errands, social visits, or recreational outings. Recognising these limitations, recent research has shifted towards employing behavioural-informed indicators. These methodologies make use of surveys, automated user-generated geographic data, and public participation geographic information systems (PPGIS) to provide a more nuanced and accurate assessment of actual exposure levels to nature [129, 130, 131, 132].

This chapter builds upon the discussion of green accessibility introduced in Chapter 3 by shifting both the focus and the nature of the contribution. While the previous chapter was limited by its structural perspective, which predominantly centred on residential locations, this chapter acknowledges that interactions that people experience with nature extend far beyond their home environments, focusing specifically on the potential for incidental exposure to nature that might happen during daily routines.

In terms of nature of the contribution, this chapter departs from the analytical type of contribution proposed in Chapter 2 – that consistend in a purely analytical study– and Chapter 3 – where the development of a computational framework was instrumental to the analytical study. Instead, it focuses on the algorithmic and programming development of a tool designed to enhance exposure to nature: a routing engine named ATGreenGO. This tool aims to increase the exposure of users to green spaces during everyday activities, such as commuting or leisure walks, by recommending routes that offer richer nature experiences. By suggesting alternatives to the conventional shortest route, ATGreenGO helps users discover more enjoyable and nature-enriching paths that they might otherwise overlook.

The routing algorithm embedded in ATGreenGO is designed to balance the maximisation of nature exposure with the minimisation of detours compared to the shortest route. This balance ensures that the suggested routes are practical alternatives that can be seamlessly integrated into daily routines without significantly altering the time required for each activity. The validation process confirms that ATGreenGO effectively achieves this balance, offering substantial improvements in nature exposure while only slightly increasing travel time. Moreover, the results indicate that the system performs consistently well across various cities and natural environments, demonstrating the robustness and reliability of the recommendations provided by ATGreenGO. This development represents a shift from purely analytical work to a practical tool capable of positively influencing daily walking behaviours and enhancing overall well-being through increased interaction with nature.

4.3 Related work

Leveraging geodata, a routing engine is software that provides tools for calculating various types of information related to possible routes within a geographical area. The most common functionality involves calculating one route between a starting location and a destination based on specific optimisation rules. Additional features

typically include the computation of isochrones, i.e. a line of equal travel time from a starting location, and the ability to compute matrices of distances between multiple origins and destinations within the same query.

In terms of underlying optimisation rules, traditional routing engines primarily prioritise efficiency metrics such as shortest distance or fastest time. Early research on route recommendation has focused on developing algorithms to optimise efficiency in static complex contexts, such as the presence of multiple users and associated congestion issues or multi-modality transportation modes. For instance, Chang et al. [156] developed a backtracking algorithm to recommend car routes that deviate from popular paths, providing variety among different users and reducing the likelihood of congestion without relying on real-time data. In the context of multi-modality, Ludwig et al. [157] used an adaptive A^*-like algorithm to suggest public transport routes that minimise walking and waiting times. More recently, routing engines that focus on efficiency metrics have been extended to incorporate dynamic factors such as real-time information on congestion [158] and delays in public transportation [159]. In this respect, one of the most widely used services is Google Maps [160].

Optimisation based on time or distance metrics may, however, be inadequate in specific contexts, particularly for non-motorized trips. A number of studies have found that street-specific features can promote active mobility behaviours, such as cycling and walking. Along this line, in Chapter 2, we discussed the preference of women for safe routes for cycling, including those with dedicated cycling infrastructure. Regarding walking preferences, numerous studies have found that the quality of the pedestrian infrastructure (e.g., the availability and width of sidewalks) is crucial in promoting walking [161, 162, 163, 164]. Other factors that effectively encourage walking include measures to reduce noise and air pollution from road vehicles [165], as well as the aesthetic and liveability features of streets [164]. These findings suggest that time- or distance-optimising routing systems might fail to identify the preferred route, especially in contexts where an array of alternative routes with similar transit times but significantly different features exist.

Acknowledging this, there has been growing interest in designing algorithms that recommend optimal urban routes based on alternative definitions of optimality, by enriching the street-network information with data on specific qualitative features. Along this line, early attempts to quantify the pleasantness of streets used indirect information based on real or online activity observed in a specific location. For example, Flickr data have been used to identify popular locations based on spatiotemporal sequences of photo uploads [166, 167]. GPS traces from mobile phones have also been used to identify interesting locations in both urban and suburban areas [21]. Quercia et al. [20] adopted a different approach, using data from online experiments to compare the pleasantness of streets in London and design a routing system that recommends emotionally pleasant routes. More recent studies have proposed routing systems that prioritise a diversity of scenarios based on Google Street View images and FourSquare data [168] or environmental quality scores, such as air quality or temperature [22]. As data from mobile social networking become more available, routing systems that are able to recommend routes based on observed user-specific preferences have also been proposed [169, 170].

This research contributes to the existing literature by introducing a novel routing

system for pedestrians that prioritises nature-rich routes while maintaining a minimal detour from the shortest path. The objective of this system is to offer users alternative routes that, with only a minor deviation from the shortest path, can be seamlessly integrated into their daily routines. This approach effectively enhances users' exposure to natural environments, promoting well-being without significantly impacting travel time.

4.4 Materials and Methods

Data extraction, processing and cleansing

To ensure its scalability to any geographical area of interest, the proposed routing system relies exclusively on publicly available data sources. Based on their role, these data sources can be classified into three main categories:

- **data on the boundary of cities**. The system was initially developed in the context of the Horizon 2020 project GoGreen Routes. As such, the initial deployment of the system has been performed for the six *cultivating cities* covered by the project, namely: Burgas (Bulgaria), Lahti (Finland), Limerick (Ireland), Tallinn (Estonia), Umea (Sweden), and Versailles (France). For testing purposes, Turin (Italy) was added due to the familiarity of the research team with the city. Subsequently, the system has been extended to the city of Barcelona (Spain) and selected boroughs of London (UK). For these latter inclusions, the system can be queried through the public API, while the integration in the web interface is currently in progress.
- **data on the street network of each urban area.** For each urban area, the street network information is extracted from the OSM [3] dump of Europe, downloaded from the GeoFabrik Download Service in May 2022.
- **data on natural environments available in each urban area.** Information on natural features available in each urban centre is derived by combining data from OSM and from the WC-ESA [6].

Boundaries of the urban centres For each urban centre, Table 4.1 reports the source of information on its boundary. Following the approach used for other research studies conducted in the context of the 2020 Horizon project GoGreenRoutes, the boundary of the six cultivating cities was extracted from the Urban Atlas of the initiative Copernicus of the European Environmental Agency [171]. For the other three cities, administrative boundaries have been used. To mitigate potential routing bias in areas near the boundary of a city, a geographical buffer of a 1500-meter radius was added to each boundary.

OpenStreetMap data ATGreenGO relies on the use of data from OSM for the extraction of street-network data and for its enrichment with green and blue attributes. Data from OSM were processed in three steps:

1. Extraction of the local osm.pbf file. For each city, the first step consisted of the extraction of a local osm.pbf dumps from the OSM dump of Europe, downloaded from the GeoFabrik Download Service on May 2022. The process followed the step described in Section 3.4.

City	Country	Definition	Source
Burgas	Bulgaria	Urban Atlas	Link
Lahti	Finland	Urban Atlas	Link
Limerick	Ireland	Urban Atlas	Link
Tallinn	Estonia	Urban Atlas	Link
Umea	Sweden	Urban Atlas	Link
Versailles	France	Urban Atlas	Link
Barcelona	Spain	Administrative boundary	Link
London	United Kingdom	Selection of boroughs	Link
Turin	Italy	Administrative boundary	Link

TABLE 4.1: Sources of information for the geographical data on city boundaries.

Type	Category	OSM key	OSM value
Green	Parks	leisure	park
		leisure	garden
		landuse	recreation_ground
		landuse	village_green
	Other green	landuse	forest
		natural	wood
		landuse	grass
		landuse	meadows
		natural	grassland
		natural	shrubbery

TABLE 4.2: OSM *key* : *value* pairs used for the identification of green features

2. Extraction of the street network. The second step consisted of the extraction of all streets for the local dump. To this scope, the Python bindings of the osmium tools were modified to query the osm.pbf file based on a customised list of OSM tags. The following OSM tags were queried: `highway`, `cycleway`, `footway`, `sidewalk`, `busway`.

3. Extraction of green features. The third step consisted of the extraction of green natural environments from the local osm.pbf file. The process followed the steps describe in 3.4 and the list of key-value pairs extracted is provided in Table 4.2

European Space Agency - World Coverage 2020 Green features from OSM have been integrated with information on land cover from the WC-ESA 2020 [6]. For each urban centre, the data were extracted following the approach described in Section 3.4. Classes `10 - Tree cover`, and `30 - Grassland` were defined as green features. Class `80 - water body` was used for the identification of blue features.

4.5 Enrichment of the street network with green and blue qualities

A key step in defining the routing engine involved enriching the street network with attributes that describe the exposure to green and blue natural environments attainable by walking on each street. We call these attributes the *green and blue qualities* of the street network of a city. The implemented system includes five green and blue

qualities and users can specify their preferred quality when querying it through the public API, thus obtaining a route that maximises exposure to the selected attribute.

To achieve an accurate characterisation of the green and blue qualities of the street network of each city, a two-step pipeline was adopted, described in detail below.

From OSM streets to micro-segments To provide a high-resolution geographical characterisation of the green and blue qualities of the street network, each street was split into segments of 20 meters in length, starting from the first node in the line string defining the street. Here, a street is defined as a unique combination of its `osm_id` and one of the extracted tags (see Section 4.4). Depending on the mapping of the raw OSM data, this means that a street in our dataset can correspond to either a toponymically-defined street or a single lane of a multi-lane toponymically-defined street. Generally, it does not correspond to a link in the street network. Since the length of a street needs not to be a perfect multiple of 20, the last segment in each street may be shorter than 20 meters. These resulting segments were then assigned the same `osm_id` and tags of the parent street, as well as a distinctive `pseudo_id`.

Characterization of micro-segments in terms of green and blue qualities For each segment, we then generated a set of attributes to define its green and blue qualities. To this scope, we used an iterative approach, testing various specifications of these attributes. These ranged from complex attributes detailing the exact geometrical relationship between the segment and nearby natural elements to simpler attributes that better align with users' preferences. Table 4.3 offers an overview of all the features tested, highlighting those ultimately implemented in the final version of the routing system.

In the first instance, we tested a set of detailed spatially aware features designed to capture the complex geometrical relationship between each micro-segment and the nearby green environments, while also distinguishing high-level characteristics of these green elements (public park vs other forms of aerial green, excluding tree coverage, vs tree coverage). For each segment, this initial set included nine features:

- `tree_coverage`: the proportion of the segment with tree coverage. The metric was constructed in two steps. First, we overlaid the tree cover polygons extracted from the WC-ESA upon the segment (buffered with a 1-meter radius). We then computed the size of the overlapping area and divided it by the overall area of the buffered segment.
- `inside_parks` and `inside_other_green`: an indicator for whether the segment falls within a park/another type of green. The segment was considered to fall within a park/another type of green if it intersects a green polygon for no less than 40% of its length.
- `inbetween_parks` and `inbetween_other_green`: an indicator for whether the segment falls within a park/another type of green (for less than 40% of its length). The segment was `inbetween` green areas if 1. it had a positive intersection with a green polygon; 2. the intersection was less than 40% of the total length of the segment.
- `surrounded_parks` and `surrounded_other_green`: an indicator for whether the segment runs parallel to a park/another type of green on both sides but without being within the green area. The segment was `surrounded` by green

areas if 1. it was not `inside` or `inbetween` a park/another type of green; 2. there was a park/another type of green within 15 meters of the streets on both sides of it. The latter condition was checked by constructing the right and left buffers of the geometry of the segment and intersecting it with the set of green polygons.

- `alongside_parks` and `alongside_other_green`: an indicator for whether the segment runs parallel to a park/another type of green on one side only. The segment was `alongside` green areas if 1. it was not `inside` or `inbetween` a park/another type of green; 2. there was a park/another type of green within 15 meters of the streets on one side only. As for the previous attributes, the latter condition was checked by constructing the right and left buffers of the geometry of the segment and intersecting it with the set of green polygons. For these two attributes, only one of the two buffers was required to intersect a green area.

The previous characterisation enabled us to inspect the complex geometrical relationship of each segment with nearby natural elements. However, it was then recognised that this level of complexity was unnecessary from the perspective of a potential user. The previous features were therefore consolidated into five new attributes. Instead of providing a comprehensive taxonomy of the geometric relationship between the micro-segment and the natural environment, these new attributes were designed to encode user-specific preferences for walking in green and blue environments. For instance, recognising that streets within a park may offer higher exposure to greenery but might be perceived as less safe (especially at night), we retained the distinction between on-street and off-street green attributes, while dropping others. The resulting five green and blue qualities available in the system are:

- `green (off)`: it measures the proportion of the segment that falls within a park or other green feature. The feature is measured by computing the proportion of the segment intersecting OSM green polygons.
- `green (on)`: it measures the proportion of the segment that –not falling directly within a park of another type of green – is adjacent to some green environments. It is measured by: 1) applying a 15 meters buffer around the portion of the segment not falling directly within a green environment; 2) measuring the size of the intersection between the area of this polygon and any green polygon in the city (hereafter: intersecting green area), 3) computing the ratio between the intersecting green area and the size of the buffered polygon. To attribute the value to the overall segment, the computed measure is then normalized by multiplying it by (1-`green (off)`). For this metric, the green information used is not limited to green features from OSM but includes the tree coverage and grassland areas from the WC-ESA.
- `blue`: the proportion of the segment that is within 50 meters of large water bodies (more than $500m^2$) or 15 meters of small (less than $500m^2$) water bodies. Practically, this is constructed by generating a buffer of the desired radius around the blue feature and measuring, for each segment, the proportion of its geometry that intersects the buffered blue area.
- `any green`: obtained as the sum of `green (off)` and `green (on)`. By construction, `any green` is bounded between 0 and 1.

Category	Name	Data type	Data range	Data source	Status
green	`tree_coverage`	Continuous	[0,1]	[6]	Tested
green	`inside_parks`	Binary	{0, 1}	[3]	Tested
green	`inside_other_green`	Binary	{0, 1}	[3]	Tested
green	`inbetween_parks`	Binary	{0, 1}	[3]	Tested
green	`inbetween_other_green`	Binary	{0, 1}	[3]	Tested
green	`surrounded_parks`	Binary	{0, 1}	[3]	Tested
green	`surrounded_other_green`	Binary	{0, 1}	[3]	Tested
green	`alongside_parks`	Binary	{0, 1}	[3]	Tested
green	`alongside_other_green`	Binary	{0, 1}	[3]	Tested
green	`green (off)`	Continuous	[0,1]	[3]	Implemented
green	`green (on)`	Continuous	[0,1]	[3, 6]	Implemented
green	`any green`	Continuous	[0,1]	[3, 6]	Implemented
blue	`blue`	Continuous	[0,1]	[6]	Implemented
green & blue	`any nature`	Continuous	[0,1]	[3, 6]	Implemented

TABLE 4.3: ATGreenGO: List of green and blue qualities

- `any nature`: obtained as the sum of `green (off)`, `green (on)` and `blue`, normalized to the unit interval. By construction, `any nature` is less or equal to each of its components. This feature was included to allow the user to select routes that are rich in both blue and green elements simultaneously. However, unlike each of its sub-components, the value of this feature cannot be interpreted as the proportion of the street length with exposure to a natural environment and should not be used to evaluate the overall exposure to nature attainable walking along a specific route (which instead should be measured for each natural environment separately).

4.6 The system

Overview of OpenTripPlanner (OTP)

OpenTripPlanner (OTP) [5] is an open-source, multimodal trip planning system designed to facilitate efficient and customizable route planning for various modes of transportation, including walking, cycling, and public transit. Based on Java, the flexibility and adaptability of OTP made it an optimal starting framework for the development of ATGreenGO. The key characteristics of OTP are:

1. **Graph Representation of the street network** - The transportation network is represented as a graph, where nodes represent geographical intersections among streets, and edges represent the possible connections between these intersections (typically portions of roads, paths, or transit routes). Each edge in the graph is associated with attributes such as distance, travel time, and transportation mode-specific information.

2. **Data Import and Processing** - OTP typically imports data from OpenStreetMap, a community-driven mapping platform. This data includes detailed information about road networks, pedestrian paths, transit stops, and other relevant geographic features. For transit routing[1], OTP incorporates schedule data, including information about routes, stops, and timetables.

3. **Routing Algorithm** - The routing problem can be thought of as a cost function problem, where the user is interested in identifying the route between

[1]This feature is not implemented in ATGreenGO. The prototype is indeed currently implemented for walking routes only.

two points that is associated with the lower cost. In a standard definition, the cost represents the transit time associated with the selected route, namely the shortest path. Depending on the exact configuration, the OTP can use two alternative shortest-path algorithms: Dijkstra's algorithm [172] or A^* algorithm [173], the latter incorporating heuristic information to guide the search efficiently. For the extension to ATGreenGO, the A^* algorithm was adopted with a modified cost function to incorporate the green and blue features of the street network (see Section 4.6).

From OTP to ATGreenGO

Aggregation of micro-segments to OpenTripPlanner (OPT) graph arcs To enable the routing, OTP arcs are constructed based on road intersections, so that each arc is the proportion of a street falling within two intersections. The inclusion of green and blue qualities within the routing requires projecting the green and blue qualities of the segments (see Section 4.5) to the arcs of the graph generated by OTP. This was performed by remapping the segments to the arcs based on the street identifier (`osm_id`) and the geographical overlap between the two, according to the following pipeline:

I For each arc in the graph, select all the micro-segments with the same `osm_id`.

II Among these arcs, select the ones that have a total or partial overlap in the geometry.

III For each arc:

- For each continuous feature, perform the weighted average of the values of the feature in the intersected micro-segments. The weighting is given by the dimension, in percentage, of the intersection to the total length of the arc.
- For each binary feature, the associated feature is the maximum value of the feature measured on all the intersecting micro-segments.

The remapping of the features from the micro-segments to the OTP arcs was performed in JAVA. The spatial overlaps were computed with the `intersection` method of the class `Geometry` of Java Topology Suite (JTS).

Tables 4.4-4.12 provide an overview of the green and blue qualities and the length of the OTP arcs for each city. The statistics for arc lengths are given in meters.

With approximately 30,000 arcs, Limerick has the smallest street network among the cities, though its average arc length (about 63 meters) is greater than that of larger cities like Barcelona (36 meters), Tallinn (51 meters), selected boroughs of London (32 meters), and a similarly sized city like Versailles (54 meters).

Regarding green attributes, northern cities show the highest overall green quality. The average value of the `any green` feature is 0.63 for Lahti and 0.57 for Umeå, indicating that arcs in these cities' street networks provide exposure to green elements for 63% and 57% of their lengths, respectively. At the other end of the spectrum, Barcelona's street network has the lowest average `any green` attribute, with a value of 0.19. Despite being a coastal city, Barcelona also shows limited blue exposure, with an average `blue` feature value of 0.1, compared to 0.2 or 0.4 for the other cities.

This seemingly surprising result can be understood by considering that blue environments include both natural water bodies and large artificial bodies (e.g., large fountains in Versailles). Moreover, the presented values are computed on arcs of varying lengths and are influenced by the configuration of each street network. Since these values are not weighted by the length of the arc, they should not be interpreted as the overall average exposure for the entire street network.

	N. arcs	mean	std	min	median	max
length	31696	104.78	231.03	0.01	54.68	7785.87
any nature	31696	0.18	0.19	0.00	0.09	1.00
any green	31696	0.33	0.37	0.00	0.16	1.00
green (on)	31696	0.22	0.30	0.00	0.07	1.00
green (off)	31696	0.11	0.30	0.00	0.00	1.00
blue	31696	0.02	0.13	0.00	0.00	1.00

TABLE 4.4: Descriptive statistics of the ATGreenGO arcs of the city of Burgas

	N. arcs	mean	std	min	median	max
length	127989	82.53	157.95	0.02	38.22	9441.11
any nature	127989	0.32	0.21	0.00	0.39	1.00
any green	127989	0.63	0.39	0.00	0.77	1.00
green (on)	127989	0.40	0.38	0.00	0.30	1.00
green (off)	127989	0.23	0.40	0.00	0.00	1.00
blue	127989	0.02	0.12	0.00	0.00	1.00

TABLE 4.5: Descriptive statistics of the ATGreenGO arcs of the city of Lahti

	N. arcs	mean	std	min	median	max
length	30106	62.79	128.16	0.64	35.34	5199.16
any nature	30106	0.28	0.21	0.00	0.30	1.00
any green	30106	0.53	0.39	0.00	0.56	1.00
green (on)	30106	0.48	0.38	0.00	0.45	1.00
green (off)	30106	0.05	0.21	0.00	0.00	1.00
blue	30106	0.04	0.18	0.00	0.00	1.00

TABLE 4.6: Descriptive statistics of the ATGreenGO arcs of the city of Limerick

	N. arcs	mean	std	min	median	max
length	197465	50.79	81.60	0.01	26.70	4453.18
any nature	197465	0.25	0.20	0.00	0.23	1.00
any green	197465	0.48	0.38	0.00	0.44	1.00
green (on)	197465	0.38	0.36	0.00	0.29	1.00
green (off)	197465	0.10	0.28	0.00	0.00	1.00
blue	197465	0.02	0.12	0.00	0.00	1.00

TABLE 4.7: Descriptive statistics of the ATGreenGO arcs of the city of Tallinn

	N. arcs	mean	std	min	median	max
length	87675	158.21	364.82	0.04	52.97	14522.75
any nature	87675	0.31	0.23	0.00	0.35	1.00
any green	87675	0.57	0.41	0.00	0.65	1.00
green (on)	87675	0.35	0.38	0.00	0.17	1.00
green (off)	87675	0.22	0.40	0.00	0.00	1.00
blue	87675	0.04	0.18	0.00	0.00	1.00

TABLE 4.8: Descriptive statistics of the ATGreenGO arcs of the city of Umea

	N. arcs	mean	std	min	median	max
length	200003	54.31	88.38	0.08	26.87	5861.08
any nature	200003	0.26	0.21	0.00	0.25	1.00
any green	200003	0.51	0.40	0.00	0.49	1.00
green (on)	200003	0.33	0.35	0.00	0.19	1.00
green (off)	200003	0.18	0.37	0.00	0.00	1.00
blue	200003	0.02	0.12	0.00	0.00	1.00

TABLE 4.9: Descriptive statistics of the ATGreenGO arcs of the city of Versailles

	N. arcs	mean	std	min	median	max
length	412484	36.20	62.52	0.03	14.48	4478.61
any nature	412484	0.10	0.17	0.00	0.00	1.00
any green	412484	0.19	0.32	0.00	0.00	1.00
green (on)	412484	0.12	0.23	0.00	0.00	1.00
green (off)	412484	0.07	0.25	0.00	0.00	1.00
blue	412484	0.01	0.09	0.00	0.00	1.00

TABLE 4.10: Descriptive statistics of the ATGreenGO arcs of the city of Barcelona

	N. arcs	mean	std	min	median	max
length	772068	32.46	47.25	0.02	16.92	3752.15
any nature	772068	0.19	0.22	0.00	0.09	1.00
any green	772068	0.33	0.38	0.00	0.14	1.00
green (on)	772068	0.22	0.31	0.00	0.04	1.00
green (off)	772068	0.11	0.31	0.00	0.00	1.00
blue	772068	0.04	0.20	0.00	0.00	1.00

TABLE 4.11: Descriptive statistics of the ATGreenGO arcs of the city of London

	N. arcs	mean	std	min	median	max
length	111667	46.21	63.73	0.07	22.28	1496.41
any nature	111667	0.16	0.21	0.00	0.04	1.00
any green	111667	0.28	0.36	0.00	0.07	1.00
green (on)	111667	0.16	0.26	0.00	0.01	1.00
green (off)	111667	0.11	0.30	0.00	0.00	1.00
blue	111667	0.04	0.18	0.00	0.00	1.00

TABLE 4.12: Descriptive statistics of the ATGreenGO arcs of the city of Turin

Definition of ATGreenGO cost function OTP uses the well-known weighted A* algorithm [173] to find the optimal path in the street network. In the pedestrian walking mode, OTP also accounts for the walkability of streets and obstacles such as stairs. For the development of ATGreenGO, we extended the cost function of the optimisation problem of OTP to incorporate the green and blue qualities of the street network, considering one quality at a time. The intuition behind our solution is that segments that are exposed to nature receive a reward, i.e., the cost of traversing them decreases relative to segments with a lower exposure to nature. This reward is proportional to the level of the exposure. The cost function is our main contribution on the routing side of the system.

More formally, let's start by recalling that the cost function of the stable OTP engine is defined as:

$$C_{OTP}(P) = \sum_{a \in P} \frac{L_a}{S_a} R_a \tag{4.1}$$

where $\mathtt{P}$ is a path between two distinct points, $\mathtt{a}$ indicates the OTP arcs composing the path $\mathtt{P}$, $\mathtt{L}_a$ is the length (in meters) of the arc $\mathtt{a}$, $\mathtt{S}_a$ is the assumed speed (in meters per seconds) on the arc $\mathtt{a}$, and R_a is a reluctance parameter used to penalize specific urban features that the user might find less appealing (e.g. stairs).

To integrate information about the presence of nature into the optimisation problem of our routing algorithm, we expanded the definition of the cost function to include these features. Before evaluating any modification to the cost function, it should be noted that, for the optimisation problem to be well defined, any suitable alternative cost function ($C_{new}(P)$) must satisfy the following three properties (PR):

Letting $\mathtt{G}_a$ be the value of the green or blue feature on the arc $\mathtt{a}$ and $\mathcal{A}_c$ the full set of arcs composing the street network of the urban center $\mathtt{c}$,

PR1 If $G_a \geq 0$ and $G_a = G \; \forall a \in \mathcal{A}$, then $C_{new}(P)$=$K(G) \times C_{OTP}(P)$ for any P, where $K(G) > 0$. That is, if all arcs grant exposure to an identical amount of green/blue (or even no exposure at all, setting $G = 0$), then the new cost function should behave as the original OTP cost function (up to a positive proportional factor, potentially dependent on the amount of green available)

PR2 $C_{new}(\cup^N a) > C_{new}(\cup^{N+1} a) \; \forall N$. That is, adding an arc to a set of arcs should always result in a positive increase in the total cost.

PR3 $C_{new}(P) > 0$ for any P. That is, all paths must have a strictly positive cost.

Among the set of cost functions satisfying the above properties, the implemented cost function for ATGreenGO is defined as:

$$C_{ATGreenGO}(P) = \sum_{a \in P} \frac{L_a}{S_a} [R_a \cdot max(1 - G_a, 0.001)] \tag{4.2}$$

Intuitively, this formula discounts the additional cost associated with an arc based on its green and blue quality –measured through one of the five green and blue attributes–, with a factor equal to $1 - G_a$. To ensure PR2 and PR3 are satisfied, the discounted factor was set not to fall below 0.001. Alternative specifications of the cost function have been evaluated and a full discussion on the comparative performance of the implemented formula compared to the alternative specifications is provided in 4.7.

Middleware Service

As the query and response formats of OTP can be complex, we implemented a lightweight middleware for ATGreenGO. The middleware was developed in Python using the Flask web framework. This middleware provides an easy-to-use API to query several routes with different optimisation goals (based on one of the five green and blue qualities). For example, one can specify which green qualities should be optimised, resulting in multiple calls to the OTP Backend. Furthermore, it provides aggregated statistics of the exposure to nature for each route, which is not available from the standard response of OTP. The middleware service is available for all cities covered by the project (including Barcelona and the selected boroughs of London) and can be queried at the following URL http://atgreengo.hpc4ai.unito.it:5005. The API parameters are available at the following link.

ATGreenGO Front-End: The web interface

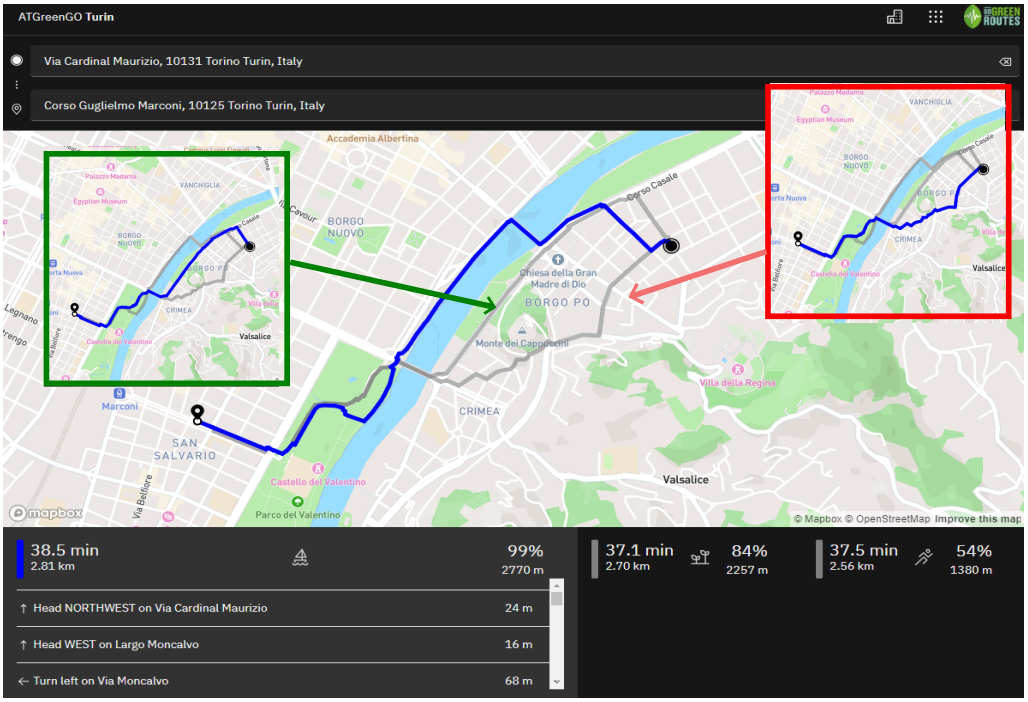

FIGURE 4.1: **ATGreenGO Front End: The web interface. An example.** The picture displays the front end of the routing system ATGreenGO. For the same OD pair in the city of Turin, the system provides the route maximising exposure to *blue* (main figure), *any green* (green insert) and the OTP shortest path (red insert).

To provide a tangible experience of the exposure to green routing, we additionally developed a web-based frontend, available at http://atgreengo.hpc4ai.unito.it/ (Figure 4.1). This allows residents of cities covered by the system to experiment by generating routes from arbitrary starting locations to destinations. To limit the information overload of a layman user and to align with the feedback received by the GoGreenRoutes' research team, the web app does not allow users to choose the underlying green and blue quality for optimisation. Instead, once an origin and destination pair is provided, the system directly generates the shortest path as well as the routes resulting from the optimisation of `any green` and `blue` qualities. The web

app then displays the routes on the map, provides detailed navigation information, and generates aggregate statistics on the (additional) green and blue exposure of the different routes in comparison to the shortest path.

4.7 Performance of ATGreenGO

ATGreenGO: performance and validation

The performance of the routing system ATGreenGO was evaluated by measuring the increase in exposure to green and blue natural environments that the recommended routes provide compared to the exposure experienced on the shortest path. Since the system aims to suggest alternative routes with only a limited level of detour, our assessment also considered the percentage increase in duration incurred by routes recommended by ATGreenGO compared to the standard distance-optimised OTP.

Definition of test set of Origin-Destination pairs For this assessment, we built a test set of Origin-Destination pairs (OD pairs) by overlaying a regular grid on the area of each city[2] and collecting all those pairs of nodes of the grid whose geodesic distance is no more than 2.5 km. The test set consisted of approximately 3,500 OD pairs across all cities. Summary information by city on the characteristics of the shortest path in terms of walking duration (top row) and green and blue qualities (bottom row) for the test set of OD pairs is provided in Figure 4.2.

[2]The boroughs of the city of London were integrated into the prototype at a later stage and are consequently excluded from this evaluation.

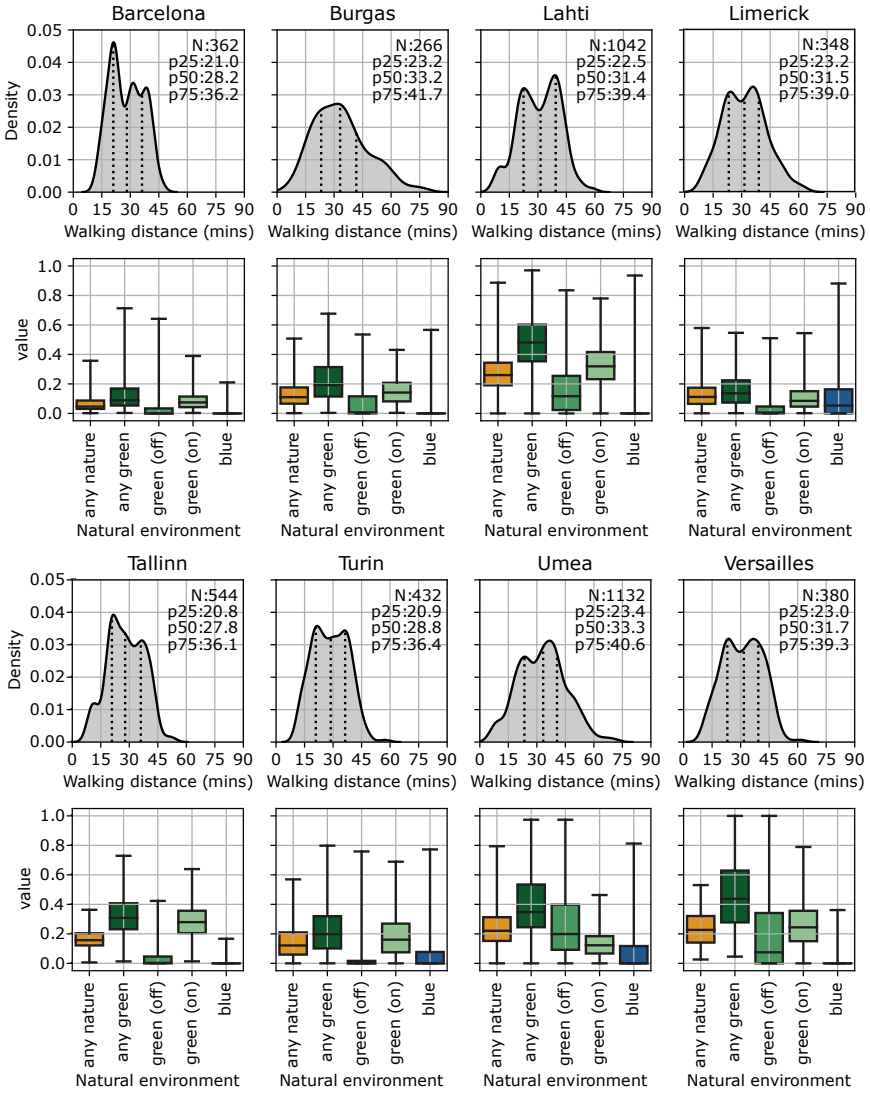

FIGURE 4.2: **Characteristics of the shortest path, for the test set of OD pairs, by city.** For each city, the panel displays summary information on the characteristics of the shortest path in terms of walking duration (top row - dotted line represents the 25th, 50th, and 75th percentiles of the distribution of walking distance) and green and blue qualities (bottom row) for the test set of OD pairs.

Evaluation of the performance of ATGreenGO on the test set To assess the performance of ATGreenGO, we analysed the absolute increment in exposure to green and blue natural features resulting from the ATGreenGO-recommended routes for the OD test set, relative to the conventional shortest path determined through the standard OTP. Figure 4.3 shows the distribution in the absolute change of exposure to each feature for the entire test sample of OD pairs. Information broken down by each city is provided in Figures C.1-C.5 in Appendix B. Considering the three green-related qualities, we observe a common behaviour:

1. The top one-fourth of routes experience a relevant increase of more than 0.21 for `any green` and `green (off)` and 0.16 for `green (on)`
2. By contrast, for approximately one-fourth of the routes, we observe almost no or very little increase in exposure.

3. By investigating the corresponding scatterplot, we notice that shortest paths with low exposures are on average associated with larger increases than shortest paths for which the exposure is already substantial.

For the feature `any nature`, it is worth noticing that the observed increase is typically much smaller than that for any green. This result is artificially low due to the normalisation of the attributes to the unit interval. Despite this, there is still an increase of approximately 0.08 for the top 25% of the origin-destination pairs. Considering the `blue` feature, we observe a skewed distribution with approximately zero impact for three-fourths of the OD pairs, a behaviour that is fully explained by the sparsity of blue natural environments within cities.

Finally, assessing the performance of ATGreenGO in terms of the total walking duration (bottom row of Figure 4.3) revealed generally stable behaviours, with increases mostly below 10% in the majority of cases.

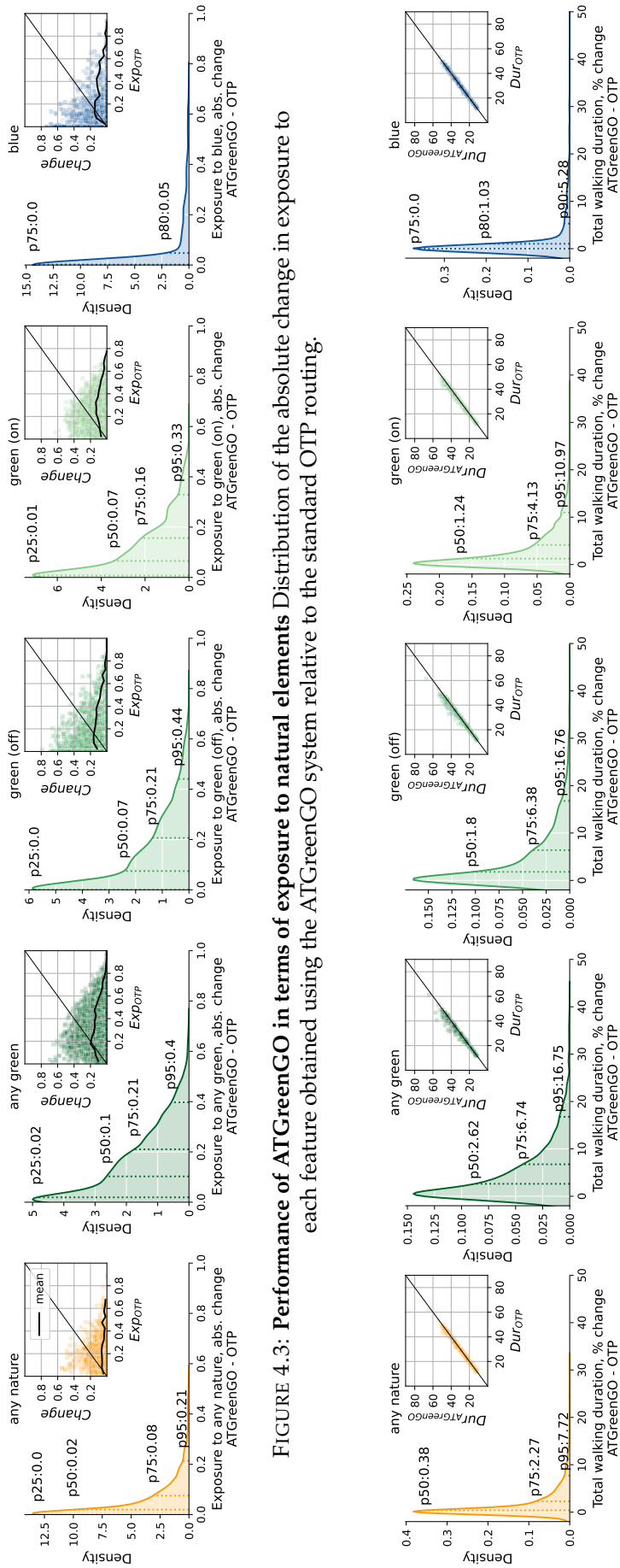

FIGURE 4.3: **Performance of ATGreenGO in terms of exposure to natural elements** Distribution of the absolute change in exposure to each feature obtained using the ATGreenGO system relative to the standard OTP routing.

FIGURE 4.4: **Performance of ATGreenGo in terms of walking duration** Distribution of the percentage change in walking duration obtained using the ATGreenGO system optimized for a specific feature relative to the standard OTP routing.

Validation of the cost function: ATGreenGO vs alternative specifications

To determine the optimal cost function to be implemented in the system, several alternative specifications have been tested. Along with the implemented function (see Eq. 4.2), the main alternative functional form tested is rooted in standard approaches adopted in transportation modelling. Instead of discounting the cost associated with an arc based on its green and blue qualities, this alternative cost function penalises arcs with little or no green/blue feature, by extending them by a stretch factor σ. More specifically, keeping the notation introduced in Section 4.6, this alternative specification was defined as:

$$C_{test}(P)^{alt} = \sum_{a \in P} \frac{L_a}{S_a}(R_a(1 + (1 - G_a)\sigma)) \tag{4.3}$$

The performance of the implemented function and the alternative specification (for several levels of σ), was then tested by evaluating the impact on the test sample of origin and destination pairs (see Section 4.7) along two dimensions:

D1 the absolute increase in exposure to the blue/green natural elements, compared to the shortest path.

D2 the percentage increase in the duration of the path (in minutes), compared to the shortest path.

Figures 4.5-4.7 offer an overview of the results of this validation process for the two dimensions, suggesting that:

D1 the increase in exposure associated with the implemented formula (named `ATGreenGO` in the labelling of the figures) is comparable to the increase observable for the alternative specifications with $\sigma = 2$. Further increasing σ for the alternative specification has a limited impact on the overall increase in exposure. This result holds for all cities and all features, except for the `blue` feature. The peculiar behaviour for the `blue` feature is a consequence of the sparsity of this natural element.

D2 the percentage increase in duration for the implemented formula is typically equal or lower to the increase obtained with the alternative specification for $\sigma = 2$. A further increase in σ above 2 leads to a substantial increase in the median value of the distribution of the percentage change of the duration of the route. Furthermore, the impact tends to become less stable across features and cities.

Overall, the validation showed that the implemented cost function is able to balance the impact on the two dimensions (high increase in exposure, limited increase in duration), also granting stability in the results across cities and features.

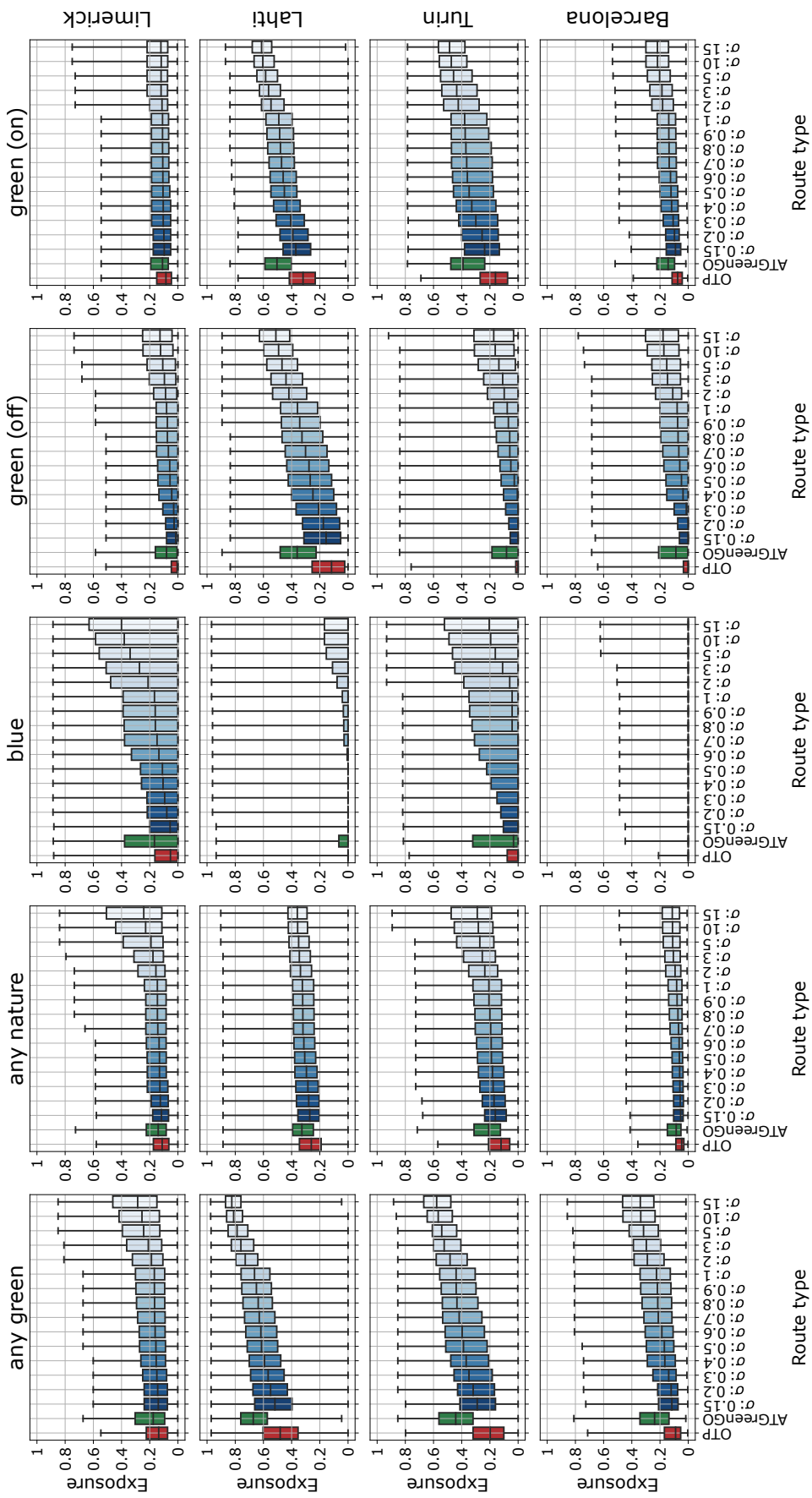

FIGURE 4.5: **ATGreenGO and alternative routing specification: exposure to natural environments (1)**Distribution of exposure to selected green and blue features, for the test set of Origin-Destination pairs and 17 cost functions, by city [Limerick, Lahti, Turin, Barcelona]

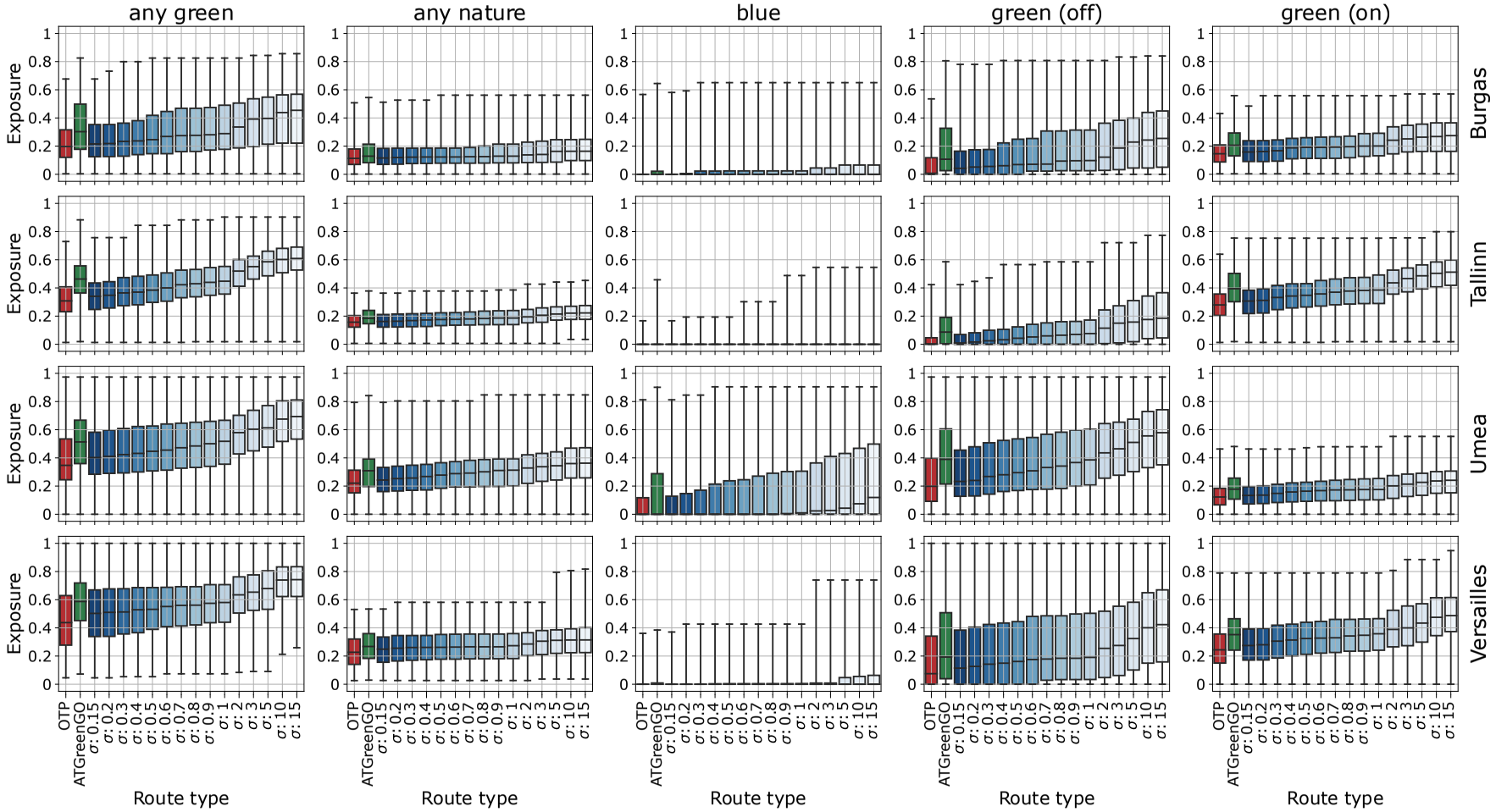

FIGURE 4.6: **ATGreenGO and alternative routing specification: exposure to natural environments (2)**Distribution of exposure to selected green and blue features, for the test set of Origin-Destination pairs and 17 cost functions, by city [Burgas, Tallinn, Umea, Versailles]

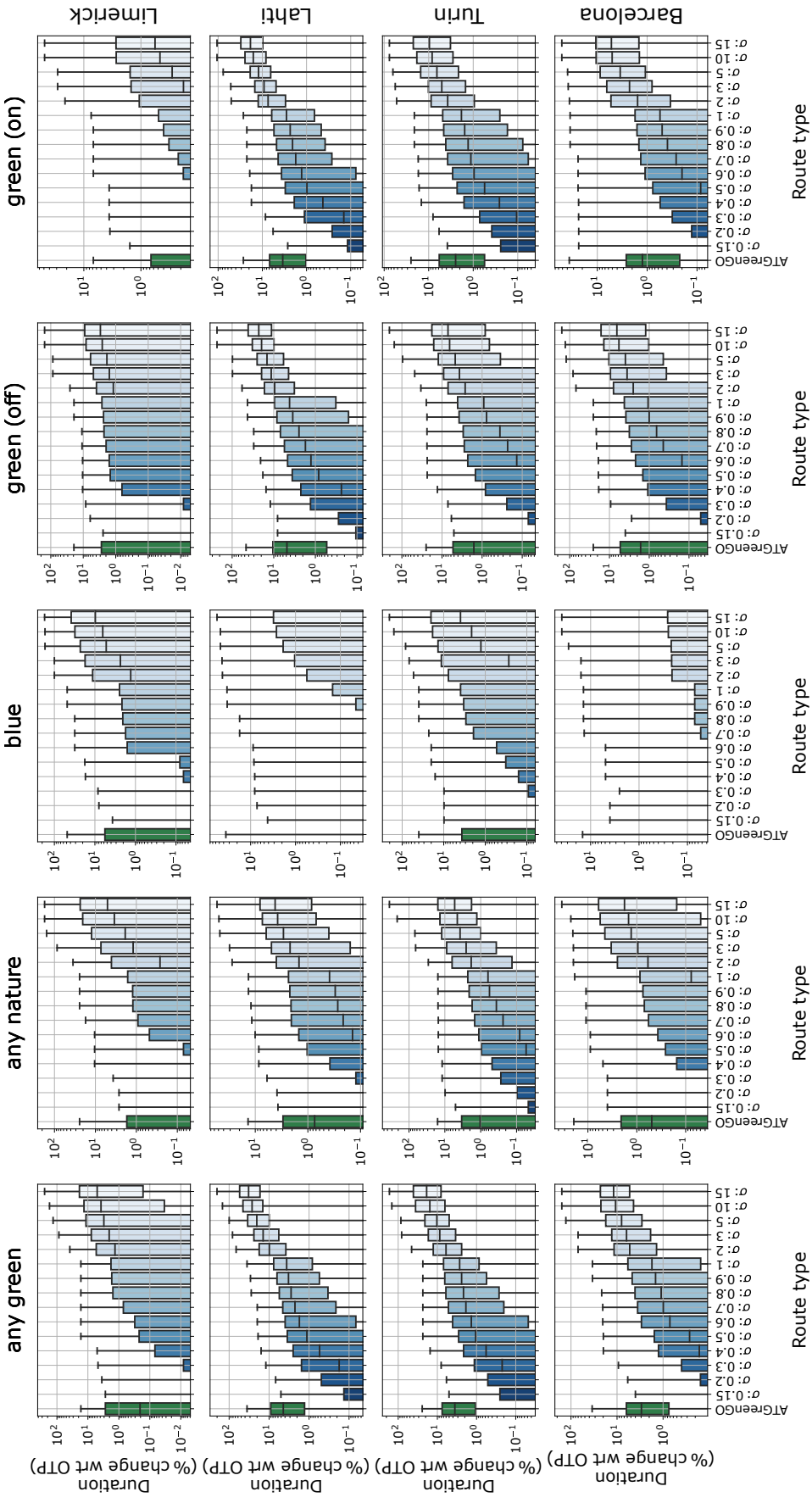

FIGURE 4.7: **ATGreenGO and alternative routing specification: walking durations (1)** Distribution of the percentage increase in duration compared to the shortest path computed using the standard OTP cost function, for the test set of Origin-Destination pairs and 16 cost functions, by city [Limerick, Lahti, Turin, Barcelona]

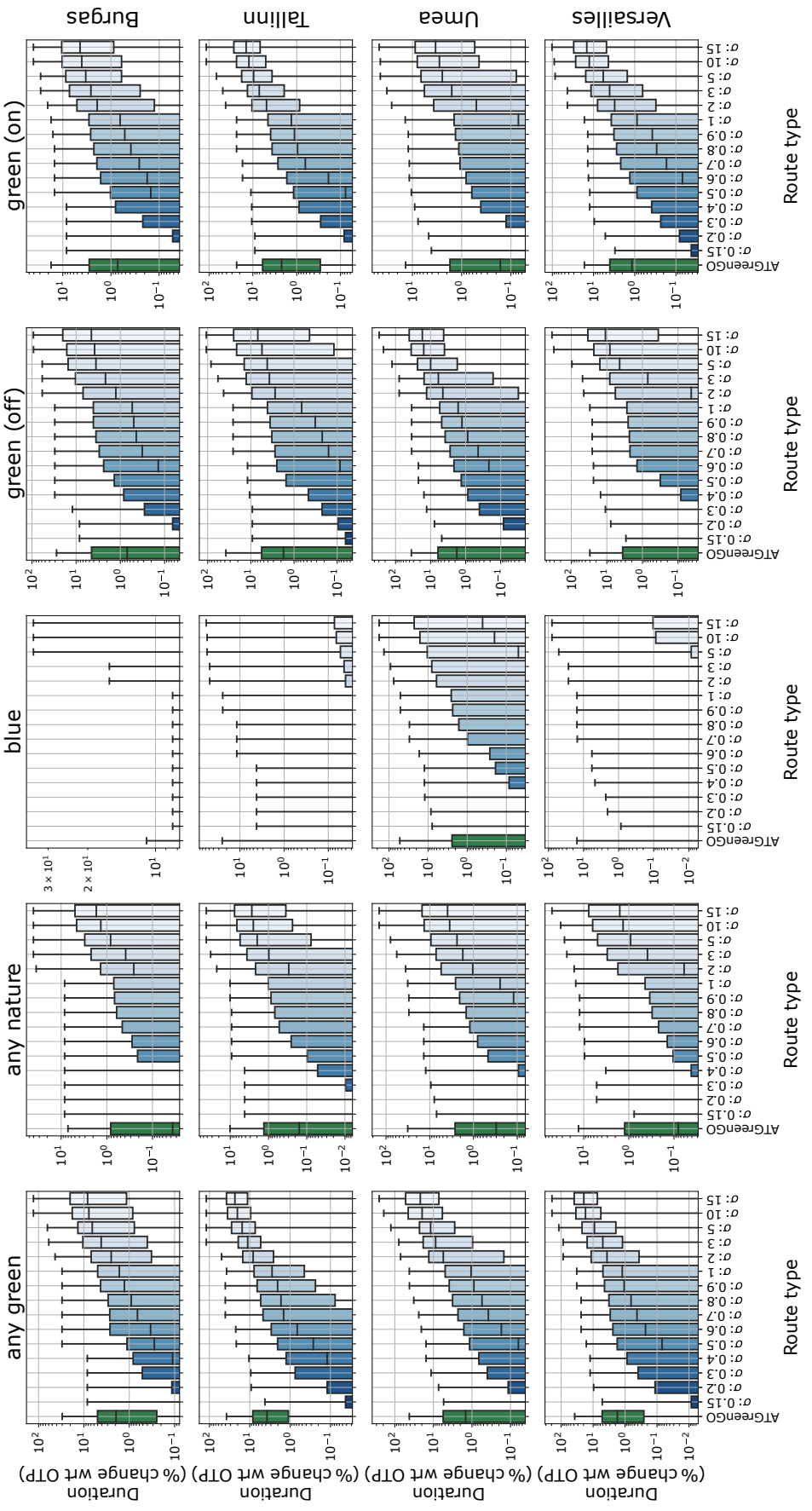

FIGURE 4.8: **ATGreenGO and alternative routing specification: walking durations (2)** Distribution of the percentage increase in duration compared to the shortest path computed using the standard OTP cost function, for the test set of Origin-Destination pairs and 16 cost functions, by city. [Burgas, Tallinn, Umea, Versailles]

4.8 Discussion

Previous research has identified the positive effects of urban greenery on health. However, these benefits can only be realised if individuals have sufficient opportunities to be exposed to nature during their day. To enhance daily exposure opportunities for residents in the cultivating cities participating in the Horizon 2020 GoGreenRoutes project, we developed an innovative routing system recommending nature-rich pedestrian routes. To maximise its usability and due to the familiarity of the research team with these urban centre, the system was then extended to three larger European cities (Turin, Barcelona and selected boroughs of the city of London). The tool was designed to help citizens plan routes that maximise their exposure to natural elements while limiting the level of detour compared to the shortest path, thus ensuring that the suggested routes are practical alternatives for daily commutes.

Our approach involved using publicly available data from OSM and satellite imagery to identify the locations of various types of natural environments. We then mapped these features onto the street networks of the cities included in the system to generate *green* and *blue* attributes. By leveraging a modified version of the open-source routing software OpenTripPlanner, we defined an optimisation algorithm that prioritises exposure to specific natural features when walking through a city.

The system has been validated by comparing the suggested routes with the shortest paths for a set of randomly selected origin-destination pairs. This comparison was performed along two dimensions: the increase in exposure to specific natural elements (green or blue) and the increase in the duration of the trip. Overall, the validation suggested that the system is able to recommend nature-rich alternatives to the shortest path while limiting the extent of the detour. The performance of the system, measured in terms of the increase in exposure to nature along the recommended routes compared to the shortest path, is in general dependent on the overall availability of the natural spaces in a specific city. For example, in nature-rich cities such as Umea, Versailles and Lahti exposure to nature is already high on the shortest path and further increases are hard to achieve. Likewise, due to the fact that the River Shannon is so central to Limerick, we can see a greater potential to walk along the river compared to e.g., Burgas and Tallinn, where the seaside does not lie on the common paths. This explains why on average, we only see modest increases of exposure to blue in this analysis, as there is often either no reasonable possibility to walk besides water bodies or the shortest path is already passing by a river. Concretely, about three in four shortest-paths routes for the `blue` feature cannot be improved either due to the absence of reasonable alternatives or already perfect exposure to this feature along the way.

To facilitate further development and usage as well as to make this tool accessible to the public, we built a straightforward software architecture that includes a middleware service and a web-based user interface. To further highlight the benefits of following the recommended route and incentivise the choice of nature-rich paths, this interface informs the users about the increase in exposure to nature that they might be able to achieve by selecting the suggested route.

While the validation confirmed the ability of the system to provide nature-rich routes, there may be other reasons for people not to be willing to choose specific routes.

Some of these concerns may stem directly from the characteristics of nature-rich areas: for instance, there may be safety reasons not to be willing to walk in a park when it is dark. Other reasons may be dependent on features not currently captured by the algorithm but whose presence may undermine the benefits of exposure to nature. For example, a qualitative investigation of the recommended routes for the city of Turin showed that nature-rich streets are often multi-lane streets where the lanes are separated by tree lines. However, these streets are also more likely to be traffic-congested and present worse air quality. To better assess the impact of these issues on the quality of the recommendations, future research requires human-based validation of ATGreenGO.

Chapter 5

Conclusions

The dissertation presented the research undertaken throughout the doctoral programme, primarily focusing on enhancing our understanding of phenomena related to urban sustainability and liveability, as well as developing tools to inform policy making and support the design of interventions to meet these objectives. The unifying theme across the three research contributions is their alignment with the objectives and principles outlined in the United Nations Sustainable Development Goal 11 [13], which sets out key objectives for cities and urban communities to be met by 2030 to ensure sustainable development. Specifically, the research contributions encompassed two major areas: 1. the investigation of gender differences in cycling preferences within Western cities (Chapter 2). 2. the development of online tools to enhance exposure to and accessibility of nature in urban environments (Chapters 3 and 4).

The significance of this research is highlighted by current global urbanisation trends. With over 50% of the world's population now living in cities and projections indicating that this figure could rise to 70% by 2050 [7], urban areas will not only accommodate more people but also become larger and more densely populated. This risks exacerbating the already disproportionate environmental impact of cities and lowering living standards for their residents [14, 7]. The expected increase in the number of mega-cities [7], defined as those with more than 10 million inhabitants, underscores the urgency of developing strategies to enhance both the sustainability and liveability of urban environments. Addressing these challenges requires a comprehensive understanding of urban phenomena but also the development of new tools to support policy making, which can inform and guide effective data-driven policy design processes [18].

Given the complexity of urban challenges, addressing these issues effectively necessitates interdisciplinary research. Combining analytical studies with tool development enables a more comprehensive approach to understanding and managing urban phenomena. Interdisciplinary research is essential for creating solutions that are both theoretically informed and practically applicable, bridging the gap between research insights and real-world applications. Recognising this, the nature of the contributions presented in this dissertation was twofold. Firstly, the analytical studies presented in Chapter 2 and Chapter 3 advance our understanding of specific urban issues, such as the gender cycling gap and urban green accessibility. These studies reveal key determinants and patterns, offering valuable insights into the factors influencing these phenomena. Secondly, the development of online tools, detailed in Chapter 3 and Chapter 4, provides practical instruments for monitoring urban

phenomena, informing policy-making, and encouraging healthy behaviours among residents.

The following sections offer a detailed summary of the key findings, contributions, and implications of these research activities. I then outline some limitations arising from the use of large-scale data to understand and model urban phenomena and suggesting future outlook of research in this area to tackle these limitations. The chapter concludes with some final reflections on the importance of a multidisciplinary approach for research in the urban science domain.

5.1 Summary of contributions

This section summarises the main results from the research activity conducted during the doctoral programme. It is organised into three sub-sections, each corresponding to one of the analytical chapters of the dissertation.

5.1.1 Chapter 2 - Investigating the determinants of the gender cycling gap using large-scale automatically-collected data

Aim of the research This study was motivated by the observation that while most research on the factors underlying the gender cycling gap relies on stated preference analyses (see Section 2.3.1), the growing availability of large-scale, automatically collected data presents a unique opportunity to deepen our understanding of this phenomenon. By leveraging such data, it is possible to expand the sample size and geographical scope of the analysis, providing more comprehensive insights into the determinants of the gender cycling gap and its large-scale characterisation.

Main findings This research consisted of a purely analytical study, with the main contributions being the empirical insights into the gender cycling gap and the identification of safety-enhancing measures as key factors in explaining its variability across and within different urban environments. The study is unique in the field of cycling behaviour and gender disparities due to its extensive geographical scale.

Firstly, the analysis confirmed the presence of a persistent gender cycling gap in recreational cycling across all cities in the sample. However, there was substantial variation in the magnitude of this gap, both across cities in different geographical areas and within the same geographical area. This suggests that factors beyond cultural influences and Strava penetration play a role in this phenomenon.

To investigate these determinants, the study introduced a set of city-level indicators inspired by the stated-preferences literature on gender differences in cycling perceptions and behaviours. A multivariate regression analysis was conducted to examine the association between these indicators and the observed gender cycling gap, with a particular focus on the role of safety-enhancing measures. The results indicated that safer urban environments, measured by the proportion of streets with a maximum speed limit of less than 20 mph (or 30 km/h), are typically associated with smaller gender cycling gaps. This finding supports existing evidence from survey-based studies that women perceive cycling as a riskier activity than men and are less likely to cycle in environments perceived as unsafe.

In the second phase of the study, the focus shifted from a city-wide analysis to examining differences in street-level characteristics within the same city. The analysis revealed that streets with dedicated cycling infrastructure are more likely to exhibit a balanced gender distribution among cyclists. For instance, in New York City, streets with protected cycleways were found to be four times more likely to have a low gender cycling gap compared to streets without such infrastructure. A smaller, but still positive, association was observed for streets with unprotected cycleways. Extending this analysis to other cities in the dataset and applying a minimal logistic regression model, similar associations were found to hold across the majority of cities. Overall, this study represented the first large-scale validation of the impact of safety-enhancing measures on sustaining cycling adoption among women, thereby corroborating findings from earlier, smaller-scale stated-preference studies [46, 47, 48].

5.1.2 Chapter 3 - ATGreen: a multi-dimensional computational framework to evaluate accessibility to urban green

Aim of the research This research was motivated by the observed lack of universally accepted standards for measuring green accessibility, despite the growing reliance on these spatial metrics for evaluation, monitoring, and informing policy design (see Section 3.3). In this context, the study aimed to assess the impact of using different metrics to inform the policy design process by comparing the accessibility outcomes produced by various indicators, whether they belong to the same class of accessibility indicators or different ones. Instrumental to this objective was the development of a computational framework capable of measuring multiple classes of indicators within a unified system, thereby enabling the proposed comparison.

Main findings and contributions The main findings and contributions of this study are twofold: first, the analytical insights gained from the empirical investigation into the interchangeability of various green accessibility indicators; and second, the algorithmic and programming advancements achieved through the development of the computational framework and its associated web interface.

Regarding the empirical investigation, I assessed the interchangeability of green accessibility outcomes generated by different green indicators. Adopting a policy perspective, the study utilised hypothetical policy interventions and established institutional targeting strategies to identify severely under-served population subgroups within each city. The investigation focused on how the identification of these groups varies depending on the chosen indicator or its parameterisation (i.e., the selection of underlying parameters in defining an indicator). The results revealed significant variability in the stability of these targeted populations depending on the specific urban context and the parametrisation or institutional targets considered. These findings suggest the necessity of adopting a multi-dimensional perspective in evaluating green accessibility.

In terms of algorithmic advancements, the study introduced a computational framework capable of efficiently measuring multiple classes of spatial green accessibility indicators at a high geographical resolution within a unified system. By utilising only open-source tools and open data, with all resources publicly released, this framework is designed to be replicable and extendable for future monitoring requirements.

The final contribution of the study is the development of the associated web interface. Going beyond a mere tool for exploring results, the tool was specifically designed for policymakers, offering customised functionalities that allow for the definition of new indicators and the simulation of specified greening interventions.

5.1.3 Chapter 4 - ATGreenGO: a routing engine recommending nature-enriching routes

Aim of the research This research stemmed from the observation that exposure to nature often occurs incidentally during daily activities. Specifically, it examined the potential for increasing such exposure during routine walking trips. Expanding and building upon recent advancements in routing engines that prioritise qualitative criteria of optimality (see Section 4.3), this study aimed to develop a routing engine that recommends nature-enriched routes with minimal detours relative to the shortest paths. The goal was to offer a tool to identify practical alternatives that can be seamlessly integrated into daily routines, thereby enhancing opportunities for regular exposure to nature.

Main contributions The main contributions of this study are both theoretical and practical through the development of an API and an interactive tool for other researchers, policymakers and the general public.

Theoretically, the contributions are twofold. First, I proposed and implemented a novel algorithm to characterise the geometrical relationship between the street network of a city and its existing green and blue infrastructure. Second, I formulated a new routing optimisation problem that balances exposure to nature with the detour required relative to the shortest routes. The performance of this proposed formulation was compared to alternative approaches commonly used in transportation literature, demonstrating greater robustness across different urban settings.

In terms of contributions to the academic and policy communities, the development of a fully open-source, data-based solution ensures that the system can be easily extended to urban settings beyond the currently deployed nine cities, thereby facilitating further research and informing policymaking. Moreover, the creation of an associated web interface accessible to the general public enables individuals in these cities to easily use the routing engine, thus enhancing opportunities for exposure to nature.

5.2 Implications

In recent decades, the shift in urban planning towards a more data-driven policy design process has led to the integration of data and computational modeling across various stages of the design process [18, 174]. This ranges from understanding phenomena to monitoring them and, ultimately, developing tools that optimise policy design processes based on desired optimal outcomes. This integration highlights the crucial importance of a multidisciplinary approach to research into sustainable and liveable cities, requiring expertise that spans urban design, data and computer science, environmental studies, as well as economics and social science.

The research presented in this dissertation exemplifies the potential contributions at various stages of the policy design process that can arise from the adoption of a multidisciplinary approach and the integration of large-scale data. In line with

this, the research offers implications across three four domains: knowledge advancements for researchers and policymakers, theoretical and algorithmic advancements, data advancements, and broader (actionable) impact of the online tools. Due to the thematic difference between the study presented in Chapter 2 and the studies presented in Chapters 3 and 4, the implications are discussed for the two macro-area of research separately. As such, in what follows, the term macro research area 1 refers to the analysis of the gender cycling gap (Chapter 2) whilst the term macro research area 2 refers to the contributions on enhancing exposure and accessibility to urban green (Chapters 3 and 4).

Knowledge advancements for researchers and policymakers In terms of macro research area 1, this study represents a significant advancement in understanding how safety-enhancing measures can promote cycling among women. While the findings align with existing research on stated preferences, the study's innovation lies in its broader geographical scope and the extensive user base contributing to the data. By examining a diverse set of cities across multiple continents, the research offers a more comprehensive and robust view of how safety-related factors impact women's cycling habits. This expanded scope enhances the generalisability of the results, demonstrating that safety measures consistently affect women's cycling behaviour across various urban environments. Moreover, this study makes a critical contribution to the research community by developing a methodology that integrates large-scale data analysis with gender-specific urban mobility concerns. This approach can be applied to other studies within the fields of transport planning, urban design, and gender studies, offering researchers a model for examining similar issues on a global scale. Additionally, the research strengthens the theoretical foundation for policymaking activities aimed at increasing cycling uptake, particularly among underrepresented groups like women. By providing empirical evidence across different contexts, the study equips policymakers with actionable insights that can be adapted to various cities and local conditions, thus bridging the gap between theory and practice.

Regarding the macro research area 2, the primary contribution is from the green accessibility study presented in Chapter 3. Unlike previous research, which typically relied on a single indicator to evaluate green accessibility, this study reveals that using a single metric may not provide a sufficiently nuanced understanding of accessibility. The findings underscore the importance of conducting sensitivity analyses to assess the potential limitations and biases introduced by the choice of indicator parameters. This approach is particularly crucial when evaluating the differential performance of green accessibility measures across various areas of a city, as it ensures a more accurate and comprehensive assessment. The study's emphasis on the necessity of multiple metrics highlights a critical gap in existing research, providing a more robust methodology for future studies. This contribution not only advances the theoretical framework for evaluating green accessibility but also offers practical guidance for urban planners and policymakers due to the design of the study and its use of fictional policy strategies.

Theoretical and algorithmic advancements The research conducted during the PhD programme in macro research area 2 has also led to notable theoretical and algorithmic advancements. A key theoretical and algorithmic contribution is the development of a computational framework for measuring various classes of green accessibility indicators. Previously, the computational complexity associated with

designing spatial indicators limited the ability to create a unified framework that could simultaneously measure multiple classes of indicators. This research contributed to the literature in two ways. Firstly, it introduced a functional form for each class of green accessibility indicators and defined a clear set of underlying parameters. These parameters can be adjusted to align the metrics with those commonly adopted in academic and policy literature. Secondly, the research presented a feasible and efficient algorithm that enables the computation of these classes of indicators, overcoming previous limitations in computational feasibility.

A further theoretical contribution is found in the formulation of the optimisation problem for the routing engine ATGreenGO, as discussed in Chapter 4. This new formulation represents a significant departure from traditional transportation models. Unlike conventional approaches in the transportation literature, it yields robust results across diverse urban settings without the need for calibration. Additionally, the research introduces an innovative algorithm for characterising the street network of a city based on its geometrical relationship with nearby natural elements. Although the final approach adopted a simplified geometrical characterisation to better align with a walker's experience, the detailed geometrical considerations are presented here for future reference. This detailed characterisation could prove valuable for research on structural street-level greening interventions, where such precision is expected to be more relevant.

Data advancements Contributions to data advancements are significant across both macro research areas addressed in this dissertation.

In macro research area 1, the study stands out as the first large-scale characterisation of gender disparities in cycling behavior. This work exemplifies the potential of utilising large-scale data for revealed preference studies within urban mobility and cycling. Methodologically, the study required extensive data processing and the development of specialised pipelines and algorithms tailored to handle Strava data. Despite concerns about the representativeness of Strava data, which may necessitate the exploration of alternative data sources, the proposed data pipeline is robust and replicable for future research in this field.

In macro research area 2, contributions to data advancements are also noteworthy. Although the research utilised publicly available data that are well-established in academic literature, the methods of processing and validating this data represent a significant improvement over previous approaches. This is particularly evident in the thorough validation of OpenStreetMap (OSM) data discussed in Chapter 3. While there is growing academic interest in enhancing the quality of OSM data and techniques for addressing data gaps, thorough validation of these data is often overlooked in applied research, with gaps and biases dismissed as study limitations. Alternatively, researchers tend to focus on a subset of geographical areas (typically large cities in the Global North) where data quality is known to be higher. Building on existing research, I developed a validation process that significantly enhanced data reliability, allowing for the inclusion of areas previously excluded due to data quality concerns. This approach not only strengthens the validity of the findings but also sets a precedent for more rigorous data handling in future studies.

Broader impact of the online tools for researchers, policymakers and urban residents Chapters 3 and 4 presented the research activities focused on the provision

of tools to enhance exposure to nature in urban environments. These tools are designed to extend the impact of the research beyond the academic community, effectively reaching policymakers and urban residents.

Recognising the computational complexity of measuring high-resolution spatial indicators and the necessity of adopting a multidimensional perspective, we constructed an online tool that can compute multiple indicators simultaneously and assess the impact of selected interventions. This tool can be used by policymakers and academic researchers alike to embrace this multidimensional approach. Moreover, all our code is made publicly available, enabling researchers or policymakers with programming expertise to easily adapt the framework to specific geographical areas not currently covered by the study.

The contributions and policy implications of the second tool—the routing engine ATGreenGO—are multifaceted. For the academic community, the tool offers a flexible architecture that can be easily expanded to include additional cities and data layers. Furthermore, the public API makes the routing data accessible to academic researchers for future studies in evaluating green exposure and walkability. The broader impact of the tool lies in its potential use by residents and policymakers in the cities currently covered by the system. One of the primary motivations for developing ATGreenGO was its ability to encourage active mobility. Research consistently shows that natural environments increase the willingness to walk, encouraging more people to choose walking over other modes of transportation. By suggesting nature-rich routes, ATGreenGO not only promotes walking but also fosters a cultural shift towards active mobility. By facilitating regular contact with natural environments, ATGreenGO effectively supports public health objectives, providing a practical tool for policymakers aiming to enhance urban health. Indirectly, ATGreenGO could also serve as a valuable resource for identifying areas deficient in green routes. Policymakers can use this information to target investments in street-level green infrastructure where it is most needed, promoting more equitable and sustainable urban spaces.

5.3 Limitations and future outlook

Despite our best attempts, the research activities presented in this dissertation have certain limitations, which also present opportunities for future research. Study-specific limitations and future research avenues have been highlighted in detail in the discussion of each contribution. However, we can here outline broader limitations related to the integration of large-scale data into urban policy design and suggest priorities for future academic research in this area.

Broadly speaking, the main limitations of the use of large-scale data pertain to the following domains: their completeness, their representativeness, and their availability.

Completeness Concerns about the completeness of large-scale data, particularly for those data collected through voluntary participation and crowd-sourced initiatives like OpenStreetMap (OSM), are well-recognised in academic research [135]. Despite its widespread use, OSM has been scrutinised for its completeness, prompting increasing efforts to develop strategies for internal [136, 137, 138] and external validation [139, 140, 141], as discussed in Section 3.4. However, despite these efforts,

the limitations of OSM data are often addressed either by dismissing them as potential study limitations or by restricting the analysis to areas with traditionally better coverage, such as those in Western countries or the Global North.

This common approach to handling completeness issues presents several challenges and risks. First, findings based predominantly on data from the Global North may not accurately reflect conditions in other parts of the world, potentially skewing our understanding of urban phenomena. Additionally, focusing on well-covered regions can lead to an over-allocation of research resources to areas that are already relatively well-equipped, while neglecting regions that may have greater needs but less available data.

In this research, we encountered similar challenges. The study of gender differences in cycling was confined to Western countries due to data availability from Strava, which offered a sufficiently large dataset only within these regions. Nonetheless, we made preliminary strides to mitigate this limitation in the development of the ATGreen tool, whose coverage is not limited to the Global North. Instead of selecting a pre-defined set of cities, we implemented extensive cleaning and processing a global dataset to address the heterogeneous quality of OSM data. This ensure that only cities with adequate data quality were included. This approach helped to improve the reliability of the findings without restricting its geographical scope .

More generally, while future research should continue to explore the integration of large data analytics in urban planning to address the complexities of rapidly growing cities, expanding the geographical scope of studies to include cities in the Global South should be a priority. This requires ongoing efforts to develop and maintain datasets with comprehensive geographical coverage, as well as to establish robust tools for assessing the quality of this coverage.

Representativeness The representativeness of large-scale automatically-collected geographic data, such as those derived from mobile phone or GPS traces, presents a substantial challenge in urban studies. While these datasets offer extensive and detailed insights, they are inherently biased due to their dependence on voluntary user participation and specific technology adoption patterns. For example, GPS data gathered through mobile applications often overrepresents individuals who are younger, urban, and higher-income—groups with higher access to smart devices and greater engagement with technology—while underrepresenting older adults, rural residents, and lower-income populations. This bias can result in a skewed spatial distribution of data points, with higher density in regions with more user activity or better infrastructure, which may distort analyses and conclusions [67]. In the context of the research presented here, concerns about the representativeness of the data were raised in relation to the behavioural data on cycling obtained from the sport-tracking application Strava. Strava data may exhibit selection bias due to differences between users who contribute data and those who do not, and the predominant use of the platform for recreational rather than commuting activities further complicates the representativeness of commuting behavior analyses. These are discussed in detail in the Section 2.8.

Addressing these representativeness concerns requires a multidisciplinary approach. Given that biases are often intrinsic to the nature of the data, supplementing quantitative analyses with qualitative methods or direct data collection can provide a more comprehensive and balanced perspective. Integrating qualitative approaches, such as interviews or focus groups, and direct data collection methods, such as targeted

surveys or observational studies, can help capture nuances and perspectives that automated data may miss. This mixed-methods approach allows for the validation and triangulation of findings, offering a richer understanding of urban phenomena and enhancing the robustness of research outcomes. By combining diverse data sources and methodologies, researchers can better address representativeness issues and ensure that their findings are more reflective of the broader population and context.

Availability In designing the research presented in this dissertation, a deliberate choice was made to use only publicly available data to ensure transparency and reproducibility. However, a growing concern in urban studies should be the proprietary nature of many large-scale datasets critical for informing urban policy development. These proprietary datasets, often controlled by private companies, present significant challenges due to restricted access and limited transparency. Such control impacts on the ability of researchers to validate, replicate, or fully comprehend the data, which can lead to potential biases or incomplete analyses. Furthermore, the restricted availability of these datasets can exacerbate inequalities in research opportunities, privileging well-resourced institutions or individuals with special access. Addressing this issue requires advocating for greater data openness and transparency, allowing the full social value of these data to be realised, but also supporting existing initiatives to develop openly accessible or collaborative datasets, thereby ensuring a more equitable and comprehensive foundation for urban policy-making.

5.4 Final thoughts

In reflecting upon the research conducted throughout my doctoral studies, it becomes increasingly evident that a multidisciplinary approach is essential for addressing the complex challenges within urban data science. The integration of diverse expertise —from data and computer science, urban planning, and social sciences to public policy— enriches the analysis and ensures a more comprehensive understanding of urban phenomena. Moreover, extending collaboration beyond the academic community to encompass local authorities, regulators, and public bodies is crucial. Such partnerships facilitate the practical application of research findings and ensure that policies are informed by a broad range of perspectives. Engaging with these stakeholders not only enhances the relevance and impact of research but also fosters the development of more nuanced and effective solutions to urban issues. A pivotal aspect of my research experience has been the collaboration within the GoGreenRoutes consortium and at the Barcelona Supercomputing Centre. These experiences provided valuable insights into the application of the research findings, helped me to familiarise with actual policy processes, and highlighted the potential impact of such collaborations in ensuring that research contributes effectively to real-world urban challenges.

Bibliography

[1] M. Schiavina, S. Feire and K. MacManus. GHS population grid multitemporal (1975, 1990, 2000, 2015) R2019A. European Commission, Joint Research Centre (JRC) [Dataset], 2019.

[2] C. Corbane M. Schiavina M. Maffenini M. Pesaresi P. Politis S. Sabo S. S. Freire D. Ehrlich A.J. Florczyk, M. Melchiorri et al. Description of the GHS urban centre database 2015. *Public Release*, 2019.

[3] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org. Information available at: https://www.openstreetmap.org, 2017.

[4] D. Luxen and C. Vetter. Real-time routing with openstreetmap data. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 513–516, New York, NY, USA, 2011. ACM.

[5] OpenTripPlanner contributors. OpenTripPlanner [Computer Software]. Information available at: https://www.opentripplanner.org/.

[6] D. Zanaga and others. ESA WorldCover 10 m 2020 v100 [Dataset]. 10.5281/zenodo.5571936, 2021.

[7] United Nations, Department of Economic and Social Affairs, Population Division. World Urbanization Prospects: The 2018 Revision, 2018.

[8] M. Batty. Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies. *CASA Working Papers (131).*, 2012.

[9] United Nations Human Settlements Programme (UN-Habitat). World Cities Report 2020: The Value of Sustainable Urbanization, 2020.

[10] B. Güneralp K. C. Seto and L. R. Hutyra. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proceedings of the National Academy of Sciences*, 109(40):16083–16088, 2012.

[11] P. Newman and J. Kenworthy. *Sustainability and Cities: Overcoming Automobile Dependence*. Island Press, Washington, D.C., 1999.

[12] E. L. Glaeser. *Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier*. Penguin Press, New York, 2011.

[13] ODDS Cf. Transforming our world: the 2030 agenda for sustainable development. *United Nations: New York, NY, USA*, 2015.

[14] IPCC. Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, 2022.

[15] S. Piaggesi A. Young N. Adler S. Verhulst L. Ferres L. Gauvin, M. Tizzoni and C. Cattuto. Gender gaps in urban mobility. *Humanities and Social Sciences Communications*, 7(1):1–13, 2020.

[16] P. Pucci C. D'Agostino, E. Piva and C. Rossi. A systematic literature review on women's daily mobility in the global north. *Transport Reviews*, pages 1–29, 2024.

[17] P. Bajardi A. Panisson A. Perotti M. Szell M. A. Battiston, L. Napoli and R. Schifanella. Revealing the determinants of gender inequality in urban cycling with large-scale data. *EPJ Data Science*, 12(1):9, 2023.

[18] B. Giles-Corti, A.V. Moudon, M. Lowe, D. Adlakha, E. Cerin, G. Boeing, C. Higgs, J. Arundel, S. Liu, E. Hinckson and others. Creating healthy and sustainable cities: what gets measured, gets done. *The Lancet Global Health*, 10(6):e782–e785, 2022.

[19] A. Battiston and R. Schifanella. On the need for a multi-dimensional framework to measure accessibility to urban green. *npj Urban Sustainability*, 4(1):1–11, 2024.

[20] R Schifanella D. Quercia and L. M. Aiello. The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 116–125, 2014.

[21] X. Xie Y. Zheng, L. Zhang and W. Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*, pages 791–800, 2009.

[22] A. EC. Redondi A. Pimpinella and M. Cesana. Walk this way! an iot-based urban routing system for smart cities. *Computer Networks*, 162:106857, 2019.

[23] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

[24] K. Jordahl. Geopandas: Python tools for geographic data. *URL: https://github.com/geopandas/geopandas*, 2014.

[25] G. Boeing. Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65:126–139, 2017.

[26] E. Seabold and J. Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, pages 10–25080. Austin, TX, 2010.

[27] A. Gramfort V. Michel B. Thirion O. Grisel M. Blondel P. Prettenhofer R. Weiss V. Dubourg J. Vanderplas A. Passos D. Cournapeau M. Brucher M. Perrot F. Pedregosa, G. Varoquaux and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[28] P. Oja, S. Titze, A. Bauman, B. De Geus, P. Krenn, B. Reger-Nash, and T. Kohlberger. Health benefits of cycling: a systematic review. *Scandinavian journal of medicine & science in sports*, 21(4):496–509, 2011.

[29] L. Leyland, B. Spencer, N. Beale, T. Jones, and C. M Van Reekum. The effect of cycling on cognitive function and well-being in older adults. *PloS one*, 14(2):e0211779, 2019.

[30] T. Cole-Hunter D. Donaire-Gonzalez M. Jerrett M. D. A. Rodriguez I. Avila-Palencia, A. de Nazelle and M. J. Nieuwenhuijsen. The relationship between bicycle commuting and perceived stress: a cross-sectional study. *BMJ open*, 7(6):e013542, 2017.

[31] S. Gössling, A. Choi, K. Dekker, and D. Metzler. The social cost of automobility, cycling and walking in the European Union. *Ecological Economics*, 158:65–74, 2019.

[32] World Cycling Alliance European Cyclists Federation. Cycling delivers of the global goals - shifting towards a better economy, society and planet for all. Available at: https://www.ecf.com/groups/cycling-delivers-global-goals, 2016.

[33] K. Hosford and M. Winters. Quantifying the bicycle share gender gap. *Findings*, page 10802, 2019.

[34] D. Funaki. *Why Don't Women Cycle?: A Case Study of Women's Perceptions of Cycling in the SOMA District of San Francisco*. PhD thesis, University of California, Berkeley, 2019.

[35] Cycling UK. Cycling statistics. Available at: https://www.cyclinguk.org/statistics, 2021.

[36] J. Pucher and R. Buehler. Making cycling irresistible: lessons from the netherlands, denmark and germany. *Transport reviews*, 28(4):495–528, 2008.

[37] J. Monk and S. Hanson. On not excluding half of the human in human geography. *The Professional Geographer*, 34(1):11–23, 1982.

[38] S. Hanson. Gender and mobility: new approaches for informing sustainability. *Gender, Place & Culture*, 17(1):5–23, 2010.

[39] Strava. Year in sport 2018. https://blog.strava.com/press/2018-year-in-sport/, 2018.

[40] C. Howard and E. K. Burns. Cycling to work in phoenix: Route choice, travel behavior, and commuter characteristics. *Transportation Research Record*, 1773(1):39–46, 2001.

[41] N. Eluru I. N. Sener and C. R. Bhat. An analysis of bicycle route choice preferences in texas, us. *Transportation*, 36:511–539, 2009.

[42] B. Van Wee E. Heinen and K. Maat. Commuting by bicycle: an overview of the literature. *Transport reviews*, 30(1):59–96, 2010.

[43] C. Pistoll and A. Goodman. The link between socioeconomic position, access to cycling infrastructure and cycling participation rates: An ecological study in melbourne, australia. *Journal of Transport & Health*, 1(4):251–259, 2014.

[44] X. Wen A. C. Lusk and L. Zhou. Gender and used/preferred differences of bicycle routes, parking, intersection signals, and bicycle type: Professional middle class preferences in hangzhou, china. *Journal of Transport & Health*, 1(2):124–133, 2014.

[45] K. Wang and G. Akar. Gender gap generators for bike share ridership: Evidence from citi bike system in new york city. *Journal of transport geography*, 76:1–9, 2019.

[46] R. Aldred, B. Elliott, J. Woodcock, and A. Goodman. Cycling provision separated from motor traffic: a systematic review exploring whether stated preferences vary by gender and age. *Transport reviews*, 37(1):29–55, 2017.

[47] A. Misra and K. Watkins. Modeling cyclist route choice using revealed preference data: an age and gender perspective. *Transportation Research Record*, 2672(3):145–154, 2018.

[48] J. Dill, T. Goddard, C. Monsere, and N. McNeil. Can protected bike lanes help close the gender gap in cycling? lessons from five cities. *Transportation Research Board 94th Annual Meeting*, pages 1–18, 2015.

[49] N. Tilahun K. J. Krizek, P. J. Johnson et al. Gender differences in bicycling behavior and facility preferences. *Research on Women's Issues in Transportation*, 2(35):31–40, 2005.

[50] Shannon K. C. Heesch, S. Sahlqvist and J. Garrard. Gender differences in recreational and transport cycling: a cross-sectional mixed-methods comparison of cycling patterns, motivators, and constraints. *International Journal of Behavioral Nutrition and Physical Activity*, 9:1–12, 2012.

[51] G. Rose J. Garrard and S. Lo. Promoting transportation cycling for women: the role of bicycle infrastructure. *Preventive medicine*, 46(1):55–59, 2008.

[52] M. Winters and M. Zanotto. Gender trends in cycling over time: an observational study in vancouver, british columbia. *Journal of Transport & Health*, 5:S37–S38, 2017.

[53] E. Sall J. Hood and B. Charlton. A gps-based bicycle route choice model for san francisco, california. *Transportation letters*, 3(1):63–75, 2011.

[54] N. Grudgings, S. Hagen-Zanker, A.and Hughes, B. Gatersleben, M. Woodall, and W. Bryans. Why don't more women cycle? an analysis of female and male commuter cycling mode-share in england and wales. *Journal of Transport & Health*, 10:272–283, 2018.

[55] J. Garrard, S. Handy, and J. Dill. Women and cycling. *City cycling*, 2012:211–234, 2012.

[56] R. Aldred R. Nakamura-L. Tatah L. M. T. Garcia B. Zapata-Diomedi T. H.. de Sa G. Tiwari A. de Nazelle R. Goel, A. Goodman et al. Cycling behaviour in 17 countries across 6 continents: levels of cycling, who cycles, for what purpose, and how far? *Transport reviews*, 42(1):58–81, 2022.

[57] N. Maizlish T. Schwanen-A. Goodman T. Götschi, M. Tainio and J. Woodcock. Contrasts in active transport behaviour across four countries: How do they translate into public health benefits? *Preventive medicine*, 74:42–48, 2015.

[58] N. Fischer G. Akar and M. Namgung. Bicycling choice and gender case study: The ohio state university. *International Journal of Sustainable Transportation*, 7(5):347–365, 2013.

[59] K. Maat E. Heinen and B. van Wee. The effect of work-related factors on the bicycle commute mode choice in the netherlands. *Transportation*, 40(1):23–43, 2013.

[60] G. Ghoshal C. R. James M. Lenormand T. Louail R. Menezes J. J. Ramasco José F. Simini H. Barbosa, M. Barthelemy and M. Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.

[61] I. Laña M. Ilardia M. N. Bilbao A. I. Torre-Bastida, J. Del Ser and S. Campos-Cordobés. Big data for transportation and mobility: recent advances, trends and challenges. *IET Intelligent Transport Systems*, 12(8):742–755, 2018.

[62] E. Eren and V. E. Uz. A review on bike-sharing: The factors affecting bike-sharing demand. *Sustainable Cities and Society*, 54:101882, 2020.

[63] K. Wang and G. Akar. Gender gap generators for bike share ridership: Evidence from citi bike system in new york city. *Journal of Transport Geography*, 76:1–9, 2019.

[64] T. S. Throndsen T. P. Uteng, H. M. Espegren and L. Böcker. The gendered dimension of multimodality: exploring the bike-sharing scheme of oslo. In *Gendering Smart Mobilities*, pages 162–187. Routledge, 2019.

[65] Y. Sun and A. Mobasheri. Utilizing crowdsourced data for studies of cycling and air pollution exposure: A case study using strava data. *International journal of environmental research and public health*, 14(3):274, 2017.

[66] W. Musakwa and K. M. Selala. Mapping cycling patterns and trends using strava metro data in the city of johannesburg, south africa. *Data in brief*, 9:898–905, 2016.

[67] E. Willberg, H. Tenkanen, A. Poom, M. Salonen, and T. Toivonen. Comparing spatial data sources for cycling studies: A review. *Transport in human scale cities*, pages 169–187, 2021.

[68] World Economic Forum. The Global Gender Gap Index 2020. Available at: https://www3.weforum.org/docs/WEF_GGGR_2020.pdf, 2020.

[69] G. Boeing. Street network models and indicators for every urban area in the world. *Geographical Analysis*, 2021.

[70] New York City Council. NYC Open Data. Available at: https://opendata.cityofnewyork.us/.

[71] V. Nicosia V. Latora and G. Russo. *Complex Networks: Principles, Methods and Applications*. Complex Networks: Principles, Methods and Applications. Cambridge University Press, Cambridge, 2017.

[72] T. P. Peixoto. The graph-tool python library. *figshare*, 2014.

[73] Ministry of Infrastructure and Netherland Water Management. Cycling facts: new insights, 2020. https://s23705.pcdn.co/wp-content/uploads/2021/03/Netherlands-Cycling-Facts-2020.pdf, 2018.

[74] Strava. Year in sport 2019. Available at: https://blog.strava.com/press/strava-releases-2019-year-in-sport-data-report/, 2019.

[75] J. W. Tukey et al. *Exploratory data analysis*, volume 2. Reading, MA, 1977.

[76] R. Mekary T.K. Courtney D.C. Christiani M. Asgarzadeh, S. Verma. The role of intersection and street design on severity of bicycle-motor vehicle crashes. *Injury prevention*, 23(3):179–185, 2017.

[77] B. Gabbe S. Braaf S. White M. Backhouse M. L. Pearson, J. Dipnall and B. Beck. The potential for bike riding across entire cities: Quantifying spatial variation in interest in bike riding. *Journal of Transport & Health*, 24:101290, 2022.

[78] W. E. Marshall and N. N. N. N. Ferenchak. Why cities with high bicycling rates are safer for all road users. *Journal of Transport & Health*, 13:100539, 2019.

[79] T. Goddard and J. Dill. Gender differences in adolescent attitudes about active travel. In *Transportation Research Board 93rd Annual Meeting. Washington, DC*, 2014.

[80] K. C Heesch, S. Sahlqvist, and J. Garrard. Gender differences in recreational and transport cycling: a cross-sectional mixed-methods comparison of cycling patterns, motivators, and constraints. *International Journal of Behavioral Nutrition and Physical Activity*, 9(1):1–12, 2012.

[81] G. Prati, F. Fraboni, M. De Angelis, L. Pietrantoni, D. Johnson, and J. Shires. Gender differences in cycling patterns and attitudes towards cycling in a sample of european regular cyclists. *Journal of transport geography*, 78:1–7, 2019.

[82] R. Aldred, J. Woodcock, and A. Goodman. Does more cycling mean more diversity in cycling? *Transport reviews*, 36(1):28–44, 2016.

[83] M. Graystone, R. Mitra, and P. M Hess. Gendered perceptions of cycling safety and on-street bicycle infrastructure: Bridging the gap. *Transportation Research Part D: Transport and Environment*, 105:103237, 2022.

[84] G. Prati. Gender equality and women's participation in transport cycling. *Journal of transport geography*, 66:369–375, 2018.

[85] T. G. Benton M. A. Goddard, A. J. Dougill. Scaling up from gardens: biodiversity conservation in urban environments. *Trends in ecology & evolution*, 25(2):90–98, 2010.

[86] X. Xue M. Shafique and X. X. Luo. An overview of carbon sequestration of green roofs in urban areas. *Urban Forestry & Urban Greening*, 47:126515, 2020.

[87] N. Shishegar. The impact of green areas on mitigating urban heat island effect: A review. *International journal of environmental sustainability*, 9(1):119–130, 2014.

[88] M. Piccardo L. Caporaso H. Taubenböck A. Cescatti G. Duveiller E. Massaro, R. Schifanella. Spatially-optimized urban greening for reduction of population exposure to land surface temperature extremes. *Nature Communications*, 14(1):2903, 2023.

[89] S. R. Kellert and E. O. Wilson. The biophilia hypothesis. In *Island Press*, 1995.

[90] S. Kaplan. The restorative benefits of nature: Toward an integrative framework. *Journal of environmental psychology*, 15(3):169–182, 1995.

[91] J. Appleton. *The experience of landscape*. Wiley Chichester, 1996.

[92] T. McKeon M. C. Kondo, J. M. Fluehr and C. C. Branas. Urban green space and its impact on human health. *International journal of environmental research and public health*, 15(3):445, 2018.

[93] A. Callaghan, G. McCombe, A. Harrold, C. McMeel, G. Mills, N. Moore-Cherry,W. Cullen, W. The impact of green spaces on mental health in urban settings: a scoping review. *Journal of Mental Health*, 30(2):179–193, 2021.

[94] G. MacKerron and S. Mourato. Happiness is greater in natural environments. *Global Environmental Change*, 23(5):992 – 1000, 2013.

[95] K. Hahn G. C. Daily G. N. Bratman, J. P. Hamilton and J. J. Gross. Nature experience reduces rumination and subgenual prefrontal cortex activation. *Proceedings of the national academy of sciences*, 112(28):8567–8572, 2015.

[96] T. Preis C. I. Seresinhe and H. S. Moat. Quantifying the impact of scenic environments on health. In *Nature Scientific Reports*, 2015.

[97] H. Sadatsafavi C. Fournier M. Shepley, N. Sachs and K. Peditto. The impact of green space on violent crime in urban environments: an evidence synthesis. *International journal of environmental research and public health*, 16(24):5119, 2019.

[98] M. G. Berman B. Cochran S. De Vries J. Flanders C. Folke H. Frumkin J. J. Gross T. Hartig G. N. Bratman, C. B. Anderson et al. Nature and mental health: An ecosystem service perspective. *Science advances*, 5(7):eaax0903, 2019.

[99] A. Cleary M. Droomers B. W. Wheeler D. Sinnett M.J. Nieuwenhuijsen R. F. Hunter, C. Cleland and M. Braubach. Environmental, health, wellbeing, social and equity effects of urban green space interventions: A meta-narrative evidence synthesis. *Environment international*, 130:104923, 2019.

[100] World Health Organisation. Urban green space interventions and health: A review of impacts and effectiveness. Available at: https://www.who.int/europe/publications/m/item/urban-green-space-and-health--intervention-impacts-and-effectiveness, 2017.

[101] European Commission, Directorate-General for Research and Innovation. Towards an EU research and innovation policy agenda for nature-based solutions & re-naturing cities - Final report of the Horizon 2020 expert group on 'Nature-based solutions and re-naturing cities'. Available at: https://data.europa.eu/doi/10.2777/479582, 2015.

[102] A. Milvoy and A. Roué-Le Gall. Aménager des espaces de jeux favorables à la santé. *La Santé en action*, 444:38–39, 2015.

[103] World Health Organization. Regional Office for Europe. Urban green spaces and health. a review of evidence., 2016.

[104] Natural England. Nature nearby: accessible natural greenspace guidance. Available at: http://www.ukmaburbanforum.co.uk/docunents/other/nature_nearby.pdf, 2010.

[105] Berlin Senate Department for Urban Development and Housing. Versorgung mit öffentlichen, wohnungsnahen grünanlagen - supply of public, green areas close to apartments. Available at: https://www.berlin.de/umweltatlas/_assets/nutzung/oeffentliche-gruenanlagen/de-texte/kc605_2020.pdf, 2020.

[106] C. Konijnendijk. The 3-30-300 rule for urban forestry and greener cities. *Biophilic Cities Journal*, 4(2):2, 2021.

[107] M. Cohen. A systematic review of urban sustainability assessment literature. *Sustainability*, 9(11):2048, 2017.

[108] M. Browning A. Rigolon and V. Jennings. Inequities in the quality of urban park systems: An environmental justice investigation of cities in the united states. *Landscape and Urban Planning*, 178:156–169, 2018.

[109] K. Oh and S. Jeong. Assessing the spatial distribution of urban parks using gis. *Landscape and urban planning*, 82(1-2):25–32, 2007.

[110] P. Fan J. Gao J. Zhang, W. Yue. Measuring the accessibility of public green spaces in urban areas using web map services. *Applied Geography*, 126:102381, 2021.

[111] K. M. Rahman and D. Zhang. Analyzing the level of accessibility of public urban green spaces to different socially vulnerable groups of people. *Sustainability*, 10(11):3917, 2018.

[112] G. Giuliani, E. Petri,E. Interwies, V. Vysna, Y. Guigoz, N. Ray and I. Dickie. Modelling accessibility to urban green areas using open earth observations data: A novel approach to support the urban sdg in four european cities. *Remote Sensing*, 13(3):422, 2021.

[113] G. Boeing, C. Higgs, S. Liu, B. Giles-Corti, J.F. Sallis, E. Cerin, M. Lowe, D. Adlakha, E. Hinckson, A. V. Moudon and others. Using open data and open-source software to develop spatial indicators of urban design and transport features for achieving healthy and sustainable cities. *The Lancet Global Health*, 10(6):e907–e918, 2022.

[114] G. Giuliani C. Chênes and N. Ray. Modelling physical accessibility to public green spaces in switzerland to support the sdg11. *Geomatics*, 1(4):383–398, 2021.

[115] E. D. Ekkel and S. de Vries. Nearby green space and human health: Evaluating accessibility metrics. *Landscape and urban planning*, 157:214–220, 2017.

[116] J. P. Hupy Y. Wang S. Huang, L. Tang and G. Shao. A commentary review on the use of normalized difference vegetation index (ndvi) in the era of popular remote sensing. *Journal of Forestry Research*, 32(1):1–6, 2021.

[117] M. Liang C. Sun J. Gao P. Xie L. Feng W. Xia H. Liu S. Ma B. Wu, X. Guo et al. Association of individual green space exposure with the incidence of asthma and allergic rhinitis: a systematic review and meta-analysis. *Environmental Science and Pollution Research*, 29(59):88461–88487, 2022.

[118] T. Christidis E. Lavigne E. M. Thomson D. L. Crouse, L. Pinault and P. J. Villeneuve. Residential greenness and indicators of stress and mental well-being in a canadian national-level survey. *Environmental research*, 192:110267, 2021.

[119] E. P. Barboza, M. Cirach, S. Khomenko, T. Iungman, N. Mueller, J. Barrera-Gómez, D. Rojas-Rueda, M. Kondo and M. J. Nieuwenhuijsen. Green space and mortality in european cities: a health impact assessment study. *The Lancet Planetary Health*, 5(10):e718–e730, 2021.

[120] J. Mcbride J. Yang, L. Zhao and P. Gong. Can you see green? assessing the visibility of urban forests in cities. *Landscape and Urban Planning*, 91(2):97–104, 2009.

[121] S. Gao H. Lin Y. Kang, F. Zhang and Y. Liu. A review of urban physical environment sensing using street view imagery in public health studies. *Annals of GIS*, 26(3):261–275, 2020.

[122] C. Zhou J. Chen and F. Li. Quantifying the green view indicator for assessing urban greening quality: An analysis based on internet-crawling street view data. *Ecological Indicators*, 113:106192, 2020.

[123] W. Li R. Ricard Q. Meng X. Li, C. Zhang and W. Zhang. Assessing street-level urban greenery using google street view and a modified green view index. *Urban Forestry & Urban Greening*, 14(3):675–685, 2015.

[124] D. Ki and S. Lee. Analyzing the effects of green view index of neighborhood streets on walking time using google street view and deep learning. *Landscape and Urban Planning*, 205:103920, 2021.

[125] W. Song B. Wu J. Zhou Y. Huang J. Wu F. Zhao S. Yu, B. Yu and W. Mao. View-based greenery: A three-dimensional assessment of city buildings' green visibility using floor green view index. *Landscape and Urban Planning*, 152:13–26, 2016.

[126] I. A. V. Sánchez and SM. Labib. Accessing eye-level greenness visibility from open-source street view images: A methodological development and implementation in multi-city and multi-country contexts. *Sustainable Cities and Society*, page 105262, 2024.

[127] V. Uscila M. Barrdahl A. Kulinkina B. Staatsen W. Swart H. Kruize I. Zurlyte M. Annerstedt Van Den Bosch, P. Mudu and A. I. Egorov. Development of an urban green space indicator and the public health rationale. *Scandinavian journal of public health*, 44(2):159–167, 2016.

[128] B. Chen, S. Wu, Y. Song, C. Webster, B. Xu and P. Gong. Contrasting inequality in human exposure to greenspace between cities of global north and global south. *Nature Communications*, 13(1):1–9, 2022.

[129] V. Heikinheimo, H. Tenkanen, C. Bergroth, O. Järv, T. Hiippala and T. Toivonen. Understanding the use of urban green spaces from user-generated geographic information. *Landscape and Urban Planning*, 201:103845, 2020.

[130] I. Caridi A. Salgado, Z. Yuan and M. C. González. Exposure to parks through the lens of urban mobility. *EPJ data science*, 11(1):42, 2022.

[131] P. Barrows M. Richardson M. Mears, P. Brindley and R. Maheswaran. Mapping urban greenspace use from mobile phone gps data. *Plos one*, 16(7):e0248622, 2021.

[132] O. Marquet G. Vich and C. Miralles-Guasch. Green exposure of walking routes and residential areas using smartphone tracking data and gis in a mediterranean city. *Urban Forestry & Urban Greening*, 40:275–285, 2019.

[133] S> L. Scott Z. S. Venter, V. Gundersen and D. N. Barton. Bias and precision of crowdsourced recreational activity data from strava. *Landscape and Urban Planning*, 232:104686, 2023.

[134] K. E. Watkins M. D. Garber and M. R. Kramer. Comparing bicyclists who use smartphone apps to record rides with those who do not: Implications for

representativeness and selection bias. *Journal of Transport & Health*, 15:100661, 2019.

[135] J. Singh S. S.Sehra and H. S. Rai. Assessment of openstreetmap data-a review. *arXiv preprint arXiv:1309.6608*, 2013.

[136] P. Mooney B. Ciepłuch and A. C. Winstanley. Building generic quality indicators for openstreetmap. *19th annual GIS Research UK (GISRUK)*, 2011.

[137] C. Keßler and R. T. A. De Groot. Trust as a proxy measure for the quality of volunteered geographic information in the case of openstreetmap. In *Geographic information science at the heart of Europe*, pages 21–37. Springer, 2013.

[138] R. Brunauer S. Gröchenig and K. Rehrl. Estimating completeness of vgi datasets by analyzing community activity over time periods. In *Connecting a digital Europe through location and place*, pages 3–18. Springer, 2014.

[139] C. Kunze R. Hecht and S. Hahmann. Measuring completeness of building footprints in openstreetmap over space and time. *ISPRS International Journal of Geo-Information*, 2(4):1066–1091, 2013.

[140] M. Haklay. How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and planning B: Planning and design*, 37(4):682–703, 2010.

[141] A. Zipf J. Jokar Arsanjani, P. Mooney and A. Schauss. Quality assessment of the contributed land use information from openstreetmap versus authoritative datasets. *OpenStreetMap in GIScience: experiences, research, and applications*, pages 37–58, 2015.

[142] M. G. Kendall. Rank correlation methods. *Griffin*, 1948.

[143] C. Gini. On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series*, 208(1):73–79, 1936.

[144] M. Lowe, J. F. Sallis, D. Salvo, E. Cerin, G. Boeing, C. Higgs, S. Liu, E. Hinckson, D. Adlakha, J. Arundel and others. A pathway to prioritizing and delivering healthy and sustainable cities. *Journal of City Climate Policy and Economy*, 1(1):111–123, 2022.

[145] D. Kalisch H. Wüstemann and J. Kolbe. Access to urban green space and environmental inequalities in germany. *Landscape and Urban Planning*, 164:124–131, 2017.

[146] P. Liu Y. Weng J. Wu, Y. Peng and J. Lin. Is the green inequality overestimated? quality reevaluation of green space accessibility. *Cities*, 130:103871, 2022.

[147] C. Albert C. Wen and C. Von Haaren. Equality in access to urban green spaces: A case study in hannover, germany, with a focus on the elderly population. *Urban Forestry & Urban Greening*, 55:126820, 2020.

[148] Y. Song, B. Chen, H.C. Ho, M. Kwan, D. Liu, F. Wang, J. Wang, J. Cai, X. Li, Y. Xu and others. Observed inequality in urban greenspace exposure in china. *Environment International*, 156:106778, 2021.

[149] D. Liu H. Zhao Y. Han, J. He and J. Huang. Inequality in urban green provision: A comparative study of large cities throughout the world. *Sustainable Cities and Society*, 89:104229, 2023.

[150] I. Anguelovski, J. JT. Connolly, H. Cole, M. Garcia-Lamarca, M. Triguero-Mas, F. Baró, N. Martin, D. Conesa, G. Shokry, C. P. Del Pulgar and others. Green gentrification in european and north american cities. *Nature communications*, 13(1):3816, 2022.

[151] A. Rigolon and J. Németh. Green gentrification or 'just green enough': Do park location, size and function affect whether a place gentrifies or not? *Urban Studies*, 57(2):402–420, 2020.

[152] I. Anguelovski,J. JT. Connolly, L. Masip, L. and H. Pearsall. Assessing green gentrification in historically disenfranchised neighborhoods: a longitudinal and spatial analysis of barcelona. *Urban geography*, 39(3):458–491, 2018.

[153] J. JT. Connolly I. Anguelovski A. T. A. Maia, F. Calcagni and J. Langemeyer. Hidden drivers of social injustice: uncovering unequal cultural ecosystem services behind green gentrification. *Environmental Science & Policy*, 112:254–263, 2020.

[154] H. Cole J. JT. Connolly L. Argüelles F. Baró S. Loveless C. Perez del Pulgar Frowein M. Garcia-Lamarca, I. Anguelovski and G. Shokry. Urban green boosterism and city affordability: For whom is the 'branded'green city? *Urban studies*, 58(1):90–112, 2021.

[155] A. Rigolon and T. Collins. The green gentrification cycle. *Urban Studies*, 60(4):770–785, 2023.

[156] M Yeh K. Chang, L. Wei and W. Peng. Discovering personalized routes from trajectories. In *Proceedings of the 3rd ACM SIGSPATIAL international workshop on location-based social networks*, pages 33–40, 2011.

[157] B. Zenker B. Ludwig and J. Schrader. Recommendation of personalized routes with public transport connections. In *International Conference on Intelligent Interactive Assistance and Mobile Multimedia Computing*, pages 97–107. Springer, 2009.

[158] K. Nellore and G. P. Hancke. A survey on urban traffic management system using wireless sensor networks. *Sensors*, 16(2):157, 2016.

[159] M. Grzenda T. Liebig, S. Peter and K. Junosza-Szaniawski. Dynamic Transfer Patterns for Fast Multi-modal Route Planning. In *Societal Geo-innovation: Selected papers of the 20th AGILE conference on Geographic Information Science*, pages 223–236, 04 2017.

[160] Google. Google maps.

[161] JL. I. Rizzi O. Strambi A. M. Larranaga, J. Arellana and H. B. B. Cybis. Using best–worst scaling to identify barriers to walkability: A study of porto alegre, brazil. *Transportation*, 46:2347–2379, 2019.

[162] M. Patil B. B. Majumdar, P. K. Sahu and N. Vendotti. Pedestrian satisfaction-based methodology for prioritization of critical sidewalk and crosswalk attributes influencing walkability. *Journal of Urban Planning and Development*, 147(3):04021032, 2021.

[163] S. Leonardi N. Distefano and N. G. Liotta. Walking for sustainable cities: Factors affecting users' willingness to walk. *Sustainability*, 15(7):5684, 2023.

[164] S. Park S. Kim and J. S. Lee. Meso-or micro-scale? environmental factors influencing pedestrian satisfaction. *Transportation Research Part D: Transport and Environment*, 30:10–20, 2014.

[165] A. CS. Mihiel G. Lovegrove P. Nunziante A. Montella, S. Chiaradonna and M. Rella Riccardi. Sustainable complete streets design criteria and case study in naples, italy. *Sustainability*, 14(20):13142, 2022.

[166] S. Amer-Yahia N. Golbandi R. Lempel M. De Choudhury, M. Feldman and C. Yu. Automatic construction of travel itineraries using social breadcrumbs. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, pages 35–44, 2010.

[167] S. NA. Van Sas A. El Ali and F. Nack. Photographer paths: sequence alignment of geotagged photos for exploration-based route planning. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 985–994, 2013.

[168] Y. Wang S. Wakamiya Y. Kawai Y. Zhang, P. Siriaraya and A. Jatowt. Walking down a different path: route recommendation based on visual and facility based diversity. In *Companion Proceedings of the the Web Conference 2018*, pages 171–174, 2018.

[169] YT. Huang W. H. Hsu AJ. Cheng, YY. Chen and HY. M. Liao. Personalized travel recommendation by mining people attributes from community-contributed photos. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 83–92, 2011.

[170] G. Irie T. Kurashima, T. Iwata and K. Fujimura. Travel route recommendation using geotags in photo sharing sites. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 579–588, 2010.

[171] Urban Atlas Land Cover/Land Use 2018 (vector), Europe, 6-yearly [Dataset]. Available at: https://doi.org/10.2909/fb4dffa1-6ceb-4cc0-8372-1ed354c285e6.

[172] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

[173] N. J. Nilsson P. E. Hart and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

[174] M. Batty. Big data, smart cities and city planning. *Dialogues in human geography*, 3(3):274–279, 2013.

Appendix A

Appendix to Chapter 2

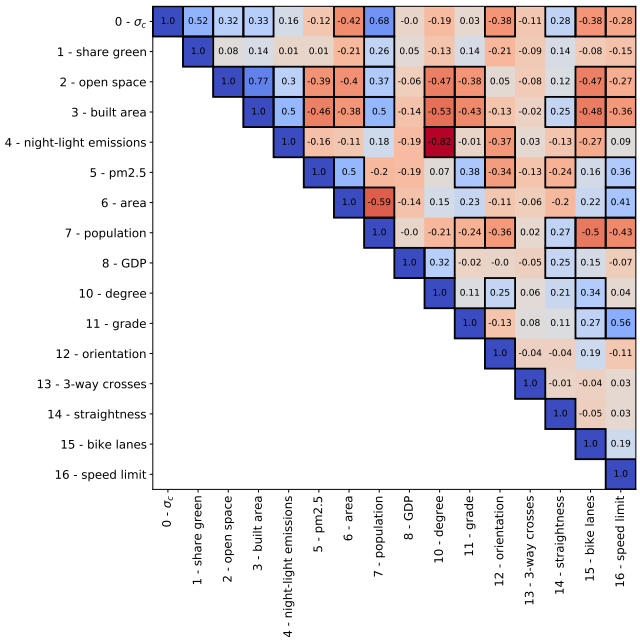

FIGURE A.1: **Correlations between σ_c and city-level features** Correlation matrix among city-level characteristics. The correlations are computed for the entire sample of 61 cities on z-score-transformed variables. Cells outlined in black indicate that the correlation is statistically different from 0 at a significance level of $\alpha = 0.05$.

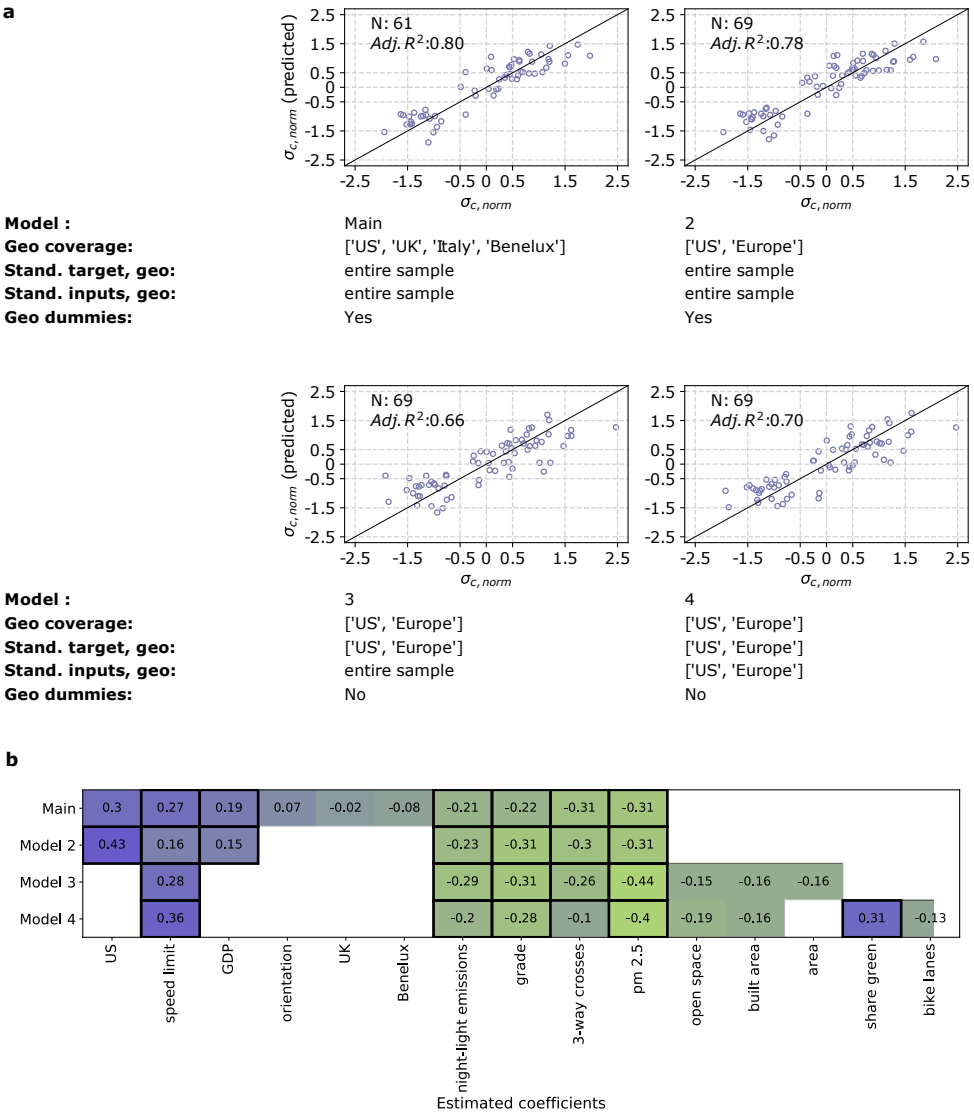

FIGURE A.2: **RQ1: sensitivity analysis** The panel provides a comparison between the preferred model (Main model) discussed in the main text and three additional models, included as sensitivity analysis. a) A summary of the characteristics of each model and scatter plots of (normalized) actual vs fitted values, for each model separately. All models share the linear formulation and the estimation technique (OLS) with the main one, but differ in terms of geographical coverage, standardization sample for target and input variables and the presence/absence of geographical dummies. b) The heatmap displaying the estimated coefficient for each model. Cells outlined in black correspond to coefficients statistically significant at 0.05 significance level.

Appendix B

Appendix to Chapter 3

Key	Value
amenity	grave_yard landuse
landuse	orchard
landuse	farmyard
natural	heath
landuse	cemetery
landuse	flowerbed
landuse	construction
natural	tree_row
amenity	school
landuse	greenfield
leisure	pitch
natural	fell
amenity	kindergarten
landuse	farmland
leisure	horse_riding
natural	wetland
amenity	education
landuse	vineyard
natural	scrub
landuse	allotments
landuse	farm
natural	moor

TABLE B.1: OSM *key* : *value* pairs used for the identification of the additional sources of green coverage

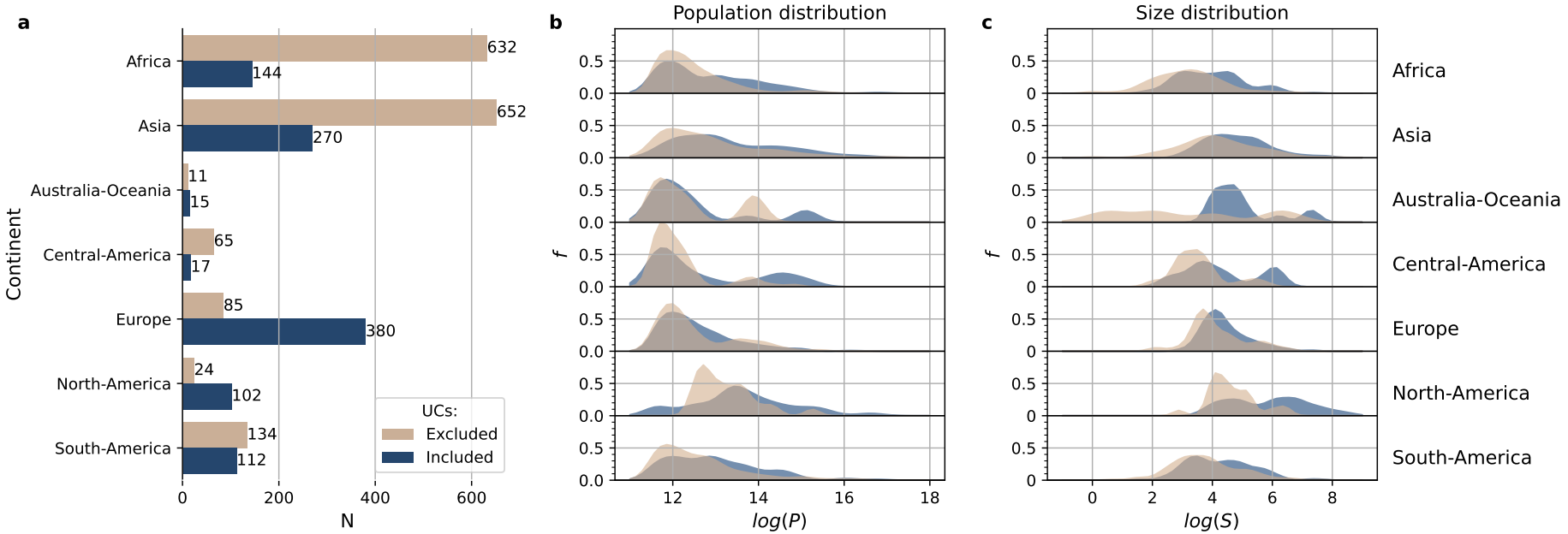

FIGURE B.1: **Characteristics of cities included and excluded from the final sample** a) The barplot depicts the number of cities included (in blue) and excluded (in pink) from the sample for each continent after applying the filtering procedure. b) The chart depicts the population distribution (in log) by continent for cities included (and in blue) from the sample and for those excluded from the sample (in pink). c) The chart depicts the area distribution (on a log scale) by continent for cities included (and in blue) from the sample and for those excluded from the sample (in pink). All cities in the Russian Federation have been attributed to Europe for this chart.

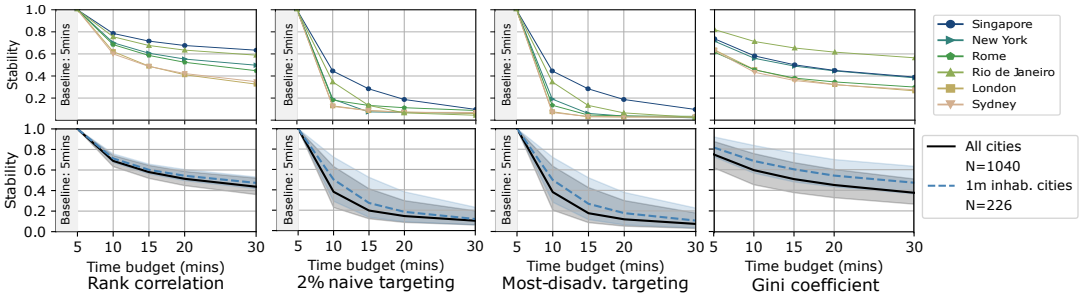

FIGURE B.2: **Stability of the exposure indicator to the time-budget** The top row depicts the exposure indicator's stability level to different time-budgets and according to different stability metrics for six cities across all continents. For the *Rank correlation*, *2% naive targeting*, and the *Most-disadvantaged targeting*, the comparison is provided with respect to the parametrization with time-budget equal to 5 mins. For the *Gini indicator*, the chart reports the indicator's value under several parametrizations. The bottom row reports the median value (solid line) and the IQR (shaded area) of the stability metrics for all cities in our sample (black) and for cities with more than 1 million inhabitants (blue).

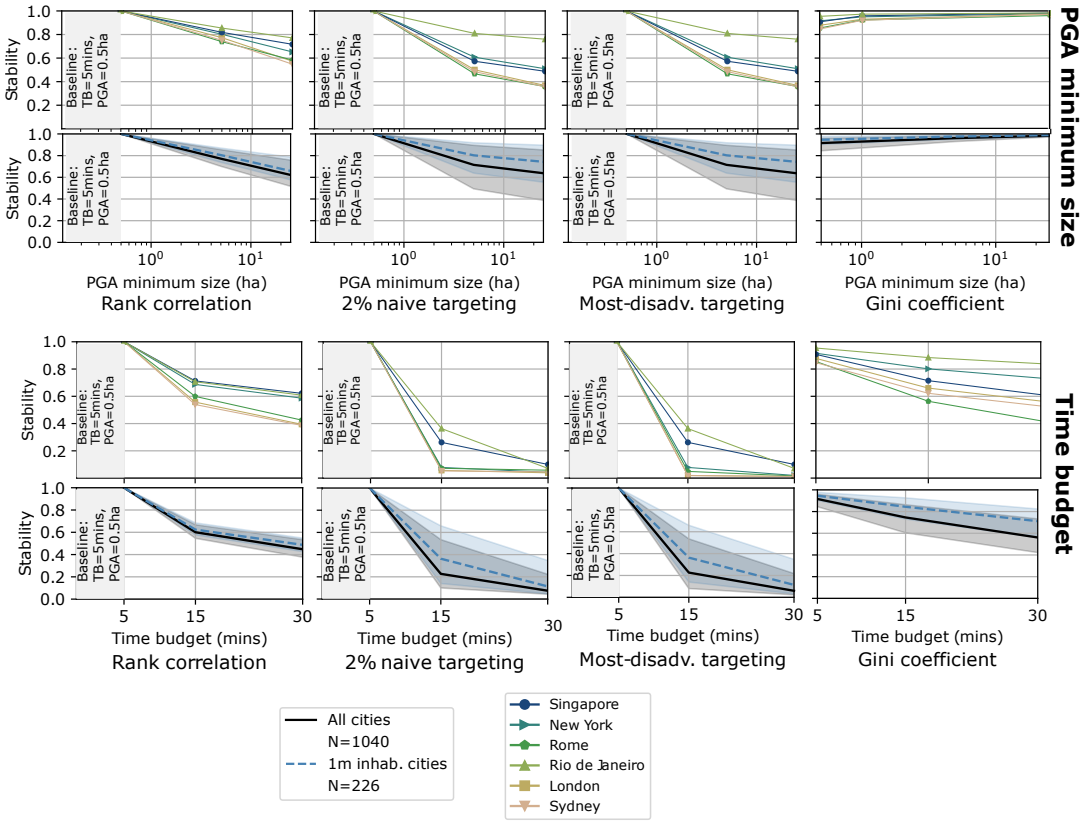

FIGURE B.3: **Stability of the per-person indicator to the minimum size of greenspaces and the time budget**. The panel depicts the level of stability of the per-person indicator to several minimum greenspaces' sizes (top sub-panel) and time budgets (bottom sub-panel). For both sub-panels, the top row displays the level of stability of the per-person indicator to the parameter and according to four stability metrics for six cities across all continents. For the *Rank correlation*, *2% naive targeting* and the *Most-disadvantaged targeting*, the comparison is provided with respect to the parametrization with time-budget equal to 5 mins and minimum greenspaces' size equal to 0.5 ha. For the *Gini indicator*, the chart reports the value of the indicator under several parametrizations; the bottom row reports the median value (solid line) and the IQR (shaded area) of the stability metrics for all cities in our sample (black) and for cities with more than 1 million inhabitants (blue).

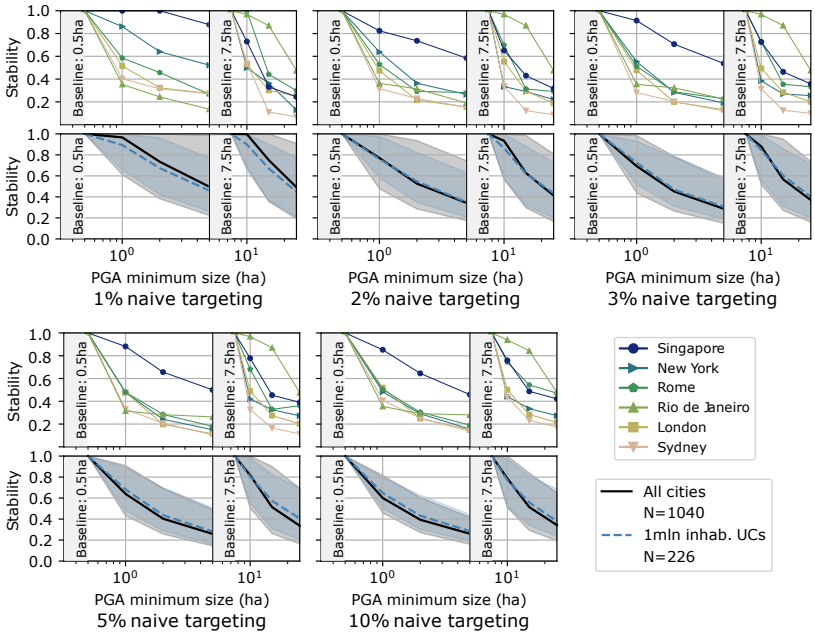

FIGURE B.4: **Stability of the minimum distance indicator to the minimum size of greenspaces: sensitivity analysis for the y% naive targeting strategy** The panel depicts the stability of the minimum distance indicator to the minimum PGA size under several naive targeting strategies [1%, 2%, 3%, 5%,10%], for six cities, the median (solid line) and IQR (shaded area) across all cities (black) and the median (solid line) and IQR (shaded area) for cities with more than 1 million inhabitants (blue)

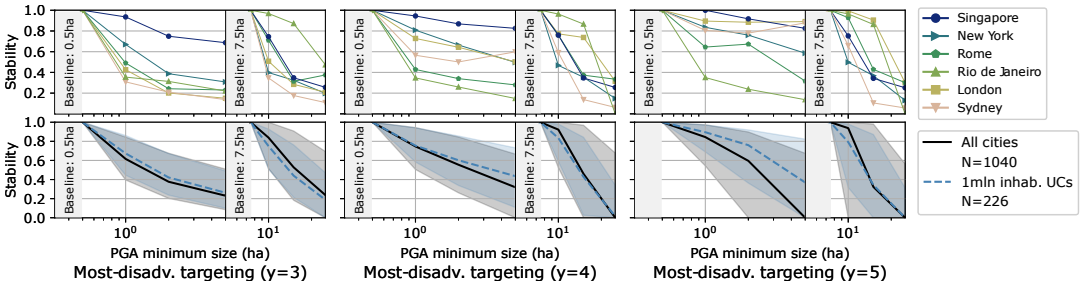

FIGURE B.5: **Stability of the minimum distance indicator to the minimum size of greenspaces: sensitivity analysis for the y-times most-disadvantaged targeting strategy** The panel depicts the stability of the minimum distance indicator to the minimum PGA size under several most-disadvantaged targeting strategies [3-times worse than the mean citizen, 4-times worse than mean citizen and 5 -times worse than mean citizen], for six cities, the median (solid line) and IQR (shaded area) across all cities (black) and the median (solid line) and IQR (shaded area) for cities with more than 1 million inhabitants (blue)

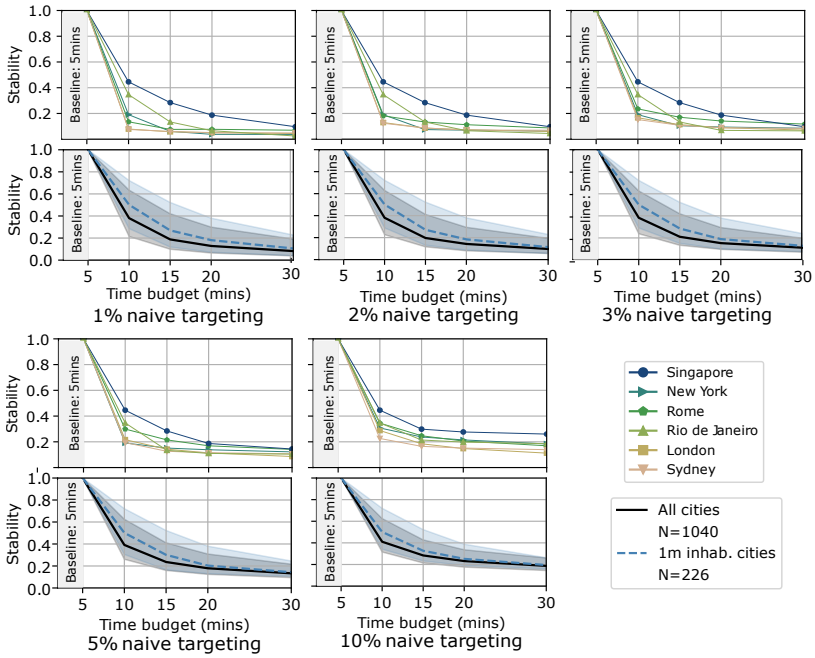

FIGURE B.6: **Stability of the exposure indicator to the time budget: sensitivity analysis for the y% naive targeting strategy** The panel depicts the stability of the exposure indicator to the time budget under several naive targeting strategies [1%, 2%, 3%, 5%,10%], for six cities, the median (solid line) and IQR (shaded area) across all cities (black) and the median (solid line) and IQR (shaded area) for cities with more than 1 million inhabitants (blue)

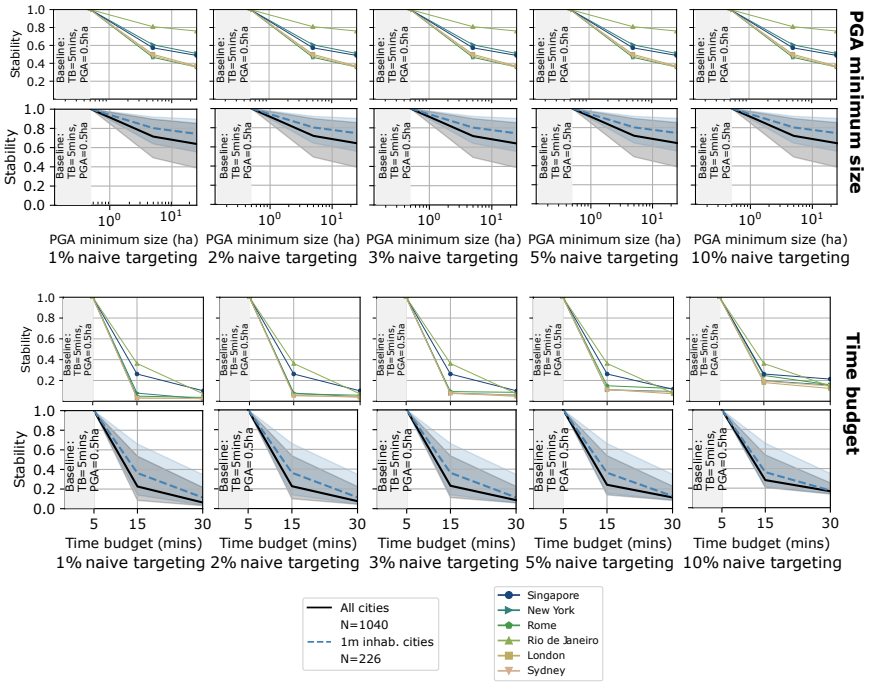

FIGURE B.7: **Stability of the per-person indicator to the minimum size of greenspaces and time budget: sensitivity analysis for the y% naive targeting strategy** The panel depicts the level of stability of the per-person indicator to several minimum PGA sizes (top sub-panel) and time-budgets (bottom sub-panel) under several naive targeting strategies [1%, 2%, 3%, 5%,10%], for six cities, the median (solid line) and IQR (shaded area) across all cities (black) and the median (solid line) and IQR (shaded area) for cities with more than 1 million inhabitants (blue).

Appendix C

Appendix to Chapter 4

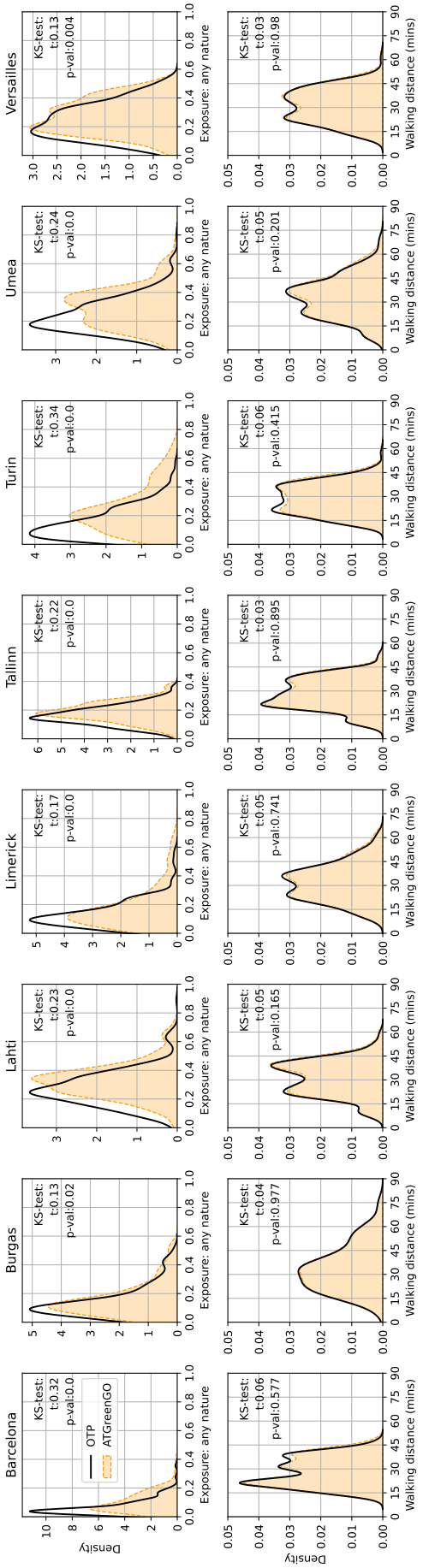

FIGURE C.1: **Performance of ATGreenGO optimised for the *any nature* feature** Distribution of walking duration and exposure to any nature on OTP shortest path (black) and ATGreenGO optimised for any nature exposure (colored filled area), by city. Results from a two-side Kolmogorov-Smirnov test are provided. The null hypothesis is that the two distributions are identical

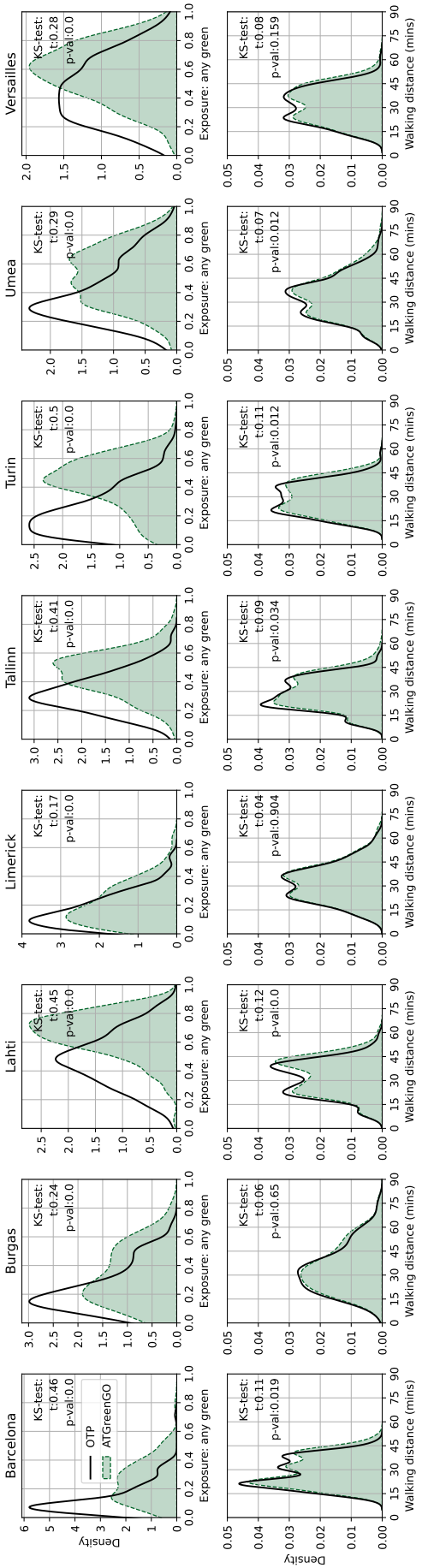

FIGURE C.2: **Performance of ATGreenGO optimised for the *any green* feature** Distribution of walking duration and exposure to any green on OTP shortest path (black) and ATGreenGO optimised for any green exposure (colored filled area), by city. Results from a two-side Kolmogorov-Smirnov test are provided. The null hypothesis is that the two distributions are identical

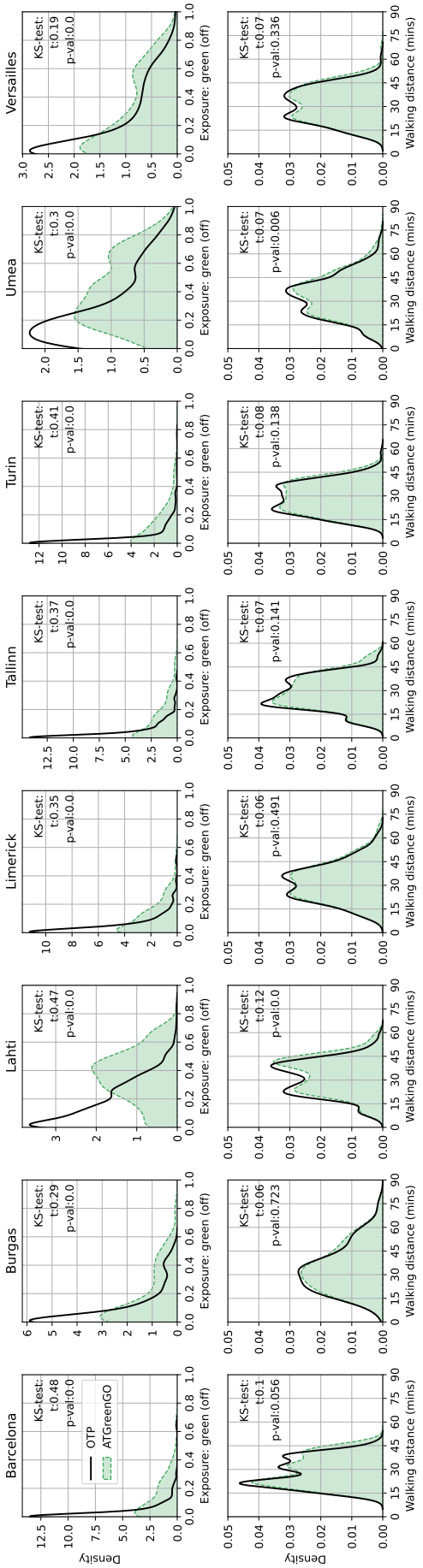

FIGURE C.3: **Performance of ATGreenGO optimised for the *green (off)* feature** Distribution of walking duration and exposure to green (off) on OTP shortest path (black) and ATGreenGO optimised for green (off) exposure (colored filled area), by city. Results from a two-side Kolmogorov-Smirnov test are provided. The null hypothesis is that the two distributions are identical

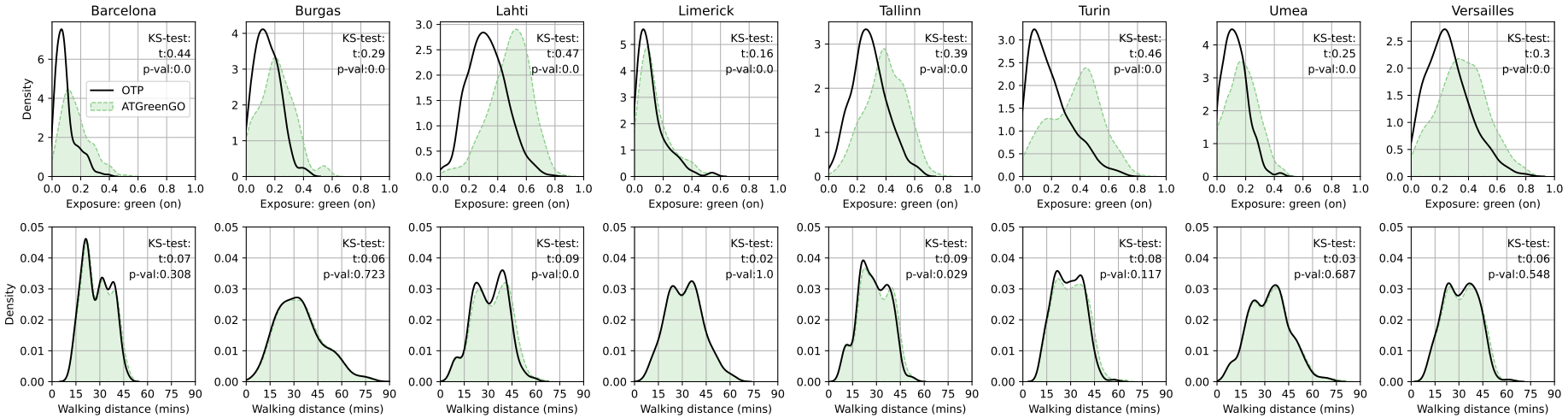

FIGURE C.4: **Performance of ATGreenGO optimised for the *green (on)* feature** Distribution of walking duration and exposure to green (on) on OTP shortest path (black) and ATGreenGO optimised for green (on) exposure (colored filled area), by city. Results from a two-side Kolmogorov-Smirnov test are provided. The null hypothesis is that the two distributions are identical

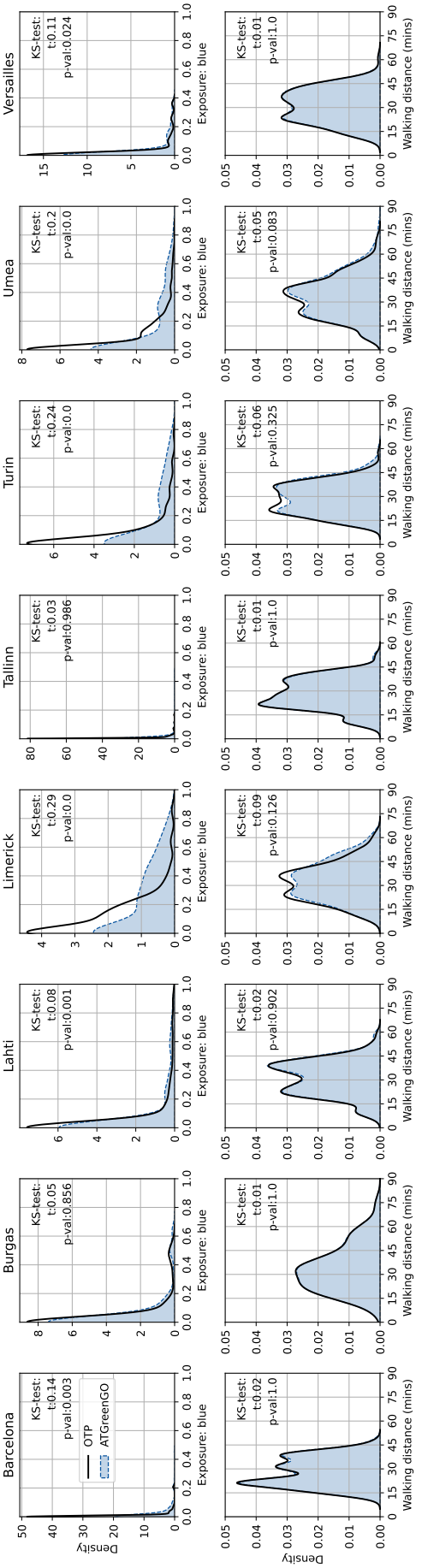

FIGURE C.5: **Performance of ATGreenGO optimised for the *blue* feature** Distribution of walking duration and exposure to blue on OTP shortest path (black) and ATGreenGO optimized for blue exposure (colored filled area), by city. Results from a two-side Kolmogorov-Smirnov test are provided. The null hypothesis is that the two distributions are identical