Battiston et al. *EPJ Data Science* (2023) 12:9
https://doi.org/10.1140/epjds/s13688-023-00385-7

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE

Open Access

Revealing the determinants of gender inequality in urban cycling with large-scale data

Alice Battiston[1*], Ludovico Napoli[4,2], Paolo Bajardi[3,4], André Panisson[3,4], Alan Perotti[3,4], Michael Szell[5,6,4] and Rossano Schifanella[1,4]

*Correspondence:
alice.battiston@unito.it
[1]University of Turin, Via Giuseppe Verdi, 8, 10124, Turin, Italy
Full list of author information is available at the end of the article

Abstract

Cycling is an outdoor activity with massive health benefits, and an effective solution for sustainable urban transport. Despite these benefits and the recent rising popularity of cycling, most countries still have a negligible uptake. This uptake is especially low for women: there is a largely unexplained, persistent gender gap in cycling. To understand the determinants of this gender gap in cycling at scale, here we use massive, automatically-collected data from the tracking application Strava on outdoor cycling for 61 cities across the United States, the United Kingdom, Italy and the Benelux area. While Strava data is particularly well-suited to describe the behavior of regular cyclists and its generalizability to occasional cyclists requires further investigation, the size of these data and their characteristics represent an unprecedented opportunity for the literature on cycling. Leveraging the associated gender and usage information, we first quantify the emerging gender gap in recreational cycling at city-level. A comparison of cycling rates of women across cities within similar geographical areas—where the penetration of Strava is assumed to be comparable—unveils a broad range of gender gaps. On a macroscopic level, we link this heterogeneity to a variety of urban indicators and provide evidence for traditional hypotheses on the determinants of the gender-cycling-gap. We find a positive association between female cycling rate and urban road safety. On a microscopic level, we identify female preferences for street-specific features in the city of New York. Assuming that the determinants of the gender-cycling-gap are similar across regular and occasional cyclists, our study suggests that enhancing the quality of the dedicated cycling infrastructure may be a way to make urban environments more accessible for women, thereby making urban transport more sustainable for everyone.

Keywords: Human mobility; Urban data science; Gender gap; Sustainable transport

1 Introduction

Cycling is an outdoor activity associated with many individual and societal benefits. From the individual perspective, cycling has a positive impact on both physical and mental health, with a strong link to improved cardio-respiratory fitness, decreased cardiovascular mortality risk, and reduced stress-levels [1–3]. From a societal viewpoint, cycling is an environmentally-friendly and highly economic commuting option, especially for typical urban trips [4]. Recently, the United Nation (UN) Sustainable Development Goals (SDG) identified it as a pivotal component of a sustainable urban-mobility system [5]. Interventions targeted at increasing the number of cyclists are recommended as one of the solutions against traffic congestion, increased emissions, poor air quality and road safety.

Despite these wide-ranging benefits, cycling is mostly a male-dominated activity with a large gap in participation rates between men and women. Cycling research and policy making that is mostly focused on improving mobility for the existing, dominant group, risks to ignore half of the population and sustainable mobility solutions for everybody [6, 7]. Data on the use of bike-sharing services in three large US cities (New York, Boston and Chicago) show that only one in four bicycle trips in the 4-year period between 2014 and 2018 was made by a woman; other modes of transport, however, do not display comparable trip-share gaps [8]. Similarly, in San Francisco, only 29% of cyclists are women [9]. Recent data for England show that on average, not only do men take more bicycle trips per week than women, but they also cover longer distances [10]. A few European countries however, such as Denmark, Germany and the Netherlands represent the main exception to this pattern, with women making up for more than 45% of all cyclists in these areas already in 2005 [11]. The evidence from this group of countries demonstrates that the reason for any kind of gender gap is not intrinsic but comes from place-specific barriers that need to be identified and, whenever possible, removed, if cycling should become a universal mode of transport.

The academic literature aimed at understanding the determinants of the gender gap in cycling links it on one hand to behavioral and psychological hypotheses. Women perceive cycling as a riskier activity compared to men, which would directly translate into a stronger preference for cycling infrastructure that is physically separated from motorized traffic [12–15]. On the other hand, physical route characteristics can play a role, for example in San Francisco, where women disfavor steep slopes, particularly for commuting [16]. In low-cycling contexts, women also report other deterring factors, such as an aversion for long distances and poor weather conditions, and a generally lower confidence in their cycling skills [17, 18]. Differences in preferences are typically stronger among occasional or non-cyclists than among regular cyclists [12], thus suggesting that policies targeting women are particularly needed to increase cycling uptake. The main limitation of these studies is that they are mostly conducted via surveys or experiments with typically low sample sizes and/or a limited geographical breadth, and therefore low statistical explanatory power—especially for the large number of possible confounders.

Recently, the emergence of new technologies for cycle-tracking and online-based services (e.g. bike-sharing) generated an unprecedented stream of automatically collected data on cycling behavior, which enlarge the potential for research in this area. In this context, data from bike-sharing services have been used to study whether interventions to the bike-sharing facilities impacted men and women differently in the city of New York [19] and, more generally, to study factors affecting the demand for these types of services [20]. Data from Strava Metro, a service provided by the sport-tracking application Strava, have been used to study exposure to air pollution for different groups of cyclists in the city of Glasgow [21] and cycling patterns and trends for the city of Johannesburg [22]. A few pioneering works have used GPS-based data to study route choices for different demographic groups in the city of San Francisco and Atlanta [14, 16].

In this study, we contribute to this strand of literature and use data from Strava to investigate the determinants of the gender gap in recreational cycling at a larger scale. With about 36 million users (2018 data) over 195 countries, Strava represents an unique data source on cycling-related behaviour [23], both in terms of the number of cyclists involved and the extent of the geographical coverage with a methodologically homogeneous data collection. For this study, we collect and use data for over 60 cities in four geographical areas across the United States and Europe, to explore the gender-cycling-gap at two different levels. First, we exploit the heterogeneity in the gender gap across the various cities in our dataset to challenge traditional hypotheses from the literature on the determinants of the gender gap in cycling. In particular, we study the strength of association between the gender gap in cycling measured at the level of urban centers and a set of urban indicators, spanning from morphological characteristics of the cities to safety indicators capturing the prevalence of cycleways and streets with low-speed limit in the road network. In the second part of the study, we move the analysis from a macro to a micro level. Here, focusing on the city of New York, we model the gender-cycling-gap measured at street-level in terms of specific urban features. By using logistic regression analysis, we investigate the association between the presence of dedicated cycling infrastructure and the volume of female cyclists on the street relative to men. The results indicate that streets with cycling infrastructures, particularly those ensuring the presence of physical separation for motorized traffic, are associated to a more balanced gender ratio, suggesting a way for policy makers to intervene to make urban environments more accessible for women.

2 Results

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

We use Strava data to measure the gender gap in recreational cycling in 61 urban centers across four geographical areas: United States, United Kingdom, Italy and Benelux. Strava is an Internet service for tracking human exercise that relies on GPS data. 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 [23]. Of these, 284 millions were cycling-activities (47%), and approximately one in five of cycling-uploads were by women (50 million) [23]. Tracking of commuting is growing in popularity on Strava [23], however the majority of uploads refers to recreational and athletic cycling. The raw data consist of a collection of Strava segments, with information on users training on these from the associated leaderboards. The data were processed to map gender and usage information from Strava segments to a network-based definition of streets and then aggregated for the entire city, following the pipeline described in the Methods.

For each city c, we define the gender-cycling-gap as the ratio σ_c between the total kilometers travelled by female cyclists and the overall kilometers 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. 1 provides an overview of

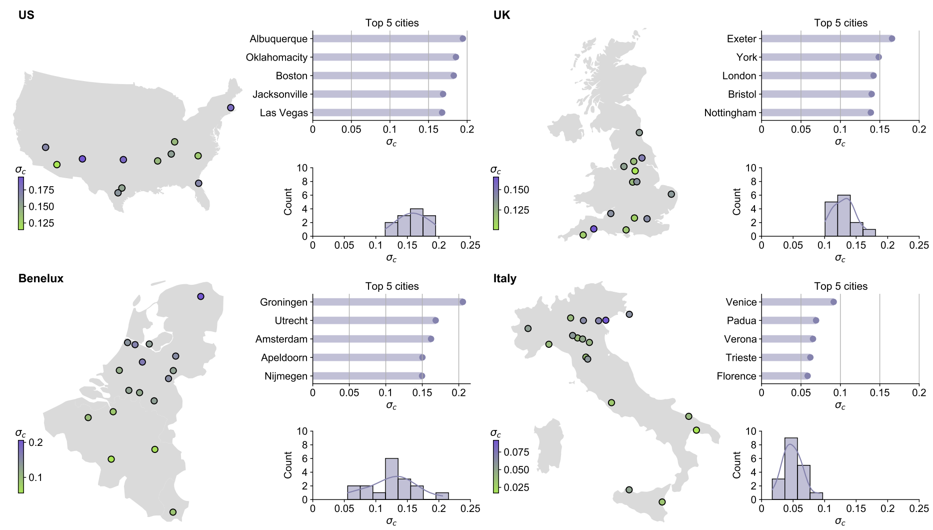

Figure 1 Gender gap in recreational cycling in Strava: overview of cities included in the study. For each of the four geographical areas covered by the study, the figure depicts: the location of cities included in the analysis, the value of the female cycling rate σ_c for the five cities displaying the highest σ_c and the distribution of σ_c in the geographical area

the cities included in the study, showing the five urban centers associated with the highest σ_c for each area, as well as the location and the distribution of σ_c of all covered cities (the full ranking is provided in the Supplementary Information (SI) in Additional file 1). 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 homogeneous geographical areas, we observe a 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 centers in the north-east displaying a lower gender gap than cities in the south and north-west. 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 noteworthy that the gender-gap measured using data from Strava may differ from official metrics on urban cycling provided from local and national administrations. This is the case for instance for cities in the Netherlands, where according to data by national authorities, men and women have a similar cycling uptake [24]. This discrepancy is likely due to the fact that Strava is mostly used for recreational purposes and as such may penetrate differently among the two groups of users (men and women). Furthermore, the degree of penetration may not be homogeneous across different geographical areas. For instance, data on the use of Strava indicates different usage patterns and adoption rates in the United States compared to other countries [25]. Our assumption throughout the study, is that the degree of penetration of the fitness app Strava for men and women is similar for cities within the same geographical area [*homogeneity assumption*]. While we do not expect the cycling behaviour tracked in Strava to be representative of the cycling behaviour of the overall population, the homogeneity assumption guarantees that the observed variation in σ_c for cities within the same geographical area is not linked to gender patterns in the app usage.

2.2 Cross-city analysis of the gender gap

The survey-based literature on the gender gap in cycling suggests that women are more-risk averse, which would result in a lower cycling rate than men in environments perceived as risky [12]. Following this hypothesis, we investigate the association between the gender ratio σ_c and two indicators of urban road safety, constructed using OpenStreetMap (OSM) data [27]. 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 equal or lower than 20 mi/h or 30 km/h. Both metrics are weighted using the length of each street. Figure 2 reports the scatter plots between the 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. For cities in Italy and in the United States, in contrast, both correlations are not statistically different from 0 (at neither a significance level of 0.05 nor 0.1). This lack of significant correlations may be due to the lower degree of development of dedicated cycling infrastructure in these areas, as opposed to cities in Benelux. This is the case, for

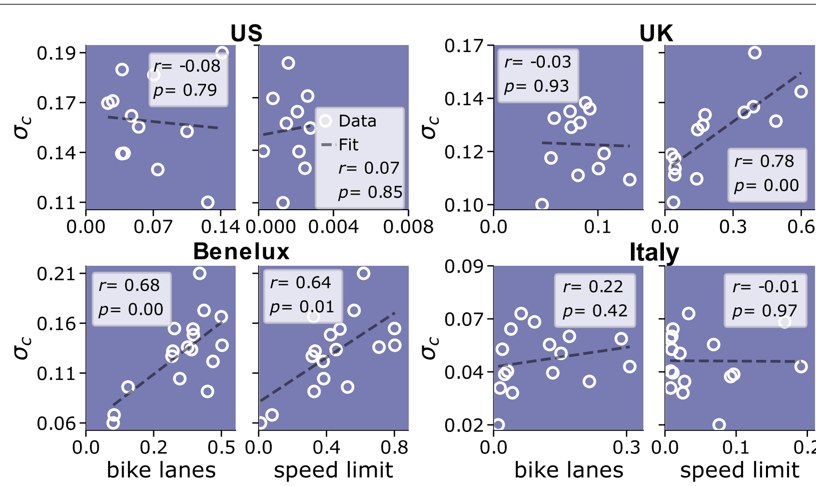

Figure 2 Correlations between gender ratio and urban road safety indicators. The scatter plots show the correlations between two urban road safety indicators and the gender ratio σ_c, for cities in the four geographical areas separately. For each area, outliers to the three distributions of σ_c, *bike lanes* and *speed limit* were identified using the IQR Score method [26] 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 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 Methods

instance, if a certain proportion of streets must be equipped with a cycleway in order for the city to be perceived as a safe cycling environment by women. In cities with a low degree of cycling-dedicated infrastructural development, the observed level of σ_c might be dependent on other concurrent factors linked—for instance—to the urban structure or the economic development of the area. Although limited to specific geographical areas, the positive correlations suggest an association between the degree of road safety and σ_c, thus supporting the hypothesis that low levels of women engagement with cycling may be explained by a greater concern for safety compared to men.

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 characterization 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 1. The correlation matrix of the indicators across the entire sample is provided in Fig. S3 in the SI. We include geographical dummies for the macro areas to account for different penetration levels of Strava worldwide.

Coefficients (and 95% confidence intervals) of a linear regression model estimated via Ordinary Least Squares (OLS) are shown in Fig. 3, with statistically significant coefficients at 0.05 level (two-tailed test) pictured in purple. The pipeline for the selection of the model is provided in the 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 means that urban centers with a relatively wider low-speed zone typically present a more balance 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 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 view-point, 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 [28]. 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 [29]. 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 [16] 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 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

Table 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	[40]
	open space	Percentage of open-spaces within the spatial domain of the Urban Centre. Ranging between 0–100	[40]
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	[40]
	light emissions	Average night time night-light emission calculated within the Urban Centre spatial domain. Expressed in nano-watt per steradian per square centimetre	[40]
	pm2.5	Total concentration of PM2.5 for reference epoch 2014, calculated over the Urban Centre. Expressed in μg/m^3	[40]
SED	area	Area of the spatial domain of the Urban Centre. Expressed in square meters	[40]
	population	Population density within the spatial domain of the Urban Centre	[40]*
	GDP	GDP per capita for year 2015 within the Urban Centre. Expressed in US dollars	[40]*
M	degree	Average node degree of street network within the spatial domain of the Urban Centre	[28]
	grade	Average absolute inclination of streets within the spatial domain of the Urban Centre. Expressed in percentage	[28]
	orientation	Orientation order of street network bearings within the spatial domain of the Urban Centre	[28]
	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	[28]
	straightness	Ratio of straightline distances to street lengths for streets in the street network within the spatial domain of the Urban Area	[28]
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 30 km/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.

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, we estimate three additional models where we adopt different strategies to account for the different levels of penetration of Strava worldwide. These strategies are described in the SI and differ in terms of: geographical coverage, specification of the geographical dummies and standardization of the input and target variables

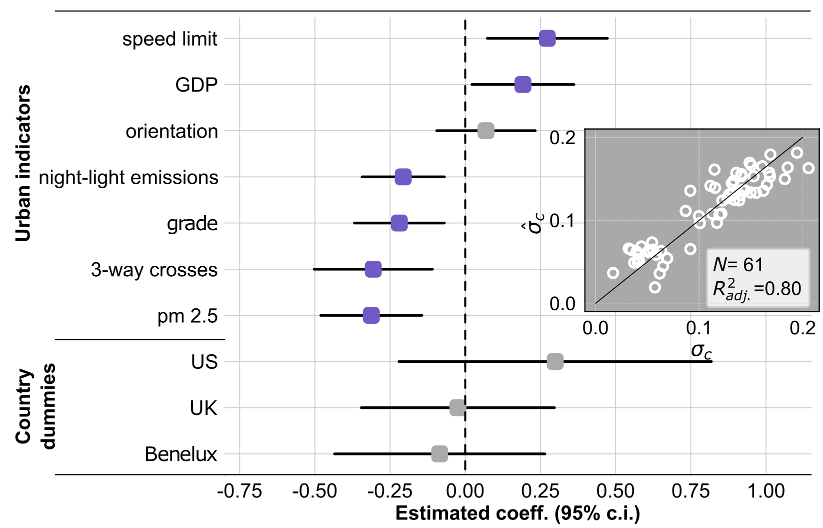

Figure 3 Regression analysis. The main plot shows the estimated coefficients (square markers) with 95% confidence intervals (black lines) for the final set of regressors included in the model. Model estimated via Ordinary Least Square, selection performed via exhaustive search. Selection criterion: Akaike Information Criterion (AIC). Statistically significant coefficient at 0.05 significance level are pictured in purple. The scatter plot displays the observed σ_c vs the fitted σ_c

(Fig. S4 in SI). The results are largely consistent with the preferred specification provided here.

2.3 Street-level analysis of the gender gap

2.3.1 New York City as a case study

The results in the previous section show that aggregated urban features model well the heterogeneity of the gender gap in cycling observed across different cities. They also confirm and provide quantitative support to traditional hypotheses from the literature, which are typically grounded on small-sample survey-based analyses. Though informative and affirmative, the previous analysis leaves open the question: Where exactly do women prefer to cycle? Also, which concrete interventions could policy makers implement to enhance cycling for women?

To answer these questions, we shift the focus from a macro-level comparison across cities, to a micro-level setting where the unit of analysis are streets within one city as opposed to the entire city itself. This shift in perspective allows us to examine the preferences of women for street-level characteristics in greater detail, thus identifying potential targets for interventions by policy makers. Among the available cities, we select as a case study the city of New York, whose large collection of administrative datasets represents an opportunity to enrich the analysis with data not otherwise available from OSM only. In particular, using OSM data, we are able to characterize each street in our sample with information on: the presence (or absence) of a protected (or unprotected) cycleway, the presence of public lighting, the type of surface (paved vs unpaved), whether the street is close to a park or to a coastline. The administrative data are instead used to measure the number of crashes (any type of vehicles or bicycle-related only) on the street (normalized by the street length) and to associate each street to a neighborhood. Finally, to proxy for traffic flow, we compute the normalized edge betweenness [30] of each street in the street

Table 2 Street-level indicators for the city of New York: description and source

Variable name	Description	Data source
Unprotected cycleway	Dummy for presence of a shared or unprotected bike-lane	OSM
Protected cycleway	Dummy for either the presence of a protected bike-lane or streets with no vehicles	OSM
Public lighting[1]	Dummy for the presence of public lighting	OSM
Unpaved surface	Dummy for unpaved surface	OSM
Park proximity	Dummy for streets next to a park (within 15 meters)	OSM*
Any-vehicle crashes	Number of crashes involving any type of vehicles per 10 m of street length	NYC*
Bike crashes	Number of bicycle crashes per 10 m of street length	NYC*
{borough_name}	Dummy for boroughs (baseline: Manhattan)	NYC
Coast proximity	Dummy for street next to the river coast	OSM*
Edge betweenness	Edge betweenness of the streets computed for streets within the largest component of the street network	

[1] Information on the presence of public lighting is very sparse in OSM. In case of missing information, we assumed that the public lighting is available.
*Indicates that the data from the original data sources required specific preprocessing (e.g. for normalization) described in the Methods.

network. The edge betweenness is a network centrality measure capturing the number of the shortest paths that go through an edge in the network. A summary of all features is provided in Table 2. As for 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, and call this metric σ_s. The indicator σ_s is a direct street-level extension of σ_c—indeed σ_c can be constructed averaging over σ_s with weights equal to the product between the length of each street and the total number of cyclists on it. The larger σ_s the greater are womens' 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 for 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 (described in the SI). 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 (Fig. 4). Stratifying the distribution by protection level of the street ('No cycleway', 'Unprotected cycleway', 'Protected cycleway'), we see that streets with no forms of dedicated-infrastructure are typically associated with lower σ_s than streets with either protected or unprotected cycleway: the median value of σ_s for streets with no cycleway roughly corresponding to the 25th percentile of both the distributions of streets with protected or unprotected cycleways. This descriptive evidence provides a first indication that streets with some form of cycling infrastructure are typically used more intensively by women than streets with no dedicated infrastructure at all. To delve deeper into women's preferences for dedicated cycling infrastructure, we study the degree of association between the presence of protected and unprotected cycleways and σ_s by means of a logistic regression analysis. We classify streets into two classes, *Low* and *High*, corresponding to the bottom, and top 33% of the distribution of σ_s and estimate

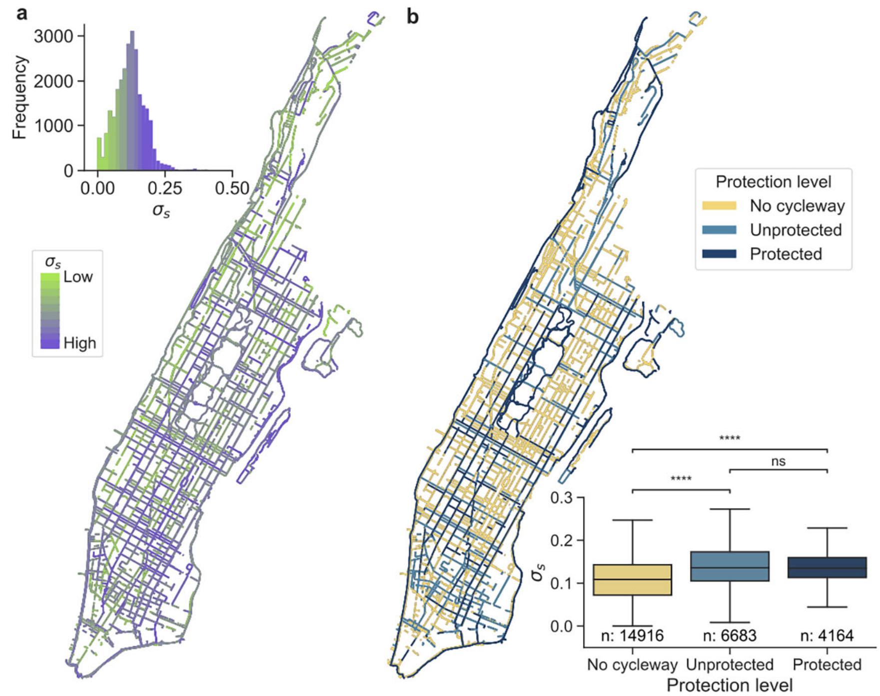

Figure 4 The case of New York City: streets characteristics and σ_s. **a**) The map displays streets in the borough of Manhattan included in the final sample. A 10-quantile color scheme has been used for the value of σ_s. The inset is the distribution of σ_s for streets included in the final sample (computed 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'

the Odds Ratios (OR) via multivariate logistic regression. To check the robustness of the results, the analysis is repeated choosing different thresholds α (0.25 and 0.40, instead of 0.33) for the classification. The results (presented in Fig. 5) are consistent across sample specifications, with generally slightly larger estimates on more extreme samples (lower values of the threshold α).

The main result pertains to the role of dedicated cycling infrastructure. With an estimated OR of around 4.08 (95% confidence interval: $[3.67, 4.54]$), the analysis indicates that the odds to be classified *High* are more than four times greater for protected cycleways than for streets with no cycleway (used as baseline). This result is largely in line with the survey-based literature on the gender-cycling-gap, according to which women would favor physical separation more than men [12–14]. Though 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 either budget or physical constraints, the use of shared unprotected cycleways would still be a way to make the urban environment more accessible for female cyclists. In light of recent findings [31, 32] which suggest that unprotected cycleways would not enhance the degree of safety of the road-network for cyclists, our results suggest that subjective safety may matter more than objective safety. In terms of other control variables, in line with the assumption that women favor more quiet streets, we estimate an OR below 1 for our

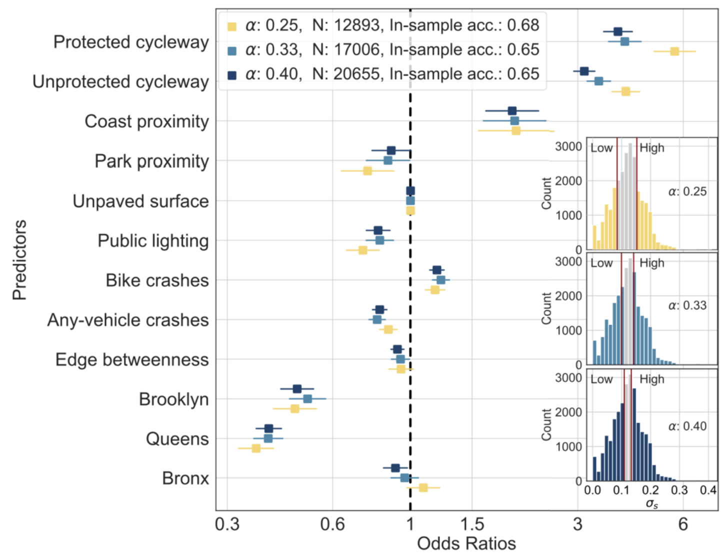

Figure 5 Odds Ratios of multivariate logistic regressions, for several levels of the threshold α **a**) the chart presents estimated ORs for a multivariate logistic regression where the target variable is the binarized σ_s and the predictors are listed in Table 2. The squared dots are the point estimates. 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)

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 to be the effect of reverse causality: a more balanced gender ratio is typically associated with a larger volume of cyclists, with an increased likelihood of bicycle crashes. The two dummies on coast and park proximity, here inserted as a proxy 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 being only statistically significant at 0.05 level for $\alpha = 0.25$). On one hand, the reason for the high coast proximity value is due to the morphology of the city of New York and the presence of a long protected cycleway along the coastline of Manhattan acting as an attractive infrastructure and impacting nearby streets too, with many cyclists riding through to reach it. On the other hand, 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 very sparse in OSM and particularly for New York City (we assume public lighting to be absent only whenever 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 hard to generalize 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).

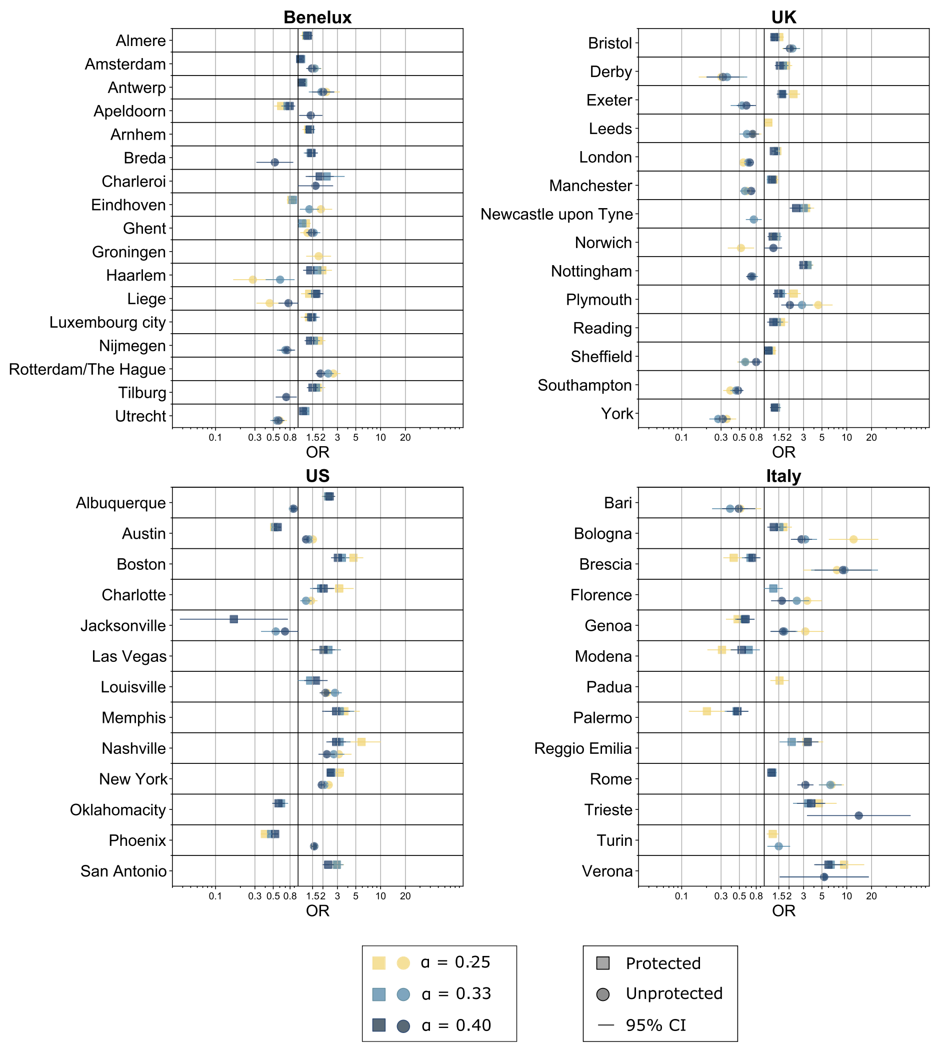

Figure 6 Odd Ratios of minimal logistic models, for all cities in the sample. The charts present estimated ORs of minimal logistic regressions, where the target variable is the binarized σ_s regressed against a categorical variable for the protection level of the street (three classes: *No protection* (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 points 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 are 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

2.3.2 *Generalization to other cities*

In the previous section, the city of New York was chosen as a case study due to the availability of additional street-level data from administrative sources. Furthermore, with its generally well-known city structure, New York represents a perfect testing environment for urban studies. In this section, we investigate to what degree the previous results on the role of protected and unprotected cycleways can be generalized to other urban environments. For all cities in our sample separately, we compute the 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 the city of New York. Results are reported in Fig. 6, where only ORs statistically different from 1 (significance level of 0.05) are pictured. 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 belong to the *high* gender ratio class, as we observe ORs above 1 for most cities in our sample. By 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 to unprotected ones is likely to depend on many additional features of the street, necessarily not encoded in this minimal model.

3 Discussion

In this study, we investigated the determinants of the gender-cycling-gap using data for over 60 cities in Europe and the United States. Unlike the vast majority of previous analyses that used survey-based data, we leveraged large automatically collected data from the online sport-tracking application Strava. While Strava represents a unique data source on cycling-related behavior—it was used by approximately 36 million users over 195 countries in 2018 [23]—it is noteworthy that it is better suited to describe the behavior of regular cyclists than occasional ones. As such the sample of this study might differ from other analyses surveying across different groups (regular cyclists, occasional cyclists, no-cyclists). Furthermore, most of the activities tracked in Strava are recreational as opposed to commuting.

In the first part of the study, we related female cycling rates in different European and American cities to macro city-level characteristics. The analysis was conducted controlling for the macro geographical area of the city, to control for differences in the penetration rate of the app among men and women across areas. We found evidence for traditional hypotheses which link the observed gender gap in cycling to gender-specific preferences on road safety. Additionally, we found higher female cycling rates in flatter than in hillier cities, also in line with the literature [16]. This is an interesting result as there may be structural, morphological or cultural [33] constraints for specific places where the cycling uptake is harder to increase for women. For urban planning, this result suggests that ad-hoc infrastructural interventions such as the provision of cycleways or the enlargement of the low-speed limit zones could have limited efficacy in these contexts and may require concurrent behavioral incentives, for instance to expand the adoption of e-bikes. A novel result concerns the strong association between the gender-cycling-gap and the air quality of a city, which however requires further research within a (quasi-)experimental setting.

In the second part of the study, we shifted the focus from a macro comparison across cities to a micro-level analysis, at the level of single streets. If the first analysis successfully provided evidence for and expanded existing hypotheses (further validating our data as a reliable source on cycling behavior), the second aims 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 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 [12–14]. We also showed that this result generalizes to most of the cities in our sample. 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 [31, 32], our result suggests that the perceived degree of safety may induce women to cycle more than the actual degree of safety in specific environments. This result, however, doesn't generalize to all cities in our study, suggesting that not all interventions in this sense are equally effective. Therefore, in contexts where no physical separation is possible (for instance for space or budget constraints), the provision of shared cycleway may still act as a way to make to urban environment perceived as more accessible by women in certain contexts. However, given that the increase in the perception of safety induced by this type of infrastructure may not always translate into a lower risk cycling environment, the planning of this type of infrastructure should be evaluated carefully by city planners, for instance favoring specific solutions associated to greater safety levels.

Overall our study validated survey-based results quantitatively using unprecedentedly large-scale automatically collected data. As previously discussed, Strava is among the major applications for sport tracking and as such, reliable information on cycling behavior for regular cyclists. The main limitation of our study pertains to the representativeness of Strava users and the purposes of Strava trips. For example, having a considerable gender gap in the Netherlands (Fig. 1), contrary to expectations [11], the Strava data are clearly not representative, and neither users nor purposes of use can be clearly inferred. We therefore stress that Strava does primarily reflect recreational cycling, and that its ability to describe other cycling behaviors should be explored with richer data sets or qualitative methods. More generally, while the use of large automatically-collected data allows us to explore phenomena at an unprecedented scale, the absence of a data collection design step might result in the inclusion of biases. The main source of bias in our analysis relates to the penetration rate of Strava, which 1—might differ across geographical areas, 2—might differ between men and women, within the same geographical areas. To account for these potential biases, we adopted two mitigation strategies. First, in the analysis at city-level, we included geographical dummies to control for differences in the penetration level of the app for the two genders depending on the geographical area of the city. The main assumption we relied upon is therefore that cities within the same geographical area have a similar bias in the use of the app among men and women. We opted for the use of geographical dummies since the number of cities in our sample was insufficient to opt for a fully interacted model. In addition, a sensitivity analysis is provided in the SI, where other mitigation strategies to control for this bias are adopted leading to highly comparable results. Second, in the analysis at street-level, the imbalance in the proportion of Strava users is taken into consideration by using the lowest and top quantiles of the gender ratio distribution to identify the 'high gender-ratio' vs 'low gender-ratio' streets rather than setting a fixed cut-off value. A second limitation of the Strava data set is the inability to extract the potentially useful information of cyclist volumes [34], 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.

It is unclear to which extent our results can be generalized to cycling for purposes other than recreational, such as transport, and to less-skilled cyclists (occasional cyclists and not cyclists). It is therefore important to find data sources that are able to reliably distinguish between such purposes and users, since gender-based constraints can differ between these categories [35]. However, since the survey-based academic literature on gender-cycling-gap indicates that cycling preferences differ less among regular cyclists than among occasional ones [12, 36], 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. Another limitation concerns the cross-sectional nature of the available cycling data. The absence of a longitudinal dimension limited the extent to which temporal variations could be analyzed in the data, hindering the use of policy-evaluation statistical tools such as diff-in-diff techniques to evaluate casual 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 [33, 34], to other environmental factors and harassment by motorists [35, 37]. 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 [38, 39]. 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 [7].

4 Methods

4.1 Strava data on recreational cycling: data collection and processing

Data collection Raw Strava data consist of a collection of Strava segments for 62 cities located in four geographical areas: the United States, United Kingdom, Benelux (Belgium, Netherlands and Luxembourg) and Italy. For the sensitivity analysis only, the dataset was extended to include 8 additional cities across other European countries. A Strava segment is a single portion of a road or a trail upon which users of Strava compete by recording their times. The performance of a user training upon a segment is automatically recorded into its leaderboard, which in turn provides a picture of the characteristics of users cycling on a specific trail. Each raw data record consists of geographic information about the segment in the form of a linestring of lat-long coordinates, enriched with the following statistics extracted from the associated leaderboard:

1. the total number of unique cyclists training on the segment. This information corresponds to the sum of the length of the female and male leaderboards (it should be noted that—irrespective of the number of training performed on the segment—each cyclist is only included once in the corresponding leaderboard, according to their best performance on the segment);
2. the gender split of users training on the segment, in terms of the length of the female and male leaderboards respectively (corresponding to the number of unique female and male cyclists training on the segment).

The data collection comprised of two phases, both undertaken in November 2018. In the first phase, we collected the whole corpus of segments ($\sim$16.4 million) available at the time through the Strava API. This step provided us with a the geographical information related to each segment, identified by its ID. The second phase consisted in the collection of summary statistics from the female and male leaderboards (with two separate queries for the same ID) associated with each segment. In particular, given a city, we made queries for the leaderboards of all the segments whose geometry is contained for at least 75% within the city boundary. The SI provides information on the characteristics of raw Strava segments for each city in our sample.

Remapping of Strava data To identify the gender of cyclists travelling upon the street network of each city, Strava data were re-projected on the street network of the corresponding city, extracted from OSM [27]. The remapping followed the six-step pipeline described below.

1. Load the Strava data for city *c*.
2. Extract the bounding box of city *c* from the Global Human Settlement—Urban Centre Database 2015, version 2019A (GHS-UCDB R2019A) [40].
3. From OpenStreetMap, extract the street network within the polygon defined in the bounding box using the OSMnx library [41]. Set: *network_type* = '*bike*', *retain_all* = *True*, *simplify* = *True*.
4. Classify streets in the street network based on OpenStreetMap attributes in: 'street with protected cycleway', 'street with unprotected cycleway' and 'street with no cycleway'. The (*key*, *value*) pairs for the classification are provided in the Table S1 in the SI. All other bikable streets are classified as 'no cycleway'.
5. Proceed with the *preferential assignment* of Strava segments as follows. Buffer with a 10 meters radius the geometries of the street network. Select all streets categorized as 'protected cycleways' 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 cycleways' first and subsequently streets with 'no cycleways'.
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, $Females_i$ ($Males_i$) 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:

$$Females_s = \sum_{i \in I} Females_i. \tag{1}$$

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

$$\sigma_s = \frac{Females_s}{Males_s + Females_s} = \frac{\sum_{i \in I} Females_i}{\sum_{i \in I} Females_i + Males_i}. \tag{2}$$

The rationale for the *preferential assignment* is that if a cycleway runs parallel to a street with no cycleway and the linestring 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.

Construction of the city-level index of the gender-cycling-gap The gender-cycling-gap of city c is measured by 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 less trips then men and the propensity to cycle shorter distance. This measure is equivalent to the weighted sum of the gender ratio on streets (σ_s) 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 city C, $Females_s$ ($Males_s$) the number of female (male) cyclists on s and l_s the length of street s expressed in kilometers, σ_c is defined as:

$$\sigma_c = \frac{\sum_{s\in S} Females_s * l_s}{\sum_{s\in S}(Females_s + Males_s) * l_s} = \frac{\sum_{s\in S} \sigma_s * l_s * (Females_s + Males_s)}{\sum_{s\in S}(Females_s + Males_s) * l_s}. \tag{3}$$

4.2 Understanding the determinants of gender-cycling-gap—a cross-cities analysis: data, and methodology

Data sources A full list of data sources used for this strand of the study is provided below.

- Data on recreational cycling at city-level from Strava. The data were processed following the steps described in the previous section.
- City-indicators from the GHS-UCDB R2019A [40]. The following information was extracted:
  1. the share of population living in green areas: data field SDG_A2G14;
  2. the percentage of open space: data field SDG_OS15MX;
  3. the built-up area per capita, data field $BUCAP15$;
  4. the average night-light emission: data field NTL_AV;
  5. the concentration levels of PM2.5: data field E_CPM2_T14;
  6. the city area: data field $AREA$;
  7. the population density: computed as $\frac{P15}{AREA}$;
  8. the GDP per person, computed as $\frac{GDP15\_SM}{P15}$.
- Street network indicators from [28]. Out of the list of available indicators, we extracted the average absolute street grade, the average degree, orientation order, the proportion of three-way intersections and the average street straightness.
- Urban safety indicators measuring the proportion of the street network with cycleways and the proportion of streets with low speed limit. These data were directly constructed from OSM [27] information following the pipeline in the following section.
- The Global Gender Gap Index (country-level) from the World Economic Forum [42]. Included in the sensitivity analysis only.

It should be noted that the final sample for this component of the analysis consists of 61 cities. The city of New York was excluded from this component of the study due to the large discrepancy between the administrative area of this city and the bounding box of the GHS.

Construction of urban road safety indicators OSM information accessed via the Python library OSMnx [41] was used to construct the two indicators on urban road safety. The indicator on the proportion of streets with max-speed limit equal or below 20 mi/h or 30 km/h (referred to as *speed limit* throughout the manuscript) was constructed according to the pipeline described below.

1. For each city c, extract the bounding box of city c from the GHS-UCDB R2019A [40].
2. Extract the street network from the polygon defined in the bounding box using the OSMnx library [41]. 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 according to the pipeline below.

1. For each city c, extract the bounding box of city c from the GHS-UCDB R2019A [40].
2. Extract the street network from the polygon defined in the bounding box using the OSMnx library [41]. Set: *network_type* = '*bike*', *retain_all* = *True*. Call this graph $G0$.
3. From OSM [27], 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 S2 in the SI.
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:

$$\sigma_c = \sum_{j=1}^{N} \beta_j z_{j,c} + \epsilon_c, \quad c = 1, \dots, 61 \tag{4}$$

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 1), the final subset of seven indicators (plus the three country-level dummies) included in the regression were selected via exhaustive search to minimize the Akaike Information Criterion (AIC) of the model. The model is estimated using the OLS function of the Python library *statsmodel* [43].

4.3 Case study on the City of New York: data and methodology

Data sources A full list of data sources used for this component of the study is provided below.

- Data on recreational cycling at street-level for the city of New York from Strava. The raw Strava data were processed and remapped to the street network of each city extracted from OSM following the steps described previously. A network definition of streets was used, which does not reflect a the toponymy of streets.

- OSM data on street-level characteristics extracted during the process of remapping of Strava data via the python library OSMnx [41]. In particular, for each street, we retained information on: the presence of public lighting, the presence of protected or unprotected cycleways, proximity with a park or with the coastline and whether the surface is paved. A list of OSM key-value pairs is provided in Table S3 in the SI. In addition, for streets in the largest component of the street network, we computed the edge-betweenness [30] via the Python library *graph-tool* [44]. Streets outside the largest component of the network (i.e. streets in the borough of Staten Island) were excluded from the sample.
- Administrative data from the OpenData Portal of the city of New York on location of all (any-vehicle) accidents and bike accidents only [45]. These data were processed to compute the number of accidents per 10 meters, for each street.
- Shapefiles of the administrative boundaries of boroughs in the city of New York [45].

Multivariate logistic regression To assess the degree of association between σ_s and the presence of cycling-dedicated infrastructure, we estimates a multivariate logistic regression model. 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 as predictors and the binarized σ_s as the target variable. Moreover, we scaled continuous predictors (*Any-vehicle crashes*, *Bike crashes* and *Edge-betweenness*) using a z-score-transformation to normalize the magnitude of the estimated coefficients. he model is estimated using the Logit function of the Python library *statsmodel* [43].

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-023-00385-7.

Additional file 1. Supplementary information (PDF 1.4 MB)

Acknowledgements

We thank Ane Rahbek Vierø for helpful discussions. RS acknowledges partial support by the European Union's Horizon 2020 research and innovation program under grant agreement No. 869764 (GoGreenRoutes). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. LN gratefully acknowledges the support from the Lagrange Project of the ISI Foundation funded by the CRT Foundation. LN, PB, APa and APe performed most of the work while at ISI Foundation.

Funding

Open Access funding provided by Università degli Studi di Torino within the CRUI-CARE Agreement.

Abbreviations

UN, United Nation; SDG, Sustainable Development Goal; OSM, OpenStreetMap; SI, Supplementary Information; OLS, Ordinary Least Squares; PM 2.5, Fine particulate matter; GDP, Gross Domestic Product; OR, Odds Ratios; AIC, Akaike Information Criterion; GHS-UCDB R2019A, Global Human Settlement—Urban Centre Database 2015, version 2019 A.

Availability of data and materials

The Python code developed for the data analysis and the datasets generated during the current study are available at the following Github repository. The raw Strava data are available from the corresponding author upon request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author contributions
AB processed the data, performed the analysis and drafted the manuscript. LN and APe performed the collection of Strava data. LN carried out the preliminary data analysis. MS, PB, APa, APe and RS designed the experiments, drafted and revised the manuscript. All authors read and approved the final manuscript.

Author details
[1]University of Turin, Via Giuseppe Verdi, 8, 10124, Turin, Italy. [2]Central European University, Quellenstraße 51, 1100 Wien, Austria. [3]CENTAI Institute, Corso Inghilterra, 3, 10138 Turin, Italy. [4]ISI Foundation, Via Chisola 5, 10126 Turin, Italy. [5]IT University of Copenhagen, Rued Langgaards Vej 7, 2300 Copenhagen, Denmark. [6]Complexity Science Hub, Josefstädter Str. 39, 1080 Wien, Austria.

Received: 14 April 2022 Accepted: 30 March 2023 Published online: 19 April 2023

References
1. Oja P, Titze S, Bauman A, De Geus B, Krenn P, Reger-Nash B, Kohlberger T (2011) Health benefits of cycling: a systematic review. Scand J Med Sci Sports 21(4):496–509
2. Leyland L-A, Spencer B, Beale N, Jones T, Van Reekum CM (2019) The effect of cycling on cognitive function and well-being in older adults. PLoS ONE 14(2):0211779
3. Avila-Palencia I, de Nazelle A, Cole-Hunter T, Donaire-Gonzalez D, Jerrett M, Rodriguez DA, Nieuwenhuijsen MJ (2017) The relationship between bicycle commuting and perceived stress: a cross-sectional study. BMJ Open 7(6):013542
4. Gössling S, Choi A, Dekker K, Metzler D (2019) The social cost of automobility, cycling and walking in the European Union. Ecol Econ 158:65–74
5. Assembly UNG (2015) Transforming our world: the 2030 agenda for sustainable development. United Nations, New York, NY, USA
6. Monk J, Hanson S (1982) On not excluding half of the human in human geography. Prof Geogr 34(1):11–23
7. Hanson S (2010) Gender and mobility: new approaches for informing sustainability. Gend Place Cult 17(1):5–23
8. Hosford K, Winters M (2019) Quantifying the bicycle share gender gap. Findings, 10802
9. Funaki D (2019) 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
10. Cycling UK (2021) Cycling statistics. Cycling UK
11. Pucher J, Buehler R (2008) Making cycling irresistible: lessons from the Netherlands, Denmark and Germany. Transp Rev 28(4):495–528
12. Aldred R, Elliott B, Woodcock J, Goodman A (2017) Cycling provision separated from motor traffic: a systematic review exploring whether stated preferences vary by gender and age. Transp Rev 37(1):29–55
13. Garrard J, Rose G, Lo SK (2008) Promoting transportation cycling for women: the role of bicycle infrastructure. Prev Med 46(1):55–59
14. Misra A, Watkins K (2018) Modeling cyclist route choice using revealed preference data: an age and gender perspective. Transp Res Rec 2672(3):145–154
15. Dill J, Goddard T, Monsere C, McNeil N (2015) Can protected bike lanes help close the gender gap in cycling? Lessons from five cities, pp 1–18
16. Hood J, Sall E, Charlton B (2011) A GPS-based bicycle route choice model for San Francisco, California. Transp Lett 3(1):63–75
17. Akar G, Fischer N, Namgung M (2013) Bicycling choice and gender case study: The Ohio State University. Int J Sustain Transp 7(5):347–365
18. Heinen E, Maat K, van Wee B (2013) The effect of work-related factors on the bicycle commute mode choice in the Netherlands. Transportation 40(1):23–43
19. Wang K, Akar G (2019) Gender gap generators for bike share ridership: evidence from citi bike system in New York City. J Transp Geogr 76:1–9
20. Eren E, Uz VE (2020) A review on bike-sharing: the factors affecting bike-sharing demand. Sustain Cities Soc 54:101882
21. Sun Y, Mobasheri A (2017) Utilizing crowdsourced data for studies of cycling and air pollution exposure: a case study using Strava data. Int J Environ Res Public Health 14(3):274
22. Musakwa W, Selala KM (2016) Mapping cycling patterns and trends using Strava metro data in the city of Johannesburg, South Africa. Data Brief 9:898–905
23. Strava: year in sport 2018. https://blog.strava.com/press/2018-year-in-sport/
24. Ministry of Infrastructure and Water Management, (2020) Cycling facts: new insights, 2020. https://s23705.pcdn.co/wp-content/uploads/2021/03/Netherlands-Cycling-Facts-2020.pdf
25. Strava: year in sport 2019. https://blog.strava.com/press/strava-releases-2019-year-in-sport-data-report/
26. Tukey JW et al (1977) Exploratory data analysis, vol 2. Addison-Wesley, Reading
27. OpenStreetMap contributors (2017) Planet dump. https://planet.osm.org
28. Boeing G (2021) Street network models and indicators for every urban area in the world. Geographical analysis
29. Asgarzadeh M, Verma S, Mekary RA, Courtney TK, Christiani DC (2017) The role of intersection and street design on severity of bicycle-motor vehicle crashes. Inj Prev 23(3):179–185
30. Latora V, Nicosia V, Russo G (2017) Complex networks: principles, methods and applications. Complex networks: principles, methods and applications. Cambridge University Press, Cambridge
31. Pearson L, Dipnall J, Gabbe B, Braaf S, White S, Backhouse M, Beck B (2022) The potential for bike riding across entire cities: quantifying spatial variation in interest in bike riding. J Transp Health 24:101290
32. Marshall WE, Ferenchak NN (2019) Why cities with high bicycling rates are safer for all road users. J Transp Health 13:100539

33. Goddard T, Dill J (2014) Gender differences in adolescent attitudes about active travel. In: Transportation Research Board 93rd annual meeting, Washington, DC
34. Aldred R, Woodcock J, Goodman A (2016) Does more cycling mean more diversity in cycling? Transp Rev 36(1):28–44
35. Heesch KC, Sahlqvist S, Garrard J (2012) Gender differences in recreational and transport cycling: a cross-sectional mixed-methods comparison of cycling patterns, motivators, and constraints. Int J Behav Nutr Phys Act 9(1):1–12
36. Prati G, Fraboni F, De Angelis M, Pietrantoni L, Johnson D, Shires J (2019) Gender differences in cycling patterns and attitudes towards cycling in a sample of European regular cyclists. J Transp Geogr 78:1–7
37. Graystone M, Mitra R, Hess PM (2022) Gendered perceptions of cycling safety and on-street bicycle infrastructure: bridging the gap. Transp Res, Part D, Transp Environ 105:103237
38. Prati G (2018) Gender equality and women's participation in transport cycling. J Transp Geogr 66:369–375
39. Garrard J, Handy S, Dill J (2012) Women and cycling. City Cycl 2012:211–234
40. Florczyk A, Melchiorri M, Corbane C, Schiavina M, Maffenini M, Pesaresi M, Politis P, Sabo S, Freire S, Ehrlich D et al (2019) Description of the GHS Urban Centre database 2015. Public release
41. Boeing G (2017) Osmnx: new methods for acquiring, constructing, analyzing, and visualizing complex street networks. Comput Environ Urban Syst 65:126–139
42. Forum WE (2020) The global gender gap index 2020. Insight report
43. Seabold S, Perktold J (2010) Statsmodels: econometric and statistical modeling with Python. In: Proceedings of the 9th Python in science conference, Austin, TX, pp 57–61
44. Peixoto TP (2014) The graph-tool Python library. figshare
45. New York City Council (2020) Administrative data from the New York City Council. https://opendata.cityofnewyork.us/. Accessed July 2021

Publisher's Note