

Analysis of Italian mobility patterns using institutional and networked data

Mauro Mazzei

`mauro.mazzei@iasi.cnr.it`

Consiglio Nazionale delle Ricerche (CNR)

Teresa Bartolomei

Consiglio Nazionale delle Ricerche (CNR)

Research Article

Keywords: Data analysis, Web scraping, Carpooling, Mobility analysis, Big data, Gravity model

Posted Date: July 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9247858/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Applied Geomatics on October 22nd, 2026. See the published version at <https://doi.org/10.1007/s12518-026-00818-8>.

Analysis of Italian mobility patterns using institutional and networked data

Mauro Mazzei^{1*} and Teresa Bartolomei¹

¹Istituto di Analisi dei Sistemi ed Informatica - LabGeoInf, Consiglio Nazionale delle Ricerche (CNR), Via dei Taurini, 19, Rome, 00185, Italy.

*Corresponding author(s). E-mail(s): mauro.mazzei@iasi.cnr.it;
Contributing authors: t.bartolomei@yahoo.it;

Abstract

Traditional data sources, such as population censuses, have several limitations, including long update times, huge costs, and measurement bias, which can lead to gaps in understanding population dynamics. The aim of this article is to assess the quality of new technologies and the usefulness of the Web as a source of massive amount of ge-ospacial data for mobility analysis. To this end, we would like to compare data obtained through web scraping with available population censuses to identify similarities or possible new trends. We will explore the travel routes offered on the carpooling site BlaBlaCar in Italy and compare them with data coming from the commuting matrix for study or work purposes elaborated by ISTAT (Istituto Nazionale di Statistica), a public research organization that produces and disseminates official statistics in Italy. Finally, we will use a gravity model by combining other open data on the web to try to estimate internal migration flows based on demographic and economic variables. While institutional data capture dense, routine commuting, BlaBlaCar data reflect longer, voluntary travel behaviours. Despite structural differences, both sources offer valuable insights into population movements.

Keywords: Data analysis, Web scraping, Carpooling, Mobility analysis, Big data, Gravity model

1 Introduction

The present study focuses on the domain of mobility, given its central role in the development of modern societies, shaping both the economic and productive framework,

but also the quality of people’s lives [1, 2]. As cities and metropolitan regions evolve, travel demand patterns shift, presenting increasingly complex challenges regarding environmental sustainability, equitable accessibility, and transport efficiency [3, 4]. Despite ongoing efforts to incentivize public transit and promote sustainable mobility options, private motorized transport continues to dominate-particularly in Italy-where car ownership rates remain among the highest in Europe, contributing significantly to congestion and emissions in urban centres [5].

In this scenario, it is necessary to promote alternative solutions, such as carpooling, a phenomenon that has gained popularity due to increasing needs for cost savings and reduced environmental impact [6]. However, despite its potential, carpooling remains less developed in Europe compared to the United States [7]. To date, European authors have primarily conducted qualitative research into the pro-files and motivations of carpoolers [8]. The results of these surveys in France and Germany seem to converge on the fact that the majority of users are students.

By exploiting the digital traces of individuals using a web-based carpooling service, it is possible to obtain a new source of data that allows both temporal analysis (e.g., distribution of trips over time) and spatial analysis at urban and national scales [9, 10]. Nonetheless, this type of data requires new methodologies for its collection, which often vary according to the platform distributor.

For this case study, the BlaBlaCar.it platform was chosen as a source of online data to be compared with official data. Founded in 2006, BlaBlaCar has become the world’s largest carpooling community, with more than 20 million users in 19 countries, offering an affordable and environmentally friendly alternative to commuting. On the other hand, the dataset on daily commuting considered official and useful for our comparison is the matrix of commuting for study or work purposes compiled by ISTAT on the occasion of its general population censuses.

2 Related Works

This chapter examines two case studies that have attempted to analyse data related to carpooling, through API-based data collection: ”Potentials and limits of (geo)digital footprints in mobility analysis: the example of the data from the BlaBlaCar carpooling platform” by Mericskay (2019) [11], and ”The territorial impacts of BlaBlaCar carpooling: Insights from user-generated data” published in 2024 by Louail, T., Commenges, H., Come, E. (2024) [12]. Although the BlaBlaCar API provides structured data (it supports formats such as json, xml, etc.), it has some limitations. Most notably, it doesn’t allow you to retrieve trips completed, but only those still being sold. Furthermore, the API has a restriction of no more than 200 records collected per day, which can only be increased if they are not shared. Given these obstacles, Project [11] had to perform multiple daily queries throughout the period of interest, while Project [12] decided to limit data collection to the 120 most significant cities.

2.1 Temporal and Spatial Analysis of Journeys to and from Rennes

The project [11] created an automated script that have collected, every day for 5 months, different types of data on trip offers to and from the French city of Rennes (geographic coordinates, departure dates and times, duration, distance, origin/destination city, number of seats available, price of the trip etc.). Analysing the frequencies of trips, it was observed that these are concentrated mainly on weekends, especially on Fridays. On a monthly basis, they reach their maximum in April and their minimum in July, and then resume at the end of August. It was also possible to observe a pattern between distance and day of the week: the average distance taken on weekend trips was higher (220 km on Friday versus 177 km on Tuesday). By geo-referencing the departure and arrival points, it is possible to explore potential territorial carpooling dynamics. For the analysis, different types of maps were created depending on the spatial scale, flow maps for the national level and heat maps for the local level. Flow maps revealed that most of the trips were concentrated between Brittany, Pays de la Loire, Normandy and le-de-France, with an area of influence around Rennes. Thus, as the theory of gravity models suggests, it was demonstrated that proximity is a determining factor for carpooling trips offer. It was also possible to observe that carpooling favours urban areas inhabitants above all, in fact more than 90% of journeys arrive in or leave from a municipality in an urban area. The georeferenced data also allowed for a more detailed analysis at the local level. Through the generation of heat maps it was possible to highlight the "hot spots" of carpooling in the city of Rennes. Most of carpooling areas were concentrated near in-termodal hubs (train stations, metro) and university areas.

2.2 Carpooling Analysis Models

The project [12] collected data on carpooling offers for the 120 most relevant French cities. The script was run daily for five different periods of the year, so that 35% were vacation days, 30% were strike days, and the remainder were regular days.

#	Hypothesis	Methods
1	As the population size increases, the number of available car-sharing offers also increases.	log-linear regression
2	The volume of offers exchanged between two locations is directly proportional to the volume of offers at each endpoint and inversely proportional to the distance that separates them.	A Pareto form of the gravity model
3	Car-sharing has the potential to facilitate connectivity between urban and rural areas, particularly in instances where public transportation services are limited.	Map of primary flows

Table 1: Hypotheses and related methods

The first model has revealed that the size of the city accounts for approximately 65% of the total variance in the BlaBlaCar offer. Indeed, cities with a population greater than 200,000 inhabitants represent more than 45% of the flows. Analysing the model’s residuals, that is, the difference between the observed (actual) values and the values predicted by the model for each observation, several situations emerged. First, the locations where the model overestimated flows were Paris and the northern area (around Lille). In the former, the lower use of carpooling is likely due to the abundance of transport services. On the other hand, most of the positive residuals, and therefore the model’s underestimates, are located in intermediate-sized cities, confirming their importance as a connecting node between large and small cities. The second model found that the number of trips to and from each city and the distance between them accounted for 70% of the variance in all trips. In this instance, the researchers found that the model overestimated the flow between Paris and other nearby major cities, such as Lyon, Lille, and Marseille. However, they observed an underestimation of the flow between Paris and Rennes, perhaps due to the latter’s high student population. Furthermore, it is interesting to note that among the underestimated flows there are also some from small tourist cities. Finally, the analysis of primary flows connecting rural areas via carpooling showed that most rural locations are located along the tourist coast, some of which are specific tourist attractions. These rural amenities are connected to intermediate cities that have grown significantly economically in recent years. Using carpooling to connect large cities with less populated towns in rural areas or tourist destinations can truly be a solution worth promoting as an alternative to uneconomical transport system development plans.

3 Materials and Methods

3.1 Big Data Collection through Web Scraping

According to Bastin and Tubaro [13], there are several methods of accessing big data: using open data, web scraping, establishing partnerships with private companies, and access through APIs. In contrast to the two papers cited above, which rely primarily on BlaBlaCar API, this case study employs a mixed approach that includes direct web scraping of the BlaBlaCar platform, in addition to querying Google APIs and using available open data sources. Web scraping from the BlaBlaCar website provided an optimization, over using the API, because it also provided completed trips, which the scraper could intercept. Web scraping refers to the automated extraction of data from websites using software that interacts with the HTML structure of web pages to retrieve specific content such as text, images, or hyperlinks. The collected data can be stored in structured formats such as databases or further processed for analytical purposes. This technique replicates, at scale, the manual browsing and copying activities typically performed by human users and is frequently employed to acquire large volumes of data (e.g., stock market indices, corporate registries, product prices). Although web scraping is generally considered legal, its use must comply with intellectual property laws regarding the data being scraped. For example, it is good practice to adhere to the Robots Exclusion Protocol, which is implemented via a website’s robots.txt file. Typically located at the root directory of the domain

(<https://example.com/robots.txt>), the file communicates the webmaster’s preferences regarding which parts of the site should not be accessed by automated agents, including scrapers and crawlers. While compliance with robots.txt is voluntary and not legally enforceable, disregarding its directives may be viewed as unethical and, in some cases, could lead to legal disputes [14]. Open data is a particular category of big data that is freely accessible, often released by public agencies or institutions for the purposes of transparency, accountability, and public reuse.

3.2 Python Tools for Web Scraping, Data Analysis, and Geospatial Visualization

Python is a high-level, general-purpose programming language developed by Guido van Rossum and first released in 1991. Due to its clean and readable syntax, it has become a cornerstone in the programming ecosystem, facilitating rapid development and data-driven applications across multiple domains, including web development, automation, artificial intelligence, and data science [15]. In the context of this research, Python was chosen for its robust ecosystem of specialized libraries that support web scraping, data manipulation, statistical modelling, and geospatial visualization. Web Scraping Libraries:

- Selenium: An open-source framework for automating browser interactions, often used for testing web applications across different browsers and platforms[16]
- Selenium driverless and selenium profiles: Educational libraries developed by Aurin Aegerter to extend Seleniums capabilities. The former removes the need for local browser drivers, while the latter allows custom browser profile management, effectively bypassing anti-bot mechanisms implemented by DNS providers such as Google, Cloudflare, and creep-js. Data Analysis Libraries:
- Pandas: A widely used library for data manipulation and analysis, particularly effective with structured datasets such as CSV, Excel, or SQL tables. The core data structure, the DataFrame, provides powerful tools for cleaning, aggregating, and transforming large volumes of data. Its integration with NumPy and its support for time-series operations make it especially suitable for empirical research in social and geographic domains. Gravity Modelling Library:
- Gravity Modeling Environment (GME): A Python-based toolkit developed by the United States International Trade Commission (USITC) for estimating gravity models using Poisson pseudo-maximum likelihood (PPML). Although originally designed for international trade flow analysis, GME can be adapted to study any type of flows, including commuting flows. The package depends on pandas and statsmodels, and in this project, modifications were made to accommodate the deprecation of the .append() method in recent versions of pandas [17]. Geospatial Data Visualization Libraries:
- Plotly: A data visualization library supporting interactive charts, including geographic visualizations such as scatter maps, heatmaps, and choropleth maps. Plotly integrates seamlessly with pandas and GeoPandas, facilitating geospatial analyses.
- Folium: A Python wrapper for the Leaflet.js JavaScript library, designed for the creation of interactive maps. It supports various data formats such as GeoJSON and

shapefiles, enabling the rendering of point, polygonal, and line features on web-based maps.

3.3 Open Data Sources

The gravity model is a method used in economics, especially in the context of international trade, to explain trade flows between countries based on their economic size and the geographical distance separating them. The first to introduce the concept that larger economies tend to trade more with each other, while trade decreases with increasing distance, was Walter Isard in 1954. Tinbergen (1962) later expanded and introduced the formula of the gravity model for the first time, based on an analogy with Newton's law of universal gravitation in the following form:

$$F = G \frac{m_1 m_2}{r^2}$$

Which states that the force of attraction between two bodies is directly proportional to their masses and inversely proportional to the square of the distance between them. Thus the Tinbergen equation was formulated as follows:

$$X_{ij} = C \frac{Y_i Y_j}{t_{ij}}$$

Where:

- X_{ij} = The amount of trade between the country I and the country J
- C = constant
- Y_{ij} = GDP of the country I and the country J
- t_{ij} = trade costs between the country I and the country J

$$\text{comm}_{m\{ij\}} = \alpha_0 + \alpha_1 \text{pop}_i + \alpha_2 \text{pop}_j + \alpha_3 \text{firms}_i + \alpha_4 \text{firms}_j + \alpha_5 \text{dist}_{\{ij\}} + \alpha_6 \text{col}_i + \alpha_7 \text{col}_j$$

where:

- $\text{pop}_i, \text{pop}_j$: population of origin/destination,
- $\text{firms}_i, \text{firms}_j$: number of active enterprises,
- dist_{ij} : road distance between municipalities,
- $\text{col}_i, \text{col}_j$: cost-of-living indexes.

Model estimation will be computed using the Poisson Pseudo Maximum Likelihood (PPML) estimator, a special case of the Generalized Linear Model (GLM) framework, following the findings of Santos Silva and Tenreyro [21]. The latter argue that this method is superior to the logarithmic linearization of the empirical model, which is obtained by the popular least squares method, because it avoids the problems associated with heteroschedasticity, multicollinearity and autocorrelations. Generalized linear models (GLMs) are an extension of the traditional linear model used in linear regression. In the classical linear model, the dependent variable is assumed to follow a normal distribution. In generalized linear models indeed, the dependent variable can follow any distribution belonging to the exponential family. This means that,

in addition to the normal distribution, other distributions of random variables can be considered, such as binomial, poissonian, gamma, inverse normal and others. Ordinary linear regression estimates the expected value of an unknown random variable, called the response variable, as a linear combination of a set of observed values, called predictors. This implies that a constant change in a predictor generates a constant change in the response variable, configuring a linear response model. An example of the application of linear regression could be the analysis of the effect of advertising budget on monthly sales. This approach is appropriate when the response variable can vary approximately in both directions without limit, or more generally, for quantities that vary relatively little with respect to the variation in predictors, as is the case. Nevertheless, these assumptions are not appropriate for response variables of certain types. For example, when the response variable is expected to be always positive and to vary over a wide range, constant changes in the predictors lead to changes in the response growing geometrically (i.e., exponentially), rather than steadily. The fundamental relationship in a GLM can be expressed by the following equation:

$$E(Y) = \mu = g^{-1}(X\beta)$$

Specifically, $E(Y)$ represents the expected value of the dependent variable Y , while μ is its mean. The g^{-1} function is the linking function that transforms the linear predictor X , a linear combination of the independent variables X and the parameters β , into a scale for interpreting the mean of the dependent variable. The Poissonian regression used in this project is a GLM with link function the logarithmic function. The unknown parameters are usually estimated by the method of maximum likelihood (or, as in this case, Pseudo Maximum Likelihood).

4 Case Study

4.1 Carpooling Scraper

In the case of BlaBlaCar.it the robots.txt file allows scraping on the data by any user-agent of the following pages: /bus, /ride-sharing, /carpool, /search-car-sharing and /scam. In other words, it is possible to extract all data about the carpooling offer (routes, prices, dates and times, etc.). As observed by the work of [12], we found that also the Italian BlaBlaCar dataset contains some errors, at least as it is displayed to a human user by the site. First of all, the semantics of the origin and destination locations is a free text provided by the drivers offering the rides. As reported further in chapter 4.3, this problem was easily solved by "cleaning up" the dirty data resulting from geocoding, i.e. the records for which Geonames had not found the spatial coordinates. Secondly, when a user queries the site by providing an origin and a destination, the offers are listed by filtering out any intermediate trips, due to the driver offering intermediate stops. Proceeding step by step, first we verified that the data of interest were uniquely identifiable by a tag and then selectable with one of the selectors of the driver.find element function (ID, name, class, xpath, css selector, etc.). As often happens for many websites, BlaBlaCar HTML class codes are assigned dynamically, probably by JavaScript frameworks or CSS libraries, this forced us to modify the code

at least once a month. Initially the script queried subpages in the /ride-sharing/all-products section of the site, but later these lists proved to be non-exhaustive. Finally, we decided to query the site by providing only the destination + date tuple as input of the manipulation of the URL of the /search-car-sharing section.. The script loads in sequence N different URL for the N Italian provinces intended as trip destinations. For example, to get all available trips to Milan on September 18, 2024, the URL to load is the following: <https://www.blablacar.it/>. In this way BlaBlaCar data lists, in addition to trips to the searched province, all trips to all municipalities related to the searched province. This allowed us to collect all current and completed trips for the date entered, up until the time of the call. The most complex issue to solve was caused by the anti-bot filter strategies of the Cloudflare web application firewall, which manages to recognise repetitive actions from a given IP address within a certain timeframe, thereby asking the user to solve a puzzle to prove it is human. To overcome this issue, we implemented two solutions. One solution was to reduce the frequency of web scraping on the designated website by inserting breaks between lines of code. It was also necessary to develop a multi-thread system that could open several browser tabs simultaneously to increase the efficiency of web scraping. The method employed was simply to divide the provinces destinations into sub-lists of equal size. The time taken by the script to extract all the trips scheduled for a single day, and save them in several files with csv format, was about 3 hours.

4.2 Google Distance Matrix API

Application Programming Interfaces (APIs) have become a fundamental component of modern software systems, both in corporate and research environments. In particular, they provide standardized access to complex datasets and services, supporting the development of data-driven applications across various domains, including transportation and geospatial analysis [22] [23]. In this study, the Google Distance Matrix API was used to get travel distances when using a car between Italian municipalities. In this study, the Google Distance Matrix API was used to obtain car travel distances between Italian municipalities. In particular, the final list of origins-destinations included both the routes of the ISTAT census and the routes extracted through web scraping. The distance calculated by Google corresponds to the minimum path between two vertices of a graph (in this case: the road network). Each arc in Google’s graph has different weights, for example distance, travel time or other factors such as traffic or road conditions. The minimum path between two vertices of a graph is that path that connects these vertices and minimises the sum of the weights or costs associated with crossing each arc. For this reason, its use allows for more realistic modelling of spatial frictions compared to straight-line (Euclidean) distances, particularly in short- to mid-range urban and regional analyses. The issue with the Google Maps Distance Matrix API is that it supports a maximum matrix size of 25 origins $\times$ 25 destinations = 625 items, that is to say for a single call. The cost per 1000 elements is 5\$, so 3.13\$ per call, about 2.9 euro. Such pricing can be unsustainable for many businesses, especially those currently in the growth stage. Nevertheless, Google offers a pay-as-you-go plan with no daily call limit, as long as you do not exceed the limits per request (max. 25 origins and 25 destinations) nor a maximum of 60.000 items per minute. In order to

avoid obtaining distances of origin-destination pairs that are not of interest to us, in our script we have chosen to group the list of routes by destination and execute a call that in input for an array with a maximum of 25 origins and only one destination. For this purpose, the code creates a temporary array in which it stores the unique destinations, with each of these it then filters the dataframe and groups the origins into arrays of up to 25 locations.

4.3 Open Data

Regarding the 'official' estimate of daily commuting flows, the most recent open-access ISTAT data refers to residents of Italy living in households or in cohabitation who were surveyed for the 15th General Population Census (reference date: 9 October 2011)[24]. Specifically, it refers to the 28.871.447 individuals who declared that they commute daily from their place of residence to their usual place of study (including nursery, kindergarten, and vocational training courses) or work. For the purposes of this study, individuals living in households and commuting outside their usual municipality were selected. These individuals account for 39% of the total, and of these, only those who use a private car as a means of transport were included. This type of travel accounts for the majority, with 63% of individuals travelling outside their usual municipality. The number of active enterprises per municipality was derived from the 2011 edition of the Statistical Register of Active Firms (ASIA - Firms), published by ISTAT [25]. The register includes economic units classified under most productive sectors, with exclusions and adjustments based on sectoral and institutional criteria, following the NACE Rev.2 classification. The data integrates multiple administrative sources, including tax authorities, social security institutions, and national banking and insurance registers. To obtain the geographic coordinates of Italian municipalities, and thus to standardize both datasets to a list of origins and destinations based on municipalities and not generic places, an open-source geolocation dataset was needed. All municipalities were then geocoded using the GeoNames database, allowing spatial analysis based on their centroid coordinates. GeoNames provides a comprehensive and regularly updated global database of geographic names, which includes more than 11 million place names and integrates data from more than 100 official and community sources. This geocoding step was also crucial to support the gravity model [26]. To provide the gravity model with an index of the cost of living in each municipality, we chose to use the average sales prices of houses (per square metre), which are available on the immobiliare.it platform at the municipality level [27]. Although this figure is only one of many factors to consider, it offers a useful and immediate overview of the economic and social situation in a locality. House prices reflect supply and demand in the housing market. An increase in house prices often indicates greater demand, which may result from population growth, job opportunities or a perceived higher quality of life. These factors contribute to a higher cost of living, as people are willing to invest more to live in a certain area. In addition, housing costs directly influence many other daily expenses. Rents and mortgages represent a significant part of the household budget. In addition, areas with more expensive houses also tend to have higher prices for goods and services, such as food, transport and health care. Consequently, house prices can be an anticipatory indicator of changes in general living costs.

Finally, unlike other cost items that can vary considerably on a monthly or seasonal basis, such as food or fuel prices, the housing market tends to be more stable over the long term. Although house prices can fluctuate, they offer a more consistent and long-term view of the cost of living in a municipality.

4.4 Exploratory Data Analysis and Spatial Patterns

Table 1 provides a comparative summary of selected descriptive statistics from two data sources: the official ISTAT commuting dataset from 2011 and web-scraped data collected from the BlaBlaCar platform of a single day of August and of a single day September 2024. The comparison reveals significant structural differences in coverage, scale and spatial extent.

Table 2 summarises selected descriptive statistics from the ISTAT dataset and from BlaBlaCar data collected on a single day in August and on a single day in September 2024.

Table 2: Summary statistics of origin-destination datasets used in the analysis.

	ISTAT (2011)	BlaBlaCar Aug 2024	BlaBlaCar Sep 2024
Total routes (unique OD pairs)	472,573	799	573
Total individuals (flow volume)	7,181,204	962	676
Number of origins	8,090	445	329
Number of destinations	8,083	217	186
Average distance	45 km	374 km	246 km
Maximum distance	498 km	1,402 km	1,314 km

The ISTAT dataset, which is derived from a national census, reports over 472,000 unique origin-destination pairs and more than seven million commuters, with almost full territorial coverage across over 8,000 municipalities. In contrast, although the BlaBlaCar data is far more limited in scale, it offers a dynamic and recent snapshot of inter-city mobility patterns, with trip counts ranging from 573 to 799, but with coverage limited to a few hundred municipalities. Notably, average travel distances differ considerably between the two sources: approximately 45 km for the ISTAT data, versus 374 km and 246 km for the BlaBlaCar data in August and September, respectively. This reflects the structural divergence between daily short-distance commuting (typical of census data) and long-distance, on-demand carpooling behaviour (characteristic of ride-sharing platforms). Further-more, the BlaBlaCar data show seasonal variation, with lower volumes recorded in September compared to the summer peak in August. These findings highlight the complementary nature of the two sources. Institutional data provides reliable, broad coverage that is suitable for long-term structural analysis, while web-scraped data capture the short-term, emerging dynamics of user-driven mobility. Integrating both can improve the spatial and temporal resolution of mobility studies, particularly in situations where official data is outdated or infrequently updated.

4.4.1 Distribution of Trip Distances

As reported in Table 2, the comparison of relative frequency distributions by distance class reveals a striking contrast between the ISTAT 2011 dataset and the BlaBlaCar data collected in August and September 2024. While official statistics indicate a predominance of short- to medium-distance daily trips (with 75

Table 3 presents relative frequency distributions by distance class.

Table 3: Distribution of daily car trips by distance class.

Distance class (km)	ISTAT 2011	BlaBlaCar Aug 2024	BlaBlaCar Sep 2024
0-5	8%	0%	0%
5-10	23%	0.10%	0.10%
10-25	44%	0.60%	2.40%
25-50	19%	2.30%	7.50%
50-100	5%	7.20%	15.70%
≥ 100	1%	89.70%	74.30%

In detail, over 89% of these results confirm that BlaBlaCar is used almost exclusively for long-distance travel, while official data capture routine, short-distance commuting, particularly within urban and suburban areas. This divergence reflects both the intended purpose of the carpooling platform and the behavioural characteristics of its users, who are likely to seek shared rides primarily for inter-city trips rather than for everyday local mobility. In addition, the month-to-month variation observed between August and September indicates a seasonal effect in the platform's usage, with shorter trips (25-100 km) becoming more common after the summer holiday period. This suggests that scraped data from mobility platforms can capture temporal dynamics and user behaviour trends that remain invisible in static institutional datasets.

4.4.2 Top 5 OD Pairs by Individual Volume: A Comparison

Table 4 lists the top five origin-destination (OD) pairs with the highest number of individuals for each dataset. The ISTAT 2011 data highlight typical short-distance, urban or suburban commuting flows such as Roma $\rightarrow$ Fiumicino and Moncalieri $\rightarrow$ Torino, which are characteristic of routine daily travel within metropolitan areas. These routes are often associated with work-related mobility within functional urban regions. In contrast, the most frequent OD pairs derived from BlaBlaCar data in both August and September 2024 reflect long-distance intercity travel. Routes such as Milano $\rightarrow$ Padova, Roma $\rightarrow$ Milano, and Lecce $\rightarrow$ Roma suggest that the carpooling platform is pre-dominantly used for medium- to long-range trips, often spanning hundreds of kilometers.

Table 4 shows the five most frequent OD pairs in each dataset.

The lack of overlap between the two rankings confirms the fundamental difference in the nature of mobility captured by institutional data versus user-generated digital traces. While ISTAT data provides insights into systematic daily commuting,

Table 4: Top five origin-destination (OD) pairs.

Rank	ISTAT 2011	BlaBlaCar Sep 2024	BlaBlaCar Aug 2024
1	Roma → Fiumicino	Milano → Padova	Lecce → Roma
2	Guidonia Montecelio → Roma	Cagliari → Sassari	Roma → Milano
3	Quartu S. Elena → Cagliari	Padova → Udine	Genova → Milano
4	Fiumicino → Roma	Torino → Milano	Trento → Milano
5	Moncalieri → Torino	Olbia → Sassari	Pisa → Milano

BlaBlaCar data capture discretionary and infrequent long-distance travel, reinforcing their complementary use in multi-source mobility analyses.

4.4.3 Flow Maps, Heatmaps, and Choropleths

This section presents the flow maps generated using the Python library Plotly, aimed at visually exploring the spatial structure of non oriented flows. Two main visualization approaches are adopted:

- The first uses weighted markers to highlight only destination points, enabling the identification of municipalities with the highest concentration of arrivals.
- The second employs weighted lines to represent the intensity of flows between municipalities, effectively constructing a network of connections that reveals spatial interaction patterns.

For the ISTAT dataset, due to the high volume of available OD pairs and the Pareto-like distribution, a data reduction was applied. Specifically, only the top 56,123 OD pairs accounting for approximately 80% of the total flow volume were included in the visualization. This filtering ensures clarity and readability while preserving the most significant patterns of territorial mobility.

Figure 1a shows the ISTAT flows using weighted destination markers. The size of each circle represents the total number of arrivals per municipality, revealing high-concentration hubs, particularly in large metropolitan areas such as Rome, Milan, and Naples. Figure 1b, based on BlaBlaCar data (August 2024), also uses weighted destination markers but reveals a significantly different structure. Flows are sparser and longer, concentrated along major North-South axes and between regional capitals. Notably, Sardinia and southern Italy have high levels of long-distance movement, which is consistent with carpooling being used for inter-regional travel.

Figure 2a employs weighted lines to represent the volume of non oriented flows between municipalities. This representation highlights key inter-municipal routes and visualizes the national commuting network. A strong spatial continuity can be observed in Northern and Central Italy, with highly connected clusters in urban regions. Figure 2b represents the flow map of BlaBlaCar trips using weighted lines to represent the intensity of mobility between municipalities. Compared to the ISTAT commuting data, the prevalence of long-distance travel becomes even more evident. Moreover, several prominent flows emerge in this representation: Rome-Lecce, Rome-Milan, and Milan-various cities in Puglia stand out due to their volume and distance. Additionally,

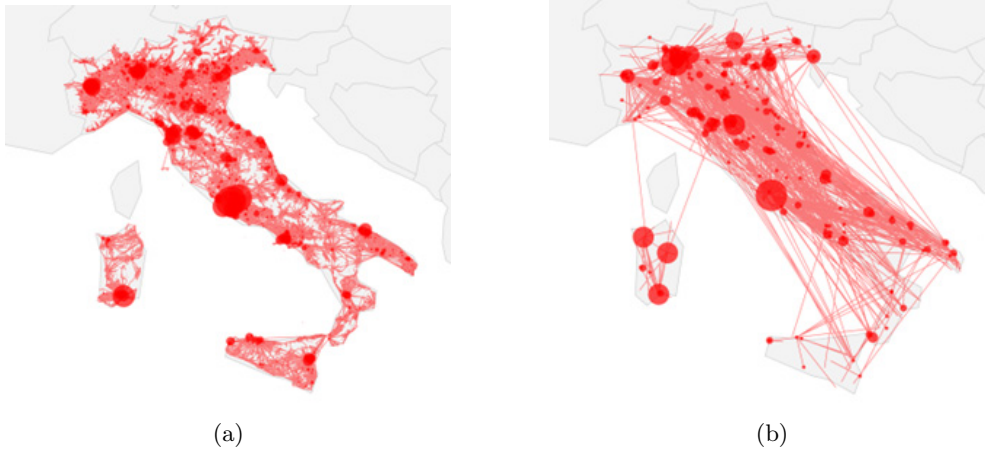


Fig. 1: Comparison of flow maps - destinations weighted markers. (a) ISTAT (2011); (b) blablacar.ite

Sardinia shows strong internal connections, particularly between Olbia, Sassari, and Cagliari. As observed previously, Milan remains the primary node in the BlaBlaCar network, acting as the most frequent origin or destination, followed by Rome and Turin. These flow maps provide a compelling visual contrast between structured, high-volume local commuting captured by official statistics and the selective, long-distance mobility typical of digital ride-sharing platforms. A heatmap is a visual representation of data in which values are expressed through colours. This tool is particularly effective for highlighting areas of high density or in-tensity within a dataset, allowing for the quick identification of patterns, trends, and anomalies. In this study, heatmaps are used to represent the areas with the highest concentrations of car travel origins and destinations. Practically, the Python library Folium was employed, using a list containing aggregated origin and destination data, each weighted by the number of individuals.

The ISTAT heatmaps show a concentration of flows primarily in urban and metropolitan areas of Central and Northern Italy, with higher intensity in the Pianura Padana, the area around Rome, and along the main mobility routes, reflecting a pattern of regular daily commuting. In contrast, the BlaBlaCar data reveals a more heterogeneous and fragmented distribution, with peaks in tourist areas and along major high-ways, particularly during the holiday season in August. The BlaBlaCar origins heatmap (Figure 3b) highlights a strong presence in the North, but also significant activity in Southern regions and on the islands. The destinations heatmap (Figure 3d) shows greater dispersion, with numerous attraction points along the coasts and in holiday destinations. This comparison thus confirms two distinct mobility models: structured and routine in the case of commuting, and more occasional and seasonally driven in the case of carpooling. A choropleth map is a type of thematic map used to represent data through shading or colouring of predefined geographic areas. Each area is shaded in proportion to the value of the variable being represented, making it easier

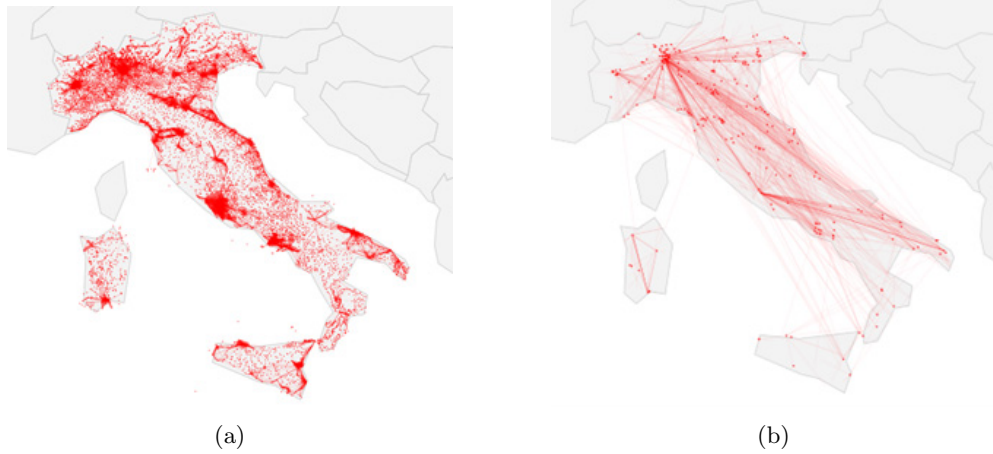


Fig. 2: Comparison of flow maps - weighted lines. (a) ISTAT (2011); (b) blablacar.it

to visualize spatial distributions and patterns. In this study, choropleth maps are created using the Python library Geopandas and are employed to depict regional inflows and outflows of individuals, based on both BlaBlaCar.it data and ISTAT 2011 data.

The maps based on BlaBlaCar data show moderate volumes of movement, with the most significant incoming flows observed in Lombardia, Lazio and Emilia Romagna indicating the importance of these regions as both sources and destinations of long-distance travel during August. In contrast, the ISTAT maps reveal a much more pronounced concentration of flows in Northern regions, particularly in Lombardia and Veneto, both in terms of departures and arrivals. Southern regions, by comparison, exhibit weaker flows in the ISTAT data, whereas they are more represented in the BlaBlaCar dataset. The comparison highlights the scale and nature of the two datasets: institutional data capture high-frequency, short-range movements tied to economic centres, while BlaBlaCar data emphasize occasional, user-driven, often long-distance trips.

5 Model Results

This chapter presents the results of the gravity model introduced earlier, estimated in multiple configurations in order to assess the robustness of the explanatory variables across different data sources and model specifications. The analysis is structured into four main applications of the model, each based on a distinct combination of input data or dependent variable. For clarity and completeness, each model specification is reintroduced with its corresponding functional form before presenting the estimation results and statistical interpretation.



(a)



(b)



(c)



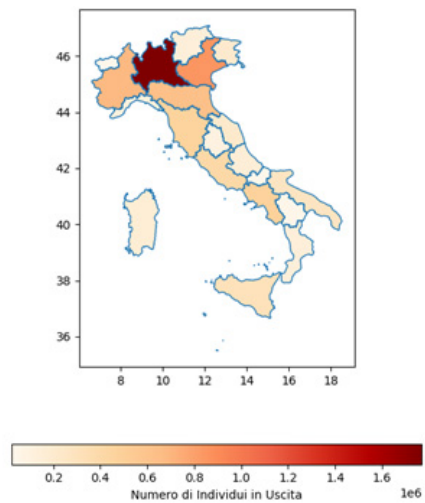
(d)

Fig. 3: Comparison of heat maps origins: (a) ISTAT (2011); (b) blablacar.it destinations: (c) ISTAT (2011); (d) blablacar.it

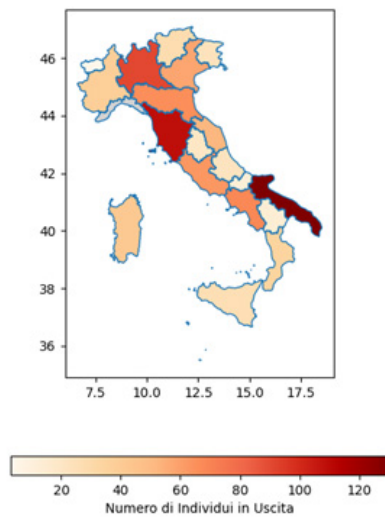
5.1 Model 1: Application to ISTAT Data - 2011

To begin, the gravity model is applied using the ISTAT 2011 commuting flows as the dependent variable. The specification includes population size, number of active enterprises, geographic distance, and cost of living indicators at both the origin and destination. Model 1 specification:

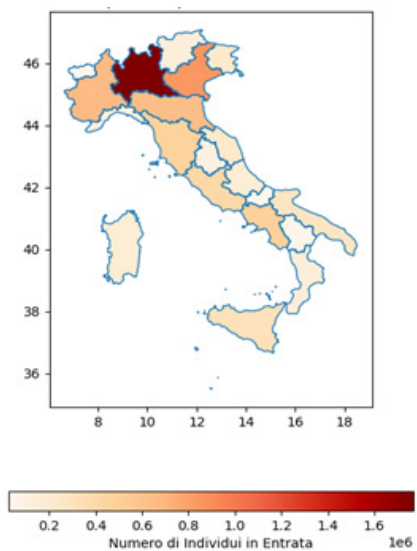
$$\text{comm}_{m\{ij\}} = \alpha_0 + \alpha_1 \text{pop}_i + \alpha_2 \text{pop}_j + \alpha_3 \text{firms}_i + \alpha_4 \text{firms}_j + \alpha_5 \text{dist}_{\{ij\}} + \alpha_6 \text{col}_i + \alpha_7 \text{col}_j$$



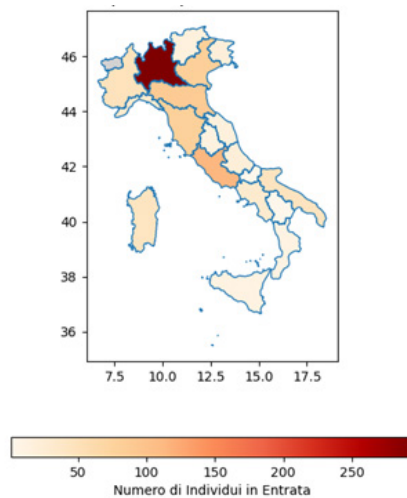
(a)



(b)



(c)



(d)

Fig. 4: Comparison of choropleth maps outflow: (a) ISTAT (2011); (b) blablacar.it
inflow: (c) ISTAT (2011); (d) blablacar.it

5.2 Model 2: Application to BlaBlaCar Data August 2024

The second model applies the same gravity framework to the BlaBlaCar dataset collected on a day in August 2024. In this case, the dependent variable represents the number of ride-sharing trips observed between each origin and destination pair. The aim is to evaluate whether the same structural variables (population, firms, distance) can explain voluntary, user-driven mobility patterns captured through web scraping. Model 2 specification:

$$blablaCarAug_{\{ij\}} = \alpha_0 + \alpha_1 pop_i + \alpha_2 pop_j + \alpha_3 firms_i + \alpha_4 firms_j + \alpha_5 dist_{\{ij\}}$$

5.3 Model 3: Application to BlaBlaCar Data September 2024

Finally, the third specification replicates the previous model but uses data from a day in September 2024 BlaBlaCar as the dependent variable. The purpose is to test the temporal stability of the model by comparing estimation results across two different months. Any variations in model performance may highlight seasonal or behavioural shifts in long-distance shared mobility. Model 3 specification:

$$blablaCarSep_{\{ij\}} = \alpha_0 + \alpha_1 pop_i + \alpha_2 pop_j + \alpha_3 firms_i + \alpha_4 firms_j + \alpha_5 dist_{\{ij\}}$$

Results comparison: Table 5 summarises the estimated coefficients for all three models.

Table 5: Model results comparison. Significant coefficients are highlighted in bold.

Variable	ISTAT 2011	BlaBlaCar Aug 2024	BlaBlaCar Sep 2024
Distance	-7.98×10^{-6}	-8.20×10^{-8}	-8.08×10^{-8}
Population (origin)	$+4.37 \times 10^{-6}$	$+4.83 \times 10^{-7}$	$+6.30 \times 10^{-9}$
Population (destination)	$+3.14 \times 10^{-6}$	-5.66×10^{-7}	-1.97×10^{-7}
Firms (origin)	-2.82×10^{-5}	-3.14×10^{-6}	-6.77×10^{-8}
Firms (destination)	-1.47×10^{-5}	$+6.80 \times 10^{-6}$	$+1.98 \times 10^{-6}$
Cost of living (origin)	$+5.00 \times 10^{-4}$	$+4.47 \times 10^{-5}$	$+6.14 \times 10^{-5}$
Cost of living (destination)	$+5.00 \times 10^{-4}$	-2.93×10^{-6}	$+9.49 \times 10^{-6}$

Model with ISTAT data shows the greatest statistical robustness. In line with the gravity model theoretical foundations, Model 1 confirms that flows between cities are particularly significant not only when cities are large and close to each other, but also when both cities are characterized by high living costs, suggesting a potential influence of economics or labour-related mobility dynamics. Models with BlaBlaCar data, with fewer observations, show lower statistical significance. In particular, Model 2 limits the most important journeys from more populous and more expensive cities to less populous and less expensive cities. These results suggest that users of this type

of carpooling could be out-of-town students returning to their hometowns or holiday-makers. In all models, origin cost of living is the most stable and significant factor, thus it is a strong predictor of mobility. The distance is significant only in the ISTAT models; in the BlaBlaCar models it is never significant, confirming that carpool-based mobility may respond to logics other than purely geographic ones. These results suggest that the spontaneous mobility observed through BlaBlaCar follows economic and structural, rather than geographic, logics. Next, we calculated the model's estimated values and the corresponding residuals (observed flows - estimated flows) and analysed them by drawing some graphs. Finally, we plotted the desire lines, which correspond to positive residuals, and the repulsive lines, which correspond to negative residuals. Both can help identify overrepresented and underrepresented areas in relation to the variables used in the model. By analysing the distribution of residuals, we can conclude whether the model is well calibrated, i.e. the mean is close to 0, and whether it has any outliers. For the BlaBlaCar data (Fig. 5b), most residuals are slightly positive or close to 0. However, there is a long tail on the right, indicating that several actual flows significantly exceed those estimated by the model. For the ISTAT data the slightly negative mean indicates that, on average, the model slightly overestimates the observed values.

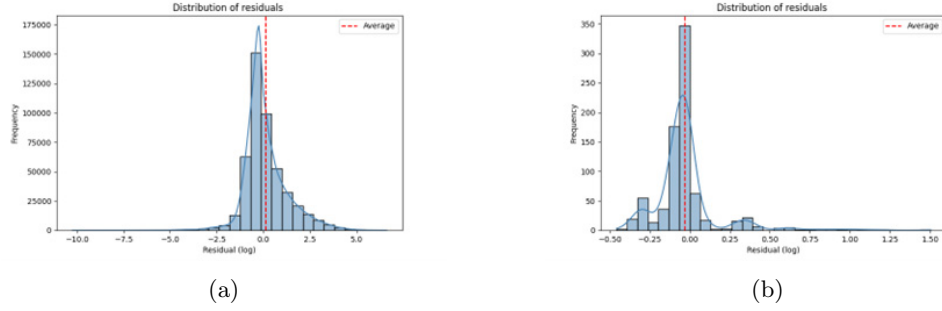
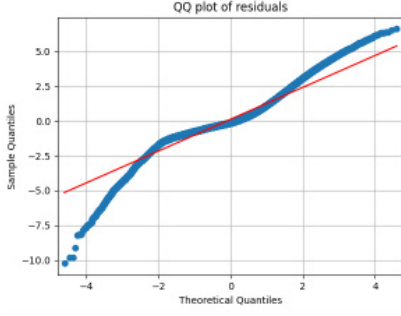


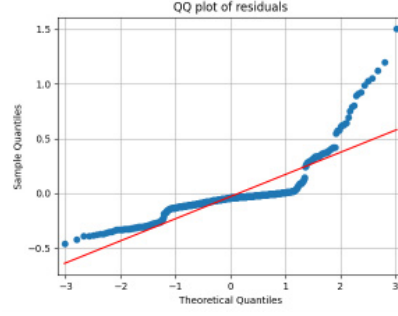
Fig. 5: Comparison of residuals distributions: (a) ISTAT, 2011; (b) blablacar.it, 2024

To assess whether the distribution of residuals follows a normal distribution, we use the QQ plot (Fig. 6). The study of the BlaBlaCar data shows that the points at the upper ends (quantiles ≥ 2) deviate significantly from the line. This indicates the presence of positive outliers, i.e., flows much larger than expected. Also for ISTAT data the tails (extremes) are heavier than expected from a normal one, however it is sufficient that the points are close to the bisector as in fact they are.

Analysing the desire lines shown in Figure 7, we can see where actual trips are more numerous than those estimated based on the variables considered (distance, population, number of companies, and cost of living). Regarding BlaBlaCar data, the connections between Milan and some large northern cities (such as Turin, Genoa, Trento, and Pisa) stand out, as does the Milan-Rome connection. One explanation for this overrepresentation is that these are cities with many university students. In this



(a)



(b)

Fig. 6: Comparison of QQ plots: (a) ISTAT, 2011; (b) blablacar.it, 2024

case, other potential variables that explain these flows should be further investigated. As for the ISTAT data, however, we note that the desired lines are consistent with the size of the dataset. Perhaps the model variables don't fully capture the phenomenon of car commuting, which could also be linked to other local dynamics (e.g., infrastructure, public transport, school campuses).



(a)



(b)

Fig. 7: Comparison of desire lines: (a) ISTAT, 2011; (b) blablacar.it, 2024

The repulsion lines are shown in Fig. 8, i.e., the connections where actual flows are lower than expected. Regarding BlaBlaCar, several connections are highlighted between very large cities in the North, such as Turin and Milan, and very distant municipalities in the South. One possible explanation is that distance in this case acts as a deterrent to car use compared to other, faster means of transport. As for

the underrepresented flows of ISTAT commuters, though, it could be that the public transport offer is preferable, since it mainly concerns the hinterland of Milan, Rome and Naples.

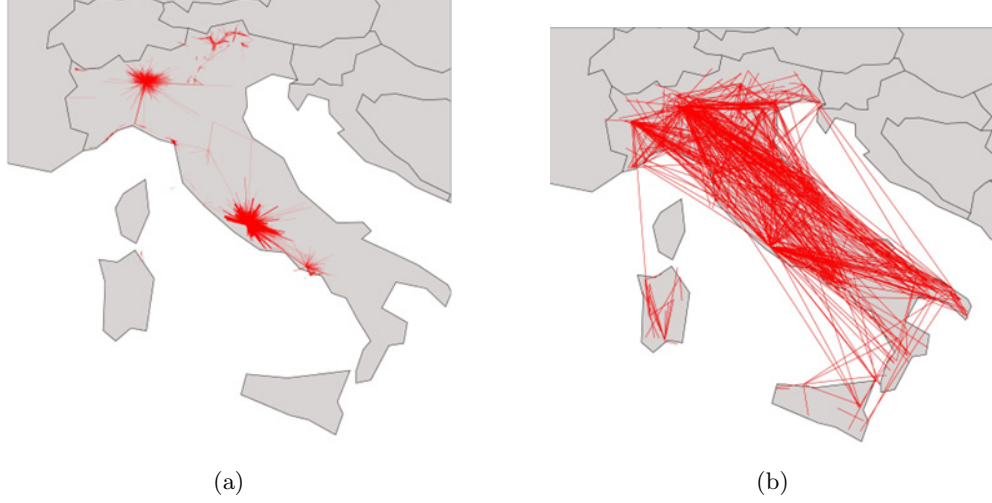


Fig. 8: Comparison of repulsive lines: (a) ISTAT, 2011; (b) blablacar.it, 2024

6 Conclusions

This study examined how official statistical data and web-based, user-generated information could be integrated to formulate a model of inter-municipal mobility in Italy. A comparison of commuting flows from the 2011 ISTAT census with digital footprints collected from the BlaBlaCar carpooling platform in 2024 revealed significant differences between the two datasets. These differences were found in terms of their structure, scale and the drivers behind population movement. ISTAT data describe dense, short-range commuting patterns associated with daily work and education-related travel. In contrast, the BlaBlaCar data capture voluntary, long-distance mobility characterised by fewer routes and seasonal or episodic behaviours. This information can be an added value for companies promoting tourism or for local accommodation facilities, which could promote sustainable transport such as carpooling especially where the public transport network is not sufficiently developed. Despite their differences in scope and granularity, both datasets offer complementary insights and together provide a more comprehensive understanding of mobility dynamics. Gravity model estimations confirmed the importance of structural variables such as population size, number of firms, and cost of living. Notably, the cost of living in the municipality of origin was a consistently significant predictor in all models. To reduce the number of outliers obtained by applying the model to car-pooling data, additional variables could be added to the traditional gravity model formulation (e.g. presence

of universities, tourist attractiveness, presence of local public transport). Methodologically, the results demonstrate the feasibility of using open-source web scraping and geospatial processing tools to generate valuable mobility insights. Further developments in web automation and accessibility of structured enrichment data (such as geocoding) could enable a fully reproducible workflow from start to finish, supporting the integration of non-traditional data into spatial interaction modelling. In the future, the analysis of carpooling data could be repeated with a greater time depth and the study could be expanded to deepen the dynamics of other captured data, such as the schedules and prices of published offers.

- ISTAT captures **dense, short-range, routine commuting** tied to daily work and study.
- BlaBlaCar reflects **voluntary, long-distance, occasional mobility**, often seasonal.

These differences highlight the complementary nature of institutional and web-generated sources. Institutional data provide comprehensive coverage and structural insight, while carpooling traces offer real-time sensitivity to behavioural patterns, tourism, seasonality, and spontaneous travel dynamics.

Gravity model estimations confirmed the relevance of population size, economic activity (number of firms), and cost of living. Notably, the **origin cost-of-living index** was a consistently significant predictor across all model specifications. In contrast, distance was highly relevant for ISTAT commuting flows but not significant for BlaBlaCar data, suggesting that shared-mobility platforms follow behavioural logics that differ from classic geographic frictions.

Future research may include:

- extension of the scraping period for deeper temporal analysis,
- integration of additional behavioural variables (e.g., student population, tourist attractiveness),
- merging of carpooling traces with public transport timetables,
- modelling price sensitivity and temporal availability of BlaBlaCar trips.

By combining traditional and digital mobility data, this study demonstrates the feasibility of reproducing mobility insights at fine spatial resolution using reproducible open-source tools.

Acknowledgements

The authors thank ISTAT for providing access to census documentation and the BlaBlaCar service (<https://www.blablacar.it/search-car-sharing>) for enabling analysis of carpooling data.

Funding

The authors declare that this research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

- [1] Banister, D. (2008). *The Sustainable Mobility Paradigm*. Transport Policy, 15(2), 73-80.
- [2] Cervero, R. (2013). *Transport Infrastructure and Global Competitiveness: Balancing Mobility and Livability*. Transport Policy, 25, 106-112.
- [3] Newman, P., Kenworthy, J. (1999). *Sustainability and Cities: Overcoming Automobile Dependence*. Island Press, Washington, DC.
- [4] Schwanen, T., Banister, D., Anable, J. (2015). *Scientific Research about Climate Change Mitigation in Transport: A Critical Review*. Transportation Research Part A: Policy and Practice, 78, 26-38.
- [5] Isfort. (2023). *Rapporto sulla Mobilit degli Italiani*. Available online: <https://www.isfort.it/pubblicazioni/rapporto-annuale-sulla-mobilita/>. Accessed 29 September 2024.
- [6] Marsden, G., Anable, J., Docherty, I. (2019). *Shared Mobility: Making Sense of the Shifting Transport Landscape*. Journal of Transport Geography, 79, 102476.
- [7] Neoh, J.G., Chipulu, M., Marshall, A. (2018). *What Encourages People to Carpool An Evaluation of Factors with Meta-Analysis*. Transportation, 45, 621-641.
- [8] Shaheen, S., Chan, N., Bansal, A., Cohen, A. (2017). *Mobility and the Sharing Economy: Potential to Facilitate the First- and Last-Mile Public Transit Connections*. Transport Policy, 53, 36-42.
- [9] Alessandretti, L., Sapiezynski, P., Lehmann, S., Baronchelli, A. (2016). *Multi-Scale Spatio-Temporal Analysis of Human Mobility*. PLOS ONE, 11(7), e0158628. Available online: <https://journals.plos.org/plosone/article-id=10.1371/journal.pone.0158628>. Accessed 29 September 2024.
- [10] Boldrini, C., Bruno, R., Laarabi, H. (2021). *Weak Signals in the Mobility Landscape: Car Sharing in Ten European Cities*. Transportation Research Procedia, 52, 13-20.
- [11] Mericskay, B. (2019). *Potentials and Limits of (Geo)digital Footprints in Mobility Analysis: The Example of the Data from the BlaBlaCar Carpooling Platform*. Cybergeog: European Journal of Geography.
- [12] Talandier, M., Duroudier, S., Andr-Poyaud, I., Chardonnel, S., Ployon, E. (2024). *The Territorial Impacts of BlaBlaCar Carpooling*. Journal of Transport Geography, 121, 104041. <https://doi.org/10.1016/j.jtrangeo.2024.104041>
- [13] Bastin, L., Tubaro, P. (2018). *Big Data and the Social Sciences: From Challenges to Opportunities*. Big Data & Society, 5(1). <https://doi.org/10.1177/>

2053951718772960. Accessed 14 June 2025.

- [14] Sellars, A. (forthcoming). *Legality and Ethics of Web Scraping*. Communications of the Association for Information Systems. Available online: https://www.researchgate.net/publication/343555462_Legality_and_Ethics_of_Web_Scraping. Accessed 01 July 2024.
- [15] Van Rossum, G., Drake, F.L. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.
- [16] SeleniumHQ. (2023). *Selenium WebDriver*. Available online: <https://www.selenium.dev/documentation/webdriver/>. Accessed 01 July 2024.
- [17] USITC Gravity Group. (2023). *Gravity Modeling Environment (GME)*. Available online: <https://usitc-gravity-group.github.io/GME>. Accessed 12 October 2024.
- [18] Anderson, J.E., van Wincoop, E. (2003). *Gravity with Gravitars: A Solution to the Border Puzzle*. American Economic Review, 93(1), 170-192. Available online: <https://www.aeaweb.org/articles-id=10.1257/000282803321455214>. Accessed 20 October 2024.
- [19] Lewer, J.J., Van den Berg, H. (2008). *A Gravity Model of Immigration*. Economics Letters, 99(1), 164-167. <https://doi.org/10.1016/j.econlet.2007.06.019>. Accessed 10 November 2024.
- [20] Strangio, D., Armenise, M., Salvucci, G. (2024). *Understanding Internal Migration Patterns Through the Lens of Gravity Models: Evidence from Italian Municipalities*. Journal of Regional Mobility Studies, 12(2), 45-66.
- [21] Santos Silva, J.M.C., Tenreyro, S. (2006). *The Log of Gravity*. The Review of Economics and Statistics, 88(4), 641-658. <https://doi.org/10.1162/rest.88.4.641>. Accessed 20 October 2024.
- [22] Fielding, R.T. (2000). *Architectural Styles and the Design of Network-based Software Architectures*. Ph.D. Thesis, University of California, Irvine. Available online: <https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm>. Accessed 14 June 2025.
- [23] Jacobson, D., Brail, G., Woods, D. (2011). *APIs: A Strategy Guide-Designing Interfaces for a Connected World*. O-Reilly Media, Sebastopol, CA.
- [24] ISTAT. (2011). *Population and Housing Census 2011 Commuter Flow Data*. Available online: <https://www.istat.it/it/censimenti-permanenti/popolazione-e-abitazioni>. Accessed 01 July 2024.
- [25] ISTAT. (2011). *Statistical Archive of Active Enterprises (ASIA)*. Available online: <https://www.istat.it/it/archivio/asia>. Accessed 01 July 2024.

- [26] GeoNames. (2024). *The GeoNames Geographical Database*. Available online: <https://www.geonames.org/>. Accessed 12 October 2024.
- [27] Immobiliare.it. (2024). *House Prices in Italy by Municipality Average Price per Square Meter*. Available online: <https://www.immobiliare.it/prezzi/>. Accessed 12 October 2024.