B Urban Analytics and City Science

EPB: Urban Analytics and City Science
2026, Vol. 0(0) 1–13

DOI: 10.1177/23998083251415040
journals.sagepub.com/home/epb
Sage

spanishoddata: A package for accessing and working with Spanish Open Mobility Big Data

Egor Kotov[1,2], Eugeni Vidal-Tortosa[3,4], Oliva G. Cantú-Ros[5], Javier Burrieza-Galán[5], Ricardo Herranz[5], Tania Gullón Muñoz-Repiso[6] and Robin Lovelace[4]

Abstract

We present `spanishoddata`, an `R` package that enables fast and efficient access to Spain's open, high-resolution origin-destination human mobility datasets, derived from anonymised mobile-phone records and released by the Ministry of Transport and Sustainable Mobility. The package directly addresses challenges of data accessibility, reproducibility, and efficient processing identified in prior studies. `spanishoddata` automates retrieval from the official source, performs file and schema validation, and converts the data to efficient, analysis-ready formats (`DuckDB` and `Parquet`) that enable multi-month and multi-year analysis on consumer-grade hardware. The interface handles complexities associated with these datasets, enabling a wide range of people – from data science beginners to experienced practitioners with domain expertise – to start using the data with just a few lines of code. We demonstrate the utility of the package with example applications in urban transport planning, such as assessing cycling potential or understanding mobility patterns by activity type. By simplifying data access and promoting reproducible workflows, `spanishoddata` lowers the barrier to entry for researchers, policymakers, transport planners or anyone seeking to leverage mobility datasets.

Keywords

human mobility flows, origin-destination data, open source, R package, Spain

[1]Department of Digital and Computational Demography, Max Planck Institute for Demographic Research, Rostock, Germany
[2]Department of Political and Social Sciences, Universitat Pompeu Fabra, Barcelona, Spain
[3]Department of Geography, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain
[4]Leeds Institute for Transport Studies, University of Leeds, UK
[5]Nommon Solutions and Technologies S.L., Madrid, Spain
[6]Subdirección de Planificación, Red Transeuropea y Logística, Ministerio de Transportes y Movilidad Sostenible, Madrid, Spain

Corresponding author:
Egor Kotov, Department of Digital and Computational Demography, Max Planck Institute for Demographic Research, Konrad-Zuse-Str. 1 18057 Rostock, Germany; Department of Political and Social Sciences, Universitat Pompeu Fabra, Jaume I Building (Ciutadella Campus), Ramon Trias Fargas, 25-27, 08005 Barcelona, Spain.
Email: kotov@demogr.mpg.de

Data Availability Statement included at the end of the article

Introduction

Releasing open-access datasets is an enabler of evidence-based decision-making (Uhlir and Schröder, 2007; Zuiderwijk and Janssen, 2014), but is not sufficient: clicking on links to large datasets is often the first step in a long process of data storage, cleaning and processing steps before analysis can begin (e.g. Quiroz et al., 2022; Souibgui et al., 2019). Data cleaning accounts for a large proportion of time spent on data science projects (Mertz, 2021). As noted by the Pragmatic Institute (2022), 'Data practitioners spend 80% of their valuable time finding, cleaning, and organising the data'. Some of the issues the open data movement has failed to solve, such as the 'reproducibility crisis', can be partly blamed on difficulties associated with actually using and understanding such datasets (Granell et al., 2018; Leek and Peng, 2015; Luthfi and Janssen, 2019; Open Science Collaboration, 2012).

Many initiatives seek to reduce the time and associated costs of 'extract, transform, load' (ETL) and 'find, clean, organise' steps. These range from 'top-down' approaches, such as governments setting data standards, to 'bottom-up' approaches led by the community. A popular bottom-up approach is the development of 'data packages' which provide functions and documentation to ease the process of accessing and using data (e.g. Botta et al., 2025; Kockmann and Panse, 2021; Lovelace et al., 2019; Ranghetti et al., 2020; Vengroff, 2022; Walker, 2023; Walker and Herman, 2024).

This article presents the `spanishoddata` `R` package, a user-friendly tool for accessing the Spanish Open Mobility Big Data (Ministerio de Transportes y Movilidad Sostenible, 2022) published by the Spanish Ministry of Transport and Sustainable Mobility (MITMS). From a review of 36 studies using the MITMS mobility data, we found that none shared data-acquisition code and that citation practices were inconsistent, which hinders reproducibility and complicates tracking of data use. Large 'csv.gz' files and the constraints of typical laptops might be limiting the adoption of this data. Our package provides a simple, reproducible access workflow that downloads data from the official source, converts it to formats that increase analysis speed 3-8 times compared to working with the provided data format and reduce processing complexity, and presents it through a data-frame-like interface familiar to R users. The package shortens the time to analysis, enables multi-month and multi-year work on consumer-grade hardware with little coding, and supplies standard wording for data citation. Together, these features improve availability, traceability of use, and reproducibility of analyses using these data. We provide an overview of the datasets, analyse their previous applications, highlight the package's key functions and features designed to address reproducibility challenges, and showcase two practical examples of its use.

Datasets and previous uses

The Spanish Open Mobility Big Data

The main objective of the Spanish Open Mobility Big Data is to support the planning and management of transport infrastructure, services and transport policy in Spain. However, through the 'open mobility data' initiative, the MITMS also makes this data publicly available to promote transparency, efficiency, citizen participation, and innovation in any other field that contributes to the public good and it has been widely used by academia, research centres, and both the private and public sectors.

The datasets are derived from Mobile Network Data (MND), covering both active events (such as calls, text messages, and Internet connections) and passive events (like periodic pings between antennas and phones) from a sample of 13 million subscribers of Orange Spain mobile

phone operator (MNO). According to Comisión Nacional de los Mercados y la Competencia (CMNC, 2024, 2025), this MNO's market share for mobile telephony fluctuated around 20-22% in 2022-2024.

Since the data come from a single mobile carrier, the sample has been expanded to represent the entire Spanish population. The datasets include origin-destination tables for Spanish residents (excluding roaming devices) at various geographic levels and spatial data on zoning boundaries. There are two versions of the datasets available, both supported by the `spanishoddata` package:

- First version (Ministerio de Transportes, Movilidad y Agenda Urbana, 2021): This covers 2020 to 2021, including data from the COVID-19 pandemic. It includes tables detailing the number of trips and distance travelled, broken down by origin and destination at various levels (districts and municipalities), as well as by activity at both the origin and destination (considering three types of activities: home, work or study, and other), residence province, time interval, distance interval, and date. In addition, it provides tables showing the daily number of individuals categorised by location (district or municipality) and the number of trips made per day (0, 1, 2, 2+). A single day of the most detailed records is approximately 70 MB in compressed `CSV` format (`csv.gz`), while the complete time series datasets exceed 22 GB. When loaded into memory in R, 1 day of data occupies approximately 500 MB.
- Second version (Ministerio de Transportes y Movilidad Sostenible, 2024): This version covers the period from 2022 onwards and enhances the first version with even more detailed spatial resolution (3792 districts compared to 2850 in the first version) and coverage of trips to and from foreign countries (detailing separately Portugal and France at the NUTS3 level). It also introduces new fields in the origin-destination tables, including variables for study-related activities, split of other activities into frequent and infrequent, and sociodemographic factors such as income, age, and sex. Additionally, it includes new tables showing the number of individuals by location of overnight stay, residence, and date. The second version of the dataset contains much richer information: the most detailed daily records file reaches approximately 180 MB in compressed `CSV` format (`csv.gz`) and occupies around 2 GB of memory when loaded into an R session. The full dataset, covering 3 years, exceeds 200 GB – and continues to grow as new data is regularly published throughout 2025.

To the best of our knowledge, these datasets are unprecedented in open access worldwide, offering exceptional spatial and temporal coverage along with a wide range of variables.

Previous uses

The Spanish Open Mobility Big Data have already been used in several studies, providing an opportunity to assess their use and identify gaps that the `spanishoddata` package could address. To this end, we searched on multiple academic platforms to identify papers that used the datasets. See details of the search strategies outlined in Supplement 1.

We analysed the full texts, supplements, and accompanying data and code (when available) to extract the research topic and the type of MITMS data used (such as years, spatial aggregation levels, and data types such as mobility flows or overnight stays). We also explored how the datasets were cited and whether the authors used primary or secondary data sources. Additionally, we recorded the extent to which the authors shared their code and the programming languages used. The findings of this analysis and implications for the development of the `spanishoddata` package are detailed in Supplement 2.

Among the 36 articles identified, most focused on COVID-19 and epidemiology, which leaves substantial scope for applications in other fields. Very few articles focused on other topics including human mobility prediction, noise pollution, and air quality. Most studies relied on the 2020–2021 dataset, frequently using detailed hourly origin-destination matrices aggregated to daily levels and analysed at the census district scale. Only two studies so far used post-2021 data.

Citation practices varied across articles, with many referencing the datasets inconsistently through different URLs, page titles, or footnote links. Some articles omitted links to MITMS webpages or directed readers to the Ministry's homepage rather than the dataset. None followed the citation required by the MITMS data licence, hindering reproducibility and limiting the government's ability to track data usage and impact.

We found little evidence of reproducible research practices. None of the articles supplied data-acquisition code. Reproducibility therefore depends on manual downloads and placing files to match the paths and formats assumed in the analysis code. Many articles provided no repository link for code or data, 20% shared pre-processing code, and even fewer provided pre-processed data. Less than one-third of the articles offered analysis or visualisation code. `Python` was used in 44% of the articles and `R` in 30%, with some studies combining both languages. In 22% of the articles, the analysis software used was not specified. Distinct from these research applications, an early effort to improve data accessibility was the *COVID-19 Flow-Maps* project (Ponce-de Leon et al., 2021). They re-hosted and provided API access (and a Python module to work with it) to the initial 2020–2021 dataset during the pandemic. However, as a static snapshot, it does not include the newer, more detailed datasets for 2022 and beyond.

The `spanishoddata R` package

The `spanishoddata R` package directly addresses key issues identified in the literature review – namely, challenges related to data sharing, code reproducibility, and citation practices.

Users can explore codebooks with `spod_codebook()`, check dataset metadata with `spod_available_data()`, and retrieve covered dates with `spod_get_valid_dates()`. Data access requires running `spod_convert()`, specifying the data type (e.g. 'origin-destination'), spatial level (e.g. 'municipalities'), and date range. For example:

```
spod_set_data_dir("d:/users/user_01/documents/spod")

distr_od_db <- spod_convert(
  type = "origin-destination",
  zones = "districts",
  dates = c(
    start = "2023-01-01",
    end = "2023-12-31"
  )
)

distr_od <- spod_connect(distr_od_db)
```

`spod_convert()` efficiently downloads raw data in `csv.gz` format to the directory set by `spod_set_data_dir()`. It supports concurrent downloads, ensures data consistency, and allows resuming of interrupted downloads. Once all requested data are downloaded, using the same base directory or another user defined path the function converts it into a `DuckDB` database via the `duckdb R` package (Mühleisen and Raasveldt, 2024).

`DuckDB` (Raasveldt and Muehleisen, 2018) is a modern, serverless database engine that can work with a folder of `CSV` or `Parquet` files, as well as its own file database format. It requires no complex setup or prior database knowledge from the user, especially when it acts as a backend in a package like `duckplyr` (Mühleisen and Müller, 2025) or our own `spanishoddata`. By streaming data in chunks, reading only the needed columns from disk when necessary, and making optimal use of random access memory (RAM) and available processor cores, `DuckDB` avoids loading entire datasets at once. This dramatically improves analysis speed and makes it possible to analyse months or years of data on a consumer-grade laptop. All of this is done automatically without user having to set up complicated workflows and optimising them manually. Therefore, this advanced backend is transparent to the user: `spod_convert()` returns the local path to the resulting `DuckDB` database file and then the user connects to it with `spod_connect()` to obtain the `distr_od` object. This object behaves like a `data.frame` or `tibble` and is compatible with `dplyr` (Wickham et al., 2023) functions familiar to many R users, such as `select()`, `filter()`, `mutate()`, `group_by()`, `summarise()` and more.

The `spod_convert()` function also supports conversion to `Parquet`, another widely used columnar data format that is slightly slower than `DuckDB`. Benchmarks in Supplement 3 show up to a 3–4x reduction in aggregation time with `DuckDB` format relative to `Parquet` and `CSV`, and additional benchmarks with different queries in the online package vignettes yielded 5–8x speedups. Performance varies by task and hardware, but across our tests `DuckDB` format was the most consistently performant general-purpose choice. Both the `Parquet` files and the `DuckDB` database file are accessible via `Python` `duckdb` module, the `duckdb` command-line client, and multiple programming languages (the up-to-date list of supported platforms and languages is available at https://duckdb.org/), allowing users to analyse the acquired and converted data outside `R`, essentially using `spanishoddata` as a tool that ensures data consistency before beginning the analysis.

`spod_convert()` can be replaced with `spod_get()` (using the same function arguments as above) to quickly analyse a few days of raw `csv.gz` origin-destination data (approximately up to 1 week) without converting it to another format. Both functions perform on-the-fly translation of variable names to English, enforce appropriate data types to dataset variables, enrich the source data with additional text labels of originally encoded variables, such as province of residence. All data transformations are documented in codebooks supplied with the package and on its website. With the same arguments, `spod_download()` can download the data without converting it or connecting it for analysis.

If a user needs to analyse total daily flows by age group over an entire year, they need to run:

```
daily_flows_by_age <- distr_od |>
  group_by(date, age) |>
  summarise(
    total_flows = sum(number_of_trips)
  ) |>
  collect()
```

To analyse the spatial aspect of the data, `spod_get_zones()` retrieves mobility zone polygons matching the origin-destination data by zone identifiers, returning an `sf` object familiar to `R` users working with spatial data. The function supports all available tessellations, checks for and corrects invalid geometries, and automatically adds columns with zone identifiers that link different tessellations both within dataset periods (2020–2021 and 2022 onwards) and across different periods.

```
distr_zones <- spod_get_zones(
  zones = "distr",
  ver = 2
)
```

The examples above demonstrate how researchers can incorporate a few lines of code into a reproducible script bundled with a research paper, simplifying data retrieval, pre-processing, and ensuring consistency while clearly documenting the data used and its source for reproducibility.

The `spod_cite()` function generates citations for the package and the underlying datasets in plain text, markdown, and BibTeX, ensuring consistent referencing and facilitating data provider tracking of usage and impact.

For more details about suggested package workflows and core functions, see Supplement 3, along with the package documentation, data codebooks, and vignettes available at https://ropenspain.github.io/spanishoddata/.

Examples of package and dataset applications

Below we present two practical examples of how the Spanish Open Mobility Big Data can be used in research. We provide the reproducible code using the `spanishoddata` package for both examples on Zenodo: https://doi.org/10.5281/zenodo.15207374 (Kotov et al., 2025).

Assessing cycling potential and infrastructure provision in Valencia

The Spanish Open Mobility Big Data do not contain information about transport modes, but they do contain a `distance` column with categories including 0.5-2 and 2-10 km. Additionally, the straight-line distances can be computed by measuring the Euclidean distance between the centroids of the origin and destination zones, providing a useful approximation of the actual travel path. This distance information can be used to estimate active travel potential based on known relationships between travel distance and mode choice (e.g. Iacono et al., 2010). The example below, applied to the city of Valencia, shows how the spanishoddata package can be used to infer cycling potential and compare it to existing infrastructure to identify opportunities for improving the cycle path network.

Figure 1 A (intra-zonal analysis) illustrates the cycling potential, considering all trips made in the first week of May 2024 within the 0.5-10 km range that occur within the same district. The cycling network from OpenStreetMap (OSM) is overlaid to visually assess how well the infrastructure aligns with cycling potential across districts. Figure 1 B (inter-zonal analysis) shows the desire lines between 0.5 and 10 km in length. The width of each line corresponds to the number of trips, while the colour indicates the connectivity ratio of the cycling network, measuring the connections of the cycling infrastructure relative to the length of the district border.

Approaches extending this illustrative example, building on methods used in a prior study in Barcelona (IERMB, 2020) with the first version of the data, could guide improvements in cycling network planning nationally.

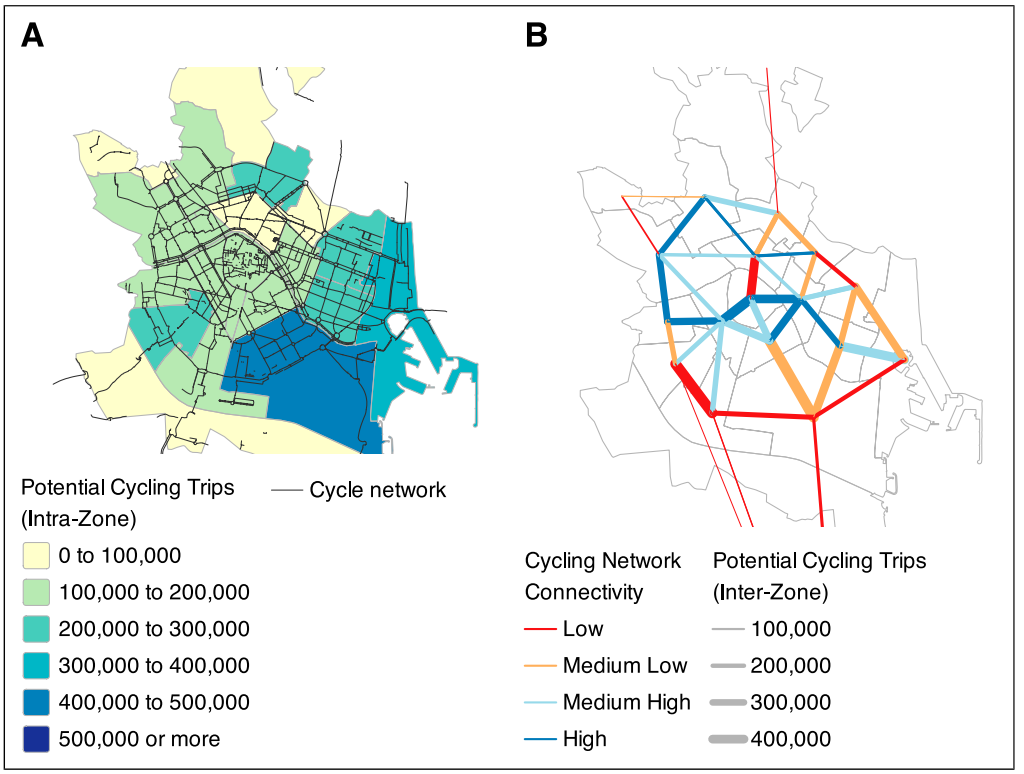

Figure 1. Cycling potential and infrastructure provision in Valencia.

Exploring work vs non-work related trips in Madrid

The Spanish Open Mobility Big Data include trip type classifications based on the activities at trip origins and destinations. These categories – '*home'*, *'work or study'*, and locations of either *'frequent'* or *'infrequent'* activity – enable detailed analyses of travel patterns. This distinction supports many use cases, including comparing work-related trips with non-work trips to inform public transport scheduling during peak travel periods.

We illustrate this by analysing mobility flows for 1 week in Madrid and its surrounding areas in February 2023 using the `flowmapper` `R` package (Mast, 2024) for visualisation. Panels A and D in Figure 2 clearly depict the expected reverse directions of commuting trips between the morning (7–11) and evening (17–21) periods. Notably, the maximum number of trips in the evening (Panel D) is lower compared to the morning (Panel A), which may indicate less predictable mobility in the evening, as individuals may take detours or engage in additional activities; evening commuting is generally known to be more spread out over time (Giménez-Nadal et al., 2021; Gorný, 2024). Panels B and E illustrate journeys between 'home' or 'work or study' and 'frequent' destinations. Because the classification rule for 'work' excludes irregular or shift-based schedules, many genuine work trips end up mixed into this 'frequent' category. At the same time, regularly attended non-work activities, such as gym visits or childcare drop-offs, may also be classified as 'frequent'. The right-hand panels (C and F) isolate trips to 'infrequent' destinations, which are lower in volume and exhibit a broader spatial spread characteristic of errands, social visits, or leisure activities. These insights have multiple applications, including planning and scheduling public transport services

Figure 2. Travel patterns in Madrid by trip type.

during peak travel times and identifying optimal locations for new facilities (e.g. shops and bars) based on the detected mobility patterns and activity types.

Discussion and conclusion

To the best of our knowledge, the Spanish Open Mobility Big Data – made available through the open mobility data initiative of the MITMS – constitute the first publicly accessible national-scale origin-destination mobility data source of its kind. However, their potential to support scientific research is limited by their size and complexity, meaning that they are out-of-reach to many researchers and costly to work with in their raw form downloaded manually from the website. The `spanishoddata` package solves these problems by providing a user-friendly, reproducible, and

performant way to access and work with them. With new technologies and data formats, such as `DuckDB` and `Parquet`, the package reduces the entry barrier for people who want to gain insight and guide policy with the data. As concrete evidence of uptake, the package recorded 2200+ downloads on the Comprehensive R Archive Network (CRAN) from December 2024 to September 2025 (this excludes some CRAN mirrors and R-universe installations, does not equal unique users, and may include automated/CI installs). If 30% of downloads correspond to real users, both working on real projects and exploring the data for their future ideas, and the package saves approximately 1 to 3 hours of initial data preparation time per project, this would redirect roughly 660 to 1980 researcher-hours from wrangling to analysis. The 1 to 3 hour assumption is intentionally conservative; writing and testing data download, processing and validation scripts often takes longer, even with the help of large language models, and the development and validation effort for the package was substantially larger. Our review identified more than 30 studies using the 2020–2021 MITMS mobility data, suggesting possibly higher future demand for the more comprehensive post-2021 data. More importantly, the package enables new research projects and a wide variety of applications in both the public and private sectors that would be impossible without it, allowing users to focus on rewarding tasks, such as analysis and modelling, rather than on data cleaning and import.

Indeed, the package has already been highlighted in several scientific papers (Pardo-Araujo et al., 2026; Renninger and Cabrera-Arnau, 2025). To ensure software sustainability and encourage contributions from the community, the package has been 'adopted' by the `rOpenSpain` project, which will help maintain and develop the package in the future. The package is stable and ready to use, and can be installed on any modern computer with the following command

```
install.packages("spanishoddata")
```

More broadly, our work represents *an approach* to enabling access to analysis-ready datasets from online sources with the following principles:

- Do not re-upload the data; always download from the official source to maintain trust, buy-in from data owners, and transparency. This prevents the data from being outdated if the authors stop the development of the project Ponce-de Leon et al. (2021).
- Convert the data to an efficient format, such as `DuckDB` or `Parquet`, to make it easier and faster to work with.
- Provide a user-friendly interface to the data, including tutorials and extensive documentation (see the package website).
- Ensure that the data is cited correctly and that the package is also cited when used.

The approach could be improved in several ways, such as porting the functionality to other languages (see #144 in the package issue tracker) and providing more features (e.g. plotting functions). Since we released `spanishoddata`, work has begun on the Python package `pySpainMobility` Beneduce et al. (2025), which follows the same approach to accessing the data. Our experiences also provide insights to data owners and raise the question: if `Parquet` is such a good file format for efficient data provision, why not provide the data in this format from the start? There are hundreds of open-access datasets provided as `CSV` files and other plain text formats in need of cleaning, however, and we are confident that the approach can usefully be applied to many other datasets. We hope that other projects will adopt this approach and look forward to seeing how the `spanishoddata` package is used by researchers, practitioners and wider open science, open source and other communities who can benefit from the work.

Acknowledgements

We thank Lluís Revilla Sancho, Iñaki Ucar, and Marta Pardo-Araujo for early beta-testing of the package and for providing valuable feedback, which helped resolve several data retrieval and scripting bugs.

ORCID iDs

Egor Kotov https://orcid.org/0000-0001-6690-5345
Eugeni Vidal-Tortosa https://orcid.org/0000-0001-5199-4103
Oliva G. Cantú-Ros https://orcid.org/0000-0002-5809-5546
Ricardo Herranz https://orcid.org/0000-0002-3663-0747
Robin Lovelace https://orcid.org/0000-0001-5679-6536

Ethical considerations

There are no human participants in this article and informed consent is not required.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

The **spanishoddata** R package can be installed from CRAN https://doi.org/10.32614/CRAN.package.spanishoddata (Kotov et al., 2025). The package source code is available at: https://github.com/rOpenSpain/spanishoddata/. Detailed examples and tutorials on acquiring the data and making static and interactive visualisations are available at the package website https://ropenspain.github.io/spanishoddata/. The mobility data accessed through the package is based on open data of The Spanish Ministry of Transport and Sustainable Mobility (Ministerio de Transportes y Movilidad Sostenible). This source data can be accessed at https://www.transportes.gob.es/ministerio/proyectos-singulares/estudio-de-movilidad-con-big-data (Ministerio de Transportes y Movilidad Sostenible, 2022). The code, data, and containerised computational environment used to generate figures in the paper and the supplement are available in the Zenodo repository at https://doi.org/10.5281/zenodo.15207374 (Kotov et al., 2025).

Supplemental material

Supplemental material for this article is available online.

References

Beneduce C, Gullón Muñoz-Repiso T, Lepri B, et al. (2025) pySpainMobility: a Python package to access and manage Spanish open mobility data. *Preprint. arXiv*. Available at: https://doi.org/10.48550/arXiv.2506.13385

Botta F, Lovelace R, Gilbert L, et al. (2025) Packaging code and data for reproducible research: a case study of journey time statistics. *Environment and Planning B: Urban Analytics and City Science* 52(4): 1002–1013. Available at: https://doi.org/10.1177/23998083241267331

Comisión Nacional de los Mercados y la Competencia (CMNC) (2024) La portabilidad móvil superó el medio millón de cambios de operador en abril. https://www.cnmc.es/prensa/datos-telecos-mensual-abril-20240626 (accessed 2025-09-26).

Comisión Nacional de los Mercados y la Competencia (CMNC) (2025) Telecomunicaciones anual mercados. https://data.cnmc.es/telecomunicaciones-y-sector-audiovisual/datos-anuales/datos-de-mercados/telecomunicaciones-anual (accessed 2025-09-26).

Giménez-Nadal JI, Molina JA and Velilla J (2021) Two-way commuting: asymmetries from time use surveys. *Journal of Transport Geography* 95: 103146. Available at: https://doi.org/10.1016/j.jtrangeo.2021.103146

Gorný D (2024) Temporal displacement and spatial unbinding of commuting in the Brno metropolitan area. *The Geographical Journal* 190(4): e12584. Available at: https://doi.org/10.1111/geoj.12584

Granell C, Nüst D, Ostermann FO, et al. (2018) Reproducible research is like riding a bike. *PeerJ Preprints*. Available at: https://doi.org/10.7287/peerj.preprints.27216v1

Iacono M, Krizek KJ and El-Geneidy A (2010) Measuring non-motorized accessibility: issues, alternatives, and execution. *Journal of Transport Geography* 18(1): 133–140. Available at: https://doi.org/10.1016/j.jtrangeo.2009.02.002

IERMB (2020) *Fluxos Potencials per a l'ús de la Bicicleta en l'àmbit Metropolità: Tractament de Dades Provinents de Telefonia Mòbil*. Technical Report. Institut d'Estudis Regionals i Metropolitans de Barcelona. https://hdl.handle.net/20.500.14439/1144 (accessed 2024-11-10).

Kockmann T and Panse C (2021) The rawrr R package: direct access to Orbitrap data and beyond. *Journal of Proteome Research* 20(4): 2028–2034. Available at: https://doi.org/10.1021/acs.jproteome.0c00866

Kotov E, Lovelace R and Vidal-Tortosa E (2025) spanishoddata: Get Spanish Origin-Destination Data. R package version 0.1.1. *CRAN*. https://doi.org/10.32614/CRAN.package.spanishoddata (accessed 2025-04-09).

Kotov E, Vidal-Tortosa E, Cantú-Ros OG, et al. (2025) Supplement data, code, and computational environment for 'spanishoddata: A package for accessing and working with Spanish Open Mobility Big Data'. *Zenodo*. https://doi.org/10.5281/zenodo.15289979 (accessed 2025-04-27).

Leek JT and Peng RD (2015) Reproducible research can still be wrong: adopting a prevention approach. *Proceedings of the National Academy of Sciences of the United States of America* 112(6): 1645–1646. Available at: https://doi.org/10.1073/pnas.1421412111

Lovelace R, Morgan M, Hama L, et al. (2019) stats19: a package for working with open road crash data. *Journal of Open Source Software* 4(33): 1181. Available at: https://doi.org/10.21105/joss.01181

Luthfi A and Janssen M (2019) Open data for evidence-based decision-making: data-driven government resulting in uncertainty and polarization. *International Journal of Advanced Science, Engineering and Information Technology* 9(3): 1071–1078. Available at: https://doi.org/10.18517/ijaseit.9.3.8846

Mast J (2024) *flowmapper: draw flows (migration, goods, money, information) on 'ggplot2' plots*. R package version 0.1.3. CRAN. https://doi.org/10.32614/CRAN.package.flowmapper (accessed 2024-11-15).

Mertz D (2021) *Cleaning Data for Effective Data Science: Doing the Other 80% of the Work with Python, R, and command-line Tools*. Packt.

Ministerio de Transportes y Movilidad Sostenible (2022) *Estudio de la Movilidad con Big Data (Study of Mobility With Big Data)*. Technical Report. Ministerio de Transportes y Movilidad Sostenible. https://www.transportes.gob.es/ministerio/proyectos-singulares/estudio-de-movilidad-con-big-data (accessed 2024-12-11).

Ministerio de Transportes y Movilidad Sostenible (2024) Estudio de movilidad de viajeros de ámbito nacional aplicando la tecnología Big Data. In: *Informe Metodológico (Study of national traveler mobility using Big Data technology. Methodological report)*. Technical Report. Ministerio de Transportes y Movilidad Sostenible. https://www.transportes.gob.es/recursos_mfom/paginabasica/recursos/a3_informe_metodologico_estudio_movilidad_mitms_v8.pdf (accessed 2024-12-11).

Ministerio de Transportes, Movilidad y Agenda Urbana (2021) *Análisis de la Movilidad en España con Tecnología Big Data Durante el estado de Alarma Para la Gestión de la crisis del COVID-19 (Analysis of mobility in Spain with Big Data Technology During the State of Alarm for COVID-19 Crisis Management)*. Technical Report. Ministerio de Transportes, Movilidad y Agenda Urbana. https://cdn.mitma.

gob.es/portal-web-drupal/covid-19/bigdata/mitma_-_estudio_movilidad_covid-19_informe_metodologico_v3.pdf (accessed 2024-07-23).

Mühleisen H and Müller K (2025) *duckplyr: a 'DuckDB'-backed version of 'dplyr'*. R package version 1.0.0. CRAN. https://doi.org/10.32614/CRAN.package.duckplyr (accessed 2025-02-07).

Mühleisen H and Raasveldt M (2024) *duckdb: DBI package for the DuckDB database management system*. R package version 1.1.2. CRAN. https://doi.org/10.32614/CRAN.package.duckdb (accessed 2024-10-30).

Open Science Collaboration (2012) An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science* 7(6): 657–660. Available at: https://doi.org/10.1177/1745691612462588

Pardo-Araujo M, Kotov E, Alonso D, et al. (2026) Understanding mosquito vector invasion pathways: synergistic effects of human mobility, climate, and natural dispersal. *Ecology Letters*, In press.

Ponce-de Leon M, del Valle J, Fernandez J, et al. (2021) COVID-19 flow-maps an open geographic information system on COVID-19 and human mobility for Spain. *Scientific Data* 8(1): 310. Available at: https://doi.org/10.1038/s41597-021-01093-5

Pragmatic Institute (2022) Overcoming the 80/20 rule in data science. https://www.pragmaticinstitute.com/resources/articles/data/overcoming-the-80-20-rule-in-data-science/ (accessed 2024-09-10).

Quiroz JC, Chard T, Sa Z, et al. (2022) Extract, transform, load framework for the conversion of health databases to OMOP. *PLoS One* 17(4): e0266911. Available at: https://doi.org/10.1371/journal.pone.0266911

Raasveldt M and Muehleisen H (2018) DuckDB. https://github.com/duckdb/duckdb (accessed 2024-08-08).

Ranghetti L, Boschetti M, Nutini F, et al. (2020) 'sen2r': an R toolbox for automatically downloading and preprocessing Sentinel-2 satellite data. *Computers & Geosciences* 139: 104473. Available at: https://doi.org/10.1016/j.cageo.2020.104473

Renninger A and Cabrera-Arnau C (2025) Spanish heat waves curb discretionary mobility and alter work behavior. *Preprint. arXiv*. Available at: https://doi.org/10.48550/arXiv.2501.03978

Souibgui M, Atigui F, Zammali S, et al. (2019) Data quality in ETL process: a preliminary study. *Procedia Computer Science* 159: 676–687. Available at: https://doi.org/10.1016/j.procs.2019.09.223

Uhlir PF and Schröder P (2007) Open data for global science. *Data Science Journal* 6: OD36-OD53. Available at: https://doi.org/10.2481/dsj.6.od36

Vengroff D (2022) censusdis: US census utilities for a variety of data loading, analysis, and mapping purposes. https://github.com/vengroff/censusdis (accessed 2024-07-10).

Walker K (2023) *Analyzing US Census Data: Methods, Maps, and Models in R. Chapman & Hall-CRC the R Series*. 1st edition. CRC Press.

Walker K and Herman M (2024) tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames. R package version 1.6.6. *CRAN*. https://doi.org/10.32614/CRAN.package.tidycensus (accessed 2024-09-20).

Wickham H, François R, Henry L, et al. (2023) *dplyr: a grammar of data manipulation*. R package version 1.1.4. CRAN. https://doi.org/10.32614/CRAN.package.dplyr (accessed 2024-10-30).

Zuiderwijk A and Janssen M (2014) Open data policies, their implementation and impact: a framework for comparison. *Government Information Quarterly* 31(1): 17–29. Available at: https://doi.org/10.1016/j.giq.2013.04.003

Author biographies

Egor Kotov a is a doctoral researcher in the Laboratory of Migration and Mobility at the Max Planck Institute for Demographic Research and in the Department of Political and Social Sciences at the Universitat Pompeu Fabra. His work centers on examining human mobility patterns to address socio-economic, environmental, and public health challenges.

Eugeni Vidal-Tortosa is a mobility researcher and consultant with a diverse academic background. He earned his PhD from the University of Leeds in November 2021, focusing on the relationship between cycling and socio-economic disadvantage. From April 2022 to March 2024, he worked as a Research Fellow in Active Travel and Road Safety Data Science at the Institute for Transport Studies (ITS), University of Leeds. Currently, he is a Postdoctoral Researcher at the Mobility, Transport and Territory Research Group (GEMOTT), Universitat Autònoma de Barcelona.

Oliva G. Cantú-Ros, Chief Research & Development Officer at Nommon, graduated in Physics from the Universidad Nacional Autónoma de México. She holds a Master in Advanced Mathematics from the University of Cambridge and a PhD in Theoretical Physics from Imperial College London. Her doctorate research focused on quantum phase transitions. From 2008 to 2012 she worked as a postdoctoral researcher in several areas of theoretical physics at Universidad Complutense de Madrid (UCM) and Universidad Carlos III (UC3M). In October 2012 she joined Nommon, where she works on the application of complex systems theory and artificial intelligence to the study of social and economic systems. She has participated in several European research projects as a coordinator and as a researcher where she has investigated the use of network theory, big data analysis and spatial interaction models to characterise regional and urban systems.

Javier Burrieza-Galán is a Civil Engineer specialized in the application of data analysis and artificial intelligence techniques to transport and mobility systems. His research interests include geolocation data analysis, data fusion techniques for characterising mobility patterns and the application of scenario analysis to transport planning.

Ricardo Herranz is co-founder and CEO of Nommon. He graduated as an Electronics and Control Engineer from the Technical University of Madrid (UPM) and as a Telecommunications Engineer from the École Supérieure d'Electricité (Supélec) and holds a Master in Quantitative Finance and Business Administration from the UPM. He has 25 years of experience as an engineer, researcher and entrepreneur in the transport and mobility industry. He has coordinated several pioneering European research projects on the use of big data for travel demand analysis and is the author of more than 20 scientific publications in this field.

Tania Gullón Muñoz-Repiso is a geodesy and cartography engineer. Currently, she is a Project Manager specializing in Geographic Information Systems (GIS) and spatial data science applied to transportation and mobility. She is in charge of the national Spanish mobility analyses using big data and is also responsible for the National Transport Infrastructure Network GIS. She is deeply involved in innovation and coordination of initiatives related to transport data governance and transport planning.

Robin Lovelace is Associate Professor of Transport Data Science at the Leeds Institute for Transport Studies (ITS) and Head of Data Science at the UK government agency Active Travel England. Robin specializes in data science and geocomputation, with a focus on modeling transport systems, active travel, and decarbonisation.