MOX-Report No. 69/2024

Estimation of dynamic Origin–Destination matrices in a railway transportation network integrating ticket sales and passenger count data

Galliani, G.; Secchi, P.; Ieva, F.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it https://mox.polimi.it

Estimation of dynamic Origin–Destination matrices in a railway transportation network integrating ticket sales and passenger count data

Greta Galliani [a], Piercesare Secchi [a], Francesca Ieva [a,b,*]

[a] MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan, 20133, Italy
[b] HDS, Health Data Science Center, Human Technopole, Viale Rita Levi-Montalcini, 1, Milan, 20157, Italy

ARTICLE INFO

Dataset link: https://github.com/GretaGalliani/dynamic-OD-estimation-railway-network

Keywords:
Data fusion
Origin–destination matrix
Railway network
Sustainable transport planning
Trip distribution modelling

ABSTRACT

Accurately estimating Origin–Destination matrices is a pressing challenge in transportation management and urban planning. However, traditional methods like travel surveys have limitations in availability and comprehensiveness, which have been further exacerbated by the recent changes in mobility patterns induced by the COVID-19 pandemic. To address this issue, we focused on the Trenord railway network in Lombardy, Italy, and developed an innovative pipeline to integrate ticket and subscription sales and Automated Passenger Counting data using the Iterative Proportional Fitting algorithm. By effectively navigating the complexities of diverse and incomplete data sources, our approach showcases adaptability across various transportation contexts. Our research offers a valuable tool for operators, policymakers, and researchers, bridging the gap between data availability and the need for precise OD matrices. Additionally, we emphasise the potential of dynamic OD matrices and showcase methods for detecting anomalies in mobility trends, interpreting them in the context of events from the last months of 2022.

1. Introduction

The accurate estimation of passenger movements within transportation networks plays a pivotal role in urban planning, infrastructure management and precision policies formulation. Origin–Destination (OD) matrices, depicting the flow of passengers between various locations within a network, serve as a fundamental tool to capture travel patterns, providing insights into the dynamics of passenger flows across different regions and modes of transport. Such matrices find applications in diverse fields, from optimising public transportation services to analysing the environmental impacts of commuting behaviours (Mohammed and Oke, 2023).

In recent years, the estimation of OD matrices went through remarkable advancements following the introduction of Automated Data Collection Systems (ADCS) in several transportation networks (Mohammed and Oke, 2023). This promoted the development of digitalised control systems able to predict flows of passengers (Yang et al., 2020). However, most works in the field have been anchored in data sources like Automated Fare Collection (AFC) and travel surveys (Ait-Ali and Eliasson, 2019; Cui, 2006; Gordon, 2012; Mohammed and Oke, 2023; Torti et al., 2021; Wang, 2010; Zhao et al., 2007), which provide valuable information for modelling transportation networks but are not always comprehensive or available for all systems. This poses significant challenges

* Corresponding author at: MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan, 20133, Italy.
E-mail address: francesca.ieva@polimi.it (F. Ieva).

https://doi.org/10.1016/j.tra.2024.104246

and limitations to complex and multifaceted transportation ecosystems, whenever a systematic and semi-automatised monitoring system is concerned.

One way to overcome the issue deriving from the missingness of AFC data might consist of developing suitable models for estimating the missing information (namely *gap filling*) and/or to rely on integration of multiple and diversified data sources (namely *data fusion*), like ticket and subscription sales. Despite helping in solving the problem, such strategies add layers of methodological complexities (Welch and Widita, 2019). In fact, while extensive works have been dedicated to trip distribution modelling (see Mohammed and Oke, 2023 and references therein), to the best of our knowledge no contribution deals with the challenge of utilising ticket and subscription sales as surrogate information when AFC systems data are unavailable.

In this paper we integrate data from ticket and subscription sales with those obtained by Automated Passenger Counting (APC) - specifically, data generated by sensors which record passengers alighting and boarding a train at each stop - supported by the evidence that their availability in transportation networks is expected to increase in the near future (Siebert and Ellenberger, 2020), acting as a potential bridge for the gap between available data sources and the demand for accurate, dynamic OD matrices for transportation networks.

In fact, this study takes on the challenge of estimating dynamic OD matrices in the context of a railway transportation network, focusing on the Trenord network operating in Lombardy, Italy. The Trenord network plays a critical role in the mobility landscape of the Lombardy region, connecting urban centres and facilitating daily commutes (Trenord, 2023). However, accurate OD matrix estimation for such intricate systems presents obstacles, such as incomplete datasets and the complex interplay between ticket and subscription sales and passenger counts. Indeed, while tickets and subscriptions are sold reporting origins and destinations for the trips, quantitative techniques are needed to assign such trips to specific time frames and correct for trips happening without purchasing a travel title, accounting for the number of passengers boarding and alighting at each station. In response to these challenges, we propose an innovative pipeline that combines ticket and subscription sales data and passenger counts collected through APC systems. Our approach addresses data limitations and integrates in a data fusion perspective heterogeneous information sources to deliver accurate dynamic OD matrices that capture the intricate patterns of passenger movement within the Trenord network.

The main contribution of this work lies in the development of a pipeline for the estimation and the on-the-fly monitoring of dynamic OD matrices, enhancing the construction of OD seeds through the data fusion of different sources of data (e.g., ticket sales data). This multi-step process involves converting ticket and subscription sales records into estimated OD trips, separating trips necessitating transfers, and predicting missing ticket data through gravity models. When combined with passenger count information collected by APC, the resultant seed matrices undergo iterative refinement using the Iterative Proportional Fitting (IPF) algorithm (Evans, 1970; Macgill, 1977) to generate dynamic OD matrices. Crucially, our pipeline transcends specific data types, offering a versatile solution applicable to various transportation contexts and adaptable to the desired network's particular characteristics and available data. We show the efficacy of our pipeline by applying it to estimate weekly OD matrices describing trips by train through six train lines of the Trenord network in the period from June to December 2022. Moreover, we showcase tools to perform almost real-time oversight of complex dynamical systems, aiming to provide methodologies to identify anomalies at global and local station levels in the complex network under analysis. Through this study, we aspire to contribute to the field of transportation network analysis, offering an adaptable approach for the dynamic estimation of OD matrices, in the intricate context of modern urban mobility. Indeed, the importance of high frequency updates of mobility information has been recently underscored by the global pandemic, which has disrupted travel patterns, altered commuter behaviour, and prompted shifts in urban dynamics (de Palma et al., 2022). As cities strive to recover and build resilient transportation systems post-pandemic, accurate mobility insights are indispensable for informed decision-making (Hu et al., 2021).

The rest of the paper is organised as follows: Section 2 analyses past literature about the estimation of static and dynamic OD matrices in transportation networks, considering various data sources; Section 3 gives some context about the Trenord network and presents the available data to tackle the OD matrices estimation problem; Section 4 recalls some theoretical notions about two well-known techniques in the field of trip distribution modelling, namely gravity models and the IPF algorithm and then proceeds to present the pipeline we developed in detail; Section 5 shows the results obtained applying this pipeline to estimate weekly OD matrices describing mobility in seven months of 2022 through six train lines of the Trenord network and presents an application of the derived OD matrices to perform anomaly detection in the network; Section 6 discusses the results obtained, highlighting their strengths and limitations and draws the conclusion of our work.

2. Literature review

This Section overviews methodologies and approaches for estimating static and dynamic OD matrices, highlighting their strengths and limitations in view of the application domain they come from, discussing recent advancements and spotting the unmet needs justifying the introduction of the approach proposed in this work.

The pursuit of estimating OD matrices in transportation networks dates back to the late 1970s (Low, 1972; Robillard, 1975; Wilson, 1970). Early efforts heavily relied on surveys and manual data collection, posing challenges due to the sporadic availability of passenger surveys, mainly because of their cost and relevance limitations (Ben-Akiva and Morikawa, 1989). Recently, advancements in Automatic Data Collection Systems (ADCS) and mobile communication data collection have transformed the landscape. An extensive literature review on the topic may be found in Mohammed and Oke (2023). Technologies like Automatic Fare Collection (AFC), Automatic Vehicle Location (AVL), and Automatic Passenger Counting (APC) have significantly enhanced the accuracy and efficiency of OD estimation. Examples of the use of such technology for OD estimation can be found in Ait-Ali and Eliasson (2019),

Cui (2006), Gordon (2012), Liu et al. (2021), Torti et al. (2021), Wang (2010), Zhao et al. (2007). Despite not initially designed for integrated OD estimation, these technologies offer rich data, reduce collection costs, provide larger sample sizes than surveys, and automate the estimation process (Zhao, 2004). As a result, the availability of OD matrices in the future is expected to significantly increase (Cui, 2006).

Among these technologies, AFC systems, particularly smart card data, have become the focal point of recent research (Wu et al., 2021; Yang et al., 2020; Chen and Fan, 2020). Smart cards are widely used in transportation networks across major cities such as Milan, Paris, London, New York, Boston, Beijing, and Hong Kong (Torti et al., 2021; Zannat and Choudhury, 2019; Hussain et al., 2021), offering substantial and almost real-time insights into people's movements. Consequently, smart cards are increasingly replacing surveys in tackling OD estimation challenges. However, estimating accurate OD data for bus or railway networks without AFC systems remains challenging. A potential solution involves leveraging mobile phone data (e.g., Wi-Fi traces, call data records, or global positioning system for mobile communication data) combined with AVL to infer boarding and alighting passengers at stops (Ge and Fukuda, 2016; Jafari Kang et al., 2020; Munizaga and Palma, 2012). Despite their usefulness, accessing such data is however hampered by privacy concerns (H° akegård et al., 2018).

Our work develops an alternative approach to estimate, with high time frequency, OD matrices. It exploits data readily available to most transportation operators equipped with APC systems but not disposing of AFC ones. Specifically, we leverage ticket and subscription data from a railway network, developing appropriate assumptions to convert this data into OD seeds. These seeds are then combined with passenger counts collected by the APC system to correct the partial estimates accounting for the totality of trips happening in the network.

In general, information provided by ADCS may be incorporated into a model for OD estimation in a twofold way. On one hand, AFC or APC counts may be integrated into iterative algorithms, in a data fusion flavour, to generate seeds of matrices estimation and/or adjust marginal counts. This approach is becoming more and more popular due to the growing availability of data coming from automated digital counting systems. On the other hand, a more traditional approach to tackle the OD estimation challenge comes from the field of trip distribution modelling. A broad review of models for human mobility is presented in Barbosa et al. (2018). Among them, growth factor, genetic algorithms and gravity models are some of the most widely applied techniques (Mohammed and Oke, 2023). Growth factor models involve updating a seed OD matrix based on growth rates calibrated to match alighting and boarding counts. In this class of models, the Iterative Proportional Fitting (IPF) (Evans, 1970; Macgill, 1977), also known as the Furness method, is the state-of-the-art approach for estimating OD flows (Ji et al., 2014) and has been employed in several works in the field (Cui, 2006; Gordon, 2012; Liu et al., 2021; Torti et al., 2021; Zhao et al., 2007). Gravity models, on the other hand, estimate OD matrices using a gravitational attraction model, incorporating population masses and distance or travel cost measures. They assume a positive association between flow volume and population size and factors in the effects of distance, space, cost, or travel time on interactions (Wheeler, 2005). Other works rely on entropy maximisation, proven to be an equivalent approach to gravity models (Ait-Ali and Eliasson, 2019; Ge and Fukuda, 2016; Wong, 2005). The field of trip distribution modelling also encompasses various other techniques, including maximum likelihood estimation (Cui, 2006; Navick and Furth, 1994; Wu et al., 2021), constrained generalised least squares (Lam et al., 2003), Bayesian estimation (H° akegård et al., 2018; Hazelton, 2010; Huo et al., 2023), neural networks (Mussone and Matteucci, 2013; Toqué et al., 2016; Pamuła and Żochowska, 2023) and Genetic Algorithms (Yun and Park, 2005).

Lastly, it is important to notice that many studies focus on static OD matrix estimation, describing trips between zones within a fixed reference period. However, the increasing abundance of data provided by the introduction of ACDS prompts interest in dynamic OD matrix estimation, introducing a temporal dimension which shifts the focus on monitoring mobility patterns instead of estimating them only (Khoshkhah et al., 2022). Literature in this area is still quite limited (Ait-Ali and Eliasson, 2019; Bierlaire and Crittin, 2004; Cerqueira et al., 2022; Zeng et al., 2015; Fujita et al., 2017) and divides into dynamic a posteriori estimates (Ait-Ali and Eliasson, 2019) and real-time estimation for short-term prediction (Bierlaire and Crittin, 2004; Zeng et al., 2015). Our proposed dynamic procedure estimates weekly OD matrices over six months, thus placing itself in the a posteriori estimates field.

3. Data

This Section introduces the Trenord network and the data employed in estimating weekly OD matrices for a specific network segment.

3.1. The Trenord network

Established on May 3, 2011, as a collaboration between Ferrovie Nord Milano and Trenitalia, Trenord is one of the major local rail transport entities in Europe. It boasts an extensive 2000-km network, interlinking 460 stations, and operates over 2170 daily trips, serving the Lombardy region in Italy and seven adjoining provinces, along with Malpensa International Airport through the Malpensa Express rail link. The network encompasses 12 suburban lines, 38 regional lines, and 3 lines connecting Lombardy to Switzerland; 77% of municipalities and 92% of Lombardy's citizens have a railway station within a 5-km radius. Trenord caters to a substantial daily ridership of more than 550,000 passengers, facilitated by 2200 train rides (Trenord, 2023).

Our study's scope focuses on six regional lines traversing the provinces of Milan, Brescia, and Bergamo, along with some stations in Lecco and Monza Brianza. Collectively, these lines comprise 46 stations, indicated by the set S. Fig. 1 visually represents the Trenord network, emphasising the lines examined in our study. Furthermore, Table 1 offers an overview of the primary stations of the six train lines under consideration.

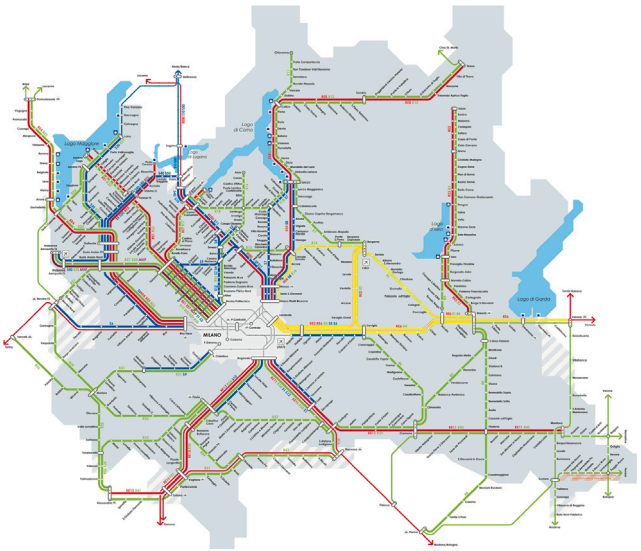

Fig. 1. Trenord map of services (Trenord, 2022d). Train lines considered in this study are coloured in yellow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Main stations of the Trenord lines involved in the study (Trenord, 2022c).

Line	Main stations
R1	Bergamo-Brescia
R2	Bergamo-Treviglio
R4	Brescia-Treviglio-Milano
R14	Bergamo-Carnate-Milano
RE2	Bergamo-Pioltello-Milano
RE6	Verona-Brescia-Milano

3.2. D^3: data sources, data fusion and data engineering

In addressing the dynamic OD matrices estimation problem, we have at our disposal a range of data provided by Trenord:

Ticket data. This dataset covers ticket and subscription sales spanning from May 1, 2022, to December 31, 2022, detailing a total of 2,676,629 transactions. The dataset encompasses journeys to and from the 46 stations within our study scope. We will refer to this dataset as the *ticket data*, including regular tickets, carnets, and weekly, monthly, and yearly subscriptions. Each record in the dataset specifies the two stations covered by the ticket, the ticket type, and the date of purchase. Notice that this dataset covers the month of May 2022, which is outside of the estimation period in our study (June–December 2022). The reason is that tickets for trips happening in June can be bought during May. It is essential to acknowledge certain intricacies within the data that necessitate consideration during the estimation process:

- Although the ticket data indicates the two stations for which the ticket is issued, information regarding the direction of travel is absent.
- The tickets only present the initial and final stations for trips involving transfers, omitting intermediate transfer points.
- Tickets encompassing *Verona Porta Nuova* station utilise an interregional fare not documented in Trenord's data, leading to a lack of information regarding Verona-related trips.
- Tickets for journeys between Milan stations and stations within the Integrated Subscription (IS) area, illustrated in Fig. 2, are not accounted for in the sales data. These trips are subject to a distinct fare structure applicable to trips to and from Milan and its environs.[1]
- Tickets sharing fares with another public transport provider operating within Milan and its surroundings are reported at half their actual quantity in the dataset. Since there is no means of deducing the exact number of these tickets sold, we round up the amounts reported in the ticket data to the smallest integers greater than or equal to the said amount.

[1] Fares regulation and corresponding areas are detailed in Trenord (2024a) and Trenord (2024b)

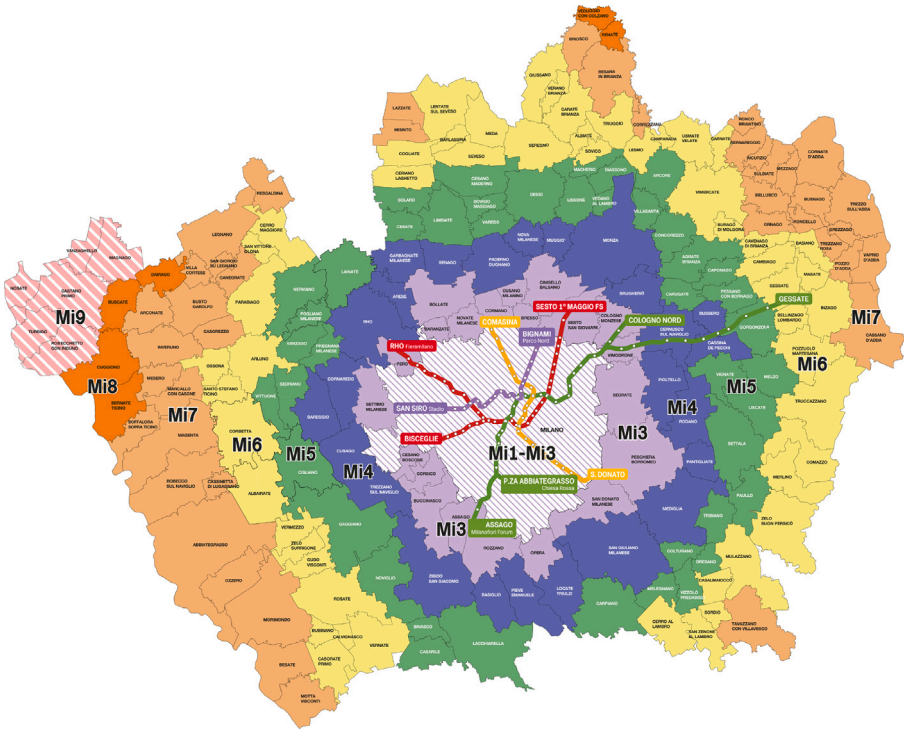

Fig. 2. Map of the municipalities belonging to the Integrated Subscriptions area, covering Milan and Monza provinces. Different colours highlight rings corresponding to different fares, as explained in official documentation and regulations provided in Trenord (2024a) and Trenord (2024b). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Counter data. The counter dataset collects information regarding passenger boarding and alighting at each station for each train ride belonging to the six study lines. The data is captured through the APC system, deployed on approximately 40% of the Trenord fleet in 2021 (Trenord, 2022b). The APC system employs sensors on train doors to count passengers, yielding an accuracy error ranging from 5% to 10%. In cases where only a train segment is equipped with APC, approximations are employed. When estimation fails, interpolations consider the same train on corresponding days of the preceding weeks, working regressively. In scenarios where reliable estimates cannot be derived, the dataset lacks passenger counts for the pertinent train rides. Each entry in the dataset provides a status denoting the outcome of the counting process for the respective train ride. The dataset covers the period between June 1, 2022, and December 31, 2022.

Timetable data. This dataset collects the actual departure and arrival times for each station and train ride pertaining to the six study lines. The data spans train rides between June 1, 2022, and December 31, 2022.

4. Methods

This Section provides an overview of the methods utilised for trip distribution modelling in this study. Two are the main methodologies leveraged in the proposed pipeline: gravity models and the IPF method. Subsequently, we delve into the development of the estimation pipeline designed for this research.

4.1. Gravity model for IPF seeds initialisation

Gravity models have gained widespread use in explaining population movements, commercial trade, and communication patterns. Inspired by Newton's law of gravity, the gravity model (Simini et al., 2012; Barbosa et al., 2018; Ortúzar and Willumsen, 2011) suggests that the movement x_{ij} of people, goods, or information between two locations, i and j, is influenced by the sizes of these places (M_i and M_j, typically related to population or economic scale) and the distance function $f(d_{ij})$ that quantifies their separation in space, time, or cost. Mathematically, this is expressed as:

$$x_{ij} = \mathcal{K} M_i^{\alpha} M_j^{\beta} / f(d_{ij}) \tag{1}$$

Here, α and β represent adjustable exponents, and the distance function $f(d_{ij})$ is often defined using power-law or exponential forms. While the gravity model has gained prominence for approximating travel flows and traffic demand based on local properties,

it must be acknowledged that it is a simplified representation and may not fully capture empirical observations (Simini et al., 2012). Moreover, this model relies on the estimation of multiple parameters, rendering it sensitive to data fluctuations or incompleteness (Simini et al., 2012).

Our pipeline uses the gravity model to estimate missing data within the ticket-estimated seed OD matrices derived from ticket and subscription sales. This approach enables us to fill missing data with inferred estimates, which are then integrated into the IPF method for merging ticket and counter information.

4.2. Iterative proportional fitting algorithm

Growth-factor models serve as key tools in trip distribution modelling, aiding the adjustment of existing OD matrices using insights into the projected growth of trips originating and terminating within specific zones (Barbosa et al., 2018; Ortúzar and Willumsen, 2011). A particularly prominent approach within this category is the IPF method (Barbosa et al., 2018; Evans, 1970; Ortúzar and Willumsen, 2011). This method functions as a doubly constrained growth-factor model, iteratively refining a seed OD matrix to align its row and column sums with data portraying the number of trips originating and ending in each network zone.

In such models, the seed OD matrix can be derived from sources like surveys, historical OD matrices within the transportation network, or insights gleaned from AFC systems, particularly smart card data. In the context of our study, we develop seed OD matrices from ticket and subscription sales data, implementing a structured procedure to translate each ticket type into one or more trips allocated to specific time frames within the seed OD matrices. Subsequently, we extract data about the volume of passengers boarding and alighting at each station during each time frame of the study period from counter data collected by the APC system.

Suppose the seed OD matrix X^* takes the form:

$$\begin{bmatrix} x^*_{11} & \cdots & x^*_{1J} & q_1 \\ \vdots & & \vdots & \vdots \\ x^*_{I1} & \cdots & x^*_{IJ} & q_I \\ b_1 & \cdots & b_J & u \end{bmatrix}$$

Here, x^*_{ij} denotes the number of trips beginning in zone i and terminating in zone j. I represents the set of zones where trips may begin, while J represents the set of zones where trips may end. Further, $q_i = \sum_{j=1}^{J} x^*_{ij}$ indicates the number of trips originating in zone i, and $b_j = \sum_{i=1}^{I} x^*_{ij}$ represents the number of trips concluding at zone j. The total number of trips is denoted as u. Since this matrix does not represent actual network movements, its row and column totals q_i and b_j do not generally match the estimates of trips starting and ending in each zone. Let $p_1, \ldots, p_I$ stand for the estimated count of actual trips originating in each zone, and let $a_1, \ldots, a_J$ denote the estimates for the number of trips ending in each zone. In our application, these estimates are derived from counter data. For consistency, the total number of trips commencing must equal the total number of trips ending, satisfying:

$$\sum_{i=1}^{I} p_i = \sum_{j=1}^{J} a_j = v \tag{2}$$

The trip distribution problem is to deduce from matrix $X^* = [x^*_{ij}]$ an estimated matrix $X = [x_{ij}]$ whose row and column totals are $p_1, \ldots, p_I$ and $a_1, \ldots, a_J$, respectively:

$$\begin{bmatrix} x_{11} & \cdots & x_{1J} & p_1 \\ \vdots & & \vdots & \vdots \\ x_{I1} & \cdots & x_{IJ} & p_I \\ a_1 & \cdots & a_J & v \end{bmatrix}$$

To derive the matrix $X = [x_{ij}]$, the IPF algorithm iteratively determines constants by which to multiply the elements of the original matrix $X^* = X^{(0)} = [x^*_{ij}]$; in each iteration $n = 1, 2, \ldots$, a matrix $X^{(n)} = [x^{(n)}_{ij}]$ is obtained by multiplying element-wise the previous matrix $X^{(n-1)} = [x^{(n-1)}_{ij}]$ by the elements of an appropriate matrix of constants $Z^{(n)} = [z^{(n)}_{ij}]$. In this formulation, for $i \in I$ and $j \in J$:

$$x^{(1)}_{ij} = z^{(1)}_{ij} x^*_{ij}$$
$$x^{(n)}_{ij} = z^{(n)}_{ij} x^{(n-1)}_{ij}, \quad \text{for } n \geq 2$$

The matrix $X = [x_{ij}]$ represents the limiting matrix whose entries $x_{ij} = \lim_{n\to\infty} x^{(n)}_{ij}$, and the value $\lim_{n\to\infty} \prod_{k=1}^{n} z^{(k)}_{ij}$ serves as the required multiplying factor for the initial value x^*_{ij}. The detailed iterative procedure is outlined in Algorithm 1, with its theoretical foundation in Evans (1970), Macgill (1977).

After computing the matrix X, the problem of evaluating its goodness of fit arises. Since we have no data available describing the ground truth of railway movements, we provide an evaluation through the maximum deviation between the generated and desired

Algorithm 1: Iterative Proportional Fitting (IPF) Algorithm

Input: Origin and destination marginal totals p_i and a_j; Initial trip matrix estimate x^*_{ij}; maximum number of iteration `max_iter`; tolerance `tol`

Output: Estimated trip matrix x_{ij}

Initialise iteration count $k \leftarrow 0$
Initialise convergence flag `converged` $\leftarrow$ `False`
Initialise OD matrix $x^{(0)}_{ij} \leftarrow x^*_{ij}$
while *not converged* **do**
  for $i \leftarrow 1$ **to** I **do**
    for $j \leftarrow 1$ **to** J **do**
      Compute $x^{(k+1)}_{ij} = x^{(k)}_{ij} \cdot (p_i / \sum_j x^{(k)}_{ij}) \cdot (a_j / \sum_i x^{(k)}_{ij})$
  Check convergence: compute $\epsilon = \max_{i,j} |x^{(k)}_{ij} - x^{(k-1)}_{ij}|$
  if $\epsilon <$ *tol* $||$ $k \geq$ *max_iter* **then**
    `converged` $\leftarrow$ True
  Update iteration count: $k \leftarrow k + 1$
Return the estimated trip matrix $x_{ij} = x^{(k)}_{ij}$

margin, expressed as:

$$\epsilon_{row} = \max_i \left| p_i - \sum_{j=1}^{J} x_{ij} \right|$$
$$\epsilon_{col} = \max_j \left| a_j - \sum_{i=1}^{I} x_{ij} \right| \tag{3}$$

During our application, we encountered specific challenges inherent to the IPF method. These challenges and their potential solutions are thoroughly discussed in Choupani and Mamdoohi (2016). We discuss the solutions we adopted to three of these problems:

- **Zero cell problem:** The IPF method cannot correct zero cells directly. Hence, it is necessary to identify and address cells that require modification by the algorithm. Although zero cells often signify the impossibility of travel between two zones, they can also stem from erroneous estimates.
  For convenience, let us indicate with S the complete set of stations belonging to the train network of the six lines under scrutiny, while $D \subset S \times S$ collects the set of direct paths, i.e. those origin–destination couples (i, j) directly connected because they both belong to the same train line.
  In our pipeline, we tackle the zero cell problem by substituting zero cells in the seed matrices corresponding to direct paths $(i, j) \in D$ with a value $x^*_{ij} = 0.1$. Cells corresponding to indirect paths are set to zero. This approach aligns with our core objective, which is to estimate the number of train trips between each pair of stations, thus differentiating between trips requiring transfers. Hence, indirect paths (i.e. paths requiring at least one transfer) should correspond to zero cells in the resulting OD matrices: $x_{ij} = 0$ for all $(i, j) \notin D$.
- **Marginals consistency:** While Eq. (2) stipulates that the total number of trips starting should equal the total number of trips ending, this assumption is disrupted in our context due to estimation errors in marginals $p_1, \ldots, p_I$ and $a_1, \ldots, a_J$. To overcome this inconsistency, (Barthélemy and Suesse, 2018) proposes a solution involving a shift towards probabilities. This entails defining, for each origin–destination couple (i, j):

$$\pi^*_{ij} = x^*_{ij} / \sum_{ij} x^*_{ij}$$
$$\rho_i = p_i / \sum_i p_i \tag{4}$$
$$\alpha_j = a_j / \sum_j a_j$$

  By performing this operation, consistency is restored, as $\sum_{i=1}^{I} \rho_i = \sum_{j=1}^{J} \alpha_j = 1$. Subsequently, the IPF method is employed using $\Pi^* = [\pi^*_{ij}]$ as the seed matrix and $\rho_1, \ldots, \rho_I$ and $\alpha_1, \ldots, \alpha_J$ as marginals. This results in the matrix Π whose elements are π_{ij}, where $\sum_{ij} \pi_{ij} = 1$. Each cell π_{ij} of Π represents the probability of a single trip occurring from zone i to zone j. To deduce the OD matrix X estimating the actual number of trips between each zone pair, we must multiply the matrix's elements π_{ij} by the total number of trips, v. The number v can be selected to be either the total number of boarded passengers ($\sum_{i=1}^{I} p_i$) or the total number of alighted ones ($\sum_{j=1}^{J} a_j$).

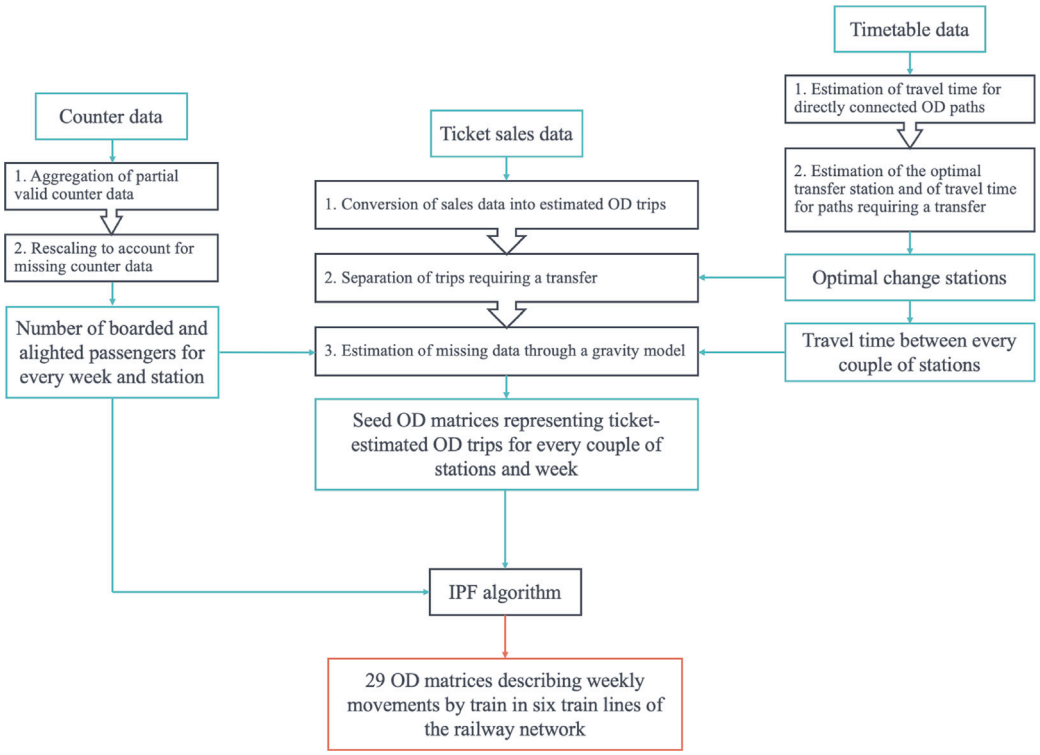

Fig. 3. Estimation pipeline developed to obtain dynamic OD matrices describing movements in the Trenord network for a given time period divided into time frames, using ticket, timetable, and counter data.

- **Integer conversion:** The matrix entries x_{ij} denote the number of trips originating from zone i and terminating in zone j. Therefore, these entries should naturally be integer values. However, the IPF method generates a matrix with non-integer values. To address this, we apply integer conversion to the output matrix. This approach has drawbacks, particularly concerning information discrimination, implying that cells containing values such as 0.501 and 0.999 are treated equally after conversion (Choupani and Mamdoohi, 2016). Moreover, the question of whether the seed x^*_{ij} and marginals p_i and a_j should be integers arises. These values hold natural interpretations in terms of estimated trips (x^*_{ij}), the number of trips originating in zone i (p_i), and the number of trips ending in zone j (a_j). We opt to round the marginal vectors p_i and a_j, while rounding is not applied to the seed OD matrices cells x^*_{ij}, as fractional values could result from the procedure used to generate these seeds, as described in Section 4.3.3.

4.3. Estimation pipeline

In this Section, we outline the pipeline used to estimate weekly OD matrices describing train movements across stations $i \in S$ of the Trenord network for a set $w \in W$ of time frames individuated in a time period. The input data are the ticket, counter, and timetable datasets described in Section 3.2. The pipeline progresses through the series of steps shown in Fig. 3.

4.3.1. Estimation of mean travel time and optimal change stations

Starting with the timetable data, which describes the actual departure and arrival times of each train ride in the considered portion of the Trenord network, we compute the mean travel time for each direct path. To ensure data quality, we eliminate values outside the [0.05, 0.95] quantile range that could stem from data entry errors or extreme delays. The computed matrix $T = [t_{ij}]$ stores mean travel times for each $(i, j) \in D$, with the fastest travel time chosen for stations connected by multiple lines. Notice that while we estimate travel times from actual timetable data, these times could alternatively be collected from the service's schedule.

This data is then used to estimate the optimal change station for paths not directly connected. For each path $(i, j) \notin D$, we identify the optimal transfer station k^* as follows:

$$k^* = k^*(i,j) = \underset{\{k:(i,k),(k,j)\in D\}}{\arg\min} (t_{ik} + t_{kj})$$

The travel time between non-directly connected stations is then estimated as follows:

$$t_{ij} = t_{ik^*} + t_{k^*j}$$

Finally, the matrix $T = [t_{ij}]$ is enriched to include travel times for directly connected stations and those that can be connected with at most one transfer. Notably, travel times involving multiple transfers are omitted, since trips requiring more than one transfer are unlikely to be observed on the Trenord network. Moreover, considering multiple transfers would introduce unnecessary uncertainty in the OD estimation pipeline. Indeed, the choice of transfer station also depends on the transfer waiting time, which we did not account for in the computation of t_{ij} since waiting times depend on the specific choice of the train rides and the time of the day when the trip happens.

4.3.2. Estimation of boarded and alighted passengers from counter data

Accurate estimates of boarded and alighted passengers for each station $i \in S$ during each time frame $w \in W$ are obtained using the counter data collected by the APC system. For each station-time frame combination (i, w), we compute partial estimates of boarded and alighted passengers ($partial\_boarded_i^{[w]}$ and $partial\_alighted_i^{[w]}$) based on valid counter data, aggregating data collected by the APC system. The coverage of each station-time frame pair, $coverage_i^{[w]}$, is calculated as the ratio of train rides with valid counter data to the total train rides in time frame w stopping at station i:

$$coverage_i^{[w]} = \#\{train\ rides\ having\ valid\ counter\ data\}_i^{[w]} / \#\{total\ train\ rides\}_i^{[w]}$$

We can estimate the total boarded and alighted passengers by computing the mean number of boarded and alighted passengers in the trains having valid counter data and then by multiplying this mean for the number of trains stopping in station i during week w.

The described procedure is then equivalent to the following estimators:

$$\begin{aligned} p_i^{[w]} &= (partial\_boarded_i^{[w]} / \#\{train\ rides\ having\ valid\ counter\ data\}_i^{[w]}) * \#\{total\ train\ rides\}_i^{[w]} \\ a_i^{[w]} &= (partial\_alighted_i^{[w]} / \#\{train\ rides\ having\ valid\ counter\ data\}_i^{[w]}) * \#\{total\ train\ rides\}_i^{[w]} \end{aligned} \tag{5}$$

Alternatively, they can be expressed as:

$$\begin{aligned} p_i^{[w]} &= partial\_boarded_i^{[w]} / coverage_i^{[w]} \\ a_i^{[w]} &= partial\_alighted_i^{[w]} / coverage_i^{[w]} \end{aligned}$$

Note that this step assumes adequate coverage of counter data for a significant fraction of train rides in the network. If this assumption is not respected, alternative methods to estimate the passenger counts should be discussed (see Section 6). While not explored in this work, regression models with appropriate design of variance covariance matrix, considering predictors like the number of trains stopping at a station during the week and the number of ticket-estimated incoming and outgoing trips or other relevant factors is one of the main option. Moreover, the rescaling procedure could be stratified to account for weekdays and time slots of train rides, aiming to obtain passenger count estimates closely resembling reality.

4.3.3. Conversion of ticket and subscription sales records into origin–destination seeds

This step generates seed OD matrices from ticket and subscription sales, to be iteratively refined using the IPF algorithm. We make a series of assumptions linking each record in the ticket sales dataset to one or more estimated trips, following the suggestions of the data provider and considering information on ticket validity and usage (Trenord, 2022a). These assumptions are detailed in Appendix A and are the cornerstone for producing OD seeds. They might evolve based on changes in ticket regulations or other factors; future work could involve robust sensitivity analysis, exploring the impact of altering them on the final OD estimates.

With these assumptions, for each week in the considered time period we obtain a seed OD matrix representing estimated trips for all possible OD paths. Notice that paths not directly connected are covered by tickets reporting the initial and final stations which do not form a couple in $\mathcal{D}$. We decouple such paths by leveraging the optimal transfer station obtained in Section 4.3.1. In fact, for each OD path $(i, j) \notin \mathcal{D}$ requiring at most one transfer, we decompose the ticket-estimated OD entries $x_{ij}^{[w]*}$ into $x_{ik^*}^{[w]*}$ and $x_{k^*j}^{[w]*}$, where k^* is the optimal transfer station. Once again, let us point out that this decoupling emphasises the goal of estimating OD matrices for train trips in the Trenord network, treating trips requiring transfers as two distinct journeys; trips needing more than one transfer are excluded, as discussed in Section 4.3.1.

We employ a gravity model to predict missing values relative to paths with missing ticket sales data. For each week w and origin–destination couple (i, j), the log-transformed gravity model of Eq. (1) reads as:

$$log(x_{ij}^{[w]*}) = log(\mathcal{K}) + \alpha\ log(p_i^{[w]}) + \beta\ log(a_j^{[w]}) + \gamma\ log(t_{ij}) + \epsilon_{ij} \tag{6}$$

Notice that, masses are chosen to be the number $p_i^{[w]}$ of passengers boarding at the origin station i, and the number $a_j^{[w]}$ of passengers alighting at the destination j, in the considered time frame $w \in W$, while the distance function is a power law of the travel time estimated in Section 4.3.1. We fit the model by Ordinary Least Squares, excluding stations with $p_i^{[w]} = 0$ or $a_j^{[w]} = 0$, since the natural entry for these cells is $x_{ij}^{[w]*} = 0$. The fitted model is then used to predict the missing values relative to paths with missing ticket sales data.

Table 2
Number of purchases and percentage of total sales for each ticket type during the seven months of 2022 considered in the study.

Ticket type	Purchases	% of total ticket sales data
Ordinary tickets	2,559,799	95.6%
Monthly subscriptions	51,076	1.91%
Weekly subscriptions	21,106	0.79%
Carnets	2,476	0.09%
Annual subscriptions	194	0.01%
Other operator	1,172	0.04%
Sanctions	1,012	0.04%
Special rates initiatives	6	<0.01%
Supplementary corrections	37,029	1.38%
Supplements	2,625	0.1%
Unpaid remains	134	0.01%

4.3.4. Summary of the iterative proportional fitting method

To obtain the dynamic OD matrices describing trip counts $X^{[w]} = [x_{ij}^{[w]}]$ for every week $w \in W$, we apply the IPF algorithm described in Section 4.2. The ticket-estimated OD matrices $X^{*[w]}$ serve as seed matrices, with boarded passengers $p_i^{[w]}$ and alighted passengers $a_j^{[w]}$ as margins.

Furthermore, to tackle the zero correction problem illustrated in Section 4.2, we replace such cells corresponding to direct paths $(i, j) \in \mathcal{D}$ with the artificial value 0.1, while we leave cells corresponding to indirect routes to 0.

Due to estimation errors and APC system issues, it might happen that $\sum_{i=1}^{I} p_i^{[w]} \neq \sum_{j=1}^{J} a_j^{[w]}$ for a given time frame $w \in W$, contradicting the consistency assumption in Eq. (2). To address this problem, we scale tickets, boarded, and alighted passengers using the probability interpretation defined by Eq. (4), yielding:

$$\pi_{ij}^{*[w]} = x_{ij}^{*[w]} \Big/ \sum_{(i,j)\in S} x_{ij}^{*[w]}$$
$$\rho_i^{[w]} = p_i^{[w]} \Big/ \sum_{k\in S} p_k^{[w]}$$
$$\alpha_j^{[w]} = a_j^{[w]} \Big/ \sum_{k\in S} a_k^{[w]}$$

This adjustment ensures $\sum_{i\in S} \rho_i^{[w]} = \sum_{j\in S} \alpha_j^{[w]} = 1$ and $\sum_{(i,j)\in S} \pi_{ij}^{*[w]} = 1$ for all $w \in W$.

Applying the IPF method results in the matrix $\pi_{ij}^{[w]}$, which is then corrected by multiplying each cell by the total number of boarded passengers as:

$$x_{ij}^{[w]} = \pi_{ij}^{[w]} \sum_{k\in S} p_k^{[w]}$$

This adjustment ensures that the final matrix $X^{[w]}$ accurately represents the trips originating in station i and ending in station j during the week w. Alternatively, the total number of alighted passengers in week w, $\sum_{k\in S} a_k^{[w]}$, could be used for scaling.

After this procedure, the final matrix $X^{[w]}$ is rounded, as its entries $x_{ij}^{[w]}$ represent the number of trips originating in station i and ending in station j during the week w.

5. Results

In this Section, we present the results obtained by applying the proposed pipeline to estimate weekly OD matrices that describe train movements across six train lines within the Trenord network for each week in the seven-month period of June-December 2022. Thus, the set of stations S covered by the estimation pipeline presented in Section 4.3 comprises 46 stations of the Trenord network and the set of timeframes W consists of the 29 weeks from June 6, 2022, to December 25, 2022. We conducted statistical analyses using the R software environment (R. Core Team, 2022) and employed the `mipfp` package to execute the IPF algorithm (Barthélemy and Suesse, 2018).

5.1. Exploratory analysis

We first present an exploratory analysis conducted on the datasets provided by Trenord, described in Section 3.2, to highlight some critical aspects of the data. Fig. 4 displays the number of tickets purchased during the study period, distinguishing ticket types. Table 2 shows the sales for each ticket type throughout the seven months of 2022 considered in the study. We can notice that ordinary tickets are by far the most purchased kind of ticket. Moreover, we observe a reduction in the purchases of all types of tickets in the summer period (July and August), followed by an increase in September after the end of the summer holidays.

Turning to the data collected by train counters, Fig. 5 shows the number of trains analysed in our study. Once again, we note a slight reduction in train rides' number during August.

Finally, Fig. 6 illustrates the distribution of the states of the APC system within train rides. The states include valid data, missing data, and cancelled train rides (where no passengers could board or alight the train). Missing data accounts for only 0.5% of train rides, excluding the cancelled ones.

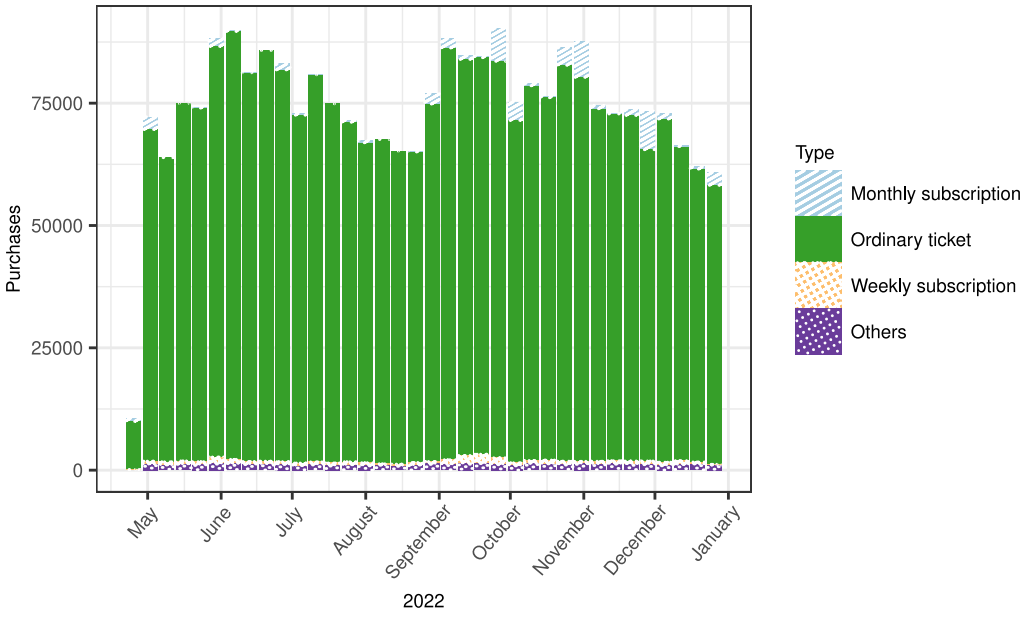

Fig. 4. Number of purchased tickets and subscriptions across seven months of 2022, stratified by type.

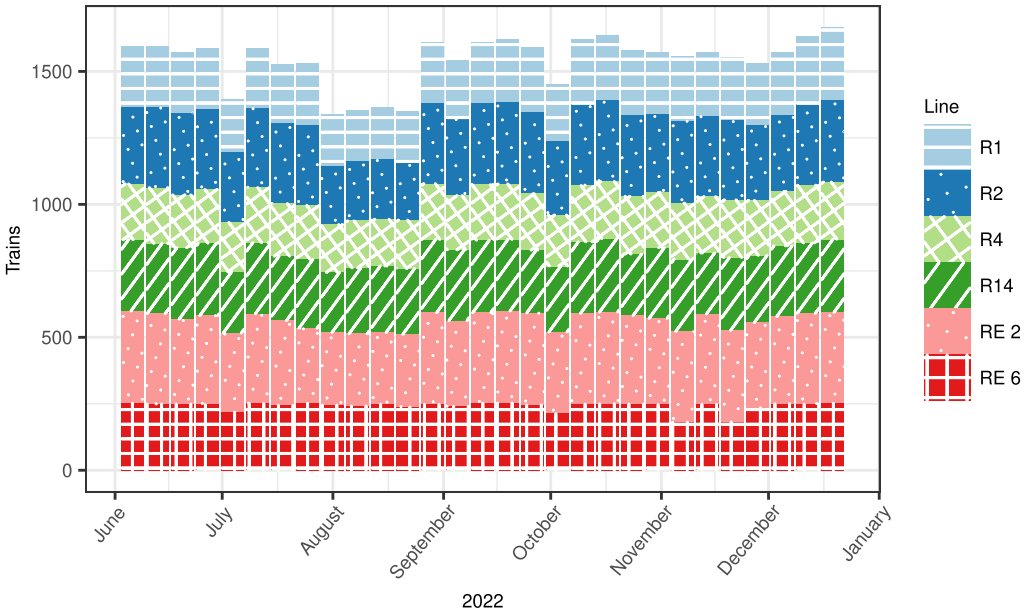

Fig. 5. Number of trains per week during seven months of 2022, stratified by the six train lines.

5.2. Counter data aggregation and estimation of missing data

Fig. 7 presents the distribution of $coverage_i^{[w]}$ across station-week pairs, showcasing that most station-week pairs have coverage exceeding 90%. Notably, even the lowest value observed in the dataset is 67%, indicating substantial coverage of train rides during the study period. This motivates us in applying Eq. (5) to estimate the number of boarded and alighted passengers during each week, as missing data are a minority in the counter dataset.

Fig. 8 shows the estimated number of passengers who boarded and alighted a train during the study period, based on partial valid counter data and following the rescaling procedure detailed in Section 4.3.2 using Eq. (5). It is evident that the rescaling has not altered the passenger trends for boarding and alighting. Passenger counts vary throughout the study period, with lower values during the summer and subsequent increases in September, followed by a decrease during the Christmas holidays.

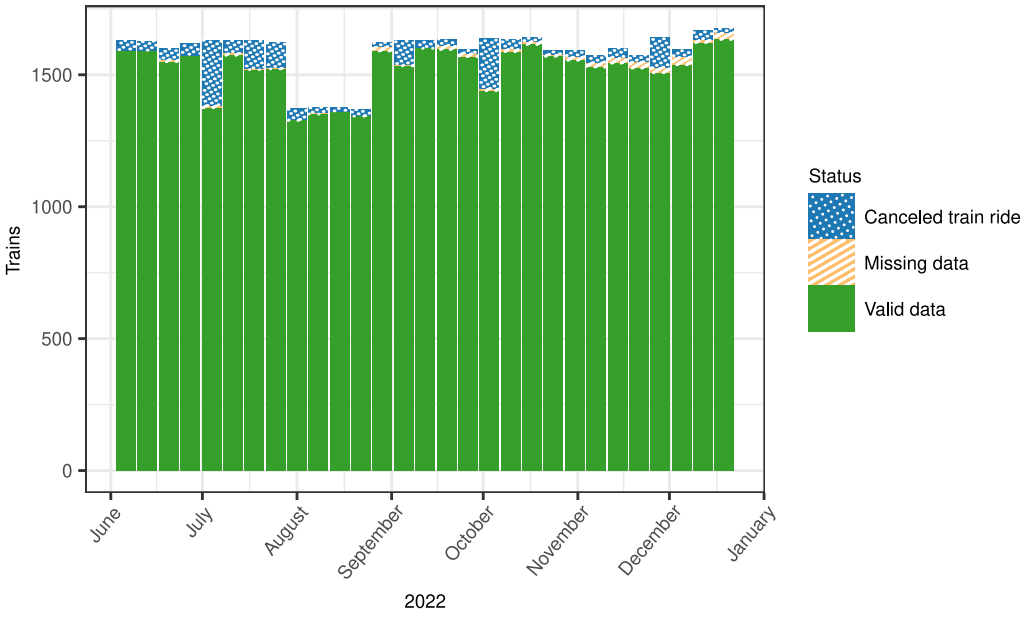

Fig. 6. Distribution of the APC system's states of train rides during seven months of 2022.

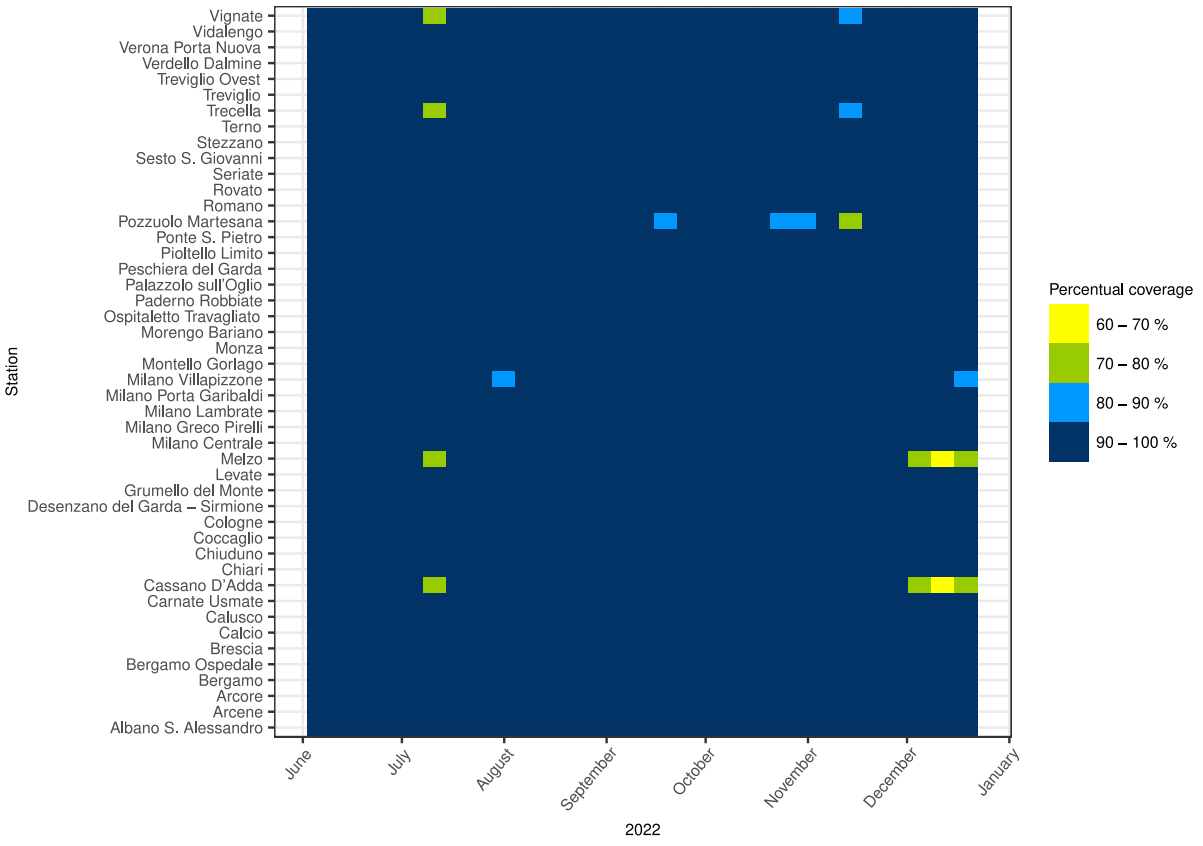

Fig. 7. Coverage per week $w \in W$ and station $i \in S$ during seven months of 2022.

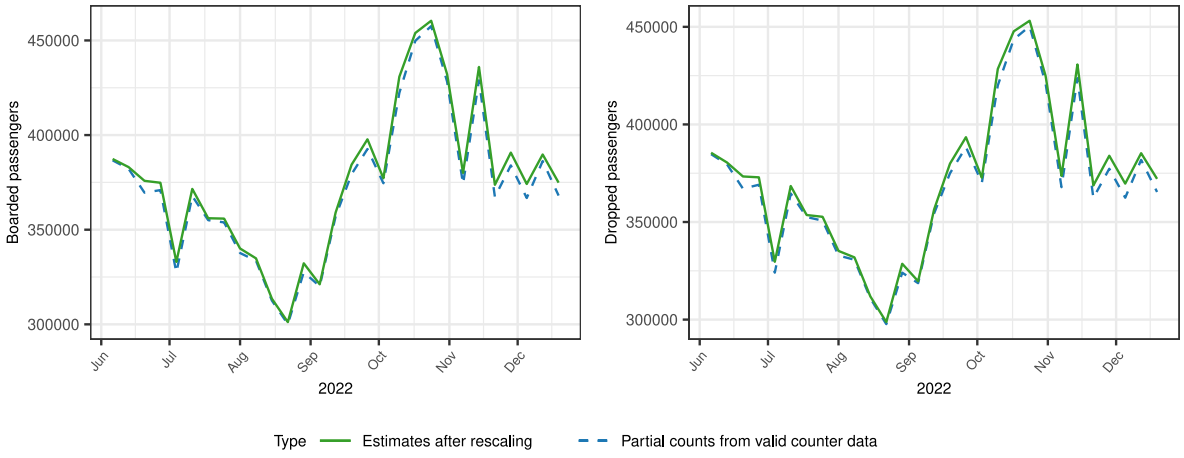

Fig. 8. Weekly counts of boarded passengers (left) and alighted passengers (right) from June to December 2022. The dashed line represents partial passenger counts considering train rides with valid estimates from the APC system, while the solid line depicts estimates for passenger counts applying the rescaling in Eq. (5).

Table 3
Number of estimated trips and percentual of total estimated trips across seven months of 2022, divided into direct paths, one transfer and more than one transfer.

Transfers' number	Estimated trips	% of total estimated trips
Direct paths	4,770,557	95.2%
One transfer	235,316	4.70%
More than one transfer	3,822	0.76%

5.3. Conversion of ticket data into seed origin–destination matrices

Applying the pipeline outlined in Section 4.3.3, we generate 29 seed OD matrices $X^{*[w]}$ for every $w \in W$ based on ticket data, portraying movements inferred from tickets and subscriptions purchased within the seven months of 2022 considered in our study and the month of May. This process is achieved through three steps, as follows:

1. Converting each record within the ticket dataset into one or multiple trips characterised by an origin $i \in S$, destination $j \in S$, and week $w \in W$.
2. Separating trips necessitating transfers according to the optimal interchange station derived from timetable data.
3. Estimating missing ticket data using a gravity model.

We will now illustrate this procedure using as example the 38th week of the year (September 19–25, 2022), showcasing the evolving ticket-estimated OD matrix ($X^{[38]*}$) after each step. Fig. 9 displays the ticket-estimated OD matrix achieved by applying the ticket-to-trip conversion assumptions. It is noteworthy the absence of ticket data to or from *Verona Porta Nuova* station and between internal Milan stations (*Milano Lambrate, Milano Centrale, Milano Porta Garibaldi, Milano Villapizzone, Milano Greco Pirelli*) and stations within the IS area (*Monza, Sesto S. Giovanni, Arcore, Carnate Usmate, Pioltello Limito, Vignate, Melzo, Trecella, Cassano d'Adda*, and *Pozzuolo Martesana*).

After constructing the ticket-estimated OD matrices, the subsequent task is to separate trips requiring a transfer based on the optimal interchange station determined in Section 4.3.1. This process omits paths necessitating more than one station change, constituting a mere 0.7% of the total estimated trips from ticket data, as shown in Table 3. Consequently, only a negligible fraction of data is lost by excluding trips with multiple transfers.

Fig. 10 exhibits the ticket-estimated OD matrix for week 38 after separating trips requiring transfers. Notably, only OD paths directly connected in D exhibit non-zero cells in $X^{[w]*}$.

The final step in ticket data processing is estimating missing OD cells, encompassing paths to and from *Verona Porta Nuova* station and routes linking Milan stations with other IS area stations. This estimation employs the gravity model described in Eq. (6). The estimates of coefficients, the corresponding p-values, and the R^2 coefficient for the gravity model are provided in Table 4. As it is natural to expect in the formulation of the gravity model, the coefficients α and β which express the influence of the number of passengers boarding and alighting at the origin and destination station on the OD flow are positive while γ, representing the effect of the separation in time between the two stations, is negative.

This model is employed to predict the 70 cells related to paths with missing ticket data for every week $w \in W$. Fig. 11 displays matrix $X^{[38]*}$ again, following the prediction of missing ticket data. Notably, cells linked to *Verona Porta Nuova station* and paths between Milan and IS area stations now contain non-zero values.

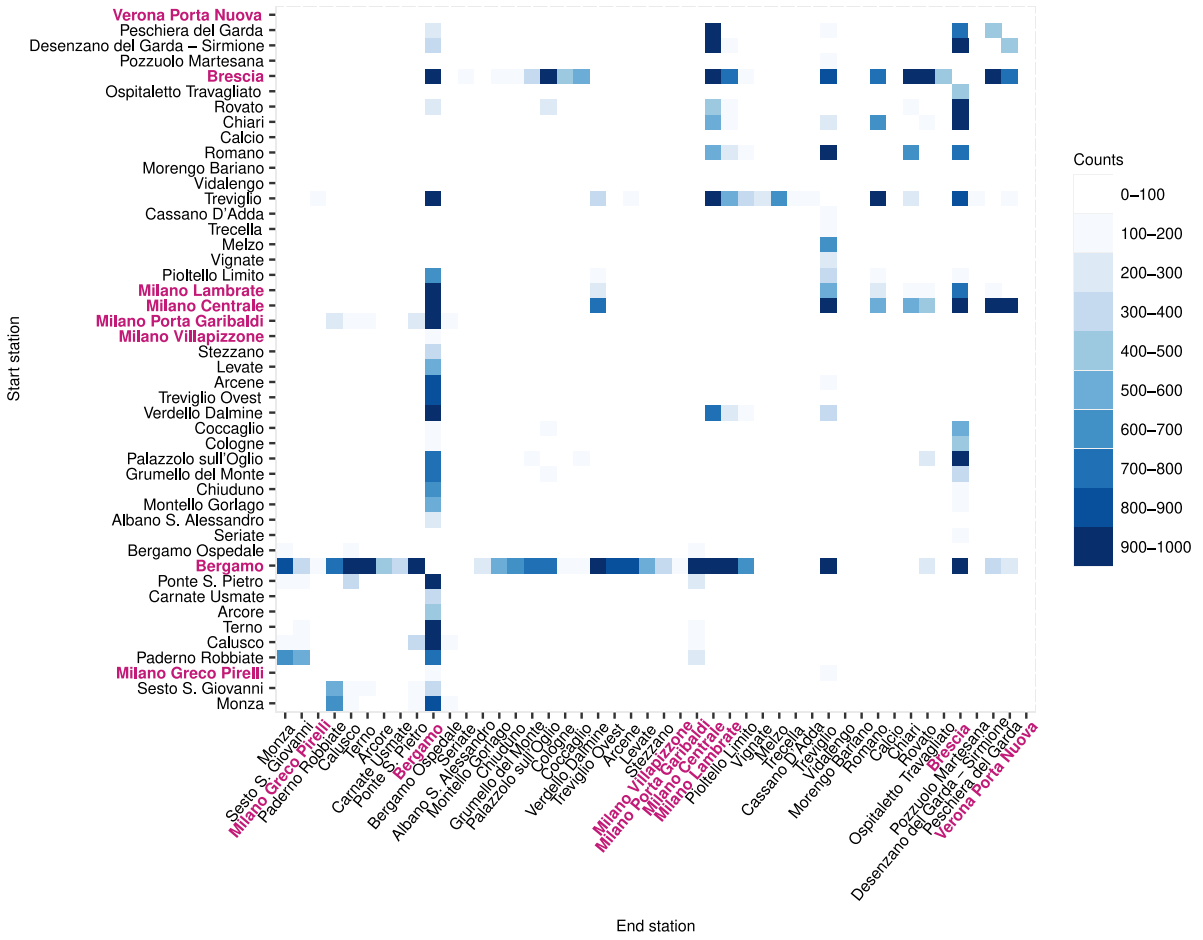

Fig. 9. Week 38: ticket-estimated OD matrix after ticket-to-trip conversion. The main stations of the network are highlighted.

Table 4
Coefficient estimates for the gravity model presented in Eq. (6), along with p-values denoting parameter significance.

Variable	Estimate	p-value
(Intercept)	−8.67	< 0.0001
α	0.68	< 0.0001
β	0.92	< 0.0001
γ	−0.35	< 0.0001
R^2	0.56	

5.4. Application of the iterative proportional fitting algorithm

Following the conversion of tickets into seed OD matrices ($X^{*[w]}$) and the computation of the margin vectors ($p_i^{[w]}$ and $a_i^{[w]}$) representing the total number of passengers boarded and alighted at each station $i \in S$ during week $w \in W$, we can now proceed to generate the estimated OD matrices ($X^{[w]}$) describing train movements from station i to station j in week w by applying the IPF algorithm as specified in Section 4.3.4.

An example of one of the finalised OD matrices is presented in Fig. 12. This matrix corresponds to week 38, already used in Section 5.3 as illustrative example.

The final OD matrices exhibit adherence to reality in the following aspects:

1. Noticeable movement centres around major hubs such as Bergamo, Brescia, Verona and Milan stations.
2. Stations linked by two or three lines display greater movement than stations with only a single connection.
3. Mobility experiences a decline during the summer, notably in August.
4. Mobility sees a reprise around the start of September.

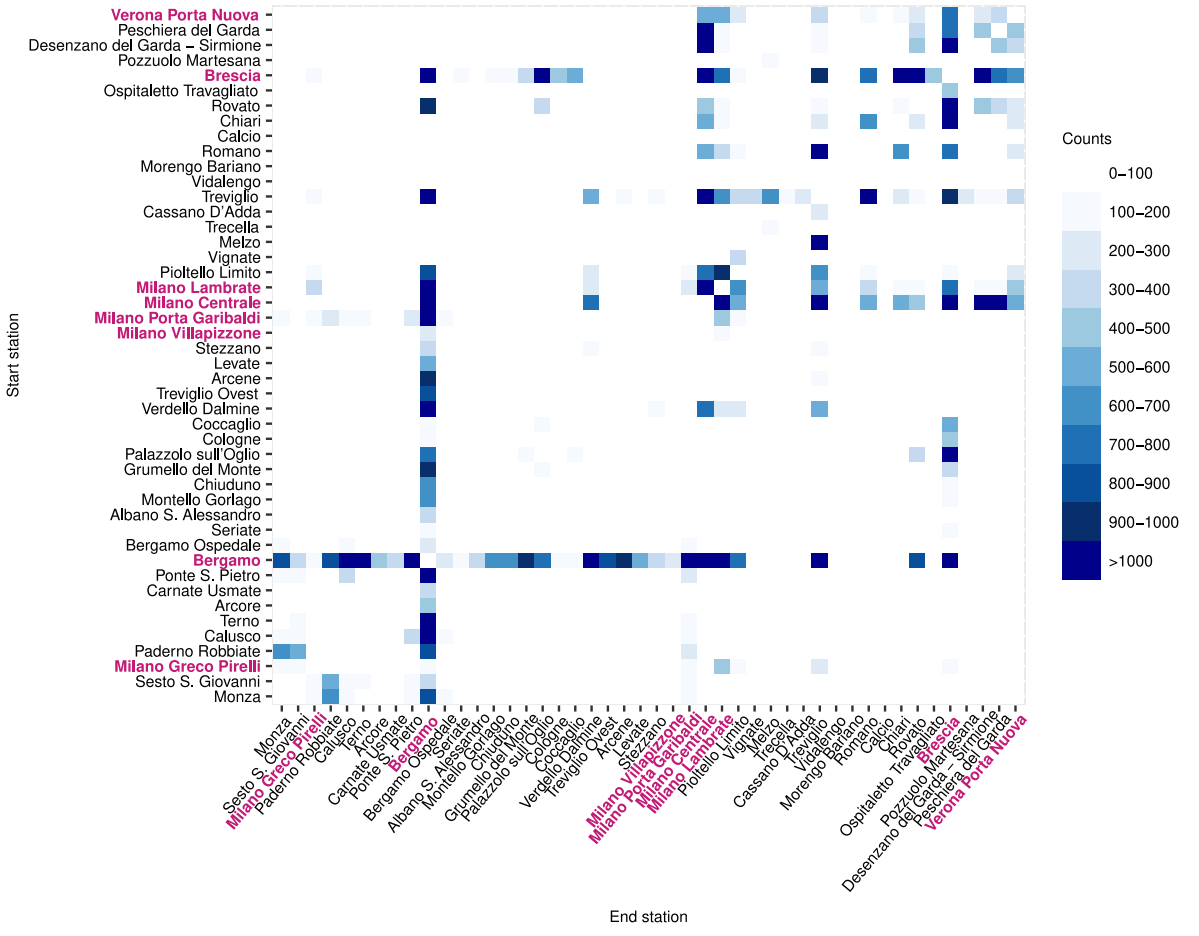

Fig. 10. Week 38: ticket-estimated OD matrix after separating trips requiring transfers. The main stations of the network are highlighted.

5. Mobility declines once more during the Christmas season.

To assess the accuracy of the proposed procedure, we quantified the alignment between the final matrix $x_{ij}^{[w]}$ and the margins $p_i^{[w]}$ and $a_j^{[w]}$, assessing margins errors ϵ_{row} and ϵ_{col} as the maximal deviation between each calculated and desired margin, as detailed in Eq. (3). It is important to note that since the final matrix results from the product of the probability matrix $\pi_{ij}^{[w]}$ and the total boarded passengers, as in Eq. (4), we should consider row errors $\epsilon_{row}^{[w]}$ to judge the fitting quality of the procedure. The row margin errors $\epsilon_{row}^{[w]}$ are consistently low, as shown in Fig. 13. This indicates that the margins of the estimated cells $x_{ij}^{[w]}$ align well with the actual margins $p_i^{[w]}$, as expected.

To further confirm the validity of our results, the assessment deriving from comparison with a ground truth would be relevant. Since none is available to this purpose, we exploited a static estimate of the OD matrix provided by the regional government of Regione Lombardia (RL) (Lombardia, 2019), as explained in Galliani et al. (2023). This has a number of drawbacks with respect to our purposes, but is the only proxy available of a ground truth. In fact, it is static, whereas our procedure produces a time-series of OD matrices. Therefore, making any assessment implies an unfair comparison between objects that evolves over time and a static one. Moreover, as the static OD matrix from Regione Lombardia was released in 2019, it did not account for the exceptional disruption to usual mobility caused by the COVID-19 pandemic. Thus, these data may not be a reliable description of mobility happening in the year 2022. Finally, the RL OD matrix refers to the spatial granularity of municipalities, while the spatial granularity of Trenord data relates to the station, implying the adoption of spatial matching techniques. Acknowledged all these issues, we found anyway a notable correlation between the two datasets, validating the pipeline's capacity to yield dynamic OD matrices approximating actual mobility.

5.5. Exploiting origin–destination matrices

Dynamic OD matrices offer several potential usages in sustainable planning as well as in complex systems maintenance, for example, anomaly detection within the Trenord network. Dynamic OD matrices may be employed in global and local identification

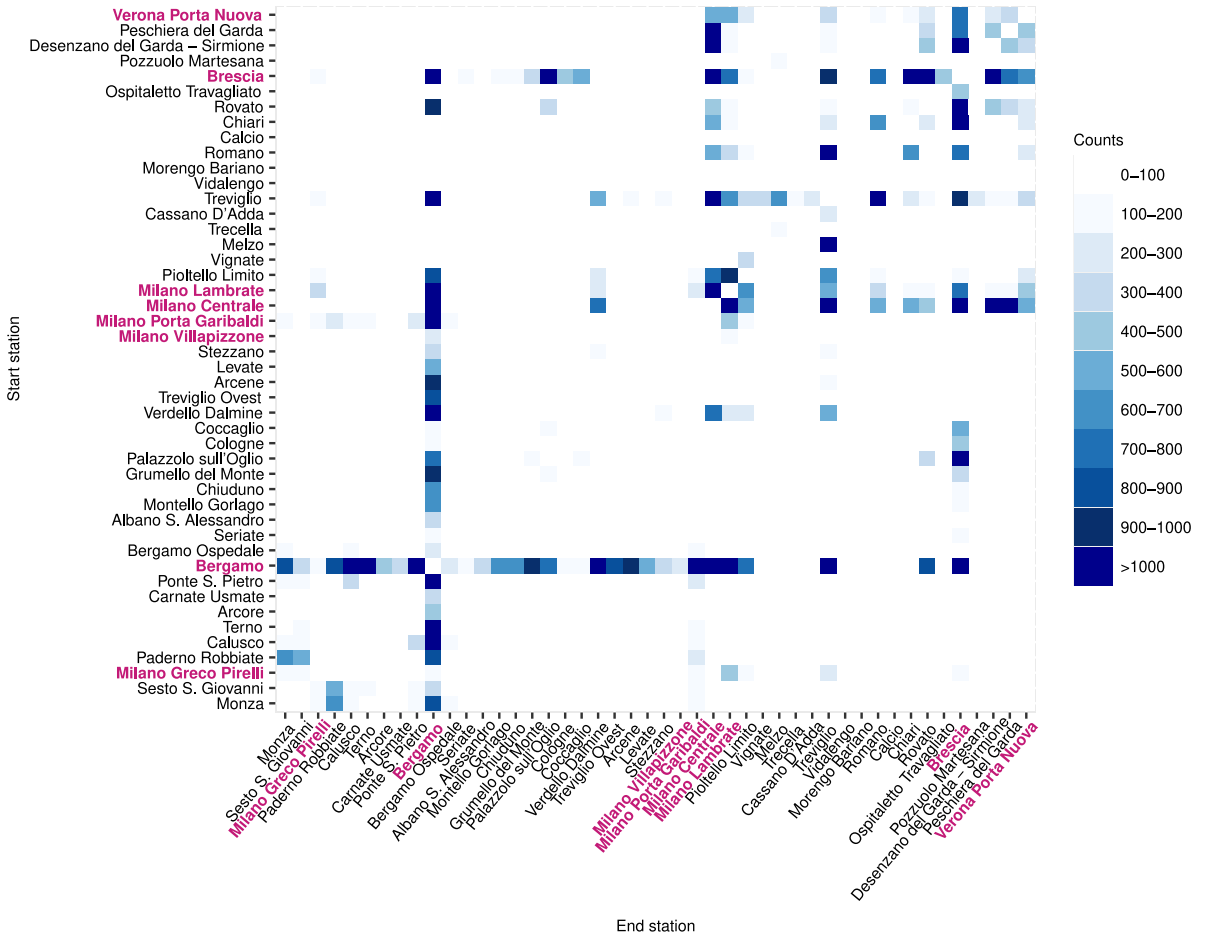

Fig. 11. Week 38: ticket-estimated OD matrix $X^{[38]*}$ after estimating missing ticket data. The main stations of the network are highlighted.

of abrupt changes in mobility expressed by the network revealing insights into the network's behaviour. To achieve this, we leverage network analysis techniques (Douglas, 2015), treating dynamic OD matrices as weighted directed dynamic networks. This, coupled with tools deriving from functional data analysis (FDA) (Ramsay and Silverman, 2006), allows for the maintenance of a time series of meaningful indicators of mobility patterns at both global and local levels.

5.5.1. Global indicators

Global indicators analyse the network characteristics and showcase the differences between consecutive weeks. We defined two global indicators and interpreted them jointly, together with some events that may have influenced the network's dynamics. These events are divided into strikes, holidays, and infrastructural construction work. Appendix B reports a description of all the events identified in the period from June to December 2022.

The global indicators we defined are:

Root mean squared difference (RMSD). This indicator, defined as

$$RMSD^{[w]} = \frac{1}{S}\sqrt{\sum_{i=1}^{S}\sum_{j=1}^{S}(x_{ij}^{[w]} - x_{ij}^{[w-1]})^2}$$

measures differences between subsequent weeks in the network. High $RMSD^{[w]}$ values mean significant global fluctuations in movements between the week w and the previous one $w-1$, while low values indicate stability between subsequent weeks. Events such as strikes, holidays, and infrastructural interventions, reported in Appendix B, were noted as potential influencers. For instance, the two major network interventions resulted in notable disruptions, leading to high values in RMSD values. These observations are depicted in Fig. 14.

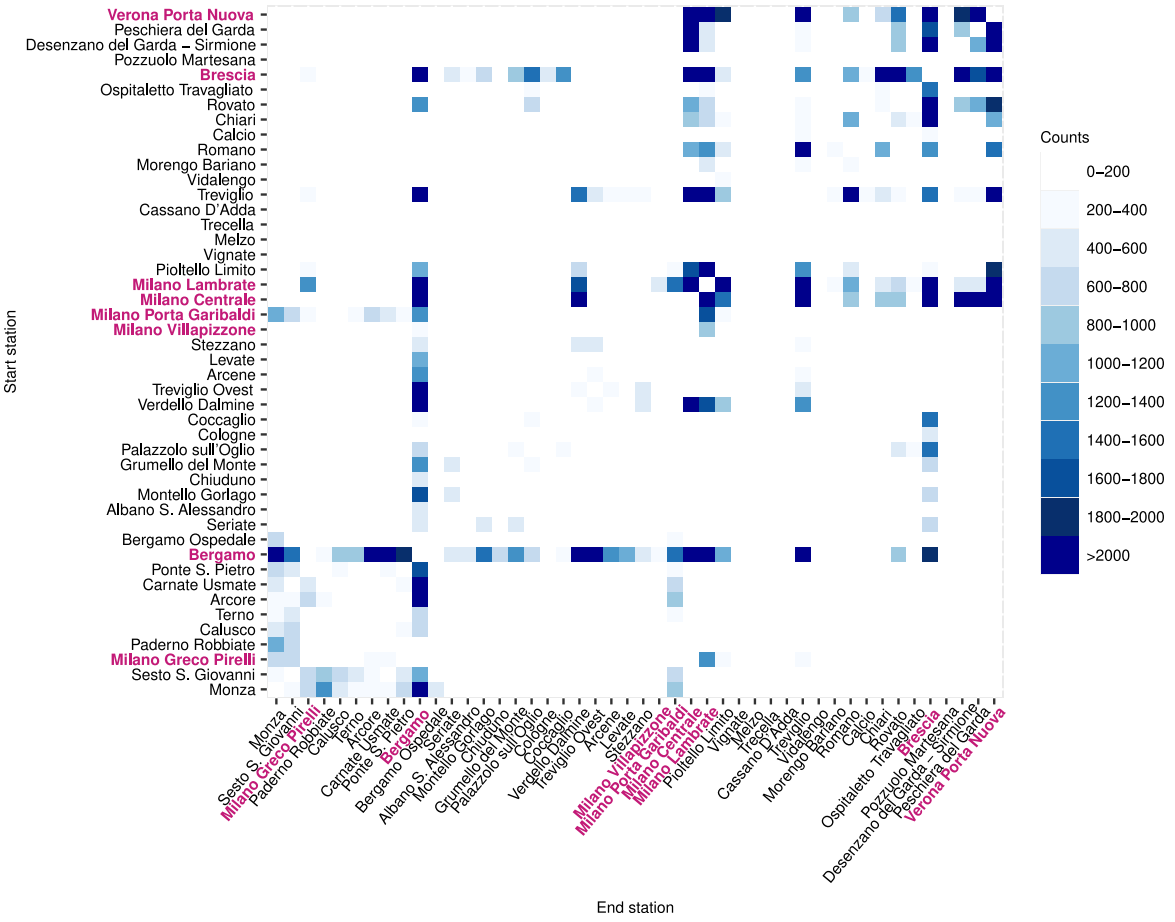

Fig. 12. OD matrix $X^{[38]}$ estimated using the IPF method, illustrating train movements in the railway network for the week starting on September 19, 2022, and ending on September 25, 2022. The main stations of the network are highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Mean strength. As a second global indicator, we compute the mean strength as the average number of passengers passing through the network's stations as:

$$\sigma_i^{[w]} = \sum_{j=1}^{S} x_{ij}^{[w]} + \sum_{j=1}^{S} x_{ji}^{[w]} \tag{7}$$

$$\bar{\sigma}^{[w]} = \frac{1}{S} \sum_{i=1}^{S} \sigma_i^{[w]} \tag{8}$$

Notice that the strength $\sigma_i^{[w]}$ could be equivalently computed directly from the number of boarded and alighted passengers as:

$$\sigma_i^{[w]} = p_i^{[w]} + a_i^{[w]}$$

Despite in the latter indicator the estimation of the entire matrix is not required, its usage has to be intended as coupled with the RMSD, where all the matrix entries are used. In fact, a small value in RMSD might indicate similar behaviour in two subsequent weeks having very high or very low strength. Only the association with values reported in Fig. 15 allows to properly interpret the meaning of Fig. 14, and vice-versa.

Also the strength indicator is influenced by events and holidays, leading to fluctuations in mobility trends. Notably, a dip in mean strength is observed in August when many Italian companies typically close for summer holidays. Fig. 15 illustrates these fluctuations and their correlation with events.

Thus, the two global indicators defined, $RMSD^{[w]}$ expressing variations in mobility in subsequent weeks and $\bar{\sigma}^{[w]}$ depicting the network's global mobility for each week, are able to identify weeks where the network's disruptions can be traced back to some events affecting railway mobility.

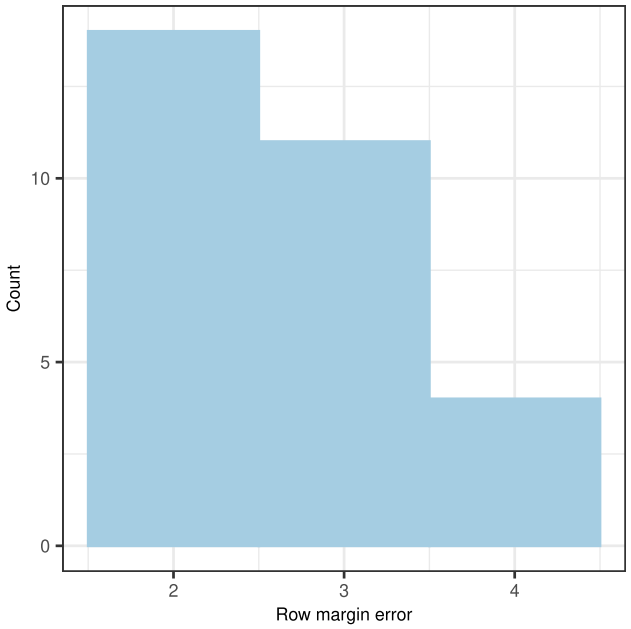

Fig. 13. Histogram of row margin errors $\epsilon_{row}^{[w]}$.

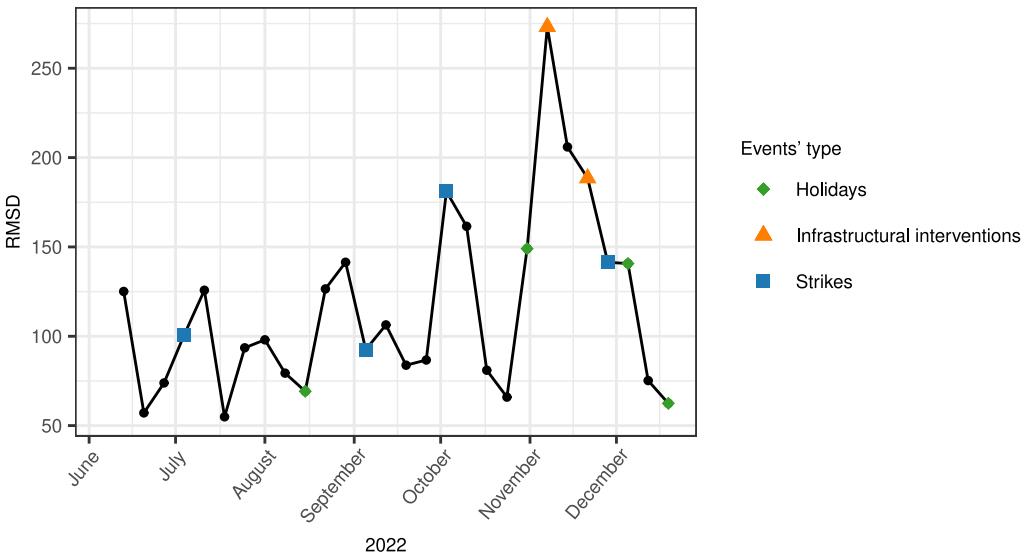

Fig. 14. Values of $RMSD^{[w]}$ over the 29 weeks of the study. Coloured points indicate weeks when one of the events reported in Appendix B has happened.

5.5.2. Local indicators

Given the ability of spotting anomalies in the mobility pattern expressed by the network, we turn to the local station level to study the evolution of network indicators and assess how each node reacts to the network's disruptions. We consider the weekly strength $\sigma_i^{[w]}$ for each station $i \in S$, as defined in Eq. (7). To remove the effect of the volume of passengers passing through each node and analyse only mobility trends, we consider the normalised version of the strength by dividing each strength value by the sum of the station's strengths through the 29 weeks of the study as:

$$\widetilde{\sigma}_i^{[w]} = \sigma_i^{[w]} / \sum_{w \in W} \sigma_i^{[w]} \tag{9}$$

Fig. 16 reports the values of $\widetilde{\sigma}_i^{[w]}$ for the 46 stations of the network. The figure allows us to notice the heterogeneity in normalised mobility trends between the stations through the seven months of the study.

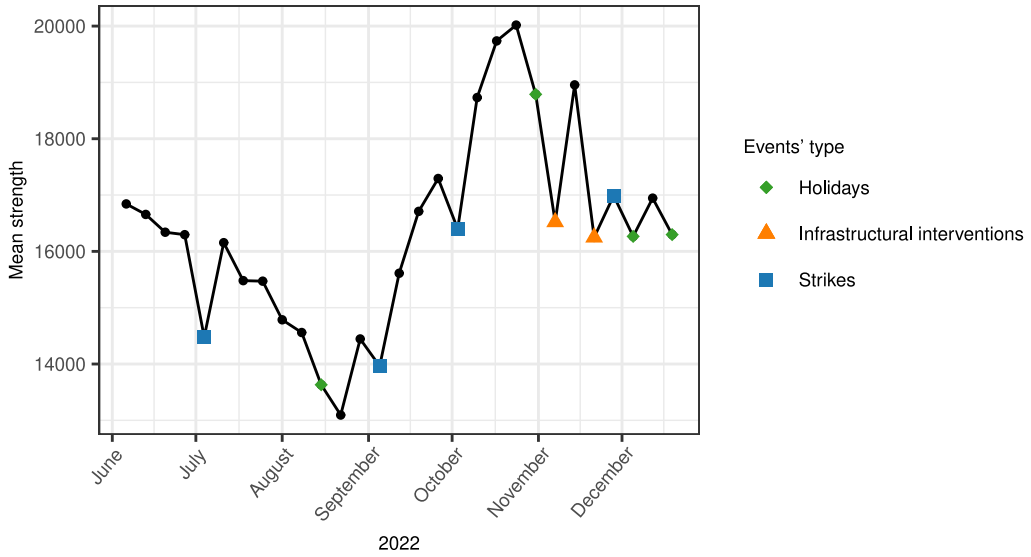

Fig. 15. Values of $\bar{\sigma}^{[w]}$ for each of the 29 weeks of the study. Coloured points indicate weeks when one of the events reported in Appendix B has happened. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

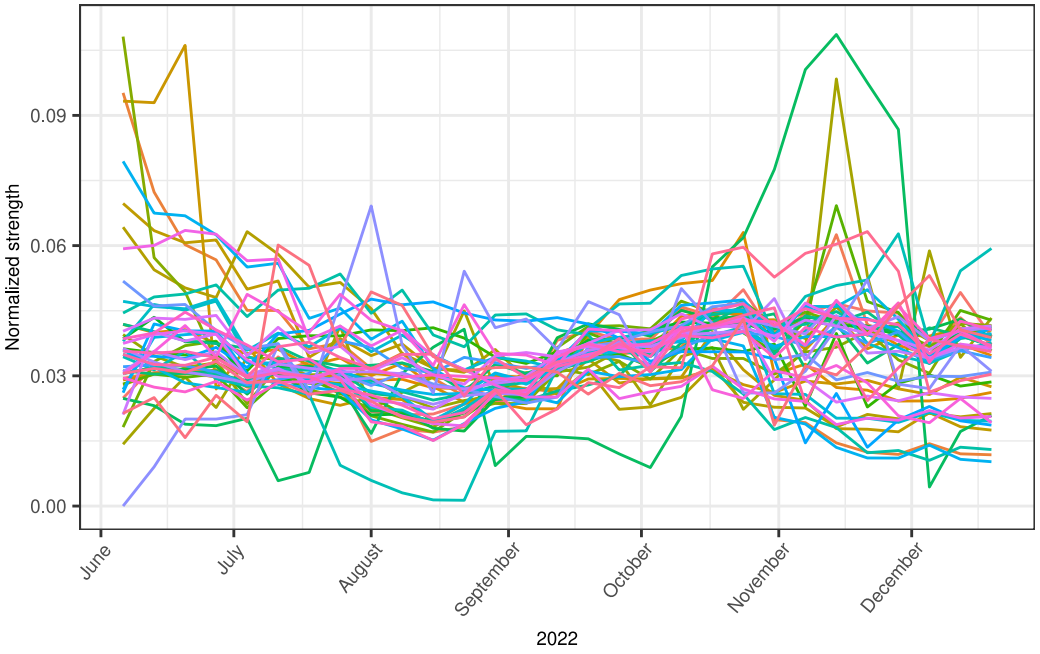

Fig. 16. Values of $\widetilde{\sigma}_i^{[w]}$ for each of the 46 stations $i \in S$ of the study.

To identify anomalous mobility trends at the station level, we apply FDA techniques. In particular, we compute the smoothed representation of $\widetilde{\sigma}_i^{[w]}$ fitting a cubic splines basis, using four basis functions and adding a roughness penalty on the second derivative of the curves. We selected the number of basis functions minimising the average generalised cross-validation error across all the curves. Fig. 17 shows the 46 smoothed functions during the 29 weeks of the study.

The smoothed functions are the starting point for further analyses involving FDA techniques, such as but not limited to functional clustering, outlier detection and principal components analysis. Since this Section focuses on highlighting anomalous mobility trends, we evaluate outliers based on the functional boxplot (Sun and Genton, 2010), which revealed 8 of the 46 stations as potential outliers. These 8 stations are shown in Fig. 18, categorised by their train lines: stations on line R14 display less heterogeneity in oscillations than the others, while those on line R4 exhibit higher amplitudes, indicating higher variability. Additionally, station *Milano Porta Garibaldi* shows oscillations in phase opposition to the average behaviour.

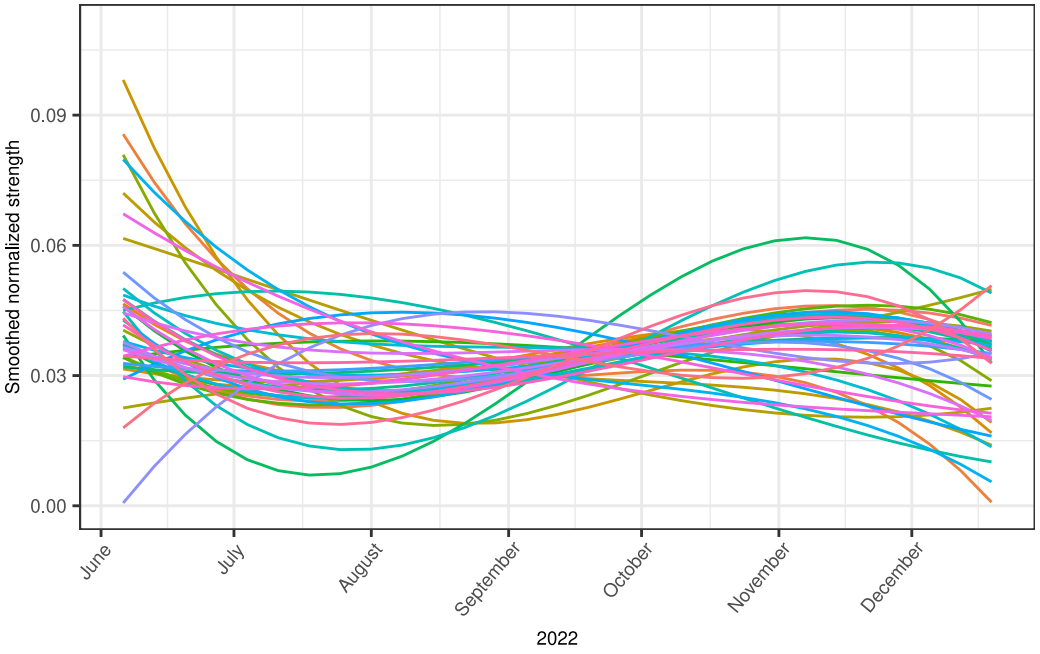

Fig. 17. Smoothed function of $\widetilde{\sigma}_i^{[w]}$ for each of the 46 stations of the study.

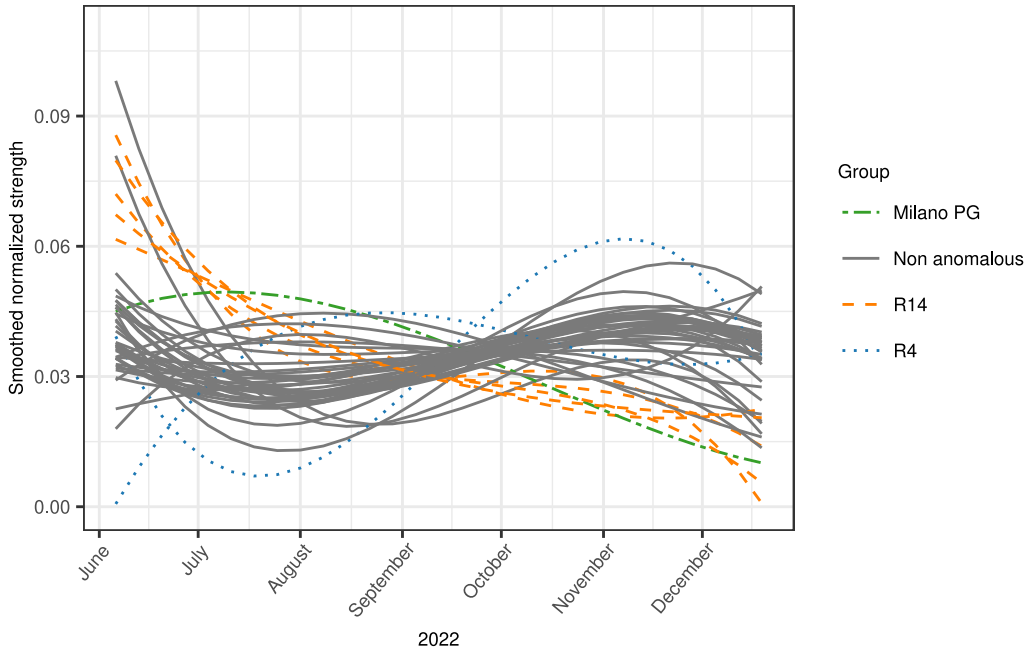

Fig. 18. Functional outliers revealed by the functional boxplot, interpreted together with the train lines whose station belongs to.

Thus, this Section provided a preliminary showcases of different tools to identify anomalies in the Trenord network on-the-fly, based on dynamic OD matrices. These insights highlight the utility of such matrices in understanding network behaviour, offering valuable applications for transportation operators and policymakers.

6. Conclusion, discussion and future directions

This work proposes a pipeline to estimate dynamic (e.g., weekly) OD matrices for a railway transportation network by combining data from ticket and subscription sales with passenger counts collected through the APC system.

While recent studies have emerged in the field of OD matrix estimation in public transport networks through ACDS systems, they often rely on specific types of AFC data, namely smart card information and travel surveys, which are not necessarily easy to get. In contrast, our approach proposes a methodology to derive accurate OD matrices estimates relying on data which are likely to become more and more easily available in the near future, due to digitalisation of services (as explained in Siebert and Ellenberger, 2020, APC data are expected to be installed on an increasingly growing proportion of transportation systems, making passenger counts readily available without the need for estimating unknown counts). Specifically, we construct OD matrices seeds using ticket and subscription sales data, which are easily available in several transportation companies, then refine the OD matrices estimates using the IPF algorithm. Thus, the primary innovation of our research lies in the procedure developed to transform ticket data into OD seeds, and to fuse the ticket-estimated OD seeds with the APC-derived passenger counts through the well-established IPF method.

The OD seeds generation is a crucial step, determining the reliability of the results of the IPF algorithm (Choupani and Mamdoohi, 2016). Nevertheless, despite several studies focus on OD matrix estimation, to the best of our knowledge, none have delved into the challenge of deriving accurate OD seeds from ticket and subscription sales data, fusing such data with passenger counts. For this reason, the actual contribution and the main novelty of our study within the realm of trip distribution modelling stands in the processes we designed to generate OD seeds from ticket and subscription sales data, including this into a very general methodological pipeline enabling estimation and monitoring of dynamic OD matrices. In fact, the APC counter-derived marginal data, representing boarded and alighted passengers at each station and week, is combined with the ticket-estimated OD seeds using the IPF algorithm. This iterative approach generates, in our application, 29 weekly OD matrices, portraying train movements across the six lines available within the Trenord network. These matrices adhere to reality, depicting high movement between stations connected by multiple lines and reduced trips during the summer and Christmas periods.

The low row margin errors, coupled with the consistent convergence of the iterative process, lend confidence to the robustness of our results. Further variability investigations and robustness assessment have been performed, for example segmenting the time frame into weekly vs week-end days, but no relevant differences nor novel insights have emerged, testifying for the reliability of our method. Unfortunately, the missingness of a ground truth does not allow for a direct assessment of the estimation error.

The pipeline we developed is fast, scalable and flexible. It can be extended to any reasonable time frame needed, leading to the opportunity to obtain estimations of recent developments in the mobility network. Thus, the accurate estimation of dynamic OD matrices serves as the foundation for post-hoc analyses of the network's behaviour. For example, such matrices may enhance operational insights for transportation operators, supporting demand study and schedule optimisation to match fluctuating demands throughout the year. Furthermore, by applying our methodology to various temporal periods, such as pre and post-COVID-19 pandemic times, these matrices offer valuable tools for exploring lifestyle changes and assessing shifts in commuting patterns. More in general, dynamic mobility analysis offers insights into the socio-economic dynamics of local communities by analysing mobility trends and anomalies. Understanding how people travel within a region provides valuable information on access to employment centres, educational facilities, healthcare services, and other essential resources. By addressing transportation barriers and improving connectivity based on data-driven insights, policymakers can support targeted interventions that strengthen the socio-economic resilience of communities and promote inclusive growth. Overall, the analysis contributes to enhancing resilience and adaptability in transportation systems.

In the framework of possible usage of dynamic matrices, we showed an example of such applications. We focused on anomaly detection, both at the global network and local station level, representing the dynamic OD matrices as temporal weighted networks. We applied a combination of network analysis and functional data analysis techniques to spot the use of our dynamic OD matrices as input for a dynamic decisional support tool. Indeed, we were able to identify anomalies at the global level and match disruptions in the global indicators to events such as strikes, holidays and network interventions. At the local station level, we identified stations showing anomalous mobility trends compared to the others in the period from June to December 2022.

Together, these highlights demonstrate the effectiveness and practical significance of the innovative algorithmic pipeline for estimating dynamic OD matrices in railway transportation networks, underscoring its contribution to data-driven decision-making and the strengthening of transportation systems for local communities, ultimately promoting sustainable development and community resilience.

6.1. Future directions

As acknowledged along the paper, despite concrete innovations have been proposed for enhancing the use of complex and diverse data in the field of transport network estimation, the proposed approach suffers from several limitations, due both to intrinsic methodological bottlenecks and application-specific issues. For this reason, looking ahead and assuming that future research efforts will lead towards extension of APC systems, some research directions may be easily identified as relevant to the general purpose of transportation network demand, estimation and monitoring.

First of all, the pipeline for initialising the IPF algorithm offers several opportunities of methodological development, especially in the direction of incorporating the domain knowledge into models. If the use of gravity model is concerned, it might be investigated how to allow parameters depend on covariates or classes of similar stations. Moreover, whenever issues arise with APC systems coverage, advanced methods for gap filling might be envisioned, e.g. regression models where suitable account for observations'

spatio-temporal dependency is allowed (for example, linear mixed-effect models with heteroscedastic residuals or even more complex models like (Lu et al., 2024)). All these methods might concur to reduce the bias and variability of the estimation, at least in the sense of providing more realistic and reliable seeds for IPF algorithm initialisation.

Another important issue comes from the lack of ground truth and then internal validation methods for assessing the accuracy of the proposed methodology. The current analysis is meaningful to the stakeholders for filling the gap of quantifying the people moving on the railway network. Since no updated estimates are available in this sense apart from Lombardia (2019), we are not able to compare our estimates with a real ground truth, providing a level of accuracy and goodness of fit. Nevertheless, we provided strategies for both internal and external validation in Section 4 for getting some proxies of the method reliability, and tested some time stratifications (for example, weekdays vs week-end) to have insight on the variability induced by time-frame adopted. Last but not least, a systematic analysis of how time granularity affects the hypotheses about OD symmetry should be developed for exploiting as a future direction the development of a system for monitoring the dynamic estimates of the OD matrices with proper control charts (Yeganeh et al., 2023).

The focus of the paper is not only on the accurate data acquisition and proper data fusion pipelines, but also on developing post-hoc analysis tools capable of harnessing the wealth of the information embedded in dynamic OD matrices. In fact, a relevant strength of the proposed approach derives from the dynamic output, consisting in a time-series of OD matrices. Such estimates enable post-hoc analysis and the dynamic monitoring of broader events related to mobility. For example, dynamic matrices may be instrumental in studying the impact of external events, such as pandemics (as was done in Galliani et al., 2023), strikes, or major public events, on transportation patterns and commuters/travellers habits. By comprehensively understanding these shifts, on large scale policymakers can make informed decisions, enhancing the resilience and adaptability of transportation systems to various challenges and allowing for operational optimisation and urban planning. Last but not least, on local scale they may have a deeper understanding of commuting patterns, peak travel times, and network congestion. Despite this potential gain, there is lack of methods for the analysis of complex temporal networks, as induced by dynamic OD matrices. By coupling these matrices with advanced analytical methods, researchers can unravel intricate mobility behaviours, spot outlying patterns and forecast travelling demand, offering crucial information for policy formulation and infrastructure planning. Techniques such those developed in Le Bail et al. (2023) may be applied to facilitate the comparative analysis of temporal networks induced by successions of dynamic OD matrices. Examples are network-based control charts (Yeganeh et al., 2023) or methods for spotting local differences in matrices (Alshaer et al., 2020). Note that all these techniques require the estimation of the full OD matrix and cannot be based on marginal counts only.

In the end, this work opens avenues for future research in the broader context of transportation network analysis. The pipeline's modular design allows for the exploration of alternative methods for different components and the application to other transportation networks beyond railways. As mobility patterns evolve, incorporating additional data sources could enrich the accuracy and relevance of dynamic OD matrices. Ultimately, our study bridges the gap between ticket sales and passenger count data, offering a comprehensive solution to estimating dynamic OD matrices in complex transportation networks and contributing to a more informed and adaptable approach to urban mobility planning.

CRediT authorship contribution statement

Greta Galliani: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation. **Piercesare Secchi:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Conceptualization. **Francesca Ieva:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization.

Data availability

Code to replicate all results in this paper can be accessed at https://github.com/GretaGalliani/dynamic-OD-estimation-railway-network, together with a synthetic version of the dataset.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT in order to edit the written article and rewrite some pieces aiming for fluency and clarity. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgements

We thank Trenord for the collaboration and for sharing the data used in this work. In particular, we thank Dr. Marta Galvani and Dr. Giovanni Chiodi for their support and insightful suggestions. The authors acknowledge the support by MUR, Italy, grant Dipartimento di Eccellenza 2023–2027. This study was funded by the European Union - NextGenerationEU, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP D43C22003110001). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

Appendix A. Assumptions to convert ticket and subscription sales data into estimated od trips

See Table A.1.

Appendix B. Events influencing the Trenord network's dynamics

See Table B.1.

Appendix C. Synthetic data description and usage

A synthetic dataset mimicking the one illustrated in the application is provided in the repository https://github.com/GretaGalliani/dynamic-OD-estimation-railway-network. For the sake of clarity, none of the information provided has any real correspondence with original Trenord Data. It aims at reproducing the dynamic observed in the Trenord data, reporting movements for six lines as it is in the study. The fictitious timeline is the period June-December 2022.

Table A.1
Assumptions needed to convert each record in the ticket and subscription sales dataset into estimated OD trips attributed to each week of the study period.

Ticket type	Conversion in the dynamic seed OD matrices
Ordinary ticket, special rate initiative, additional exaction	We extract a random day in the 7 days following the purchase and attribute 0.5 trips between origin and destination and 0.5 between destination and origin to the extracted day. This is because each ticket can be used in either direction between the two stations for which it has been emitted, so we split the number of trips evenly in the two directions.
Carnet	For carnets, we randomly extract 5 days in the 30 days following the carnet's purchase. We suppose a round trip is made in the 5 days drawn and then aggregate weekly. We chose the period of 30 days to extract the trips because it was previously the validity period of the carnet, while now carnets do not have an expiration date.
Weekly subscription	For weekly subscriptions, we suppose 5 round trips attributed to the current week if the subscription is bought between Monday and Wednesday, to the following week if the subscription is purchased between Thursday and Sunday.
Monthly subscription	For monthly subscriptions, round trips are distributed into the month's weeks starting from the day of selling. The month of usage is the current month if the subscription is bought before the 22nd of the month or the following month if it is purchased on the 22nd or the days after. We suppose 5 round trips for full weeks (i.e., entirely belonging to the subscription month). For partial weeks, we use the correspondences obtained by computing and rounding the proportion $\frac{5\ round\ trips}{7\ days} * n\ partial\ days$.
Yearly subscription	For yearly subscriptions, we attribute 5 round trips to each complete week starting from the day of purchasing and ending the last day of the 12th month after purchase, applying the same convention to uncomplete weeks (if any) used for monthly subscriptions

Table B.1
Events influencing Trenord's network dynamics in 2022.

Event type	Date	Notes
Strikes	July 10–11, 2022	Strike of Trenord's train crews
	September 9, 2022	Strike of Trenord's train crews
	October 8–9, 2022	Strike of Trenord's train crews
	December 2, 2022	Strike of Trenord's train crews
Holidays	August 15, 2022	Assumption Day - National Italian holiday
	November 1, 2022	All Saints' Day - National Italian holiday
	December 8, 2022	Immaculate Conception - National Italian holiday. Moreover, December 7, 2022, is Saint Ambrose Day, which is Milan's holiday.
	December 25, 2022	Christmas - National Italian holiday
	December 26, 2022	St. Stephen's Day - National Italian holiday
Infrastructural interventions	November 12–14, 2022	Infrastructural construction work on line RE_6, causing the closure of stations *Desenzano del Garda - Sirmione*, *Peschiera del Garda* and *Verona Porta Nuova*
	November 26–28, 2022	Infrastructural construction work on line RE_6, causing the closure of stations *Desenzano del Garda - Sirmione*, *Peschiera del Garda* and *Verona Porta Nuova*

Table C.1
Description of the variables contained in the *ticket* dataset.

Variable	Explanation
`COD_DES`	Code of destination station.
`COD_ORI`	Code of origin station.
`Date`	Date in the period May–December 2022,
`Type`	A ticket type randomly extracted between the ones shown in Fig. 4.
`Quantity`	A quantity randomly extracted, assuming values in $\{0.5, 1, 1.5, 2, 2.5, \dots\}$. This quantity expresses the number of tickets of the considered type purchased in the specified date for the couple of origin and destination stations. Fractionary values indicate rates shared with another public transport operator.

The repository is composed by two synthetic datasets (namely *ticket.csv* e *train.csv*), whose variables are illustrated in Table C.1 and Table C.2. In particular, dataset *ticket.csv* is used to build the seed OD matrices induced by ticket data, while dataset *train.csv*

Table C.2
Description of variables contained in the *train* dataset.

Variable	Explanation
`MissionDate`	A date randomly extracted in the period June-December 2022.
`TrainCode`	Train ride numerical code, identifying the train ride for the considered day. Train rides are uniquely identified by the couple (`MissionDate`, `TrainCode`)
`Line`	A code in the set $\{L1, L2, L3, L4, L5, L6\}$ describing 6 train lines.
`DepartureStation`	Code describing the initial departure station of the train ride. Correspondence with the codes in dataset `ticket` is guaranteed.
`ArrivalStation`	Anonymised code describing the final arrival station of the train ride. Correspondence with the codes in dataset `ticket` is guaranteed.
`DepartureTime`	Information about the year, month, day, hour and minute at which the train ride has left the initial departure station.
`ArrivalTime`	Information about the year, month, day, hour and minute, at which the train ride has arrived at the final arrival station.

(*continued on next page*)

Table C.2 (*continued*).

Variable	Explanation
`StopIndex`	Integer number describing the order in which the ride stops at the station. It starts from 1 for the initial departure station. Missing values in the sequence indicate that the train ride identified by (`MissionDate`, `TrainCode`) has not stopped at every station in the line.
`StopStation`	Code of the stop station of the train ride. Correspondence with the codes in dataset `ticket` is guaranteed.
`EntryTime`	Information about the year, month, day, hour and minute, at which the train ride has entered the station identified by `StopStation`. This field is empty when the stop station is the initial departure station, as this information is reported in variable `DepartureTime`.
`ExitTime`	Information about the year, month, day, hour and minute, at which the train ride has left the station identified by `StopStation`. This field is empty when the stop station is the final arrival station, as this information is reported in variable `ArrivalTime`.
`PassengersBoarded`	Number of passengers boarded at the considered stop station.
`PassengersExited`	Passengers disembarking at the considered stop station.[a]
`Status`	A code describing the status of the APC system for the train ride. "-1" identifies a cancelled train ride, "0" identifies a train ride with missing APC data and "1" identifies a ride with reliable APC data.

[a] Coherence between the total number of boarded passengers for the train ride and the total number of passengers disembarking is not assumed.

contains the needed information to derive the partial counts of passengers boarding and disembarking at each station of the network and the travel times between each couple of station.

References

Ait-Ali, A., Eliasson, J., 2019. Dynamic origin-destination estimation using smart card data: An entropy maximisation approach. http://dx.doi.org/10.48550/arXiv.1909.0282, ArXiv Preprint.

Alshaer, M., et al., 2020. Detecting anomalies from streaming time series using matrix profile and shapelets learning. In: 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence. ICTAI, IEEE, http://dx.doi.org/10.1109/ICTAI50040.2020.00066.

H° akegård, J., et al., 2018. Statistical modelling for estimation of OD matrices for public transport using wi-fi and APC data. In: 2018 21st International Conference on Intelligent Transportation Systems. ITSC, pp. 1005–1010. http://dx.doi.org/10.1109/ITSC.2018.8570009.

Barbosa, H., et al., 2018. Human mobility: Models and applications. Phys. Rep. (ISSN: 0370-1573) 734, 1–74. http://dx.doi.org/10.1016/j.physrep.2018.01.001, URL https://www.sciencedirect.com/science/article/pii/S037015731830022X.

Barthélemy, J., Suesse, T., 2018. Mipfp: An R package for multidimensional array fitting and simulating multivariate Bernoulli distributions. J. Stat. Softw. Code Snippets 86 (2), 1–20. http://dx.doi.org/10.18637/jss.v086.c02, URL https://www.jstatsoft.org/index.php/jss/article/view/v086c02.

Ben-Akiva, M., Morikawa, T., 1989. Data fusion methods and their applications to origin-destination trip tables. In: Transport Policy, Management & Technology Towards 2001: Selected Proceedings of the Fifth World Conference on Transport Research, Vol. 4.

Bierlaire, M., Crittin, F., 2004. An efficient algorithm for real-time estimation and prediction of dynamic OD tables. Oper. Res. 52 (1), 116–127.

Cerqueira, S., et al., 2022. Inference of dynamic origin-destination matrices with trip and transfer status from individual smart card data. Eur. Transp. Res. Rev. 14 (1), 42. http://dx.doi.org/10.1186/s12544-022-00562-1.

Chen, Z., Fan, W., 2020. Extracting bus transit boarding and alighting information using smart card transaction data. J. Public Transp. 22 (1), 40–56. http://dx.doi.org/10.5038/2375-0901.22.1.3.

Choupani, A., Mamdoohi, A., 2016. Population synthesis using iterative proportional fitting (IPF): A review and future research. Transp. Res. Procedia 17, 223–233. http://dx.doi.org/10.1016/j.trpro.2016.11.078.

Cui, A., 2006. Bus Passenger Origin-Destination Matrix Estimation Using Automated Data Collection Systems (thesis). Massachusetts Institute of Technology.

de Palma, A., et al., 2022. An overview of effects of COVID-19 on mobility and lifestyle: 18 months since the outbreak. Transp. Res. A (ISSN: 0965-8564) 159, 372–397. http://dx.doi.org/10.1016/j.tra.2022.03.024, URL https://www.sciencedirect.com/science/article/pii/S0965856422000714.

Douglas, L., 2015. A User's Guide to Network Analysis in R, first ed. Springer, http://dx.doi.org/10.1007/978-3-319-23883-8.

Evans, A., 1970. Some properties of trip distribution methods. Transp. Res. 4, 19–36.

Fujita, M., et al., 2017. Time coefficient estimation for hourly origin-destination demand from observed link flow based on semidynamic traffic assignment. J. Adv. Transp. 22 (6495861), http://dx.doi.org/10.1155/2017/6495861.

Galliani, G., et al., 2023. The impact of public transport on the diffusion of COVID-19 pandemic in lombardy during 2020. Med. Res. Arch. (ISSN: 2375-1924) 11 (9), http://dx.doi.org/10.18103/mra.v11i9.4356, URL https://esmed.org/MRA/mra/article/view/4356.

Ge, Q., Fukuda, D., 2016. Updating origin-destination matrices with aggregated data of GPS traces. Transp. Res. C 69, 291–312.

Gordon, J., 2012. Intermodal Passenger Flows on London's Public Transport Network: Automated Inference of Full Passenger Journeys Using Fare-Transaction and Vehicle-Location Data (thesis). Massachusetts Institute of Technology.

Hazelton, M., 2010. Statistical inference for transit system origin-destination matrices. Technometrics 52 (2), 221–230.

Hu, T., et al., 2021. Human mobility data in the COVID-19 pandemic: characteristics, applications, and challenges. Int. J. Digit. Earth 14 (9), 1126–1147. http://dx.doi.org/10.1080/17538947.2021.1952324.

Huo, J., et al., 2023. Simulation-based dynamic origin–destination matrix estimation on freeways: A Bayesian optimization approach. Transp. Res. E 173, 103108. http://dx.doi.org/10.1016/j.tre.2023.103108.

Hussain, E., et al., 2021. Transit OD matrix estimation using smartcard data: Recent developments and future research challenges. Transp. Res. C 125, http://dx.doi.org/10.1016/j.trc.2021.103044.

Jafari Kang, M., et al., 2020. A procedure for public transit OD matrix generation using smart card transaction data. Public Transp. 13 (1), 81–100. http://dx.doi.org/10.1007/s12469-019-00216-y.
Ji, Y., et al., 2014. Estimating transit route OD flow matrices from APC data on multiple bus trips using the IPF method with an iteratively improved base: Method and empirical evaluation. J. Transp. Eng. 140 (5), 04014008. http://dx.doi.org/10.1061/(ASCE)TE.1943-5436.0000639.
Khoshkhah, K., et al., 2022. Real-time system for daily modal split estimation and OD matrices generation using IoT data: A case study of tartu city. Sensors (Basel) 22 (8), 3030. http://dx.doi.org/10.3390/s22083030.
Lam, W., et al., 2003. Estimation of transit origin-destination matrices from passenger counts using a frequency-based approach. J. Math. Model. Algorithms 2 (4), 329–348.
Le Bail, D., et al., 2023. Flow of temporal network properties under local aggregation and time shuffling: a tool for characterizing, comparing and classifying temporal networks. http://dx.doi.org/10.48550/arXiv.2310.09112.
Liu, X., et al., 2021. Optimization models for estimating transit network origin-destination flows with big transit data. J. Big Data Anal. Transp. 3 (3), 247–262. http://dx.doi.org/10.1186/s42466-021-00115-5.
Lombardia, R., 2019. Matrice OD2020 - passeggeri. Data retrieved from Regione Lombardia Open Data, URL https://www.dati.lombardia.it/Mobilit-e-trasporti/Matrice-OD2020-Passeggeri/hyqr-mpe2. (Last accessed 13 April 2023).
Low, D.E., 1972. New approach to transportation systems modeling. Traffic Q. 26 (3).
Lu, J., et al., 2024. Spatial-temporal memory enhanced multi-level attention network for origin-destination demand prediction. Complex Intell. Syst. http://dx.doi.org/10.1007/s40747-024-01494-0.
Macgill, S., 1977. Theoretical properties of biproportional matrix adjustments. Environ. Plan. A: Econ. Space 9 (6), 687–701.
Mohammed, M., Oke, J., 2023. Origin-destination inference in public transportation systems: A comprehensive review. Int. J. Transp. Sci. Technol. (ISSN: 2046-0430) 12 (1), 315–328. http://dx.doi.org/10.1016/j.ijtst.2022.03.002, URL https://www.sciencedirect.com/science/article/pii/S2046043022000223.
Munizaga, M., Palma, C., 2012. Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from santiago, Chile. Transp. Res. C 24, 9–18. http://dx.doi.org/10.1016/j.trc.2011.10.001.
Mussone, L., Matteucci, M., 2013. OD matrices network estimation from link counts by neural networks. J. Transp. Syst. Eng. Inf. Technol. (ISSN: 1570-6672) 13 (4), 84–92. http://dx.doi.org/10.1016/S1570-6672(13)60117-8, URL https://www.sciencedirect.com/science/article/pii/S1570667213601178.
Navick, D., Furth, P., 1994. Distance-based model for estimating a bus route origin-destination matrix. Transp. Res. Rec. 16.
Ortúzar, J., Willumsen, L., 2011. Modelling Transport, fourth ed. Wiley, Hoboken, http://dx.doi.org/10.1002/9781119993308.
Pamuła, T., Żochowska, R., 2023. Estimation and prediction of the OD matrix in uncongested urban road network based on traffic flows using deep learning. Eng. Appl. Artif. Intell. 117, 105550. http://dx.doi.org/10.1016/j.engappai.2022.105550.
R. Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.
Ramsay, J., Silverman, B., 2006. Functional Data Analysis, second ed. Springer, http://dx.doi.org/10.1007/b98888.
Robillard, P., 1975. Estimating the OD matrix from observed link volumes. Transpo. Res. 9 (2–3), 123–128.
Siebert, M., Ellenberger, D., 2020. Validation of automatic passenger counting: introducing the t-test-induced equivalence test. Transportation 47, 3031–3045. http://dx.doi.org/10.1007/s11116-019-09991-9.
Simini, F., et al., 2012. A universal model for mobility and migration patterns. Nature 484 (7392), 96–100. http://dx.doi.org/10.1038/nature10856.
Sun, Y., Genton, M., 2010. Functional boxplot. J. Comput. Graph. Statist. 20, http://dx.doi.org/10.2307/23110490.
Toqué, F., et al., 2016. Forecasting dynamic public transport origin-destination matrices with long-short term memory recurrent neural networks. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems. ITSC, pp. 1071–1076. http://dx.doi.org/10.1109/ITSC.2016.7795689.
Torti, A., et al., 2021. Analysing Transportation System Reliability: The Case Study of the Metro System of Milan. Technical Report, MOX-Report No. 84/2021, URL https://www.mate.polimi.it/biblioteca/add/qmox/84-2021.pdf.
Trenord, 2022a. Biglietti e Abbonamenti. URL https://www.trenord.it/biglietti/. (Last accessed 13 April 2023).
Trenord, 2022b. Il bilancio di sostenibilità 2021. URL https://www.trenord.it/chi-siamo/bds/. (Last accessed 13 April 2023).
Trenord, 2022c. Linee e Orari. URL https://www.trenord.it/linee-e-orari/. (Last accessed 13 April 2023).
Trenord, 2022d. Linee regionali. URL https://www.trenord.it/linee-e-orari/il-nostro-servizio/linee-regionali/. (Last accessed 13 April 2023).
Trenord, 2023. Chi siamo. URL https://www.trenord.it/chi-siamo/. (Last accessed 13 April 2023).
Trenord, 2024a. Fares rules. URL https://normelombardia.consiglio.regione.lombardia.it/NormeLombardia/Accessibile/main.aspx?iddoc=rr0020140610000004&view=showdoc. (Last accessed 28 June 2024).
Trenord, 2024b. Fares tables. URL https://www.trenord.it/fileadmin/contenuti/TRENORD/4-Info_e_assistenza/Informazioni_utili/Condizioni_di_trasporto/Condizioni_di_trasporto_IN_VIGORE/AvvisoA3Trenord_2024_062_Tariffe_dall_11_aprile.pdf. (Last accessed 28 June 2024).
Wang, W., 2010. Bus Passenger Origin-Destination Estimation and Travel Behavior Using Automated Data Collection Systems in London, UK (thesis). Massachusetts Institute of Technology.
Welch, T.F., Widita, A., 2019. Big data in public transportation: a review of sources and methods. Transp. Rev. 39 (6), 795–818. http://dx.doi.org/10.1080/01441647.2019.1616849.
Wheeler, J., 2005. Geography. In: Kempf-Leonard, K. (Ed.), Encyclopedia of Social Measurement. Elsevier, New York, ISBN: 978-0-12-369398-3, pp. 115–123. http://dx.doi.org/10.1016/B0-12-369398-5/00277-2, URL https://www.sciencedirect.com/science/article/pii/B0123693985002772.
Wilson, A.G., 1970. The use of the concept of entropy in system modelling. J. Oper. Res. Soc. 21, 247–265.
Wong, K.o., 2005. Estimation of origin-destination matrices for a multimodal public transit network. J. Adv. Transp. 39 (2), 139–168.
Wu, L., et al., 2021. Inferring origin-destination demand and user preferences in a multi-modal travel environment using automated fare collection data. Omega (ISSN: 0305-0483) 101, 102260. http://dx.doi.org/10.1016/j.omega.2020.102260, URL https://www.sciencedirect.com/science/article/pii/S0305048319313490.
Yang, Y., et al., 2020. Dynamic origin-destination matrix estimation based on urban rail transit AFC data: Deep optimization framework with forward passing and backpropagation techniques. J. Adv. Transp. 2020, http://dx.doi.org/10.1155/2020/8846715.
Yeganeh, A., et al., 2023. A network surveillance approach using machine learning based control charts. Expert Syst. Appl. 19, http://dx.doi.org/10.1016/j.eswa.2023.119660.
Yun, I., Park, B., 2005. Estimation of dynamic origin destination matrix: a genetic algorithm approach. In: Proceedings. 2005 IEEE Intelligent Transportation Systems. pp. 522–527. http://dx.doi.org/10.1109/ITSC.2005.1520080.
Zannat, K., Choudhury, C., 2019. Emerging big data sources for public transport planning: A systematic review on current state of art and future research directions. J. Indian Inst. Sci. 99 (4), 601–619.
Zeng, Q., et al., 2015. Development of application for estimating daily boarding and alighting counts on New York City buses: Implementation of daily production system. Transp. Res. Rec. 2535 (1), 1–14. http://dx.doi.org/10.3141/2535-01.
Zhao, J., 2004. The Planning and Analysis Implications of Automated Data Collection Systems: Rail Transit OD Matrix Inference and Path Choice Modeling Examples (thesis). Massachusetts Institute of Technology.
Zhao, J., et al., 2007. Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. Comput.-Aided Civ. Infrastruct. Eng. 22 (5), 376–387. http://dx.doi.org/10.1111/j.1467-8667.2007.00500.x.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

Galliani, G.; Secchi, P.; Ieva, F.
Estimation of dynamic Origin–Destination matrices in a railway transportation network integrating ticket sales and passenger count

68/2024 Gambarini, M.; Ciaramella, G.; Miglio, E.
A gradient flow approach for combined layout-control design of wave energy parks

64/2024 Cavazzutti, M.; Arnone, E.; Ferraccioli, F.; Galimberti, C.; Finos, L.; Sangalli, L.M.
Sign-Flip inference for spatial regression with differential regularization

65/2024 Possenti, L.; Vitullo, P.; Cicchetti, A.; Zunino, P.; Rancati, T.
Modeling Hypoxia Induced Radiation Resistance and the Impact of Radiation Sources

63/2024 Vitullo, P.; Franco, N.R.; Zunino, P.
Deep learning enhanced cost-aware multi-fidelity uncertainty quantification of a computational model for radiotherapy

62/2024 Roknian, A.A.; Scotti, A.; Fumagalli, A.
Free convection in fractured porous media: a numerical study

60/2024 Temellini, E.; Ferro, N.; Stabile, G.; Delgado Avila, E.; Chacon Rebollo, T.; Perotto, S.
Space - time mesh adaptation for the VMS - Smagorinsky modeling of high Reynolds number flows

61/2024 Speroni, G.; Ferro, N.
A novel metric - based mesh adaptation algorithm for 3D periodic domains

59/2024 Carbonaro, D.; Ferro, N.; Mezzadri, F.; Gallo, D.; Audenino, A.; Perotto, S.; Morbiducci, U.; Chiastra, C.
Easy-to-use formulations based on the homogenization theory for vascular stent design and mechanical characterization

Possenti, L.; Vitullo, P.; Cicchetti, A.; Zunino, P.; Rancati, T.
Modeling Hypoxia-Induced Radiation Resistance and the Impact of Radiation Sources