Transportation Research Part A 199 (2025) 104550

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

TRANSPORTATION RESEARCH

Part A: Policy and Practice

Resilience of public transport in the face of disruptions: Insights from explainable machine learning

Benjamin Cottreau [a,b,*], Mehmet Güney Celbiş [a], Ouassim Manout [a], Louafi Bouzouina [a]

[a] *LAET ENTPE-CNRS-University of Lyon, France*

[b] *Keolis Lyon, France*

ARTICLE INFO

Keywords:
Disruptions
Resilience
Public transit
Demand management
Smart card
Machine learning

ABSTRACT

Disruptions in public transport (PT) can have a major impact on passenger activities and on the attractiveness of the service, particularly when they are not absorbed by the network as a whole. The present study aims to detect the presence of disruption and assess the contribution of existing alternative bus or tramway stops to the resilience of the PT network, using explainable machine learning techniques. The detection task is formulated as a supervised classification problem performed using Random Forest (RF) for 39 different subway stations, using Automatic Fare Collection (AFC) data and Service Disruption logs (SD-logs). Furthermore, the SHapley Additive exPlanation (SHAP) interpretation method is implemented to retrieve the magnitude and the direction of each alternative stop's contribution to PT resilience. Results show that the proposed modeling framework has high prediction performance, can minimize false alarm rates, and can foresee the occurrence of disruptions 5 min before their registered beginning in SD-logs. Findings also indicate where demand is reallocated, resulting in 5 different resilience clusters for subway stations. Density and connectivity emerge as two major attributes of resilience that have a central role in the design of disruption management (tactical) and development (strategical) plans. The proposed approach has been applied to the PT network of Lyon (France) and is replicable by adapting the hyperparameters to the observed use in other PT networks.

1. Introduction

Disruptions in public transport (PT) systems have critical consequences on their functionality and attractiveness, particularly when the network as a whole fails to absorb them. Such disruptions hinder the system's ability to provide reliable user accessibility, and increase the risk of network degradation challenging the resilience of the system. The impacts of disruptions on PT systems in general include delays, trip cancellations, and significant economic losses (Cox et al., 2011; Mo et al., 2023). For service providers, disruptions incur additional costs such as overtime payments for staff, fare reimbursements to passengers, and fines under certain contractual agreements with PT authorities (Yap and Cats, 2021). Over time, repeated disruptions can erode passenger confidence in the system, leading to perceived unreliability resulting in declining ridership, which in turn, further diminishes revenue (Yap and Cats, 2021).

Disruptions also substantially reduce satisfaction levels and undermine the overall resilience of the transportation network (Rahimi et al., 2019). For passengers, disruptions often force unplanned adjustments, such as finding alternative routes or coping with

[*] Correspondence to: LAET-ENTPE, Université de Lyon, CNRS, Rue Maurice Audin, 69518 Vaux-en-Velin Cedex, France.
E-mail address: benjamin.cottreau@entpe.fr (B. Cottreau).

https://doi.org/10.1016/j.tra.2025.104550
Received 8 January 2025; Received in revised form 22 May 2025; Accepted 24 May 2025

Available online 17 June 2025

severe delays (Müller et al., 2020). When such incidents disrupt lines or stations, they lead to extended travel times and altered travel behaviors (Sun et al., 2016). Improved planning and implementation of both preventive measures and restorative actions can mitigate the negative effects of disruptions (Vodopivec and Miller-Hooks, 2019).

If disruption is defined as an event that causes a substantial deviation from the usual performance levels (Bešinović, 2020), resilience refers to the disruptions that can be absorbed (Reggiani et al., 2015) and the ability of the PT system to maintain functions during these interruptions (Liu et al., 2024). Considering railway transport systems, Bešinović (2020) identified three stages to characterize disruptions. First, the performance of the PT system involved translates from regular to disrupted performance levels. Hence, the first step of disruption management strategies is **the detection of the disruption**. Then, the system reaches a steady state, where the performance is substantially lower than usual. In this context, the purpose of disruption management strategies is to simultaneously reduce the intensity and duration of the disruption. To compensate for the induced lack of performance, the operator can provide temporary supply alternatives (i.e. **supply management**), and passengers can be redirected to existing alternative supply (i.e. **demand management**). Finally, the PT system recovers and switches from disrupted to regular performance levels.

Disruptions have heterogeneous profiles and can be characterized by their source (endogenous or exogenous) and predictability (planned or unplanned). Planned disruptions, endogenous or exogenous, refer to *special events* (Noursalehi et al., 2018) that require specific planning policies to be handled in advance. Most planned endogenous disruptions provoke a supply reduction (e.g. planned maintenance), while most planned exogenous disruptions provoke an increase in demand (e.g. concerts). These disruptions result in temporary adaptation handled by PT operators, who adjust PT schedules to new service availability (service-based management) or provide relevant information to users to reschedule their trips themselves (demand-based management). In this work, we are specifically interested in **unplanned disruptions**, which are more challenging in terms of operations because they need to be handled in real-time. These disruptions are mostly provoked by service issues, such as abandoned baggage, power failure or informatic failure (Cottreau et al., 2025), but also can be related to external and global issues such as pandemics (El Zein et al., 2022; Cottreau et al., 2023) or safety issues (Cox et al., 2011).

The specificity of urban PT systems lies in their high spatial density and service connectivity. Considered as “the ability to create and maintain a connection between two or more points in a spatial system”, connectivity is a key element for “the development of resilience in transport systems” (Reggiani et al., 2015). In terms of spatial density, the distance between stations (or stops) in urban PT systems is in the range of walkable distance, with dedicated infrastructures that might ease the switch from one alternative to another. In terms of time, urban PT schedules are built to provide a supply in the range of a few minutes for a specific origin–destination (OD) trip, while time intervals are broader when dealing with national or regional train lines. Consequently, when a disruption occurs in urban PT systems, numerous reliable and relevant alternatives are available, leading to spontaneous flow redistribution. Therefore, demand management is a powerful tool to help urban PT operators enhance the redistribution process.

In recent years, PT demand management has been enhanced by increasing computing capabilities and increasing data availability. More specifically, the fine-grained data collected by operators through Automatic Fare Collection (AFC), Automatic Vehicle Location (AVL), and Service Disruptions (SD) logs are now used to track the performance of PT systems in real time. Data-driven management strategies have already been explored and have been proven to improve operators’ ability to face disruptions compared to the use of operational reports (Jasperse, 2020). In this process, the detection step is crucial because it is the first step of any disruption management plan. Once the event is detected, operational responses are brought in to restore the supply and reallocate PT flows. Several studies use a topological approach (Zhang et al., 2015; Cats, 2016; Massobrio and Cats, 2024) to measure the contribution of stops or stations to the resilience of the PT system. The methodology used in these works consists of retrieving the shortest paths in a given graph and combining them with service data to assess several aspects of resilience (e.g. robustness or recoverability). Performance is often tracked through passenger travel time, and disruption effects are assessed by removing links from the graph. Compared to this approach, the advantage of AFC data is to work with real demand flows, and therefore propose even more specific recommendations to build resilient PT networks. Along these lines, Yap and Cats (2021) used demand data to predict stations’ exposure to disruption and clustered them based on their criticality, i.e. their contribution to the system’s vulnerability. In return, our study aims to analyze the contribution of alternative stops to the system’s resilience. In this regard, this paper aims to answer the two following research questions: *To what extent can data-driven methods be implemented to detect PT disruptions? How can we measure the contribution of existing alternative options to the resilience of PT systems?* We approach these research questions from a data partitioning-based machine learning perspective, applied to the resilience of urban PT systems to disruptions in subway service. This work proposes a short-term and demand-oriented framework that provides useful insights to improve the real-time management of disruptions and to enhance the resilience of PT networks at the operational level. It notably uses an anomaly detection framework to constantly track the probability of facing disruptions, which aims to improve the responsiveness of PT operators in the event of disturbances. In addition, the use of explainable machine learning with SHapley Additive eXplanation (SHAP) enables to measure the contribution of existing stops in the demand reallocation process provoked by disruptions. This approach helps identify potential crowding issues and suboptimal flow redistribution that could be improved with targeted real-time information. To the best of our knowledge, this study is the first to combine disruption (i.e. anomaly) detection algorithms with a measurement of the contribution of alternative stops to the resilience of PT systems.

The next Section 2 provides a general literature review (Background). The methodology and data description are explained in Section 3. The results are presented and discussed in Section 4.

2. Background

In this section, we address the problem of disruption detection from the perspective of PT systems, but also with a generic methodological approach. Section 2.1 gives an overview of how the notion of resilience is treated in PT literature and what kinds of data are at hand. However, there is a rich literature on disruption detection outside the scope of PT. More generally, the notion of *disruption* is referred to as *anomaly* in the literature. In addition, the detection problem can be translated into a binary classification task, where one predicts the presence or the absence of a disruption. For these reasons, Section 2.2 focuses on the existing extensive literature on *anomaly detection*, while Section 2.3 deals with the well-studied topic of classification algorithms applied to anomaly detection problems.

Throughout the literature review, three main objectives are defined and broken down into several working directions to build our modeling framework.

2.1. Resilience and performance assessment of public transport systems

In the field of PT, most studies have tackled the problem of resilience from the perspective of graph theory. In that regard, vulnerability and robustness are two corollary concepts often studied in the literature. *Vulnerability* refers to the extent to which system performance is impacted by a disruption (Ge et al., 2022), while *robustness* is defined as the ability of a system to withstand disruptions with minimal impact on performance (Cats, 2016). This approach can be used with topological data (Zhang et al., 2015), demand-based data such as OD matrices (Rodríguez-Núñez and García-Palomares, 2014), service-based data (Massobrio and Cats, 2024) or costs (Khaled et al., 2015).

Other works investigate the impact of disruptions at the station levels thanks to service data. Delays are a measurement of the deviation between planned and actual performance, which is often used in the literature (Weng et al., 2014; Louie et al., 2017; Marra and Corman, 2019; Zhang et al., 2022). However, it does not provide complete information on how the system reacts to a disruption. The unavailability of local supply measured by delays provokes flow redistribution which needs to be measured with demand data.

Several demand-based performance measures can be used to characterize disruption. When Origin–Destination (OD) matrices are available, passenger flows (Yap et al., 2018), travel times (Mo et al., 2023) or passenger delays (Sun et al., 2016) can be retrieved and used to characterize the response of PT systems to disruptions. This information can also be translated into generalized costs (Yap and Cats, 2022), which help to appraise the impact of disruption with revenue loss. These user-oriented metrics provide useful insight into the flow redistribution process and the preparedness of PT operators toward disruption scenarios. The main drawback of OD matrices is that they are less suitable for real-time disruption management strategies. Considering tap-in-only PT systems, the destination inference procedure is time-consuming, and OD matrices cannot always be retrieved quickly. Consequently, an approach using rapidly-available demand data might be more universal, applicable for tactic (i.e. short-term) and strategic (i.e. long-term) purposes, and for a large variety of PT systems.

Raw ridership data answers these needs. First, it is available for all PT networks equipped with an Automatic Fare Collection system (AFC), including tap-in only systems. Second, it is suitable for real-time operations, as very few prepossessing steps are needed to be able to work with it. Once aggregated at a chosen spatiotemporal level, substantial information can be extracted from it (De Nailly et al., 2022), and the resulting time series are the ground for many disruption detection techniques. More generally, we will refer to **anomaly detection**, being a broader term that includes all forms of unexpected or unwanted event detection.

2.2. Anomaly detection models for public transport systems

Anomaly detection techniques refer to identifying inconsistencies in a set of data. They are used in several fields, such as water distribution and treatment (Chen et al., 2022) medical (Wang et al., 2020), human activity (Sagha et al., 2013), energy consumption (Oprea et al., 2021), or cyber-security (Illiano et al., 2018; Shahinzadeh et al., 2021). Transportation literature also refers to these methods, but there is a gap between their use in road traffic (Weil et al., 1998; Talpade et al., 1999; Jiang and Papavassiliou, 2006; Davis et al., 2020) and PT. Most recent studies in road traffic focus on computer vision methods (Rathee et al., 2023) or multisource data, including image and multivariate times-series captured by sensors (Driss Laanaoui et al., 2024). These studies also use various kinds of machine learning models, such as tree-based, Recurrent Neural Networks (RNN) for capturing time structures or Convolutional Neuronal Networks (CNN) when working with image data. In the following, we gather the main contributions made in the field of PT using time series.

Working on the PT network of Paris, Tonnelier et al. (2018) have used AFC data for 13 consecutive weeks. They have proposed a first model based on the periodicity of the time series which aims to provide an anomaly score based on the distance to the periodic signal. A second model uses non-negative matrix decomposition to extract the main component of the signal, allowing it to consider more complex patterns than the periodicity defined in the first model. Results show that non-negative matrix decomposition is best at minimizing the false alarm rate. Using machine learning and deep learning models such as Random Forest (RF) and Long Short-Term Memory (LSTM), Pasini et al. (2022) have detected anomalies from 50 different subway stations in Montreal, from 2015 to 2017. They are notably based on the forecasting ability of such models to detect anomalies and provide contextual anomaly scores. Noteworthy, the authors show that LSTM outperforms RF in the prediction task, while RF is better than LSTM at detecting anomalies.

Interestingly, Tonnelier et al. (2018) have also developed a continuous user-based model based on the selection of the 10% least likely validation logs for every cardholder. Once these logs are selected, new time series are compiled, and the detection task is performed. The underpinning assumption of this approach is that anomalous events may lead a large number of users to have unusual behaviors, increasing the number of low-probability validations to be observed. Findings show that the model is effective for detecting fine-grained changes in time series.

Other work described time-series anomalies from a statistical point of view. Using AFC data collected on the Shanghai subway system, Gu et al. (2020) reconstructed a 5 min interval time series and clustered it at the day level. Resulting clusters correspond to calendar attributes: weekdays, weekends, and holidays, and for each of them, the authors retrieve interquartile ranges (IQR) related to the time series. This information is used to calculate indexes that evaluate different types of anomalies. The authors proposed a 3-level detection system based on four different indexes, which helps trigger different disruption management strategies. Briand et al. (2017) used a similar approach, using hierarchical clustering to group days and measure the difference in the number of validations between the observed and nominal data every 15 min. Using a simpler approach, Sun et al. (2016) assumed that ridership levels were normally distributed, and applied the thrice standard-error principle to identify anomalies in demand levels.

Data on real anomalies are always hard to obtain and the level of confidence that one can have in such data is often controversial. One of the main objectives of anomaly detection is, therefore, to reduce the uncertainty related the limited knowledge we have about anomalies. Two different approaches are identified in the literature: unsupervised settings and supervised settings with label pre-processing.

On one hand, most anomaly detection approaches found in the literature are unsupervised. Indeed, most studies acknowledge that raw disruption databases are not trustworthy enough to be used as labels in supervised settings. To overcome this issue, Ji et al. (2018) used the subway delay disruption dataset from Washington Metropolitan Area Transit Authority as labels in their multi-task learning model, using data extracted from social media as features to detect anomalies. The study provides positive insights into using social media data to detect disruptions. In addition, Zhang et al. (2022) used a Gaussian Mixture Model (GMM) to detect service disruptions in metro headways and used service disruption logs to perform a post-hoc comparison of their results. Other works used unsupervised settings with a post-hoc evaluation, such as Tonnelier et al. (2018) used Twitter data, Pasini et al. (2022) used service disruption logs, or Jasperse (2020) used delays.

On the other hand, supervised settings require labeled datasets. However, the well-known problem of low quality labels has led academics to develop other learning processes such as soft-labeling learning (Nguyen et al., 2014; de Vries and Thierens, 2024) or confidence learning (Northcutt et al., 2019). For instance, in the case of manual implementation of disruption logs, these techniques could be based on the operator's confidence level to label an event as a disruption. It could result in *soft-labels*, where the output of the procedures is not binary, but offers a range of possibilities from high confidence in the presence of disruptions or conversely, high confidence in its absence. Another approach is to focus on a subset of the most trustworthy labels, for which the supervised detection task can be performed. For instance, only delays above a certain threshold can be considered as an anomaly (Marra and Corman, 2019), and therefore, can be used as labels.

In this work, we are interested in minimizing the uncertainty related to SD logs which releases some of the constraints for executing a supervised detection algorithm using this data. The following section aims to explore what models and settings are used in the literature to handle anomaly detection problems, and also investigate data processing techniques strengthening the confidence in labels.

2.3. Supervised settings for imbalanced learning

As explained by Primartha and Tama (2017), anomaly detection can be summarized as a binary classification problem, where one class stands for the presence of anomaly and the other stands for its absence. As anomalies are seldom, their corresponding class is largely underrepresented in the dataset: this problem is known as **imbalanced classification**. This raises two main challenges: first, as few anomalies can be used to train the classifier, high false alarm rates are often observed (Primartha and Tama, 2017). Second, the usual performance metrics used for classification problems are not adapted to class imbalance. For example, the percentage of correct predictions (i.e. the accuracy) is expected to be very high because the model will easily predict the outcome for the majority class, regardless of the outcome for the minority class. Considering both statements, the modeling framework needs to be improved to optimize its performance, and suitable metrics need to be found for anomaly detection.

To tackle the first objective, several authors have shown that an ensemble classifier performs better than a single classifier (Woźniak et al., 2014; Fernández et al., 2018b) with methods such as bagging (Breiman, 1996) or boosting (Schapire, 1990). The intuition behind ensemble classifiers is to combine several classifiers into a single classifier, therefore reducing the prediction's variance and/or bias. In addition, in the case of bagging, each bootstrap sample can be drawn so that the sample contains the same proportion of objects belonging to the majority and minority classes (Chen et al., 2004): this method is known as balanced bagging.

As mentioned in Fernández et al. (2018b), ensemble classifiers are often embedded with data pre-processing techniques in the case of imbalanced learning. Most techniques rely on data sampling "in which the training instances are modified in such a way as to produce a more balanced class distribution, that allows classifiers to perform in a similar manner to standard classification" (Fernández et al., 2018a). Oversampling consists of creating or replicating instances from the minority class, while undersampling consists of removing existing instances from the majority class. When data is highly skewed, undersampling is not recommended as it may delete too many data points to reach the desired class ratio. Among oversampling methods, the Synthetic Minority Oversampling Technique (SMOTE) is very popular due to its ability to minimize overfitting, unlike other simpler sampling techniques (Chawla et al., 2002; Elor and Averbuch-Elor, 2022).

To address the second objective, we will consider two types of output the model can provide: nominal class and numerical scoring predictions (Fernández et al., 2018c). In binary classification settings, nominal class means that the model's task is to detect the presence or absence of an anomaly. Dealing with nominal class prediction, the F1-score is the most used metric. F1-score is the harmonic mean of precision (i.e. the number of relevant items among the ones retrieved by the model) and recall (i.e. the number of retrieved items among the relevant ones). In addition, some improved F measures on class imbalance can be found in the literature (Nguyen, 2019). Numerical scoring is the assignment of a score to each output. In some cases, this score can be interpreted as the probability of belonging to the target class (i.e. the anomaly). Several studies argued in favor of the use of the Precision–Recall (PR) curve in the case of imbalanced data (Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015; Branco et al., 2015), which uses continuous outcomes such as numerical scores. The Receiver-Operating Characteristic (ROC) curve is also used for imbalanced learning, but can be inaccurate for highly skewed datasets (Hanczar et al., 2010).

3. Methods and data

General problem specifications are given before introducing methodological materials. This work aims to find a model h predicting the probability that a disruption – also called anomaly – occurs at a given time t and a given subway station. Let us denote y the binary variable taking the value 1 in case of disruption, 0 otherwise. In addition, let us denote $\mathbf{X}$ the matrix of the K variables used to explain the occurrence of disruptions until time T.

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & \cdots & x_{t,1} & \cdots & x_{T,1} \\ \vdots & \ddots & & & \vdots \\ x_{1,k} & & x_{t,k} & & x_{T,k} \\ \vdots & & & \ddots & \vdots \\ x_{1,K} & \cdots & x_{t,K} & \cdots & x_{T,K} \end{pmatrix}, \quad t \in [1, T] \quad \text{and} \quad k \in [1, K] \tag{1}$$

From matrix $\mathbf{X}$, we can retrieve different information that will be used in further steps of the modeling framework.

- The values of a given explanatory variable can be extracted column-wise. The resulting vector is denoted $\mathbf{X_k} = (x_{1,k}, \ldots, x_{T,k})$ and will be referred to as **variable** in the following.
- The values of all explanatory variables at a given time step t can be extracted row-wise. The resulting vector is denoted $\mathbf{X_t} = (x_{t,1}, \ldots, x_{t,K})$, and will be referred to as **instance** in the following.
- A single value $x_{t,k}$ taken from matrix $\mathbf{X}$ is called a **realization** of variable k at time t.

Section 3.1 focuses on explanatory variables ($\mathbf{X}$) and Section 3.2 focuses on the model specifications (h). Section 3.3 deals with measuring the contribution of alternative stops to the resilience of PT systems. Finally, the data used in this study is introduced in Section 3.4

3.1. Variables embedding

In this study, we distinguish temporal from spatial features.

Temporal variables are categorical calendar variables. They are used to capture the seasonality observed in PT demand, which might influence the occurrence of disruptions. Intra-day variability is taken into account by affecting an ID to each of the 288 five-minute time bins included in a single day. Similarly, week variability is taken into account using days of the week (from 1 to 7) and annual variability is implemented using a boolean variable for the presence (1) or the absence (0) of holidays.

Spatial variables correspond to the time series of least probable validation logs at stops near the studied subway station, also called Spatial User-Based Metric (SUBM) in this work. The intuition is that when a disruption occurs at a subway station, many users will validate at a nearby stop, where they are usually unlikely to validate (this is a property of the flow reallocation process). Therefore, SUBM is expected to increase in the surroundings of a disrupted subway station. These stops are called **alternative stops** in the following, as they are considered the preferred alternative solution in case of subway disruptions.

Specifically, each validation is characterized by the triplet (u, s, t), where u denotes the user ID, s the PT stop and t the time. Let us denote $V_{u',s'}$ the set of validations made by a given user u' at a specific stop s', and $V_{u'}$ the set of all validations made by the same user u'. For each validation, we calculate the probability that user u' tap-in at stop s' according to Eq. (1) (Tonnelier et al., 2018) :

$$\mathbb{P}_{u'}(s') = \frac{\text{Card}(V_{u',s'})}{\text{Card}(V_{u'})} \tag{2}$$

Then, the 10% most unlikely logs (i.e. having the lowest probability) are selected and aggregated into a 5 min time series. The resulting time series is then filtered to avoid noise-related problems. Although Tonnelier et al. (2018) used a Gaussian Kernel, this filter is irrelevant for real-time purposes because calculating the filtered value involves considering validations after the observed time t. To mimic a real-time data inflow, we set the value of the convolution cell to 0 after time t. This adapted kernel results in a filter called *Half-Gaussian* in this work.

Spatial variables selection - For each subway station considered in this study, the set of alternative stops is selected according to the following two steps. First, a distance threshold is used to select the closest stops. We choose a 600-meter buffer, which is used by Egu and Bonnel (2020) to delimit the walkable distance for the transfer between the PT stops in Lyon. Second, the remaining alternative stops are selected according to one supply-based and one demand-based indicator. The chosen supply-based indicator is the number of days with at least one validation recorded. It notably helps to separate regular supply from event-specific supply, such as bridging buses. Indeed, adding information on bridging buses in the model would be considered as a **data leakage**, because they are specifically and exclusively implemented to overcome disruptions. Then, the total number of validations at each alternative stop is used as a demand-based indicator. It aims to distinguish alternative stops with sufficient data available from **uninformative stops**. To automatically select the relevant stops, all couples of indicators are clustered using spectral clustering (Von Luxburg, 2007). The number of clusters is set to two, and the one that maximizes both indicators is selected.

3.2. Model specifications

Model - In this work, we use RF as a classifier for several reasons. First, we use RF due to its ability to cope with imbalanced classification problems (Breiman, 2001). Gradient Boosting algorithms have the same properties regarding imbalanced datasets and have also been tested, giving similar results but having higher computing needs during the tuning stage. For this reason, we only present RF in this work. For this application, we divide our dataset into training data (80%) and testing data (20%). In the training stage, a 5-fold cross-validation procedure is applied to validate the relevance of our model and avoid overfitting.

Second, RF and tree-based models in general, are machine learning models which are interpretable. Indeed, global explanations can be retrieved from RF given how often and how well a given variable explanatory variable will contribute to building new branches in the decision trees. When combined with SHAP, we are also able to calculate exact local contribution of explanatory variables (see Section 3.3).

Finally, RF are structured in such a way that terminal nodes in each decision tree will contain a fraction of anomalous instances, that can be interpreted as the probability for these instances to be anomalous. This is in line with the numerical scoring approach commonly used for anomaly detection in imbalanced settings (see Section 2.3).

Outputs - To give an anomaly score to each sample, we use the RF probability score as defined in Olson and Wyner (2018). Considering instance $\mathbf{X_t}$ in a single tree θ and falling into the terminal node τ^*, the probability for $\mathbf{X_t}$ to be predicted as an anomaly can be defined as the fraction $f_1(\theta, \mathbf{X_t})$ of instances being labeled as an anomaly ($y = 1$) in τ^*. In the case of RF, N_θ trees are used to make the prediction such as the ensemble of trees can be written as $\{\theta_i\}_{i\in[1,N_\theta]}$. Consequently, the probability for the instance $\mathbf{X_t}$ to be an anomaly is defined as the average of the probabilities for single trees, as given in Eq. (3).

$$\mathbb{P}_{\mathbf{X_t}}(y = 1) = \frac{1}{N_\theta} \sum_{i=1}^{N_\theta} f_1(\theta_i, \mathbf{X_t}) \tag{3}$$

Performance - Precision (Pr) and Recall (Re) are two common metrics used to assess the performance of classification models. They can be calculated when the model's output $\hat{y}$ is binary. In this case, we can define the number of true positives (TP, $y = 1$, $\hat{y} = 1$), false positives (FP, $y = 0$, $\hat{y} = 1$), true negatives (TN, $y = 0$, $\hat{y} = 0$) and false negatives (FN, $y = 1$, $\hat{y} = 0$), and precision and recall can be calculated as in Eq. (4).

$$Pr = \frac{TP}{TP + FP} \qquad and \qquad Re = \frac{TP}{TP + FN} \tag{4}$$

Hard labels (i.e. binary outputs) are retrieved using a probability threshold p that optimizes a given performance measure. The Precision–Recall curve is produced by varying the probability threshold so that precision and recall can be calculated for each step. This curve is particularly useful in imbalanced datasets where one class is much more frequent than the other, as it provides a clearer picture of the model's performance across different threshold levels. The aggregated metric used at the end of this process is the Area Under the Curve (AUC), which is best approximated using the Average Precision (AP) score as shown in Eq. (5) (Manning et al., 2008).

$$AP = \sum_p (Re_p - Re_{p-1}) Pr_p \tag{5}$$

The AP score summarizes the Precision–Recall curve by taking the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight. This approach gives a single-figure measure of quality across the range of threshold values, making it a robust metric for evaluating model performance.

Specificity due to imbalance - The particularity of imbalanced settings is that the fraction α of instances for which $y = 1$, also called contamination rate in anomaly detection, is very low. To overcome this problem, we implement the Synthetic Minority Oversampling Technique (SMOTE) algorithm to create new instances and reach a desired ratio α_S, which aims to improve the learning capacity of the model (Elor and Averbuch-Elor, 2022).

The notion of Balanced Random Forest (BFR) has been developed in Chen et al. (2004) to deal with imbalanced settings. It differs from RF in that the bootstrap sample used for each tree in the RF is balanced. More precisely, each sample is drawn from the disrupted instances ($y = 1$), and the same number of samples is drawn with replacement from the regular instances ($y = 0$). Results of this procedure are given in Appendix A

Parameters tuning - Each tree in RF is trained on a random sample drawn from the whole dataset. In the case of bootstrap sampling, samples are drawn with replacement, which is proven to reduce the variance of the model (Breiman, 2001). Another common regularization is restricting the algorithmic feature selection to a randomly selected subset of explanatory variables for each node. For a set of K variables, a random subsample of size $\sqrt{K}$ is often used (Breiman, 2001).

The literature often studies two other parameters, namely the number of trees in the RF and the tree's depth. Most research suggests that trees in RF should be unpruned, i.e. that trees should be built to full depth (Breiman, 2001). However, there is no clear consensus on the number of trees to use in RF. One of the most recommended techniques to find the optimal parameter is to perform a sensitivity analysis using out-of-bag errors (Hastie et al., 2009). In this study, we will perform a grid search analysis using average precision as the performance metric, in a 5-fold cross-validation setting. The tested number of trees lies from 10 to 500 and the corresponding tuning curves are given in Appendix A.

3.3. SHapley Additive exPlanation (SHAP) to measure the contribution of explanatory variables to the resilience of public transport systems

Definition - Shapley Additive Explanation (SHAP) is a method inspired by game theory to explain the output of any machine learning or deep learning model (Lundberg and Lee, 2017). SHAP algorithm helps to unveil each explanatory variable's contribution to the constitution of the predicted output, which most AI models are unable to explain. RF models inherently have this ability to assess the contribution of explanatory variables: importance scores can be assessed thanks to the measure of impurity that is used to split nodes (Breiman, 2001; Menze et al., 2009). However, importance scores do not give information on the direction of the variable's effect and are inefficient when trying to explain the interaction between different variables. For this reason, we choose to compute SHAP values instead of using the classic RF importance scores. In the transportation field, SHAP has been used for traffic applications, especially to explore line-changing behaviors (Ali et al., 2022; Sun et al., 2024), and also to a lesser extent for PT applications (Lee, 2022; Xi et al., 2024).

Mathematical formulation - The SHAP theory is based on game theory, where "a prediction can be viewed as a coalitional game by considering each [variable] value of an instance as a player in a game" (Molnar, 2023). Let us denote $S \subseteq X$ the coalition of variables considered, and $C = X \backslash S$ the remaining set of variables. For a given model h and time step t, the SHAP value function v can be written as in Eq. (6).

$$v_{h,t}(S) = \int h(x_{t,S} \cup X_C) d\mathbb{P}_{X_C} - \mathbb{E}[h(X)] \tag{6}$$

The term $\int h(x_{t,S} \cup X_C) d\mathbb{P}_{X_C}$ means that for the selected coalition of variables S, we consider the realization of corresponding random variables at time t, denoted $x_{t,S}$. Then, we integrate over all possible values of the remaining variables X_C according to their distribution in the dataset $\mathbb{P}_{X_C}$, to retrieve the contribution of variables in S to the prediction using model h. Finally, the average prediction $\mathbb{E}(h(X))$ is subtracted from the integral term. In our work, $\mathbb{E}(h(X)) \simeq 0$ because disruption occurs rarely (i.e. the class imbalance is high).

Marginal contribution- The marginal contribution of variable X_k to coalition S can be written as $v_{h,t}(S \cup \{X_k\}) - v_{h,t}(S)$. The final SHAP value $\phi_{t,k}$ of variable X_k at time t is the average marginal contribution of a realization $x_{t,k}$ to all possible coalition of variables (Molnar, 2023), as in Eq. (7).

$$\phi_{t,k} = \sum_{S \subseteq X \backslash \{X_k\}} \frac{|S|!(K - |S| - 1)!}{K!} \left(v_{h,t}(S \cup \{X_k\}) - v_{h,t}(S) \right) \tag{7}$$

Importance score - Similar to the RF importance score, SHAP importance scores can be computed for each variable by summing absolute values of SHAP values of the kth variable over the studied period ($t \in [1, T]$), as shown in Eq. (8).

$$\Phi_k = \sum_{t=1}^{T} |\phi_{t,k}| \tag{8}$$

As SHAP is a very computationally demanding process, we decided not to perform the SHAP algorithm on the full dataset. The fraction of data used in this process is a subsample of the testing data set, which is chosen according to the following conditions. First, as the data is very imbalanced, all disruption instances are included in the sample. Second, the time distribution of the sample must follow the time distribution of the testing data set. In doing so, the size of the data to be analyzed is reduced by 12 for the most demanding stations while the accuracy of feature importance is maintained. This study uses the Tree SHAP algorithm to compute SHAP values.

Interaction - For two variables X_k and X_l where $(k, l) \in [1, K]^2, l \neq k$, the SHAP interaction value can be calculated as in Eq. (10) (Lundberg et al., 2020).

$$\phi_{t,k,l} = \sum_{S \subseteq X \backslash \{X_k, X_l\}} \frac{|S|!(K - |S| - 2)!}{2(K - 1)!} . \nabla_{k,l}(h, t, S) \tag{9}$$

where

$$\nabla_{k,l}(h, t, S) = \left[[v_{h,t}(S \cup \{X_k, X_l\}) - v_{h,t}(S \cup \{X_l\})] - [v_{h,t}(S \cup \{X_k\}) - v_{h,t}(S)] \right] \tag{10}$$

In this work, we are particularly interested in the interaction between time and space variables. Stops have different timetables depending on the time of the year, hence we believe the interaction between SUBM measured at each stop and time varies depending on the stop considered.

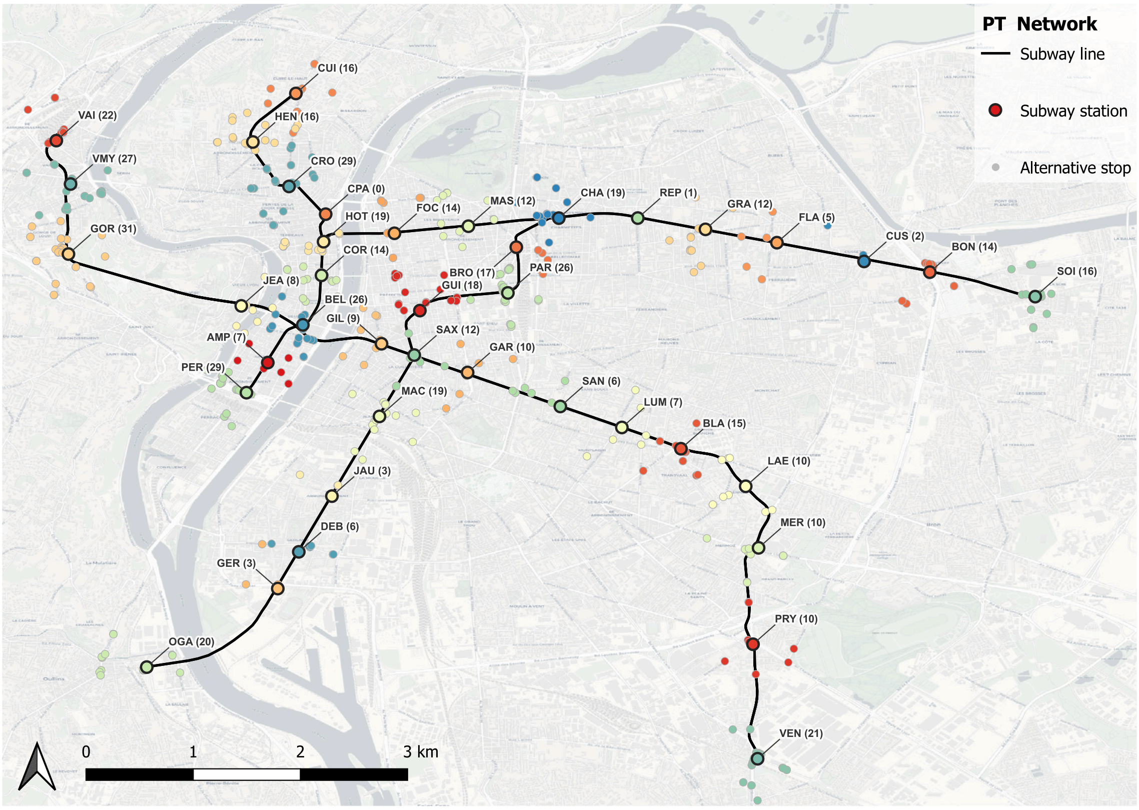


Fig. 1. Public Transport network of Lyon, France - *The stops shown in this map are restricted to the one used in the study. The numbers in parentheses indicate the number of alternative stops associated with the corresponding subway station, and colors are used to distinguish the service area of each subway station.* Map tiles: OpenStreetMap.

3.4. Data

The PT network of Lyon is composed of 4 subway lines with a total of 40 stations. Two additional stations have been put into service on October 20th 2023, but are not considered in this study which focuses on 2021, 2022, and 2023. More than 2 million trips are made each day, 75% of them are made using smart-card and almost 50% of them involve the subway. In addition to subway stations, the study focuses on alternative stops, which are all other PT stops included in a 600-meter buffer around every subway station. Other PT stop includes tramway, Bus Rapid Transit (BRT) and bus. The selection procedure for these stops is described in Section 3.1 and Fig. 1 shows the number of alternative stops finally selected for each station.

We use two data sources in this work: Automatic Fare Collection (AFC) and Service Disruption logs (SD logs).

Automatic Fare Collection (AFC) data are a good proxy for PT demand. These data are comprehensive and very granular, allowing fine spatio-temporal analysis of mobility behaviors. They can be used for longitudinal analysis, which is crucial in our approach to calculate the SUBM, and to target unlikely validations. Access to the PT system in Lyon requires a ticket or a smart card validation (also called fare transaction). These detailed validations collected by the operator are available for 2021, 2022, and 2023. In this paper, validations are used to calculate SUBM as defined in Section 3.1.

Service Disruptions logs (SD-logs) provided by the operator contain 3 years of data: 2021, 2022, and 2023. These dataset are usually very noisy and require several pre-processing steps to be used in a supervised setting. Once cleaned, a spatiotemporal match with AFC data can performed, allowing to measure the impact of service disruptions on PT demand (Cottreau et al., 2025). From this procedure, we are able to distinguish disruption from minor disturbances, which considerably strengthens the confidence in the labels for the detection task. In this work, we focus on the subset of disruptions that have the most intense impact on demand (3,916 observations). When distributing the disruption temporally over the 3 years of the studied period, we end up with disruption rates between 1‰and 6‰and lower (< 1‰) for stations CUI, HEN and CRO. Station CPA has not experienced any disruptions during the study period. SMOTE (Chawla et al., 2002) is performed to balance these low disruption rates and improve the detection algorithm, as mentioned in Section 3.2. Four different desired ratios are tested: $\alpha_s = \{0, 0.05, 0.10, 0.20\}$.

4. Results and discussions

4.1. The detection of public transport disruptions

Fig. 2 shows the results of the 39 models that have been run to detect disruptions at the subway station level. The tuning stage oriented our choice on a RF over a BRF. Also, the values of parameters have been set to $\alpha_s = 0$ for SMOTE (i.e. SMOTE has not been implemented). RF inherent parameters are found through grid search analysis, the results of which are introduced in Appendix A. The minimum number of elements a terminal node can contain is fixed to 1, the minimum number of elements required to split a node is set to 2 and the number of trees is set to 200 for all stations. The efforts to correct imbalance (i.e. SMOTE and BRF) did not significantly affect the model's performance, which is consistent with Yap and Cats (2021).

Model's performance - Average Precision (AP) scores given in Fig. 2(a) indicate that 35 stations have AP scores superior to 0.79, followed by CUS (0.69), LAE (0.60) and REP (0.00). Station REP is associated with only one alternative stop and the poor model performance suggests that this stop does not play a major role in the reallocation process. Therefore, the corresponding variable is not informative and the model cannot detect disruptions. To a lesser extent, CUS faces the same issue with only two alternative stops in its scope. They bring useful information for the detection process, but the lack of additional variables leads to model underperformance. Station LAE experiences very few disruptions (250 min), which also leads to underperformance. Consequently, the performance of the detection model is primarily driven by the number of alternative stops and the number of disruptions that a station experiences.

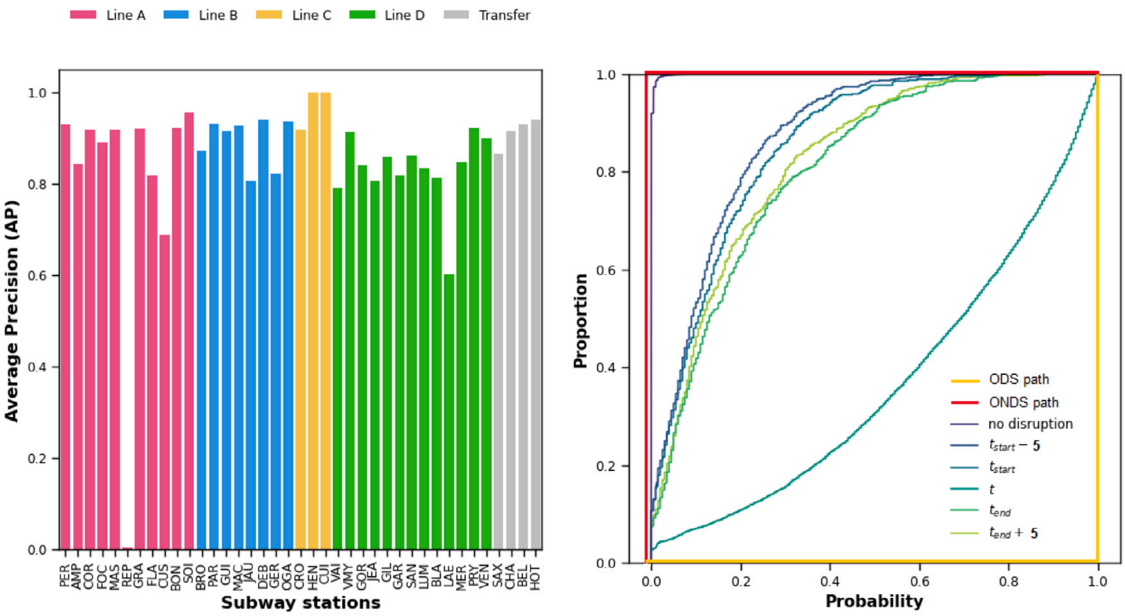
Given the data imbalance, the AP score of a random classifier is expected to be close to zero. From this perspective, the ability of RF to detect disruption is good. The strength of our modeling approach resides in the consideration of the SUBM. The Spatial User-Based Metric (SUBM) is a probabilistic approach that helps to target unlikely fare transactions that are supposedly more frequent in case of disruption (Tonnelier et al., 2018). SUBM helps distinguish an increase in demand, which is not necessarily correlated with a disruption, from an increase in abnormal behaviors that is more likely to be caused by a disruption. What is more, the redundancy brought by the consideration of SUBM for several stops simultaneously enhances the explanatory power of the model. Noteworthy, this process avoids spending too much time defining a reference state, based on the most observed behaviors (Jasperse, 2020). In addition, the proposed methodology is grounded on the use of a supervised classification problem, which is not used in the literature because SD-logs are deemed unreliable, or are simply unavailable (Pasini et al., 2022; Tonnelier et al., 2018; Jasperse, 2020). Here, we select the set of most intense disruptions to perform the detection task, which allows us to build a robust supervised modeling framework.

Time-wise probability - Fig. 2(b) shows the cumulative probability that instances are classified as a disruption 5 min before it starts ($t_{start} - 5$) when it starts (t_{start}), during a disruption (t), when it ends (t_{end}) and 5 min after it ends ($t_{end} + 5$). If 100% of instances in a given time frame have 0% chance of being a disruption according to the model, then the curve will rapidly reach a plateau, starting from probability = 0 and reaching proportion =1. This case is considered as the *optimal **no-disruption** scenario (ONDS)*. On the other hand, if 100% of instances in a given time frame have 100% chances of being a disruption, then the curve will follow the x-axis, until it reaches probability = 1 for which proportion = 1. This case is considered as the *optimal **disruption** scenario* (ODS).

Results first show that the probability of detecting a disruption when there is actually no disruption is very close to zero at any time, and is nearly identical to *ONDS*. It means that the proposed model is good at minimizing false alarms. Then, the cumulative distribution of probabilities during disruptions (t) shows that 70% of these instances have probabilities higher than 0.5 to be considered a disruption. This curve is closer to ODS, even if the model rarely attribute a 100% probability of being a disruption. For the same probability threshold (0.5), less than 10% of instances at t_{end} and $t_{end} + 5$ are concerned, less than 5% for t_{start} and t_{start}+5, and 0% for the remaining non-disruption instances. Therefore, the four corresponding curves that are displayed in the center of Fig. 2(b), which we would expect to be closer to the *ODNS*, show a less stereotyped behavior. We can conclude that there are clues in the data to target a disruption before its registered beginning in SD-logs. This means that an alarm can be raised when the flow reallocation process has started but the operator did not start to implement management strategies. This is a useful insight in favor of the implementation of early and automatic detection models, using different levels of detection such as in Gu et al. (2020). In addition, the cumulative distributions for t_{end} and $t_{end} + 5$, show that the effect of the disruption remains after its ending time as registered in SD-logs. Again, such probabilities are useful to extend disruption management plans when needed, by providing information or additional supply to PT users until full recovery of the PT system. This result is in line with Malandri et al. (2018), who found that the effect of a disruption that involves a station closure can last up to 6 times its corresponding closing time.

4.2. The contribution of alternatives stops to public transport resilience

In this study, we used the number of 10% least likely validations (i.e. the SUBM metric) at alternative stops to determine the probability that a disruption occurs at a given time and a subway station. Consequently, we estimate the contribution of an alternative stop to the system's resilience by its ability to absorb unusual flows during disruptions, which is ultimately measured by SHAP importance scores. However, optimal disruption management should also consider flow redistribution among alternatives. For instance, let us consider a single alternative stop having a high SHAP importance score among 10 stops in total. This stop will certainly contribute to the system's resilience but if the 9 left are low contributors, PT operators can also expect overcrowding, denied boarding, or even user dropout. This is why we consider the most resilient station should be defined by high but also evenly distributed importance scores. For this reason, we consider both the raw values of SHAP importance scores (**absolute contribution**)



(a) Average Precision (AP) scores

(b) Cumulative probability curves

Fig. 2. Model performance - ***(a)** Stations are classified by line. One bar corresponds to one RF model applied to one station. - **(b)** shows the cumulative probability of instances being classified as a disruption 5 min before it starts ($t_{start} - 5$) when it starts (t_{start}), during a disruption (t), when it ends (t_{end}), 5 min after it ends ($t_{end} + 5$), and when there is no disruption. In addition, illustrative curves are introduced for understanding purpose. The red curve refers to the optimal non-disruption scenario (ODNS), that is, a model predicting no-disruptions instances without false alarms. The yellow curve refers to optimal disruption scenario (ODS), that is, a model predicting disruption instances with 100% accuracy.* The test data are used to build these graphs.

and their relative importance in the set of alternatives linked to a single subway station (**relative contribution**). At the alternative stop level, we consider the raw importance score (*stop absolute contribution*) and its share in the total importance scores associated with the corresponding subway station (*stop relative contribution*), as shown in Fig. 3. At the subway station level, we consider the sum (*aggregated absolute contribution*) and the standard deviation (*aggregated relative contribution*) of the importance scores of the corresponding alternative stops, as shown in Fig. 4.

We empirically clustered the stations into 5 groups corresponding to 5 different types of resilience. An additional group comprises *uninformative* stations for which the lack of data on the disruption makes it difficult to conclude. In the following, we use the data on the stop contribution of Fig. 3 in the form of *stop relative contribution* (expressed in %). In addition, we use *aggregated absolute* (sum) *and relative* (std) *contributions* as mentioned in Fig. 4.

Low resilience: these stations are characterized by low absolute (sum< 100) and low relative (std< 5) contribution. They also experience fewer disruptions than others (from 5 to 107), which explains their low absolute contributions. For the subset of stations that experience less than 30 disruptions (CRO, CUI, HEN, LAE), the standard deviation of SHAP importance scores is below 1, meaning that the contribution of each alternative stop to the system's resilience is deemed equal. The level of information from these stations is too low to conclude their resilience abilities: we call them *uninformative* stations. For other stops (BRO, GER, GOR, JEA, PRY, SAX, VAI, VEN), one or several alternative stops dominate others regarding contribution to the system's resilience.

High resilience: these stations are characterized by high absolute contribution (sum> 350) and experience more disruptions than others (from 291 to 337). This group gathers the city hall of Lyon (HOT) and the two national train stations of Lyon (PAR and PER). They have low standard deviations regarding the high importance scores attributed to their alternative stops. Having more stop alternatives, PER (std=9) and PAR (std=8) are more resilient than HOT (std=13) in the sense they have more capacity to absorb the shock spatially. The reallocation process is more constrained at HOT, where lines C3 and C14 stand out as popular alternative options, while others (e.g. C18) are neglected.

Critical resilience: this group gathers high relative contribution (std> 30) stations such as CUS and JAU. They have a low number of alternative stops (respectively 2 and 3) with one of them standing out as the most reliable alternative solution in case of disruptions. For example, line C7 represents 57% of the total importance scores attributed to JAU, and line C17 represents 73% of the total importance scores attributed to CUS.

Hub resilience: *hubs* stations are characterized by low relative contribution (std< 6.5). These are stops capable of absorbing the demand due to a high number of alternatives on average (18). In this group, we can notably distinguish terminal hubs (BON, OGA, SOI, VMY), which are designed for modal shift, from inner-city hubs (AMP, BEL, BLA, CHA, COR, GUI, MAC), which benefit from the

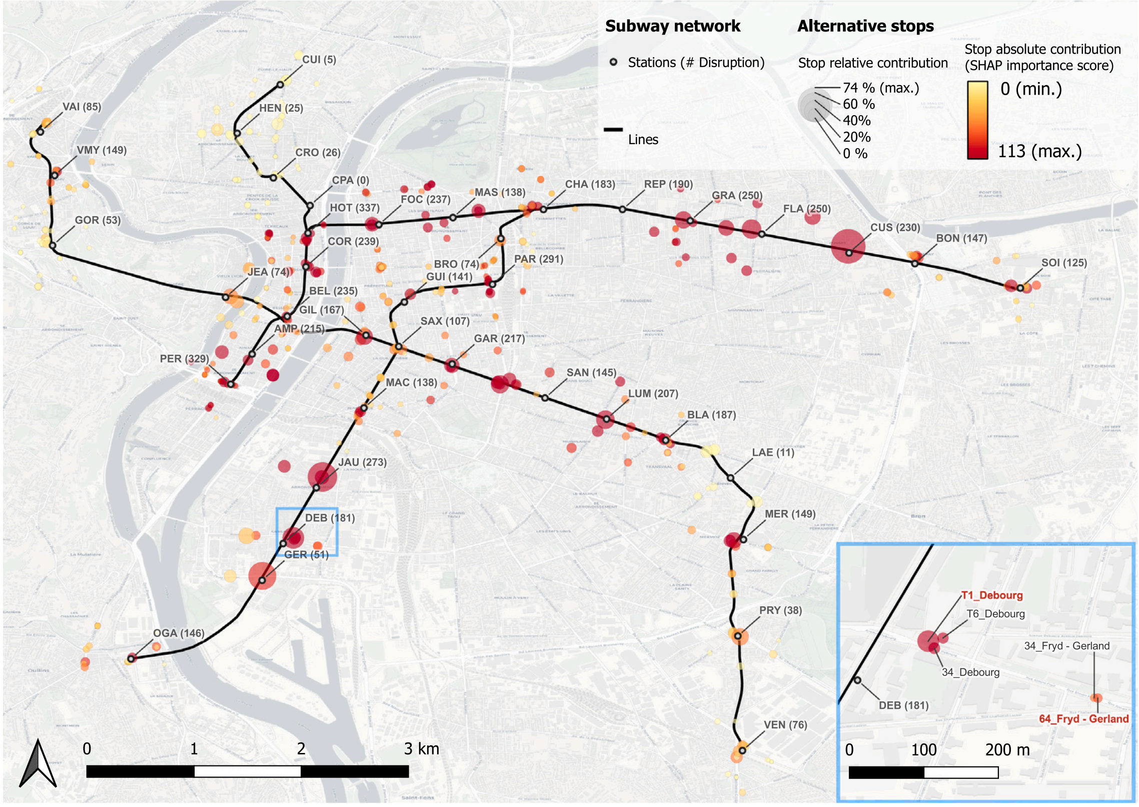


Fig. 3. Contribution of alternative stop to PT resilience - *The stop absolute contribution is given by raw SHAP importance scores and displayed using a color scale from yellow (0) to red (113). The relative contribution is the percentage of SHAP importance score of an alternative stop relative to the sum of all SHAP importance scores of a set of alternative stops associated with a given subway station. It is displayed using the point size on the map. Map tiles: Open Street Map.*

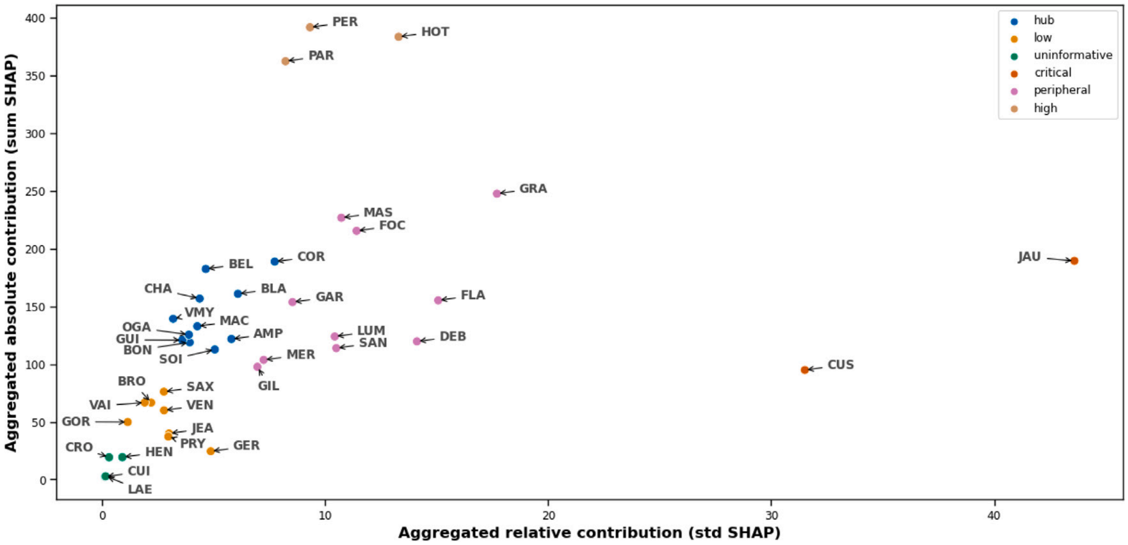


Fig. 4. Empirical clusters of resilience - *Subway station are displayed in this graph according to their aggregated absolute (sum of SHAP values of the corresponding alternative stops) and relative (standard deviation of SHAP values of the corresponding alternative stops) contributions. Five clusters result from this analysis, with an additional group of uninformative stops for which* $std < 1$ *and* $sum < 30$.

high density of PT network in central areas. On a larger scale, PAR, PER and HOT have similar properties and could be considered as *hubs* stations.

Peripheral resilience: *peripheral* stations are characterized by high standard deviations (std> 6.5) and two times fewer alternatives than hubs on average (9). One or several stops dominate others in terms of SHAP importance score (i.e. at least one alternative stop contributes from 20% to 40% to the aggregated absolute resilience). The resilience of these stations (DEB, FLA, FOC, GAR, GIL, GRA, LUM, MAS, MER, SAN) relies on a few alternatives, which can be risky in the case of disruption propagation and increase the chance of overcrowding issues at the corresponding alternative stops. Their properties are close to single alternative stations such as CUS and JAU (*critical resilience*), or *low resilience* stations even though they face fewer disruptions.

In addition, Fig. 4 helps to identify stations that are on the edge of a cluster. For instance, stations DEB, FLA, and GRA are *peripheral resilient*, but their aggregated relative contribution makes them closer to the *critical* cluster. Indeed, Fig. 3 shows that these stations have one alternative station that plays a higher role in their resilience, namely *T1_Debourg* for DEB (40%), *69_Flachet Alain Gilles* for FLA (35%) and *C26_Gratte Ciel Metro*for GRA (29%). The same conclusion can be drawn for GER, which is *low resilient*, but also close to *critical* with stop *60_Stade de Gerland* accounting for 56% of the total SHAP importance score at this station.[1]

Furthermore, the absolute contributions of alternative stops show that tramway stops (mean SHAP = 13) contribute significantly more than BRT (mean SHAP = 11) or bus stops (mean SHAP = 7) to the resilience of PT systems. However, when tramways are not available, BRT contribution increases (mean SHAP = 13), and when neither tramways or BRT are available, bus contribution increases in return (mean SHAP = 9). This indicates that the mode of alternative stops is a key factor in the design of PT resilience.

The aggregated analysis shown in Fig. 4 can be compared with the results from Yap and Cats (2021), who analyze the criticality of subway stations according to their disruptions' exposure and impact. For instance, Yap et al. (2018) have shown that the cluster of most critical stations comprises subway-to-subway and subway-to-train transfer stations. In addition, our study demonstrates that stations having comparable attributes (HOT, PAR, and PER) also have the best resilience properties. Therefore, they are more exposed to disruptions but are more able to handle the flow redistribution than other stations. Noteworthy, Yap and Cats (2021) also note that the line-specific context can play a role in criticality, which can explain the few disruptions experienced by line C (CPA, CRO, HEN and CUI) in our study. More generally, criticality and resilience analyses are complementary and could be performed together to fully appraise the endurance of a station against disruptions.

The stop-level analysis helps to understand which stop solely contributes to the resilience of its corresponding station, or if the stop is included in a set of alternatives that also contribute to the station's resilience. For instance, bus-bridging strategies should target the *critical resilience* cluster which involves alternative stops that are more likely to experience overcrowding. Noteworthy, Ge et al. (2022) have noticed that bus bridging is one of the hottest topics in the literature when dealing with disruptions management plans, but collaboration with taxi companies, and more recently micro-mobility could also be solutions to explore. Also, development plans should consider the vulnerability of stations belonging to the *critical resilience* cluster, and propose more alternative solutions to improve their resilience. From this perspective, we are in line with the strategy that Cats (2016) has observed in the case of the Stockholm development plan, which prioritized the densification of the PT network over its expansion. The *high resilience* and *hub resilience* clusters share the property of density, and also connectivity to the rest of the network, especially for inner-city hubs. These results are consistent with Zhang et al. (2015), for who compactness and connectivity are two essential attributes of PT network resilience.

4.3. Interactions

In this section, we will focus on the example of subway station Debourg (DEB) to analyze the interaction between time and space features. This station was chosen because it clearly illustrates the results of a major tramway line and a small bus line with interesting patterns that can be observed elsewhere in the network. Fig. 5 shows the SHAP values distribution for each variable, sorted by importance. Color is used to display the original value of a feature. SHAP for time, SUBM, and interaction between them for stop *T1_Debourg* and SHAP interaction for stop *64_Fryd-Gerland* are shown in Fig. 6. The location of the corresponding alternative stops is indicated in the insert of Fig. 3.

Fig. 5 displays SHAP values for all explanatory variables. *T1_Debourg* has the highest importance score, while *Time (summer holidays)* has the lowest. Regarding spatial variables (SUBM), we identify clear patterns for *T1_Debourg*, *34_Debourg*, and *T6_Debourg*, for which variable values seem positively correlated with SHAP values. For *64_Fryd-Gerland*, *34_ENS Lyon*, and *34_Fryd-Gerland*, patterns are fuzzier. With respect to time variables, the 5 min time level has the highest explanatory power by far. On the contrary, *Holidays* and *Summer Holidays* have almost no explanatory power. Fig. 6 aims to show a clearer visualization of the time (5 min) variable, in interaction with two different contributors: *T1_Debourg* and *64_Fryd-Gerland*.

Fig. 6(a) shows the SHAP values for time, with time on the x-axis and values of SUBM for line T1 on the y-axis. Fig. 6(b) shows the SHAP values for SUBM at the stop *T1_Debourg*, with time on the x-axis and values of SUBM for line T1 on the y-axis. The first insight that comes out of both graphs is that SUBM follows a stereotyped distribution over time, with a peak during evening hours (15:00–21:00). Also, disrupted data points mainly correspond to the highest values of SUBM for *T1_Debourg*, at each given time.

On one hand, Fig. 6(b) indicates that SHAP values are substantially high (0.30–0.40) when the SUBM is also high (1.50–3.50), notably during the evening peak hours. However, the SHAP values for *T1_Debourg* are lower during morning hours (6:00–12:00). The SUBM values for *T1_Debourg* and disrupted data points during the morning hours are in the range of SUMB values observed for

[1] An interactive map similar to Fig. 3 is available online and on demand.

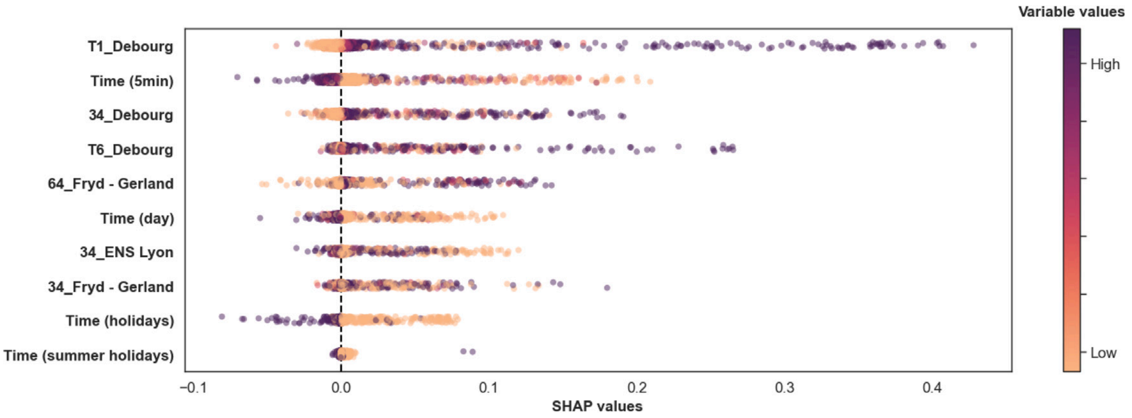
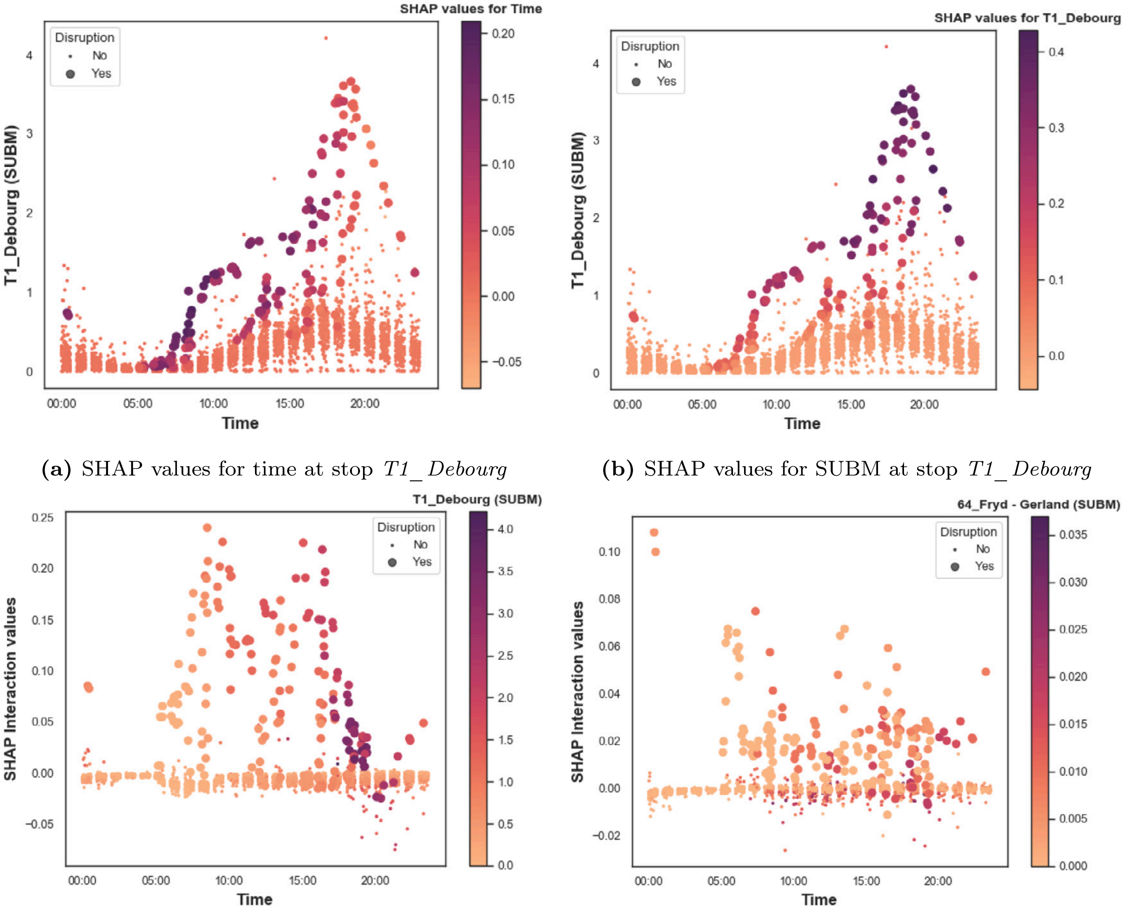


Fig. 5. Global SHAP values distribution for station DEB - *The SHAP values distribution is shown for each variable, sorted by importance (T1_Debourg having the highest importance score). Color is used to display the original value of a feature.*



(a) SHAP values for time at stop *T1_ Debourg*

(b) SHAP values for SUBM at stop *T1_ Debourg*

(c) SHAP interaction values for SUBM and time at stop *T1_ Debourg*

(d) SHAP interaction values for SUBM and time at stop *64_ Fryd - Gerland*

Fig. 6. Examples of SHAP and SHAP interaction values - *T1_Debourg and 64_Fryd - Gerland are two alternative stops associated with subway station DEB. T1_Debourg shows a strong interaction between time and SUBM while 64_Fryd - Gerland does not..*

regular data points during evening hours (0.50–1.50). Therefore, it is clear that the variable *T1_Debourg* on its own will contribute to detecting disruptions during evening hours. On the other hand, Fig. 6(a) indicates that SHAP values for time are higher in the morning (0.15–0.20) than in the evening (0.00–0.15) when looking at disrupted data points. Hence, the time variable seems to compensate for the weakness of variable *T1_Debourg* and help to detect disruption happening in the morning, i.e. there is a high interaction between time and *T1_Debourg*.

Fig. 6(c) displays the SHAP interaction values on the *y*-axis, and time on the *x*-axis. Values of variable *T1_Debourg* are indicated through point coloration. This figure shows that the interaction is high during morning peak hours, reaching its highest point (0.24) around 9:00. Interaction is also high during evening peak hours, reaching its maximum (0.22) at 15:00. For both peaks observed, it means that a high amount of low-probability validations is more likely to indicate the presence of a disruption. However, a negative interaction is observed between 19:00 and 22:00. These data points also correspond to a high amount of low-probability validations, and are regular data points. This phenomenon seems typical of post-work activities (e.g. leisure or purchase), which occur mainly at this time of the day (Egu and Bonnel, 2020) and are less predictable in space than other trip purposes such as commuting.

Fig. 6(d) shows SHAP interaction values similar to Fig. 6(c) for alternative stop *64_Fryd-Gerland*. The pattern that comes out of this graph is not stereotyped. The highest values of SUBM for this stop correspond to interaction values of 0. The low frequency of line 64 (one bus every 30 to 60 min) can explain this lack of interaction with time. Also, line 64 has a very local coverage which can explain its low contribution to the resilience of the PT system, being used by a specific population and accounting for few trips.

The results displayed for Debourg station can be generalized to any subway station. The strength of the model lies in its ability to deal with different kinds of information coming from different alternatives stops. Event or activity-specific bus lines provide useful information to make the difference between low probability validation related to a disruption and low probability validation related to other kinds of expected events. Commuting lines, that have more stereotyped patterns, allow to identify the general probability distribution of validations (e.g. high-probability validations during peak hours). These lines play the role of baseline features, as they feed the model with common mobility patterns. They usually have the highest SHAP importance scores and mostly correspond to strongest modes such as tramways or BRT.

The effect of time of the day on disruptions is known, and existing works have shown that the most intense disruptions mostly occur during peak hours (Yap and Cats, 2022). The majority of disruption detection models use different representations of the time of the day as an explanatory variable (Pasini et al., 2022), or use time to operate separated analyses of specific periods (Jasperse, 2020). This work goes one step further by analyzing the interaction between time and space variables, thanks to SHAP algorithm which is a major recent development in the field of explainable IA. To the best of our knowledge, this study is the first to implement SHAP in combination with disruption detection in the field of transportation research. This shows how machine learning models can interpret different kinds of unexpected behaviors according to time, and therefore make the difference between low-probability fare transactions due to disruption and post-work activities. Further work could include other calendar variables, such as expected events (e.g. football games), that regularly impact specific subway stations and are expected to be associated with low values of SUBM as well.

5. Conclusions

Using a sample of 3 years of subway Automatic Fare Collection (AFC) and Service Disruptions (SD) logs provided by the PT operator of Lyon, this study aims to answer the following questions: *To what extent can data-driven methods be implemented to detect PT disruptions? How can we measure the contribution of existing alternative options to the resilience of PT systems?*

In this work, the detection task is assimilated to a supervised classification problem, performed using RF for 39 subway stations of the PT system of Lyon, France. The explanatory variables are twofold: time variables are based on calendar attributes (time of the day, day of the week, and holidays), while space variables are embedded in a Spatial User-Based Metric (SUBM). The SUBM represents the time series of the 10% least likely fare transactions, which value is expected to reflect the flow reallocation process in case of disruptions.

Also, SHalpey Addiditive exPlaination (SHAP) is implemented to retrieve the importance of explanatory variables and assess the magnitude and the direction of their interaction. We are particularly interested in assessing the contribution of space variables to PT resilience thanks to SHAP importance scores, and the interaction between time and space variables thanks to SHAP interaction values.

First, the detection model is trained to compute the probability that a disruption occurs at each 5 min interval time step. This model has high prediction performance, with Average Precision (AP) scores over 0.79 for 35 stations. The analysis of the probabilities shows that we can detect disruptions (70% of the disruption data points have probabilities over 0.5), but also that there are some clues in the data that inform on flow reallocation processes emerging 5 min before and lasting 5 min after the registered period of disruptions in SD-logs. At the same time, the model is good to minimize the false alarm rate. These are relevant insights for PT operators, that could automatize and enhance their disruption detection procedure with different confidence intervals based on different probability thresholds.

Second, the contribution of alternative stops measured with SHAP values helps to target where the flow reallocation process happens and therefore can be used to improve disruption management plans by implementing temporary services where it is the most needed, or orienting the flow towards better existing alternatives using high-quality information. Noteworthy, five clusters of station resilience emerge from this work, for which connectivity and density seem to play a role.

Finally, this work focuses on the interaction between time and space and shows how RF can make the difference between low-probability fare transactions caused by disruptions and those caused by evening activities. This result indicates the relevance of the SHAP algorithm in better interpreting machine learning (ML) models for the disruption detection task. With additional time explanatory variables, such as planned events (e.g. football games) that are expected to cause low-probability fare transactions as well, the detection algorithm could be improved in specific contexts. Indeed, SHAP interaction algorithm is a powerful tool to understand overlapping events using ML.

This study faces several limitations. Unlike works using network theory, the notion of alternative stops does not consider Origin–Destination trips. One development of the proposed model could be to take into account two directions at least for each stop, to have a finer approach in the recommendation for service reinforcement for instance, and answer questions like: *Is it better to implement bridging buses towards the city center or terminal stops of a subway line?*. In addition, if PT operators would like to implement such a model for real-time operations, some adjustments would be needed to switch from an *off-line* to *on-line* detection model. Data should be considered time-wise, and using lagged variables or deep learning models such as Long-Short Term Memory (LSTM) might be more appropriate to build a proper *on-line* detection model. Moreover, since deep learning models are also proven efficient for short-term forecasting (Pasini et al., 2022; Wang et al., 2023), further works could investigate the relevance of integrated models, executing both detection and forecasting tasks. In addition, comparing SHAP values between different models, having different distributions of binary target variables can lead to serious issues when interpreting the data (e.g. the issue of uninformative stops). We assume in this study that the data is representative of the distribution of disruptions because we use a large dataset (i.e. 3 years of 5 min time series for 39 subway stations and more than 500 alternative stops), but this assumption needs to be permanently verified and actualized in the view of an operation use of this model. For instance, although SMOTE does not improve the detection performance, it could be used to reach similar levels of contamination rate for each station and therefore have more comparable results. Finally, the short-term approach introduced in this study could be complemented with a longitudinal analysis of individual long-term adaptation. Being repeatedly exposed to disruption, in the context of infrastructure projects for instance, could durably change PT users' behaviors. Moreover, the use of multimodal data could greatly improve the understanding of these changes over time, as shown by Shateri Benam et al. (2025).

The detection framework proposed in this work is fully replicable for PT network having tap-in validation systems and access to SD-logs. Particular attention should be paid to the selection of alternative stops through the choice of the distance threshold. The procedure defined in Appendix B is also replicable to determine the distance threshold in any PT network. Further works could investigate the opportunity to have a variable distance threshold, which might more adapted to different urban contexts, having different local stop density for instance.

CRediT authorship contribution statement

Benjamin Cottreau: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mehmet Güney Celbiş:** Writing – review & editing, Validation, Methodology, Formal analysis, Conceptualization. **Ouassim Manout:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Conceptualization. **Louafi Bouzouina:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition.

Acknowledgments

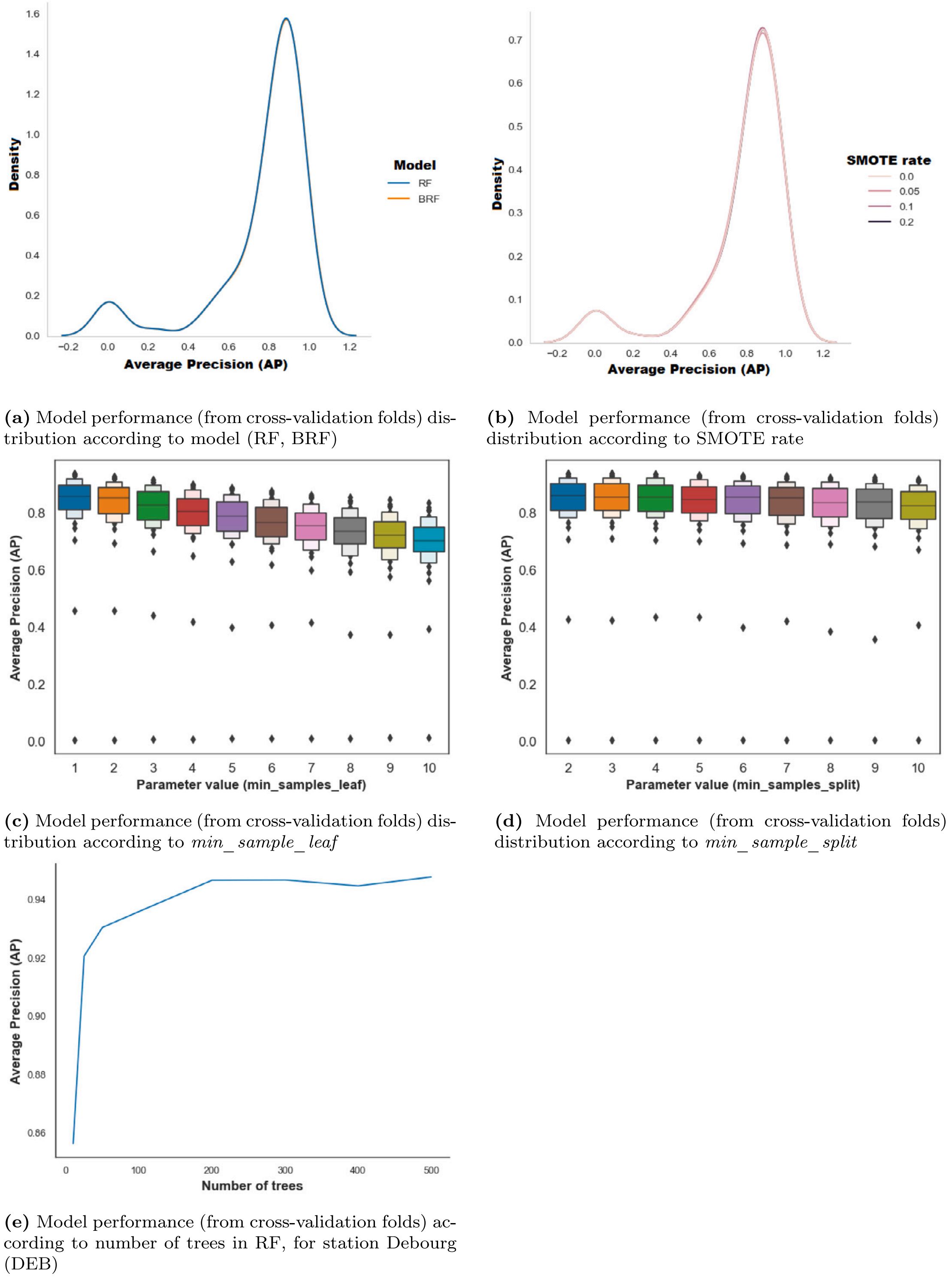
This paper is part of the "Mobility in the Big Data Era" chair. The authors are thankful to KEOLIS for providing the data on PT use in Lyon.

Appendix A. Parameters tuning

For each station, a grid search analysis has been performed, taking five parameters into account, namely the type of model (RF, BRF), the SMOTE rate (0.00, 0.05, 0.10, 0.20), the minimum number of elements that a terminal node – also called a leaf – is allowed to contain (*min_samples_leaf*), the minimum number of elements that a node is required to have to split it (*min_samples_split*), and the number of trees in the RF. The grid search is performed using a five-fold cross-validation procedure on the train set. The resulting model performances have been stored and displayed in Fig. 7.

Figs. 7(a) and 7(b) show that attempts to rectify data imbalance did not significantly change the model performance, as curves overlap for each modality of the parameters considered. Figs. 7(c) and 7(d) show that the best performance is reached for lowest values of *min_sample_split* (=2) and *min_sample_leaf* (=1). The number of trees was set to 200, as no significant gain in performance was observed after this threshold. Fig. 7(e) shows a use case for station Debourg (DEB).

Finally, for computational efficiency, the following parameters have been chosen: model = RF, SMOTE rate = 0.00, and number of trees = 200.



(a) Model performance (from cross-validation folds) distribution according to model (RF, BRF)

(b) Model performance (from cross-validation folds) distribution according to SMOTE rate

(c) Model performance (from cross-validation folds) distribution according to *min_ sample_ leaf*

(d) Model performance (from cross-validation folds) distribution according to *min_ sample_ split*

(e) Model performance (from cross-validation folds) according to number of trees in RF, for station Debourg (DEB)

Fig. 7. Parameters tuning - *The efforts to correct imbalance (i.e. SMOTE* ***[a]*** *and BRF* ***[b]****) did not significantly affect the model's performance. According to best performance observed, the minimum sample leaf was fixed to 1, minimum samples split was fixed to 2, number of trees was fixed to 200* ***[e]****.*

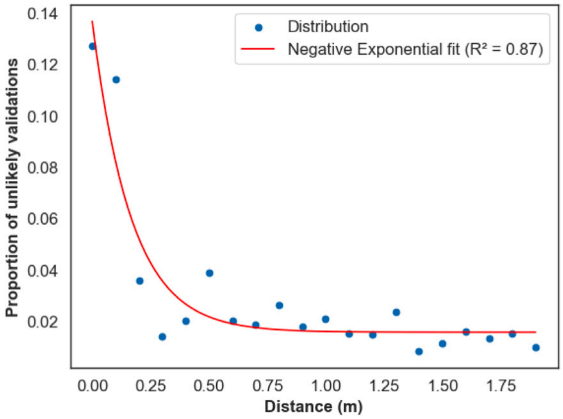


Fig. 8. Sensitivity of the proportion of least-likely validation to distance.

Appendix B. Sensitivity to distance

Considering the 520 alternative stops of the study, the distribution of unlikely validations according to the distance to the closest disrupted subway station is shown with blue dots in Fig. 8. This distribution is approximated by a negative exponential function, we can be described as follows :

$$f(x) = ae^{-bx} + c \tag{11}$$

The best regression is associated with the following coefficients: $a = 0.12$, $b = 6.00$, and $c = 0.015$. The corresponding R-squared value is $R^2 = 0.87$.

In accordance with the value found in the literature (Egu and Bonnel, 2020), the buffer distance is set to 600 m, where the regression curve also starts to reach a plateau, meaning that the number of least likely validations does not increase after this distance threshold.

References

Ali, Y., Hussain, F., Bliemer, M.C., Zheng, Z., Haque, M.M., 2022. Predicting and explaining lane-changing behaviour using machine learning: A comparative study. Transp. Res. Part C: Emerg. Technol. 145, 103931. http://dx.doi.org/10.1016/j.trc.2022.103931, Retrieved 2024-11-27, from https://linkinghub.elsevier.com/retrieve/pii/S0968090X22003448.

Bešinović, N., 2020. Resilience in railway transport systems: a literature review and research agenda. Transp. Rev. 40 (4), 457–478. http://dx.doi.org/10.1080/01441647.2020.1728419, Retrieved 2024-04-22, from https://www.tandfonline.com/doi/full/10.1080/01441647.2020.1728419.

Branco, P., Torgo, L., Ribeiro, R., 2015. A survey of predictive modelling under imbalanced distributions. http://dx.doi.org/10.48550/ARXIV.1505.01658, Retrieved 2023-02-28, from https://arxiv.org/abs/1505.01658, (Publisher: arXiv Version Number: 2).

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140. http://dx.doi.org/10.1007/BF00058655, Retrieved 2024-10-01, from http://link.springer.com/10.1007/BF00058655.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. http://dx.doi.org/10.1023/A:1010933404324, Retrieved 2024-10-03, from http://link.springer.com/10.1023/A:1010933404324.

Briand, S., Come, E., Trépanier, M., Oukhellou, L., 2017. Smart card clustering to extract typical temporal passenger habits in transit network. Two case studies: Rennes in France and gatineau in Canada. In: 3rd International Workshop and Symposium: Research and Applications on the Use of Passive Data from Public Transport (TransitData).

Cats, O., 2016. The robustness value of public transport development plans. J. Transp. Geogr. 51, 236–246. http://dx.doi.org/10.1016/j.jtrangeo.2016.01.011, Retrieved 2024-08-30, from https://linkinghub.elsevier.com/retrieve/pii/S0966692316000120.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority over-sampling technique. J. Artificial Intelligence Res. 16, 321–357. http://dx.doi.org/10.1613/jair.953, Retrieved 2024-11-28, from https://www.jair.org/index.php/jair/article/view/10302.

Chen, Z., Chen, D., Zhang, X., Yuan, Z., Cheng, X., 2022. Learning graph structures with transformer for multivariate time-series anomaly detection in IoT. IEEE Internet Things J. 9 (12), 9179–9189. http://dx.doi.org/10.1109/JIOT.2021.3100509, Retrieved 2024-01-09, from https://ieeexplore.ieee.org/document/9497343/.

Chen, C., Liaw, A., Breiman, L., 2004. Using Random Forest to Learn Imbalanced Data. Technical Report No. 666, Department of Statistics, UC Berkley, Retrieved from http://xtf.lib.berkeley.edu/reports/SDTRWebData/accessPages/666.html.

Cottreau, B., Adraoui, A., Manout, O., Bouzouina, L., 2023. Spatio-temporal patterns of the impact of COVID-19 on public transit: An exploratory analysis from lyon, France. Reg. Sci. Policy Pr. 15 (8), 1702–1721. http://dx.doi.org/10.1111/rsp3.12718, Retrieved 2024-05-29, from https://rsaiconnect.onlinelibrary.wiley.com/doi/10.1111/rsp3.12718.

Cottreau, B., Manout, O., Bouzouina, L., 2025. Spatio-temporal impacts of unplanned service disruptions on public transit demand. Transp. Res. Interdiscip. Perspect. 30, 101354. http://dx.doi.org/10.1016/j.trip.2025.101354, Retrieved from https://www.sciencedirect.com/science/article/pii/S2590198225000338.

Cox, A., Prager, F., Rose, A., 2011. Transportation security and the role of resilience: A foundation for operational metrics. Transp. Policy 18 (2), 307–317. http://dx.doi.org/10.1016/j.tranpol.2010.09.004, Retrieved 2024-12-28, from https://linkinghub.elsevier.com/retrieve/pii/S0967070X10001137.

Davis, J., Goadrich, M., 2006. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. ICML '06, ACM Press, Pittsburgh, Pennsylvania, pp. 233–240. http://dx.doi.org/10.1145/1143844.1143874, Retrieved 2023-02-28, from http://portal.acm.org/citation.cfm?doid=1143844.1143874.

Davis, N., Raina, G., Jagannathan, K., 2020. A framework for end-to-end deep learning-based anomaly detection in transportation networks. Transp. Res. Interdiscip. Perspect. 5, 100112. http://dx.doi.org/10.1016/j.trip.2020.100112, Retrieved 2024-08-20, from https://linkinghub.elsevier.com/retrieve/pii/S2590198220300233.

De Nailly, P., Côme, E., Samé, A., Oukhellou, L., Ferriere, J., Merad-Boudia, Y., 2022. What can we learn from 9 years of ticketing data at a major transport hub? A structural time series decomposition. Transp. A: Transp. Sci. 18 (3), 1445–1469. http://dx.doi.org/10.1080/23249935.2021.1948626, Retrieved 2024-04-22, from https://www.tandfonline.com/doi/full/10.1080/23249935.2021.1948626.

de Vries, S., Thierens, D., 2024. Learning with confidence: Training better classifiers from soft labels. http://dx.doi.org/10.48550/ARXIV.2409.16071, Retrieved 2024-09-30, from https://arxiv.org/abs/2409.16071, Version Number: 1.

Driss Laanaoui, M., Lachgar, M., Mohamed, H., Hamid, H., Gracia Villar, S., Ashraf, I., 2024. Enhancing urban traffic management through real-time anomaly detection and load balancing. IEEE Access 12, 63683–63700. http://dx.doi.org/10.1109/ACCESS.2024.3393981, Retrieved 2025-04-13, from https://ieeexplore.ieee.org/document/10508783/.

Egu, O., Bonnel, P., 2020. How comparable are origin-destination matrices estimated from automatic fare collection, origin-destination surveys and household travel survey? An empirical investigation in lyon. Transp. Res. Part A: Policy Pr. 138, 267–282. http://dx.doi.org/10.1016/j.tra.2020.05.021, Retrieved 2024-11-12, from https://linkinghub.elsevier.com/retrieve/pii/S0965856420306030.

El Zein, A., Beziat, A., Pochet, P., Klein, O., Vincent, S., 2022. What drives the changes in public transport use in the context of the covid-19 pandemic? Highlights from lyon metropolitan area. Reg. Sci. Policy Pr. 14, 122–142. http://dx.doi.org/10.1111/rsp3.12519.

Elor, Y., Averbuch-Elor, H., 2022. To SMOTE, or not to SMOTE?. http://dx.doi.org/10.48550/ARXIV.2201.08528, Retrieved 2024-08-27, from https://arxiv.org/abs/2201.08528, Version Number: 3.

Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F., 2018a. Data level preprocessing methods. In: Learning from Imbalanced Data Sets. Springer International Publishing, Cham, pp. 79–121. http://dx.doi.org/10.1007/978-3-319-98074-4_5, Retrieved 2024-10-01, from http://link.springer.com/10.1007/978-3-319-98074-4_5.

Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F., 2018b. Ensemble learning. In: Learning from Imbalanced Data Sets. Springer International Publishing, Cham, pp. 147–196. http://dx.doi.org/10.1007/978-3-319-98074-4_7, Retrieved 2024-10-01, from http://link.springer.com/10.1007/978-3-319-98074-4_7.

Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F., 2018c. Performance measures. In: Learning from Imbalanced Data Sets. Springer International Publishing, Cham, pp. 47–61. http://dx.doi.org/10.1007/978-3-319-98074-4_3, Retrieved 2024-10-01, from http://link.springer.com/10.1007/978-3-319-98074-4_3.

Ge, L., Voß, S., Xie, L., 2022. Robustness and disturbances in public transport. Public Transp. 14 (1), 191–261. http://dx.doi.org/10.1007/s12469-022-00301-8, Retrieved 2024-10-01, from https://link.springer.com/10.1007/s12469-022-00301-8.

Gu, J., Jiang, Z., Fan, W.D., Wu, J., Chen, J., 2020. Real-time passenger flow anomaly detection considering typical time series clustered characteristics at metro stations. J. Transp. Eng. Part A: Syst. 146 (4), 04020015. http://dx.doi.org/10.1061/JTEPBS.0000333, Retrieved 2024-08-19, from https://ascelibrary.org/doi/10.1061/JTEPBS.0000333.

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., Dougherty, E.R., 2010. Small-sample precision of ROC-related estimates. Bioinformatics 26 (6), 822–830. http://dx.doi.org/10.1093/bioinformatics/btq037, Retrieved 2024-10-01, from https://academic.oup.com/bioinformatics/article/26/6/822/244957.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning. Springer Series in Statistics, Springer New York, New York, NY, http://dx.doi.org/10.1007/978-0-387-84858-7, Retrieved 2024-08-14, from http://link.springer.com/10.1007/978-0-387-84858-7.

Illiano, V.P., Paudice, A., Muñoz-González, L., Lupu, E.C., 2018. Determining resilience gains from anomaly detection for event integrity in wireless sensor networks. ACM Trans. Sens. Netw. 14 (1), 1–35. http://dx.doi.org/10.1145/3176621, Retrieved 2024-08-23, from https://dl.acm.org/doi/10.1145/3176621.

Jasperse, F., 2020. Automated Offline Detection of Disruptions Using Smart Card Data: A Case Study of the Metro Network of Washington DC. (Master's thesis). Delft University of Technology, Retrieved from https://resolver.tudelft.nl/uuid:251de9e9-5f83-45c8-a5b7-dc682c2102d7, Master's thesis, Faculty of Civil Engineering and Geosciences.

Ji, T., Fu, K., Self, N., Lu, C.-T., Ramakrishnan, N., 2018. Multi-task learning for transit service disruption detection. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ASONAM, pp. 634–641. http://dx.doi.org/10.1109/ASONAM.2018.8508799.

Jiang, J., Papavassiliou, S., 2006. Enhancing network traffic prediction and anomaly detection via statistical network traffic separation and combination strategies. Comput. Commun. 29 (10), 1627–1638. http://dx.doi.org/10.1016/j.comcom.2005.07.030, Retrieved 2024-08-26, from https://linkinghub.elsevier.com/retrieve/pii/S0140366405002744.

Khaled, A.A., Jin, M., Clarke, D.B., Hoque, M.A., 2015. Train design and routing optimization for evaluating criticality of freight railroad infrastructures. Transp. Res. Part B: Methodol. 71, 71–84. http://dx.doi.org/10.1016/j.trb.2014.10.002, Retrieved 2024-08-30, from https://linkinghub.elsevier.com/retrieve/pii/S0191261514001714.

Lee, E.H., 2022. Exploring transit use during COVID-19 based on XGB and SHAP using smart card data. J. Adv. Transp. 2022, 1–12. http://dx.doi.org/10.1155/2022/6458371, Retrieved 2024-11-27, from https://www.hindawi.com/journals/jat/2022/6458371/.

Liu, L., Porr, A., Miller, H.J., 2024. Measuring the impacts of disruptions on public transit accessibility and reliability. J. Transp. Geogr. 114, 103769. http://dx.doi.org/10.1016/j.jtrangeo.2023.103769, Retrieved 2024-12-28, from https://linkinghub.elsevier.com/retrieve/pii/S0966692323002417.

Louie, J., Shalaby, A., Habib, K.N., 2017. Modelling the impact of causal and non-causal factors on disruption duration for toronto's subway system: An exploratory investigation using hazard modelling. Accid. Anal. Prev. 98, 232–240. http://dx.doi.org/10.1016/j.aap.2016.10.008, Retrieved 2024-04-05, from https://linkinghub.elsevier.com/retrieve/pii/S0001457516303694.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2 (1), 56–67. http://dx.doi.org/10.1038/s42256-019-0138-9, Retrieved 2024-10-30, from https://www.nature.com/articles/s42256-019-0138-9.

Lundberg, S., Lee, S.-I., 2017. A unified approach to interpreting model predictions. http://dx.doi.org/10.48550/ARXIV.1705.07874, Retrieved 2024-10-21, from https://arxiv.org/abs/1705.07874, Version Number: 2.

Malandri, C., Fonzone, A., Cats, O., 2018. Recovery time and propagation effects of passenger transport disruptions. Phys. A 505, 7–17. http://dx.doi.org/10.1016/j.physa.2018.03.028, Retrieved 2023-05-31, from https://linkinghub.elsevier.com/retrieve/pii/S0378437118303480.

Manning, C.D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press, New York, OCLC: ocn190786122.

Marra, A.D., Corman, F., 2019. From delay to disruption: The impact of service degradation on public transport network. p. 8 p.. http://dx.doi.org/10.3929/ETHZ-B-000368811, Retrieved 2024-08-19, from http://hdl.handle.net/20.500.11850/368811, Artwork Size: 8 p. Medium: application/pdf Publisher: ETH Zurich.

Massobrio, R., Cats, O., 2024. Topological assessment of recoverability in public transport networks. Commun. Phys. 7 (1), 108. http://dx.doi.org/10.1038/s42005-024-01596-8, Retrieved 2024-08-30, from https://www.nature.com/articles/s42005-024-01596-8.

Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., Hamprecht, F.A., 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics 10 (1), 213. http://dx.doi.org/10.1186/1471-2105-10-213, Retrieved 2024-10-03, from https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-213.

Mo, B., Koutsopoulos, H.N., Shen, Z.-J.M., Zhao, J., 2023. Robust path recommendations during public transit disruptions under demand uncertainty. Transp. Res. Part B: Methodol. 169, 82–107. http://dx.doi.org/10.1016/j.trb.2023.02.004, Retrieved 2023-07-26, from https://linkinghub.elsevier.com/retrieve/pii/S0191261523000176.

Molnar, C., 2023. Interpreting Machine Learning Models with SAP: a Guide with Python Examples and Theory on Shapley Values, first ed. Chistoph Molnar c/o MUCBOOK, Heidi Seibold, München, Germany.

Müller, S.A., Leich, G., Nagel, K., 2020. The effect of unexpected disruptions and information times on public transport passengers: a simulation study. Procedia Comput. Sci. 170, 745–750. http://dx.doi.org/10.1016/j.procs.2020.03.161, Retrieved 2024-12-28, from https://linkinghub.elsevier.com/retrieve/pii/S1877050920306219.

Nguyen, M.H., 2019. Impacts of unbalanced test data on the evaluation of classification methods. Int. J. Adv. Comput. Sci. Appl. 10 (3), http://dx.doi.org/10.14569/IJACSA.2019.0100364, Retrieved 2024-10-01, from http://thesai.org/Publications/ViewPaper?Volume=10&Issue=3&Code=ijacsa&SerialNo=64.

Nguyen, Q., Valizadegan, H., Hauskrecht, M., 2014. Learning classification models with soft-label information. J. Am. Med. Inform. Assoc. 21 (3), 501–508. http://dx.doi.org/10.1136/amiajnl-2013-001964, Retrieved 2024-09-30, from https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2013-001964.

Northcutt, C.G., Jiang, L., Chuang, I.L., 2019. Confident learning: Estimating uncertainty in dataset labels. http://dx.doi.org/10.48550/ARXIV.1911.00068, Retrieved 2024-09-30, from https://arxiv.org/abs/1911.00068, Publisher: arXiv Version Number: 6.

Noursalehi, P., Koutsopoulos, H.N., Zhao, J., 2018. Real time transit demand prediction capturing station interactions and impact of special events. Transp. Res. Part C: Emerg. Technol. 97, 277–300. http://dx.doi.org/10.1016/j.trc.2018.10.023, Retrieved 2024-04-09, from https://linkinghub.elsevier.com/retrieve/pii/S0968090X18301797.

Olson, M.A., Wyner, A.J., 2018. Making sense of random forest probabilities: a kernel perspective. http://dx.doi.org/10.48550/ARXIV.1812.05792, https://arxiv.org/abs/1812.05792, Version Number: 1.

Oprea, S.-V., Bâra, A., Puican, F.C., Radu, I.C., 2021. Anomaly detection with machine learning algorithms and big data in electricity consumption. Sustainability 13 (19), 10963. http://dx.doi.org/10.3390/su131910963, Retrieved 2024-08-23, from https://www.mdpi.com/2071-1050/13/19/10963.

Pasini, K., Khouadjia, M., Samé, A., Trépanier, M., Oukhellou, L., 2022. Contextual anomaly detection on time series: a case study of metro ridership analysis. Neural Comput. Appl. 34 (2), 1483–1507. http://dx.doi.org/10.1007/s00521-021-06455-z, Retrieved 2024-01-09, from https://link.springer.com/10.1007/s00521-021-06455-z.

Primartha, R., Tama, B.A., 2017. Anomaly detection using random forest: A performance revisited. In: 2017 International Conference on Data and Software Engineering. ICoDSE, IEEE, pp. 1–6. http://dx.doi.org/10.1109/ICODSE.2017.8285847.

Rahimi, E., Shamshiripour, A., Shabanpour, R., Mohammadian, A., Auld, J., 2019. Analysis of transit users' waiting tolerance in response to unplanned service disruptions. Transp. Res. Part D: Transp. Environ. 77, 639–653. http://dx.doi.org/10.1016/j.trd.2019.10.011, Retrieved 2024-12-28, from https://linkinghub.elsevier.com/retrieve/pii/S1361920919303785.

Rathee, M., Bačić, B., Doborjeh, M., 2023. Automated road defect and anomaly detection for traffic safety: A systematic review. Sensors 23 (12), 5656. http://dx.doi.org/10.3390/s23125656, Retrieved 2025-04-13, from https://www.mdpi.com/1424-8220/23/12/5656.

Reggiani, A., Nijkamp, P., Lanzi, D., 2015. Transport resilience and vulnerability: The role of connectivity. Transp. Res. Part A: Policy Pr. 81, 4–15. http://dx.doi.org/10.1016/j.tra.2014.12.012, Retrieved 2024-12-28, from https://linkinghub.elsevier.com/retrieve/pii/S0965856414003048.

Rodríguez-Núñez, E., García-Palomares, J.C., 2014. Measuring the vulnerability of public transport networks. J. Transp. Geogr. 35, 50–63. http://dx.doi.org/10.1016/j.jtrangeo.2014.01.008, Retrieved 2024-06-28, from https://linkinghub.elsevier.com/retrieve/pii/S0966692314000180.

Sagha, H., Bayati, H., Millán, J.D.R., Chavarriaga, R., 2013. On-line anomaly detection and resilience in classifier ensembles. Pattern Recognit. Lett. 34 (15), 1916–1927. http://dx.doi.org/10.1016/j.patrec.2013.02.014, Retrieved 2024-08-23, from https://linkinghub.elsevier.com/retrieve/pii/S0167865513000585.

Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. In: Brock, G. (Ed.), PLOS ONE 10 (3), e0118432. http://dx.doi.org/10.1371/journal.pone.0118432, Retrieved 2024-08-02, from https://dx.plos.org/10.1371/journal.pone.0118432.

Schapire, R.E., 1990. The strength of weak learnability. Mach. Learn. 5 (2), 197–227. http://dx.doi.org/10.1007/BF00116037, Retrieved 2024-10-01, from http://link.springer.com/10.1007/BF00116037.

Shahinzadeh, H., Mahmoudi, A., Moradi, J., Nafisi, H., Kabalci, E., Benbouzid, M., 2021. Anomaly detection and resilience-oriented countermeasures against cyberattacks in smart grids. In: 2021 7th International Conference on Signal Processing and Intelligent Systems. ICSPIS, IEEE, Tehran, Iran, Islamic Republic of, pp. 1–7. http://dx.doi.org/10.1109/ICSPIS54653.2021.9729386, Retrieved 2024-08-23, from https://ieeexplore.ieee.org/document/9729386/.

Shateri Benam, A., Furno, A., El Faouzi, N.-E., 2025. Unraveling urban multi-modal travel patterns and anomalies: a data-driven approach. Urban Plan. Transp. Res. 13 (1), 2481962. http://dx.doi.org/10.1080/21650020.2025.2481962, Retrieved 2025-04-13, from https://www.tandfonline.com/doi/full/10.1080/21650020.2025.2481962.

Sun, H., Cheng, Q., Wang, P., Huang, Y., Liu, Z., 2024. Lane change decision prediction: an efficient BO-XGB modelling approach with SHAP analysis. Transp. A: Transp. Sci. 1–38. http://dx.doi.org/10.1080/23249935.2024.2372020, Retrieved 2024-11-27, from https://www.tandfonline.com/doi/full/10.1080/23249935.2024.2372020.

Sun, H., Wu, J., Wu, L., Yan, X., Gao, Z., 2016. Estimating the influence of common disruptions on urban rail transit networks. Transp. Res. Part A: Policy Pr. 94, 62–75. http://dx.doi.org/10.1016/j.tra.2016.09.006, Retrieved 2023-02-27, from https://linkinghub.elsevier.com/retrieve/pii/S0965856416303913.

Talpade, R., Kim, G., Khurana, S., 1999. NOMAD: traffic-based network monitoring framework for anomaly detection. In: Proceedings IEEE International Symposium on Computers and Communications. Cat. No.PR00250), IEEE Comput. Soc, Red Sea, Egypt, pp. 442–451. http://dx.doi.org/10.1109/ISCC.1999.780942, Retrieved 2024-08-20, from http://ieeexplore.ieee.org/document/780942/.

Tonnelier, E., Baskiotis, N., Guigue, V., Gallinari, P., 2018. Anomaly detection in smart card logs and distant evaluation with Twitter: a robust framework. Neurocomputing 298, 109–121. http://dx.doi.org/10.1016/j.neucom.2017.12.067, Retrieved 2023-02-28, from https://linkinghub.elsevier.com/retrieve/pii/S0925231218302170.

Vodopivec, N., Miller-Hooks, E., 2019. Transit system resilience: Quantifying the impacts of disruptions on diverse populations. Reliab. Eng. Syst. Saf. 191, 106561. http://dx.doi.org/10.1016/j.ress.2019.106561, Retrieved 2024-12-28, from https://linkinghub.elsevier.com/retrieve/pii/S0951832018309797.

Von Luxburg, U., 2007. A tutorial on spectral clustering. Stat. Comput. 17 (4), 395–416. http://dx.doi.org/10.1007/s11222-007-9033-z, Retrieved 2024-08-01, from http://link.springer.com/10.1007/s11222-007-9033-z.

Wang, W., Lee, J., Harrou, F., Sun, Y., 2020. Early detection of parkinson's disease using deep learning and machine learning. IEEE Access 8, 147635–147646. http://dx.doi.org/10.1109/ACCESS.2020.3016062, Retrieved 2024-01-09, from https://ieeexplore.ieee.org/document/9165732/.

Wang, B., Vu, H.L., Kim, I., Cai, C., 2023. Distributional prediction of short-term traffic using neural networks. Eng. Appl. Artif. Intell. 126, 107061. http://dx.doi.org/10.1016/j.engappai.2023.107061, Retrieved 2024-11-27, from https://linkinghub.elsevier.com/retrieve/pii/S0952197623012459.

Weil, R., Wootton, J., García-Ortiz, A., 1998. Traffic incident detection: Sensors and algorithms. Math. Comput. Modelling 27 (9–11), 257–291. http://dx.doi.org/10.1016/S0895-7177(98)00064-8, Retrieved 2024-08-20, from https://linkinghub.elsevier.com/retrieve/pii/S0895717798000648.

Weng, J., Zheng, Y., Yan, X., Meng, Q., 2014. Development of a subway operation incident delay model using accelerated failure time approaches. Accid. Anal. Prev. 73, 12–19. http://dx.doi.org/10.1016/j.aap.2014.07.029, Retrieved 2024-04-19, from https://linkinghub.elsevier.com/retrieve/pii/S0001457514002322.

Woźniak, M., Graña, M., Corchado, E., 2014. A survey of multiple classifier systems as hybrid systems. Inf. Fusion 16, 3–17. http://dx.doi.org/10.1016/j.inffus.2013.04.006, Retrieved 2024-10-01, from https://linkinghub.elsevier.com/retrieve/pii/S156625351300047X.

Xi, Y., Hou, Q., Duan, Y., Lei, K., Wu, Y., Cheng, Q., 2024. Exploring the spatiotemporal effects of the built environment on the nonlinear impacts of metro ridership: Evidence from Xi'an, China. ISPRS Int. J. Geo- Inf. 13 (3), 105. http://dx.doi.org/10.3390/ijgi13030105, Retrieved 2024-11-27, from https://www.mdpi.com/2220-9964/13/3/105.

Yap, M., Cats, O., 2021. Predicting disruptions and their passenger delay impacts for public transport stops. Transportation 48 (4), 1703–1731. http://dx.doi.org/10.1007/s11116-020-10109-9, Retrieved 2023-02-28, from https://link.springer.com/10.1007/s11116-020-10109-9.

Yap, M., Cats, O., 2022. Analysis and prediction of ridership impacts during planned public transport disruptions. J. Public Transp. 24, 100036. http://dx.doi.org/10.1016/j.jpubtr.2022.100036, Retrieved 2024-03-28, from https://linkinghub.elsevier.com/retrieve/pii/S1077291X22017362.

Yap, M., Nijënstein, S., van Oort, N., 2018. Improving predictions of public transport usage during disturbances based on smart card data. Transp. Policy 61, 84–95. http://dx.doi.org/10.1016/j.tranpol.2017.10.010, Retrieved 2023-02-28, from https://linkinghub.elsevier.com/retrieve/pii/S0967070X16307648.

Zhang, N., Graham, D.J., Bansal, P., Hörcher, D., 2022. Detecting metro service disruptions via large-scale vehicle location data. Transp. Res. Part C: Emerg. Technol. 144, 103880. http://dx.doi.org/10.1016/j.trc.2022.103880, Retrieved 2022-12-12, from https://linkinghub.elsevier.com/retrieve/pii/S0968090X22002935.

Zhang, X., Miller-Hooks, E., Denny, K., 2015. Assessing the role of network topology in transportation network resilience. J. Transp. Geogr. 46, 35–45. http://dx.doi.org/10.1016/j.jtrangeo.2015.05.006, Retrieved 2024-08-30, from https://linkinghub.elsevier.com/retrieve/pii/S0966692315000794.