A Causal Machine Learning Approach for Analyzing the Heterogeneous Treatment Effects of Weather Conditions on Arrival Delay Predictions in Public Rail Transportation

Elham Ahmadi*
elham.ahmadi@klu.org
Kuehne Logistics University
Hamburg, Hamburg, Germany

André Ludwig
ludwig@wifa.uni-leipzig.de
Leipzig University
Leipzig, Germany

Henrik Leopold
henrik.leopold@klu.org
Kuehne Logistics University
Hamburg, Germany

Abstract

Effective logistics disruption management is crucial to ensure reliable transportation. This study employs a causal machine learning methodology, namely Double Machine Learning, to assess the causal effect of extreme weather on train arrival delays, while controlling for significant confounders including passenger occupancy, temporal variables, and station attributes. Using a rich dataset of regional rail operations in central Germany, we estimate average and heterogeneous treatment effects of extreme weather events. The DML model identifies a statistically significant Average Treatment Effect, indicating that extreme weather leads to substantial delay increases, and reveals strong heterogeneity in treatment effects across lines, months, weekdays, and hours with Group Average Treatment Effects. Robustness assessments validate the consistency and dependability of the causal estimations. In addition, the use of passenger-related variables enhances the balance and overlap of the model, highlighting their importance in the modeling of delays. This study demonstrates that causal machine learning offers valuable information on the mechanisms behind train delay, providing a data-driven foundation for targeted interventions, adaptive scheduling, and improved resource allocation in rail transport systems.

CCS Concepts

• **Computing methodologies** → **Machine learning approaches**.

Keywords

Public Transportation Delays Weather Impact Passenger Occupancy Causal Effect Analysis Double Machine Learning

ACM Reference Format:
Elham Ahmadi, André Ludwig, and Henrik Leopold. 2025. A Causal Machine Learning Approach for Analyzing the Heterogeneous Treatment Effects of Weather Conditions on Arrival Delay Predictions in Public Rail Transportation. In *Conference on KDD (KDD '25), 2025, Toronto, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

KDD '25, Toronto, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Public rail transportation systems are increasingly vulnerable to delays from disruptions such as extreme weather, infrastructure constraints, system inefficiencies, and fluctuating passenger demand. These delays reduce service reliability, cause economic losses, lower customer satisfaction, and weaken network resilience [1, 24]. As climate change intensifies and urban populations increase, identifying and mitigating the causes of rail delays has become a critical priority for transport agencies and policymakers [6].

Traditionally, prediction of train delays has relied on statistical models or machine learning algorithms focused on pattern recognition and correlation-based forecasting [29, 35]. These typically estimate average effects, assuming homogeneous impacts across time, routes, or conditions. However, the influence of factors such as extreme weather or passenger load varies between contexts [17]. For example, snowfall may disrupt specific lines during peak hours but have little effect under lighter traffic. As a result, average-effect models risk obscuring important variation and limiting predictive usefulness.

To address these gaps, recent advances in causal inference, particularly the Neyman-Rubin Causal Model (RCM), provide tools to identify not just whether disruptions cause delays, but how much, for whom, and under which conditions [21, 22, 26]. Estimating heterogeneous treatment effects (HTE) enables a better understanding of causal mechanisms and more precise policies. Causal Machine Learning (CML) methods such as Double Machine Learning (DML) are well-suited for estimating these effects in high-dimensional, observational data [4, 7, 34].

Past transportation studies have applied causal inference to areas like flight delays and road safety, using methods such as Propensity Score Matching and Inverse Propensity Weighting [7, 19, 25, 33]. Yet, the complexity of urban rail systems and dynamic weather patterns demand models that can flexibly capture non-linear relationships and confounding bias. DML meets these needs by combining machine learning with robust causal inference, while Causal Forests allow for group-level or individualized causal estimation [4, 7, 34].

This study applies causal machine learning to assess the impact of extreme weather on train delays. Using DML, we aim to (1) estimate average and heterogeneous weather effects on delays and (2) explore how these effects vary across time, passenger load, and rail lines. Our contributions are:

- Developing a causal framework using DML to estimate Average Treatment Effects, accounting for high-dimensional confounding factors.

- Conduct robustness checks, including overlap diagnostics, exogeneity testing, and sensitivity analyzes.

- Analyzing a large-scale operational data set from regional rail services in central Germany, providing causal information to improve scheduling, resource allocation, and disruption management in adverse weather.

In summary, this study shows the value of causal machine learning in understanding delay mechanisms and supporting more resilient, adaptive rail transport planning.

The subsequent sections of this work are organized as follows: Section 2 reviews the literature on causal machine learning and its application to transportation delays. Section 3 introduces the methodological framework and provides the mathematical background underlying the DML. Section 4 describes the experimental setup, including the data sources, variable selection, and implementation details, followed by a presentation of the empirical results. Finally, Section 5 concludes the paper with a summary of insights and recommendations for future research.

2 Literature Review

Extreme weather events, increasingly frequent due to climate change, pose significant challenges to the reliability and efficiency of transportation networks. Disruptions caused by snowstorms, heavy rainfall, heatwaves, and strong winds can lead to infrastructure damage, reduced visibility, safety concerns, and significant schedule deviations [6, 13, 32]. These effects translate into both direct and indirect economic losses, affecting passengers, operators, and broader logistics systems.

In support of these observations, Gössling et al. [14] provide a comprehensive review of how phenomena such as heatwaves, storms, heavy precipitation, flooding, and wildfires are increasingly disrupting transport operations across all modes. Their study highlights that these disruptions affect both the supply side, through infrastructure degradation and service interruptions, and the demand side, by altering traveler behavior and reducing reliability perceptions. Moreover, they emphasize the cascading nature of such impacts, where delays in one transport mode can ripple across interconnected systems.

In the context of rail transportation, weather-induced delays have been studied across various areas. Chen and Wang [6] found that while both high-speed rail (HSR) and air travel are affected by extreme weather, HSR tends to be less vulnerable, particularly in regions with better infrastructure and adaptive operational strategies. Similarly, Zhang and Chen [36] demonstrated that weather events such as rain and snow significantly deteriorate travel time reliability and exacerbate congestion in urban traffic systems. While earlier studies emphasized the correlation between general weather conditions and train delays, more recent efforts have focused on specific phenomena and real-time mitigation. Focusing specifically on wind-related disruptions, Fu and Easton [12] developed a predictive model for the British railway network that identified wind gust speed, direction, temperature, and humidity as key determinants of incident probability. Building on this, Saeednia et al. [28] introduced a real-time integration framework that uses weather forecasts to dynamically manage freight trains' temporary speed restrictions (TSRs). Their approach, tested through the Shift2Rail platform, allows precise targeting of TSRs, minimizing unnecessary delays and improving communication with stakeholders through an API-driven forecast.

Aviation research also supports the significant role of weather as a disruption factor. Similar challenges have been observed in aviation. For instance, de Oliveira et al. [10] attributed delays in Brazilian domestic air travel to low ceiling, visibility, and wind gusts. Choi et al. [8] applied binary classification to predict weather-induced flight delays, while Sridhar and Chen [30] used autoregressive models with weather and traffic inputs to forecast airspace delays. These studies reinforce the need for proactive planning and data-driven strategies to enhance transport system resilience. More importantly, it reveals a critical research gap in quantifying the varied impacts of extreme weather across operational contexts, which this study addresses by applying causal machine learning methods to estimate heterogeneous treatment effects of weather disruptions on rail delays. Machine learning (ML) has become a widely used tool for predicting transportation delays, particularly due to its ability to capture complex, non-linear patterns in large datasets. For instance, Wang and peng Zhang [35] utilized ML algorithms to forecast train delays under various weather conditions, showing that extreme weather alters delay patterns significantly. Similarly, Sajan and Kumar [29] applied regression models incorporating temperature, rainfall, and wind to estimate rail delays in India. Beyond passenger transport, ML has also been applied to optimize last-mile delivery, train scheduling, and resource allocation by analyzing historical travel time data [2, 9, 20].

While effective at pattern recognition, traditional ML methods fall short when it comes to providing interpretable results or uncovering causal relationships, both of which are essential for decision-making and intervention design. This has led to growing interest in causal machine learning (CML), which focuses on identifying cause-and-effect relationships rather than mere correlations. As noted by Arti et al. [3], CML supports more transparent and actionable insights, enabling transportation agencies to test "what-if" scenarios and make proactive adjustments, such as targeted scheduling changes or resource deployments in anticipation of extreme weather.

Huenermund et al. [16] distinguishes between predictive and causal approaches, placing regression, time series forecasting, and tree-based methods in the former, and identifying causal forests [34], generalized random forests [4], DML (DML) [7], and modified causal forests (MCF) [38] as key causal tools. These methods estimate heterogeneous treatment effects (HTEs) while accounting for high-dimensional confounding, enabling more nuanced policy recommendations.

CML has shown promise in fields like supply chain risk management [5], defense logistics [37], and collaborative transport planning [18]. Within the transportation domain, Truong [33] applied structural causal models to analyze air traffic delays, while Srivastava et al. [31] investigated causal delay factors in rail systems. Li et al. [19] further demonstrated the value of doubly robust causal ML in estimating the effects of highway incidents.

Despite these advances, causal ML remains underutilized in rail and road transportation, where interactions among infrastructure, environmental conditions, and passenger behavior are inherently

complex. Broader adoption of causal inference methods could uncover the true drivers of delays and inform more effective, data-driven interventions, particularly under increasing risks associated with climate change. This study contributes to filling this gap by applying causal machine learning to evaluate how extreme weather affects rail delays in a heterogeneous manner.

3 Methodology

This section provides an overview of the methodological framework adopted to estimate the causal effect of extreme weather on train arrival delays, as illustrated in Figure 1.

The subsequent sections delineate the fundamental elements of the methodology. Causal assumptions are initially introduced as the basis for impartial causal inference. The DML framework is delineated, employing machine learning methodologies to estimate the effects of treatment on output. This foundation enables a comprehensive and robust causal analysis.

3.1 Assumptions

We adopt a causal machine learning framework to estimate the causal effects of extreme weather conditions on train arrival delays. Within this framework, each unit (a train journey) is associated with two potential outcomes: $Y(1)$, representing the outcome (arrival delay) if the unit is exposed to the treatment (extreme weather), and $Y(0)$, representing the outcome if it is not exposed to the treatment. The causal effect for a unit is then defined as the difference between these two potential outcomes:

$$\tau_i = Y_i(1) - Y_i(0) \tag{1}$$

However, the fundamental challenge arises from the fact that for any individual unit, only one of these outcomes is observed, while the other remains counterfactual. This issue, known as the Fundamental Problem of Causal Inference, necessitates additional assumptions to identify and estimate causal effects from observational data [26]. To ensure the validity of our causal claims, we make the following standard assumptions:

Assumption 1: Conditional Independence Assumption (CIA)

$$(Y(1), Y(0)) \perp T \mid X \tag{2}$$

This assumption implies that, conditional on observed covariates X (e.g., calendar features, occupancy levels, and weather conditions), the treatment assignment T (e.g., presence of extreme weather) is independent of the potential outcomes $Y(1)$ and $Y(0)$ (arrival delays). This ensures that there are no unobserved confounders influencing both the treatment and the outcome[22, 34]. To reduce potential biases, we include a rich set of pre-treatment variables such as precipitation, temperature, passenger boarding and alighting counts, wind speed, and time-related attributes.

Assumption 2: Stable Unit Treatment Value Assumption (SUTVA)

There is no interference between units, and each train trip's potential outcome depends only on its treatment status. In other words, the treatment status of one trip does not affect the delay outcome of another. This is a reasonable assumption given that individual trips and their delays are considered independent in the dataset [27].

Assumption 3: Positivity (Overlap)

$$0 < P(T = 1 \mid X = x) < 1 \quad \forall x \tag{3}$$

This assumption ensures that for every combination of covariates, there is a non-zero probability of receiving both treatment and control. To satisfy this, we perform trimming of extreme propensity scores (e.g., removing scores below 0.01 or above 0.99) to enforce common support and retain only samples with sufficient overlap [25].

Assumption 4: No Post-Treatment Bias

All covariates X used for adjustment are measured prior to the treatment (e.g., weather conditions are recorded before delay outcomes), avoiding the risk of conditioning on variables that are themselves affected by the treatment [23].

Assumption 5: Consistency

The observed outcome Y corresponds to the potential outcome under the observed treatment. That is, if a trip experienced extreme weather (treatment = 1), the observed delay is the same as the potential delay under treatment [15].

Adhering to these assumptions, we apply causal machine learning techniques, such as DML , to estimate the Average Treatment Effect (ATE) and Conditional Average Treatment Effect (CATE), also known as Group Average Treatment Effect (GATE), when conditioning on a subgroup.

Average Treatment Effect (ATE):

$$\tau = \mathbb{E}[Y(1) - Y(0)] \tag{4}$$

Conditional Average Treatment Effect (CATE / GATE):

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x] \tag{5}$$

Or for a subgroup $Z = z$:

$$\tau(z) = \mathbb{E}[Y(1) - Y(0) \mid Z = z] \tag{6}$$

This method yields a singular overall impact, GATEs enable the identification of certain rail lines or time periods that are most influenced. This information is crucial for formulating targeted and effective policy actions, like timetable adjustments, increased buffer times, or more efficient resource allocation in preparation for unfavorable weather conditions.

3.2 Double Machine Learning

We employ two advanced causal machine learning techniques, L-DML and CF-DML, to estimate the impact of extreme weather on rail delays while addressing high-dimensional confounding. In the following, we outline the general procedure of DML, which provides the foundation for both estimators used in this study. We then detail how the methodology is implemented using the L-DML and CF-DML frameworks, each tailored to estimate average and heterogeneous treatment effects.

DML, introduced by Chernozhukov et al. [7], is a robust semi-parametric estimation approach for causal inference from observational data. DML is particularly effective in managing high-dimensional covariates and leverages machine learning algorithms to estimate nuisance parameters. The key advantage of DML is its robustness against biases due to model misspecification, achieved through Neyman orthogonalization.

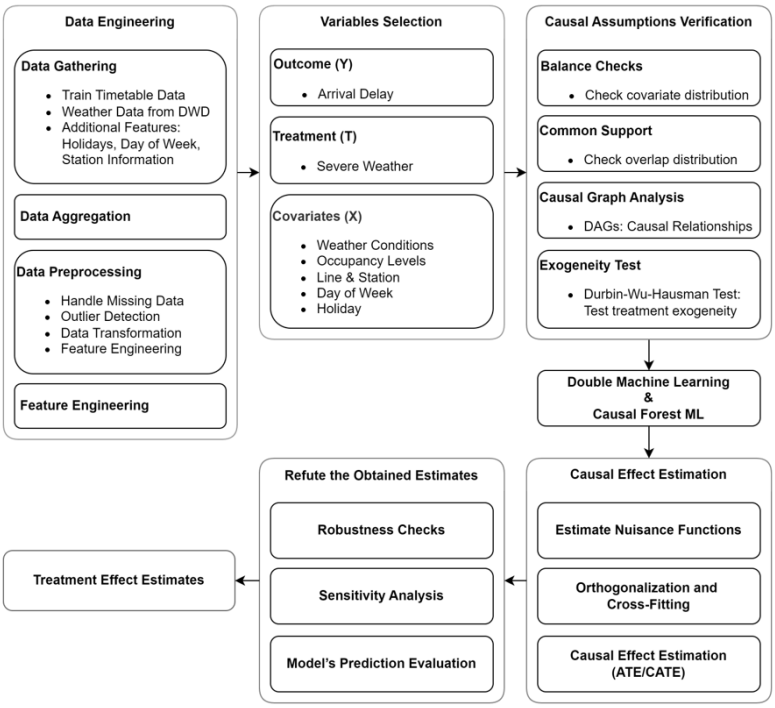

Figure 1: The whole structure of the Causal ML model

We consider the structural causal inference framework given by:

$$Y = g(D, X) + U, \quad D = m(X) + E, \tag{7}$$

$$\text{with} \quad E[U \mid X, D] = 0, \quad E[E \mid X] = 0 \tag{8}$$

where:

- Y is the observed outcome,
- D represents the treatment variable,
- X is a vector of covariates,
- U and E denote unobserved error terms.

Our primary interest lies in estimating the Average Treatment Effect (ATE), defined as:

$$ATE = E[g(1, X) - g(0, X)] \tag{9}$$

where $g(1, X)$ and $g(0, X)$ denote the expected outcomes (train delays) under treatment (extreme weather) and control (normal weather), respectively, conditional on covariates X. The function $g(d, X)$ represents the outcome regression function for treatment level $d \in \{0, 1\}$. This expression captures the average difference in expected delays between the treated and control groups. The treatment assignment process is described by the equation $D = m(X)+E$, where $m(X)$ is the treatment model, representing the conditional probability of receiving the treatment given covariates X. The term E denotes the residual variation in D that is not explained by X. This reflects the observational nature of the data, where treatment is not randomly assigned but depends on observed characteristics. By modeling D as a function of X, we can account for confounding and identify the causal effect of D on Y under the assumptions of unconfoundedness and overlap.

3.2.1 Estimation Procedure. DML employs a two-stage estimation procedure:

Stage 1: Estimation of nuisance functions. In the first stage, two nuisance functions are estimated using machine learning methods, Random Forests:

- **Conditional outcome model:** Estimates the expected outcome given treatment and covariates:

$$\hat{\mu}_d(X) = \hat{E}[Y \mid D = d, X], d \in \{0, 1\} \tag{10}$$

- **Propensity score model:** Estimates the probability of receiving the treatment given covariates:

$$\hat{p}(X) = \hat{E}[D \mid X] \tag{11}$$

These estimates are obtained through K-fold cross-validation to minimize overfitting and ensure robust out-of-sample performance.

Stage 2: Orthogonalized Estimation of the ATE. In the second stage of the DML procedure, we estimate the Average Treatment Effect (ATE) using a Neyman-orthogonal score function. This approach mitigates sensitivity to errors in the nuisance parameter estimates obtained in the first stage. Specifically, the orthogonalized score function for each observation is defined as:

$$\psi(W_i; \theta, \eta) = \frac{(D_i - \hat{p}(X_i))(Y_i - \hat{\mu}_{D_i}(X_i))}{\hat{p}(X_i)(1 - \hat{p}(X_i))} + (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) - \theta), \tag{12}$$

where $\hat{p}(X_i)$ is the estimated propensity score and $\hat{\mu}_d(X_i)$ denotes the predicted outcome under treatment level $d \in \{0, 1\}$. The Average Treatment Effect estimator ($\widehat{ATE}$) is then computed by solving

the empirical moment condition:

$$\frac{1}{N}\sum_{i=1}^{N}\psi(W_i;\widehat{ATE},\hat{\eta})=0. \tag{13}$$

This yields the following closed-form estimator:

$$\widehat{ATE}=\frac{1}{N}\sum_{i=1}^{N}\left[\hat{\mu}_1(X_i)-\hat{\mu}_0(X_i)+\frac{(D_i-\hat{p}(X_i))(Y_i-\hat{\mu}_{D_i}(X_i))}{\hat{p}(X_i)(1-\hat{p}(X_i))}\right]. \tag{14}$$

This estimator combines outcome regression and inverse probability weighting in a doubly robust manner and leverages Neyman orthogonality to ensure robustness to small estimation errors in the nuisance functions. The method enables valid inference even in high-dimensional settings and forms the foundation of modern semiparametric causal inference techniques [7].

Stage 3: Estimation of Heterogeneous Treatment Effects (CATE and GATE). The average treatment effect (ATE) provides a population-level summary of causal impact; many real-world applications, especially in transportation systems, require understanding how treatment effects vary across different contexts or subgroups. To address this, we estimate heterogeneous treatment effects, specifically Conditional Average Treatment Effects (CATEs) and Group Average Treatment Effects (GATEs). The CATE is defined as:

$$\tau(x)=\mathbb{E}[Y(1)-Y(0)\mid X=x], \tag{15}$$

where X represents a vector of observed covariates, and $Y(1)$ and $Y(0)$ are the potential outcomes under treatment and control, respectively. $\tau(x)$ captures the causal effect for individuals with characteristics $X=x$, offering a more granular understanding of treatment heterogeneity.

To obtain interpretable summaries of these effects across relevant groups, for instance months, train lines, and hours, we compute Group Average Treatment Effects (GATEs), which are defined as:

$$\tau(z)=\mathbb{E}[Y(1)-Y(0)\mid Z=z], \tag{16}$$

where Z denotes a discrete grouping variable such as season, weekday, or location. GATEs are computed by averaging the estimated CATEs within each group:

$$\widehat{\tau}(z)=\frac{1}{n_z}\sum_{i:Z_i=z}\hat{\tau}(X_i), \tag{17}$$

where n_z is the number of observations in group z.

This approach allows us to identify subpopulations most affected by the treatment, extreme weather, enabling the formulation of targeted and efficient policy interventions. For example, GATEs by time of day or rail line can inform resource allocation and scheduling strategies under adverse weather conditions.

3.2.2 Refute the Obtained Estimates. To evaluate the robustness of the estimated causal effect of extreme weather on arrival delay, we applied multiple falsification and validation techniques. These procedures test whether the estimated treatment effects are reliable, or possibly driven by spurious correlations or model overfitting. The main steps include:

- **Cross-Fitting** To further enhance robustness, DML incorporates cross-fitting. Data are partitioned into K folds; nuisance functions are estimated in separate training folds and evaluated on distinct hold-out folds. The final cross-fitted DML estimator is the average across folds:

$$\widehat{ATE}_{K-fold}=\frac{1}{K}\sum_{k=1}^{K}\widehat{ATE}^{(k)} \tag{18}$$

  where $\widehat{ATE}^{(k)}$ is estimated using the k^{th} fold as a validation set and the remaining data as the training set.
- **Placebo Test:** We conducted a placebo treatment test based on the principles of randomization inference. In this procedure, the treatment assignment indicating exposure to extreme weather was randomly permuted across observations while maintaining the original proportion of treated and control units. This method is rooted in the statistical framework by Fisher [11], where under the sharp null hypothesis of no treatment effect, any random reassignment of treatment labels should yield estimates centered around zero.
  Formally, the placebo test evaluates the hypothesis:

$$H_0: Y_i(1)=Y_i(0)\quad\forall i \tag{19}$$

  where $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes for unit i under treatment and control conditions, respectively. We re-estimated the causal effect using the permuted treatment to assess whether the original effect was likely due to a true causal link.
- **Sensitivity Analysis:** To examine the influence of individual covariates on the estimated treatment effects, we iteratively removed potential confounders (e.g., Occupancy) and re-estimated the Average Treatment Effect (ATE). This process helped identify which variables were essential for maintaining effect stability.
- **Out-of-Sample Validation:** We applied cross-validation along with standard evaluation metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) to assess the predictive performance of the underlying models used in the causal estimation. These metrics provide insight into the accuracy and generalizability of the model across different data splits.

4 Experiment and Results

This section presents the experimental design, dataset characteristics, and empirical findings of the study. We begin by describing the data sources and preprocessing steps, followed by the specification of treatment and covariates. We then assess the validity of causal assumptions and proceed with the estimation of average and heterogeneous treatment effects using the Double Machine Learning framework. Finally, we conduct robustness checks to verify the consistency of our findings and evaluate the performance of the model.

4.1 Data Preparation

The core operating dataset was obtained from a regional rail logistics provider and includes train-level and station-level data covering scheduled and actual arrival and departure times, train lines, and station identifiers. Passenger-related features such as occupancy,

boarding counts, and alighting counts were also included, offering insights into demand and dwell time per stop. To capture environmental influences, this operational data was augmented with daily weather records from the German Weather Service (DWD). Specifically, we used DWD's Bad Weather Days dataset, which categorizes each day based on weather severity: 1) A (very difficult conditions), 2) B (difficult), 3) C (damaging). These classifications are derived from thresholds on wind, temperature, and precipitation. For modeling purposes, we generated a binary extreme weather indicator (Extreme Weather = 1 for categories A, B, or C) to identify environmentally disruptive days.

We also integrated calendar information, including German public holidays specific to the federal state of Hessen, and one-hot encoded day-of-week indicators (Monday to Sunday). These temporal signals are crucial in transportation modeling, as they capture routine patterns in passenger volume, scheduling practices, and potential disruptions (e.g., weekend service reductions or holiday crowding).

The datasets were merged using date and station-level identifiers. Temporal alignment ensured that weather and holiday information matched the operational records on a daily basis. Passenger statistics were aggregated per train stop to match the granularity of the delay data. The final dataset spans the years 2017 to 2022 and includes multiple lines operating in central Germany.

4.2 Treatment and Covariate Specification

Extreme Weather has been identified as the treatment variable. The outcome of interest is Arrival Delay. To account for factors that might simultaneously affect both weather conditions and arrival delays, we selected a set of confounding variables based on domain understanding and empirical associations. To identify appropriate covariates for evaluating the causal impact of extreme weather on train arrival delays, we initially analyzed the correlations among all variables, the treatment (Extreme Weather), and the outcome (Arrival Delay) to detect potential confounders. We created a Pearson correlation matrix, Figure 2, as a preliminary step in our variable selection procedure to assess potential multicollinearity among the variables. This matrix indicated that the majority of variables display weak correlations, implying that each provides distinct information to the model. We identified a moderate correlation of 57% between "year" and "line," suggesting a potential temporal trend associated with operational changes, and a 48% correlation between alighting and boarding counts, illustrating the inherent link in passenger flow data.

We eliminated post-treatment variables, including rainfall, snow depth, and temperature measurements, as they either define or follow the treatment and would distort the estimation. Similarly, Planned Arrival and Planned Departure times were not included as covariates, as these variables are directly involved in the construction of the outcome variable, arrival delay. We retained only pretreatment variables that could influence both the likelihood of extreme weather and arrival delays, including temporal variables (Year, Hour, Holidays), operational variables (Station ID, Line), and passenger metrics (Occupancy, Boarding Count, Alighting Count). Categorical variables such as Line, Month, and Day of week were encoded with one hot spot to make them suitable for machine learning models.

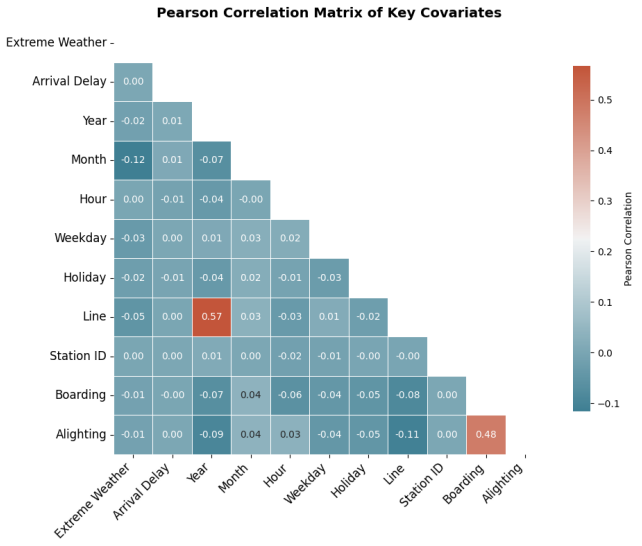

Figure 2: Pearson Correlation Matrix

4.3 Validating Causal Assumptions

This section evaluates the foundational assumptions of causal inference to ensure that the estimated causal effects can be interpreted as unbiased and reliable.

4.3.1 Balance Check. Figure 3 presents the standardized mean differences (SMDs) for individual covariates and grouped dummy variables, assessing balance between the treatment group (extreme weather) and the control group (normal weather). An SMD near zero indicates good balance, while values beyond ±0.1 suggest imbalance, and values exceeding ±0.2 are considered problematic.

Most operational covariates, including Year, Hour, Station ID, and Boarding/Alighting Counts, exhibit strong balance, with SMDs well below ±0.1. Holidays also remain within acceptable limits. In contrast, Occupancy displays a moderate imbalance, suggesting that it may be a confounding factor that warrants careful control in the modeling.

Regarding grouped dummy variables, the month group reveals the most critical imbalance: Multiple individual months, such as January, June, and August, exceed the ±0.2 threshold, indicating problematic imbalance related to seasonality. The weekday group also shows an uneven distribution, particularly on Saturdays and Tuesdays, which implies that weather extremes may not be uniformly distributed throughout the week. Similarly, the Line group shows that certain routes, especially Line RB8 (−0.1294), are overrepresented in either the treated or control group, although most lines remain within mild to moderate imbalance.

These findings emphasize the need for rigorous adjustment. We applied a Double Machine Learning (DML) framework to flexibly control for these covariates and correct potential biases. This approach is particularly suited to settings with high-dimensional

confounders and moderate imbalance, ensuring more reliable estimation of treatment effects under extreme weather conditions.

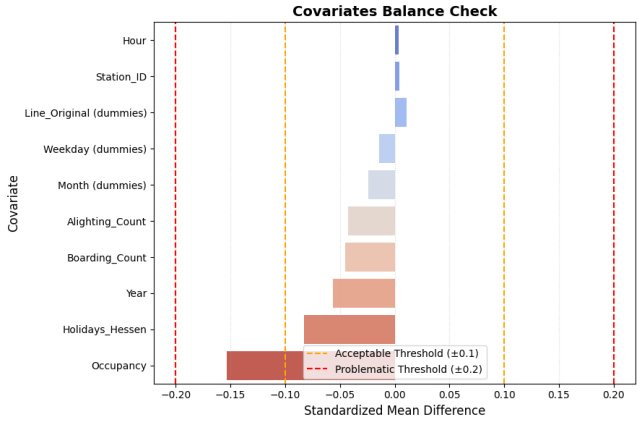

Figure 3: Covariate Balance Check Between Extreme Weather and Normal Weather Groups

4.3.2 Overlap Check. Figure 4, the overlap plot of the propensity score, illustrates the distribution of the estimated propensity scores for the treated group (Extreme Weather) and the control group (Normal Weather), with the propensity score derived from a logistic regression model using selected covariates. The propensity score represents the predicted probability of experiencing extreme weather, conditional on the selected covariates. The substantial overlap between the two distributions ensures that the treated and control units are comparable across the full range of propensity scores, satisfying the common support condition. This validates the common support condition, which is crucial for finding causal effects by ensuring the existence of a comparable control unit for each treated unit.

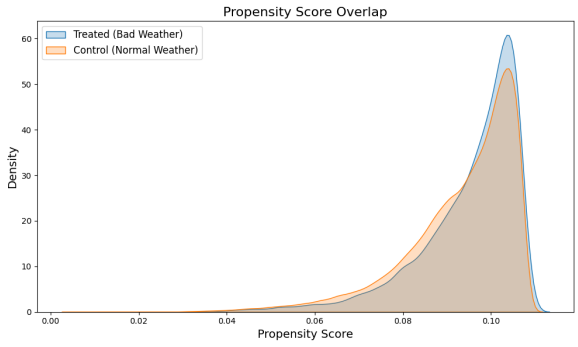

Figure 4: Distribution of estimated propensity scores

4.3.3 Exogeneity Test. To evaluate the exogeneity of the treatment variable, the Durbin-Wu-Hausman (DWH) test was performed. This test assesses the correlation between treatment and error term in the outcome equation, suggesting potential endogeneity and bias in causal estimations. The methodology entailed a two-phase regression procedure: initially, the treatment was regressed on all variables to provide expected values and residuals; subsequently, the outcome variable, Arrival Delay, was regressed on both the treatment and the residuals from the initial phase. The primary coefficient of interest is that of the residual component. If statistically significant, it indicates that the treatment is endogenous. Table 1

Table 1: Durbin-Wu-Hausman Test Results for Exogeneity of *Extreme Weather*

Test	Coefficient	Std. Error	t-Statistic	p-Value
DWH	8.63×10^{-7}	1.69×10^{-6}	0.510	0.610

highlights the findings of the DWH test. The results indicate that the residuals are not statistically significant (p = 0.610), supporting the exogeneity assumption for the treatment variable. This outcome supports the validity of proceeding with the estimation of the average treatment effect (ATE) and conditional average treatment effects (CATE) using causal inference methodologies.

4.3.4 Causal Graph Analysis. To understand the causal relationship between arrival delays and extreme weather events, we apply a causal graph analysis based on a Directed Acyclic Graph (DAG). This graphical model encodes our assumptions about the underlying data-generating process and provides a transparent framework to identify direct causal effects, mediating mechanisms, and potential confounding structures. The DAG developed for this study is

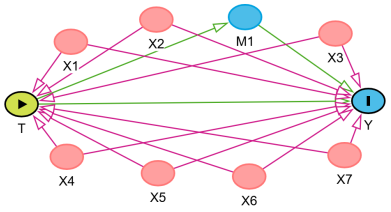

Figure 5: Causal Directed Acyclic Graph (DAG)

shown in Figure 5. The outcome variable is Arrival Delay (Y), and the treatment variable is Extreme Weather (T). We hypothesize that weather affects delays both directly and indirectly, with Passenger Occupancy (M_1) acting as a mediator. Both treatment and outcome are assumed to be influenced by a set of observed confounders (X_1 to X_7), which include Month, Day of the Week, Station ID, Line, Holiday, Year, and Hour.

To identify the total causal effect of extreme weather on arrival delays, it is essential to adjust for these observed confounders. Conditioning in this set of covariates satisfies the backdoor criterion, ensuring that any spurious paths from T to Y through common causes are appropriately blocked, allowing for unbiased estimation of the total effect of treatment.

4.4 Treatment Effects Analysis

Table 2: ATE Inference Results for the Treatment Effect of Extreme Weather

Model	ATE	Std. Error	z-Statistic	95% CI Lower	95% CI Upper
DML	20.546	9.652	2.129	1.628	39.465

Table 2 presents the estimated Conditional Average Treatment Effects (CATEs) derived from the DML model, evaluating the causal impact of extreme weather conditions on train arrival delays. The DML model estimates an average treatment effect (ATE) of 20.55 minutes, with a standard error of 9.65. The corresponding z-statistic of 2.129 and p-value of 0.033 indicate statistical significance at the 5% level. The 95% confidence interval spans from 1.63 to 39.47 minutes. These results suggest a consistent and moderately strong positive effect of extreme weather on train delays, with robust statistical support.

4.5 Grouped Average Treatment Effects (GATEs)

Examining GATEs demonstrates a consistent and statistically significant causal relationship between extreme weather conditions and train arrival delays across all stratification groups: train lines, calendar months, and weekdays. Each GATE has been estimated by aggregating estimated individual treatment effects derived from a machine learning-based causal inference framework, enabling the identification of heterogeneous effects while controlling for high-dimensional confounders.

In this approach, for each group, we present the point estimate of the effect (GATE), the corresponding standard error, the 96% confidence interval, and the p-value obtained from a two-sided z-test. A group is considered severely affected by extreme weather if the 96% confidence interval excludes zero, indicating a p-value below 0.04. As shown in Figure 6, the analysis reveals that extreme weather significantly affects all train lines, with the most severe delays on Line RB83 (41.6 minutes) and a surprising negative effect on RE50 (11.4 minutes), possibly due to compensatory operational strategies. Across months, the strongest impacts are seen in December and November (55.4 and 50.4 minutes), consistent with seasonal weather challenges, while September uniquely shows a negative effect (34.8 minutes), potentially reflecting transitional conditions or lower demand. Weekday results show the largest delays on Tuesday (51.9 minutes) and notable effects on Sunday, whereas Saturday stands out with a small but significant negative effect (1.5 minutes), suggesting that reduced weekend services may help limit disruption. Together, these findings highlight pronounced heterogeneity in how extreme weather affects rail operations across lines, seasons, and days of the week.

4.6 Refute the Obtained Estimates

4.6.1 Models Validation. We assessed the predictive performance of the outcome model within the causal framework using a Random Forest regressor to determine its accuracy. The model attained a Root Mean Squared Error (RMSE) of 3.72 minutes and a Mean Absolute Error (MAE) of 2.59 minutes. The findings suggest that

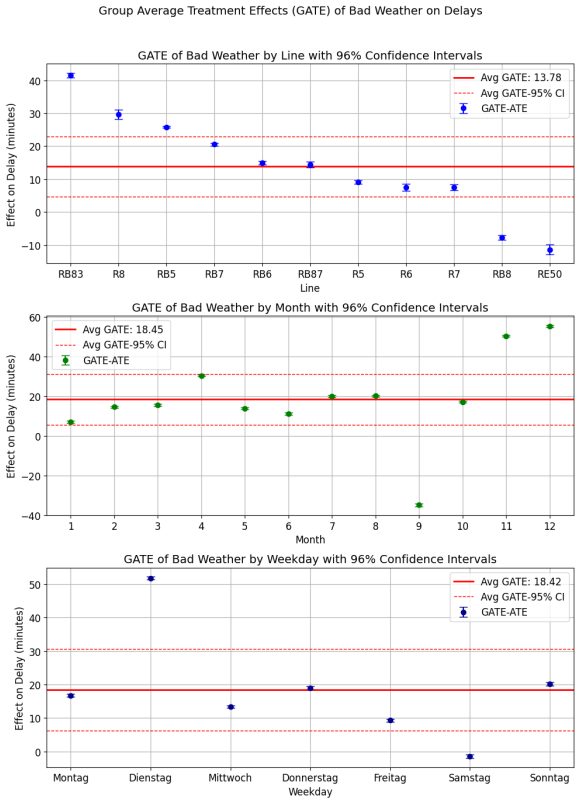

Figure 6: Grouped Average Treatment Effects by Line, Month, and Weekday

the model effectively captures significant patterns in the data and demonstrates reasonable accuracy in predicting train arrival delays, thus establishing a reliable basis for subsequent causal effect estimation.

To ensure the robustness of treatment effect estimation, we compared two widely used predictive models, Random Forest (RF) and Gradient Boosting Regressor (GBR), for modeling the outcome variable. Both methods are tree-based ensemble learners commonly employed in causal machine learning frameworks. In these settings, the accuracy of the outcome model plays a pivotal role, as it directly influences the quality and reliability of treatment effect estimates. If this model is poorly specified, it can introduce bias or inefficiency into the estimated ATE. As shown in Table 3, RF achieved a lower prediction error compared to GBR, suggesting that RF provides a more suitable basis for subsequent causal inference.

Table 3: Comparison of Predictive Models for Outcome Estimation

Model	MAE	RMSE
Random Forest Regressor	2.59	3.72
Gradient Boosting Regressor	3.00	4.23

The treatment model exhibits high predictive accuracy in distinguishing between extreme weather and normal conditions, as

presented in the table. 4. The model achieves an overall accuracy of 96.4%, with particularly high precision and recall for the control class (normal weather). While performance on the treated class (extreme weather) is slightly lower in recall (71.6%), the precision remains high (86.7%), indicating that predicted treatment instances are generally reliable. The macro-averaged F1-score of 0.88 reflects good balance between precision and recall across both classes, while the weighted average F1-score of 0.96 highlights the model's robustness despite class imbalance. These results support the adequacy of the treatment model for propensity score estimation in subsequent causal inference steps.

Table 4: Classification Performance of Treatment Model

Metric	Control Group	Treated Group	Overall / Avg.
Precision	0.972	0.867	0.919
Recall	0.989	0.716	0.852
F1-Score	0.980	0.784	0.882
Accuracy			0.964

4.6.2 Robustness Check. To assess the robustness of the causal effect estimates, a placebo test was conducted by randomly permuting the treatment assignment and re-estimating the ATE using the same causal model specification. The resulting placebo ATE was approximately 0.0066, suggesting that there was no significant treatment effect under a random treatment assignment. This supports the validity of the original findings and indicates that the estimated effects are unlikely to be driven by random chance or misspecification of the model.

4.6.3 Sensitivity Analysis. To assess the robustness of our estimates of causal effects, we performed a sensitivity analysis excluding passenger-related covariates, namely occupancy, boarding, and alighting counts, from the model specification. This allowed us to evaluate how the omission of potentially important confounders influences both the estimated average treatment effects (ATE) and the quality of the covariate balance, as reflected in the propensity score distribution.

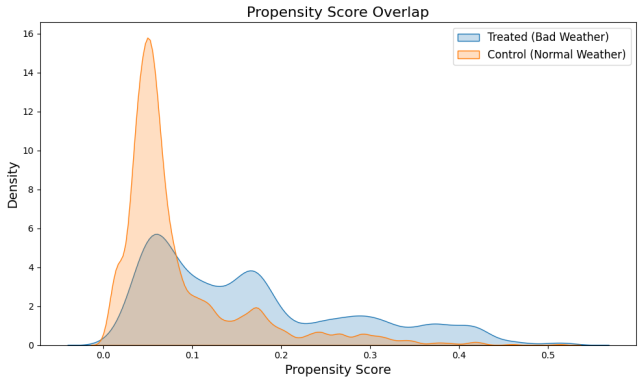

Figure 7: Propensity Score Overlap Without Passenger-related Covariates

Figure 7 illustrates the overlap in propensity scores between the treated (extreme weather) and control (normal weather) groups. When occupancy and boarding/alighting counts are excluded, the overlap between the two groups becomes visibly poor. The distribution for the control group has sharply peaked at low propensity values (close to 0.05), whereas the treated group has a more spread distribution with a much heavier right tail. This separation indicates a poor common support and suggests that the covariates are insufficient to balance the two groups. Without adequate overlap, treatment effect estimates may become biased or unstable due to extrapolation in regions without comparable treated and control observations. Our findings underscore the importance of integrating passenger-related data in causal delay analysis. Excluding these factors resulted in implausibly elevated ATE, 92.72 with DML, and inadequate overlap in propensity scores, indicating possible misspecification due to unmeasured confounding.

5 Conclusion

This research utilizes causal machine learning techniques to assess the influence of extreme weather and passenger load on train delays in public rail systems. By employing DML, we systematically estimate both average and heterogeneous treatment effects while mitigating significant confounding variables. The findings indicate extreme weather substantially exacerbates delays, with variations observed among train lines, time intervals, and passenger densities. These findings highlight the necessity of transcending conventional predictive methods to adopt causal inference techniques that can reveal significant differences in treatment effects.

From an applied perspective, the study yields valuable implications for rail operators and decision-makers. By identifying instances and areas where weather and occupancy factors collectively intensify delays, stakeholders can modify scheduling methods, enhance resource allocation, and establish proactive mitigation strategies. This indicates an increasing focus in transportation research on utilizing causal insights to guide adaptive and data-driven operations.

Despite the comprehensive causal modeling framework, the study faces limitations. The weather data used were categorized rather than continuous, which may obscure finer-grained effects of specific weather events. Future studies could benefit from integrating more granular meteorological data or incorporating real-time weather monitoring to improve causal precision.

In conclusion, this application of causal machine learning contributes to a deeper understanding of delay mechanisms in public rail transport. It underscores the importance of explicitly modeling causal relationships among operational drivers of delay, enabling more accurate diagnostics and service reliability improvements. As climate variability and transport demand continue to increase, integrating causal methods into transportation planning and operations will be essential for building more resilient and efficient systems.

6 Acknowledgments

Acknowledgments

The authors would like to acknowledge the CargoSurfer project partners and the rail logistics provider for their invaluable data contributions, technical support, and funding that made this research possible.

References

[1] A. Abdi and C. Amrit. 2021. A review of travel and arrival-time prediction methods on road networks: classification, challenges and opportunities. *PeerJ Computer Science* 7 (2021), e689. doi:10.7717/peerj-cs.689

[2] M. Arshad and M. Ahmed. 2019. Train delay estimation in Indian railways by including weather factors. *Open Transportation Journal* (2019). doi:10.2174/2666255813666190912095739

[3] Shindy Arti, Indriana Hidayah, and Sri Suning Kusumawardhani. 2020. Research Trend of Causal Machine Learning Method: A Literature Review.

[4] Susan Athey, Julie Tibshirani, and Stefan Wager. 2019. Generalized Random Forests. *The Annals of Statistics* 47, 2 (2019), 1148–1178. doi:10.1214/18-AOS1709

[5] Frank Bodendorf, Max Sauter, and Jörg Franke. 2022. A mixed method approach to analyze and predict supply disruptions by combining causal inference and deep learning. *International Journal of Production Economics* (2022).

[6] Zhenhua Chen and Yuxuan Wang. 2019. Impacts of severe weather events on high-speed rail and aviation delays. *Transportation Research Part D: Transport and Environment* 69 (2019), 168–183. doi:10.1016/j.trd.2019.01.030

[7] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21, 1 (2018), C1–C68. doi:10.1111/ectj.12097

[8] Sun Choi, Young Jin Kim, Simon Briceno, and Dimitri Mavris. 2016. Prediction of weather-induced airline delays based on machine learning algorithms. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*. 1–6. doi:10.1109/DASC.2016.7777956

[9] Hongrui Chu, Wensi Zhang, Peng Fei Bai, and Yahong Chen. 2021. Data-driven optimization for last-mile delivery. *Complex Intelligent Systems* 9 (2021), 2271–2284.

[10] McWillian de Oliveira, Ana Beatriz Rebouças Eufrásio, Marcelo Xavier Guterres, Mayara Condé Rocha Murça, and Rogéria de Arantes Gomes. 2021. Analysis of airport weather impact on on-time performance of arrival flights for the Brazilian domestic air transportation system. *Journal of Air Transport Management* 91 (2021), 101974. doi:10.1016/j.jairtraman.2020.101974

[11] Ronald Aylmer Fisher. 1935. *The Design of Experiments*. Oliver and Boyd, Edinburgh.

[12] Qian Fu and John Easton. 2018. Prediction of Weather-Related Incidents on the Rail Network: Prototype Data Model for Wind-Related Delays in Great Britain. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems Part A Civil Engineering* 4 (06 2018). doi:10.1061/AJRUA6.0000975

[13] Sarah Greenham, Emma Ferranti, Andrew Quinn, and Katherine Drayson. 2020. The impact of high temperatures and extreme heat to delays on the London Underground rail network: An empirical study. *Meteorological Applications* 27, 3 (2020), e1910. doi:10.1002/met.1910

[14] Stefan Gössling, Christoph Neger, Robert Steiger, and Rainer Bell. 2023. Weather, climate change, and transport: a review. *Natural Hazards* 118 (2023), 1341–1360. doi:10.1007/s11069-023-06054-2

[15] Miguel A. Hernán and James M. Robins. 2020. *Causal Inference: What If*. Chapman and Hall/CRC. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/ Available at https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

[16] Paul Huenermund, Jermain Kaminski, and Carla Schmitt. 2021. Causal Machine Learning and Business Decision Making. *SSRN Electronic Journal* (2021).

[17] Md. Kamrul Islam. 2011. Impact of Occupancy Profile for Waiting Time of Bus Transit Users. In *Proceedings of the Eastern Asia Society for Transportation Studies*. https://api.semanticscholar.org/CorpusID:106457439

[18] Yasanur Kayikci and Volker Stix. 2014. Causal mechanism in transport collaboration. *Expert Systems with Applications* 41, 4 (2014), 1561–1575. doi:10.1016/j.eswa.2013.08.053

[19] Shuang Li, Ziyuan Pu, Zhiyong Cui, Seunghyeon Lee, Xiucheng Guo, and Dong Ngoduy. 2024. Inferring heterogeneous treatment effects of crashes on highway traffic: A doubly robust causal machine learning approach. *Transportation Research Part C: Emerging Technologies* 160 (2024), 104537. doi:10.1016/j.trc.2024.104537

[20] Nikola Marković, Sanjin Milinković, Konstantin S. Tikhonov, and Paul Schonfeld. 2015. Analyzing passenger train arrival delays with support vector regression. *Transportation Research Part C: Emerging Technologies* 56 (2015), 251–262. doi:10.1016/j.trc.2015.04.004

[21] Stephen L. Morgan and Christopher Winship. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (2 ed.). Cambridge University Press, Cambridge.

[22] Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*. Cambridge University Press.

[23] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

[24] Juan Pineda-Jaramillo and Francesco Viti. 2023. Identifying the rail operating features associated to intermodal freight rail operation delays. *Transportation Research Part C: Emerging Technologies* 147 (2023), 103993. doi:10.1016/j.trc.2022.103993

[25] Paul R. Rosenbaum and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55. doi:10.1093/biomet/70.1.41

[26] Donald B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 5 (1974), 688–701.

[27] Donald B. Rubin. 1980. Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *J. Amer. Statist. Assoc.* 75 (1980), 591.

[28] Mahnam Saeednia, Stefan Wegele, Rolf Goossmann, Cem Ormesher Hussein, and Scott Heath. 2023. Management of extreme weather impact on railway operations. *Transportation Research Procedia* 72 (2023), 2644–2651. doi:10.1016/j.trpro.2023.11.803 TRA Lisbon 2022 Conference Proceedings.

[29] Gill Varghese Sajan and Priyanka Kumar. 2021. Forecasting and Analysis of Train Delays and Impact of Weather Data using Machine Learning. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. 1–8. doi:10.1109/ICCCNT51525.2021.9580176

[30] Banavar Sridhar and Neil Y. Chen. 2008. Short-Term National Airspace System Delay Prediction Using Weather Impacted Traffic Index. *Journal of Guidance Control and Dynamics* 32 (2008), 657–662.

[31] Nishtha Srivastava, Bhavesh N. Gohil, and Suprio Ray. 2024. Rail transit delay forecasting with Causal Machine Learning. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Spatiotemporal Causal Analysis*. 1–10. doi:10.1145/3681778.3698784

[32] Iraklis Stamos, Evangelos Mitsakis, Josep Maria Salanova, and Georgia Aifadopoulou. 2015. Impact assessment of extreme weather events on transport networks: A data-driven approach. *Transportation Research Part D: Transport and Environment* 34 (2015), 168–178. doi:10.1016/j.trd.2014.11.002

[33] Dothang Truong. 2021. Using causal machine learning for predicting the risk of flight delays in air transportation. *Journal of Air Transport Management* 91 (2021), 101993. doi:10.1016/j.jairtraman.2020.101993

[34] Stefan Wager and Susan Athey. 2018. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242. doi:10.1080/01621459.2017.1319839

[35] Pu Wang and Qing peng Zhang. 2019. Train delay analysis and prediction based on big data fusion. *Transportation Safety and Environment* (2019).

[36] Xu Zhang and Mei Chen. 2019. Quantifying the Impact of Weather Events on Travel Time and Reliability. *Journal of Advanced Transportation* (2019). https://api.semanticscholar.org/CorpusID:115762221

[37] Ying Zhao, Jacob Jones, and Douglas J. MacKinnon. 2019. Causal Learning to Discover Supply Chain Vulnerability. In *International Conference on Knowledge Discovery and Information Retrieval*.

[38] Meihong Zhu. 2023. The Effect of Political Participation of Chinese Citizens on Government Satisfaction: Based on Modified Causal Forest. *Procedia Computer Science* 221 (2023), 1044–1051. doi:10.1016/j.procs.2023.08.086