

<https://doi.org/10.1038/s44333-026-00097-1>

A causal discovery and inference framework for on-demand food delivery delays

Check for updates

Miaojia Lu¹, Rui Liu¹, Zhicheng Jin^{1,2}✉ & Quan Yuan¹

The on-demand food delivery (OFD) industry has seen growth yet grapples with arrival delays. For major platforms like DoorDash and Uber Eats, over one-third of orders arrive late, highlighting the severity of this challenge. Currently, few studies distinguish true drivers of these delays from mere correlations. This study addresses this gap by developing an innovative framework combining Bayesian causal discovery with double machine learning. From 405,180 OFD records in China, we found that 16.7% of orders experienced delays. Pickup and transport durations exhibited the strongest causal effects to these delays. In addition, delay propagation was first identified within OFD services, where delays in preceding orders significantly increase the length of subsequent delays. The findings offer practical insights for OFD platforms to reduce order delays, such as optimizing courier pickup processes and mitigating delay propagation. By targeting these root causes, platforms can enhance operational efficiency and make their services more sustainable.

The on-demand food delivery (OFD) market is experiencing explosive growth and is expected to surge by USD 470.5 billion from 2024 to 2029, with a compound annual growth rate of 26.9%¹. This rapid expansion places increasing demands on the operational efficiency of OFD platforms. However, delivery delays remain a critical operational challenge, undermining the platform's core promise of speed and efficiency. In recent reports, over a third of OFD orders are delayed, with DoorDash recording a 38% delay rate and Uber Eats an even higher 44%². Such high delay rates not only inconvenience customers but also diminish the operational efficiency of OFD platforms and courier earnings: customers rank delays among their top complaints, eroding trust and reducing repeat purchases³; platforms suffer reputational damage, negative reviews, and churn⁴; and couriers face penalties, intense pressure, and incentives for unsafe riding⁵. The root of this problem lies in the system's inherent complexity, where a multitude of stochastic variables can affect the final arrival time, such as merchant food preparation, courier pickup, and transport. Therefore, it is essential to identify the primary drivers of OFD delays and to investigate their respective contributions and impacts. This will enable platforms to develop targeted interventions aimed at reducing both the frequency and duration of delivery delays.

OFD delays stem from both endogenous and external factors. Endogenous factors are rooted in the sequential OFD process—order processing, courier pickup, and final transport—where bottlenecks like slow assignment or prolonged preparation lead to late arrival. External factors include spatio-

temporal variables (e.g., order timing, location) and dynamic conditions (e.g., weather and traffic)⁶. Beyond these known factors, a critical interaction remains largely underexplored in OFD literature: the propagation of delays between consecutive orders. While documented in public transit^{7,8}, it is unclear if a delay from one delivery reliably affects the next in a courier's delivery wave.

Existing research has primarily focused on predicting delivery times using machine learning models^{9,10} or on ranking influential variables using methods like XGBoost and SHAP^{11,12}. However, a significant gap remains: these approaches only uncover correlation, not causation⁷. This distinction is vital; a factor highly correlated with delay (like delivery speed or distance) may simply be a symptom, while the direct causal driver is the resulting transport time. Intervening based on correlation alone risks targeting symptoms, leading to ineffective strategies. To enable the development of robust, effective interventions, recent research has advocated for causal analysis, which is specifically designed to distinguish true causal relationships from mere statistical associations^{7,8}.

This study collected a total of 405,180 OFD order records in China to account for independent variables causing delivery delays and estimate each variable's causal contribution. Variables, such as platform processing duration, merchant preparation time, courier transport duration, time-of-day factors, and the delays of preceding orders, were incorporated into a causal graph to infer their causal relationships with delay. By explicitly modeling these dependencies, this research offers actionable insights for

¹The Key Laboratory of Road and Traffic Engineering, College of Transportation, Tongji University, Jiading district, Shanghai, China. ²Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China. ✉e-mail: zhicheng.jin@connect.polyu.hk

mitigating delay propagation and improving system reliability. The causal analysis in this study provides a roadmap for optimizing last-mile logistics to support more economically and socially sustainable urban transport systems. Reducing delays enhances the economic sustainability of delivery platforms by boosting operational efficiency, courier productivity, and customer retention. In addition, mitigating the systemic pressures that cause delays can reduce incentives for risky courier behavior, thereby promoting safer working conditions and social sustainability. The primary contributions of this study are as follows:

- We develop an end-to-end framework that integrates Bayesian causal discovery with DML-based estimation to identify the structural drivers of delivery delays. This approach allows us to move beyond correlation-based analysis and to infer the underlying causal mechanisms that shape delays.
- Our analysis provides clear evidence that delays propagate across orders handled by the same courier. Longer processing and pickup durations, as well as delays originating from preceding delivery orders, are shown to have measurable causal impacts on the delay of subsequent orders.
- By comparing our causal approach against traditional correlation-based methods, we demonstrate the value of causal inference techniques in distinguishing true delay drivers from variables that merely co-vary with delays.

This study is structured as follows. Section Literature review reviews the literature on delay analysis and causal methods. Section Methods introduces the collected OFD data and details the proposed causal analysis framework. Section Results presents the causal graph of delays and estimates causal effects for different variables. Section Discussion discusses some policy implications based on the findings, and Section Conclusions concludes with a summary of key results, limitations, and future directions.

Literature review

Factors related to OFD delays

Delays in OFD arise from a combination of endogenous operational factors and external environmental conditions. These factors interact throughout the service process and jointly shape final delivery performance.

Endogenous factors originate within the OFD workflow and are directly influenced by platform operations. Delays may occur at any stage, from order creation to final delivery. A key factor is the duration of order processing and assignment. Slow matching between couriers and incoming orders, especially during peak demand, can create initial backlogs. Platforms continually refine these algorithms, incorporating complex decisions regarding order bundling to enhance efficiency¹³. Another critical factor is merchant preparation time. Kitchen layouts, preparation complexity, and

unexpected surges in orders can extend preparation times and force couriers to wait. Prior studies have emphasized that fixed or lengthy preparation cycles can dominate the customer’s total waiting time and impose a hard constraint on subsequent stages^{14,15}. In addition, the delivery distance between the merchant and customer, along with the courier’s transport time, represents another important set of endogenous factors. A substantial body of research has focused on developing fast routing algorithms to help couriers determine the order of deliveries and select efficient transport paths. The primary objective of these optimization methods is to reduce transport time and improve overall delivery efficiency^{16,17}.

The propagation of delays between consecutive orders is also an endogenous variable. Distinct from the endogenous factors mentioned above, it remains underexplored in OFD literature, even though it is well established in public transit research^{7,8}. Within OFD, a courier’s tasks are often batched into delivery waves. When an order is delayed, regardless of whether the cause lies in preparation, pickup, or transport, the remaining orders in the wave may inherit this delay. Existing studies have acknowledged this cascading phenomenon, but they have not quantified its magnitude or operational implications¹⁸.

External factors, beyond the platform’s direct control, also exert significant impacts on the final delivery time of an order. Temporal variables, including lunch and dinner peaks, weekends, and holidays, tend to trigger abrupt increases in order volume. This increased demand intensifies the delivery burden on couriers, ultimately leading to delayed deliveries¹⁹. Regarding spatial factors, delivery durations are typically prolonged in suburban areas compared to dense metropolitan centers¹⁹. Additionally, adverse weather conditions like heavy rain or snow hinder courier speeds and affect riding behavior, potentially leading to widespread delays¹⁹.

Table 1 summarizes the factors identified in the literature. Although numerous studies have discussed endogenous and external determinants of OFD timeliness, most focus on prediction or optimization rather than delay-specific causal mechanisms. Only a limited number of empirical works have examined delays directly, underscoring the need for a systematic causal analysis such as the one developed in this study.

Methods applied in OFD delays

Existing research has primarily focused on predicting delivery times using machine learning models or on ranking influential variables using methods like XGBoost and SHAP. Salari et al.⁹ identified delivery time as a key factor in customer satisfaction and developed an enhanced tree-based model to more accurately estimate the delivery time distribution. Using a Difference-in-Differences approach, Cui et al.²⁰ found that quicker deliveries boosted sales and profits. Chen et al.²¹ and Mai and Le²² designed online delivery scheduling algorithms to realize faster delivery speed. Although these existing studies have developed a large number of predictive models and route optimization algorithms for instant delivery, there is a scarcity of

Table 1 | Factors related to OFD delays

Category	Factors	Impact mechanism	References
Endogenous	Order processing duration	Slow or suboptimal matching of orders to couriers, especially during peak demand, creates initial waiting time bottlenecks.	13,14
Endogenous	Merchant preparation time	Variability in kitchen capacity, order complexity, and in-house demand leads to courier waiting times at restaurants.	15
Endogenous	Pickup duration	Insufficient number of active couriers in a specific zone or at a specific time leads to longer order acceptance and pickup times.	18,56
Endogenous	Delay propagation	A delay in one order (e.g., late preparation) cascades to other co-batched orders or subsequent assignments for the same courier.	18
External	Spatio-Temporal Demand	High temporal volatility (peak hours) and spatial heterogeneity (demand hotspots) lead to system overload and resource imbalance.	57,58
External	Traffic congestion	Unpredictable road traffic, accidents, or closures are a primary driver of increased travel time and delivery uncertainty.	13,19
External	Weather conditions	Rain, snow, and fog reduce courier speed, increase travel risks, and can damage packages, causing significant delays.	19

empirical studies utilizing large-scale order data to systematically explore the variables related to delivery delays. This gap is especially salient in the context of OFD, where the expectations for timeliness are considerably higher. To date, few studies have conducted a comprehensive correlation analysis on the determinants of delivery delays in OFD services, let alone a causal analysis.

In prior literature, regression analysis is widely employed as a fundamental tool to explore the relationship between independent variables and the dependent variable, thereby identifying significant factors^{23,24}. Nevertheless, these approaches may lead to misleading spurious correlations, while the inherent inability to distinguish between mere associations and genuine causal pathways prevents the differentiation of direct and indirect effects within the regression analysis. To address these issues, recent studies began to adopt causal discovery and inference methods to infer causal relationships from observational data^{7,8,25,26}. Zhang et al.⁷ constructed a causal graph to model bus arrival delays and further applied the DML method to estimate the causal contribution of each variable. Their findings evidenced the delay propagation of bus arrivals at successive stops. Zhang et al.⁸ further replaced correlation-based models with causality-based frameworks and uncovered that delays at a bus stop tend to propagate along the route network, with highly connected stops exerting stronger influences on subsequent delays.

Causal analysis typically comprises causal discovery and causal inference. Causal discovery aims to uncover the underlying causal structure among variables using observational data. Classical approaches include constraint-based methods such as the PC (Peter-Clark) algorithm²⁷, score-based methods like the Greedy Equivalence Search (GES)²⁸, and functional causal models such as LiNGAM²⁹. While these methods have demonstrated effectiveness in discovering DAGs representing causal relationships, they often rely on strong assumptions such as causal sufficiency, acyclicity, or linearity, which may not hold in real-world settings. In addition, the performance of causal discovery methods is sensitive to sample size, noise, and latent confounders. Recently, advances in deep learning have led to the development of neural network-based approaches, including NOTEARS³⁰ and DAG-GNN³¹. These learning-based methods can reframe DAG discovery as a continuous optimization problem³⁰ and neural networks leverage non-linear activation functions and hierarchical representations to model intricate interactions³¹.

After causal discovery, causal inference methods estimate the effects of known causal relationships. Traditional techniques include regression with covariate adjustment, instrumental variable (IV) methods³², propensity score matching³³, and difference-in-differences³⁴. More recent developments integrate machine learning into causal inference, such as causal forests³⁵ and DML³⁶, which enhance estimation robustness and flexibility in high-dimensional settings.

Traditional causal analyses usually divide causal discovery and causal inference into two separate steps. Causal discovery methods first infer causal graphs from observed data, and then causal inference is carried out based on the known causal graphs. This separation can lead to missing information or mismatched model assumptions. For example, errors in the causal discovery stage may be passed on to the inference stage, affecting the accuracy of the results. To address these issues, this study developed a unified deep learning framework to learn both causal graphs and causal effects directly from OFD order data, avoiding the above issues. An end-to-end framework that combines Bayesian causal discovery with DML for inference was developed. This framework is particularly suitable for scenarios where causality is unknown and is therefore used in this study to explore the drivers of OFD delays and to quantify their importance.

Methods

Data description

This data was collected over a 1-week period in October 2022 from one of the major domestic OFD platforms in China (Meituan). The city of study, while anonymized due to a Non-Disclosure Agreement with the platform, is a major Tier-3 city in Northern China, positioned as a key

node within the expansive Beijing-Tianjin-Hebei (Jing-Jin-Ji) urban cluster. With a permanent population exceeding 9 million, the city's scale ensures the data captures significant metropolitan logistics complexity. The city's mature and competitive instant delivery market—with full coverage by all major platforms around its commercial centers, universities, and large residential areas—validates the dataset's capacity to represent the full spectrum of operational and behavioral challenges faced in modern urban OFD logistics. The empirical analysis utilizes a detailed dataset comprising 405,180 records of OFD orders. The dataset includes detailed order information such as order ID, courier ID, geographic coordinates of sender and recipient locations, and critical timestamps (e.g., order creation time, order dispatch time, order assignment time, courier acceptance time, estimated meal preparation time, courier pickup time, and actual arrival time). A sample of the data is shown in Table 2. In the data processing step, orders with extreme meal preparation time, courier pickup time, or transport time were excluded using the interquartile range (IQR) method³⁷. Outliers were defined as trips with times beyond $Q_3 + 3 \times \text{IQR}$ or below $Q_1 - 3 \times \text{IQR}$, where Q_1 and Q_3 are the 25th and 75th percentiles, respectively, and $\text{IQR} = Q_3 - Q_1$. After filtering, 405,180 valid order records remained. To identify delivery waves, we first sorted all orders accepted by each courier in a day according to acceptance time. Two consecutive orders were classified as belonging to the same wave if the acceptance time of the later order occurred before the delivery time of the earlier one. This criterion allows us to quantify the number of orders within each wave, which indirectly reflects the courier's workload³⁸. This operational definition is consistent with the platform's batch-dispatching practice, in which orders are dynamically bundled based on real-time factors such as demand density, courier proximity, and route compatibility³⁹. Because new tasks may be assigned before previous ones are completed, the resulting schedule within a delivery wave becomes interdependent. As a result, a delay in one order can be carried over to subsequent tasks, forming the structural basis for delay propagation.

The entire delivery process is divided into three main phases. Figure 1a reveals the proportion of total time occupied by each phase, with the transport duration accounting for 52.1%, the pickup duration for 32.7%, and the processing duration for 15.1%. This implies that strategies aimed at significantly reducing overall delivery delays should prioritize the transport and pickup durations. Figure 1b indicates that 16.7% of OFD orders, roughly one in every six, are delayed. This non-trivial delay rate underscores the persistent challenge of service timeliness in the OFD industry and highlights the importance of this study.

Figure 2a illustrates the temporal patterns of order volume, delay rate, and the number of available couriers. It is evident that the trends in order volume and courier availability are closely aligned, both displaying pronounced midday (12:00–13:00) and evening (18:00–19:00) peaks. Specifically, the midday and evening peaks experience the highest order volume, peaking at approximately 8000 orders, followed by a sharp decline after 20:00. The number of couriers similarly reaches comparable peaks around midday and evening, with roughly 4000 couriers active, but dramatically reduces overnight (from 00:00 to 06:00) to fewer than 500. The delay rate exhibits three distinct peaks throughout the day. It reaches 20% at noon and surpasses that level during the evening peak. Interestingly, the rate remains notably high during late-night hours (22:00–01:00), at approximately 17%. Figure 2b further illustrates the distribution of delay durations for overdue orders. While more than half of these orders exceed their deadlines by relatively short margins (under 4 min), nearly a quarter experience delays exceeding 6 min, with a notable 11.77% delayed by more than 8 min. These findings underscore a dual challenge, i.e., addressing both the high frequency of minor delays and the significant impact of less common, severe ones.

Variable selection

The dependent variable in this study is order delay, defined as the difference between the platform's estimated arrival time (communicated to the

Table 2 | OFD order samples

Order ID	Courier ID	Sender locations	Recipient locations	Order creation time	Order dispatch time	Order assignment time	Courier acceptance time	Estimated meal preparation time	Courier pickup time	Actual arrival time	Estimated arrival time
107524	1	115.443352, 38.884127	115.463891, 38.871188	2022/10/ 17 11:09	2022/10/ 17 11:09	2022/10/17 11:11	2022/10/17 11:11	2022/10/17 11:22	2022/10/ 17 11:23	2022/10/ 17 11:38	2022/10/ 17 11:40
125682	1	115.4479329, 38.892922	115.5157439, 38.881109	2022/10/ 17 11:10	2022/10/ 17 11:10	2022/10/17 11:15	2022/10/17 11:16	2022/10/17 11:25	2022/10/ 17 11:30	2022/10/ 17 12:01	2022/10/ 17 12:00
352887	2	115.507222, 38.892116	115.462797, 38.855911	2022/10/ 20 22:55	2022/10/ 20 23:01	2022/10/20 23:06	2022/10/20 23:07	2022/10/20 23:13	2022/10/ 20 23:15	2022/10/ 20 23:19	2022/10/ 20 23:20

consumer) and the courier’s actual arrival time. As for independent variables, order processing time, courier pickup, and transport time are selected. Considering that the meal preparation time varies across different orders (e.g., fast food and fine dining), this study also incorporates this variable. To investigate delay propagation effects, some attributes from the preceding order, including its phase durations and delay, are collected. Other factors include delivery distance, delivery speed, and the total number of orders within a delivery wave.

The dependent variable and independent variables are defined in Table 3. The k th order of courier i during his/her j th delivery wave is denoted as (i, j, k) . Considering that some variables may be interrelated, it is necessary to check for excessive multicollinearity among these independent variables to ensure the causal analysis produces valid and reliable results. Variance inflation factor (VIF) testing is a widely used method to detect multicollinearity⁴⁰. In general, explainable variables with a VIF greater than 10 should be excluded⁴¹. The VIF values for all independent variables are shown in supplementary information, and no variable had a value greater than 10. Therefore, all independent variables were placed into the subsequent causal analysis framework.

Causal analysis framework

Let $\mathbf{X} = X_1, X_2, \dots, X_n$ denote the set of independent variables, and Y denote the delivery delays. The dataset comprises observations $(\mathbf{x}^{(i)}, y^{(i)})$ for $i = 1, 2, \dots, m$, where $m = 405,180$ is the number of delivery orders associated with 4950 couriers. The goal of causal discovery is to infer the structure of \mathcal{G} from observation data $(\mathbf{x}^{(i)}, y^{(i)})$, capturing how each variable interacts with others and contributes to delays. In this DAG, nodes correspond to the variables $\mathbf{X} \cup \{Y\}$, and directed edges signify direct causal relationships. For example, an edge $X_j \rightarrow Y$ indicates that variable X_j directly drives the delivery delay Y . Once the causal graph \mathcal{G} is established, causal inference quantifies the total causal effect of each variable X_j on Y . The effect measures the expected change in Y when intervening on X_j while allowing other variables to adjust according to the causal relationships in \mathcal{G} . Using do-calculus, it can be expressed as $\mathbb{E}[Y|do(X_j = x_j)] - \mathbb{E}[Y|do(X_j = x'_j)]$, where x_j and x'_j are different values of X_j .

The integrated framework of causal discovery and causal inference is illustrated in Fig. 3, with the corresponding algorithms employed listed as follows:

- Causal discovery with a Bayesian approach. The study employs a Bayesian approach to learn the causal graph \mathcal{G} from the observational data. This approach utilizes deep neural networks to model complex, non-linear relationships among variables, making it well-suited for the intricate variables of OFD services. By fitting a structural causal model to the data, the Bayesian approach infers both the graph structure (edges) and functional dependencies, accounting for confounding factors that bias correlation analyses²⁶. The output is a DAG \mathcal{G} that maps how independent variables interact and influence Y .
- Causal Inference with DML. Using the learned causal graph \mathcal{G} , this step performs causal inference to estimate the total causal effect of each factor X_j on Y . DML is a method that helps estimate causal effects while controlling for high-dimensional confounding variables³⁶. It uses machine learning models to flexibly predict both the treatment and the outcome based on covariates. The core idea is to “partial out” the effects of confounders, and then estimate the causal effect using the residuals. DML applies cross-fitting to avoid overfitting and ensure valid inference. This approach is particularly useful when dealing with complex, nonlinear relationships in observational data^{7,8}.
- Comparison with correlation-based methods. To validate the results of the above causal analysis, this study compares them against the outputs of traditional correlation-based methods. After ranking variables by feature importance, some key variables are selected to generate partial dependence plots (PDPs) to analyze their non-linear relationships with delivery delays.

Fig. 1 | Statistics of OFD orders. **a** Decomposes the end-to-end delivery process into three sequential phases: processing duration, pickup duration, and transport duration, and reports the share of total delivery time contributed by each phase (transport: 52.1%, pickup: 32.7%, processing: 15.1%). **b** Shows the proportion of delayed versus on-time orders, where an order is defined as “delayed” if its actual arrival time is later than the platform-estimated arrival time; the overall delay rate is 16.7%.

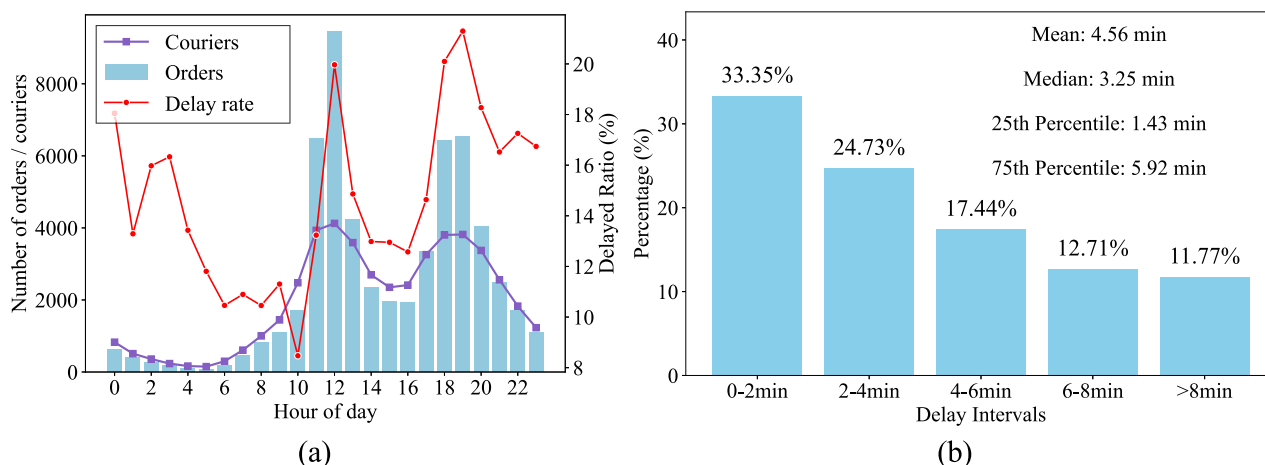
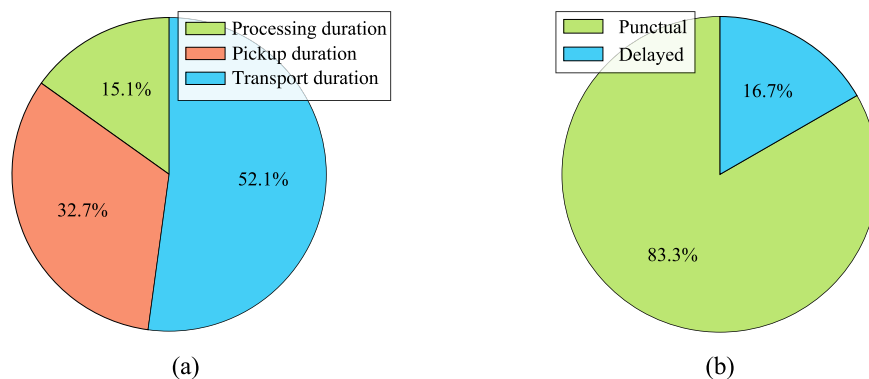


Fig. 2 | Ratio and distribution of delays. **a** Visualizes the hourly profiles of order volume, delay rate, and the number of active couriers across a day. **b** Reports the distribution of delay durations among overdue orders, showing that many delays are minor (e.g., under 4 min) while a non-trivial share are severe (e.g., over 6–8 min).

Table 3 | Definition of independent and dependent variables

Variables	Notation	Description
Dependent variables		
Delivery delay	$d_{i,j,k}$	For order (i,j,k) , it is defined as the difference between its actual arrival time and its estimated arrival time
Independent variables		
Processing duration	$t_{i,j,k}^1$	For order (i,j,k) , it is defined as the difference between its order creation time and its courier acceptance time.
Pickup duration	$t_{i,j,k}^2$	For order (i,j,k) , it is defined as the difference between its courier acceptance time and its courier pickup time.
Transport duration	$t_{i,j,k}^3$	For order (i,j,k) , it is defined as the difference between its courier pickup time and its actual arrival time.
Merchant preparation time	$p_{t_{i,j,k}}$	For order (i,j,k) , it is defined as the difference between its order creation time and its estimated meal preparation time.
Delivery distance	$dis_{i,j,k}$	The road network distance from the merchant to the customer.
Speeding	$s_{i,j,k}$	Binary, if courier i is speeding when delivering order (i,j,k) , it is 1; otherwise 0.
Delivery load	$l_{i,j,k}$	The total number of orders in the delivery wave of order (i,j,k) .
Previous processing duration	$t_{i,j,k}^{1'}$	For order (i,j,k) , it is defined as the length of the processing duration of its preceding order.
Previous pickup duration	$t_{i,j,k}^{2'}$	For order (i,j,k) , it is defined as the length of pickup duration of its preceding order.
Previous transport duration	$t_{i,j,k}^{3'}$	For order (i,j,k) , it is defined as the length of transport duration of its preceding order.
Previous delivery overtime	$o_{i,j,k}'$	Binary, if the preceding order of (i,j,k) exceeded its scheduled arrival time, it is 1; otherwise 0.
Previous delivery delay	$d_{i,j,k}$	For order (i,j,k) , it is defined as the delivery delay of its preceding order.
Lunch rush	$m_{i,j,k}$	Binary, if order (i,j,k) occurs in lunchtime, it is 1; otherwise 0.
Dinner rush	$n_{i,j,k}$	Binary, if order (i,j,k) occurs in dinner time, it is 1; otherwise 0.
Weekend	$w_{i,j,k}$	Binary, if order (i,j,k) occurs in the weekend, it is 1; otherwise 0.

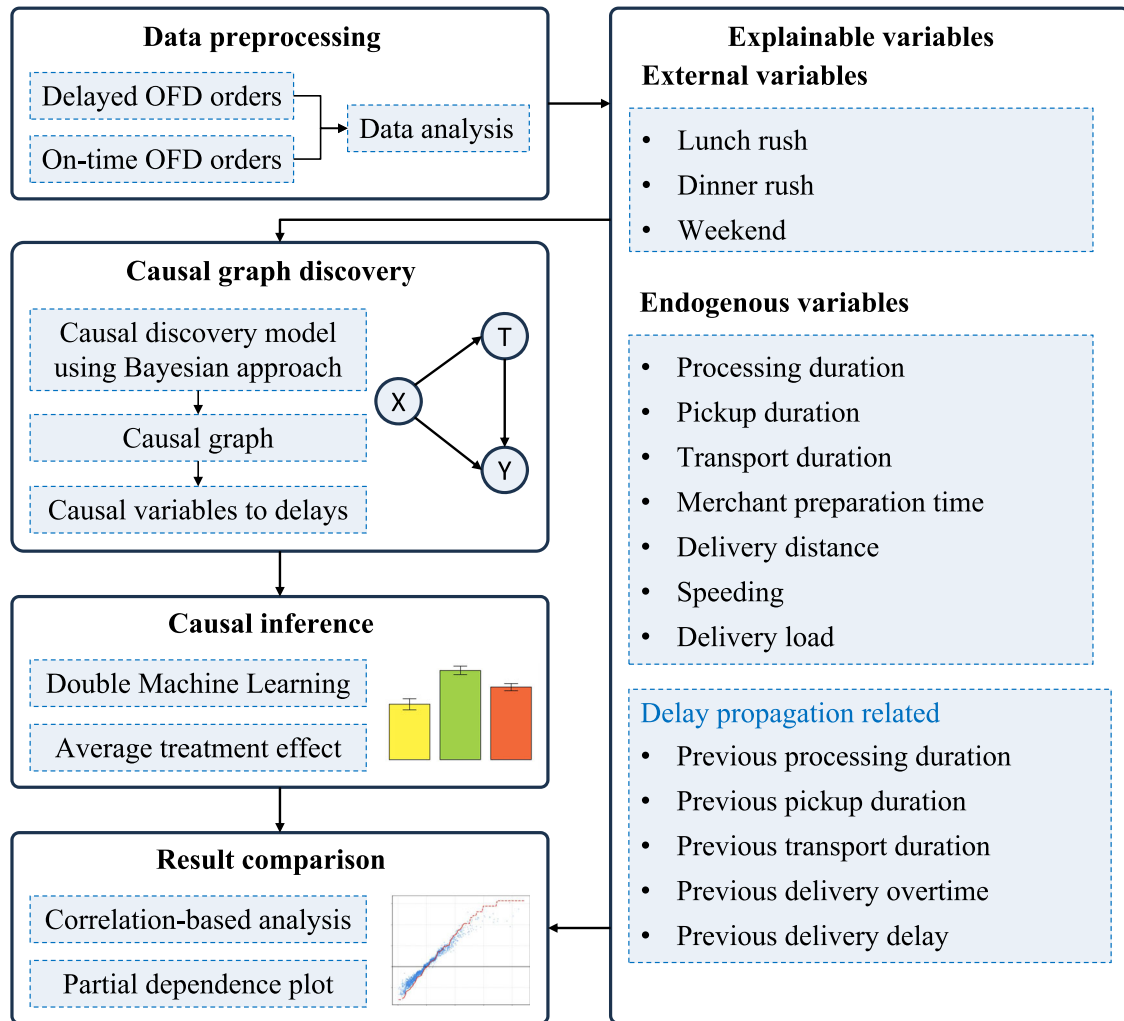


Fig. 3 | The framework of causal analysis. The figure summarizes the full workflow from data preprocessing and variable construction, to Bayesian causal graph discovery that yields a DAG, to causal inference using DML for estimating ATEs, and finally to result comparison with correlation-based analyses.

Causal graph discovery

To identify the causal structure among n variables $\mathbf{X} = X_1, X_2, \dots, X_n$, this study employs a Bayesian causal discovery approach. First, we adopt a structural equation model (SEM) to represent the causal data-generating process. An SEM expresses each variable as a function of its direct causes in the DAG, plus an exogenous noise term. In this study, each structural equation is parameterized as a non-linear additive noise model (ANM), a widely used functional form that ensures noise terms remain independent of parent variables and thereby supports causal identifiability. Specifically, the ANM defines how the data is generated, while the Bayesian approach defines how we learn the graph structure from that data.

Given a DAG \mathcal{G} with n nodes, ANM allows us to express the value of each variable X_i as a non-linear function of its parent variables plus an independent noise term:

$$X_i = f_i(\mathbf{X}_{pa(i;\mathcal{G})}) + z_i, \quad i = 1, \dots, n, \tag{1}$$

where $\mathbf{X}_{pa(i;\mathcal{G})}$ denotes the parents of node i in \mathcal{G} , f_i is a non-linear function (parameterized by neural networks in this study) capturing the causal dependency. z_i is an exogenous noise variable, with $z_i \perp z_j$ for $i \neq j$. In vector form, (1) becomes

$$\mathbf{X} = f_{\mathcal{G}}(\mathbf{X}) + \mathbf{z}, \tag{2}$$

where $f_{\mathcal{G}}$ respects the adjacency structure of \mathcal{G} . This distinct functional form of ANM provides the necessary structural constraints to distinguish the true causal graph from other statistically equivalent graphs.

Building upon this ANM structure, the Bayesian approach treats the causal graph \mathcal{G} as a latent variable and aims to learn its posterior distribution $p(\mathcal{G}|\mathbf{X})$ rather than a single fixed graph. This allows for a robust estimation of uncertainty. Assuming that there are m observations $\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^m$, we model the joint distribution of the observational data and the causal graph as

$$p_{\theta}(\mathbf{X}^1, \dots, \mathbf{X}^m, \mathcal{G}) = p(\mathcal{G}) \prod_{j=1}^m p_{\theta}(\mathbf{X}^j|\mathcal{G}), \tag{3}$$

where $p(\mathcal{G})$ is a prior over DAGs, $p_{\theta}(\mathbf{X}^j|\mathcal{G})$ is the likelihood of the data given the graph. Once (3) is fitted, $p_{\theta}(\mathbf{X}^1, \dots, \mathbf{X}^m, \mathcal{G})$ is a characterization of the causal structure. These parameters are gradually increased during training following an augmented Lagrangian scheme, ensuring only DAGs remain at convergence.

To ensure the graph \mathcal{G} is a DAG, the continuous acyclicity constraint proposed by³⁰ is adopted here:

$$h(\mathcal{G}) = \text{tr}(e^{\mathcal{G} \odot \mathcal{G}}) - n, \tag{4}$$

where $h(\mathcal{G}) \geq 0$ only if \mathcal{G} is a DAG, $\text{tr}(\cdot)$ stands for the trace of a matrix, \odot denotes the Hadamard product, n is the number of variables.

Eq. (4) is incorporated into the graph prior $p(\mathcal{G})$ as

$$p(\mathcal{G}) \propto \exp(-\lambda_s \|\mathcal{G}\|_F^2 - \rho h(\mathcal{G})^2 - \alpha h(\mathcal{G})), \quad (5)$$

where α means the proportional relationship. $\|\mathcal{G}\|_F^2$ is the squared Frobenius norm of \mathcal{G} . $h(\mathcal{G})$ is the DAG penalty function from (4). λ_s, ρ, α are scalar hyperparameters controlling the regularization strength, with λ_s controlling sparsity, and ρ, α (increased during training) ensuring acyclicity.

The observational likelihood $p_\theta(\mathbf{X}^j|\mathcal{G})$ can be derived from the ANM structure⁴². (2) can be rearranged as $\mathbf{z} = \mathbf{X} - f_{\mathcal{G}}(\mathbf{X})$. The components of \mathbf{z} are independent. If we have a distribution p_{z_i} for component z_i , then we can write the observational likelihood as:

$$p_\theta(\mathbf{X}^j|\mathcal{G}) = p_z(g_{\mathcal{G}}(\mathbf{X}^j; \theta)) = \prod_{i=1}^n p_{z_i}(g_{\mathcal{G}}(\mathbf{X}^j; \theta)_i) \quad (6)$$

where the Jacobian determinant is unity for DAGs, simplifying the expression. n is the number of variables. p_{z_i} denotes the noise distribution, and this study considers a Gaussian distribution $p_{z_i}(\cdot) = \mathcal{N}(\cdot|0, \sigma_i^2)$. If there is no edge $j \rightarrow i$, the i th component of the output of $f_{\mathcal{G}}(\mathbf{X})$, $f_i(\mathbf{X})$, must satisfy $\partial f_i(\mathbf{X})/\partial X_j = 0$. A flexible parameterization is proposed to represent $f_i(\mathbf{X})$:

$$f_i(\mathbf{X}) = \zeta_i \left(\sum_{j=1}^n \mathcal{G}_{j,i} \ell_j(X_j) \right), \quad (7)$$

where $\mathcal{G}_{j,i}$ indicates the presence of the edge $j \rightarrow i, i, j = 1, 2, 3, \dots, n, \zeta_i$ and ℓ_j are multi-layer perceptrons (MLPs) with shared weights across nodes, modulated by node-specific embeddings. For more details about ζ_i and ℓ_j , refer to²⁶.

Considering that the true posterior over $p_\theta(\mathbf{X}^1, \dots, \mathbf{X}^m, \mathcal{G})$ in (3) is intractable, this study introduces a variational distribution $q_\phi(\mathcal{G}) = \prod_{j,i} \text{Bernoulli}(\phi_{j,i})$ and maximizes the evidence lower bound (ELBO):

$$\begin{aligned} \text{ELBO}(\theta, \phi) &= \mathbb{E}q_\phi(\mathcal{G})[\log p(\mathcal{G}) + \sum_{j=1}^m \log p_\theta(\mathbf{X}^j|\mathcal{G})] \\ &+ H(q_\phi) \leq \log p_\theta(\mathbf{X}^1, \dots, \mathbf{X}^m, \mathcal{G}). \end{aligned} \quad (8)$$

where $H(q_\phi)$ represents the entropy of the distribution $q_\phi(\mathcal{G})$ and $p_\theta(\mathbf{X}^j|\mathcal{G})$ is the form of (6). $q_\phi(\mathcal{G}) = \prod_{j,i} \text{Bernoulli}(\phi_{j,i})$ is the product of independent Bernoulli distributions for each potential directed edge in \mathcal{G} . The edge existence and edge orientation are parametrized separately, using the efficient neural causal discovery parametrization⁴³. The SEM parameters θ and variational parameters ϕ are trained by maximizing (8), the evidence lower bound (ELBO). The Gumbel-softmax trick facilitates the stochastic gradient estimation for ϕ , the full optimization procedure can refer to²⁶. It has been proven that maximizing the ELBO in (8) can recover the true causal graph \mathcal{G}_0 in the infinite data limit.

Causal inference

After obtaining the true causal graph \mathcal{G}_0 , this study adopts the DML method to conduct causal inference. The method offers a robust solution by leveraging flexible machine learning methods to estimate nuisance parameters while maintaining valid inference on the parameter of interest³⁶. Without loss of generality, suppose \mathcal{G}_0 has n variables $\mathbf{X} = X_1, X_2, \dots, X_n$, where $Y \in \mathbf{X}$ denotes the outcome variable, $T \in \mathbf{X}$ denotes a treatment variable (a causal variable directly connected to Y in \mathcal{G}_0), $\mathbf{W} \subseteq \mathbf{X}$ represents the vector of confounders identified by the causal graph \mathcal{G}_0 (variables indirectly connected to Y in \mathcal{G}_0). The goal is to estimate the ATE of T on Y , defined as:

$$\tau = \mathbb{E}[Y|\text{do}(T = t_1)] - \mathbb{E}[Y|\text{do}(T = t_0)] \quad (9)$$

where t_1 and t_0 are specific values of the treatment variable, and $\text{do}(\cdot)$ denotes the do-operator from causal inference, representing an intervention on T .

The DML approach combines machine learning flexibility with statistical rigor to estimate τ . The steps are as follows. A partially linear model is first adopted to express the effect of the treatment T on the outcome Y :

$$Y = \tau T + g_0(\mathbf{W}) + \epsilon_Y, \mathbb{E}[\epsilon_Y|T, \mathbf{W}] = 0, \quad (10)$$

$$T = m_0(\mathbf{W}) + \epsilon_T, \mathbb{E}[\epsilon_T|\mathbf{W}] = 0, \quad (11)$$

where τ is the ATE to estimate. $g_0(\mathbf{W})$ captures the effect of covariates \mathbf{W} on Y . $m_0(\mathbf{W})$ models the relationship between \mathbf{W} and T . ϵ_Y and ϵ_T are error terms, assumed to be mean-independent of T and \mathbf{W} .

Then, this study uses flexible machine learning methods (e.g., random forests, gradient boosting framework) to estimate $g_0(\mathbf{W})$ and $m_0(\mathbf{W})$. In this study, the XGBoost model is adopted to estimate them and utilizes Optuna to automate and optimize the hyperparameter tuning process of the XGBoost model⁴⁴. XGBoost is a powerful machine learning method known for its capacity to handle complex, nonlinear relationships via boosted trees. More information about XGBoost can be referenced in ref. 45. We regress T on \mathbf{W} to obtain $\hat{m}(\mathbf{W})$, and regress Y on \mathbf{W} to obtain $\hat{g}(\mathbf{W})$. These estimates are "nuisance parameters" because they are not the primary focus but are necessary for isolating τ . With (10), we further compute residuals to remove the influence of \mathbf{W} .

$$\tilde{Y} = Y - \hat{g}(\mathbf{W}), \quad (12)$$

$$\tilde{T} = T - \hat{m}(\mathbf{W}). \quad (13)$$

where these residuals represent the components of Y and T unexplained by \mathbf{W} .

τ can be estimated by regressing the residualized outcome \tilde{Y} on the residualized treatment \tilde{T} :

$$\hat{\tau} = \left(\frac{1}{m} \sum_{j=1}^m \tilde{T}_j^2 \right)^{-1} \left(\frac{1}{m} \sum_{j=1}^m \tilde{T}_j \tilde{Y}_j \right). \quad (14)$$

where m is the number of observations. This is essentially a least-squares estimator applied to the residuals, yielding an unbiased estimate of the ATE under the DML assumptions.

To perform statistical inference on $\hat{\tau}$, compute its variance:

$$\hat{\sigma}^2 = \frac{1}{m} \left(\frac{1}{m} \sum_{j=1}^m \tilde{T}_j^2 \right)^{-2} \sum_{j=1}^m \tilde{T}_j^2 (\tilde{Y}_j - \hat{\tau} \tilde{T}_j)^2. \quad (15)$$

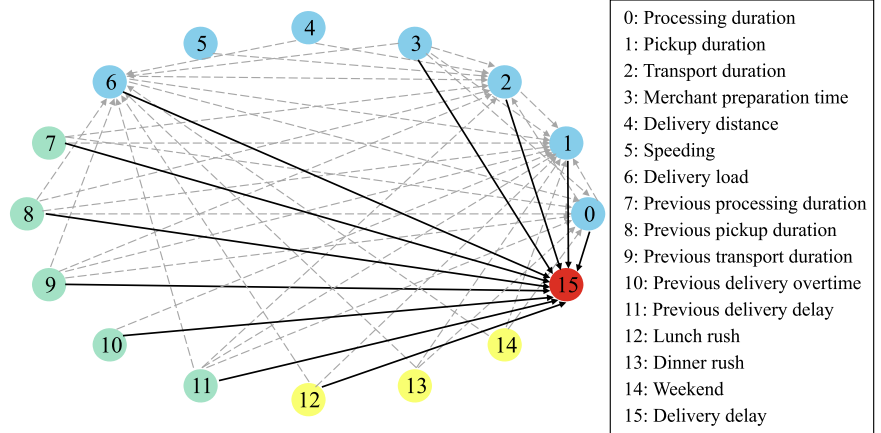
This variance accounts for the uncertainty in estimating the nuisance parameters. Confidence intervals for $\hat{\tau}$ can then be constructed as:

$$\hat{\tau} \pm z_{\alpha/2} \sqrt{\hat{\sigma}^2} \quad (16)$$

where $z_{\alpha/2}$ is the critical value from the standard normal distribution for a desired confidence level, which is set to 95% in this study.

As an end-to-end framework, the DML method is integrated with causal graph discovery. First, the causal graph \mathcal{G}_0 informs the DML by identifying T as a variable directly connected to Y . Then, specifying \mathbf{W} as all variables directly connected to Y except T , ensuring proper control for confounding. By combining causal graph discovery with DML, this study provides a robust framework for causal inference. The DML method estimates the ATE τ efficiently, using machine learning to handle complex relationships while grounding the analysis in the true causal structure of \mathcal{G}_0 . This methodology is particularly valuable for observational data, where confounding and non-linearity are prevalent.

Fig. 4 | Causal graph on delivery delays. The DAG contains 16 nodes, with the red node representing the dependent variable (current order delivery delay) and the remaining nodes representing candidate drivers. Node colors distinguish variable types: blue nodes denote current order operational characteristics (e.g., pickup and transport durations), green nodes denote delay-propagation factors from the preceding order, and yellow nodes denote external factors (e.g., rush hours and weekends). Directed edges indicate the inferred causal direction from cause to effect.



Results

Causal graph on delivery delays

Using the aforementioned causal discovery method, a DAG was constructed to elucidate the causal relationships underlying delivery delays, as shown in Fig. 4. The DAG consists of 16 nodes. The first 15 nodes represent different types of independent variables. Specifically, blue nodes correspond to variables about current order characteristics, such as the lengths of pickup and transport durations. Green nodes capture variables related to delay propagation, while yellow nodes represent external variables, such as rush hours and weekends. The red node denotes the dependent variable, the delay of the current order. These nodes were interconnected by directed edges, with arrows indicating the causal direction from cause to effect. Solid black arrows signify direct causal influences on delivery delays, while dashed grey arrows denote indirect effects.

According to Fig. 4, the durations of processing, pickup, and transport are causes of delays. A change in these durations directly influences the delay risk. For example, an extended order processing duration delays the start of pickup and transport, increases consumer waiting times, and compresses the buffer time for subsequent phases, ultimately leading to arrival delays. Additionally, merchant preparation time is also a driver of delays. One possible explanation is that longer preparation times likely lead to extended courier waiting at the restaurant, which in turn reduces the available time for transport and increases delay probability. Delivery load, the number of orders within the same delivery wave, also has a causal relationship with delays. It is possible that the number of orders determines the complexity of route planning and detour time. This effect is compounded by prior findings that a high delivery load raises couriers' psychological stress⁴⁶, thus negatively affecting their efficiency in time allocation and route optimization⁴⁷.

Figure 4 further offers compelling evidence of delay propagation in OFD, where a delay in one order directly impacts subsequent orders within a courier's delivery wave. Both the occurrence of a prior delay (Node 11) and its duration (Node 10) are identified as significant causal drivers for delays of subsequent orders. The underlying mechanism may be that an initial delay disrupts the courier's schedule and operational efficiency, consuming the time buffer allocated for subsequent deliveries and thus increasing their likelihood of also being late. This cascading effect is particularly pronounced under adverse conditions, such as heavy rain, where orders later in a delivery sequence often experience severe overruns⁴⁸. In addition, delays originating in the processing (Node 7), pickup (Node 8), or transport (Node 9) phase of a preceding order significantly disrupt the schedule and arrival time for the next one. These effects may create a domino effect, i.e., an initial delay during the pickup of a preceding order pushes back the courier's arrival at the next restaurant, which in turn triggers a longer pickup duration for the current order and erodes the timeline for its transport duration, ultimately leading to arrival delays.

As for external variables, Fig. 4 indicates a unique causal link between the lunch rush and delays, which is not observed for the evening rush or

weekend period. This distinction likely stems from demand patterns. As shown in Fig. 2a, demand surges during the lunch rush hour from 11:00 to 12:00, then drops rapidly. This intense and concentrated demand strains the meal-preparation time and courier workload in a very short time, thus elevating the risk of delays. In contrast, demand during evening and weekend periods tends to be more distributed over a longer duration (from 17:00 to 20:00), providing greater time flexibility that appears to absorb the pressure and prevent a direct causal impact on delays.

In addition to direct effects, some variables indirectly affect delivery delays. Figure 4 reveals that delivery distance, speeding, dinner rush, and weekend variables lack direct causal links to delays but exert significant influence by altering delivery load and processing, pickup, and transport durations. The increase or decrease of delivery distance can constrain courier delivery load in a short time and, in turn, impact delivery delays (path: Node 4 → Node 6 → Node 15). A longer delivery distance also prolongs the transport duration, increasing the likelihood of delays (path: Node 4 → Node 2 → Node 15). Conversely, speeding can partially offset delays by reducing the transport duration (path: Node 5 → Node 2 → Node 15). Furthermore, both the evening rush and weekend periods trigger a surge in delivery load, which indirectly extends the processing and pickup durations, ultimately exacerbating delays (paths: Node 13/Node 14 → Node 0/Node 1 → Node 15). Mechanistically, these indirect variables influence delays by modulating other causal variables. Therefore, despite lacking a direct causal link, they play a critical role in determining final delay outcomes via these intermediate pathways.

Causal contribution to delivery delays

This study further utilizes the DML approach to quantify the causal contributions of variables that have a direct impact on delivery delays, as shown in Fig. 5. The results are expressed as ATE values, which quantify the average causal effect of each variable on delivery delay, i.e., the expected difference in average delivery delay if we were to "switch" the variable from its control level to its treatment level while holding all other factors constant. Additionally, the ATE estimates are accompanied by 95% confidence intervals, with their narrow widths indicating high reliability.

As illustrated in the Fig. 5, most variables exhibit positive ATEs, indicating that their increase tends to prolong delays. The pickup duration has the highest ATE (nearly 4 min). This may stem from uncertainties like merchant cooking and packaging times, which can consume significant buffer and amplify subsequent delays. The transport duration ranks second (about 3 min). With longer transport distances and times, couriers may face detours and unexpected congestion, particularly during peak hours or in city-center areas, which further exacerbate their late arrivals. Ranked third, the delay of the preceding order reveals a strong propagation effect, where the late completion of one task compresses the timeline for the next (less than 2 min), creating a cascade of delays. According to Fig. 1a, although the

Fig. 5 | Causal contribution to delivery delays. The figure reports ATE estimates for variables that have direct causal links to delivery delay in the discovered graph. Each effect quantifies the expected change in delivery delay under an intervention that shifts the focal variable from a control level to a treatment level while holding other factors appropriately controlled, and 95% confidence intervals are shown to reflect estimation uncertainty.

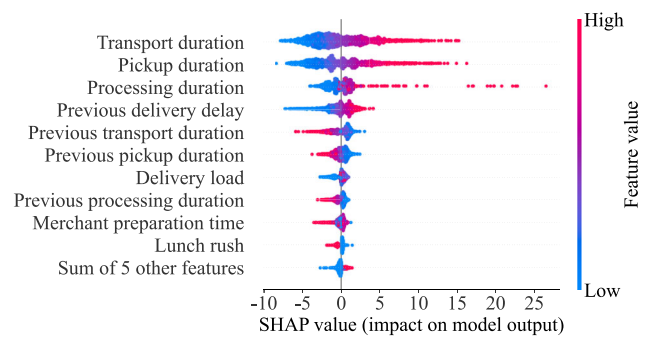
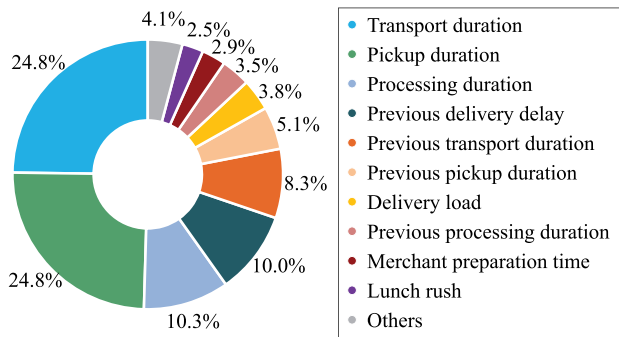
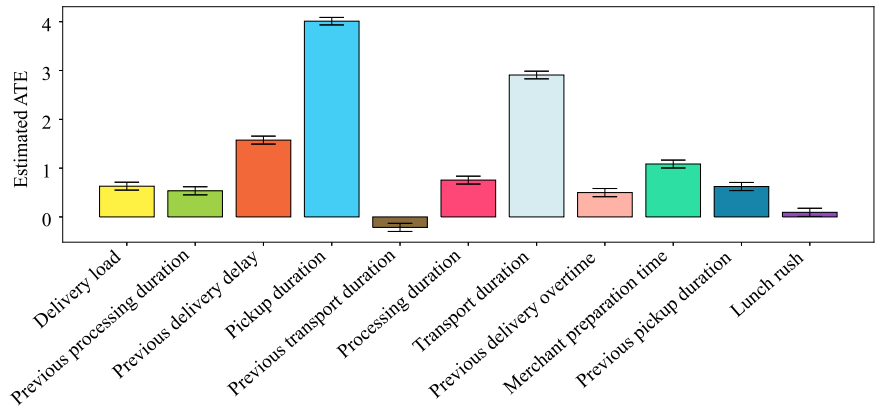


Fig. 6 | Correlation-based importance ranking and SHAP explanations. **a** Presents the relative feature-importance ranking from the correlation-based predictive model, interpreted as each variable’s relative contribution to predicted

delays. **b** Shows SHAP beeswarm diagrams, where each point corresponds to an order-level SHAP value; the *x*-axis is the SHAP contribution to predicted delay, and color encodes the feature value from low to high.

transport duration accounts for a larger share of total delivery time (52.1% vs. 32.7%), the pickup duration exerts a stronger causal influence on the formation of delays. This finding suggests that the pickup stage, despite being shorter, is the most critical point of leverage for delay mitigation. Other factors like delivery load and the midday peak show positive but more modest ATEs (less than 1 min). In addition, the previous Transport duration has a minimal negative ATE. This may be because when the platform detects that a courier is running late, it will reallocate their subsequent orders or extend the deadlines to mitigate the delivery pressure and ensure safety^{5,49}.

Comparison with correlation-based methods

This section compares the variable importance rankings generated by our proposed causality-based framework and a commonly used correlation-based method. For the latter, this study selected the CatBoost model, which is highly effective at capturing the complex, non-linear relationships and intricate patterns in data that traditional regression models might overlook^{11,50}. To interpret the model’s output and rank variable importance, we integrated CatBoost with Shapley Additive Explanations (SHAP), a prominent explainability method, following the approach of recent studies^{7,12}. Further details on SHAP value calculation are available in supplementary information.

Figure 6 displays the variable importance ranking of the correlation-based method. It is obvious to observe that both methods identify the lengths of the pickup and transport durations as the two most critical factors. The figure shows their contributions are almost identical, 25.2% and 25.1%, respectively. Furthermore, both methods rank the delay of the preceding order as the third most significant factor, confirming the strong effect of

delay propagation. Regarding the direction of effects, the positive SHAP values for these top variables align perfectly with their positive ATEs, confirming their role in increasing delays.

As for the ranking of merchant preparation time, the proposed causality-based analysis identifies it as the fourth most influential factor, whereas the SHAP analysis assigns it minimal importance (3.0%). This discrepancy likely arises because correlation-based methods struggle to disentangle the interdependence between merchant preparation and courier pickup time, leading to an underestimation of the former’s true impact. In addition, the two methods report contradictory effects for the previous pickup duration. SHAP values suggest a negative correlation (a longer previous pickup is associated with a shorter current delay), while its ATE indicates a longer previous pickup causes a longer current delay. These contradictions powerfully demonstrate that even sophisticated, interpretable machine learning models may produce findings that misrepresent a variable’s true causal impact. This stems from the fundamental distinction between correlation and causation. Correlation-based methods capture the net outcome of a complex web of interactions, which can mask, or in some cases even reverse, the underlying causal contribution of a variable.

Discussion

The causal analysis in Section Causal contribution to delivery delays identified the pickup and transport durations, as well as delay propagation factors, as the most dominant drivers of delays (exhibiting the highest ATEs). However, understanding the average effect alone is insufficient for operational optimization. It is crucial to determine how the risk of delay varies as these variables increase and to identify specific tolerance thresholds. To this end, this study employs PDPs to visualize the marginal effects of

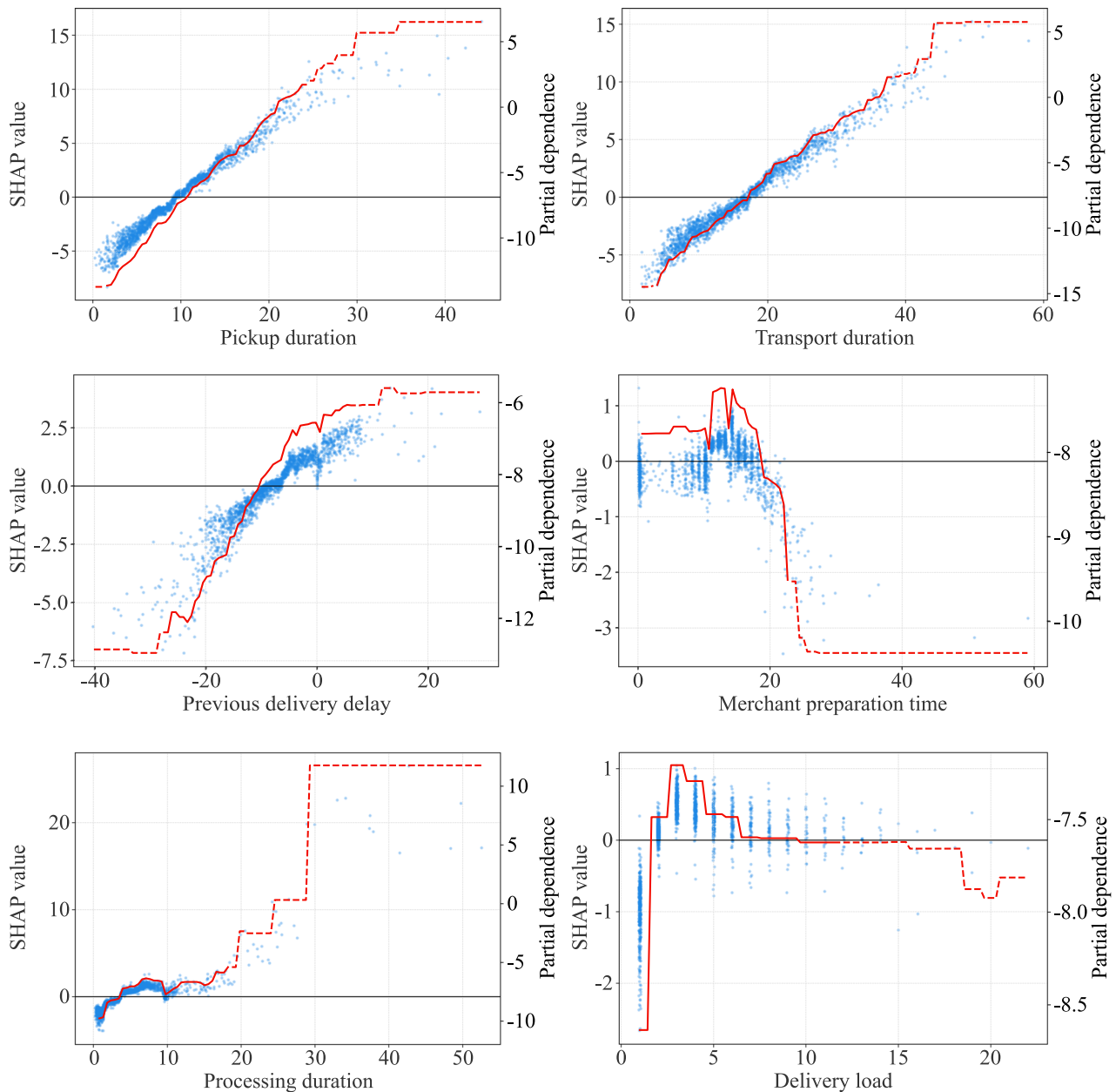


Fig. 7 | PDPs of key variables. The figure visualizes the marginal effects of the top causal variables on delivery delay using PDPs. In each panel, the PDP curve summarizes the model-averaged relationship between the focal variable and predicted delay while averaging over the distribution of other features; SHAP points are

overlaid to reflect the distribution and heterogeneity of effects across observations. Interpretation focuses on the central data range (1st–99th percentile) to avoid extrapolation from extreme values.

the top six causal variables on delivery delays^{7,51}. Methodologically, the PDP estimates the relationship between a specific feature and the outcome by averaging the model’s predictions over the distribution of all other variables, thereby isolating the feature’s influence from confounding interactions. Details of the PDP calculation are provided in supplementary information. In Fig. 7, the PDP for each variable is depicted by a red curve, overlaid with SHAP values (blue points) to corroborate the distribution of effects. The analysis focuses on the solid lines representing the core data range (1st–99th percentile) to ensure the reliability of the interpreted non-linear patterns¹¹.

The PDP of pickup duration exhibits an approximately linear trend with a positive slope. The curve crosses the zero-delay threshold at approximately the 10-min mark. This finding suggests that the pickup duration exceeding 10 min becomes a significant and consistently growing contributor to delays. A similar linear trend is observed for the transport

duration. The zero-crossing point for the PDP curve is around 17 min. This means the transport duration above this threshold is likely to increase delays.

The PDP for the previous order’s delay also shows a clear upward trend, indicating delay propagation. The curve crosses zero at approximately –10 min. This suggests that completing the preceding order about 10 min early is typically sufficient to neutralize, on average, its causal contribution to the next order’s delay. Notably, even when the previous delivery is on time (delay = 0), the SHAP value remains strongly positive (≈ 2.5), indicating persistent time pressure and tight scheduling margins within delivery waves.

A nonlinear, hump-shaped trend is observed for merchant preparation time. The effect on delay is close to zero for the first 10 min, but it rises sharply to a peak between 10 and 15 min. This peak is likely due to

prolonged courier waiting at restaurants, which delays their subsequent deliveries. Interestingly, beyond 15 min, a downward trend is observed. This suggests that for exceptionally long preparation times (e.g., made-to-order bakery items), the platform may extend the courier's delivery window or adjust the dispatch schedule to reduce delay risk⁵².

The PDP for the processing duration is nearly flat, exhibiting only a slight positive slope. This indicates that variations of the variable only have a minimal, non-critical causal impact on delays. Regarding the delivery load (number of orders in a delivery wave), its PDP is also hump-shaped. The delay peaks significantly at three orders and then declines. This suggests that platforms may deploy targeted strategies (e.g., better routing, higher incentives) to prevent delays when a courier's load exceeds this threshold.

Based on the above non-linear effects and the ATE quantified in Section Causal contribution to delivery delays, we propose targeted operational strategies and policy implications to mitigate delays effectively. Regarding the delay propagation phenomenon, some platforms have issued operational guidelines to detect delivery anomalies. For instance, algorithms often check if a new assignment causes overtime for existing orders and provide order reassignment when requested by couriers^{39,53}. However, these measures are often reactive. They only address delays after they have manifested or rely on manual courier inputs. In contrast, our causal analysis identifies the quantitative thresholds and structural sources of delay propagation embedded within delivery waves. Therefore, the following recommendations aim to transition from reactive mitigation to proactive prevention:

Optimizing pickup and merchant preparation. The PDP analysis indicates that delay risks increase sharply when pickup durations exceed 10 min and when merchant preparation times fall within the 10–15 min range. Platforms should therefore deploy real-time preparation prediction systems to align courier arrivals with meal readiness and reduce idle waiting. For orders whose estimated merchant preparation time surpasses 10 min, delivery promises should be proactively adjusted, and courier dispatch schedules should be modified to reduce the potential delay and prevent it from propagating to subsequent tasks.

Mitigating delay propagation. Our PDP analysis indicates that a previous order must be completed more than 10 min early to neutralize delay risks for subsequent tasks. To mitigate this propagation effect, platforms should continuously monitor courier progress. If a courier is projected to finish a current task with less than a 10-min buffer before the estimated arrival time, the algorithm should automatically inject dynamic slack time into the downstream orders. Where such adjustments remain insufficient, downstream orders may need to be reassigned, effectively severing the propagation chain before delays accumulate.

Managing delivery load. Delivery load exhibits a hump-shaped effect, with delay risk peaking when couriers handle three orders in a delivery wave. To mitigate this, platforms may adopt dynamic order caps that temporarily restrict new assignments once a courier reaches a high load, except in cases where real-time route conditions are feasible or spatial compatibility with the existing trajectory. We acknowledge that this strategy may partially hinder couriers' earnings. However, it also has potential economic benefits for both platforms and couriers. On the one hand, once delays occur, platforms may reduce a courier's maximum orders in a delivery wave or even temporarily suspend their delivery tasks⁵⁴. On the other hand, most OFD platforms compensate customers when orders exceed specified delay thresholds. These compensations increase with the duration of the delay (e.g., 2 RMB for less than 10 min, 4 RMB for 10–20 min, and 6 RMB for above 20 min)⁵⁵. Thus, we propose that dynamic capping can maintain courier motivation and protect service quality.

Implementing dynamic wait-time compensation and penalties. ATE results show that the pickup duration is the most significant contributor to delivery delays. On average, an increase of 1 min in pickup duration leads to an increase of approximately 4 min in final delivery delay. To tackle the paramount issue of pickup duration, platforms could implement automatic "paid waiting time" for couriers that kicks in after a reasonable grace period (e.g., 5 min) at the merchant, accurately compensating them for delays. Conversely, merchants consistently causing excessive wait times (e.g., above

10 min on more than 15% of weekly orders) should face temporary visibility throttling on the app during peak hours until their metrics improve.

Optimizing transport duration. ATE results indicate that transport duration is the second dominant contributor to delivery delays. On average, an increase of 1 min in transport duration results in an increase of about 2.9 min in final delivery delay. Optimizing transport duration requires more intelligent routing and batching. Platforms should integrate real-time traffic data to flexibly calibrate routing algorithms, systematically applying data-driven buffer times during known congestion periods. Rather than static estimates, these buffers should be dynamically calculated based on historical delay probability distributions. Regarding order batching, while it offers cost efficiencies, it often prolongs delivery for at least one customer. A hard constraint should be imposed: a second order cannot be batched to a courier if it adds more than 8 min of additional transport time to the first customer's predicted arrival time, balancing efficiency with customer experience.

These strategies provide a path toward a more sustainable OFD ecosystem. Regarding social sustainability, dynamic order caps help reduce the physical and mental strain on couriers. By decoupling earnings from tight deadlines, these measures decrease incentives for risky behaviors like speeding, thereby enhancing public road safety. In terms of economic sustainability, shifting from reactive compensation to proactive mitigation preserves platform margins by minimizing delay-triggered payouts, which typically range from 2 to 6 RMB. This approach encourages merchants to improve operational efficiency and ensures long-term economic viability for all stakeholders.

To further validate the practical impact of the identified causal drivers and quantify the potential benefits of the proposed strategies, we conducted counterfactual simulations based on the observational dataset. Specifically, in each simulation scenario, we applied a systematic time shift $\Delta \in \{0, 1, 2, 3\}$ min to specific timestamps (estimated meal preparation timestamp, courier pickup timestamp, and actual arrival timestamp in Table 2) in the historical data. In addition, these simulations account for delay propagation. A delay incurred in an earlier order (e.g., the first in a wave) delays the arrival times of all subsequent orders in the same delivery wave. We then calculated the average increase in delivery delay across all orders compared to the baseline ($\Delta = 0$). The three counterfactual simulations are as follows.

Pickup deferral. This scenario simulates delays occurring specifically during the courier's pickup process. We uniformly extend the actual courier's pickup timestamp by Δ minutes while holding other phase durations constant. This shift propagates forward, delaying the arrival times of the current order and the subsequent orders in the wave.

Transport deferral. This scenario simulates changes in the transport duration. We uniformly extend the actual arrival timestamp by Δ minutes. This directly pushes back the arrival time for the current order and, through the wave linkage, propagates the delay to the courier's subsequent orders.

Meal preparation deferral. This scenario simulates the impact of merchant efficiency by extending the estimated meal preparation time by Δ minutes. Unlike the other scenarios, the impact here is conditional on the interaction between the courier and the merchant. If the courier originally arrived before the food was ready, this shift directly prolongs the courier's wait time, thereby delaying the actual pickup timestamp. If the food was originally ready well before the courier arrived (i.e., the courier was the bottleneck), the additional preparation time is absorbed by this buffer. In this case, the shift does not necessarily delay the actual pickup timestamp.

The simulation results, stratified by the baseline duration of each phase (bins: [0, 10], [10, 20], and 20+ min), are presented in Fig. 8. The simulation results closely echo the findings from the ATE values and PDP analysis. Consistent with the highest ATE observed for the pickup duration, pickup deferral exhibits the most drastic sensitivity. We observe that a mere 1-min shift ($\Delta = 1$) results in an average delay increase ranging from 2.62 to 4.06 min across bins. This magnitude is remarkably consistent with the ATE value (≈ 4 min) quantified in Section Causal contribution to delivery delays, confirming the robustness of our DML estimates. Moreover, the impact of these shifts amplifies non-linearly with the baseline duration. As shown in Fig. 8a, for orders with already long pickup durations (10+ min), a 3-min

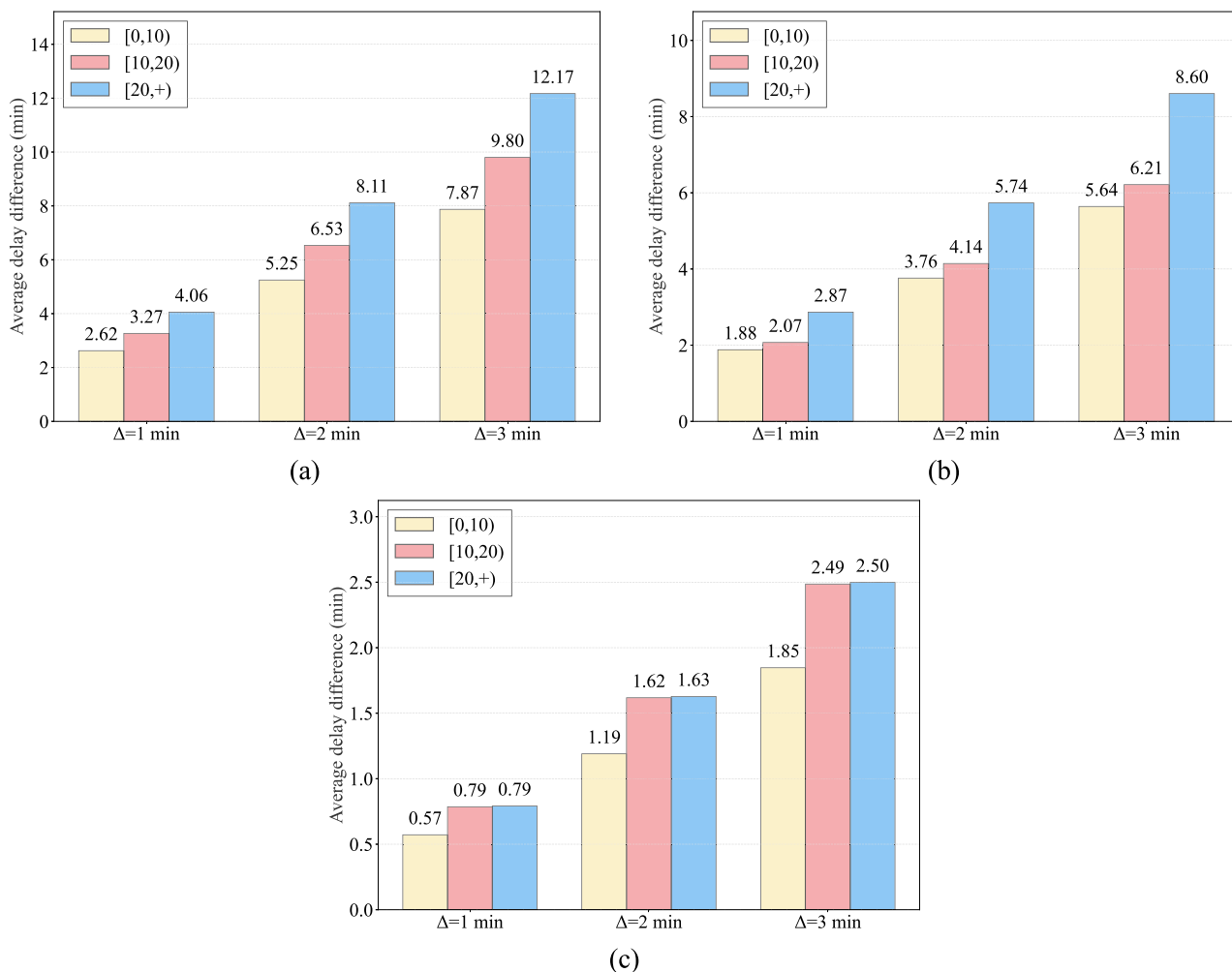


Fig. 8 | Results of counterfactual simulations. The figure reports simulation outcomes where specific timestamps are systematically shifted by $\Delta \in \{1, 2, 3\}$ min to emulate delays arising in a pickup, b transport, or c meal preparation. Simulations explicitly account for delay propagation within a delivery wave, so a delay in an

earlier order can shift the arrival times of subsequent orders handled by the same courier. Results are stratified by baseline phase duration bins ([0, 10), [10, 20), 20+ min) and summarize the average increase in delivery delay relative to the baseline ($\Delta = 0$).

shift triggers a disproportionate delay increase of nearly 10 min or even more, which is higher than the 7.87-min increase observed for pickups below 10 min. This finding echoes our PDP analysis, which showed that delay risk escalates sharply once pickup duration exceeds the critical 10-min threshold.

Figure 8b confirms that transport deferral also translates into arrival delays, with a $\Delta = 3$ min shift resulting in an average delay increase of approximately 5–8 min. This validates the substantial ATE (≈ 3 min) found for the transport duration. Similarly, the amplification effect is more pronounced for orders with longer transport durations, where the same shift results in substantially larger delays. Interestingly, the shift on merchant preparation time shows a relatively dampened impact compared to the others. As shown in Fig. 8c, a 3-min shift in preparation time results in a smaller increase in final delay (approximately 2.5 min). This suggests that the system has some natural buffering capacity against preparation variance (e.g., courier travel time absorbs some waiting), but once the courier is involved (pickup duration), the buffer evaporates. Overall, these simulation results could provide quantitative support for our policy recommendations.

Conclusions

This study thoroughly investigated the true causal drivers of delays in the OFD industry by developing and applying a novel causality-based framework. Leveraging a large dataset of 405,180 OFD orders from a major city in China, the framework integrates a deep Bayesian causal discovery model

with DML for causal inference, allowing us to disentangle true drivers from mere correlations and quantify each factor’s contribution through the ATE. The key findings and contributions are summarized as follows.

First, we empirically confirmed that the pickup and transport durations are the most significant causal drivers of delays, with ATE exceeding 2 min. While correlation-based methods (SHAP) also highlight these factors, our causal framework provides a more accurate quantification by controlling for confounding variables. Second, a novel contribution of this study is the quantification of delay propagation within OFD services. We demonstrated that both the occurrence and duration of a preceding order’s delay significantly exacerbate delays for subsequent orders, creating a systemic “domino effect” previously established only in public transit literature^{7,8}. Building upon the causal analysis, this study further uses PDPs to reveal critical non-linear thresholds for operational intervention. Specifically, we identified a “10-min critical window” for merchant preparation and pickup durations, beyond which delay risks escalate sharply, and a “hump-shaped” impact of delivery load, where risks peak at three orders per delivery wave. Finally, by benchmarking against traditional CatBoost-SHAP models, we demonstrated that the proposed causal framework effectively corrects for misleading correlations, such as the underestimation of merchant preparation time delay impact, providing a more robust basis for policy-making.

This study is subject to several limitations that warrant careful consideration. First, regarding data generalizability, the dataset is confined to

OFD orders from a single city in China. However, we posit that this city serves as a robust representative case for modern OFD ecosystems. Situated as a key node within the Beijing-Tianjin-Hebei urban agglomeration, the city houses a permanent population exceeding 9 million. Its urban morphology features a typical mix of dense central business districts, residential clusters, and suburban zones, providing sufficient statistical power to model complex interactions and rare delay events. Second, platform-specific practices may introduce bias. Our analysis is based on data from a single major platform (Meituan). As dispatching algorithms and operational rules (e.g., punishment mechanisms for delays) vary significantly across platforms like Uber Eats or DoorDash, the specific magnitude of causal effects observed here may differ in other platforms. Finally, the analysis did not account for weather conditions, as there were no significant meteorological changes during the study period.

To address these limitations, future research could broaden the geographic and platform scope to include diverse cities and countries, thereby enhancing the applicability of the findings. Incorporating additional external factors, such as adverse weather conditions and traffic congestion indices, would provide a more comprehensive and robust understanding of the influences on OFD delays. Furthermore, future work could integrate these causal findings into dynamic optimization models to simulate the impact of the proposed policy interventions (e.g., establishing preparation time thresholds or implementing dynamic order caps) on overall system performance.

Data availability

The datasets generated and analyzed during the current study are not publicly available due to legal/ethical reasons, but are available from the corresponding author on reasonable request.

Code availability

The custom code that supports the findings of this study is available from the corresponding author upon reasonable request.

Received: 20 July 2025; Accepted: 9 March 2026;

Published online: 25 March 2026

References

1. Technavio. Online on-demand food delivery services market analysis: APAC, North America, Europe, Middle East and Africa, South America - US, China, Canada, UK, Japan, South Korea, India, Germany, Australia, France - size and forecast 2025–2029. <https://www.technavio.com/report/online-on-demand-food-delivery-services-market-size-industry-analysis> (2025).
2. Intouch Insight & Technomic. The path to 3rd party delivery excellence: an Intouch insight study on 3rd party delivery and technomic trends data. https://kioskindustry.org/wp-content/uploads/2024/09/The-Path-to-3rd-Party-Delivery-Excellence_-_An-Intouch-Insight-Study-on-3rd-Party-Delivery-and-Technomic-Trends-Data-Cameron-Watt-Robert-Byrne.pdf (2024).
3. Mao, W., Ming, L., Rong, Y., Tang, C. S. & Zheng, H. Faster deliveries and smarter order assignments for an on-demand meal delivery platform. *J. Oper. Manag.* **71**, 220–245 (2025).
4. Amat-Lefort, N. & Barnes, S. J. An inconvenient truth: understanding service inconvenience in digital platforms. *J. Serv. Res.* <https://doi.org/10.1177/10946705241254735> (2024).
5. Zheng, Y., Ma, Y., Guo, L., Cheng, J. & Zhang, Y. Crash involvement and risky riding behaviors among delivery riders in China: the role of working conditions. *Transp. Res. Rec.* **2673**, 1011–1022 (2019).
6. Hu, Y. Research on the prediction model of arrival time of takeout o2o orders based on feature extraction. *China Prices* 120–123 (2022).
7. Zhang, Q., Ma, Z., Wu, Y., Liu, Y. & Qu, X. Quantifying variable contributions to bus operation delays considering causal relationships. *Transp. Res. Part E Logist. Transp. Rev.* **194**, 103881 (2025).
8. Zhang, Q., Wang, W., She, J. & Ma, Z. Understanding bus network delay propagation: Integration of causal inference and complex network theory. *J. Transp. Geogr.* **123**, 104098 (2025).
9. Salari, N., Liu, S. & Shen, Z.-J. M. Real-time delivery time forecasting and promising in online retailing: when will your package arrive?. *Manuf. Serv. Oper. Manag.* **24**, 1421–1436 (2022).
10. Wen, H. et al. A survey on service route and time prediction in instant delivery: Taxonomy, progress, and prospects. *IEEE Trans. Knowl. Data Eng.* **36**, 7516–7535 (2024).
11. Zhang, X., Zhou, Z., Xu, Y. & Zhao, X. Analyzing spatial heterogeneity of ridesourcing usage determinants using explainable machine learning. *J. Transp. Geogr.* **114**, 103782 (2024).
12. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In (eds Guyon, I. et al.) *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, 2017).
13. Liu, Y., Shang, Y. & Li, S. Joint infrastructure planning and order assignment for on-demand food-delivery services with coordinated drones and human couriers. Preprint at <https://doi.org/10.48550/arXiv.2501.14325> (2025).
14. Seghezzi, A. & Mangiaracina, R. On-demand food delivery: investigating the economic performances. *Int. J. Retail Distrib. Manag.* **49**, 531–549 (2021).
15. Seghezzi, A., Winkenbach, M. & Mangiaracina, R. On-demand food delivery: a systematic literature review. *Int. J. Logist. Manag.* **32**, 1334–1355 (2021).
16. Liu, S. & Luo, Z. On-demand delivery from stores: dynamic dispatching and routing with random demand. *Manuf. Serv. Oper. Manag.* **25**, 595–612 (2023).
17. Wang, X., Ji, C., Xu, H. & Guo, K. Research on dynamic optimization of takeout delivery routes considering food preparation time. *Sustainability* **17**, 2771 (2025).
18. Li, X., Wang, X., Liu, Z., Zhang, J. & Tang, J. Real-time demands, restaurant density, and delivery reliability: An empirical analysis of on-demand meal delivery. *J. Oper. Manag.* **71**, 246–292 (2025).
19. Garg, A., Ayaan, M., Parekh, S. & Udandara, V. Food delivery time prediction in indian cities using machine learning models. Preprint at <https://doi.org/10.48550/arXiv.2503.15177> (2025).
20. Cui, R., Lu, Z., Sun, T. & Golden, J. M. Sooner or later? promising delivery speed in online retail. *Manuf. Serv. Oper. Manag.* **26**, 233–251 (2024).
21. Chen, J., Fan, T., Gu, Q. & Pan, F. Emerging technology-based online scheduling for instant delivery in the O2O retail era. *Electron. Commer. Res. Appl.* **51**, 101115 (2022).
22. Mai, D. & Le, T. T. Improving Instant Delivery with Consideration of Traffic Congestion. Available at SSRN 3990596.
23. Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* 3rd edn. (Routledge, 2013).
24. Chatterjee, S. & Hadi, A. S. *Regression Analysis by Example* 5th edn (John Wiley & Sons, 2013).
25. Glymour, C., Zhang, K. & Spirtes, P. Review of causal discovery methods based on graphical models. *Front. Genet.* **10**, 524 (2019).
26. Geffner, T. et al. Deep end-to-end causal inference. Preprint at <https://doi.org/10.48550/arXiv.2202.02195> (2022).
27. Spirtes, P., Glymour, C. N. & Scheines, R. *Causation, Prediction, and Search*. Bradford Books; *Adaptive Computation and Machine Learning* 2nd edn (MIT Press, 2000).
28. Chickering, D. M. Optimal structure identification with greedy search. *J. Mach. Learn. Res.* **3**, 507–554 (2002).
29. Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A. & Jordan, M. I. A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* **7**, 2003–2030 (2006).
30. Zheng, X., Aragam, B., Ravikumar, P. K. & Xing, E. P. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In Bengio, S.

- et al. (eds.) *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 9472–9483 (Curran Associates, 2018).
31. Yu, Y., Chen, J., Gao, T. & Yu, M. DAG-GNN: DAG structure learning with graph neural networks. In Chaudhuri, K. & Salakhutdinov, R. (eds.) *Proc. 36th International Conference on Machine Learning*, vol. 97, 7154–7163 (PMLR, 2019).
 32. Angrist, J. D., Imbens, G. W. & Rubin, D. B. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* **91**, 444–455 (1996).
 33. Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).
 34. Card, D. & Krueger, A. B. Minimum wages and employment: a case study of the fast food industry in New Jersey and Pennsylvania. NBER Working Paper 4509, National Bureau of Economic Research (1993).
 35. Wager, S. & Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* **113**, 1228–1242 (2018).
 36. Chernozhukov, V. et al. Double/debiased machine learning for treatment and structural parameters. *Econ. J.* **21**, C1–C68 (2018).
 37. Tukey, J. W. *Exploratory Data Analysis*, vol. 2 of *Addison-Wesley Series in Behavioral Science: Quantitative Methods* (Addison-Wesley Publishing Company, 1977).
 38. Zheng, Q., Zhan, J. & Feng, X. Working safety and workloads of Chinese delivery riders: the role of work pressure. *Int. J. Occup. Saf. Ergon.* **29**, 869–882 (2023).
 39. Meituan. How are orders assigned to couriers? <https://www.meituan.com/zh-HK/news/NN250825125002496?source=relativeNews> (2025).
 40. Farrar, D. E. & Glauber, R. R. Multicollinearity in regression analysis: the problem revisited. *Rev. Econ. Stat.* **49**, 92–107 (1967).
 41. Jin, Z., Sun, X., Xu, Z. & Tu, H. A data-driven approach to uncovering the charging demand of electrified ride-hailing services. *Transp. Res. Part D Transp. Environ.* **139**, 104599 (2025).
 42. Khemakhem, I., Monti, R., Leech, R. & Hyvarinen, A. Causal autoregressive flows. In Banerjee, A. & Fukumizu, K. (eds.) *Proc. 24th International Conference on Artificial Intelligence and Statistics*, vol. 130, 3520–3528 (PMLR, 2021).
 43. Lippe, P., Cohen, T. & Gavves, E. Efficient neural causal discovery without acyclicity constraints. Preprint at <https://doi.org/10.48550/arXiv.2107.10483> (2021).
 44. Agrawal, T. *Optuna and autoML*, 109–129 (Springer, 2020).
 45. Deka, P. P. & Weiner, J. *XGBoost for Regression Predictive Modeling and Time Series Analysis: Learn how to build, evaluate, and deploy predictive models with expert guidance* (Packt Publishing Ltd, 2024).
 46. Chen, C.-F. Investigating the effects of job stress on the distraction and risky driving behaviors of food delivery motorcycle riders. *Saf. Health Work* **14**, 207–214 (2023).
 47. Kancharla, S. R., Van Woensel, T., Waller, S. T. & Ukkusuri, S. V. Meal delivery routing problem with stochastic meal preparation times and customer locations. *Netw. Spat. Econ.* **24**, 997–1020 (2024).
 48. Yao, W., Zhao, H. & Liu, L. Weather and time factors impact on online food delivery sales: a comparative analysis of three Chinese cities. *Theor. Appl. Climatol.* **153**, 1425–1438 (2023).
 49. Jin, H. Order allocation strategy for instant delivery: from modeling to optimization. <https://tech.meituan.com/2017/10/11/o2o-intelligent-distribution.html> (2017).
 50. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. Catboost: unbiased boosting with categorical features. In (eds Bengio, S. et al.) *Advances in Neural Information Processing Systems*, vol. 31 (Curran Associates, 2018).
 51. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
 52. Meituan Company. First algorithm disclosure: Making food delivery algorithms more transparent and allowing more voices to participate in change. <https://www.meituan.com/news/NN250103064009007> (2021).
 53. Meituan. How is the “estimated delivery time” calculated?. <https://www.meituan.com/news/NN250825131002446?source=relativeNews> (2025).
 54. Xinhuanet. Further improvements to courier dispatching algorithms. <https://www.xinhuanet.com/tech/20250828/ed2b35ce7a5c40aa81a3a8d73b9063cb/c.html> (2025).
 55. Meituan. Meituan delivery service guarantees. <https://rules-center.meituan.com/customer-rights/1> (2025).
 56. Chen, A. H., Lee, J. Z.-H. & Ho, Y.-L. Influential factors for online food delivery platform drivers’ order acceptance. *Information Discovery and Delivery.* **54**, 113–128 (2026).
 57. Bhat, R. L. & Gillani, I. A. Spatio-temporal demand prediction for food delivery using attention-driven graph neural networks. Preprint at <https://doi.org/10.48550/arXiv.2507.15246> (2025).
 58. Bezerra, H. & Cancho, V. Spatial forecasting of online food delivery demand. Preprint at <https://doi.org/10.21203/rs.3.rs-3903945/v1> (2024).

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China [72101188] and the Funds for International Cooperation and Exchange of the National Natural Science Foundation of China [72361137005] and SINERGI Project. This research was supported by data provided by Meituan.

Author contributions

M.L. supervised the study, contributed to the investigation and methodology, and reviewed and edited the manuscript. R.L. conducted the literature review and formal analysis and prepared the original draft. Z.J. conceptualized the study, developed the methodology and software, performed formal analysis, prepared the original draft, and contributed to reviewing and editing the manuscript. Q.Y. provided resources and contributed to reviewing and editing the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

Author Quan Yuan is an editorial board member (Associate Editor) for *npj Sustainable Mobility and Transport*, but was not involved in the editorial review of, or the decision to publish, this article. The authors declare no other competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s44333-026-00097-1>.

Correspondence and requests for materials should be addressed to Zhicheng Jin.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026