



Research article

Data-driven optimization for airline service quality improvement under delay and cancellation risk

Jinho Cha¹ and Eun-Young Kwon^{2,*}

¹ Department of Computer Science, Gwinnett Technical College, Lawrenceville, GA, USA

² AI Industry Research Center, Kwangwoon University, Seoul, Korea

* **Correspondence:** Email: eykwon@kma.ac.kr, eykwonclass0524@gmail.com.

Appendix

A. Data processing and reproducibility

This appendix documents the reproducible construction of the airline service-quality data through three layers: raw official files, standardized intermediate files, and analysis-ready panels.

A.1. DOT Air Travel Consumer Report panels

We first process official U.S. Department of Transportation Air Travel Consumer Reports into carrier-month, airport-month, and carrier-airport-month panels. The DOT processing pipeline produces analysis-ready panels at three levels:

$$(a, t), \quad (r, t), \quad (a, r, t),$$

where a indexes the marketing carrier, r indexes the airport, and t indexes the calendar month.

The carrier-month DOT panel is the primary official service-quality outcome panel. It contains carrier-month measures of arrival on-time performance, cancellation rates, scheduled operations, canceled operations, completed operations, and service-harm indices. The corrected DOT service-harm measure uses the mutually exclusive decomposition

$$L_{a,t}^{nc} = 100 - OT_{a,t} - CR_{a,t}, \quad L_{a,t}^{can} = CR_{a,t},$$

and

$$H_{a,t}^{DOT}(\lambda) = L_{a,t}^{nc} + \lambda L_{a,t}^{can}.$$

Thus, cancellations are excluded from the non-cancellation service-loss component and receive exactly the intended scenario weight λ . In the baseline specification, we set $\lambda = 2$.

The airport-month DOT panel records arrival and departure reliability at major airports. The carrier-airport-month panel records carrier-specific arrival performance at reported airports. Rows with zero reported arrivals are retained for official completeness but are excluded from ranking and worst-case service-loss calculations.

A.2. BTS flight-level data standardization

We next process monthly BTS T_ONTIME_MARKETING files into standardized flight-level parquet files. The raw BTS files are official monthly ZIP files from the Marketing Carrier On-Time Performance table. The processed period covers

$$2024-01 \leq t \leq 2026-02.$$

The main sample is 2024-01 to 2025-12, and 2026-01 to 2026-02 is retained as a robustness sample.

The raw BTS downloads contain two naming conventions. Earlier files use uppercase field names such as

FL_DATE, MKT_UNIQUE_CARRIER, ARR_DEL15, CANCELLED,

whereas later files use names such as

FlightDate, Marketing_Airline_Network, ArrDel15, Cancelled.

We standardize both formats into a common schema before aggregation. The standardized monthly parquet files retain flight date, marketing carrier, operating carrier, origin and destination airports, origin and destination states, delay indicators, delay minutes, cancellation and diversion indicators, distance, and delay-cause variables.

The resulting standardized parquet layer contains 26 monthly files. This intermediate layer is preserved so that all analysis-ready panels can be regenerated from standardized flight-level records without repeatedly parsing the raw ZIP files.

A.3. Domestic route definition

The analysis-ready BTS panels use a U.S. domestic-route filter. A flight is retained if both the origin and destination state codes correspond to U.S. states, the District of Columbia, or U.S. territories included in BTS reporting. Formally, define

$$D_i = \mathbf{1} \left\{ s_i^{orig} \in \mathcal{S}_{US} \text{ and } s_i^{dest} \in \mathcal{S}_{US} \right\},$$

where s_i^{orig} and s_i^{dest} denote the origin and destination state codes of flight i , and $\mathcal{S}_{US}$ contains the 50 states, DC, and included U.S. territories. Only flights with $D_i = 1$ are used in the canonical BTS aggregation panels.

This restriction aligns the flight-level operational data with the DOT service-quality panels, which focus on U.S. airline service performance reported through official DOT channels.

A.4. BTS aggregated risk panels

From the standardized domestic flight-level records, we construct four BTS risk panels:

$$(a, t), \quad (a, o, t), \quad (a, d, t), \quad (a, r, t),$$

where o is the origin airport, d is the destination airport, and r denotes the arrival airport for the carrier-airport arrival panel.

For each aggregation unit, we compute:

$$N_{g,t} = \text{number of scheduled flights,}$$

$$C_{g,t} = \text{number of canceled flights,}$$

$$A_{g,t}^{15} = \text{arrival-delay rate of at least 15 minutes,}$$

$$D_{g,t}^{15} = \text{departure-delay rate of at least 15 minutes,}$$

where g denotes the aggregation unit. Arrival-delay rates are computed only over non-canceled and non-diverted flights with valid arrival outcomes. Departure-delay rates are computed over non-canceled flights with valid departure outcomes.

For the corrected mutually exclusive reconstruction, let $D_{g,t}$ denote delayed operated arrivals and $V_{g,t}$ diverted flights. The BTS risk index is

$$R_{g,t}^{BTS}(\lambda) = 100 \frac{D_{g,t} + V_{g,t} + \lambda C_{g,t}}{N_{g,t}},$$

with $\lambda = 2$ in the baseline specification. The corresponding exposure-weighted measure is

$$E_{g,t}^{BTS}(\lambda) = D_{g,t} + V_{g,t} + \lambda C_{g,t}.$$

Delayed operated arrivals, diversions, and cancellations therefore enter as mutually exclusive event categories.

The canonical BTS carrier-month panel contains 258 carrier-month observations over 26 months. The main sample contains 240 observations, corresponding to 24 months and 10 marketing carriers. The robustness sample contains 18 observations, corresponding to 2 months and 9 marketing carriers.

A.5. Carrier mapping and DOT–BTS merge

The DOT–BTS merge requires a carrier-name harmonization step because the two data sources use different carrier identifiers. BTS flight-level records report marketing carriers using two-letter carrier codes, whereas the DOT Air Travel Consumer Report reports official carrier or carrier-network names. We therefore construct a deterministic carrier crosswalk before aggregating and merging the two datasets.

In this crosswalk, AA is mapped to American Airlines Network, AS to Alaska Airlines Network, B6 to JetBlue Airways, DL to Delta Air Lines Network, F9 to Frontier Airlines, G4 to Allegiant Air, HA to Hawaiian Airlines, NK to Spirit Airlines, UA to United Airlines Network, and WN to Southwest Airlines. All carrier names are standardized to a common uppercase format before merging in order to avoid mismatches caused by capitalization, spacing, or naming conventions.

After applying the carrier crosswalk, we aggregate the BTS flight-level delay and cancellation risk measures to the carrier-month level. The resulting BTS carrier-month panel is then merged with the DOT carrier-month service-quality panel using the common carrier-month key (a, t) , where a denotes the harmonized carrier identifier and t denotes the reporting month.

The final merged DOT–BTS carrier-month panel contains 258 matched observations. All DOT and BTS carrier-month records match exactly, so there are no DOT-only or BTS-only rows after harmonization. The main sample contains 240 matched observations, and the robustness sample contains 18 matched observations. This merged panel provides the basis for the DOT–BTS alignment diagnostics, the construction of service-harm measures, and the subsequent resource-allocation experiments.

A.6. Alignment between DOT outcomes and BTS reconstructions

The DOT and BTS measures are not mechanically identical. DOT values come from official report tables, whereas BTS values are reconstructed from flight-level records. Therefore, small differences can arise from rounding, table-release timing, revision status, and aggregation conventions.

Under the corrected mutually exclusive construction, the main-sample rate-based DOT–BTS Pearson and Spearman correlations are 0.999983 and 0.999919, respectively; the corresponding exposure-weighted correlations are 0.999998 and 0.999985. These statistics are interpreted as reconstruction consistency between linked reporting layers, not as independent-source validation.

A.7. Canonical files

Table 1 identifies the canonical data objects required to regenerate the empirical analysis, organized by role and unit of analysis.

Table 1. Canonical DOT–BTS data files

File role	Unit	Purpose
Monthly BTS parquet folder	Flight	Standardized monthly flight-level intermediate files.
BTS domestic carrier-month risk panel	Carrier-month	Main BTS risk aggregation panel.
BTS domestic carrier-origin-month panel	Carrier-origin-month	Departure-side airport risk panel.
BTS domestic carrier-destination-month panel	Carrier-destination-month	Arrival-side airport panel.
BTS domestic carrier-airport arrival panel	Carrier-airport-month	Arrival-airport panel aligned with DOT airport reporting.
Merged DOT–BTS carrier-month panel	Carrier-month	Main empirical backbone for the paper.
Carrier mapping table	Carrier	Mapping between BTS marketing carrier codes and DOT report carrier names.
Merge quality-check table	Sample	Documents matched rows and source consistency.

The merged DOT–BTS carrier-month panel is the empirical backbone for the alignment, service-harm, and allocation analyses.

B. Supplementary alignment figure

Figure S1 summarizes the DOT–BTS cross-layer measurement-consistency diagnostics used in the main text. Because both sources describe the same underlying airline operations, the figure is interpreted as a visual audit of reconstruction consistency across levels, time, and rankings, not as independent-source validation or evidence of incremental information.

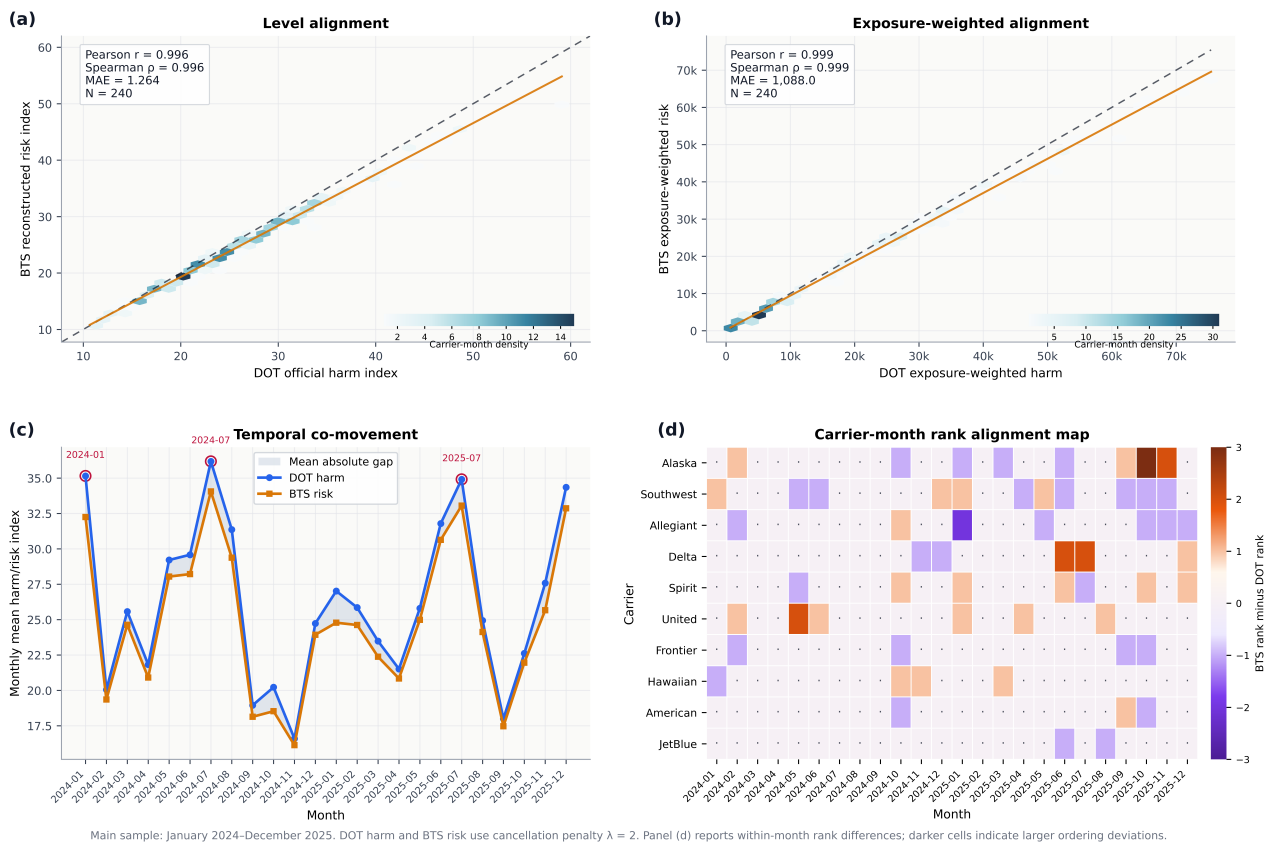


Figure S1. Supplementary DOT–BTS cross-layer alignment diagnostics. (a) DOT-published harm index versus the BTS reconstructed risk index at the carrier-month level. (b) Exposure-weighted DOT and BTS measures. (c) Monthly co-movement across the study period. (d) Within-month carrier-rank differences. The figure is interpreted as a reconstruction and measurement-consistency diagnostic rather than as independent-source validation or evidence of incremental information from blending.

C. Forecast-to-prescription reproducibility

The forecast-to-prescription evaluation is locked to 120 carrier-month observations covering 10 marketing carriers from January through December 2025. The target is the corrected mutually exclusive exposure-weighted DOT service harm at $\lambda = 2$, and the primary forecast score is generated by the Gradient Boosting model. The allocation score does not use the realized current-month target. Primary decisions use the diminishing-return water-filling solution with $\beta = 1$ and budget fractions 0.05, 0.10, 0.20, 0.30; the linear priority-index rule is retained as a secondary check. Month-block bootstrap uncertainty uses 1,000 replications with seed 20260824. Reproducibility outputs include locked forecast scores, monthly rank-fidelity summaries, diminishing-return and linear policy summaries, row-level allocations, month-level realized values, bootstrap draws, and bootstrap percentile intervals.

D. Heterogeneous cost/capacity reproducibility

The heterogeneous-cost/capacity analysis reuses the locked 2025 Forecast-PTO panel from the forecast-to-prescription experiment. Relative costs are anchored to the percentile rank of the 2024 coefficient of variation of scheduled operations, and relative capacity limits are anchored to the percentile rank of 2024 mean scheduled operations. The Homogeneous, Mild, Moderate, and Strong scenario-strength pairs are (0, 0), (0.20, 0.25), (0.35, 0.40), and (0.50, 0.60), with mean-one normalization after scenario-specific clipping. The primary diminishing-return model uses $\beta = 1$, budget fractions 0.05, 0.10, 0.20, and 0.30, and the scenario-specific mapping $B_t(b) = b \sum_a c_a \bar{x}_a$. A secondary linear diagnostic compares s_a/c_a against the raw forecast score under identical costs, capacities, and budgets. Month-block bootstrap uncertainty uses 1,000 replications with seed 20260824. The design audit verifies that current-month realized harm is not used in the forecast score or in the construction of cost/capacity profiles; it is used only for ex-post realized value and hindsight-oracle normalization.

E. Uncertainty-aware robust allocation reproducibility

The uncertainty-aware analysis reuses the locked 2025 forecast-score panel from the forecast-to-prescription experiment. Forecast disagreement is calculated as 1.4826 times the cross-model median absolute deviation across Gradient Boosting, Historical Average, Lag-1 DOT, Lag-1 BTS, Seasonal-12 DOT, Ridge, and Rolling-3 DOT. Each carrier-month contains all seven forecast scores. The realized 2025 DOT target is not used to construct the uncertainty radius or allocation score.

The deterministic baseline is $\kappa = 0$, and the robust grid is $\kappa \in \{0, 0.5, 1.0, 1.5, 2.0\}$ with $\kappa = 1$ designated as primary. The experiment uses homogeneous normalized costs and capacities to isolate forecast uncertainty, $\beta = 1$, and budget fractions 0.05, 0.10, 0.20, and 0.30. Month-block bootstrap uncertainty uses 1,000 paired replications with seed 20260824. The deterministic reproduction, leakage, budget-feasibility, and robust-counterpart optimality audits all pass. Saved outputs include the uncertainty-score panel, carrier-level uncertainty diagnostics, row- and month-level allocations, full robust summaries, paired $\kappa = 1$ comparisons, bootstrap draws and intervals, and the robust-counterpart optimality audit.

F. Volume-controlled DOT–BTS alignment reproducibility

The volume-controlled alignment experiment uses the balanced January 2024–December 2025 main panel with 240 carrier-month observations, 10 carriers, and 24 months. It evaluates raw rate and exposure correlations, exposure per 100 scheduled operations, within-carrier and within-month demeaning, carrier-month two-way demeaning, partial correlation controlling log scheduled operations, and residual correlation after separate log-volume, carrier fixed-effect, and month fixed-effect regressions for DOT and BTS measures. Scheduled-operation quartiles provide an additional scale-stratified check. Month-block bootstrap uncertainty uses 1,000 replications with seed 20260824. Saved outputs include carrier-level correlations, monthly correlations, volume-quartile correlations, volume-explanation summaries, residualized row-level panels, manuscript-ready correlation summaries, and month-block bootstrap draws and percentile intervals. The data/computation audit passes.

G. Carrier-airport-month robustness reproducibility

The finer-resolution robustness experiment merges the DOT carrier-airport-month arrival panel with the BTS carrier-airport-month arrival-risk panel for the 10 study carriers over January 2024–December 2025. After matching and positive-volume filtering, the main sample contains 5,807 observations. The forecast split uses 2,929 observations from 2024 for training and 2,878 observations from 2025 for out-of-sample evaluation. The primary forecast target is DOT arrival-delay exposure, and all features use only information available before month t , including one-month lags, three-month rolling means, expanding historical averages, lagged volume terms, and calendar seasonality. Allocation uses the same diminishing-return water-filling model with $\beta = 1$, homogeneous unit costs and capacities, and budget fractions 0.05, 0.10, 0.20, and 0.30. The primary inclusion rule requires at least one reported arrival, with 100- and 500-arrival thresholds used for thin-cell robustness checks. Saved outputs include manuscript-ready alignment, forecast-metrics, rank-fidelity, primary-allocation, allocation-table, and threshold-sensitivity CSV files, together with publication-quality main and supplementary figures. The experiment is arrival-side only and does not impute airport-specific cancellation harm.

H. Computational audit

The computational analyses were rerun from the corrected mutually exclusive carrier-month harm panel wherever the carrier-month target enters the forecast or allocation calculation. The canonical 2025 prospective panel contains 120 carrier-month observations (10 carriers over 12 months). Experiment 07 trains only on months strictly earlier than the forecast month and excludes current-month DOT and BTS outcomes from all forecasting features. Experiment 08 uses the resulting ex-ante Forecast-PTO score and reserves realized month- t DOT harm for ex-post value calculation and hindsight-oracle normalization. Experiment 09 anchors heterogeneous cost and capacity profiles only to 2024 information. Experiment 10 constructs its model-disagreement radius from seven ex-ante forecasts and reproduces deterministic Forecast-PTO at $\kappa = 0$ to numerical precision. Experiment 04R reproduces the corrected Experiment 08 results at $\beta = 1$ to within 4.54×10^{-7} . Experiment 12 is an arrival-side carrier-airport-month robustness analysis and does not impute airport-specific cancellation harm.

I. Finer-resolution robustness supplementary figure

Figure S2 provides the visual audit for Experiment 12 at the finer carrier-airport-month resolution. The four labeled views summarize DOT–BTS alignment, Airport-GBR calibration, monthly priority-rank fidelity, and highest-priority-set recovery. The analysis remains arrival-side and does not claim a complete airport-network recovery model.

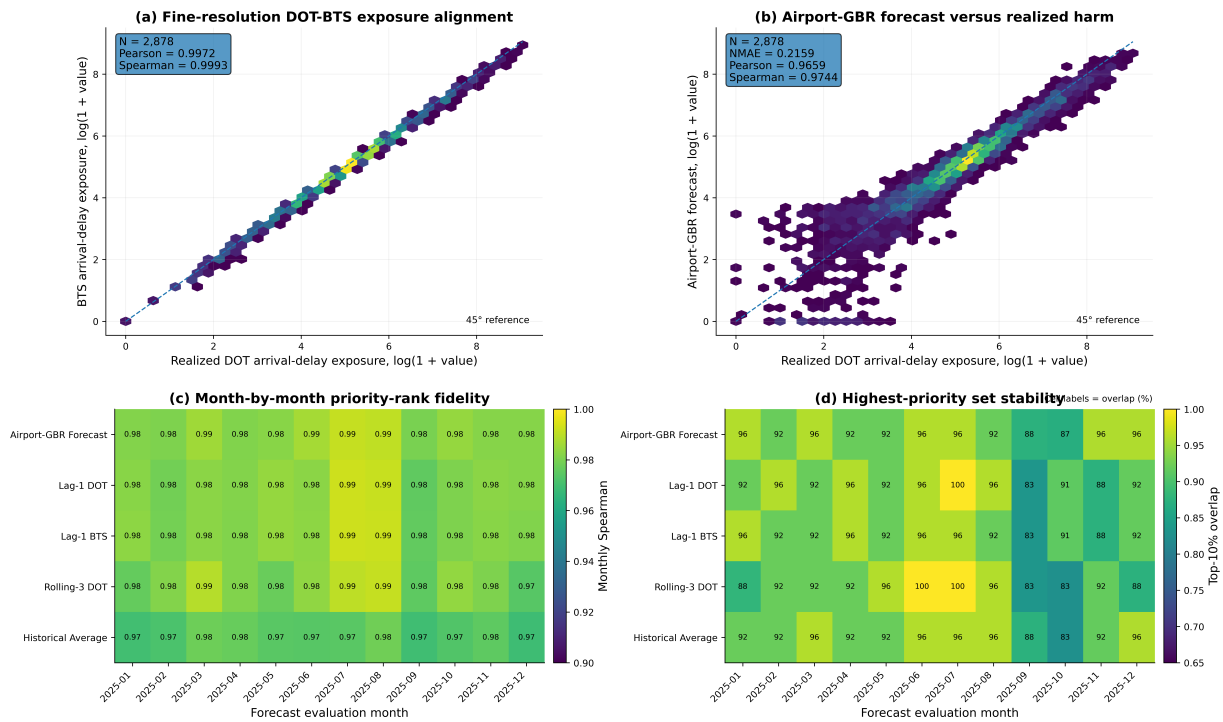


Figure S2. Supplementary finer-resolution diagnostics for Experiment 12. (a) DOT and BTS arrival-delay exposure on the locked 2025 carrier-airport-month evaluation panel with a 45-degree reference. (b) Airport-GBR forecasts versus realized DOT arrival-delay exposure. (c) Month-by-month Spearman priority-rank fidelity across the five implementable forecasting rules. (d) Month-by-month top-10% priority-set overlap with the realized highest-harm set. The $\log(1 + \text{value})$ transformation is used only for visualization in (a) and (b); all reported accuracy and correlation statistics are calculated on the original scale.



AIMS Press

©2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>)