



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Regime-Gated Explainable AI for Philippine Tourism Source-Market Decisions: Evidence from a 2008–2023 Multi-Source Panel

Karen Razelle M. Duyan^{1,2*}, Juan Moshe M. Duyan^{1,2}, Valerie A. Abesamis Paloy¹, Eric A. Paloy¹, Melanie S. Manuel¹, Mark Joby Aguilar¹, Joe Robert G. Lucena^{2,3}, Rissa A. Lasap^{2,4*}, Luigi Carlo M. De Jesus^{2,3,5}

¹ Kalinga State University, Tabuk City, Kalinga, Philippines; kmduyan@ksu.edu.ph; jmduyan@ksu.edu.ph; vaabesamis@ksu.edu.ph; epaloy@ksu.edu.ph; msmanuel@ksu.edu.ph

² Naveen Jindal Youth Fellowship Program, O.P. Jindal Global University, Sonapat, Haryana, India; karen.razelle_fs@jgu.edu.in; juan.moshe_fs@jgu.edu.in; luigi.carlo_fs@jgu.edu.in; joe.robert_fs@jgu.edu.in

³ School of Engineering, Asia Pacific College, Makati City, Metro Manila, Philippines; luigid@apc.edu.ph; joerobertpaul@apc.edu.ph

⁴ Biological Science Department, Centro Escolar University, Manila, Metro Manila, Philippines

⁵ Department of Computer, Electronics, and Electrical Engineering, De La Salle University, Manila, Metro Manila, Philippines

Abstract

International-arrival forecasts are often evaluated as if one model should remain optimal across normal demand, border closure, and recovery. That assumption is consequential for the Philippines, where source markets recovered at sharply different rates after COVID-19. This study develops a regime-gated explainable artificial intelligence (XAI) framework for source-market planning. Monthly visitor-arrival reports from the Philippine Department of Tourism were harmonized with World Development Indicators for 14 markets from January 2008 to October 2023, producing 2,646 market-month observations and covering 84.46% of 2019 arrivals. Persistence, seasonal-naïve, ridge, random-forest, extra-trees, and gradient-boosting variants were evaluated with strictly temporal holdouts for 2019, 2022, and 2023. Exact grouped baseline Shapley values translated the best stable-regime model into auditable market-level explanations, while a recovery-volume matrix converted the outputs into portfolio actions. In 2019, residual Extra Trees achieved 7.25% weighted absolute percentage error (WAPE), versus 13.13% for persistence; a month-cluster bootstrap placed the improvement at 5.88 percentage points (95% CI: 3.56–8.47). The ranking reversed in transition: persistence outperformed the same model in 2022 (23.87% versus 32.30%) and 2023 (15.28% versus 28.25%). A transparent gate that used machine learning only in stable demand and persistence during transition reduced mean fold WAPE from 17.43% to 15.47%. Shapley analysis showed that prior demand level dominated stable forecasts, followed by recent momentum, seasonality, and market identity. The novelty is not another claim that AI universally improves tourism forecasting; it is an empirically validated, explanation-linked rule for deciding when AI should and should not be trusted, combined with a Philippine source-market decision portfolio.

Keyword: tourism demand forecasting; explainable artificial intelligence; regime change; source-market portfolio; Philippines; Shapley value.

1. Introduction

Tourism-demand forecasts are decision instruments: they influence route-development priorities, market budgets, staffing, capacity, and the timing of recovery interventions. Forecasting research has consequently moved from classical time-series models toward mixed-frequency, search-enhanced, ensemble, and deep-learning methods [1–12]. The attraction of machine learning is clear when demand contains nonlinear seasonality and heterogeneous market responses. Yet the operational question is not merely whether a complex model can fit historical arrivals. A destination manager needs to know whether the model remains dependable when the data-generating process changes, why a forecast differs across source markets, and which action follows from that explanation.

These questions are acute in an archipelagic destination such as the Philippines. International visitation depends heavily on air access and a concentrated set of Asian and long-haul markets. The pandemic then created a discontinuity that was not an ordinary seasonal disturbance: border restrictions reduced arrivals nearly to zero, reopening restarted markets at different speeds, and the historical relationship between recent lags and future demand became unstable. The Philippine Statistics Authority's tourism satellite accounts confirm that tourism remains material to national output and employment, making forecast governance—not forecast accuracy alone—a public and managerial concern [33].

The literature provides strong foundations for tourism forecasting [1–12] and a growing account of machine learning in tourism [13]. Philippine work has also demonstrated the feasibility of artificial neural networks for aggregate arrivals [14]. However, three gaps remain. First, aggregate national forecasts conceal the source-market heterogeneity that tourism boards actually manage. Second, most accuracy comparisons reward a single winner across the test period, even though stable demand and post-shock transition may require different forecasting rules. Third, high-performing models are rarely joined to an explanation and prioritization layer that can be audited by decision makers. General XAI methods such as local surrogate explanations, tree-based Shapley values, and

accumulated local effects improve transparency [20–23], but their decision value depends on careful feature grouping, temporal validation, and an explicit boundary between prediction and causation.

This study therefore asks: (RQ1) How does forecast performance change across stable, reopening, and recovery regimes? (RQ2) Which feature groups drive accurate forecasts in the stable regime? (RQ3) How can forecast performance, explanations, and observed recovery gaps be combined into source-market actions? The purpose is to develop and evaluate a regime-gated XAI framework using official Philippine secondary data. The analysis combines monthly Department of Tourism (DOT) arrival reports [31] with annual World Development Indicators (WDI) [32] and uses Philippine Tourism Satellite Account evidence for decision context [33]. Fourteen source markets are modeled separately but within a common panel.

NOVELTY AND CONTRIBUTIONS. To the authors' knowledge, this is the first Philippine tourism study to combine (i) a 14-market monthly panel covering 84.46% of pre-pandemic international arrivals, (ii) out-of-time tests that deliberately separate stable, reopening, and recovery regimes, (iii) an explicit rule that determines when an explainable residual-tree model should yield to a simple persistence forecast, and (iv) exact grouped baseline Shapley explanations connected to a recovery-volume portfolio. The contribution is thus a model-governance mechanism, not merely a new accuracy score. It also tests and documents a negative result—AI underperformance during transition—that is usually obscured by full-period averages. Figure 1 summarizes the integrated framework.

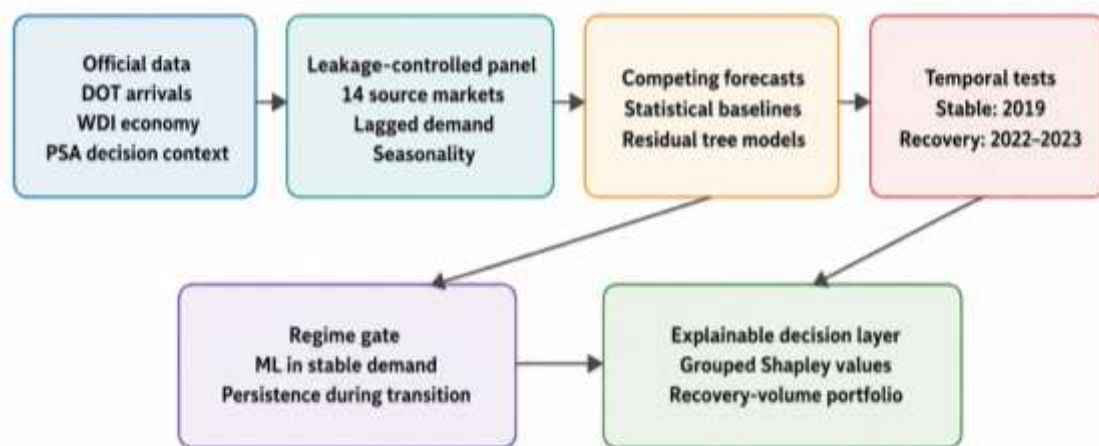


Fig. 1. Regime-gated explainable-AI framework linking official data, temporal evaluation, model governance, explanations, and source-market action

2. Related Work and Decision Context

2.1 Tourism Demand Forecasting Under Structural Change

Reviews of tourism-demand forecasting consistently emphasize the value of both benchmark discipline and data relevance [1–3]. Univariate models often remain difficult to beat, while multivariate specifications can benefit when economic or behavioral predictors add information unavailable in the arrival history [2,4]. Search-query and mixed-data sampling studies demonstrate incremental value from high-frequency signals [5–7,10,11], and recent group-pooling, decomposed deep-learning, and ensemble methods show that information can be shared across related series [8,9,12]. Those advances nevertheless assume that relationships learned from the training sample remain sufficiently stable in the evaluation period.

COVID-19 made that assumption fragile. Tourism research described the shock as a global systems disruption and called for resilience, redesign, and explicit attention to recovery paths [24–30]. A forecast trained largely on normal mobility may interpret near-zero closure observations or rapid reopening growth as ordinary variation. Conversely, a model trained through closure can become conservative when pent-up demand returns. This motivates evaluation by regime and the inclusion of simple baselines: a sophisticated model should earn operational use in each decision environment rather than inherit it from pre-shock performance.

2.2 Explainable AI for Forecast Governance

Explainability is most useful when it supports a concrete governance choice. LIME provides locally faithful surrogate explanations [20], while TreeSHAP derives additive attributions for tree ensembles [21]. Methodological work also warns that model-agnostic explanations can be unstable or misleading when correlated variables, unrealistic counterfactuals, and causal language are ignored [22,23]. In a lag-rich tourism panel, individual-feature attributions can fragment one managerial concept across many correlated variables. The present study therefore groups features into seven interpretable blocks and computes exact Shapley values over those groups relative to a fixed empirical baseline. The result answers a predictive question—what moved this forecast away from the baseline—not a causal question about what would happen after a policy intervention.

2.3 Philippine Source-Market Decision Gap

The available Philippine evidence is dominated by descriptive reporting and aggregate forecasting. The national arrival total is useful for macro planning, but it cannot distinguish a large market that remains far below its 2019

level from a smaller market that has already recovered. The DOT's annual reports provide the required monthly market detail [31], while WDI offers comparable source-country economic indicators [32]. Their combination supports a panel architecture in which seasonality and lagged demand are learned jointly but market identity remains visible. This is important for explainability: a tourism authority can review whether a recommendation follows from recent momentum, persistent scale, seasonal timing, market identity, or the source economy.

3. Methodology

3.1 Research Design and Data Sources

The study uses a retrospective predictive design with public, aggregated secondary data. Table 1 lists the sources and their roles. DOT PDF reports were programmatically parsed and manually spot-checked across report layouts. Monthly arrivals were harmonized for South Korea, China, the United States, Japan, Australia, Canada, the United Kingdom, Singapore, Malaysia, India, Germany, France, Indonesia, and Vietnam. These markets represented 84.46% of the DOT's 8,260,913 total arrivals in 2019. The resulting panel contains 2,646 market-month observations from January 2008 to October 2023.

Table 1. Secondary data sources and analytical roles

Source	Coverage used	Variables	Role in study
Philippine DOT visitor-arrival reports [31]	Jan 2008–Oct 2023; monthly; 14 markets	Visitor arrivals by source market	Prediction target, lags, momentum, seasonality, recovery ratios
World Development Indicators [32]	2008–2023; annual; source markets and Philippines	Real GDP per capita, population, official exchange rate, CPI	Lagged economic scale and bilateral price-competitiveness features
Philippine Tourism Satellite Accounts [33]	National annual statistical context	Tourism direct gross value added, expenditure, employment	Decision context; not used as a model predictor

A material measurement caveat was retained rather than silently harmonized away. Reports for 2008–2021 and the country-of-residence section of the 2023 report define the market geographically by residence. The February–December 2022 report is tabulated by nationality. The 2022 fold is therefore treated as a reopening stress test and definitional bridge, not as a directly exchangeable extension of the residence series. No individual traveler records or sensitive personal information were used.

3.2 Leakage-Controlled Feature Engineering

All predictors available for a month t were constructed only from information dated before t or from calendar values known at prediction time. Demand history comprised the logarithms of arrivals lagged by 1, 3, and 12 months; rolling three- and twelve-month means shifted by one month; recent log momentum; and year-over-year log change. Calendar seasonality was represented with sine and cosine transforms of month. Annual source-market real GDP per capita and population were lag-aligned to avoid using future releases. A bilateral exchange-rate competitiveness index combined source and Philippine official exchange rates and CPI, normalized to 2019. Categorical market and regional identifiers allowed partial pooling through one-hot encoding. Closure/reopening flags and months since reopening were known calendar-policy states. Continuous variables were median-imputed within the training pipeline; categorical variables used the training mode.

Figure 2 shows the annual trajectories of the eight largest 2019 markets. The synchronized collapse is visually distinct from normal fluctuations, while the uneven 2022–2023 slopes motivate both the regime gate and market-level decision layer. Annual values represent the months available in each source report; 2023 ends in October.

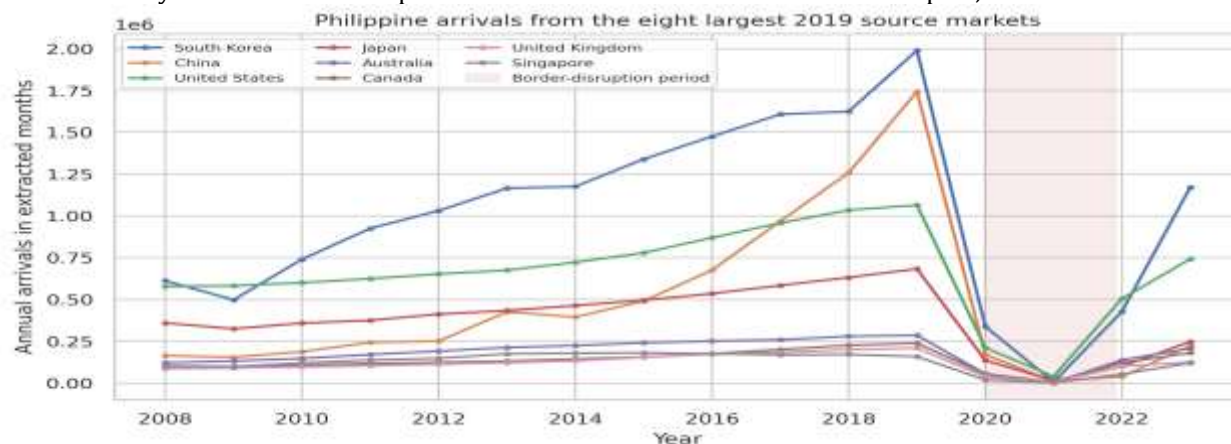


Fig. 2. Annual arrivals in the extracted months for the eight largest 2019 Philippine source markets; 2023 ends in October

3.3 Forecast Models and Regime Gate

The comparison included two transparent benchmarks and seven machine-learning specifications. Persistence predicts the previous month's arrivals; seasonal naive predicts the same month one year earlier. Direct models predict log arrivals using ridge regression, random forest, Extra Trees, and histogram gradient boosting. Residual variants predict the log change from the previous month using ridge, Extra Trees, or gradient boosting and then add that change to the lag-one anchor. Extra Trees was selected as the focal nonlinear method because randomized split thresholds reduce variance across many correlated lag features while retaining nonlinear interactions [19]. Hyperparameters were fixed before holdout evaluation: 500 trees, minimum leaf size 2, square-root feature sampling, and a reproducible random seed. The objective was an auditable comparison rather than a broad hyperparameter tournament.

The regime-gated policy is deliberately simple and available *ex ante*. In a stable, unrestricted demand regime it uses the residual Extra Trees forecast. During reopening or early recovery it uses persistence. This is a model-governance rule rather than another fitted estimator. Its purpose is to prevent a historically accurate nonlinear model from receiving automatic authority when recent observations reflect a structural transition. Table 2 details the temporal folds.

Table 2. Expanding-window temporal evaluation folds

Fold	Training window	Test window	n train	n test	Decision interpretation
2019 pre-shock	Jan 2008–Dec 2018	Jan–Dec 2019	1,680	168	Stable normal-demand regime
2022 reopening	Jan 2008–Feb 2022	Mar–Dec 2022	2,198	140	Stress test; nationality-based bridge
2023 recovery	Jan 2008–Dec 2022	Jan–Oct 2023	2,338	140	Residence-based early recovery

3.4 Evaluation and Uncertainty

Accuracy was assessed with mean absolute error (MAE), root mean squared error (RMSE), weighted absolute percentage error (WAPE), and symmetric mean absolute percentage error (sMAPE). WAPE is the sum of absolute errors divided by the sum of actual arrivals; sMAPE averages twice the absolute error divided by the sum of actual and predicted arrivals. These measures avoid the instability of conventional percentage errors near zero and provide complementary scale-dependent and percentage views [15]. The evaluation used expanding temporal holdouts, because random cross-validation can leak future time-series structure [16,17].

To quantify whether the focal model's advantage over persistence was robust to month-level shocks shared across markets, the analysis used a paired cluster bootstrap. Months within each holdout were resampled with replacement 10,000 times, all 14 markets within a selected month were retained, and the WAPE difference between residual Extra Trees and persistence was recomputed. This resampling unit preserves contemporaneous cross-market dependence. The intervals are empirical 2.5th and 97.5th percentiles; they test predictive error differences and are not causal confidence intervals.

3.5 Exact Grouped Baseline Shapley Explanations

The 2019 residual Extra Trees model was explained because that is the regime in which it earned deployment. Features were grouped into seven managerial concepts: prior demand level; recent momentum; seasonality; source-market economy; exchange-rate competitiveness; disruption and reopening; and market identity and region. For each holdout observation, every one of the 2^7 possible group coalitions was evaluated. Features in a present group took the observation's value; absent groups took the training median or modal baseline. The marginal contribution of each group was averaged over coalition orders using exact Shapley weights. The sum of group attributions plus the baseline reproduced every model prediction to a maximum numerical error of 8.73×10^{-11} arrivals. Because a single baseline can be off the data manifold, results are interpreted as transparent predictive decompositions and accompanied by that limitation [21–23].

3.6 Source-Market Decision Portfolio

The decision layer compares January–October 2023 arrivals with the same months of 2019. Pre-pandemic volume captures strategic scale, while the recovery ratio captures remaining distance from the benchmark. Median splits provide four descriptive action zones: defend recovered high-volume markets, accelerate high-volume shortfalls, reinforce smaller markets with strong momentum, and rebuild or monitor smaller lagging markets. Bubble size represents current 2023 arrivals. This portfolio is a prioritization aid; it does not estimate the causal return on marketing expenditure.

4. Results

4.1 Forecast Accuracy Changes by Regime

The central empirical finding is a reversal rather than a universal machine-learning advantage. In the stable 2019 holdout, residual Extra Trees produced MAE of 3,011 arrivals, RMSE of 6,099, WAPE of 7.25%, and sMAPE of 7.71%.

Persistence produced MAE of 5,455, RMSE of 9,560, WAPE of 13.13%, and sMAPE of 16.58%. The nonlinear residual model therefore reduced WAPE by 5.88 percentage points, or 44.8% relative to persistence. The advantage disappeared after the disruption. In the 2022 reopening fold, residual Extra Trees returned 32.30% WAPE versus 23.87% for persistence. In the 2023 recovery fold, it returned 28.25% versus 15.28%. The result is not that Extra Trees is intrinsically weak: it was the strongest stable-regime model. Rather, learned nonlinear corrections were less dependable than carrying forward the most recent observed market level when demand was rapidly re-anchoring. Figure 3 and Table 3 show this reversal and the resulting gate.

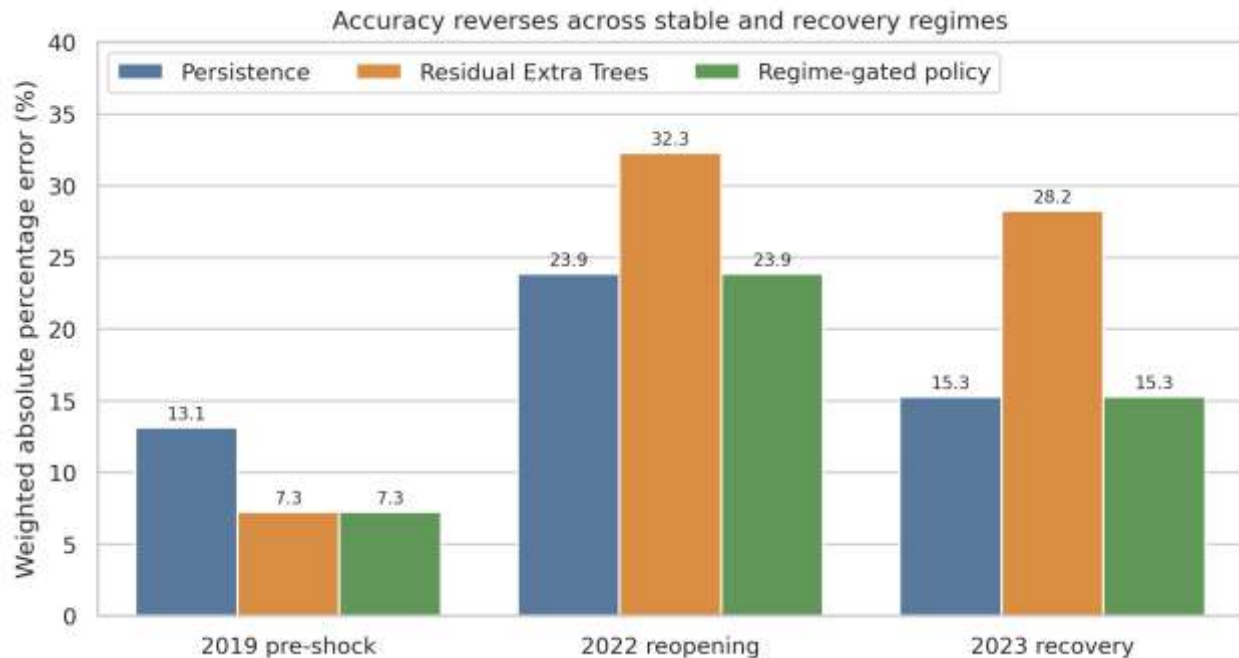


Fig. 3. WAPE of persistence, residual Extra Trees, and the regime-gated policy across stable and transition holdouts

Table 3. Out-of-time accuracy of focal models

Fold	Model	MAE	RMSE	WAPE (%)	sMAPE (%)
2019 pre-shock	Persistence	5,455	9,560	13.13	16.58
2019 pre-shock	Residual Extra Trees	3,011	6,099	7.25	7.71
2019 pre-shock	Regime-gated policy	3,011	6,099	7.25	7.71
2022 reopening	Persistence	2,880	6,083	23.87	36.60
2022 reopening	Residual Extra Trees	3,897	7,727	32.30	48.18
2022 reopening	Regime-gated policy	2,880	6,083	23.87	36.60
2023 recovery	Persistence	3,665	6,663	15.28	18.49
2023 recovery	Residual Extra Trees	6,774	13,735	28.25	24.40
2023 recovery	Regime-gated policy	3,665	6,663	15.28	18.49

Across the three folds, the simple mean WAPE was 17.43% for persistence and 15.47% for the regime-gated policy, an 11.25% relative reduction. This average is descriptive because the folds differ in horizon and definition; the more important result is that the gate never selects the empirically weaker focal method within the prespecified regime logic. Other machine-learning models displayed the same broad pattern: stable-period errors were much lower than transition-period errors, and no complex model consistently displaced persistence in 2022–2023.

4.2 Bootstrap Evidence for the Reversal

The paired month-cluster bootstrap supports the regime interpretation. Table 4 reports the residual Extra Trees WAPE minus persistence WAPE. Negative values favor machine learning. The entire 2019 interval is below zero, while the 2022 and 2023 intervals are above zero. Thus, the reversal is not driven by a single idiosyncratic market-month in the holdouts.

Table 4. Paired month-cluster bootstrap comparison

Fold	Persistence WAPE (%)	Residual ET WAPE (%)	Difference (pp)	95% bootstrap CI (pp)
2019 pre-shock	13.13	7.25	−5.88	[−8.47, −3.56]
2022 reopening	23.87	32.30	+8.43	[+6.23, +11.22]

Fold	Persistence WAPE (%)	Residual ET WAPE (%)	Difference (pp)	95% bootstrap CI (pp)
2023 recovery	15.28	28.25	+12.96	[+2.70, +24.59]

4.3 What the Stable-Regime Model Learned

Global grouped Shapley importance for the 2019 holdout is reported in Table 5 and Figure 4. Prior demand level dominated, with a mean absolute contribution of 31,554 arrivals per market-month. Recent momentum contributed 1,697, seasonality 1,113, and market identity and region 1,053. Source-market economy contributed 190. Exchange-rate competitiveness and disruption/reopening contributed zero in this fitted stable model. The zero values do not establish that exchange rates or border policies are economically irrelevant; they show that the selected trees did not use those blocks to adjust 2019 predictions beyond the other features.

Table 5. Exact grouped baseline Shapley importance for the 2019 holdout

Feature group	Mean absolute contribution (arrivals/month)	Predictive interpretation
Prior demand level	31,554	Persistent market scale and recent arrival anchor
Recent momentum	1,697	Acceleration or deceleration relative to recent history
Seasonality	1,113	Month-specific recurring demand pattern
Market identity and region	1,053	Persistent cross-market heterogeneity
Source-market economy	190	Incremental annual economic context
Exchange-rate competitiveness	0	Not used by fitted 2019 focal model
Disruption and reopening	0	Constant/inactive in the stable holdout

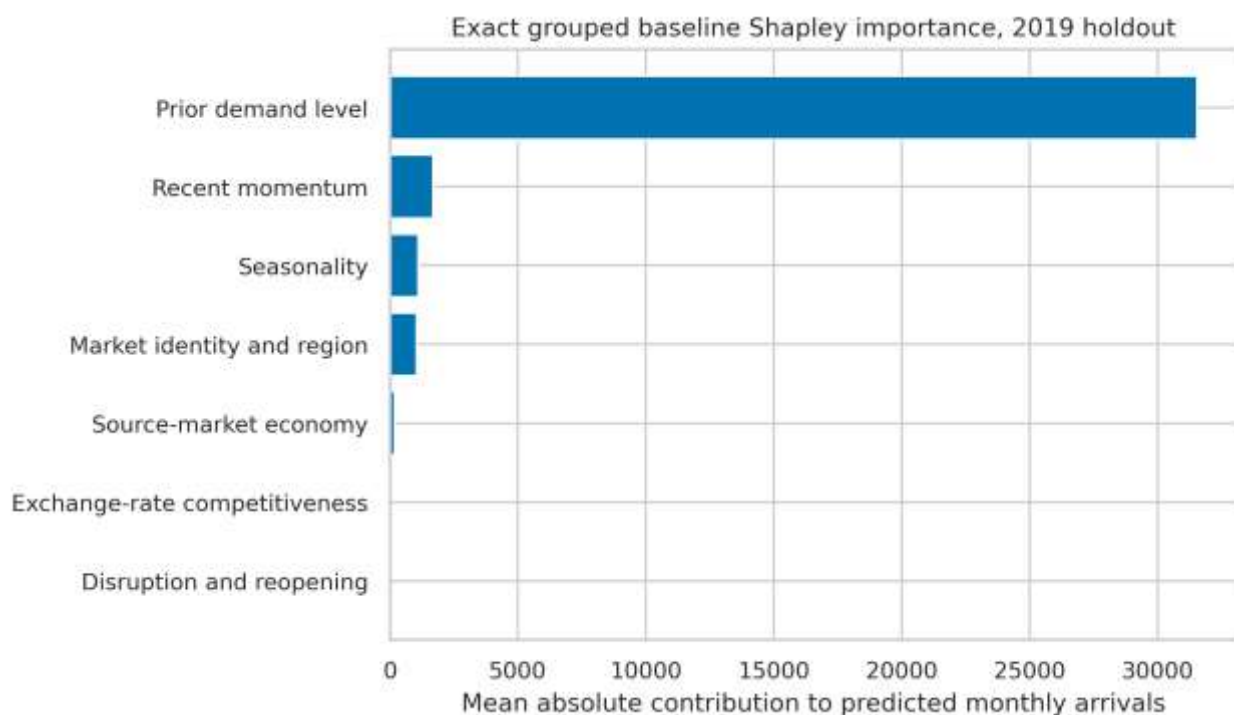


Fig. 4. Global mean absolute grouped baseline Shapley contributions for residual Extra Trees in the 2019 holdout

The explanation profile is also heterogeneous by market. South Korea and China receive large positive prior-demand contributions because their baseline scale is far above the median training observation; several smaller markets receive negative level contributions. Momentum is negative for South Korea, China, Japan, and several European markets but positive for Singapore, Malaysia, Indonesia, and the United States. Figure 5 displays these signed profiles. A key governance benefit is that managers can distinguish a large forecast caused by durable scale from one elevated by recent momentum.

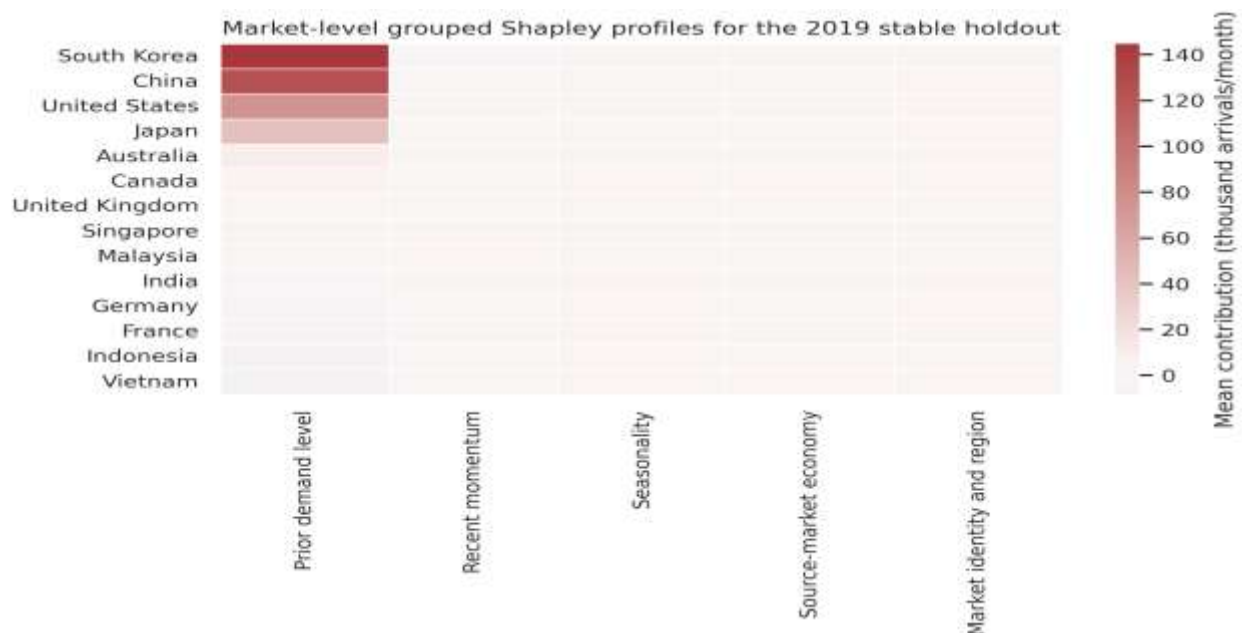


Fig. 5. Mean signed grouped baseline Shapley profiles by market for the 2019 stable holdout

4.4 Recovery-Volume Portfolio

The January–October recovery comparison reveals a second form of heterogeneity. South Korea remained the largest 2023 market in the panel but recovered to 72.74% of its 2019 level. The United States recovered to 85.09%; Australia, Canada, and Singapore exceeded 91%. China, formerly comparable in scale to South Korea, recovered to only 14.90%; Japan reached 43.48%. Vietnam was the only modeled market above its 2019 benchmark, at 106.14%. Figure 6 locates these markets by pre-pandemic volume and recovery ratio, and Table 6 translates the pattern into descriptive action zones.



Fig. 6. Philippine source-market recovery portfolio using January–October 2019 volume and the corresponding 2023 recovery ratio

Table 6. Source-market recovery evidence and descriptive action zones

Market	Jan–Oct 2019	Jan–Oct 2023	Recovery (%)	Decision zone
South Korea	1,609,172	1,170,498	72.74	Accelerate high-volume shortfall
China	1,499,524	223,411	14.90	Accelerate high-volume shortfall
United States	872,335	742,295	85.09	Defend recovered high-volume market
Japan	569,625	247,674	43.48	Accelerate high-volume shortfall
Australia	226,167	207,366	91.69	Defend recovered high-volume market
Canada	187,794	180,625	96.18	Defend recovered high-volume market
United Kingdom	168,954	125,009	73.99	Defend/rebuild near median
Singapore	130,841	119,632	91.43	Reinforce strong recovery

Malaysia	115,619	79,609	68.85	Rebuild and monitor
India	112,247	58,504	52.12	Rebuild and monitor
Germany	82,121	58,001	70.63	Rebuild and monitor
France	73,705	41,839	56.77	Rebuild and monitor
Indonesia	56,488	45,063	79.77	Reinforce momentum
Vietnam	54,316	57,652	106.14	Reinforce above-baseline momentum

The matrix does not imply that the largest shortfall should automatically receive the largest budget. China and Japan warrant investigation because their scale-weighted gaps are large, but channel access, traveler sentiment, seat capacity, and marginal acquisition cost must be examined before intervention. By contrast, the United States, Australia, Canada, and Singapore call for retention and yield management because their recovery is comparatively advanced. Vietnam's above-baseline result suggests a smaller but expanding market in which reinforcement and product adaptation may be more appropriate than broad recovery messaging.

5. Discussion

5.1 Interpretation of the Findings

The study answers RQ1 by showing that model ranking is regime-dependent. Residual Extra Trees captured stable nonlinearities that persistence missed in 2019, yet those learned corrections became liabilities in 2022–2023. The result aligns with the broader forecasting principle that out-of-sample evaluation must reflect the intended deployment environment [15–18]. It also gives tourism resilience arguments [24–30] an operational form: a destination should change not only its marketing strategy after a shock but also the authority granted to its forecasting model.

RQ2 is answered by the grouped Shapley analysis. Stable forecasts were primarily scale forecasts. Prior level carried roughly an order of magnitude more mean absolute contribution than any other group, while momentum, seasonality, and market identity provided smaller corrections. This hierarchy is managerially plausible and easy to audit. It also cautions against attributing a high forecast to economic conditions merely because macro variables are present in the model. In this fit, the economy block was weak and the exchange-rate block inactive after lagged demand and market identity were supplied.

RQ3 is answered by linking explanation to the recovery-volume portfolio. The forecast explains the stable demand mechanism; the portfolio shows where post-shock recovery stands. Together they prevent two common errors: treating every large forecast as recent momentum and treating every low recovery ratio as equally important. The output is a layered decision aid rather than a single ranked list.

5.2 Theoretical and Methodological Contributions

The first contribution is model governance under structural change. Tourism forecasting studies commonly compare algorithms, but the deployable decision can be framed as model selection conditional on an observable regime. The gate is transparent, falsifiable, and operational: machine learning is used only where it has demonstrated stable-regime value. This moves the research question from “Which algorithm wins?” to “Under which demand regime should this algorithm be trusted?”

The second contribution is explanation at a managerial level of abstraction. Exact Shapley computation over seven groups avoids distributing attention across many highly related lags. It also provides a numerical efficiency check, making the additive explanation auditable. The third contribution is the source-market panel itself: national Philippine totals are decomposed into 14 strategically meaningful markets and linked to comparable international indicators. The fourth contribution is the explicit publication of AI failure during transition. Negative performance evidence is essential for responsible decision support because it defines the model's boundary of competence.

5.3 Practical Implications

For the DOT and destination-management organizations, the first implication is to maintain a champion–challenger process by regime. During stable operations, residual Extra Trees can serve as champion and persistence as challenger. During reopening, schedule shocks, or abrupt policy change, that relationship should reverse until rolling evidence shows that nonlinear corrections again add value. The gate can be monitored using trailing WAPE and a documented transition flag.

Second, forecast reports should show group explanations beside point estimates. A source-market recommendation supported mainly by prior scale calls for capacity and retention planning; one supported by momentum calls for investigation of recent drivers; one supported by seasonality calls for timing rather than permanent reallocation. Third, recovery budgets should separate scale-weighted gaps from growth opportunities. China and Japan are high-volume shortfalls; Vietnam is a smaller momentum opportunity; the United States, Australia, Canada, and Singapore are retention priorities. Each action should subsequently be tested against air-seat capacity, spending yield, and campaign cost.

5.4 Limitations and Future Research

Several limitations bound the findings. The 2022 report classifies markets by nationality rather than residence, so that fold is a stress test with a definitional break. The panel ends in October 2023 and does not establish whether the persistence advantage continued after recovery matured. Annual WDI variables are low frequency and subject to release lags and revision; the lag alignment reduces but cannot eliminate vintage uncertainty. The 14-market

panel covers most, not all, arrivals. The regime gate is a prespecified rule and was evaluated on only three holdouts; it is not yet a learned change-point detector.

The explanations are predictive and baseline-dependent. Correlated lag features can share attribution, and counterfactual coalitions may be unrealistic. No Shapley value should be interpreted as the causal effect of marketing, exchange rates, or policy. Future research should extend the panel with 2024–2026 data, harmonize residence-based definitions, add airline seat capacity and digital-intent variables, compare learned change-point gates, and evaluate probabilistic forecasts. A prospective test should pre-register the regime rule and measure whether explanation-supported decisions improve budget allocation, not only forecast error.

6. Conclusions

This study developed a regime-gated XAI framework for Philippine tourism source-market decisions using 2,646 observations from 14 markets over 2008–2023. Residual Extra Trees was clearly superior to persistence in the stable 2019 holdout, but clearly inferior during 2022 reopening and 2023 recovery. Exact grouped Shapley values showed that the stable model primarily learned prior market scale, with smaller adjustments from momentum, seasonality, and market identity. A transparent gate—machine learning in stable demand, persistence during transition—converted this evidence into a model-governance policy and reduced mean fold WAPE by 11.25% relative to always using persistence.

The novelty is the integration of regime-specific validation, exact grouped explanations, and a recovery-volume portfolio for Philippine source markets. The paper does not argue that explainable AI is always better. It establishes when the AI model earned trust, documents when it did not, and converts that boundary into a practical decision rule. For intelligent decision-making systems, this is a more consequential form of explainability: the system explains both its forecasts and the conditions under which managers should decline to use them.

Author Contributions

Karen Razelle M. Duyan: Conceptualization, investigation, supervision, project administration, writing—original draft, and writing—review and editing. Juan Moshe M. Duyan: Conceptualization, investigation, data curation, validation, writing—original draft, and writing—review and editing. Valerie A. Abesamis Paloy: Methodology, investigation, data curation, validation, writing—original draft, and writing—review and editing. Eric A. Paloy: Investigation, data curation, formal analysis, validation, writing—original draft, and writing—review and editing. Melanie S. Manuel: Investigation, data curation, formal analysis, validation, writing—original draft, and writing—review and editing. Mark Joby Aguilar: Software, data curation, visualization, validation, writing—original draft, and writing—review and editing. Joe Robert G. Lucena: Methodology, software, formal analysis, data curation, visualization, writing—original draft, and writing—review and editing. Rissa A. Lasap: Conceptualization, methodology, supervision, project administration, writing—original draft, and writing—review and editing. Luigi Carlo M. De Jesus: Conceptualization, methodology, software, formal analysis, writing—original draft, and writing—review and editing. All authors have read and approved the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable. The study used publicly available, aggregated secondary data and did not involve human participants or personally identifiable information.

Informed Consent Statement

Not applicable.

Data Availability Statement

The underlying arrival reports are publicly available from the Philippine Open Data Portal and Department of Tourism [31]. Macroeconomic indicators are publicly available through the World Bank API [32]. The harmonized analytical panel and reproducibility code are available from the corresponding authors upon reasonable request.

Conflicts of Interest

The authors declare no conflict of interest.

Declaration of Generative AI and AI-Assisted Technologies

During manuscript preparation, the authors used ChatGPT by OpenAI for language refinement, document formatting, and assistance drafting reproducible analysis code. The authors reviewed and verified the data extraction, analysis, results, interpretations, and references and take full responsibility for the content of the manuscript.

Acknowledgment

The authors acknowledge the Philippine Department of Tourism, the Philippine Statistics Authority, and the World Bank for maintaining public statistical resources used in this study.

References

- [1] Song, H., Qiu, R. T. R., & Park, J. (2019). A review of research on tourism demand forecasting: Launching the Annals of Tourism Research Curated Collection on tourism demand forecasting. *Annals of Tourism Research*, 75, 338–362. <https://doi.org/10.1016/j.annals.2018.12.001>
- [2] Song, H., & Li, G. (2008). Tourism demand modelling and forecasting—A review of recent research. *Tourism Management*, 29(2), 203–220. <https://doi.org/10.1016/j.tourman.2007.07.016>
- [3] Witt, S. F., & Witt, C. A. (1995). Forecasting tourism demand: A review of empirical research. *International Journal of Forecasting*, 11(3), 447–475. [https://doi.org/10.1016/0169-2070\(95\)00591-7](https://doi.org/10.1016/0169-2070(95)00591-7)
- [4] Gunter, U., & Önder, I. (2015). Forecasting international city tourism demand for Paris: Accuracy of uni- and multivariate models employing monthly data. *Tourism Management*, 46, 123–135. <https://doi.org/10.1016/j.tourman.2014.06.017>
- [5] Bangwayo-Skeete, P. F., & Skeete, R. W. (2015). Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management*, 46, 454–464. <https://doi.org/10.1016/j.tourman.2014.07.014>
- [6] Wen, L., Liu, C., & Song, H. (2019). Forecasting tourism demand using search query data: A hybrid modelling approach. *Tourism Economics*, 25(3), 309–329. <https://doi.org/10.1177/1354816618768317>
- [7] Wen, L., Liu, C., Song, H., & Liu, H. (2021). Forecasting tourism demand with an improved mixed data sampling model. *Journal of Travel Research*, 60(2), 336–353. <https://doi.org/10.1177/0047287520906220>
- [8] Zhang, Y., Li, G., Muskat, B., & Law, R. (2021). Tourism demand forecasting: A decomposed deep learning approach. *Journal of Travel Research*, 60(5), 981–997. <https://doi.org/10.1177/0047287520919522>
- [9] Sun, S., Li, Y., Guo, J., & Wang, S. (2022). Tourism demand forecasting: An ensemble deep learning approach. *Tourism Economics*, 28(8), 2021–2049. <https://doi.org/10.1177/13548166211025160>
- [10] Sun, S., Wei, Y., Tsui, K.-L., & Wang, S. (2019). Forecasting tourist arrivals with machine learning and internet search index. *Tourism Management*, 70, 1–10. <https://doi.org/10.1016/j.tourman.2018.07.010>
- [11] Li, X., Li, H., Pan, B., & Law, R. (2021). Machine learning in internet search query selection for tourism forecasting. *Journal of Travel Research*, 60(6), 1213–1231. <https://doi.org/10.1177/0047287520934871>
- [12] Zhang, Y., Li, G., Muskat, B., Law, R., & Yang, Y. (2020). Group pooling for deep tourism demand forecasting. *Annals of Tourism Research*, 82, 102899. <https://doi.org/10.1016/j.annals.2020.102899>
- [13] Sancho Núñez, J. C., Gómez-Pulido, J. A., & Robina Ramírez, R. (2024). Machine learning applied to tourism: A systematic review. *WIREs Data Mining and Knowledge Discovery*, 14(5), e1549. <https://doi.org/10.1002/widm.1549>
- [14] De Jesus, N. M., & Samonte, B. R. (2023). AI in tourism: Leveraging machine learning in predicting tourist arrivals in Philippines using artificial neural network. *International Journal of Advanced Computer Science and Applications*, 14(3). <https://doi.org/10.14569/IJACSA.2023.0140393>
- [15] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688. <https://doi.org/10.1016/j.ijforecast.2006.03.001>
- [16] Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120, 70–83. <https://doi.org/10.1016/j.csda.2017.11.003>
- [17] Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4), 437–450. [https://doi.org/10.1016/S0169-2070\(00\)00065-0](https://doi.org/10.1016/S0169-2070(00)00065-0)
- [18] Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20(1), 134–144. <https://doi.org/10.1198/073500102753410444>
- [19] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- [20] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- [21] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67. <https://doi.org/10.1038/s42256-019-0138-9>
- [22] Molnar, C., König, G., Herbringer, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M., & Bischl, B. (2022). General pitfalls of model-agnostic interpretation methods for machine learning models. In *xxAI—Beyond Explainable AI* (pp. 39–68). Springer. https://doi.org/10.1007/978-3-031-04083-2_4
- [23] Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4), 1059–1086. <https://doi.org/10.1111/rssb.12377>
- [24] Gössling, S., Scott, D., & Hall, C. M. (2021). Pandemics, tourism and global change: A rapid assessment of COVID-19. *Journal of Sustainable Tourism*, 29(1), 1–20. <https://doi.org/10.1080/09669582.2020.1758708>
- [25] Sigala, M. (2020). Tourism and COVID-19: Impacts and implications for advancing and resetting industry and research. *Journal of Business Research*, 117, 312–321. <https://doi.org/10.1016/j.jbusres.2020.06.015>

- [26] Ioannides, D., & Gyimóthy, S. (2020). The COVID-19 crisis as an opportunity for escaping the unsustainable global tourism path. *Tourism Geographies*, 22(3), 624–632.
<https://doi.org/10.1080/14616688.2020.1763445>
- [27] Sharma, G. D., Thomas, A., & Paul, J. (2021). Reviving tourism industry post-COVID-19: A resilience-based framework. *Tourism Management Perspectives*, 37, 100786. <https://doi.org/10.1016/j.tmp.2020.100786>
- [28] Assaf, A., & Scuderi, R. (2020). COVID-19 and the recovery of the tourism industry. *Tourism Economics*, 26(5), 731–733. <https://doi.org/10.1177/1354816620933712>
- [29] Yang, Y., Zhang, H., & Chen, X. (2020). Coronavirus pandemic and tourism: Dynamic stochastic general equilibrium modeling of infectious disease outbreak. *Annals of Tourism Research*, 83, 102913.
<https://doi.org/10.1016/j.annals.2020.102913>
- [30] Zenker, S., & Kock, F. (2020). The coronavirus pandemic—A critical discussion of a tourism research agenda. *Tourism Management*, 81, 104164. <https://doi.org/10.1016/j.tourman.2020.104164>
- [31] Department of Tourism. (2023). Tourism demand statistics on visitor arrivals to the Philippines, 2008–2023 [Data set]. Open Data Philippines. <https://data.gov.ph/index/public/dataset/tourism-demand-statistics-on-visitor-arrivals-to-the-philippines/pqj5mu1b-5tsg-7dcb-oj3y-s1jcu1ys2h3a>
- [32] World Bank. (2026). World Development Indicators [Data set].
<https://databank.worldbank.org/source/world-development-indicators>
- [33] Philippine Statistics Authority. (2026). Philippine Tourism Satellite Accounts.
<https://psa.gov.ph/statistics/philippine-tourism-satellite-accounts>