MERCon 2026
MORATUWA ENGINEERING RESEARCH CONFERENCE

The Effect of Search Intent Data on Predicting Post-Crisis Daily Tourist Arrivals in Sri Lanka

Tharusha Perera, Janith Mahanama, Maleesha Kumarasinghe, Bimsara Udurawana, Praveen Nawarathna, Pataliee Narasinghe, Nathali Athukorala, Sandaneka Wickramanayake, Nisansa de Silva
Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka

Abstract

- Daily arrival forecasting is hard during Sri Lanka's **post-crisis recovery (2023–2025)**.
- Existing methods use only Google Trends, missing markets where **Yandex** dominates (e.g. Russia).
- We fuse arrivals with **Yandex** + Google search intent, exchange rate, and weather; features screened by **Granger + CCF** on training data only.
- **Best model: SVR (RBF), 8.94% MAPE, 736 RMSE.** Yandex queries beat (all-flat) Russian Google Trends.

Post-Crisis Context & Contributions

- **Polycrisis:** 2019 Easter bombings (−70%), COVID-19 (near-zero for 18 months), 2022 economic collapse, a highly unstable series.
- Multi-source daily panel (2023–2025): SLTDA counts [1,2] + macro, weather, Google & **Yandex** search intent.
- Leakage-free **structural-break** pipeline (Chow [3] + CUSUM) and **Granger** [4] **+ CCF** lag audit proving Yandex's superiority for Russia.
- Formal **Data–Model Alignment** (V1/V2/V3 matrices) + cross-model **SHAP** [5] verifying signal drivers.

Dataset: Multi-Source Daily Panel

- **Span:** Jan 2023 – Dec 2025 (≈1,095 days); five public streams left-joined on the SLTDA date index.
- **Arrivals (SLTDA) [1,2]:** total daily international arrivals, ∼2k–12k/day (the target).
- **USD/LKR rate:** daily macro cost-of-travel signal.
- **Weather:** Colombo temperature (°C) + precipitation (mm) via Open-Meteo.
- **Search intent:** localized weekly Search Volume Indices per market (Google + **Yandex**).
- **Integration [6]:** weekly series forward- then backward-filled to daily; *no future leakage*. Target $y_t = \log_{10}(\text{Count}_t + 10)$; split at **2024-06-29**.

Market	Engine	Example queries
Global	Google	travel, flights, visa
UK	Google	"Sri Lanka ETA", holidays
Germany	Google	"Urlaub", "Flüge Colombo"
India	Google	tour package, Colombo flight
Russia	**Yandex**	"Sri Lanka tour", flights to Colombo

Table 1. Search-intent markets, engines and representative queries.

Structural Break Detection

- Non-stationary: ADF p=0.138, KPSS p<0.01.
- Rolling **Chow** [3] F-scan (7-day grid, 90-day buffer, BH-FDR [7] α=0.05) + **CUSUM** (3.26, $p = 1.15\times10^{-9}$).
- ⇒ **four breaks, five economic regimes** used as stress-test dimensions.

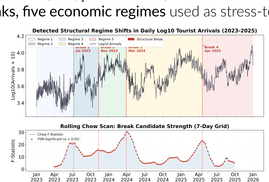


Figure 1. Regime shifts (top) and rolling Chow F-scan with FDR-significant candidates (bottom).

Temporal Dependence Structure

- ACF/PACF of y_t show *daily inertia + weekly seasonality*.
- Lags $\{1, 2, 7, 14\}$ and seasonal period 7 enter every model.

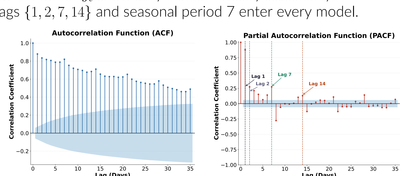


Figure 2. Autocorrelation (left) and partial autocorrelation (right); key lags 1, 2, 7, 14.

Search-Intent Bias: Yandex vs. Google

- Retain feature if peak $|r| \geq 0.10$ and Granger [4] $p < 0.20$.
- **Anglocentric bias:** Russian *Google Trends* [8] is a flat, zero-variance constant, *no* predictive power [9].
- **Yandex tour** leads arrivals by ∼27 days (r=0.423) [10].
- Retained: `Yandex_RU_Tour`, `Yandex_RU_Flights`, `IN_SriLankatourpackage`.

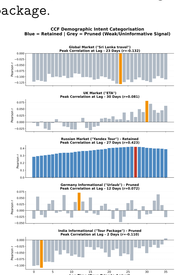


Figure 3. CCF by market/intent. Blue = retained, grey = pruned; peak lag per panel.

Feature–Model Alignment (V1/V2/V3)

Not every model can safely share one feature set, so three geometry-aligned matrices are built [11,12]:

- **V1 All-Weather** (SVR [13], XGBoost [14], RF): lags 1/2/7/14 + 3 exogenous + macro/weather + day-of-week; 1,065 rows.
- **V2 Window** (LSTM [15]): $\pm\frac{w-1}{2}$-day window; w=7 (13.79% MAPE) > w=1 (18.54%); 26 features.
- **V3 Strictly Independent** (SARIMAX [16]): `Yandex_RU_Tour` + `IN_pkg` only, since Yandex Tour/Flights are collinear (r=0.92).

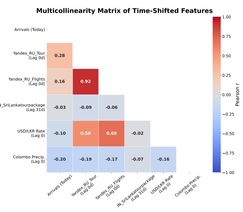


Figure 4. Pearson correlation of lag-aligned features (training window).

Model Leaderboard (550-Day Rolling Test)

- One-step-ahead rolling forecast; weekly retraining on accumulated data.
- **Online learning is required:** static Pure SARIMA 24.57% vs. 9.48% with retraining (∼15 pp).

Model	MAPE (%)	RMSE	MAE
SVR (All-Weather, RBF)	**8.94**	736	552
SARIMAX	9.08	742	561
Grid-Search SVR	9.16	751	563
Hybrid (SARIMAX + XGBoost)	9.25	743	565
Pure SARIMA	9.48	750	578
XGBoost	9.59	822	601
LSTM V2	10.58	903	678

Table 2. Final leaderboard. SVR wins on all metrics; SARIMAX close and best-calibrated.

Regime Stress Test & Uncertainty

- **SVR** degrades least across regimes (8.70% → 9.20%).
- Tree ensembles fail by extrapolation (XGBoost 8.78% → 10.52%).
- For planning, **SARIMAX** gives the *narrowest* 95% interval (width 2,833, coverage 94.9%).

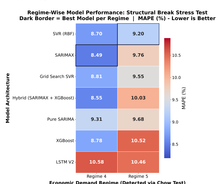


Figure 5. Regime-wise MAPE; dark border = best model per regime.

Explainability via SHAP

- `Arrivals_Lag_1` dominates every ML model (daily inertia).
- `Yandex_RU_Tour` is the top *exogenous* driver (importance 1.00 in the hybrid corrector).
- Weather variables contribute negligible SHAP [5]: statistical noise.

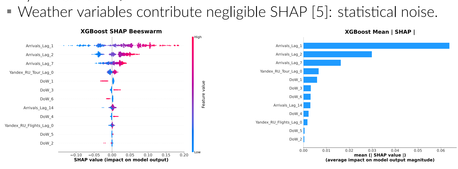


Figure 6. SHAP beeswarm (left) and mean |SHAP| importance (right) for XGBoost.

Managerial Implications & Conclusion

- **27-day early-warning system:** Russian Yandex demand leads arrivals by ∼3 weeks, giving time to adjust pricing, staffing and fuel/food stockpiles.
- **Market-specific campaign timing** follows each market's distinct booking horizon.
- **Plan for uncertainty:** SARIMAX confidence bounds enable safe operational buffers over point forecasts.

Conclusion: *data engineering* (breaks, exact lags, search-bias audit, feature-geometry alignment) drives accuracy; model choice is secondary. Target the Russian market with Yandex, not Google.

References

[1] SLTDA, *Annual Report* 2020, 2021.
[2] SLTDA, *Annual Report* 2023, 2024.
[3] Chow, "Tests of equality between sets of coefficients in two linear regressions," *Econometrica* 591–605, 1960.
[4] Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica* 424–438, 1969.
[5] Lundberg & Lee, "A unified approach to interpreting model predictions," *NeurIPS* 30, 2017.
[6] Snrantha, "Sri Lanka document datasets," a00v2510.04124, 2025.
[7] Benjamini & Hochberg, "Controlling the false discovery rate," *J. R. Stat. Soc. B* 57(1):289–300, 1995.
[8] Dvder, "Forecasting tourism demand with Google Trends," *Int. J. Tourism Res.* 19(6):648–660, 2017.
[9] Wickramasinghe & Ratnasiri, "Disaggregated search data in tourism forecasts: Sri Lanka," *Current Issues in Tourism* 24(19):2740–2754, 2021.
[10] Hewapathirana, "Advancing tourism demand forecasting in Sri Lanka," *J. Tourism Futures* 11(2):243–265, 2025.
[11] Song & Li, "Tourism demand modelling and forecasting," *Tourism Mgmt.* 29(2):203–220, 2008.
[12] Yu & Schwartz, "Forecasting short time-series tourism demand with AI models," *J. Travel Res.* 45(2):194–203, 2006.
[13] Vapnik, *The Nature of Statistical Learning Theory*, Springer, 2013.
[14] Chen & Guestrin, "XGBoost: A scalable tree boosting system," *KDD* 785–794, 2016.
[15] Hochreiter & Schmidhuber, "Long short-term memory," *Neural Comput.* 9(8):1735–1780, 1997.
[16] Box, Jenkins, Reinsel & Ljung, *Time Series Analysis: Forecasting and Control*, Wiley, 2015.

ORGANIZED BY
FINANCIAL SPONSOR
TECHNICAL CO-SPONSORS