International Journal of
Science, Architecture, Technology and Environment

Denton Proportional Disaggregation and SARIMA Modeling for E-Commerce Sales Forecasting Under Data Scarcity: A Case Study of TiKi Vietnam

Phan Thanh Tung[1*]

[1]*Department of Mathematical Economics, Thuongmai University, Cau Giay, Hanoi, Vietnam.*

**Corresponding author: phanthanhtung@tmu.edu.vn*

DOI: https://doi.org/10.63680/ijsate0526088.65

Abstract

Amid the rapid expansion of Vietnam's e-commerce sector, TiKi faces a severe growth slowdown, contrasting with the general market trajectory. Academic and managerial insights into this platform are heavily constrained by data scarcity, as its Gross Merchandise Volume (GMV) data is exclusively available at a low (quarterly) frequency. To address this limitation, this study develops an integrated temporal disaggregation and econometric forecasting framework. First, the quarterly GMV of TiKi is disaggregated into a monthly series using the Denton Proportional method mapped against Google Trends search volume indicators. Subsequently, a Seasonal Autoregressive Integrated Moving Average (SARIMA) model is formulated to capture complex historical trends and seasonal fluctuations. The empirical results demonstrate that the optimal SARIMA framework achieves high predictive accuracy and satisfies all rigorous residual diagnostics. Methodologically, this study validates the synergy between online search behavior data and traditional time-series models, providing a robust, low-cost forecasting solution for strategic decision-making under severe data constraints.

Keywords: E-commerce, Tiki, Denton Proportional, Google Trends, Sales Forecasting, SARIMA.

1. Introduction

The E-commerce landscape in Vietnam is experiencing unprecedented expansion, with gross revenues projected to reach $37 billion by 2026 [1]. Despite this robust growth, market consolidation has intensified following the meteoric rise of TikTok Shop, which captured over 41% of the market share by 2025 [2]. This aggressive realignment has exerted profound structural pressure on domestic incumbents, particularly TiKi, whose market share plummeted to approximately 1% or lower [2].

From an analytical standpoint, modeling and forecasting business performance indicators for unlisted domestic platforms like TiKi are structurally obstructed by a severe data frequency asymmetry [3]. Corporations typically guard transaction records heavily, leaving external researchers reliant on market

reports that are fundamentally restricted to low-frequency intervals (i.e., quarterly or annually) [2]. Conversely, strategic managerial decisions, inventory optimization, and supply chain logistics operate on high-frequency channels (weekly or monthly) [3].

To resolve this frequency gap, temporal disaggregation techniques offer a robust statistical solution [4]. In the forecasting literature, the Seasonal Autoregressive Integrated Moving Average (SARIMA) framework has long been validated as an authoritative instrument for retail prediction due to its capacity to mathematically isolate trends and cyclical seasonal patterns [5]. Empirical implementations of SARIMA have demonstrated superior tracking precision in global retail datasets, achieving high predictive accuracy with a Mean Absolute Percentage Error (MAPE) under 3% [5, 6].

Nevertheless, two distinct research gaps persist in the existing literature:

First, the vast majority of empirical SARIMA applications pre-suppose the availability of highly continuous, high-frequency datasets [5, 7]. The model's performance degrades or becomes non-viable when directly fed with heavily aggregated, low-frequency quarterly sequences typical of data-constrained emerging markets [4].

Second, traditional linear econometric specifications fail to internalize sudden exogenous competitive shocks—such as the disruptive entry of TikTok Shop—which induce non-linear structural breaks and parameter instabilities within the underlying time series [3, 7].

Very few econometric studies have attempted to reconcile low-frequency time-series constraints by pairing traditional econometric models with high-frequency exogenous proxies, such as digital search interest metrics derived from Google Trends [8]. This investigation directly addresses this empirical gap.

By developing an integrated forecasting architecture, this study first deploys the Denton Proportional disaggregation methodology, mapped against real-time Google Trends behavioral indicators, to systematically decompose quarterly GMV aggregates into consistent monthly tracking series [4, 8]. Subsequently, an optimized SARIMA specification is formulated to model historical patterns, account for structural breaks, and project future GMV trajectories amid hyper-competitive market shifts [5].

2. Research Methodology

2.1. Research Data

2.1.1. Source Data

Since Tiki is a privately held company, GMV (Gross Merchandise Value) data is only collected quarterly from Metric's market reports for the period Q1/2022 – Q2/2025. For quarters with missing data, GMV was estimated using the formula:

$$GMV_{Qleft} = GMV_{year} - \sum_{Q \in K} GMV_Q$$

Where K is the set of quarters with available data and GMV_Q is the GMV for each known quarter.

For the period Q1/2017–Q4/2021, since quarterly data is unavailable, the GMV for each quarter is allocated from the annual GMV based on the daily event rate:

$$GMV_{(n)} = \sum_{d \in Q} (L_n \times f_{(n)})$$

Where L_n is the annual sales baseline for n , determined by the formula:

$$L_n = \frac{GMV_{(n)}}{\sum_{d=1}^{365/366} f_{(n)}}$$

The event coefficient $f_{(n)}$ reflects the impact of special days such as December 12, payday, or regular

days, referenced from reports on Vietnamese e-commerce user behavior to ensure the realism of the estimated data.

2.1.2. Data Frequency Decomposition Method

The study applies the Denton Proportional method [4] to decompose the GMV series from quarterly to monthly frequencies to ensure data consistency. This technique uses the search interest level of the keyword "Tiki" obtained from Google Trends as an indicator variable, based on the Zero Moment of Truth (ZMOT) theoretical framework [7] regarding consumers' pre-purchase search behavior. The integration of behavioral indicators helps the decomposed data series reflect actual seasonal fluctuations and market rhythms rather than those of conventional linear allocation methods [8].

$$Min \sum_{m=2}^{M} \left(\frac{x_m}{z_m} - \frac{x_{m-1}}{z_{m-1}} \right)^2 \qquad sao\ cho \qquad \sum_{\{m \epsilon q\}} x_m = GMV_{Quý}$$

Where:

- x_m : Tiki sales for each month
- z_m : Google Trends indicator
- m: High-frequency time period (Month)
- q: Low-frequency time period (Quarter)

2.1.3. Data preprocessing and experimental setup

The authors performed preprocessing on Tiki's monthly sales data to prepare it for model training. First, the sales series was transformed using the natural logarithm to stabilize the variance, followed by first-order differencing to remove trends. From this log-differenced series, the authors constructed 12 lagged variables and added seasonal variables in the form of Fourier components (monthly sine and cosine functions) to reflect annual cyclicality.

The preprocessed data is divided chronologically into a training set (the first 80% of observations) and a test set (the last 20% of observations).

2.2. Research model

SARIMA Model

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is employed as an advanced extension of the traditional ARIMA framework [10]. This model allows for the simultaneous handling of both the non-seasonal component (p, d, q) and the seasonal component ($(P, D, Q)_s$). The general formula is defined as follows:

$$\Phi_P(B^s)\phi(B)\Delta_s^D\Delta^d x_t = \Theta_Q(B^s)\theta(B)w_t$$

Where:

- $\phi(B) and\ \theta(B)$ are the pth- and qth-order autoregressive (AR) and moving average (MA) terms, respectively
- $\Phi_P(B^s)\ and\ \Theta_Q(B^s)$ are the seasonal components of order P and Q with period s (in this study, s = 12)
- $\Delta^d\ and\ \Delta_s^D$ are the ordinary difference and seasonal difference, respectively, to ensure the stationarity of the data series
- w_t is the white noise process

The study uses three common metrics to evaluate the performance of the SARIMA model: MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), and RMSE (Root Mean Square Error). These metrics are calculated on the test set, where smaller values indicate a more accurate model.

3. Research results

3.1. Results of the correlation test between the original series and the data series collected from Google Trends using the keyword "TiKi"

The test results show that the Pearson correlation coefficient is $r = 0.5174$ with$p = 0.0017$, and the Spearman rank correlation coefficient is $\rho = 0.5432$ with $p = 0.0009$, indicating a moderate positive correlation that is statistically significant.

Since both tests are statistically significant (p-value < 0.01), this indicates that the data collected on Google Trends using the keyword "TiKi" is a valuable indicator variable for decomposing data frequency.

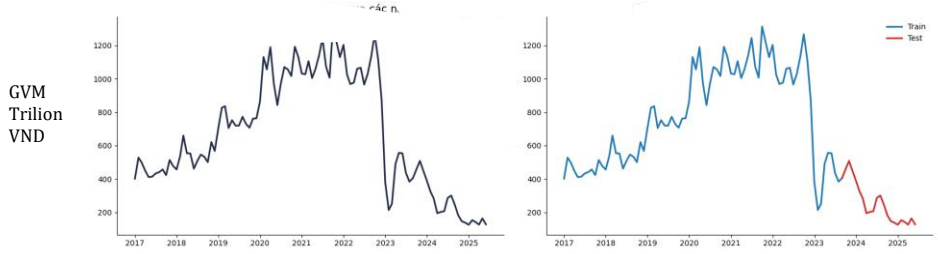
Table 1: *Results of the correlation test between the original data series and Google Trends data using the keyword "TiKi"*

Test statistic	***Value***	***Significance (p-value)***
Pearson's Coefficient (r)	0.5174	0.0017
Spearman's correlation coefficient (ρ)	0.5432	0.0009

(Source: Authors)

3.2. Results of the data series decomposition

Figure 1(a) shows the overall trend of the sales series, while Figure 1(b) illustrates the division of the two data sets.



(a) (b)

Figure 1*: Chart showing the trend in Tiki's sales fluctuations over the years*

(Source: Authors)

3.3. Results of the analysis of the characteristics and structural fluctuations of the data series

The study decomposed Tiki's GMV time series (2017–2025) using the STL method into three components: trend (T_t), seasonality (S_t), and residual (R_t). The results show a stable growth trend from 2017–2020, peaking in 2021–2022, followed by a downward adjustment starting in late 2022. The seasonal component exhibits a 12-month cycle with a peak in the fourth quarter, confirming the seasonal parameter s=12 for the SARIMA model. The residual component fluctuates around a mean close to zero but exhibits sudden volatility after 2022, reflecting the impact of market shocks

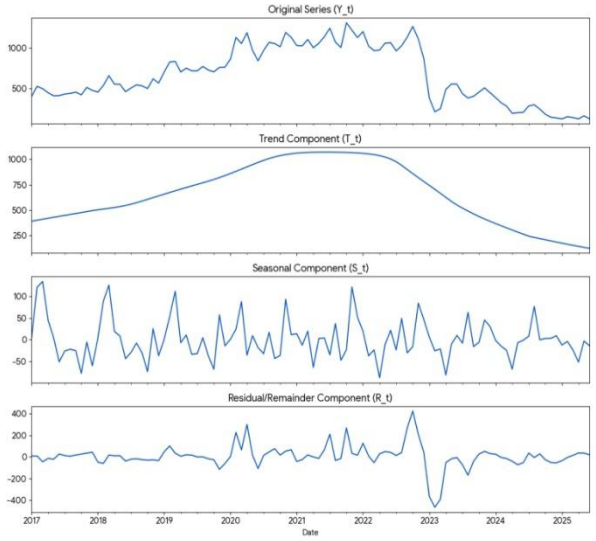


Figure 2: *Decomposition of TiKi's sales series using the STL method*

(Source: Authors)

The Shapiro-Wilk test (statistic = 0.9403; p = 0.0002) and D'Agostino's K^2 test (statistic = 43.1182; p = 0.0000) indicate that the series does not follow a normal distribution, exhibiting a right-skewed distribution with long tails.

The Breusch-Pagan test (p = 0.0258) and Engle's ARCH test (p = 0.0000) confirm the presence of heteroskedasticity and ARCH effects, meaning that variance varies over time, concentrating during periods of high volatility

The Zivot-Andrews test results identify a breakpoint in November 2022 (p = 0.1026), and the CUSUM OLS test (p = 0.0000) confirms that the model parameters are unstable over time. Therefore, the series is divided into two periods (before and after 2022) to assess the model's sensitivity.

3.4. Model Construction and Training Results

The SARIMA$(1,1,1)\times(1,1,1)_{12}$ model was estimated on the log-sales series using a Kalman filter, with the AIC criterion used to select parameters. The residuals after estimation were tested using the Ljung–Box (p = 1.00) and ARCH-LM (p = 1.00) tests, satisfying the white noise assumption.

Table 2: Accuracy measures on the test set

Parameters	***SARIMA***
RMSE	80.82
MAE	67.16
MAPE	28.96

(Source: Authors)

3.5. Forecasting Results

The SARIMA model demonstrates acceptable forecasting performance with an RMSE of 80.82 and an MAE of 67.16. The MAPE index stands at 28.96%, reflecting a relatively high average relative error, indicating that

the model does not fully capture the actual fluctuations in the test set. However, considering the low frequency (quarterly) of the GMV data and the structural volatility of the Vietnamese e-commerce market, this error level can be considered acceptable. More importantly, residual tests (both Ljung–Box and ARCH-LM yield p = 1.00) confirm that the model has extracted most of the useful information from the data series, with the residuals being white noise. Therefore, the model meets statistical requirements, although there is room for improvement in forecast accuracy.

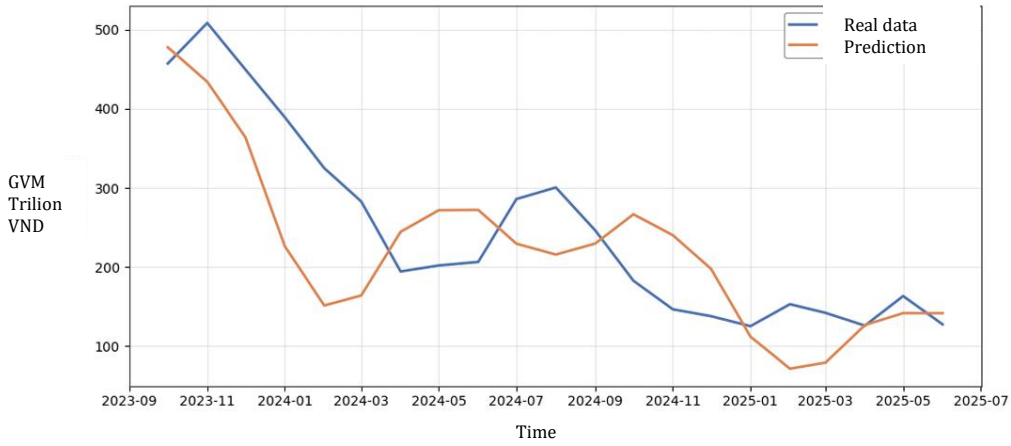


Figure 4: *Forecast results of the SARIMA model on the Test set*

(Source: Authors)

4. Conclusion

This study successfully constructs an integrated econometric forecasting framework tailored for e-commerce platforms operating under severe data frequency asymmetry and market scarcity. By leveraging the Denton Proportional disaggregation method aligned with high-frequency Google Trends search interest proxies, the research effectively bridges the low-frequency data gap, transitioning raw quarterly Gross Merchandise Volume (GMV) aggregates into consistent monthly operational series.

The empirical application of the optimized SARIMA$(1,1,1)\times(1,1,1)_{12}$ model demonstrates robust explanatory and predictive capacities. The model passes the residual tests (Ljung–Box p=1.00; ARCH-LM p=1.00) and achieves RMSE=80.82, MAE=67.16, MAPE=28.96%—acceptable error levels given the low-frequency data and volatile market conditions. The study identifies a structural break at the end of 2022, marking a shift from growth to adjustment due to the rise of TikTok Shop. Methodologically, the Denton Proportional model combined with Google Trends proves effective in addressing data frequency asymmetry (quarterly → monthly).

Based on the empirical trajectories forecasted, two actionable strategic initiatives are proposed for TiKi's executive decision-making:

Implementation of Rolling Forecast Architectures: TiKi should transition away from static annual budgeting toward a dynamic, monthly updated rolling forecast system. This continuous data infrastructure will enable agile, real-time recalibrations of supply chain logistics, inventory optimization, and digital marketing spending in response to prompt market shifts.

Value-Driven Competitive Positioning: Instead of engaging in unsustainable, capital-intensive price wars with dominant social commerce networks like TikTok Shop, TiKi must solidify its market footprint through alternative value-driven channels. Strategic resource allocation should focus on

enforcing product quality guarantees, elevating post-purchase technical support, optimizing fulfillment reliability, and expanding personalized retention algorithms to lock in the remaining premium customer segments.

Declaration of Conflicting Interests

A inherent limitation of this research stems from the strict non-disclosure policies surrounding corporate transactional databases, which mandated the deployment of secondary market estimations. To advance predictive resolution, future extensions should explore the integration of high-frequency macroeconomic covariates, real-time pricing indices of direct competitors, and non-linear machine learning architectures capable of endogenizing structural shocks.

References

1. Ministry of Industry and Trade, *E-commerce Industry Summary Report*, Hanoi: Industry and Trade Publishing House, 2026.
2. Metric, *Vietnam E-commerce Market Report*, Metric.vn, 2026.
3. Vietnam E-Commerce Association (VECOM), *Vietnam E-Commerce Index Report 2026 (EBI 2026)*, 2026.
4. A. M. Bloem, R. J. Dippelsman, and N. O. Maehle, *Quarterly National Accounts Manual: Concepts, Data Sources, and Compilation*, Washington, D.C.: International Monetary Fund, 2001.
5. Balpreet Singh et al., "Sales Forecast for Amazon Sales with Time Series Modeling," *First International Conference on Power, Control and Computing Technologies (ICPC2T 2020)*, 2020.
6. Yiding Zhao, "Research on E-Commerce Retail Demand Forecasting Based on SARIMA Model and K-means Clustering Algorithm," *Academic Journal of Science and Technology*, 2024.
7. J. Lecinski, *Winning the Zero Moment of Truth*, Google, 2011.
8. Dai Wei, Peng Geng, Liu Ying, and Shuaipeng Li, "A Prediction Study on E-commerce Sales Based on Structure Time Series Model and Web Search Data," *26th Chinese Control and Decision Conference (CCDC)*, 2014.
9. Anupriya Jain et al., "Demand Forecasting for E-Commerce Platforms," *IEEE International Conference on Innovation in Technology (INOCON 2020)*, 2020.
10. G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, Revised Edition, San Francisco, CA: Holden-Day, 1976.