Time Series Analysis for Sales and Demand Forecasting

Table of Content

1. Introduction

This project applied time series analysis to UK weekly book sales data to develop accurate demand forecasts for select titles. The objective was to identify and model sales patterns that could inform inventory management and restocking decisions. This effort supports Nielsen's initiative to enhance their service offering for small and medium-sized independent publishers, enabling data-driven investments in new titles and better stock control.

Using historical sales data, we explored various forecasting techniques—including statistical, machine learning, and hybrid models—focusing particularly on two well-known books: *The Alchemist* and *The Very Hungry Caterpillar*. These books were chosen for their distinct sales profiles and long market presence.

2. Data Overview and Preparation

Two datasets were provided: the **ISBN metadata** and the **UK weekly trended timeline**. Each contained multiple tabs based on book categories—Fiction, Educational, Trade, and Specialist. Data spanned from **2001-01-06 to 2024-07-20**, with weekly sales volume, value, and metadata like publisher, author, and pricing.

Summary Statistics

Category	Rows	Volume (avg)	ASP (£)	RRP (£)	Missing (Author)
Educational	55,286	530	5.56	6.92	5,173
Trade	65,344	376	9.21	12.78	4,689
Specialist	32,827	87	13.48	16.30	4,750

Fiction	73,767	381	7.16	9.57	267

Time series were resampled to weekly intervals and missing weeks filled with zeros to accurately reflect periods with no sales activity.

3. Exploratory Data Analysis

Across all categories, initial sales trends show a **strong launch period** followed by a **gradual decline**. Post-2012, most titles reached a stable or flat phase. Books with academic relevance exhibited **seasonal trends**, likely tied to academic calendars. Notably, some seasonal patterns became **less pronounced** over time, possibly due to waning academic use or declining overall demand.

4. Focus Books for Forecasting

We selected two books with long-standing popularity for detailed time series analysis:

- **The Alchemist** (Fiction, stable seasonal trends)
- **The Very Hungry Caterpillar** (Trade/Children, more irregular but high-volume sales)

Sales data from **2012 onwards** was retained for both, yielding **655 weekly observations** per title. These selections allowed for comparison of different modelling approaches on data with distinct structures.

5. Time Series Analysis

5.1. Decomposition and Seasonality Detection

STL decomposition with a 52-week period confirmed the presence of annual seasonality in both titles. However, residual components were not purely random, suggesting STL could not fully isolate trend/seasonality.

- ACF plots confirmed **yearly seasonality**.
- PACF suggested **lag-1 autocorrelation** for both books.

Stationarity checks using Augmented Dickey-Fuller tests showed that *The Alchemist* required **first-order differencing**, while *The Very Hungry Caterpillar* was already stationary.

6. Modelling Approaches and Results

6.1 Auto ARIMA

Auto ARIMA models were fitted to both books.

Book	MAE	MAPE
The Alchemist	155	30%
The Very Hungry Caterpillar	353	19%

While ARIMA effectively captured trends, it struggled with sudden peaks and dips. Residuals were not homoscedastic, indicating variable forecast confidence.

6.2 XGBoost and LSTM

Machine learning and deep learning models were applied next.

Book	LSTM MAE	LSTM MAPE	XGBoost MAE	XGBoost MAPE
The Alchemist	196	29%	404	83%
The Very Hungry Caterpillar	574	26%	387	20%

LSTM performed better for *The Alchemist*, likely due to its ability to model sequential dependencies. Conversely, **XGBoost** excelled for *The Very Hungry Caterpillar*, perhaps because of its strength in handling structured non-linearities and outliers.

6.3 Hybrid Models

Both sequential (ARIMA → LSTM) and parallel (combined ARIMA and LSTM) hybrid models were tested.

Book	Sequential MAE	Sequential MAPE	Parallel MAE	Parallel MAPE
The Alchemist	172	32%	118	24%
The Very Hungry Caterpillar	517	30%	346	17%

Parallel hybrids outperformed sequential models for both titles. This suggests combining predictions allows each model to compensate for the other's weaknesses. For example, ARIMA captures trend/seasonality well, while LSTM handles irregular bursts more effectively.

6.4 Monthly Aggregation

Aggregating weekly data to monthly enabled broader trend analysis.

Book	XGBoost MAE	XGBoost MAPE	ARIMA MAE	ARIMA MAPE
The Alchemist	699	34%	701	27%
The Very Hungry Caterpillar	2349	24%	1997	19%

In both cases, **Auto ARIMA outperformed XGBoost** in relative accuracy (MAPE), indicating that statistical models are better suited for **low-frequency** forecasting with stable seasonal structures.

7. Key Insights and Recommendations

1. **Model Selection Depends on Data Complexity**:
    a. For smoother, regular patterns (*The Alchemist*), LSTM and ARIMA both perform well.
    b. For more erratic or high-variance data (*The Very Hungry Caterpillar*), ensemble or boosting methods like XGBoost handle irregularities better.
2. **Hybrid Modelling is Optimal**:
    a. Parallel hybrid models showed the best performance across all metrics, demonstrating the value of combining complementary model strengths.
3. **Weekly vs Monthly Forecasting**:
    a. Weekly data supports high-resolution planning (e.g., promotions, short-term stocking).

b. Monthly forecasts, although less granular, provide a clearer long-term view, useful for procurement and reprint decisions.

4. **Seasonality is Crucial**:
    a. Books tied to academic calendars or seasonal events benefit from models that capture periodicity (e.g., SARIMA).
5. **Sales Decline After Initial Launch**:
    a. Most titles demonstrate a declining trend post-launch, reaffirming the importance of early demand prediction for stock optimization.

8. Conclusion

This project has demonstrated the practical application of time series analysis in the publishing sector. Leveraging historical sales data, we implemented a variety of models to forecast book sales and evaluated their performance. The insights obtained will help small to medium-sized publishers make better-informed decisions on reordering, inventory control, and publication planning.

Future work may include using external variables (e.g., holidays, academic schedules) and implementing multivariate models to further refine predictions. The hybrid modelling approach, especially parallel integration, presents a promising path for scalable forecasting in this domain.