AI6123 - Time Series Analysis
Group Project

Ravi Shwetha[†]
Matric No: G2304191F
shwetha002@e.ntu.edu.sg

Maheep[†]
Matric No: G2303665G
maheep001@e.ntu.edu.sg

Aradhya Dhruv[†]
Matric No: G2303518F
ar0001uv@e.ntu.edu.sg

UWINEZA Joseph[†]
Matric No: G2303477F
joseph005@e.ntu.edu.sg

Maheswaran RohinKumar[†]
Matric No: G2303513K
rohinkum001@e.ntu.edu.sg

NTAMBARA Etienne[†]
Matric No: G2304253K
ntam0001@e.ntu.edu.sg

Mukanyandwi Joselyne 7[†]
Matric No: G2304244J
joselyne001@e.ntu.edu.sg

Adnan Azmat [†]
Matric No: G2303265K
adnan002@e.ntu.edu.sg

Shinu Abdulvahab [†]
Matric No: G2203403C
shinu001@e.ntu.edu.sg

Sivanandan Vijayakumary Abhilash[†]
Matric No: G2203094G
abhilash005@e.ntu.edu.sg

Introduction

Our group project centers around analyzing Johnson Johnson (JJ) sales data and Drug sales data to develop accurate forecasts using ARIMA (AutoRegressive Integrated Moving Average) models. The JJ sales data spans quarterly earnings per share from 1960 to 1980, while the drug sales data covers monthly anti-diabetic drug sales in Australia from 1992 to 2008.

Drawing inspiration from a reference project on internet usage behavior, we aim to understand sales patterns by visualizing Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots. These insights will guide us in initializing our ARIMA models effectively, refining them iteratively based on ACFs of residuals and statistical metrics like Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

Our report will detail our approach, from data analysis to model refinement, providing valuable forecasts to support decision-making processes.

Dataset Description

- **JJ Sales Data:** The dataset covers the period from 1960 to 1980, illustrating the earnings per share of the company over time. It exhibits a discernible upward trend throughout the years, accompanied by a subtle seasonal influence.
- **Drug Sales Data:** This dataset portrays the sales of anti-diabetic drugs in Australia from 1992 to 2008. It encompasses total monthly prescriptions for pharmaceutical products categorized under ATC code A10, as reported by the Australian Health Insurance Commission. Each data point comprises a date and the corresponding total sales value.

[1]† equal contribution

Experiments on JJ Dataset

Exploratory Data Analysis

The graph below represents earnings per share of JJ over the years

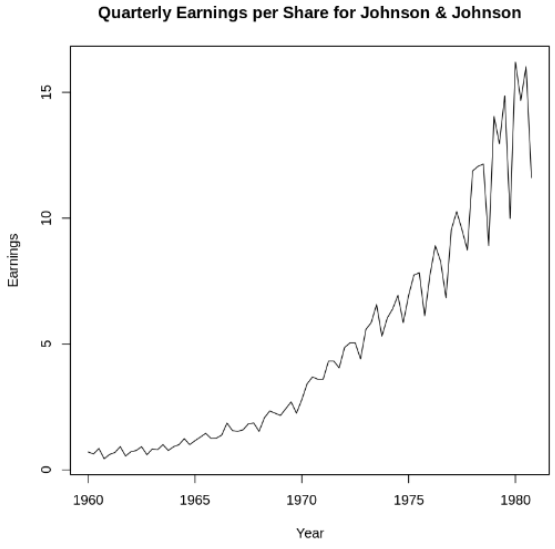


Figure 1: Earning per share of Johnson and Johnson

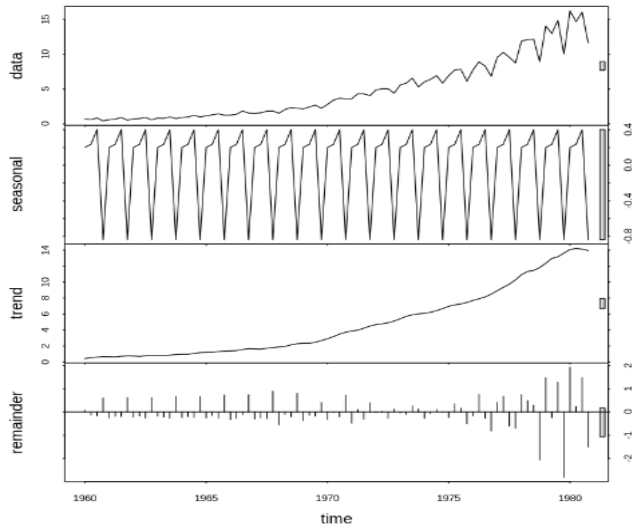


Figure 2: JJ Dataset when decomposed to seasonal, trend and remainder components

Upon decomposition of the Johnson Johnson earnings-per-share dataset, distinct characteristics emerge in each component. The seasonal aspect unveils periodic troughs and crests, suggestive of recurring patterns within specific time intervals. In contrast, the trend component showcases a consistent linear progression, steadily increasing in value across the entire time span. Notably, the residual components exhibit heightened prominence towards the dataset's conclusion, particularly evident between 1970 and 1975, signifying deviations from both the seasonal and trend patterns during this period.

The decomposition of additive time series is also shown in the graph below:

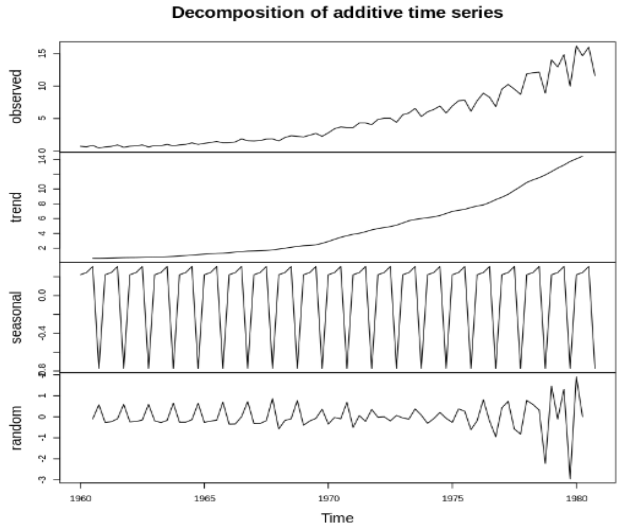


Figure 3: Decomposition of Additive Time Series

The graph shows that the trend is going up and that variance is not constant as it goes. There is also quarterly seasonality/

Initial Model Fitting

In time series analysis, ensuring stationarity in the data is paramount for accurate modeling and forecasting. Stationarity implies that the statistical properties of the data, such as mean, variance, and auto-correlation, remain constant over time. Nonstationary data, on the other hand, exhibits trends, seasonality, or other time-dependent patterns, making it challenging to model accurately. To confirm whether the data has seasonality to or not we carried out the augmented dicky-fuller test.

```
        Augmented Dickey-Fuller Test

data:  log_ts_data
Dickey-Fuller = -1.1543, Lag order = 4, p-value = 0.9087
alternative hypothesis: stationary
```

Figure 4: Augmented Dickey-Fuller Test Results

Since, the ADF shows that the data is non stationary, we had to perform differencing on the data

Following steps were followed in order to make the data non-stationary:

- Applied **log transformation** to stabilize the variance of the data, addressing its non-constant nature.
- Utilized **differencing** to remove the observed linear trend post-log transformation, ensuring the data's stationarity.
- Detected persistent **seasonal fluctuations** in the data post-differencing.
- Employed additional **differencing** to effectively eliminate the remaining seasonal components, achieving a more stable time series.

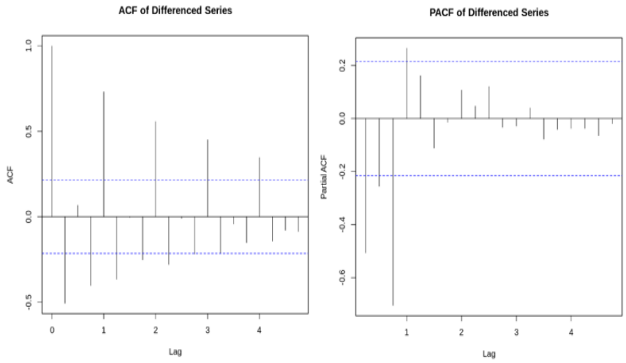


Figure 5: ACF PACF Plots on Transformed Data

Model Improvement Strategies

- Identified the optimal ARIMA parameters through careful analysis:
  - Utilized **D=1** for differencing to achieve stationarity.

- Determined **Q=4** based on significant autocorrelation at lag 4 in the autocorrelation function (ACF), indicating the presence of seasonality.
- Chose **P=4** based on significant partial autocorrelation at lag 4 in the partial autocorrelation function (PACF), capturing additional autocorrelation patterns.
- Considered the inclusion of **P=3** due to the presence of a large negative correlation, enhancing the model's predictive capacity.

– Selected the **ARIMA(4,1,3)** model as the most suitable representation of the data, effectively capturing its underlying patterns and ensuring accurate forecasting capabilities.

Model Selection

In our analysis, we implemented three different models to identify the optimal parameters for forecasting our time series data. Here's how we describe it in the report:

1. Model:

– We utilized the SARIMA (Seasonal AutoRegressive Integrated Moving Average) model with manually specified parameters.

– The chosen parameters were ARIMA(4,1,3) with seasonal components SARIMA(4,1,3). Figure[6]

– We assessed the model's performance through a summary of its key statistics and diagnostic plots generated by the checkresiduals function.

– Subsequently, we forecasted future values using the forecast function, specifying a horizon of 12 time points.

– Finally, we visualized the forecasted values alongside the historical data using the plot function.

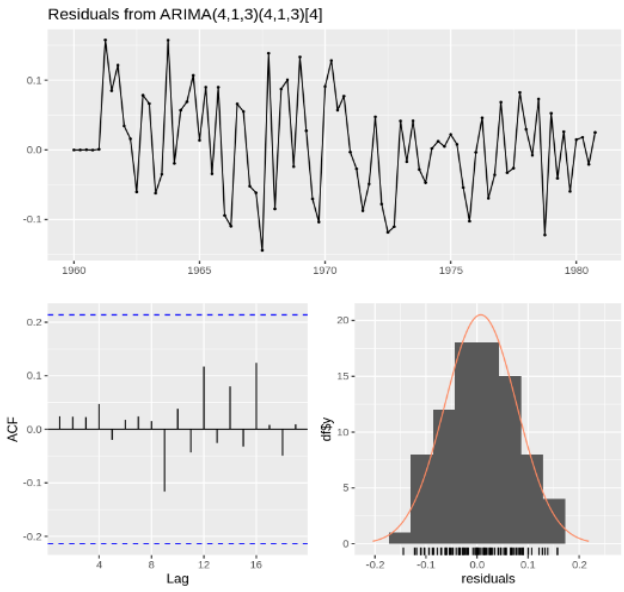


Figure 6: Residuals from ARIMA(4,1,3)(4,1,3)[4]

2. Model:

– We repeated the process with a different set of manually specified parameters.

– This time, the chosen parameters were ARIMA(4,1,4) with seasonal components SARIMA(4,1,4). Figure[7]

– Similar to Model 1, we evaluated the model's performance through summary statistics and diagnostic plots, followed by forecasting and visualization.

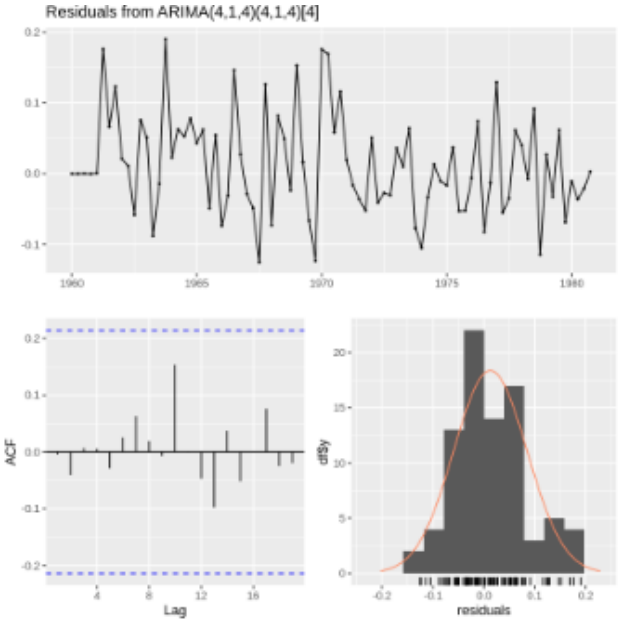


Figure 7: Residuals from ARIMA (4,1,4)(4,1,4)[4]

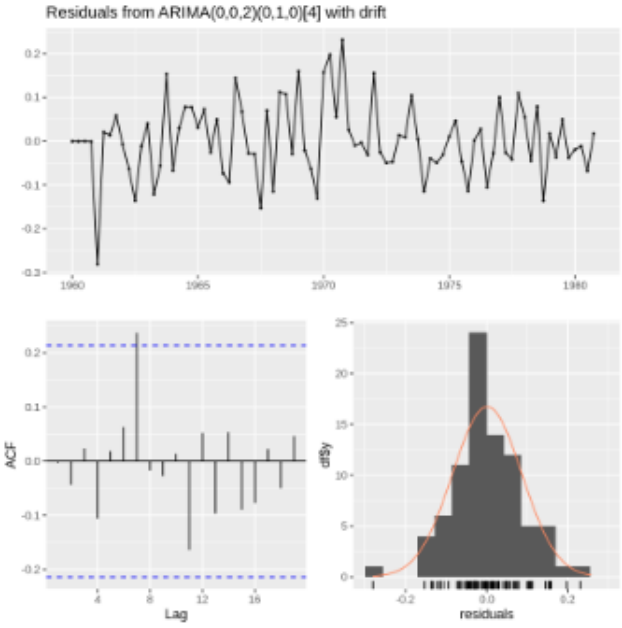


Figure 8: Residuals from ARIMA (0,0,2)(0,1,0)[4] with drift

3. Model:

- We employed an automated approach using the auto.arima function to select the optimal parameters.
- This function iteratively explores various combinations of parameters to find the best-fitting SARIMA model.
- We again assessed the selected model's performance through summary statistics, diagnostic plots, forecasting, and visualization. Figure[8]

In addition to these steps, we calculated the Root Mean Squared Error (RMSE) scores for each model to quantify their forecasting accuracy. These scores provide valuable insights into the reliability of the forecasts generated by each model.

Forecasting

The table below summarizes the RMSE values for different parameter settings, and the forecast showcases the best observed prediction with parameter values: (4,1,4)(4,1,4).

Table 1: Performance Metrics for Selected ARIMA Models in Forecasting Monthly Anti-Diabetic Drug Sales

ARIMA Value	ME	RMSE	MAE
(0,0,2)(0,1,0)	0.001493826	0.08493652	0.06402484
(4,1,4)(4,1,4)	**0.01296632**	**0.07265574**	**0.05674175**

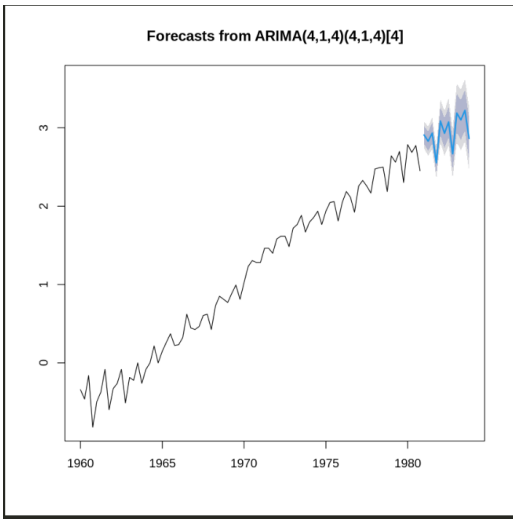


Figure 9: Observed Forecast with best parameters

Experiments on Drug Sales Dataset

Exploratory Data Analysis

In this exploratory analysis of the monthly anti-diabetic drugs sales dataset, many necessary measures were taken to better understand the underlying patterns and prepare the data for time series modeling.

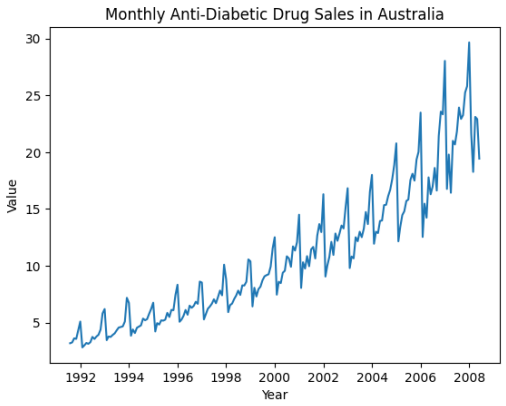


Figure 10: Monthly Anti-Diabetic Drug Sales in Austrailia

The first step in visualizing drugs sales was to plot a time series across several months. This demonstrated a substantial increase in anti-diabetic medicine sales over the examined time period. Stationarity tests were used to investigate the data's characteristics in further detail. Both the Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests were used. The results showed that the original series was not stationary, indicating the necessity for transformation.

Initial Model Fitting

To select the initial model for forecasting monthly anti-diabetic drug sales, a thorough exploration of the dataset was conducted. The process began with an examination of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots, which revealed potential patterns and seasonality in the data. Subsequently, the series was differenced to stabilize variance and remove trends, followed by a logarithmic transformation to mitigate fluctuations. These steps aimed to achieve stationarity, a prerequisite for ARIMA modeling.

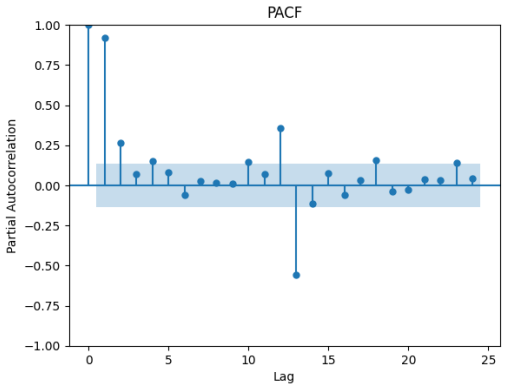


Figure 11: Partial Autocorrelation Function (PACF)

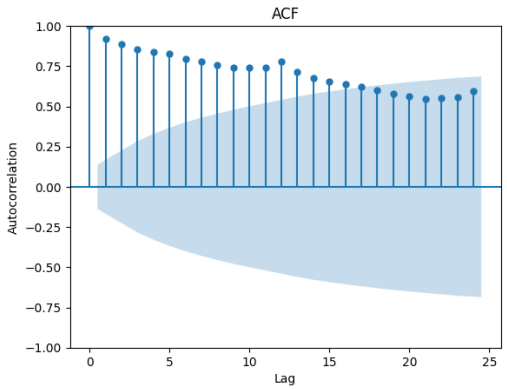


Figure 12: Autocorrelation Function (ACF)

ACF and PACF aid in determining the MA and AR terms in the ARIMA model, respectively. Significant spikes or cutoffs in these plots suggest potential values for the p (AR) and q (MA) parameters. To achieve stationarity, the series underwent differencing, stabilizing variance by subtracting each observation from its previous value. A logarithmic transformation further reduced fluctuations and improved normal distribution.

Model Improvement Strategies

In our efforts to enhance the initial ARIMA model, we engaged in a systematic approach to iteratively refine its parameters. Initially built with an order of (2, 1, 1), we delved into the examination of the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the model's residuals. By scrutinizing these plots, we aimed to pinpoint optimal parameters for the ARIMA model. This iterative process led us to explore various ARIMA orders, progressively moving from (2, 1, 1) to (5, 1, 1), (10, 1, 1), and finally to (13, 1, 1). At each step, we fitted the model, assessed the ACF and PACF of the residuals, and conducted statistical tests such as the Ljung-Box, Augmented Dickey-Fuller (ADF), and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests to ensure model adequacy and stationarity. Additionally, we applied a logarithmic transformation followed by differencing to stabilize the variance of the time series.

Final Model Selection

– Several ARIMA models were fitted to the altered dataset to start the process. The results of autocorrelation functions (ACF and PACF) and stationary tests (ADF and KPSS) were used to iteratively modify key parameters. The goal of this iterative refinement was to get the model's parameters as close to the underlying patterns in the dataset as possible.

  ARIMA(2, 1, 1) and ARIMA(13, 1, 1) was taken as two primary models to be candidates. The ARIMA(13, 1, 1) model performed better and had a lower AIC value after a thorough comparison using the AIC, indicating that it was more appropriate for forecasting the series with lower error and higher predictive accuracy.

– ARIMA(13, 1, 1) was the final chosen model. The model's robustness and forecasting reliability were further validated by looking at the residuals for autocorrelation and non-stationarity. The results showed that there were no serious problems.

  With a strong foundation for accurate and dependable forecasting, this thorough methodology guarantees that the ARIMA(13, 1, 1) model is statistically valid and accurately represents the dynamics of the monthly sales of anti-diabetic drugs.

Forecasting

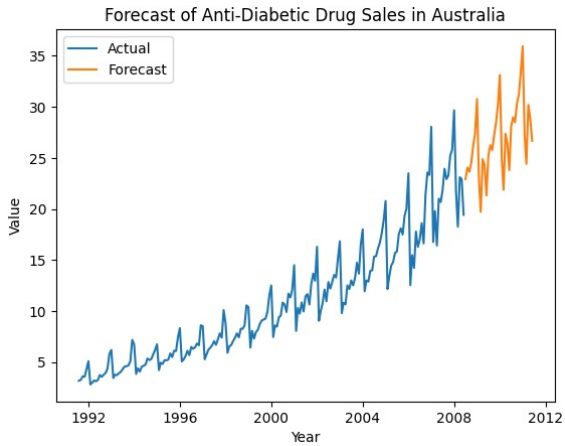
Table 2: Summary of Model Statistics (Part 1)

Model	AIC	Ljung-Box Test Statistic	Ljung-Box p-value
ARIMA(2, 1, 1)	-127.52	322.64	1.67e-53
ARIMA(5, 1, 1)	-136.68	320.43	4.68e-53

Table 3: Summary of Model Statistics (Part 2)

Model	ADF Statistic	ADF p-value	KPSS Statistic	KPSS p-value
ARIMA(2, 1, 1)	-3.82	0.0027	0.03998	0.1
ARIMA(5, 1, 1)	-4.50	0.0002	0.01059	0.1
ARIMA(10, 1, 1)	-4.50	0.0002	0.01059	0.1
ARIMA(13, 1, 1)	-5.29	0.0000	0.04058	0.1

Figure 13: Forecast of Anti-Diabetic Drug Sales in Australia



REFERENCES

1 arXiv, ”What Can Large Language Models Tell Us about Time Series Analysis”, 2024-02-05. [Online]. Available: https://arxiv.org/html/2402.02713v1

2 Vanessa Freitas Silva, ”Time Series Analysis via Network Science: Concepts and Algorithms”, 2021-10-11. [Online]. Available: https://arxiv.org/abs/2110.09887

3 arXiv, ”A Survey on Deep Learning based Time Series Analysis with Frequency Transformation”, 2023-02-04. [Online]. Available: https://arxiv.org/abs/2302.02173

4 arXiv, ”MTSA-SNN: A Multi-modal Time Series Analysis Model Based on Spiking Neural Network”, 2024. [Online]. Available: https://arxiv.org/abs/2402.02713v1

5 Fredric S. Roberts and Paul J. Ding, ”Time Series Analysis”, 2000. [Online]. Available: https://citeseerx.ist.psu.edu/document?doi=fc5146ccbb6b3eb469b71b7ede0dd295058ad2a9&repid=rep1&type=pdf

6 arXiv, ”Time Series Analysis and Modeling to Forecast: a Survey”, 2021-03-31. [Online]. Available: https://arxiv.org/abs/2104.00164

7 Ali Ebadi, ”Comparing Time-Series Analysis Approaches Utilized in Research Papers to Forecast COVID-19 Cases in Africa: A Literature Review”, 2023-10-05. [Online]. Available: https://arxiv.org/abs/2310.03606

8 Desikan, P., Srivastava, J., *Time Series Analysis and Forecasting Methods for Temporal Mining of Interlinked Documents*, University of Minnesota, 2000. [Online]. Available: https://citeseerx.ist.psu.edu/document?doi=fc5146ccbb6b3eb469b71b7ede0dd295058ad2a9&repid=rep1&type=pdf

9 ScienceDirect, *Time series analysis with explanatory variables: A systematic literature review*, Environmental Modelling and Software, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S136481521730542X

10 ScienceDirect, *Time series analysis with explanatory variables: A systematic literature review*, Environmental Modelling and Software, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136481521730542X

11 ScienceGate, *Time series Latest Research Papers*, 2022. [Online]. Available: https://www.sciencegate.app/keyword/965