ADM4307 Forecasting Project – Overseas Vacationers' Tourism Expenditure

Cia Cui

University of Ottawa

ADM4307

Professor:Firouz Fallahi

12/13/2023

Outline

Executive Summary

This article designs an in depth evaluation of overseas vacationers' tourism expenditure in Canada, specializing in the principle supply of international locations, tourism regions and expenditure categories. Notably, the USA emerged as the primary source of tourists, accompanied by means of countries which include China, Japan and the UK. Spending classes consist of lodging, dining, transportation inside Canada, entertainment and enjoyment, and garb and presents. In addition to these info, the records illuminates nuances among areas, permitting a nuanced knowledge of the economic effect of tourism on one-of-a-kind components of the us of a. This data has vital implications for policymakers to suggest tailored advertising strategies, focused infrastructure investments and coverage issues to optimize monetary possibilities and sell sustainable growth of Canada’s tourism industry, and via 3 forecast models: STL Model, ETS (Error, Trend, Seasonal) Model and ARIMA（Autoregressive Integrated Moving Average） Model ensure predictions to attract powerful conclusions.

Introduction of dataset

Overview:

The data set analyzed is derived from Canada's Statistics Table 24 10 0047 01, offering an all-encompassing illustration of spending by non-resident visitors to the country. The dataset was thoughtfully designed to capture intricate details such as country of origin, travel destination, and various expenditure categories. This includes significant source countries such as the United States, Australia, China, Japan, South Korea, India, the United Kingdom, France, Germany, Mexico, and other international destinations. By delving into this dataset, one can gain a comprehensive understanding of the diverse landscape of international tourism. It consists of 16 columns, each representing a crucial aspect of tourism spending, including pertinent information like the date of reference, geographical location, a unique identifier, and a numerical value reflecting the amount spent.

Pre - tasking

In order to create an accurate monthly forecast, we first start with the quarterly estimates as our foundation. Through a statistical model that captures the intricate monthly dynamics of our service components, we project these values onto a monthly timeline during the preprocessing stage. This crucial step ensures that our forecast reflects the most up-to-date and detailed information possible.

This article selects three different prediction models

1. STL Model

2. EST Model

3. (ARIMA) Model

Methodology

STL DECOMPOSITION METHOD

STL Model is a robust method of time series decomposition often used in economic and environmental analyses. The STL method uses locally fitted regression models to decompose a time series into trend, seasonal, and remainder components.

The reason for choosing to use the STL decomposition method is that STL's versatility and robustness make it stand out, especially compared to other methods such as X-12-ARIMA. The fact that it can handle any type of seasonality is a huge win, and the ability to have seasonal components change over time at a user-controlled rate of change adds extra flexibility.

Its control over the smoothness of trend cycles is a key advantage, allowing users to tailor the decomposition to their specific needs. Its robustness to outliers is another advantage, ensuring that outlier data points do not affect the overall prediction process.

In some cases, the lack of trading days or calendar adjustments can be a limitation, but depending on the forecasting requirements, it can also be viewed as simplicity that avoids unnecessary complexity.

The option of additive and multiplicative decompositions, as well as the flexibility of using Box-Cox transformations for different decompositions, add to its appeal. This adaptability makes it suitable for a wide range of forecasting scenarios, making it a popular choice among analysts and forecasters.

ETS MODEL

ETS is a flexible modeling framework that encompasses a wide range of exponential smoothing methods. It decomposes a time series into three components: level (l), trend (b), and seasonality (s). The ETS framework allows for different combinations of these components, resulting in various models, including additive (A), multiplicative (M), and damped trend (D). It has different types:

- Additive Model (ETS-A)

In the additive model, the components are combined linearly. It assumes that the magnitude of the seasonality remains constant regardless of the level or trend. The additive model is suitable when the seasonal fluctuations are consistent across different levels of the time series.

- Multiplicative Model (ETS-M)

The multiplicative model assumes that the components are combined multiplicatively, meaning the seasonality is proportional to the level or trend. This model is appropriate when the seasonal patterns change in proportion to the level or trend of the time series.

- Damped Trend Model (ETS-D)

The damped trend model extends the additive or multiplicative model by introducing a damping parameter (phi) to dampen the trend over time. This model is useful when the trend is expected to decrease or converge towards a certain value.

The reason why I choose ETS techniques is due to multiple reasons. One key advantage of ETS is its flexibility. ETS provides a modeling framework that allows for different combinations of components (level, trend, and seasonality), enabling analysts to choose the most suitable model for their data. This flexibility makes ETS applicable to a wide range of time series patterns and data characteristics. In addition to its flexibility, ETS is known for its simplicity. The models are relatively easy to understand and implement, making them accessible to users with varying levels of statistical expertise. This simplicity allows for quick model building and interpretation, saving time in the analysis process. Another benefit of ETS is its adaptability to changing data. ETS models can automatically adjust to evolving patterns and update their forecasts and parameter estimates as new observations become available. This adaptability makes ETS well-suited for analyzing time series data with dynamic and evolving patterns. Scalability is another advantage of ETS. The models are scalable and can efficiently handle large datasets. Since ETS operates on individual observations rather than requiring the entire dataset to be loaded into memory, it is computationally efficient for analyzing large-scale time series data. ETS models have demonstrated good forecasting accuracy, particularly when the underlying patterns in the data align with the assumptions of the chosen ETS model. By considering the level, trend, and seasonality components, ETS models can capture complex patterns and generate reliable forecasts. Furthermore, ETS models provide interpretable results. The decomposition of the time series into understandable components allows for a clear understanding of the contributions of each component to the overall behavior of the time series. This interpretability makes it easier to explain the results to stakeholders. ETS models are robust and can handle missing data points and outliers effectively. The smoothing techniques used in ETS consider all available data, including missing observations, and are less sensitive to outliers compared to some other time series analysis methods. Finally, ETS models find applications in various domains, including demand forecasting, sales analysis, financial forecasting, resource planning, and anomaly detection. The versatility of ETS models allows them to be applied to different types of time series data across industries. It is important to note that while ETS offers many benefits, the choice of a time series analysis technique depends on the specific characteristics of the data and the objectives of the analysis. It is always recommended to consider alternative techniques and assess their suitability for the given context.

We fit an ETS model into our data. And we measure the average performance of the ETS model in the following table. The prediction result shows a mean squared error (MSE) of 1,908,210,402, an adjusted mean squared error (AMSE) of 4,881,730,480, and a mean absolute error (MAE) of 2,455. The MSE and AMSE metrics indicate the average squared difference between the predicted values and the actual values. A lower MSE and AMSE value suggests better accuracy and a closer fit of the predictions to the true values. In this case, the high values of MSE and AMSE indicate a significant amount of variability and a larger discrepancy between the predicted and actual values. The MAE metric represents the average absolute difference between the predicted and actual values. It provides a measure of the average magnitude of the errors in the predictions. The MAE value of 2,455 indicates that, on average, the predictions deviate by approximately 2,455 units from the actual values. Overall, based on these metrics, the prediction result appears to have a relatively high level of error and discrepancy between the predicted and actual values. It may be necessary to further investigate the model and explore potential improvements to enhance the accuracy and reliability of the predictions.

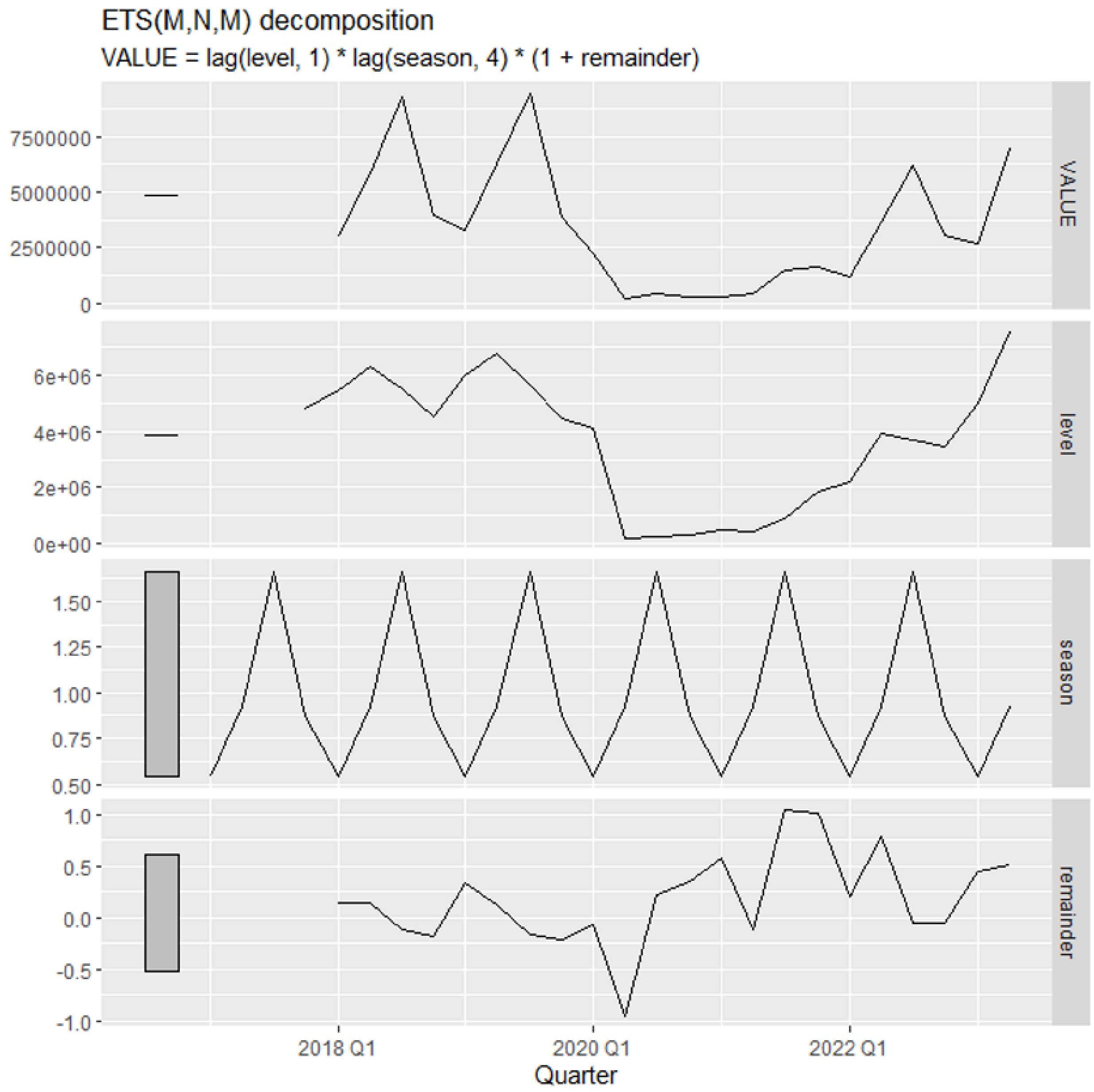
MSE	AMSE	MAE_mean
1908210402.	4881730480.	2455

We also investigate performances of some examples. For VECTOR="v1139468013", the detailed characteristics are listed below.

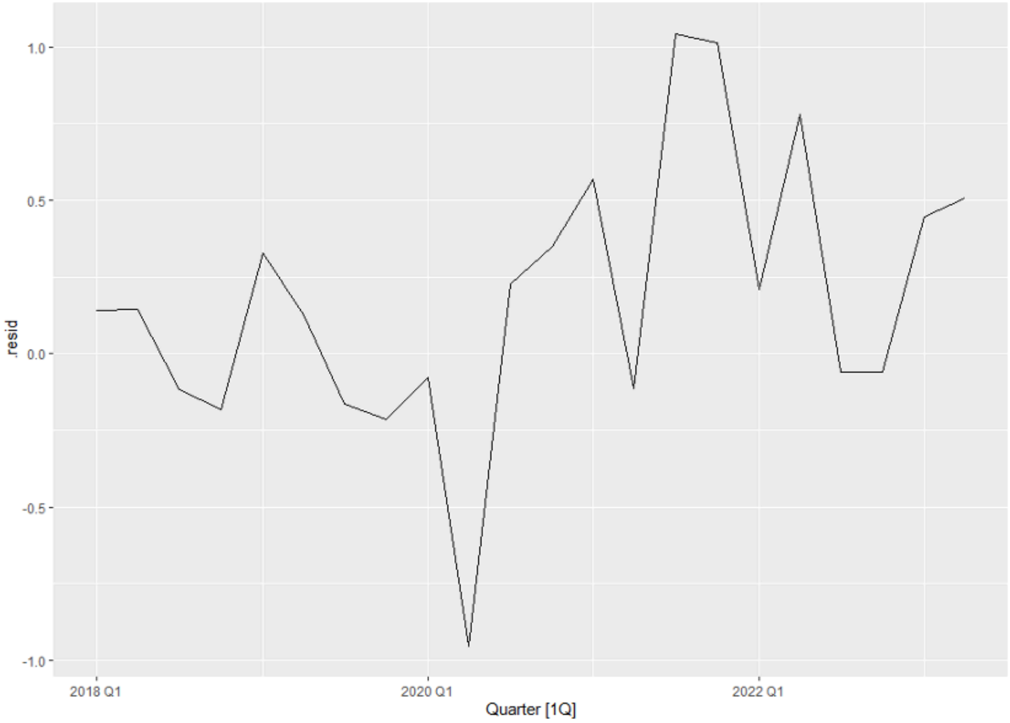
GEO	Area of residence	Type of expenditures	Status
Canada	Australia	Total expenditures	A

The smoothing parameters in this model for the series are alpha and gamma. The alpha value of 0.9990737 suggests a high level of smoothing, indicating that recent observations have a strong influence on the forecasts. The gamma value of 0.000100731 indicates a relatively low level of seasonality smoothing, implying that seasonality patterns have a minimal impact on the forecasts. The initial states represent the starting values for the level component (l) and the seasonal components (s[-3] to s[0]). These initial states are essential for initializing the model and capturing the initial values of the components. In this case, the initial level (l[0]) is 4805904, and the initial seasonal components (s[-3] to s[0]) have specific values. The sigma^2 value of 0.3087 represents the estimated variance or the squared standard deviation of the forecast errors. It provides an indication of the overall variability or volatility of the series. The AIC (Akaike Information Criterion), AICc (corrected AIC), and BIC (Bayesian Information Criterion) are model selection criteria that consider the trade-off between goodness-of-fit and model complexity. Lower AIC, AICc, and BIC values indicate better model fit. In this case, the AIC is 687.9333, AICc is 695.9333, and BIC is 695.5706.

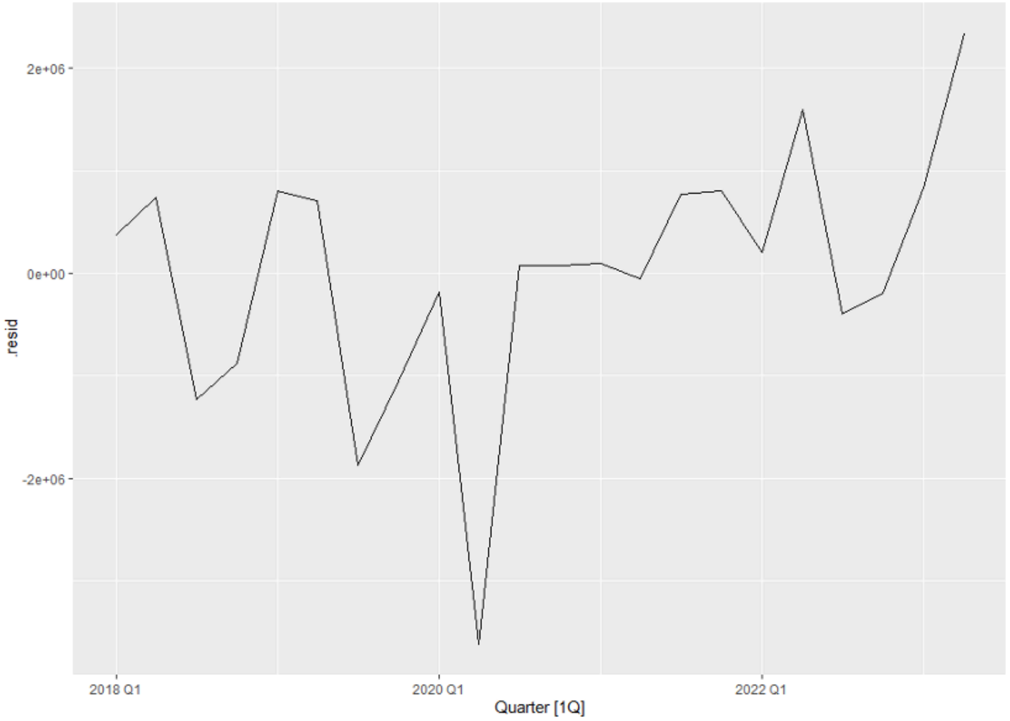
The ETS components are drawn in the following figure.



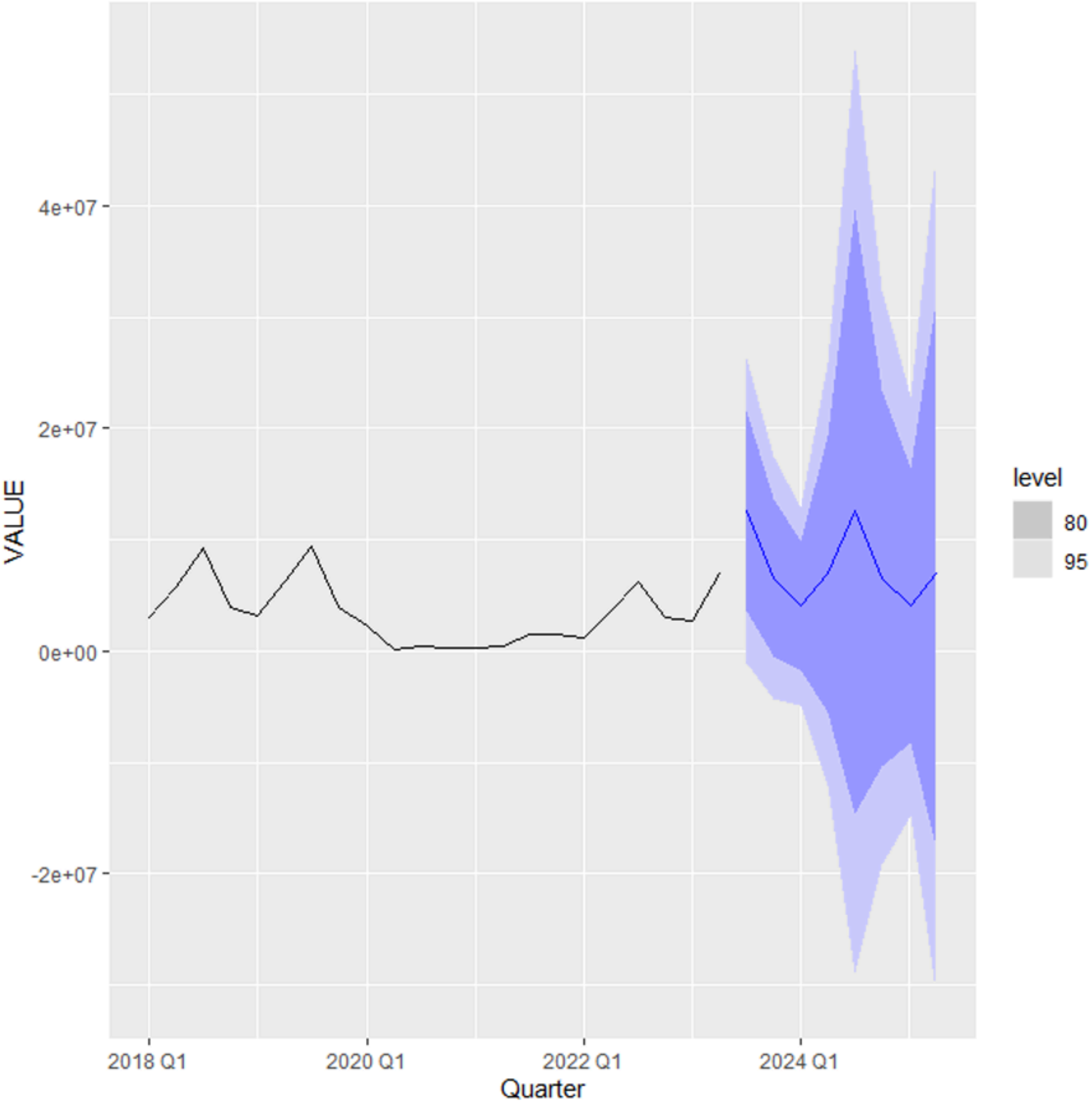
The innovation residuals are plotted below.



The response residuals are plotted below.



We can use the model to predict the upcoming figures.



the specific statistic reported is the Ljung-Box statistic, which is commonly used to assess the presence of autocorrelation in the residuals of a model. It helps determine whether the residuals exhibit any significant patterns or information that the model has not captured. In this case, the Ljung-Box statistic is reported as 7.45, with a corresponding p-value of 0.489. The p-value represents the probability of observing a test statistic as extreme as, or more extreme than, the one calculated, assuming the null hypothesis of no autocorrelation in the residuals. With a p-value of 0.489, there is no strong evidence to reject the null hypothesis of no autocorrelation in the residuals. This suggests that the residuals may not exhibit significant autocorrelation patterns beyond what would be expected by chance.

ARIMA

The Autoregressive Integrated Moving Average (ARIMA) model is a widely used time series forecasting technique. This report aims to provide a comprehensive explanation of the ARIMA model, including its components, working principles, and applications. By understanding the ARIMA model, analysts can leverage its capabilities to make accurate predictions and gain valuable insights from time series data. The ARIMA model is a powerful tool for analyzing and forecasting time series data. It combines autoregressive (AR), differencing (I), and moving average (MA) components to capture the underlying patterns and dynamics of the data. The ARIMA model is particularly useful when the data exhibits trends, seasonality, or other non-stationary characteristics. There are multiple components of ARIMA:

l Autoregressive (AR) Component: The AR component models the dependence of the current value on past values of the series. It assumes that the current value is a linear combination of its lagged values, weighted by autoregressive coefficients.

l Integrated (I) Component: The I component involves differencing the data to make it stationary. Differencing removes trends and seasonality, transforming the series into a stationary process. The number of differencing steps required is indicated by the "I" parameter.

l Moving Average (MA) Component: The MA component captures the dependency between the current value and past forecast errors. It models the series as a linear combination of past errors, weighted by moving average coefficients.

Selecting the appropriate order of the ARIMA model is crucial. It involves determining the values of the parameters (p, d, q) that define the AR, I, and MA components, respectively. Various techniques, such as autocorrelation and partial autocorrelation plots, information criteria (AIC, BIC), and grid search, can aid in identifying the optimal order. Once the order is determined, the ARIMA model is estimated using maximum likelihood estimation or other suitable methods. Diagnostic checks, including residual analysis, normality tests, and autocorrelation of residuals, are performed to evaluate the model's goodness of fit and identify any remaining patterns or deficiencies. ARIMA models can generate forecasts by extrapolating historical patterns into the future. The model takes into account the estimated coefficients and the initial observations. Forecast accuracy can be assessed using measures such as mean absolute error (MAE), mean squared error (MSE), or root mean squared error (RMSE). ARIMA models find extensive applications in diverse fields, including finance, economics, sales forecasting, demand planning, and resource allocation. They are particularly useful for short to medium-term forecasting, where trends, seasonality, and other patterns need to be captured.

ARIMA (Autoregressive Integrated Moving Average) offers several advantages over other time series analysis techniques. It provides flexibility by accommodating various types of time series data, including those with trends, seasonality, and complex patterns. The model's ability to handle non-stationary data through differencing allows for robust modeling. ARIMA effectively captures autocorrelation, enabling better understanding and prediction of future observations based on past values. The model has demonstrated strong forecasting accuracy, particularly for short to medium-term predictions. ARIMA models yield interpretable results, allowing analysts to understand the contributions of different components and gain insights into the underlying dynamics of the data. The methodology behind ARIMA is well-established, and there is extensive guidance available for model selection, parameter estimation, and diagnostic checks. The availability of software packages and libraries further facilitates the implementation of ARIMA models. ARIMA's versatility allows for extensions and combinations with other techniques to handle specific scenarios, such as seasonal patterns or incorporating exogenous variables. While considering the benefits of ARIMA, it is essential to remember that the choice of a time series analysis technique should align with the data characteristics and analysis objectives, necessitating the evaluation of alternative techniques.

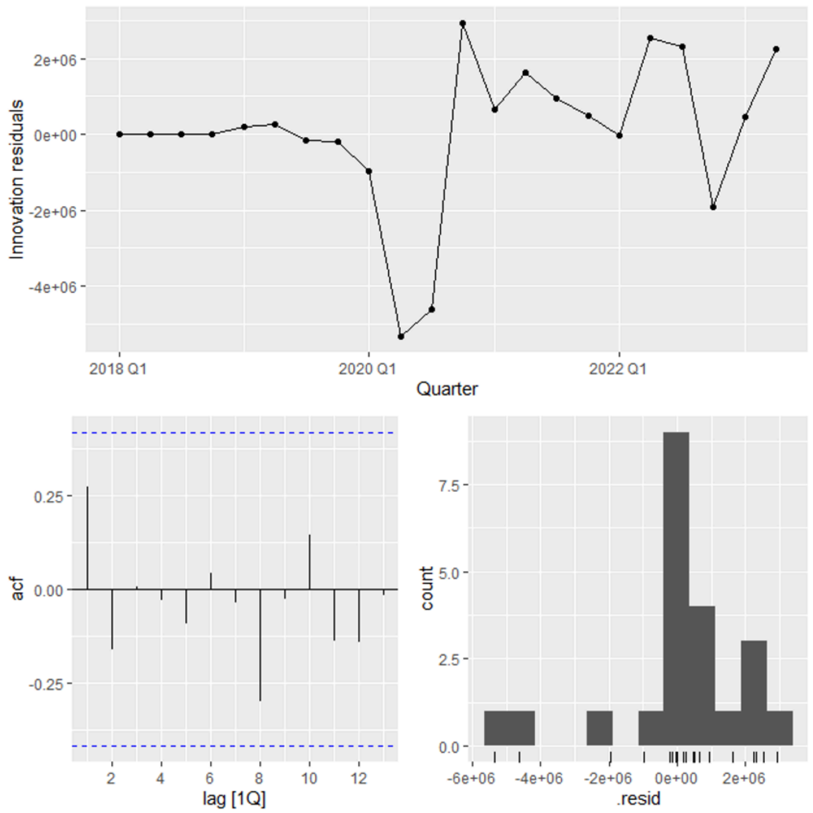
We fit ARIMA models with our data. The Bayesian Information Criterion (BIC) is a criterion that penalizes model complexity more heavily than the Akaike Information Criterion (AIC). A lower BIC value indicates a better-fitting model with lower complexity. In this case, the BIC_mean value of 144 suggests a relatively good fit. The Akaike Information Criterion (AIC) is another model selection criterion that balances the goodness of fit with the number of parameters in the model. Lower AIC values indicate better-fitting models. The AIC_mean value of 144 suggests a good fit as well. The corrected Akaike Information Criterion (AICc) is a modification of AIC that corrects for small sample sizes. It provides a more reliable estimation when the number of observations is relatively small. Like AIC, lower AICc values indicate better-fitting models. The AICc_mean value of 144 suggests a good fit, considering the correction for sample size.

We also investigate performances of some examples. For VECTOR="v1139468013", the detailed characteristics are listed below.

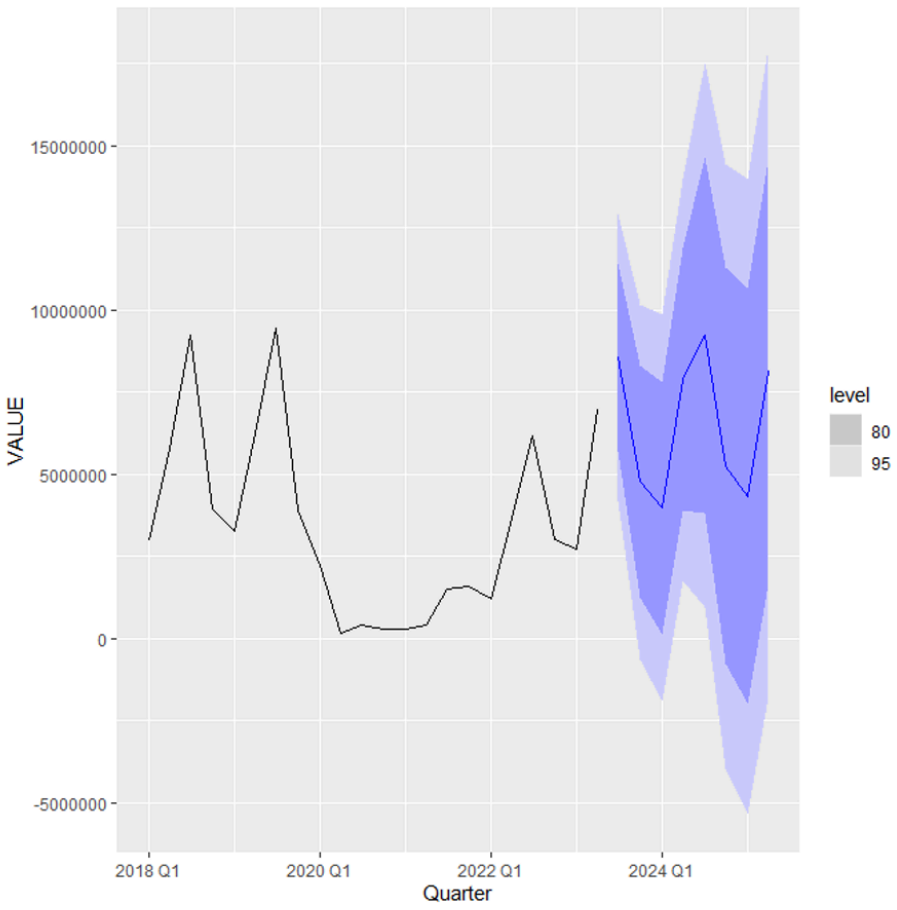
GEO	Area of residence	Type of expenditures	Status
Canada	Australia	Total expenditures	A

The model's order is specified as (1,0,0)(0,1,0)[4], which indicates an autoregressive component of order 1 (ARIMA(1,0,0)), and a seasonal differencing component of order 1 with a seasonal period of 4 (ARIMA(0,1,0)[4]). The coefficient for the autoregressive term (AR) is estimated to be 0.7245, with a standard error of 0.1507. This coefficient represents the weight assigned to the lagged value of the series in predicting the current value. The estimated variance (sigma^2) is reported as 4.999e+12, resulting in a log likelihood of -288.56. The Akaike Information Criterion (AIC) is calculated as 581.12, while the corrected AIC (AICc) is 581.92. The Bayesian Information Criterion (BIC) is computed as 582.9. These information criteria, AIC, AICc, and BIC, are commonly used for model selection. Lower values indicate better-fitting models, considering both goodness of fit and model complexity. In this case, the reported values suggest that the model may provide a reasonable fit to the data.

The residuals plot is shown below.



We use the model to predict the values.



The Ljung-Box statistic is a test used to determine if the residuals of a model exhibit significant autocorrelation. It helps assess whether there are any remaining patterns or information in the residuals that the model has not captured. In this case, the Ljung-Box statistic for the residuals is reported as 6.35, with a corresponding p-value of 0.608. The p-value represents the probability of observing a test statistic as extreme as, or more extreme than, the one calculated under the null hypothesis of no autocorrelation in the residuals. A p-value of 0.608 suggests that there is no strong evidence to reject the null hypothesis of no autocorrelation in the residuals. This indicates that the residuals may not exhibit significant autocorrelation patterns beyond what is expected by chance.

Conclusion

In this project, we conduct a time-series-forecast task upon travel spending data. We complete the task with three different methods: STL, ETS and ARIMA methods. We find that STL is the one with the fastest training time and is easy to fit. However, it doesn’t perform well. We find that ARIMA performs the best but with the longest running time. In practice, we should choose the best model based on our needs.