Data Overview

This section examines the core characteristics and statistical properties of the time series. Understanding these attributes is important for assessing data quality and gaining a preliminary context. We explore the temporal structure, summary statistics, and distribution patterns to create a baseline understanding of your data.

Time Series Plot

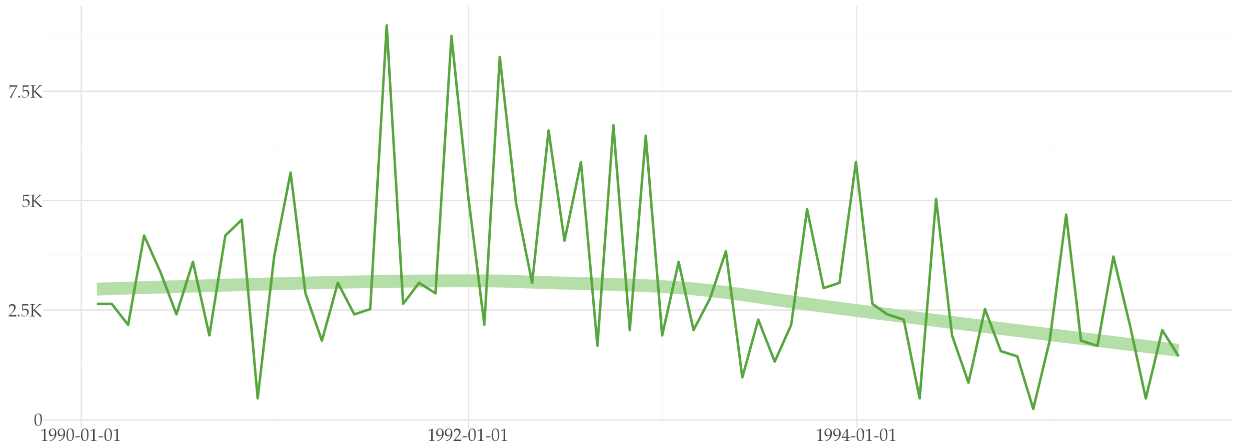
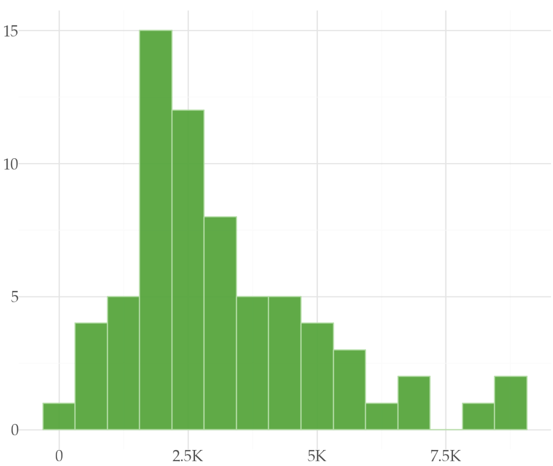


Figure 1: Time series line plot.

- A total of 68 monthly observations which span from January 1990 to August 1995.
- The mean value of the series is 3185.29 (median equal to 2640), with a standard deviation of 1932.54. The data ranges from a minimum of 240 to a maximum of 9000.

Data Distribution



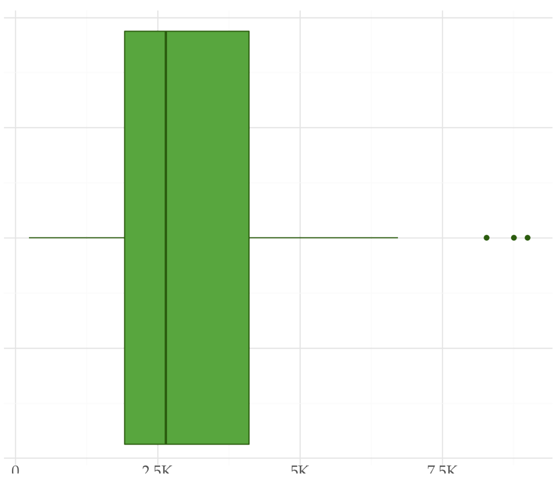


Figure 2: Distribution of the time series using an histogram (left) and a boxplot (right).

- The Kolmogorov-Smirnov test rejects the hypothesis that the series is distributed according to the following distributions: Exponential, Pareto, Power-law, and Chi-squared
- The distribution with largest p-value is Exponentially Modified Gaussian distribution (p-value equal to 0.95). But, we cannot reject the hypothesis that the data follows the following distributions (ordered decreasingly by p-value): Log-Normal, Gamma, Logistic, Cauchy, and Gaussian.
- There are 3 outliers in the data, all of which are upper outliers. The outliers represent 4.41% of the complete data set.
- The excess kurtosis is equal to 1.08. This value is similar to that found from data following a Gaussian distribution
- The skewness is equal to 1.13, indicates that the right tail is long relative to the left tail.

Trend and Seasonality

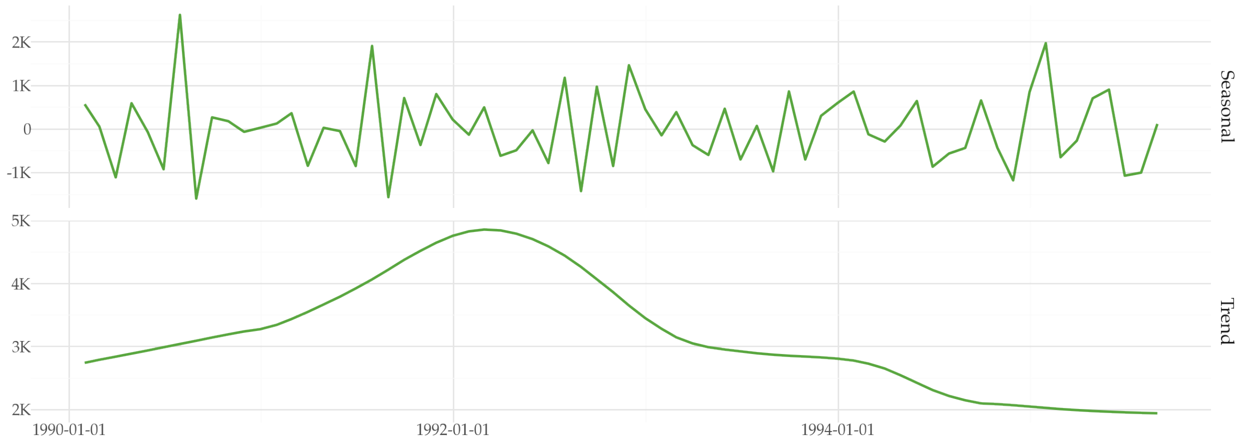


Figure 3: Seasonal and Trend components after decomposition using the STL (Season-Trend decomposition using LOESS) method.

- The following tests indicate that the time series is non-stationary in trend/level: KPSS. On the other hand, other tests (Augmented Dickey-Fuller and Philips-Perron) fail to reject the hypothesis that the data is stationary
- The following tests indicate that the time series is non-stationary in seasonality for the specified period: OCSB. On the other hand, other tests (Wang-Smith-Hyndman) fail to reject the hypothesis that the data is stationary

Auto-Correlation

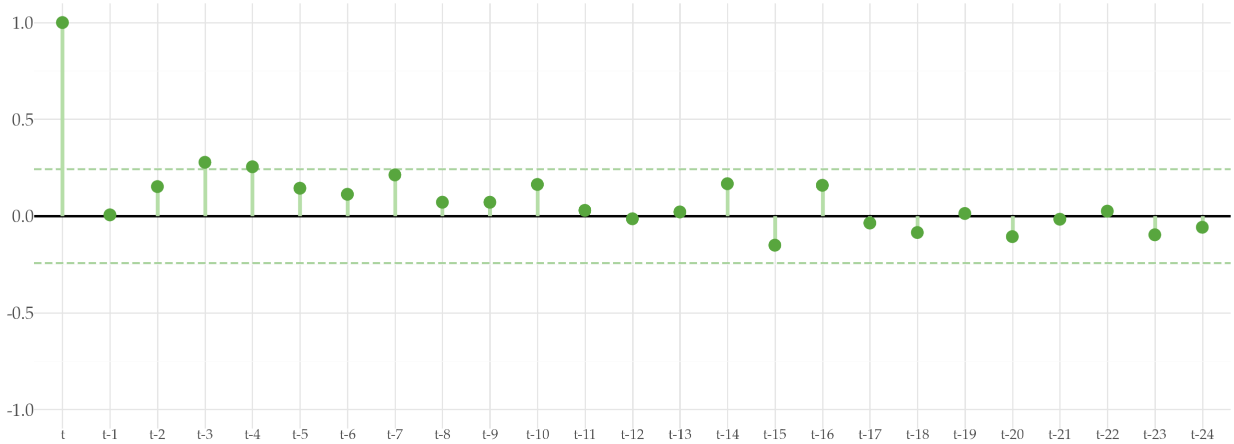


Figure 4: Auto-correlation plot up to 24 lags.

- The following lags show significant autocorrelation: t-3 and t-4. The autocorrelation is positive for all lags with a significant value.
- None of the lags relative to the seasonal period (t-12 and t-24) show any significant autocorrelation.

Partial Auto-Correlation

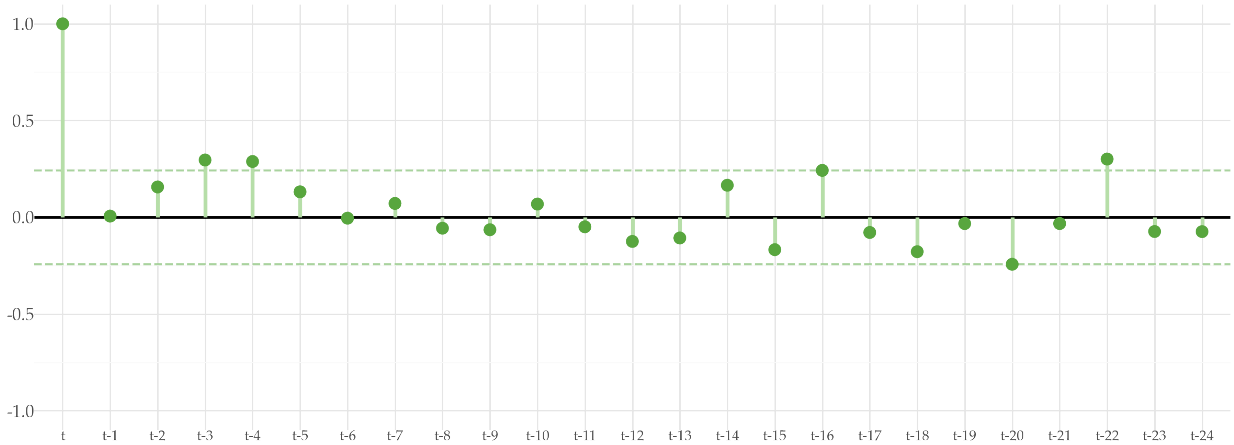


Figure 5: Partial Auto-correlation plot up to 24 lags. At each lag, the partial auto-correlation takes into account the previous correlations.

- The following lags show significant partial autocorrelation: t-3, t-4, t-20, and t-22.
- None of the lags relative to the seasonal period (t-12 and t-24) show any significant partial autocorrelation.

Report Organization

The following sections dive deeper into some relevant aspects of the time series. The remaining sections address the following topics. Some analysis are based on forecasting experiments. These were carried out using a ridge regression model trained for one-step ahead forecasting.

- Section 2 details the analysis of trend. This analysis is split into two parts. The statistical analysis that described whether the time series is trend-stationary or not are described. Besides, we test whether some typical transformation used to deal with trend lead to better forecasting performance.
- The seasonal component is analised in Section 3. Several statistical tests are carried out for different seasonal periods. These evaluate not only seasonal stationarity but also statistical differences among seasonal groups. Finally, tentative experimental results are reported which indicate possible directions for modeling seasonality.
- Section 4 presents the analysis of the variance. Different heteroskedasticity tests are carried out. Besides these, experiments are performed to assess whether transforming the data improves forecasting performance.

Other analysis were carried out besides the ones shown in the report. But, these were not deemed significant enough to be included.

- Finally, no change points were found in the time series.

Trend

Trend refers to the long-term change in the mean level of a time series. It reflects systematic and gradual changes in the data over time. Understanding the trend is important for identifying long-term growth or decline, structural changes, and making informed modeling decisions. This section examines the characteristics of the trend of the time series.

Trend Line Plot

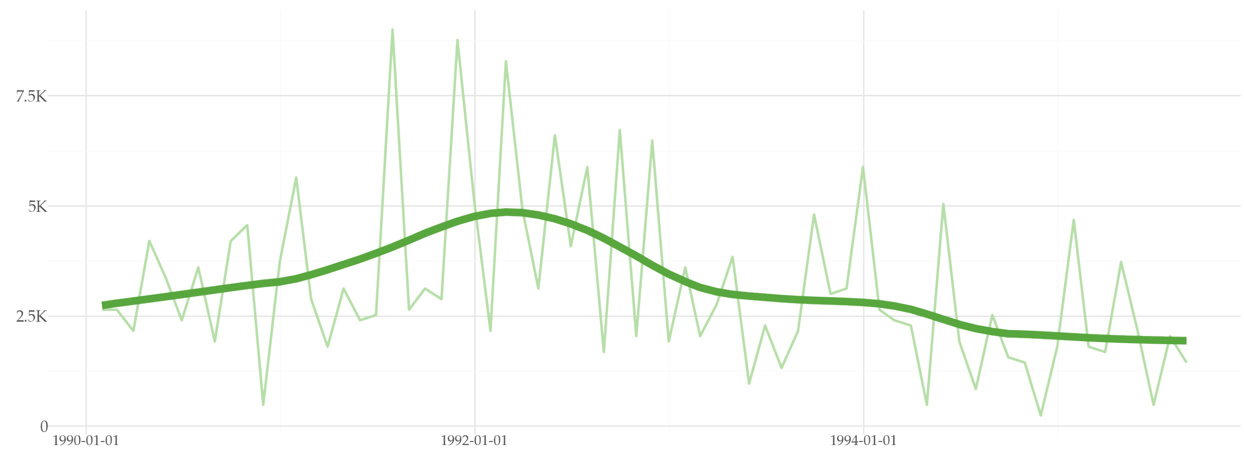


Figure 6: Time series trend plot.

- The time series has non-stationary trend according to the statistical test(s): KPSS. There is a slight downward trend.
- On the other hand, the methods Augmented Dickey-Fuller and Philips-Perron did not find evidence for the presence of trend.
- The same tests were applied to analyse whether the time series is stationary around a constant level. The method(s) KPSS and Augmented Dickey-Fuller reject this hypothesis. But, the test(s) Philips-Perron fail to reject.
- Including a trend explanatory variable which denotes the position of each observation improves forecasting performance.

Distribution of Differences

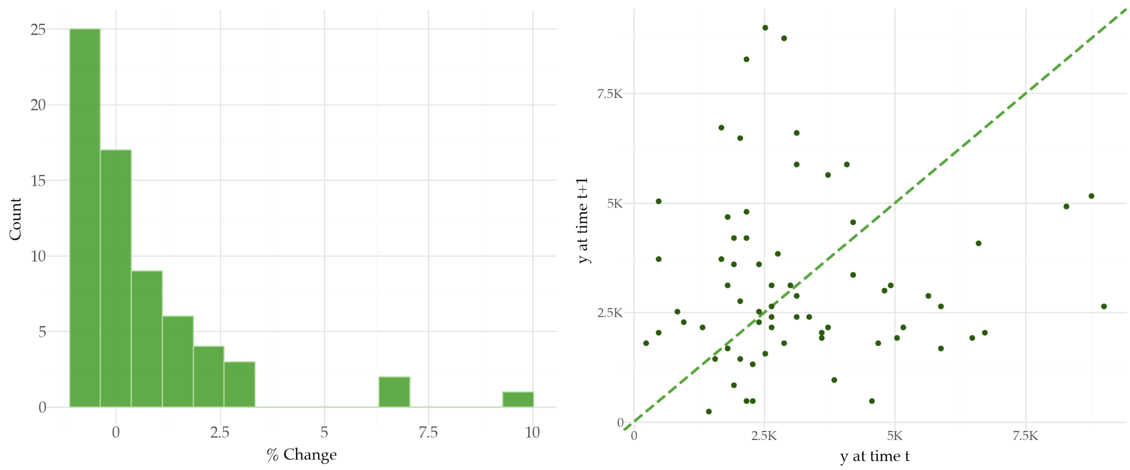


Figure 7: Distribution of percentage changes (left), and a Lag-plot (right). These plots help to understand how the data changes over consecutive observations. The histogram shown the distribution of these changes. The lag-plot depicts the randomness in the data. The time series shows greater randomness as the points deviate from the dotted line.

- The Kolmogorov-Smirnov test rejects the hypothesis that the differenced series is distributed according to the following distributions: Power-law, Exponential, Pareto, Log-Normal, and Chi-squared.
- The distribution with largest p-value is Logistic (p-value equal to 1.0). But, we cannot reject the hypothesis that the differenced series follows the following distributions (ordered decreasingly by p-value): Exponentially Modified Gaussian distribution, Gamma, Gaussian, and Cauchy.
- The excess kurtosis of the differenced series is equal to -0.14. This value is similar to that found from data following a Gaussian distribution
- The skewness of the differenced series is equal to 0.26, which is close to zero. This indicates a symmetric distribution, though there is a slight right skewness.

- **Forecasting experiments**: Taking first differences does not improve forecasting performance.

Seasonality

Seasonality represents recurring patterns or cycles that appear at regular intervals in time series data. These are predictable fluctuations that reflect periodic influences such as monthly, quarterly, or yearly cycles. Understanding seasonal patterns is crucial for forecasting, trend analysis, and identifying anomalies. This section examines the presence, strength, and characteristics of seasonal components in the input time series.

Seasonal Line Plot (Monthly)

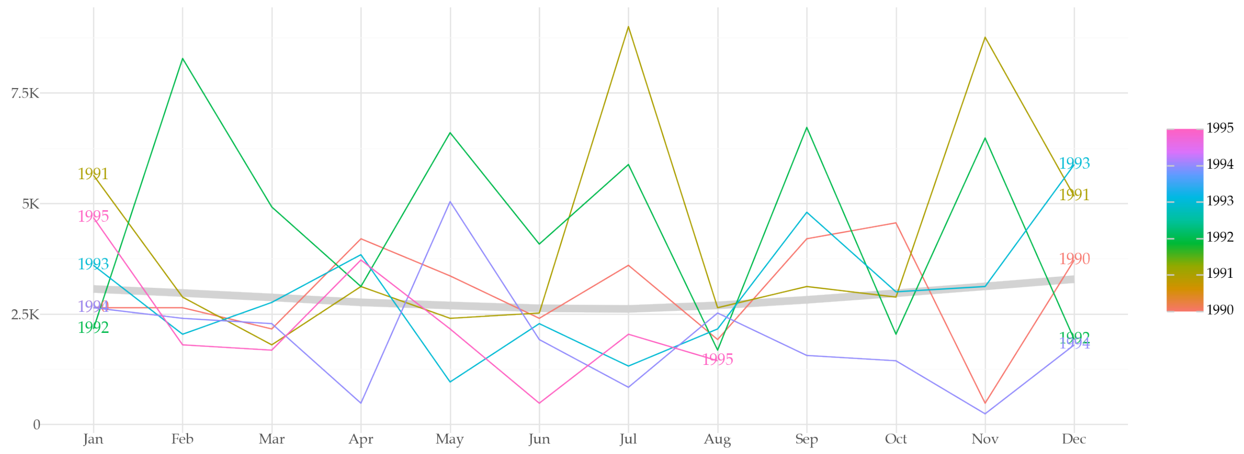


Figure 8: Seasonal plot of monthly values grouped by year.

- The following tests indicate that the time series is non-stationary in seasonality for a yearly period: OCSB. On the other hand, other tests (Wang-Smith-Hyndman) fail to reject the stationary null hypothesis.

- **Forecasting experiments**: Including monthly information in the predictive model decreases forecasting performance. This information was included as Fourier terms and repeating basis function terms in the explanatory variables.

Seasonal Sub-series Plot (Monthly)

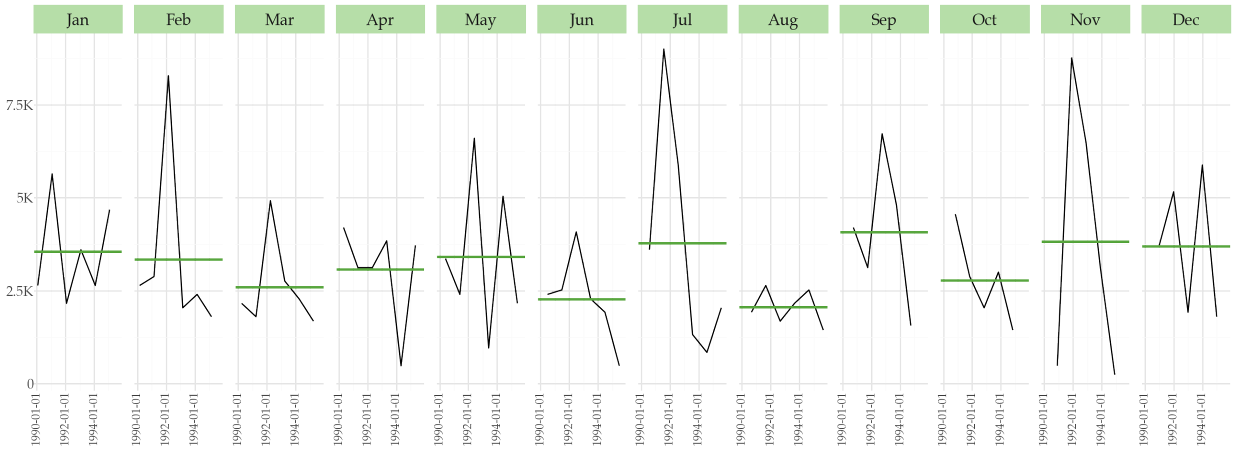


Figure 9: Monthly seasonal sub-series. This plot helps to understand how the data varies within and across monthly groups.

- Statistical tests were carried out to check for differences among means and variances across months. No significant differences were found.
- Overall, there is a reasonable evidence that the time series is not stationary around a constant level. But, the data is constant around a level in each Month.
- **Forecasting experiments**: There is evidence for a yearly seasonal pattern from statistical tests. Yet, including information about this period in the forecasting model decreased its performance.

Seasonal Sub-series Plot (Quarterly)

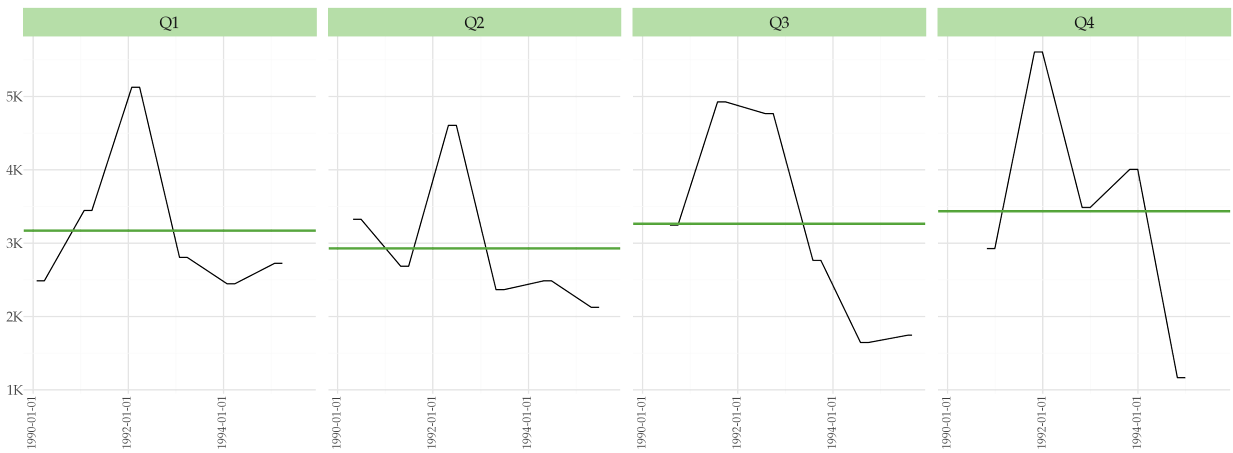


Figure 10: Quarterly seasonal sub-series. This plot helps to understand how the data varies within and across quarterly groups.

- Statistical tests were carried out to check for differences among means and variances across quarters. No significant differences were found.

- The following tests indicate that the time series is non-stationary in seasonality for a quarterly period: OCSB. On the other hand, other tests (Wang-Smith-Hyndman) fail to reject the stationary null hypothesis.
- Overall, there is a reasonable evidence that the time series is not stationary around a constant level. But, the data is constant around a level in each Quarter.
- **Forecasting experiments**: There is evidence for a quarterly seasonal pattern from statistical tests. Yet, including information about this period in the forecasting model decreased its performance.

Variance

Variance measures how data points spread around the average value in your time series. This section examines whether the variability remains stable (homoskedastic) or changes (heteroskedastic) over time. Understanding variance patterns is crucial for selecting appropriate modeling techniques, which can have a significant impact on forecasting accuracy.

Heteroskedasticity Testing

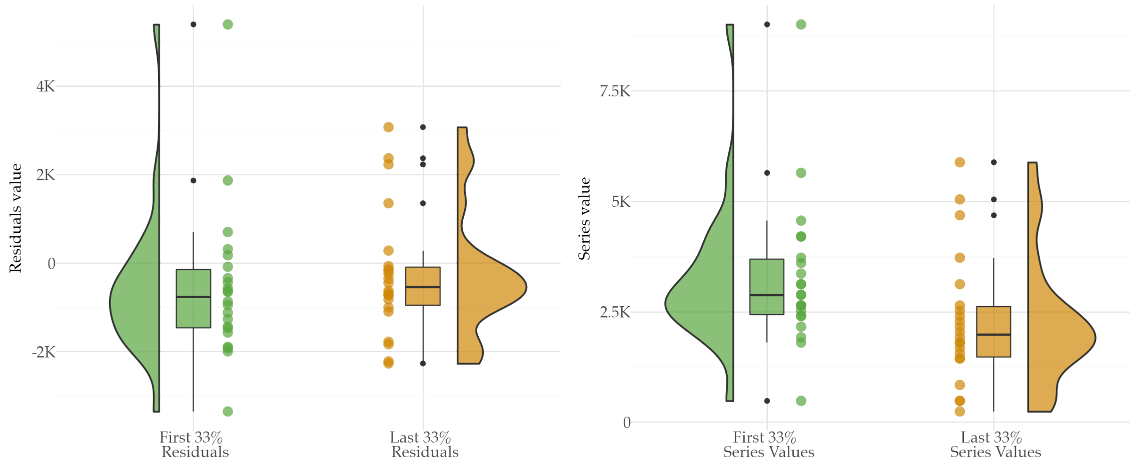


Figure 11: Time series residuals analysis. Difference in the distribution of the residuals (left) and the series (right) in the first and last thirds of the series, following a Goldfeld-Quand partition.

- In the analysis of seasonality we did not find significant differences in the dispersion among periodic groups of observations.
- No statistical evidence was found for the hypothesis that the time series is heteroskedastic, according to the White, Breusch-Pagan, and Goldfeld-Quandt tests.
- **Forecasting experiments**: Transforming the series with either the logarithm or the Box-Cox method improved forecasting performance.
- A Logarithm transformation did not affect the most appropriate distribution to fit the data (Exponentially Modified Gaussian distribution distribution).