ECON 723 Econometrics II

ECON 723 Group Project 1

name

id

22 Jun 2023

Abstract

Contents

1 Overview

2 Literature Review

3 Data Acquisition and Management

3.1 Data Sources

Our dataset comprises the monthly yields for 1, 2, 5, and 10-year periods for both New Zealand (NZ) and the United States (US). Additionally, it includes the exchange rates and 90 day bank bill yields from both countries.

3.2 Data Enrichment

Being a time series dataset, our primary focus is to ascertain the presence of trends, seasonality, autocorrelation, and its stationarity. For this, we decomposed the time series data into its constituent elements: the seasonal, trend, and random (or irregular) components.

To achieve this, we utilized the decompose model. An analysis of the resultant table suggests an absence of a distinct trend. However, the data exhibits clear seasonality while appearing somewhat random.

$$Y_t = Trend_t + Seasonal_t + Random_t$$

From the table, we can see that there is not a clear trend and clearly the data is seasonal, and it looks random.

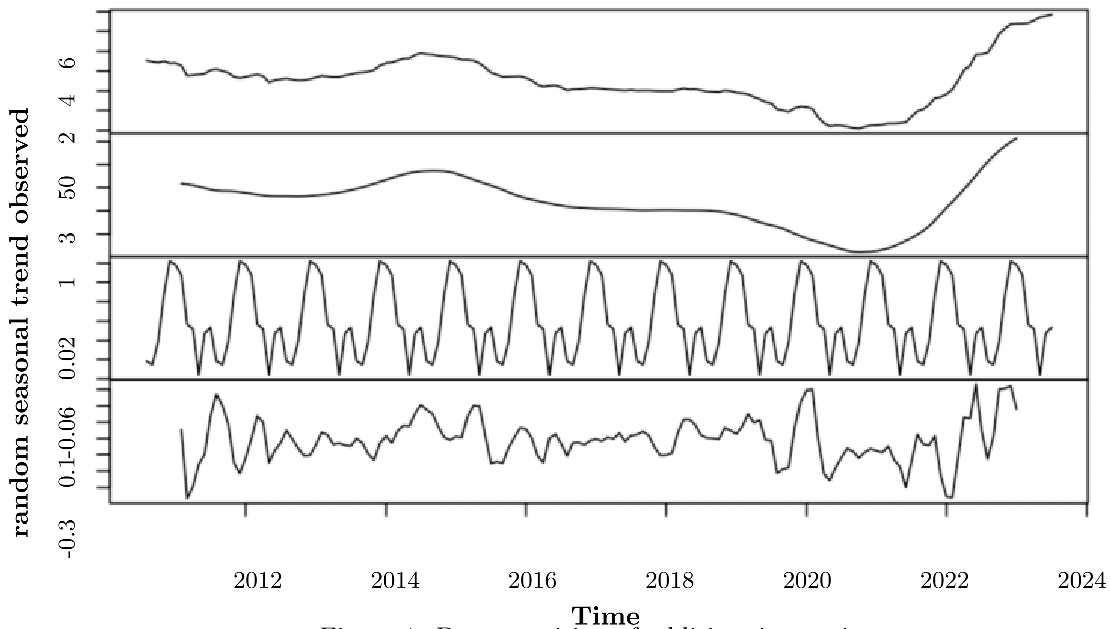

Figure 1: Decomposition of additive time series

It's worth noting that a non-stationary time series will always contain the a random, volatile component. This component encapsulates minor fluctuations around the overarching trend, cycle, and seasonal elements. While unpredictable, this can typically be modeled as random observations deriving from a certain distribution, often represented as $N(0, \sigma^2)$.

3.2.1 ACF test

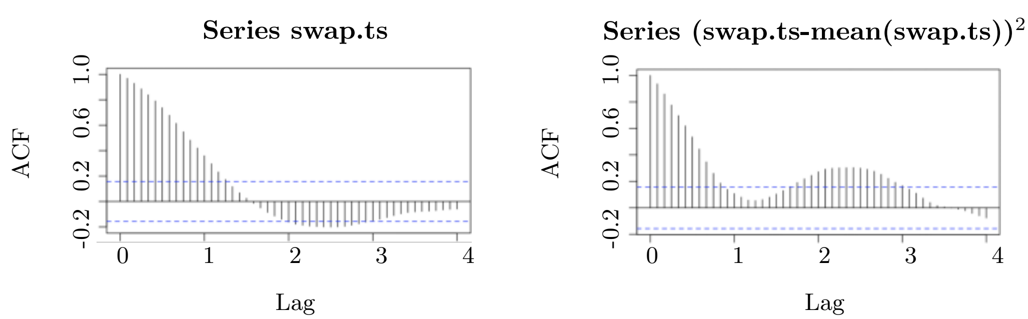

Figure 2:

The Autocorrelation Function (ACF) points to a robust autocorrelation at lag(16) in the swap rate, hinting at seasonality. It is observed that as time progresses, the positive correlation tends to wane until it turns negative. However, this correlation also diminishes with increasing lag. Our primary focus is on data points outside the blue region, given their high statistical significance. An analysis of heteroscedasticity corroborates these findings.

```
        Augmented Dickey-Fuller Test

data: swap.ts
Dickey-Fuller  =-1.0953, Lag order  =5, p -value  =0.9203
alternative hypothesis: stationary
```

3.2.2 The ADF test:

For a time series to be classified as stationary, it must be devoid of trends or seasonal impacts. If the procured p-value surpasses the significance level of 0.05 and the ADF statistic is greater than any of the established critical values, we have no compelling reason to reject the null hypothesis. As such, our conclusion is that the time series is non-stationary.

Furthermore, a similar analysis was conducted for the two-year period swap rate. The findings paralleled the observations made previously.

```
        Augmented Dickey-Fuller Test

data: swap.ts1
Dickey-Fuller  = -0.96763, Lag order  = 5, p-value  = 0.9405
alternative hypothesis: stationary
```

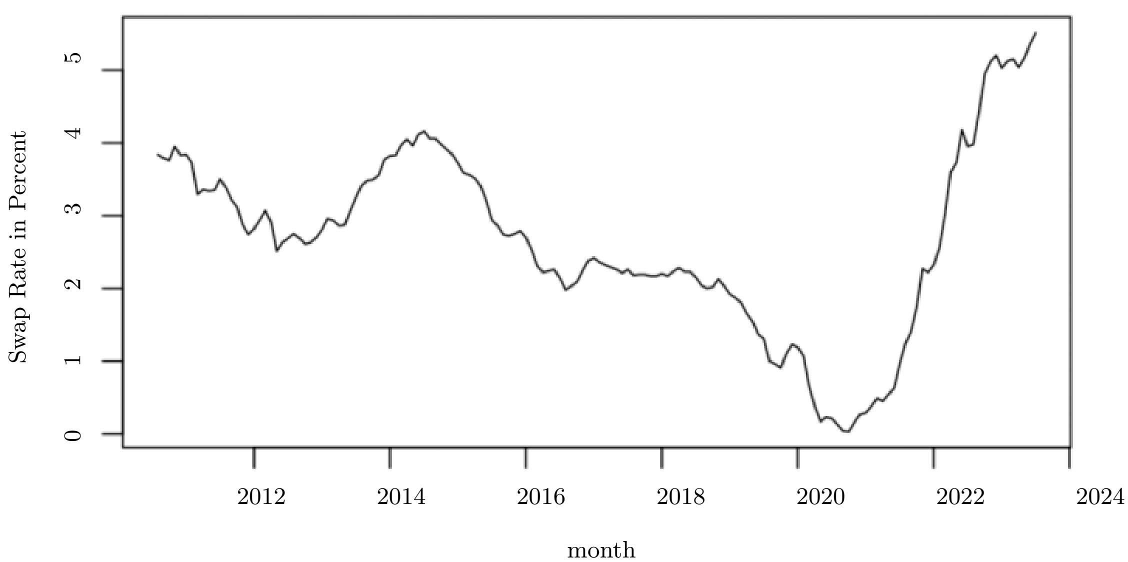

Figure 3: Swap Rate for 2 Year Period

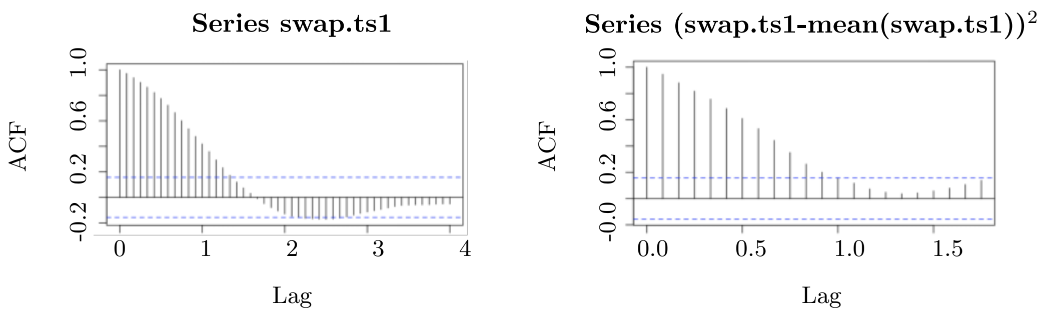

Figure 4:

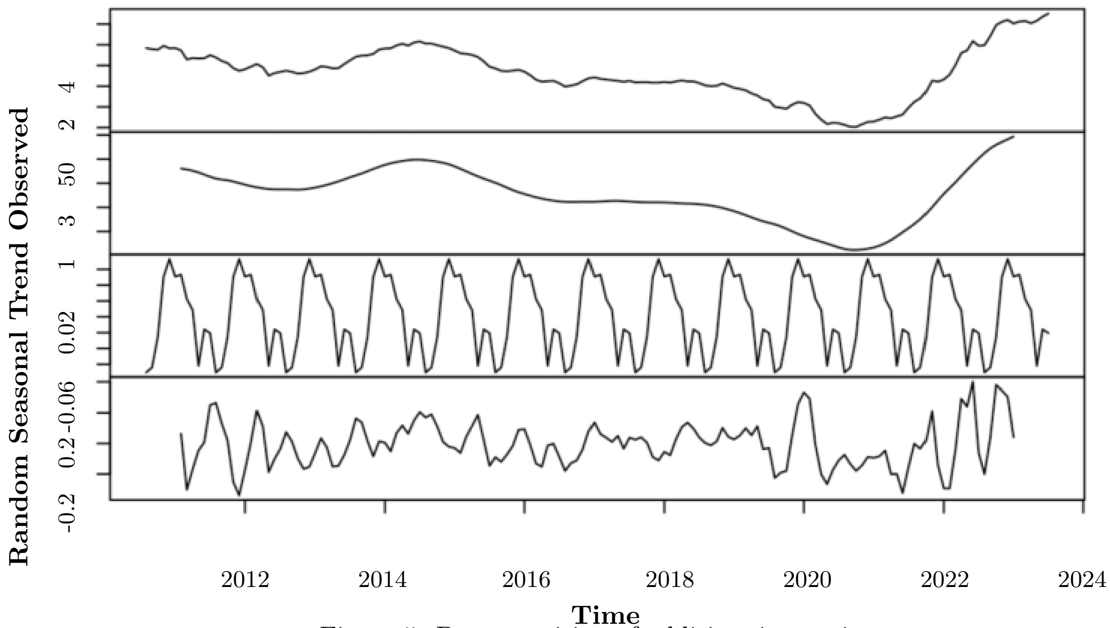

Figure 5: Decomposition of additive time series

3.3 Predicting Yield Rates Using Swap Rates

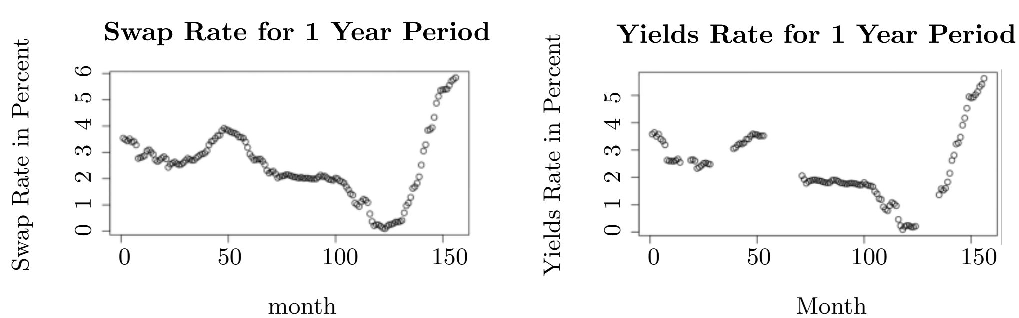

Figure 6:

3.3.1 Correlation Between Swap and Yield Rates

We initiated our analysis by graphing the trends between the swap and yield rates. A visual inspection suggests a potential correlation. This observation led us to employ regression using data from both the yield and swap rates. The following regression model was proposed:

$$Yield\_1year_i = c + \beta \cdot Swap\_rate\_1year_i + \epsilon_i$$

```
Call:
 lm(formula = data$bond_closing_yields_1year ~data$Swap_rates_1year)

Residuals:
     Min      1 Q    Median      3Q     Max
-0.34417  -0.06211 -0.02351  0.05513  0.42781


Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            -0.047444   0.027415  -1.731   0.0863 .
data$Swap_rates_1year   0.936676   0.009457  99.050   <2 e-16 ***
---
Signif. codes: 0 '* * *', 0.001 '* *', 0.01 '*', 0.05 '.', 0.1  '' 1
Residual standard error: 0.1328 on 113 degrees of freedom
  (41 observations deleted due to missingness)
Multiple R-squared: 0.9886,  Adjusted R-squared: 0.9885
F-statistic: 9811 on 1 and 113 DF, p-value: < 2.2e-16
```

3.3.2 Omitted Variable Bias

Though the model appeared to be significant, closer scrutiny of the one-year yield rate revealed gaps in the data for specific years. Relying solely on simple linear regression for prediction in such cases can lead to omitted variable bias. This necessitates the exploration of alternative, more reliable methods. The issue of missing data persisted when considering the two-year period.

```
Call:
lm(formula = data$bond_closing_yields_2year ~ data$Swap_rates_2year)

Residuals:
      Min       1Q    Median      3Q     Max
 -0.42686  -0.10933  -0.01985  0.06238  0.54959

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)              -0.13826    0.03648   -3.79 0.000229 ***
data$Swap_rates_2year     0.96271    0.01245   77.33 <2 e-16 ***

Residual standard error: 0.1628 on 130 degrees of freedom
(24 observations deleted due to missingness)
Multiple R-squared: 0.9787,  Adjusted R-squared: 0.9786
F-statistic: 5980 on 1 and 130 DF, p-value: < 2.2e-16
```

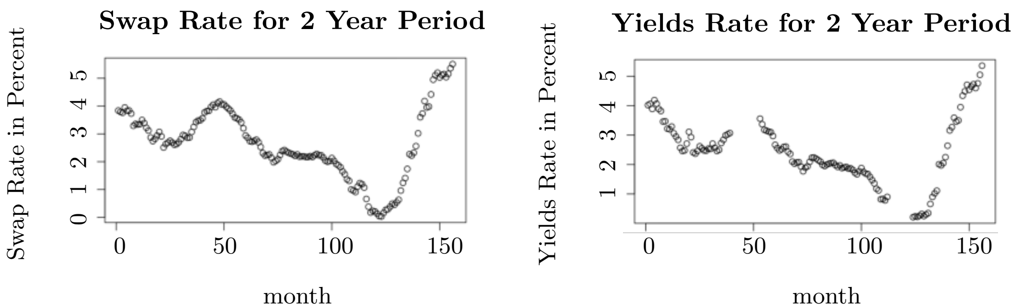

Figure 7:

3.4 Machine Learning – The Random Forest Model

Random Forest is an ensemble learning technique widely used for both classification and regression tasks.[1] It operates by constructing numerous decision trees during the training phase and then produces the mode of the classes (for classification) or the mean prediction (for regression) of the individual trees when provided with an input.[2] In our study, we trained the model using swap rate data to predict yield data.

One of the principal advantages of the Random Forest model is its foundation on multiple decision trees. Every tree is trained on a unique, randomly chosen subset of the data and offers its predictions. The model then compiles these predictions to yield a final outcome. This method leverages bagging (bootstrap aggregation) to bolster the model's robustness. When each decision tree is trained, a random data sample is chosen with replacement. Furthermore, Random Forest introduces an element of randomness during node splitting. Instead of selecting the best feature from all available ones during this process, it chooses from a random subset, leading to a diverse tree ensemble. This design helps in mitigating overfitting and maintains data accuracy even in the presence of missing values.[3]

After the data training:

[1] Boehmke and Greenwell 2020
[2] Breiman 2001
[3] Hastie et al. 2001

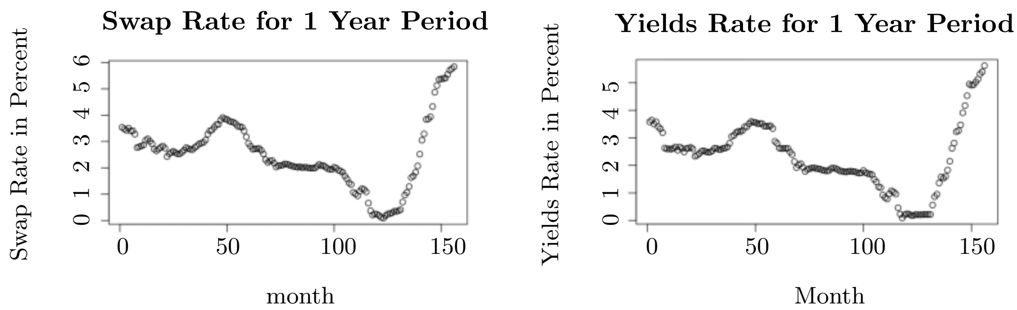

Figure 8:

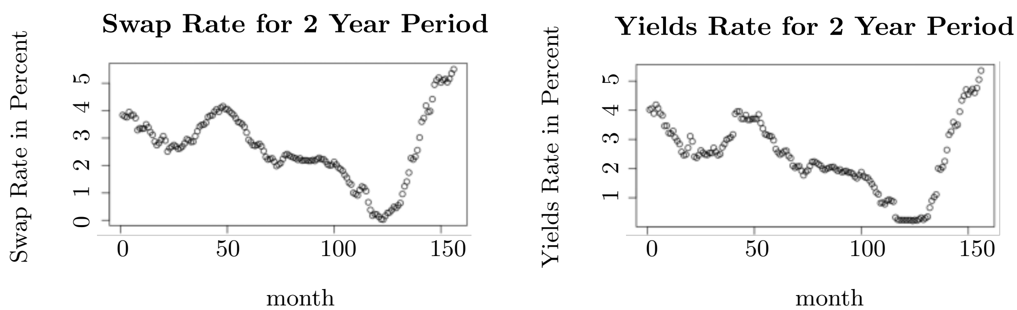

Figure 9:

Test of filled variable

```
        Augmented Dickey-Fuller Test

data: Yield.ts
Dickey-Fuller = -0.14344, Lag order = 5, p -value = 0.99
alternative hypothesis: stationary
```

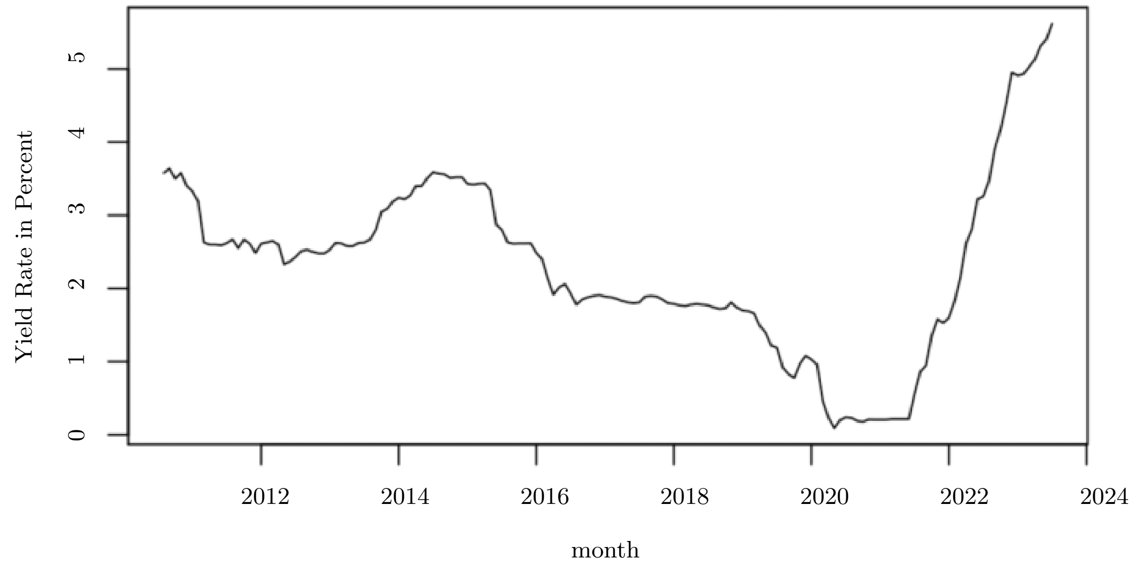

Figure 10: Yield Rate for 1 Year Period

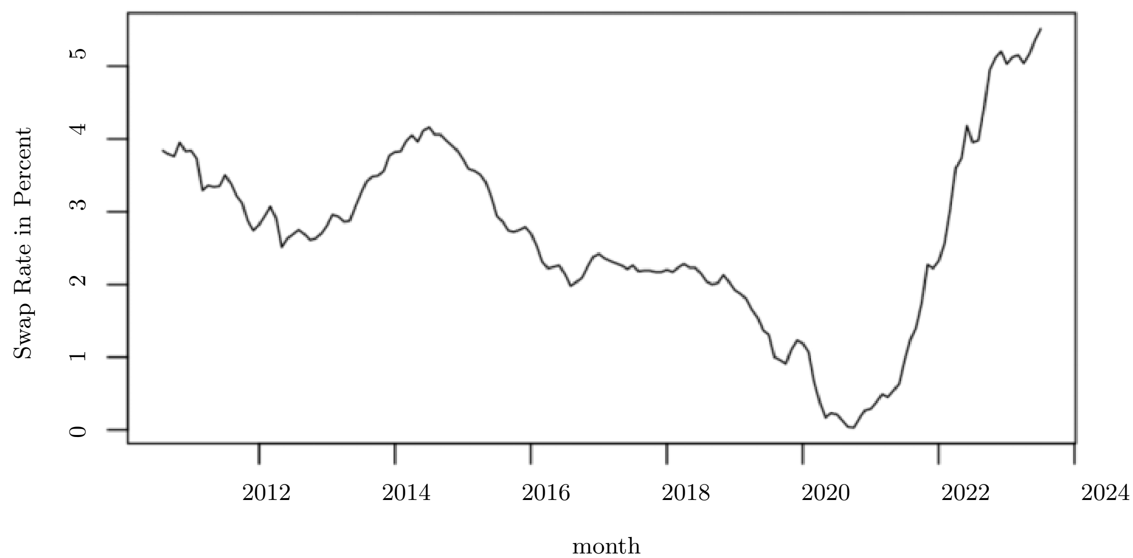

Figure 11: Swap Rate for 2 Year Period

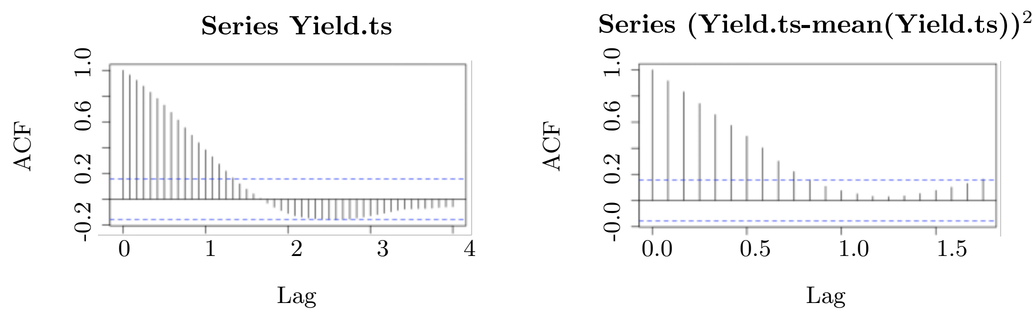

Figure 12:

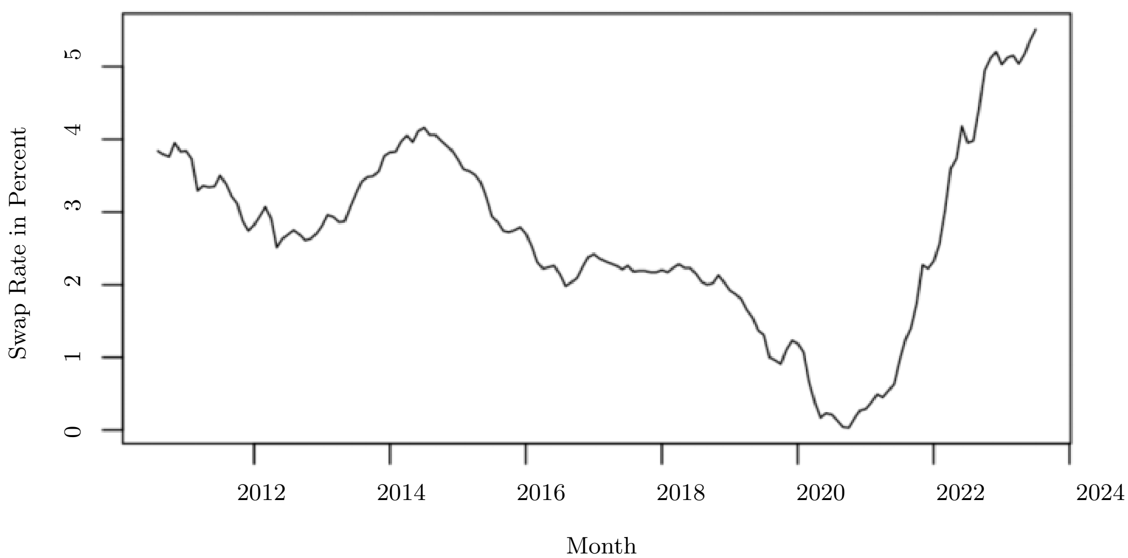

Figure 13: Swap Rate for 2 Year Period

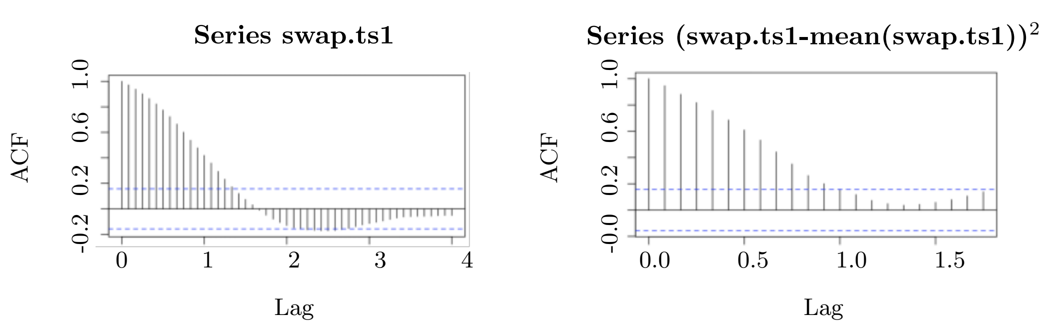

Figure 14:

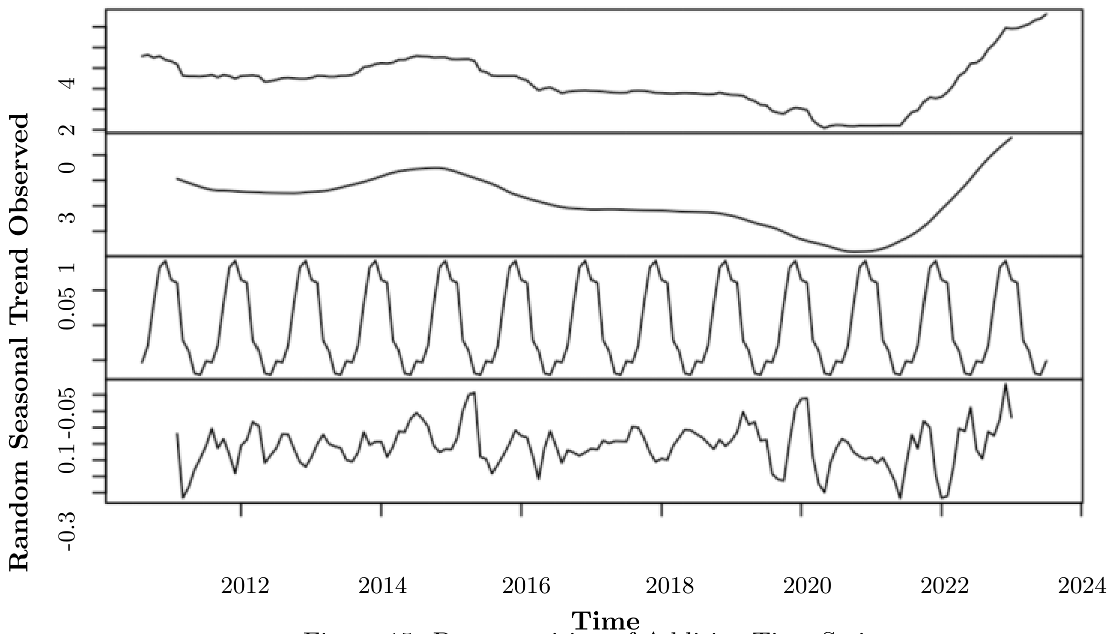

Figure 15: Decomposition of Additive Time Series

For 2 year period:

```
        Augmented Dickey-Fuller Test

data: Yield.ts1
Dickey-Fuller = -0.29267, Lag order = 5, -value = 0.9899
alternative hypothesis: stationary
```

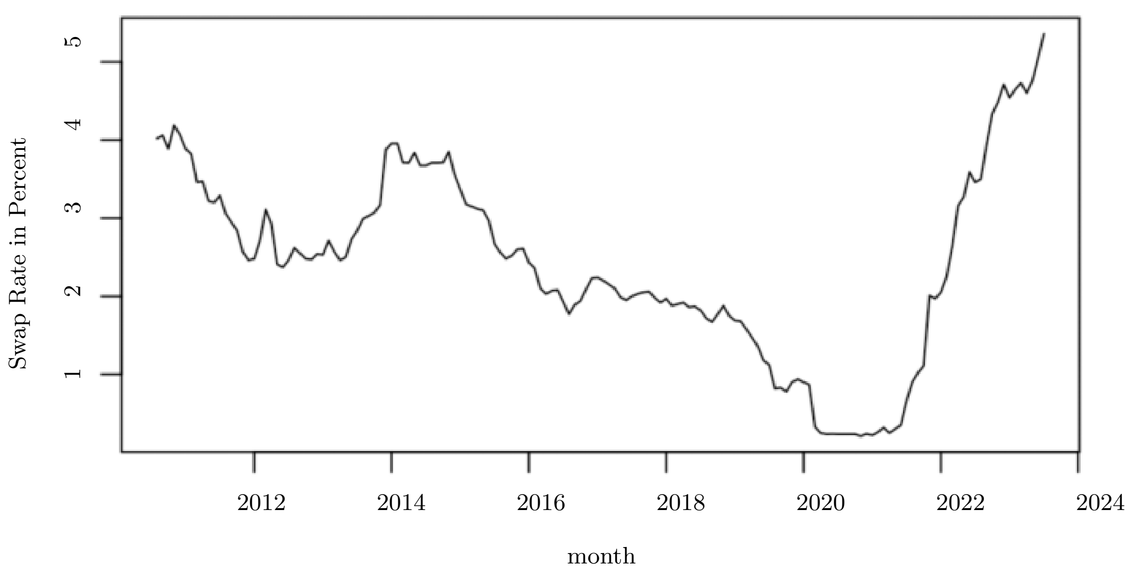

Figure 16: Yield rate for 2 Year Period

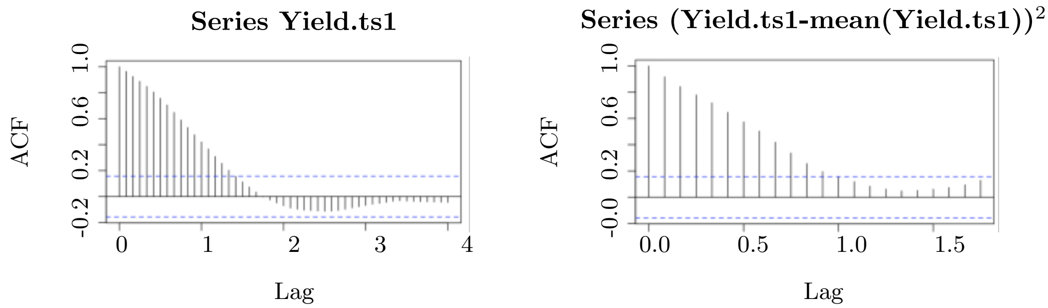

Figure 17:

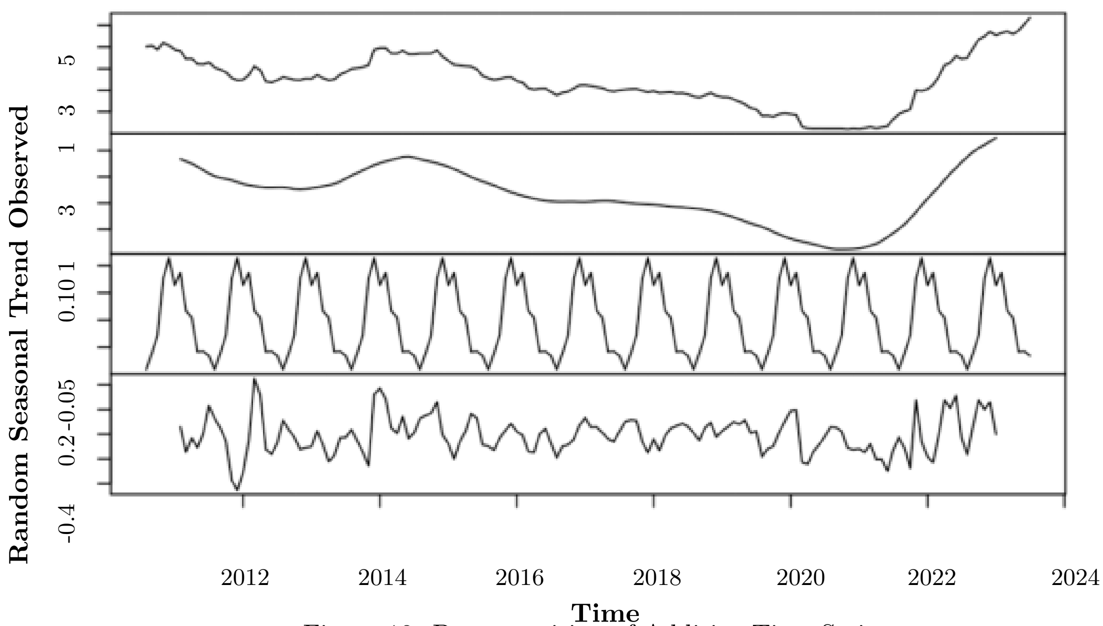

Figure 18: Decomposition of Additive Time Series

For the absent data preceding August 2010, we leveraged the interbank rate for predictions. The method resembled our previous predictions. Given the continuous data absence in 2009 – a tumultuous year post the financial crisis – a straightforward linear regression based on the interbank rate was deemed unsuitable.

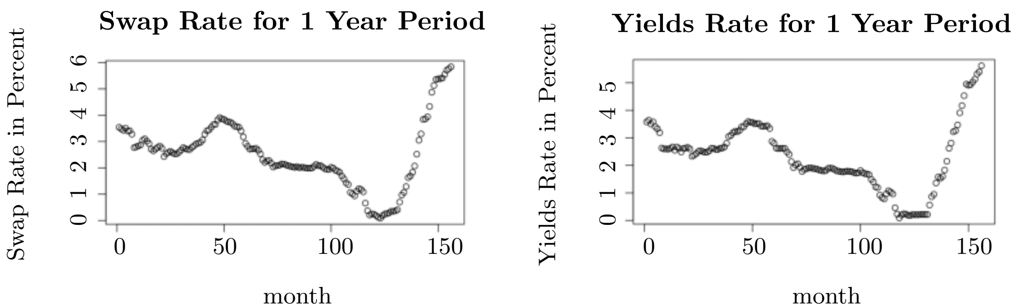

Figure 19:

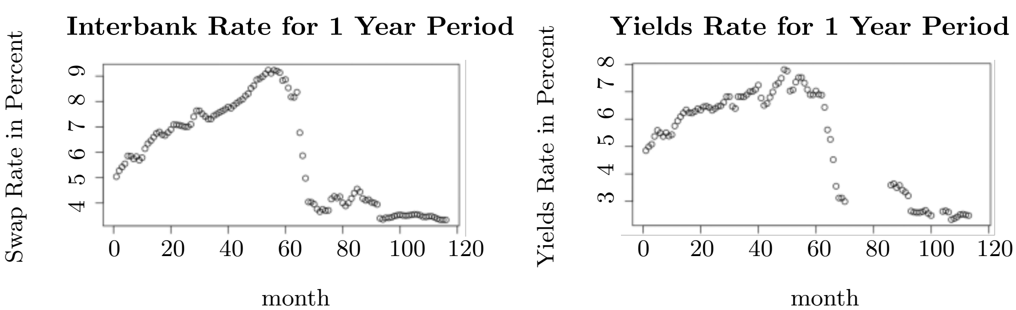

Figure 20:

After filled

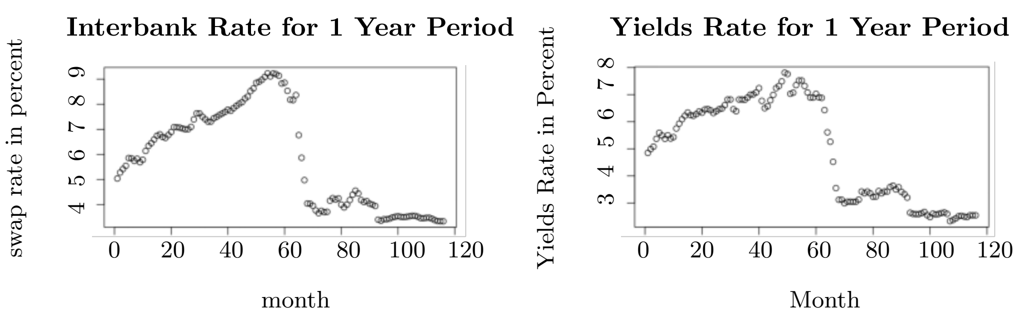

Figure 21:

Final data:

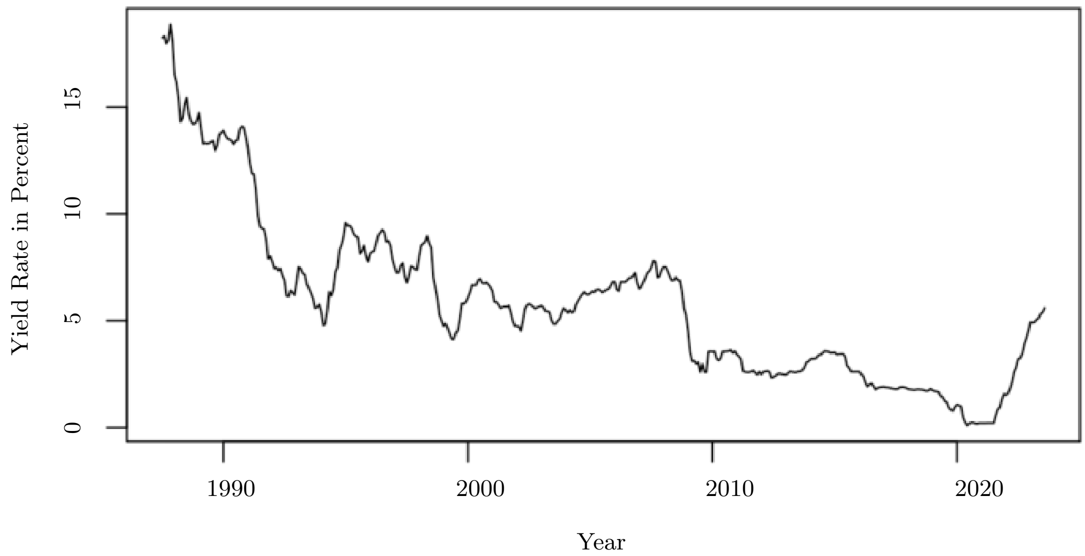

Figure 22: Yields Rate for 1 Year Period

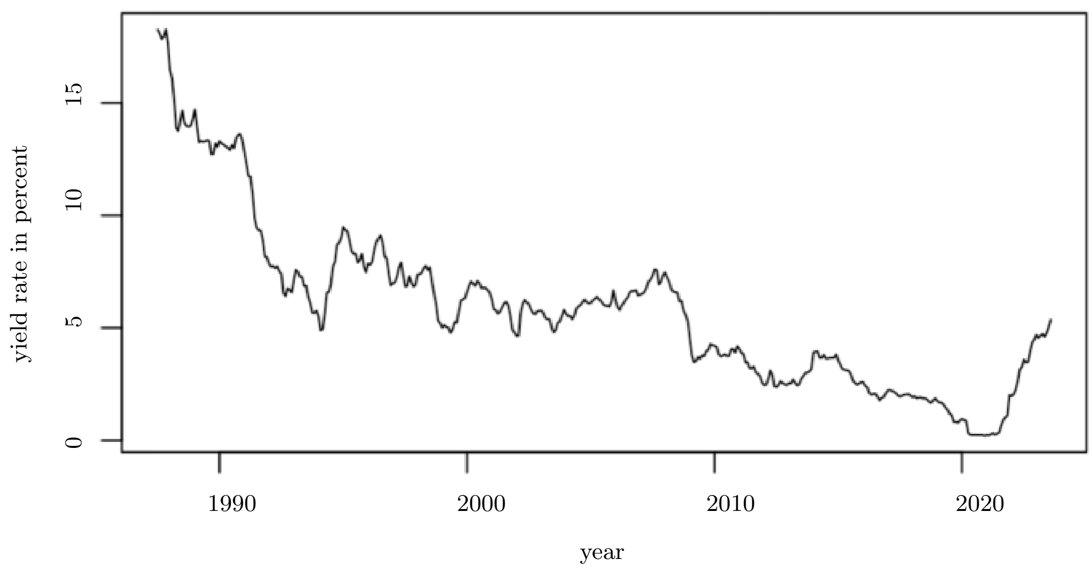

Figure 23: Yields rate for 2 Year Period

3.5 Conversion of Daily Data to Monthly Data:

Certain datasets, such as the US yield data, were provided on a daily basis. To standardize this with our monthly datasets, I utilized Pivot Tables and the average function to transform these daily figures into monthly averages.

References

Boehmke, B. and B. Greenwell (Feb. 2020). *Hands-on machine learning with R*. URL: `https://bradleyboehmke.github.io/HOML/random-forest.html`.

Breiman, L. (2001). “Random forests”. In: *Machine learning* 45, pp. 5–32.

Hastie, T., R. Tibshirani, and J. Friedman (2001). “The elements of statistical learning. Springer series in statistics”. In: *New York, NY, USA*.

A Appendix A - Figure

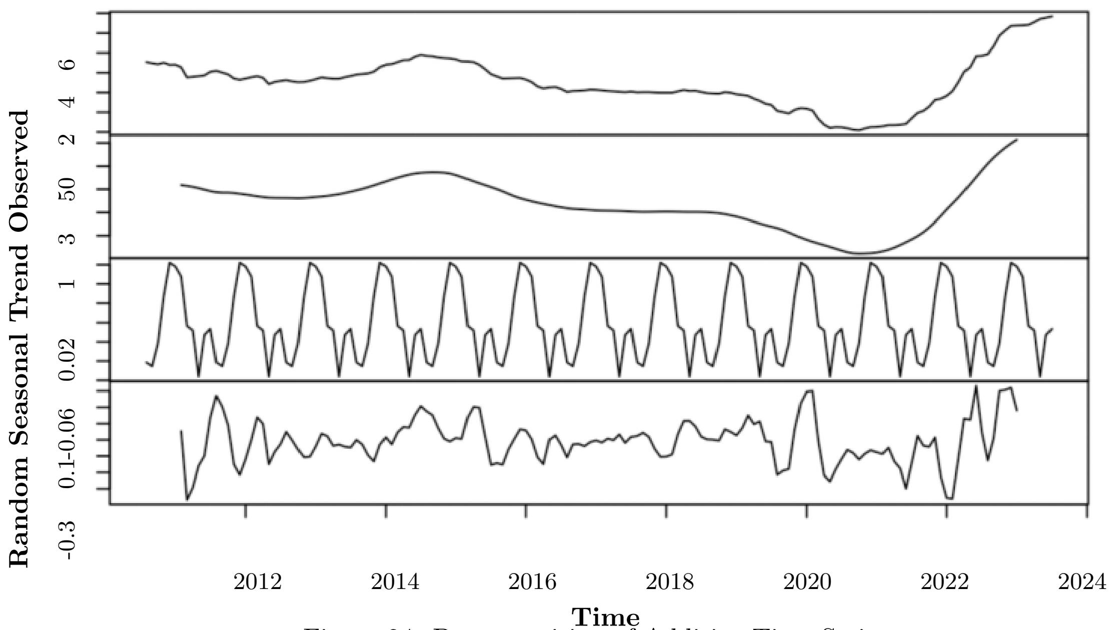

Figure 24: Decomposition of Additive Time Series

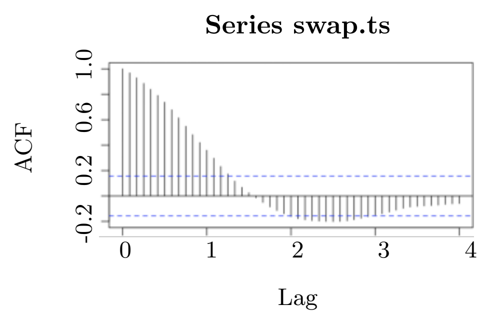

Figure 25: Series swap.ts

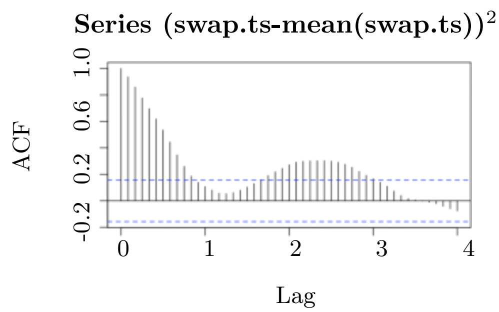

Figure 26: **Series (swap.ts-mean(swap.ts))**^{2}

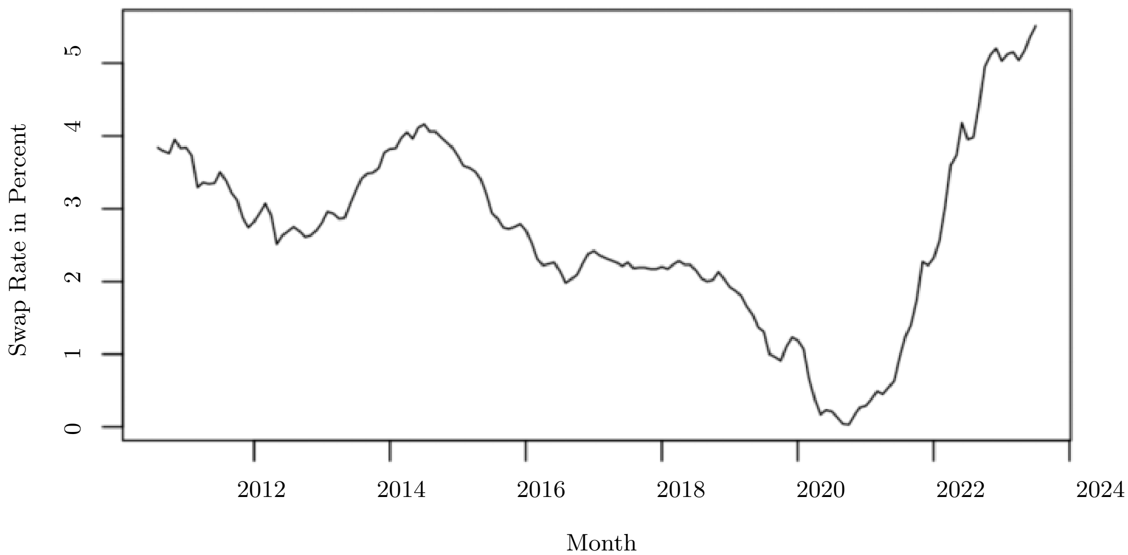

Figure 27: Swap Rate for 2 Year Period

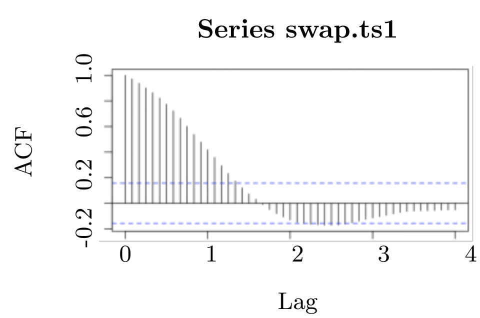

Figure 28:

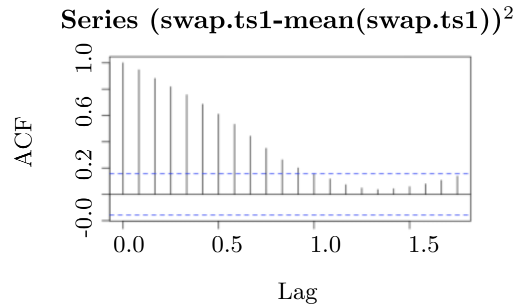

Figure 29:

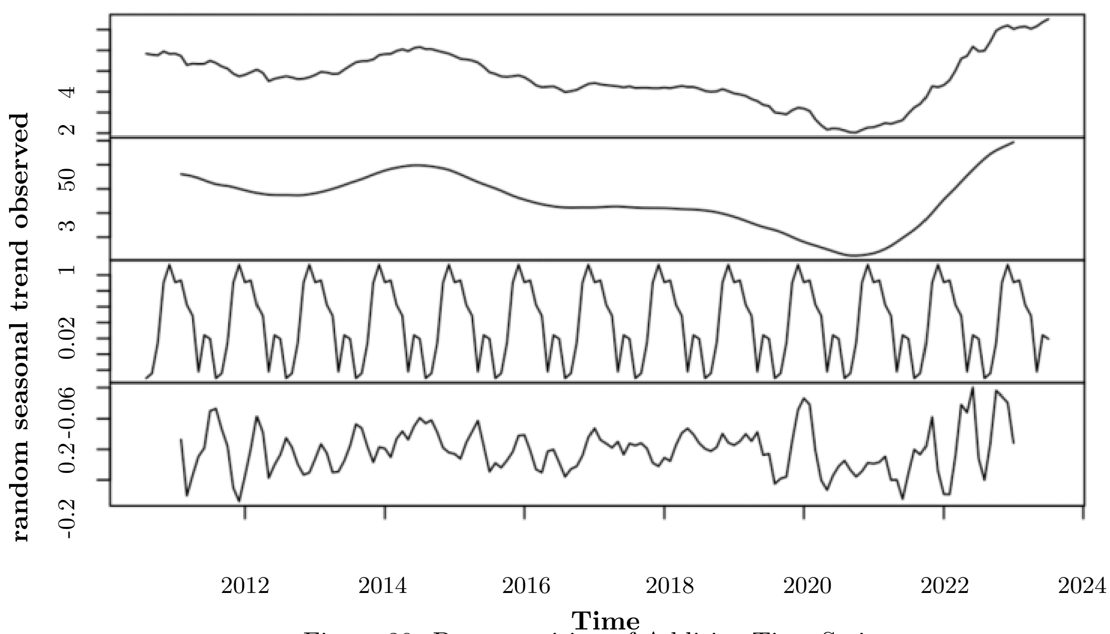

Figure 30: Decomposition of Additive Time Series

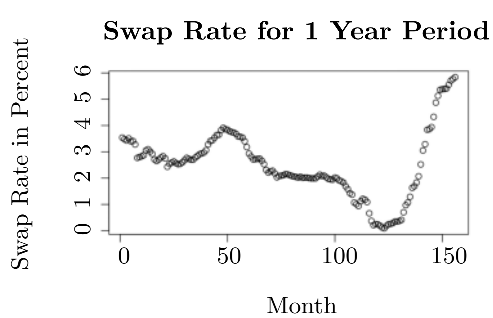

Figure 31:

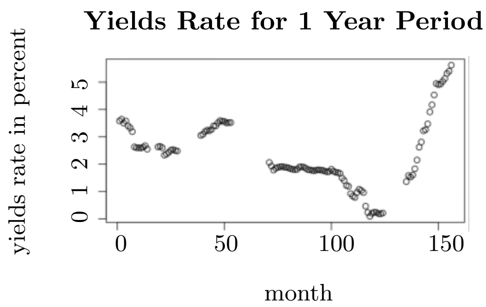

Figure 32:

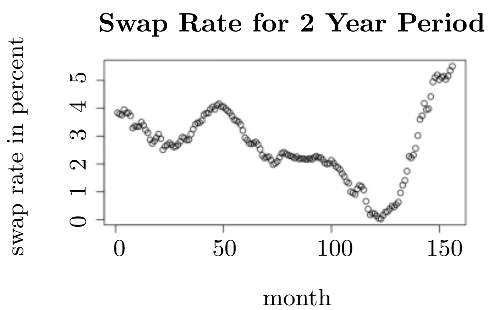

Figure 33:

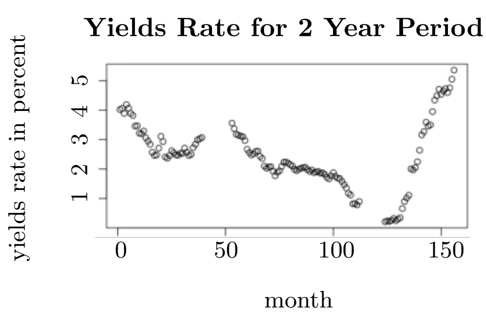

Figure 34:

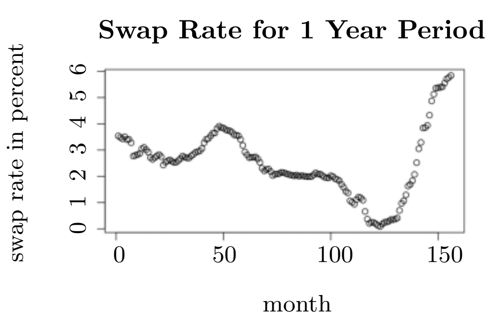

Figure 35:

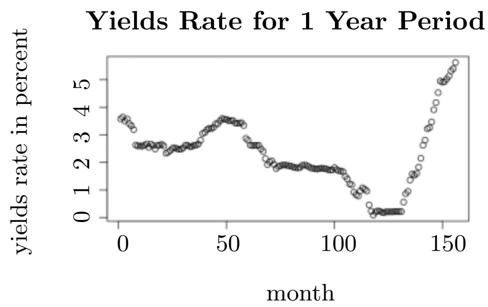

Figure 36:

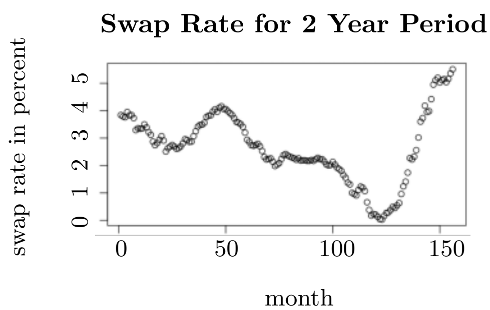

Figure 37:

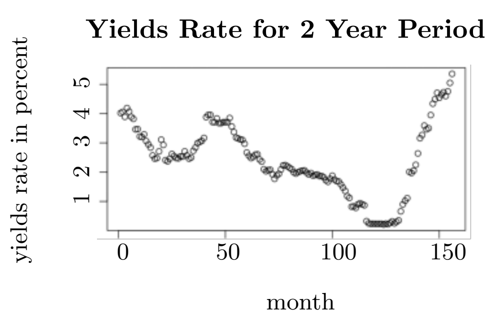

Figure 38:

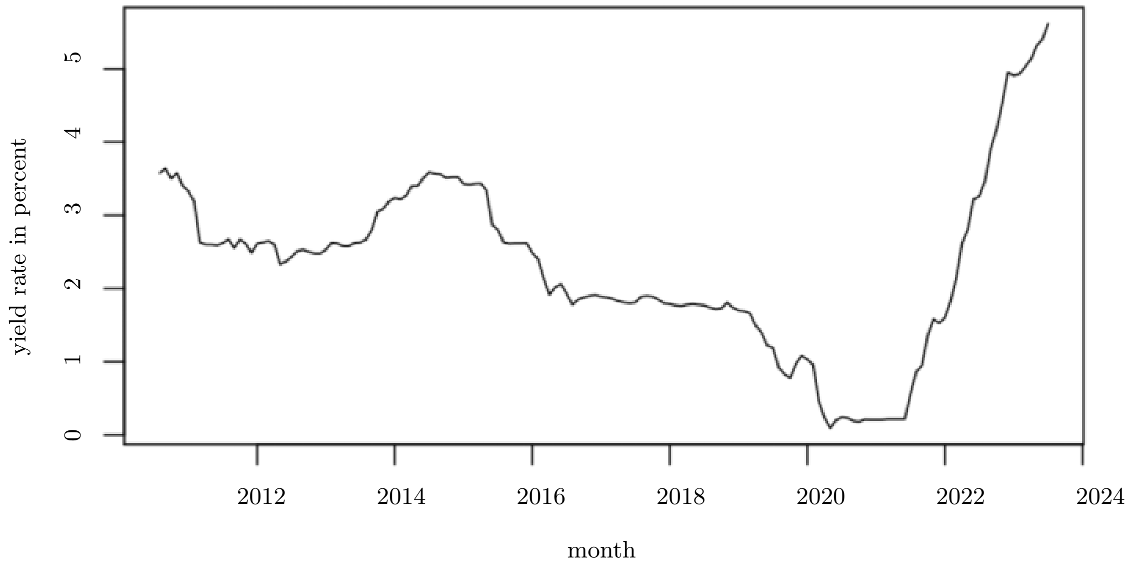

Figure 39: Yield rate for 1 Year Period

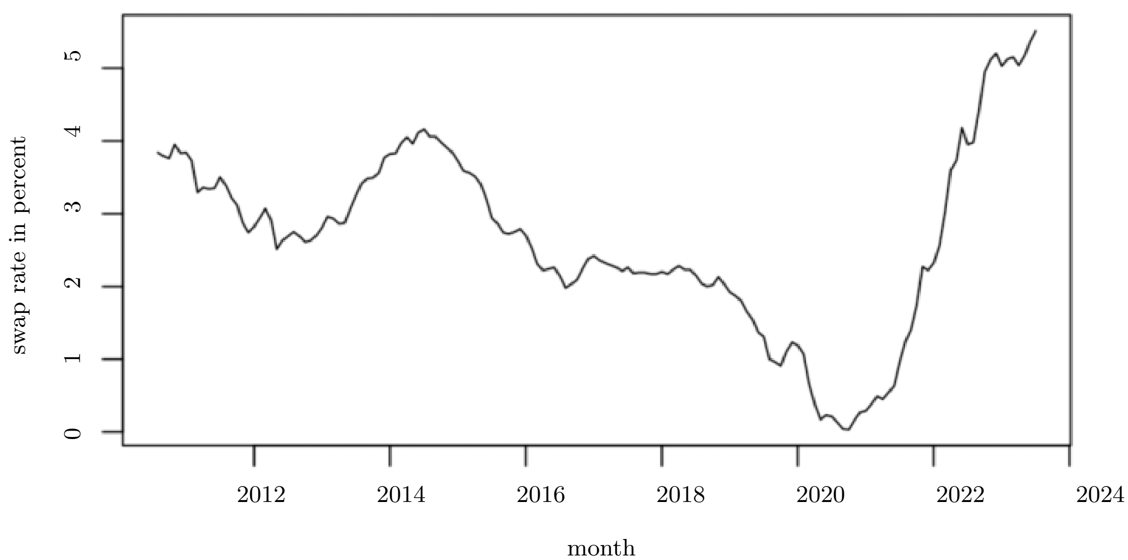

Figure 40: Swap rate for 2 Year Period

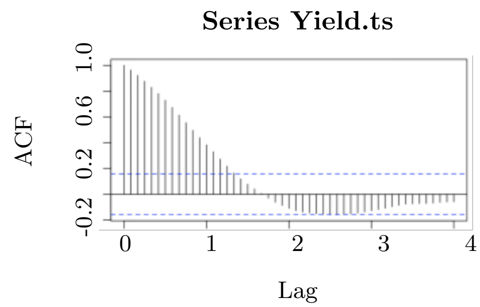

Figure 41:

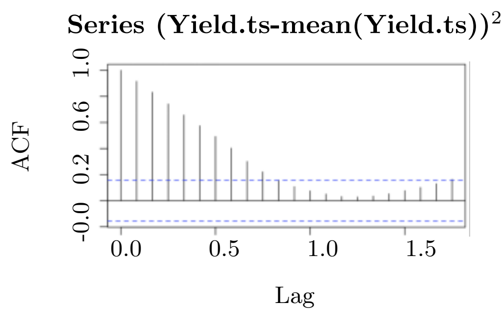

Figure 42:

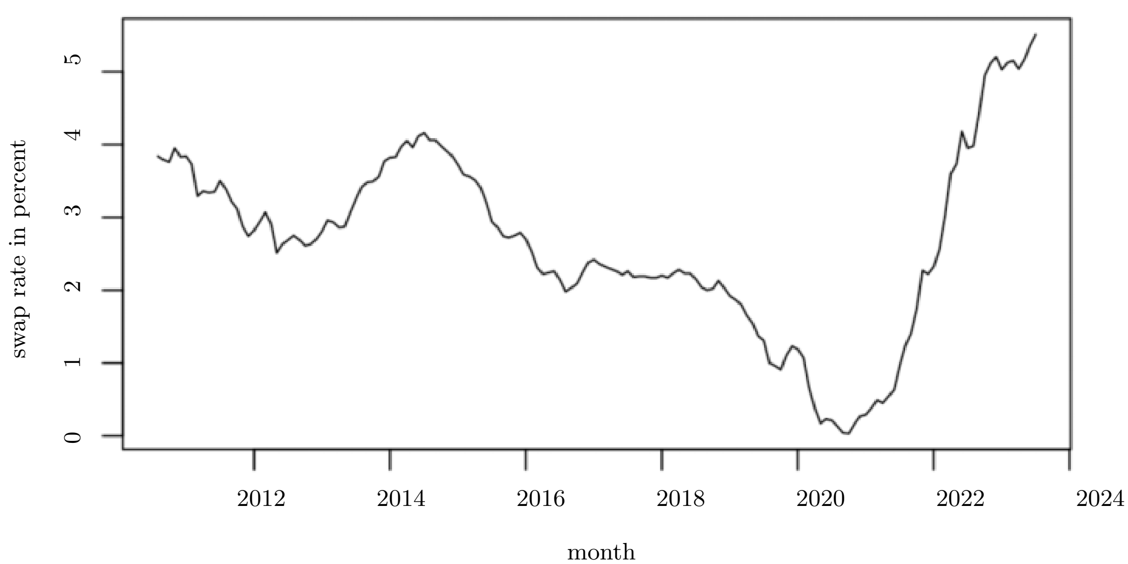

Figure 43: Swap rate for 2 Year Period

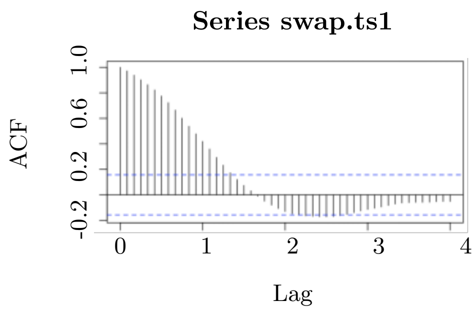

Figure 44:

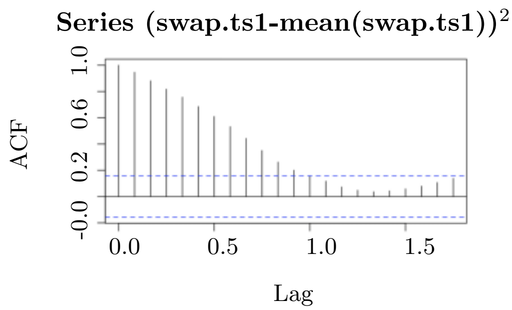

Figure 45:

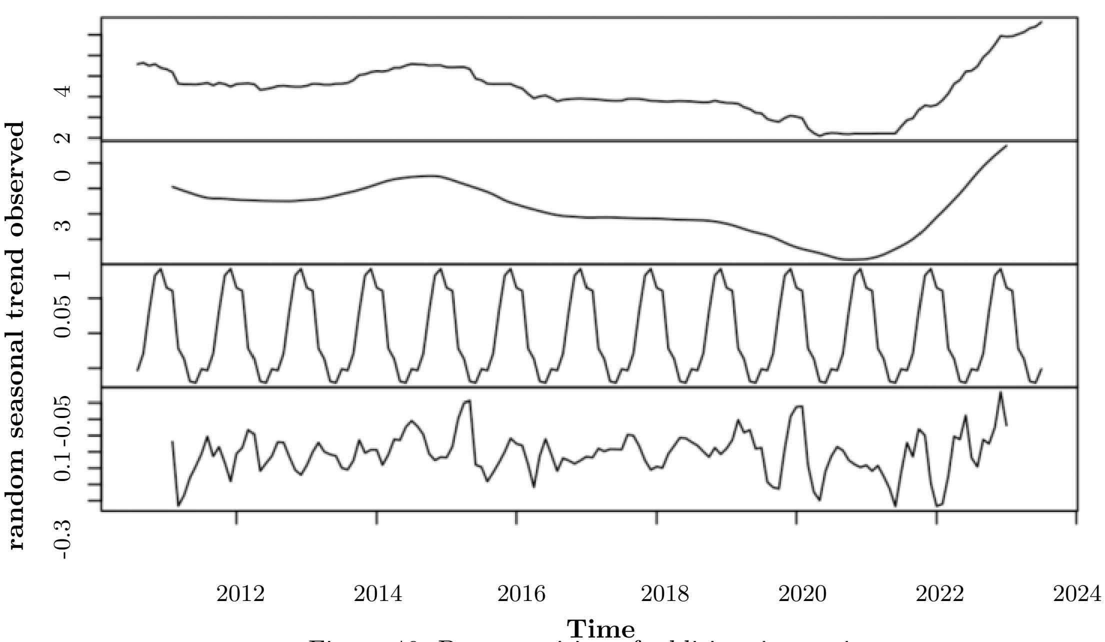

Figure 46: Decomposition of additive time series

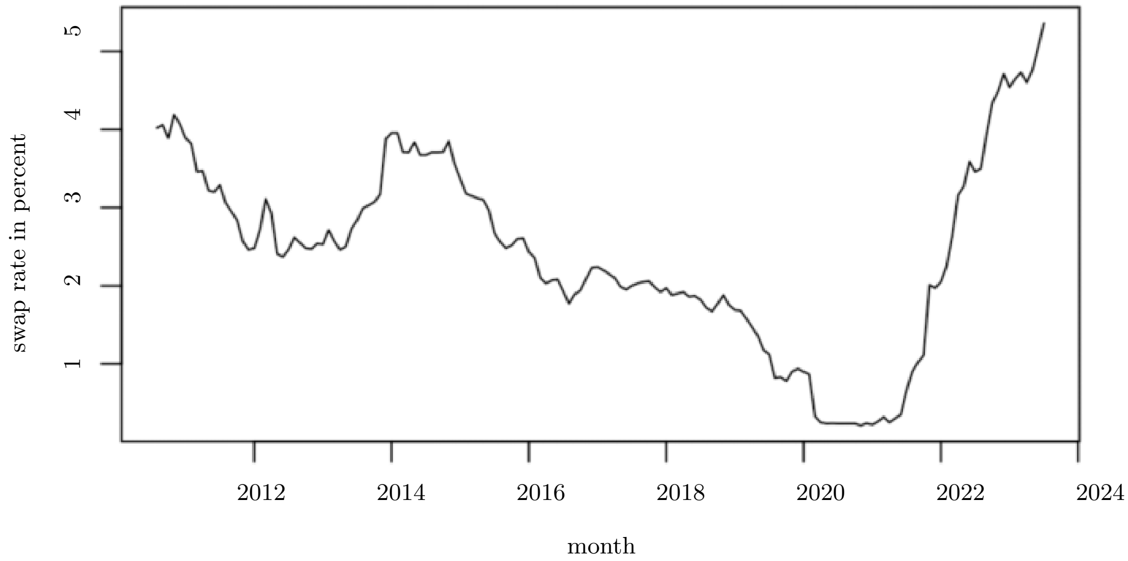

Figure 47: Yield rate for 2 Year Period

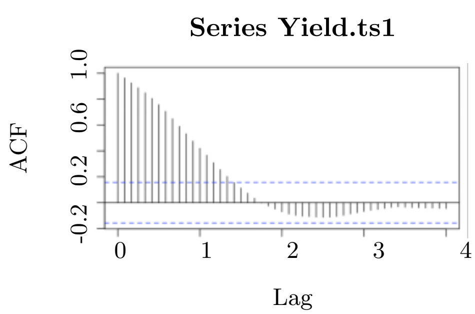

Figure 48:

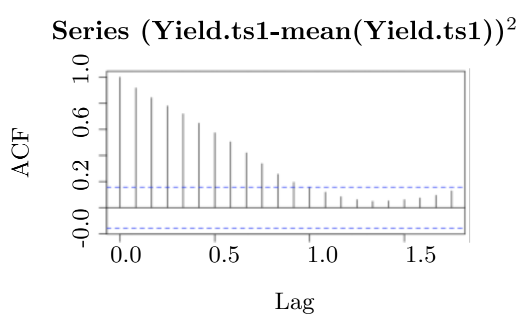

Figure 49:

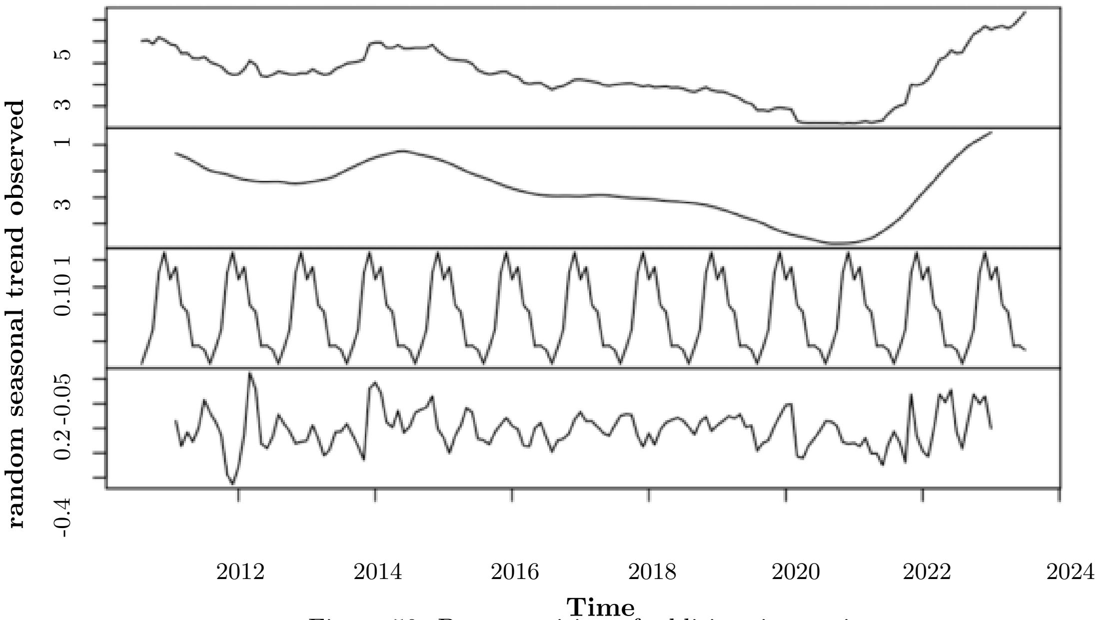

Figure 50: Decomposition of additive time series

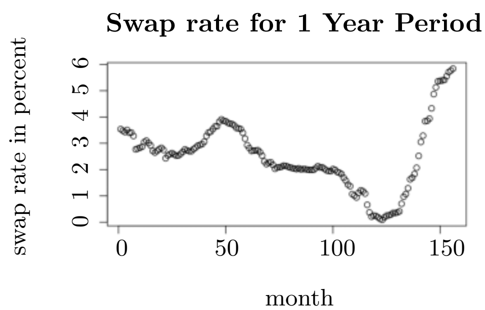

Figure 51:

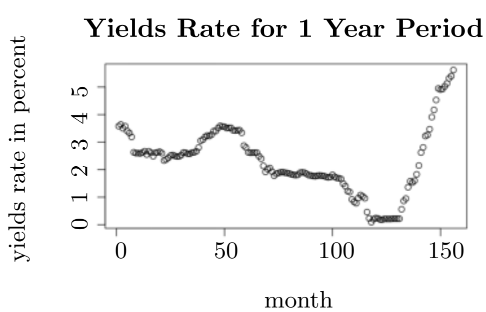

Figure 52:

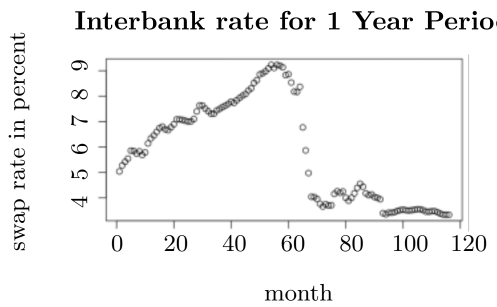

Figure 53:

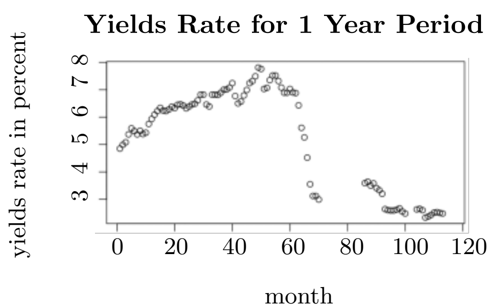

Figure 54:

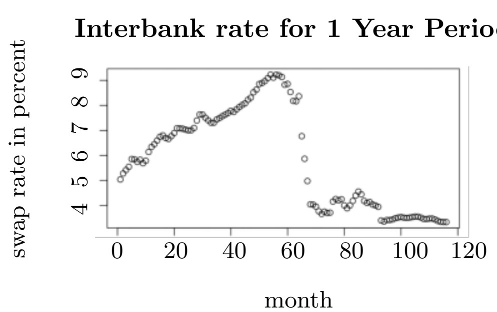

Figure 55:

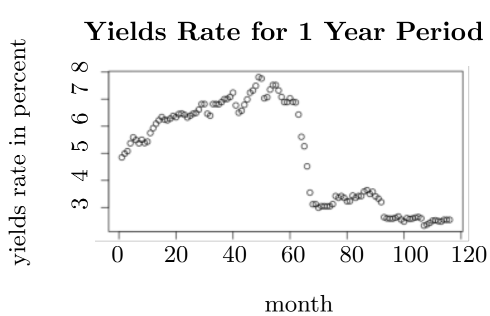

Figure 56:

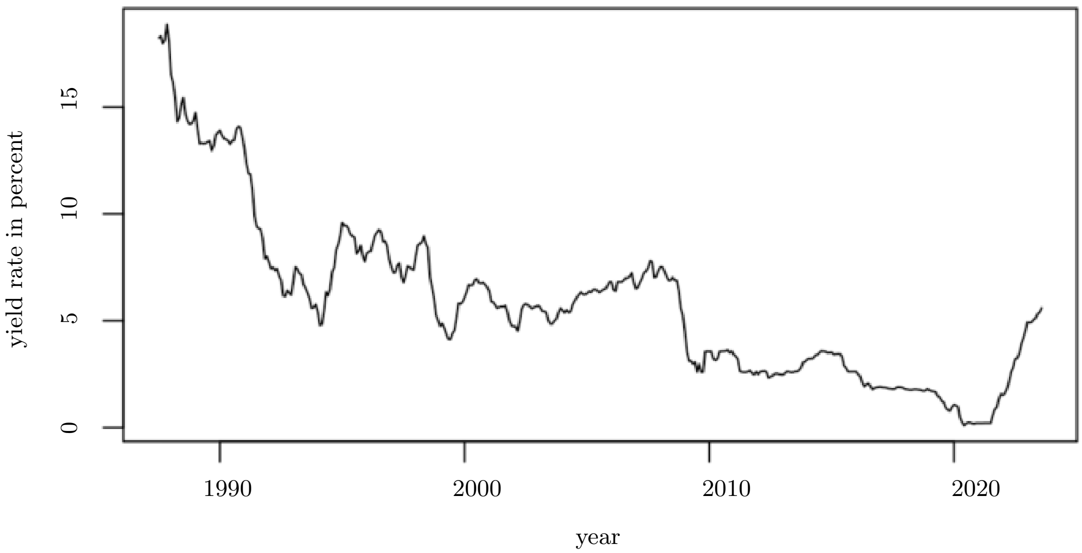

Figure 57: Yields rate for 1 Year Period

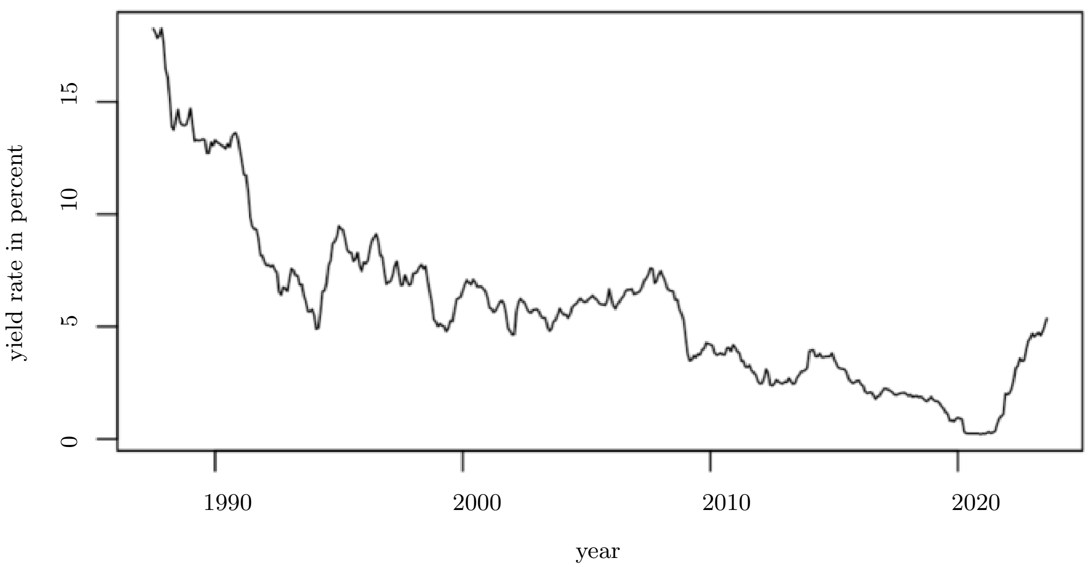

Figure 58: Yields rate for 2 Year Period