# Case Study: Estimating Historical Volatility of SP500 Stock Index

Peter J. Kempthorne

2024-11-11

## Contents

# 1. Measuring Volatility with Daily OHLC Data

In the following sections we introduce alternative measures of volatility based on daily bar price data. The notation for different symbols are given in the following table

| *Symbol* | *Description* |
|---|---|
| $\sigma$ | Volatility |
| $N$ | Number of closing prices in a year |
| $n$ | Number of historical prices used for the volatility estimate |
| $O_i$ | The opening price |
| $H_i$ | The high |
| $L_i$ | The low |
| $C_i$ | The close |

We illustrate the different measures with the S&P 500 Index. First, import prices of S&P500 Index and SP500 Stock using the function tq_get() from the library tidyquant.

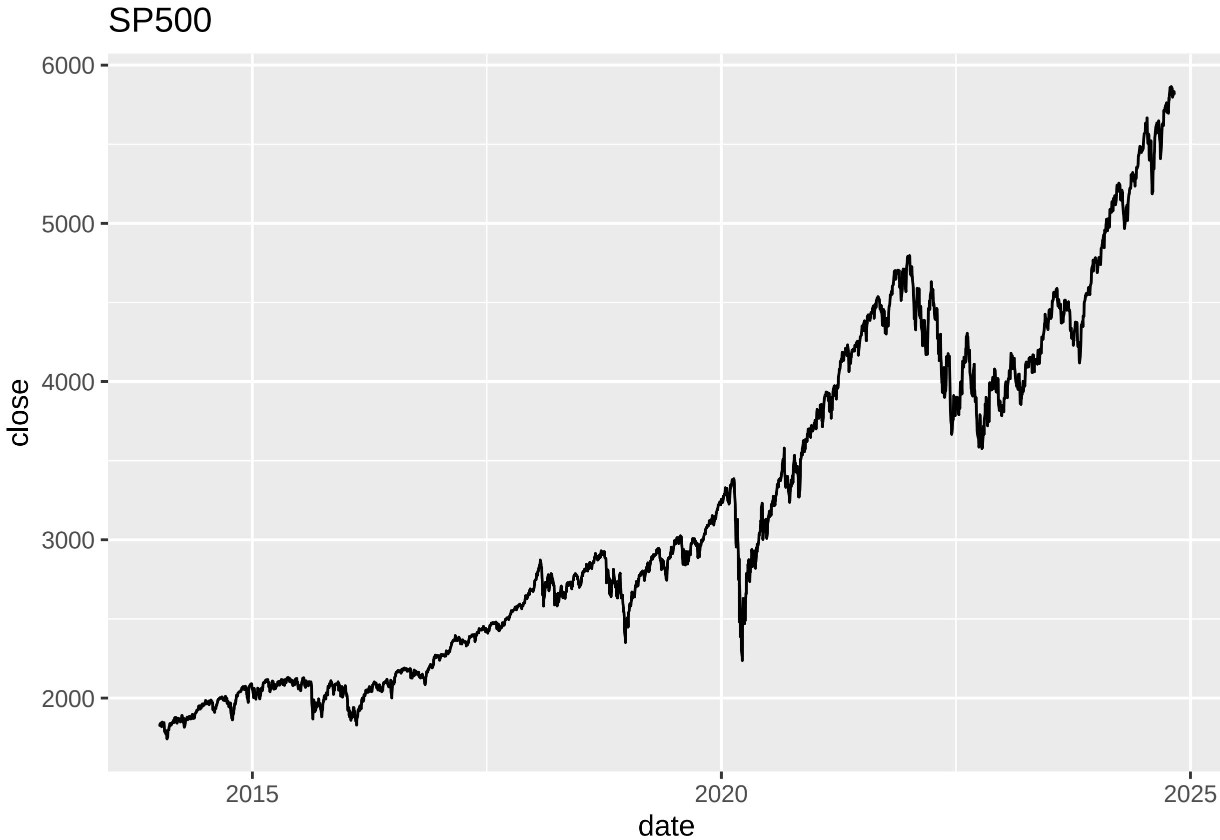
```r
# To make code reproducible, fix the end.date of the price data

SP500<-tq_get("^GSPC", to="2024-10-31")

symbolname0<-"SP500"
```

The object SP500 consists of daily bar data, collected from finance.yahoo.com

```
## [1] 2726    8
```

```
## # A tibble: 6 x 8
##   symbol date        open  high   low close     volume adjusted
##   <chr>  <date>     <dbl> <dbl> <dbl> <dbl>      <dbl>    <dbl>
## 1 ^GSPC  2014-01-02 1846. 1846. 1828. 1832. 3080600000    1832.
## 2 ^GSPC  2014-01-03 1833. 1838. 1829. 1831. 2774270000    1831.
## 3 ^GSPC  2014-01-06 1832. 1837. 1824. 1827. 3294850000    1827.
## 4 ^GSPC  2014-01-07 1829. 1840. 1829. 1838. 3511750000    1838.
## 5 ^GSPC  2014-01-08 1838. 1840. 1831. 1837. 3652140000    1837.
## 6 ^GSPC  2014-01-09 1839  1843. 1830. 1838. 3581150000    1838.
```

```
## # A tibble: 6 x 8
##   symbol date        open  high   low close     volume adjusted
##   <chr>  <date>     <dbl> <dbl> <dbl> <dbl>      <dbl>    <dbl>
## 1 ^GSPC  2024-10-23 5834. 5835. 5762. 5797. 3532650000    5797.
## 2 ^GSPC  2024-10-24 5818. 5818. 5785. 5810. 3543030000    5810.
## 3 ^GSPC  2024-10-25 5827. 5863. 5800. 5808. 3501280000    5808.
## 4 ^GSPC  2024-10-28 5834. 5843. 5823. 5824. 3691280000    5824.
## 5 ^GSPC  2024-10-29 5820. 5847. 5802. 5833. 3879100000    5833.
## 6 ^GSPC  2024-10-30 5833. 5851. 5811. 5814. 3851120000    5814.
```




The following sections provide the definitions of different measures of historical volatility measures.

We set $n = 21$, the number of periods for the volatility estimate which corresponds to the typical number of business days in a month.

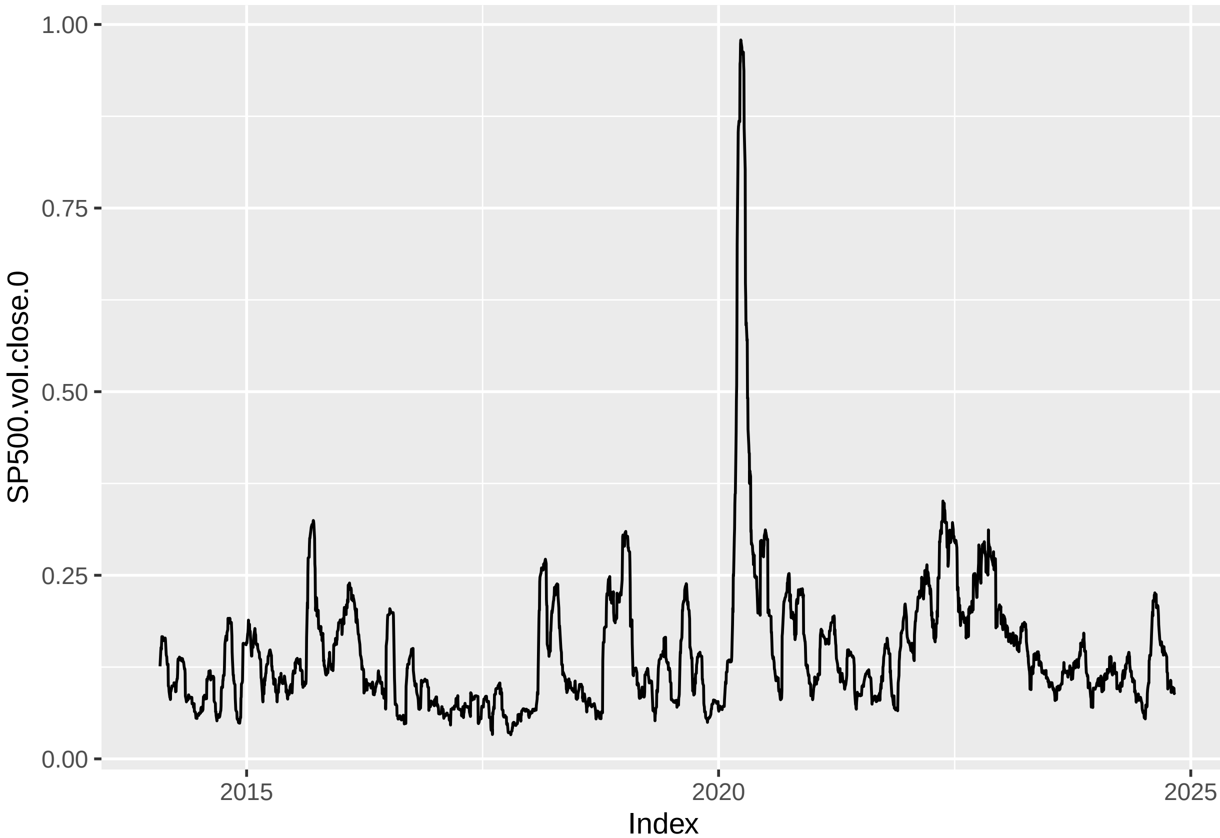
### 1.1 Close-to-Close Volatility

Historical volatility calculation using close-to-close prices.

$$
\begin{aligned}
r_i &= \ln\left(\frac{C_{i+1}}{C_i}\right) \\
\bar{r} &= \frac{r_1 + r_2 + \ldots r_{n-1}}{n-1} \\
\sigma &= \sqrt{\frac{N}{n-2}\sum_{i=1}^{n-1}(r_i - \bar{r})^2}
\end{aligned}
$$

Note that the historical volatility is computed from a time series of $n$ closing prices. The annualized volatility is computed using an unbiased estimate of the (n-1) period variance of the close-to-close returns.

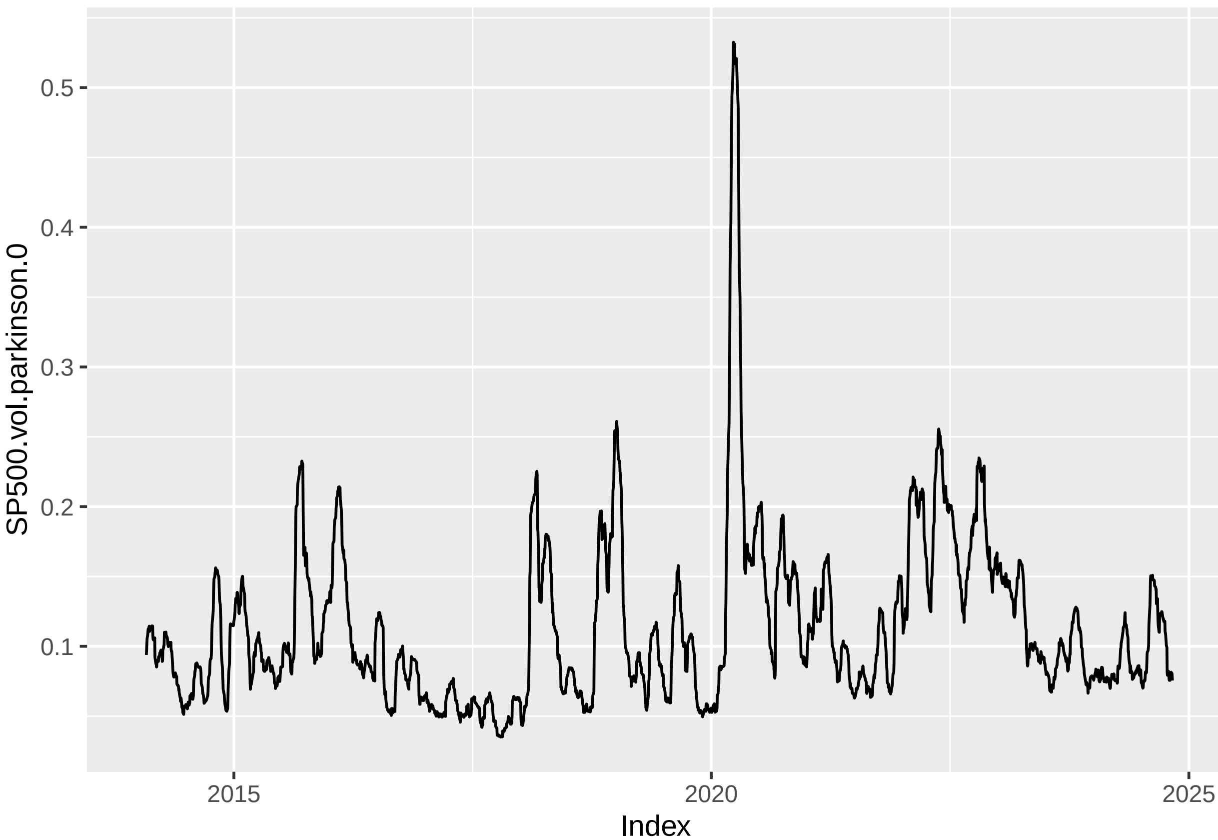
The 21-day variance is scaled up to an N=252 day horizon corresponding to typical annual period of days.



## 1.2 Parkinson High-Low Volatility

The Parkinson formula for estimating the historical volatility of an underlying is based on only the daily high and low prices.

$\sigma = \sqrt{\frac{N}{n4\ln 2}\sum_{i=1}^{n}\left(\ln\frac{H_i}{L_i}\right)^2}$

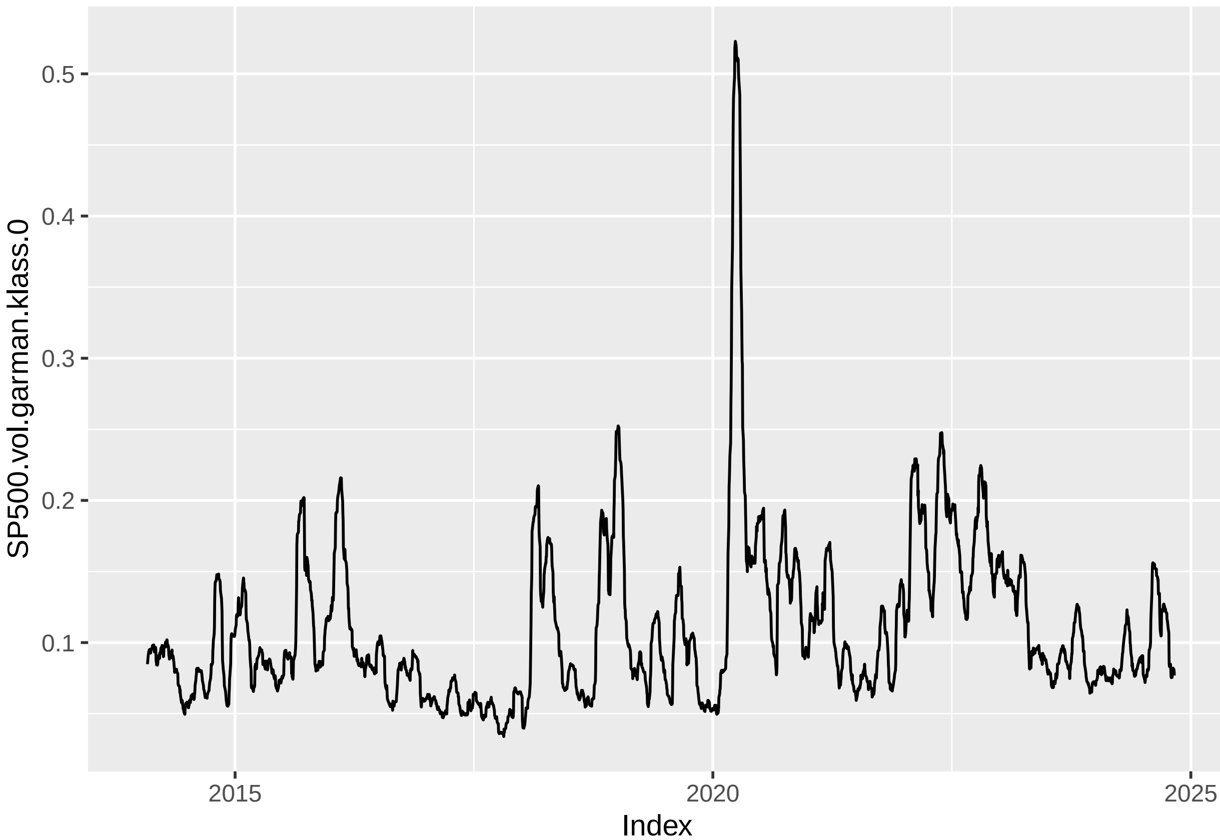



Note that the Parkinson volatility component for a day will be positive if the daily range (High minus Low) is positive, even if the Close-to-Close return is zero.

### 1.3 Garman-Klass OHLC Volatility

The Garman and Klass estimator for estimating historical volatility assumes Brownian motion with zero drift and no opening jumps (i.e. the opening price equals the close of the previous period). This estimator is 7.4 times more efficient than the close-to-close estimator.

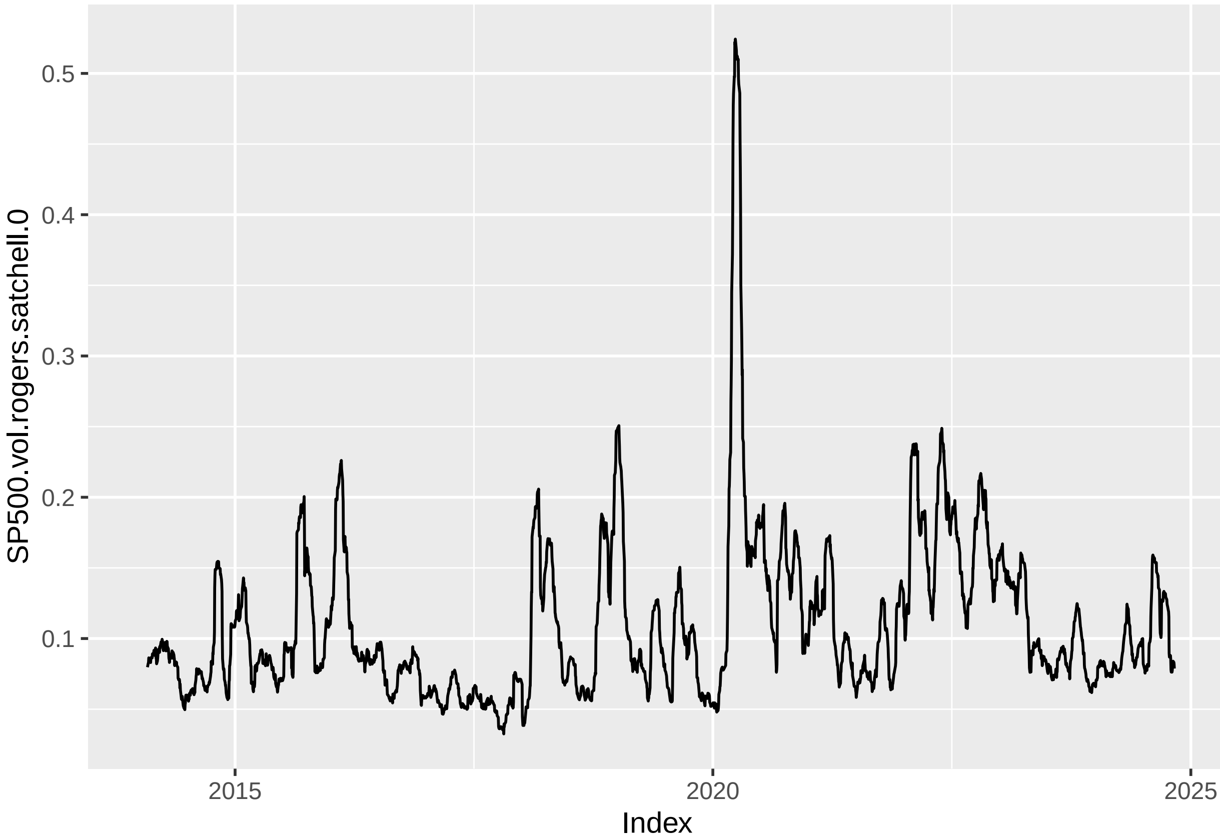
$$\sigma = \sqrt{\tfrac{N}{n} \sum_{i=1}^{n} \left[ \tfrac{1}{2} \left( \ln \frac{H_i}{L_i} \right)^2 - (2 \ln 2 - 1) \left( \ln \frac{C_i}{O_i} \right)^2 \right]}$$



## 1.4 Rogers and Satchel OHLC Volatility

$\sigma = \sqrt{\frac{N}{n} \sum_{i=1}^{n} \left[ (\ln \frac{H_i}{C_i})(\ln \frac{H_i}{O_i}) + (\ln \frac{L_i}{C_i})(\ln \frac{L_i}{O_i}) \right]}$

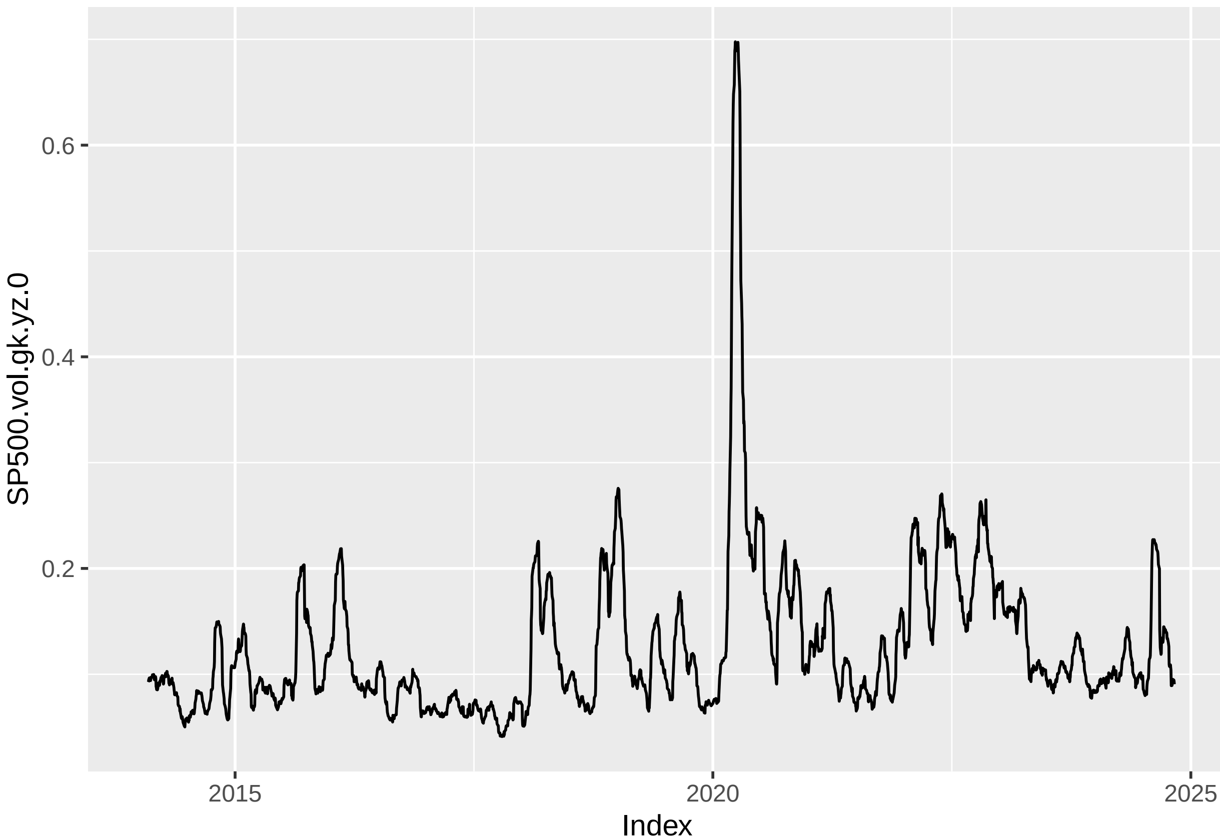
The Roger and Satchell historical volatility estimator allows for non-zero drift, but assumed no opening jump.



### 1.5 Garman-Klass Yang and Zhang Historical Open-High-Low-Close Volatility

Yang and Zhang derived an extension to the Garman-Klass historical volatility estimator that allows for opening jumps. It assumes Brownian motion with zero drift. This is currently the preferred version of open-high-low-close volatility estimator for zero drift and has an efficiency of 8 times the classic close-to-close estimator. Note that when the drift is nonzero, but instead relative large to the volatility, this estimator will tend to overestimate the volatility.

$$\sigma = \sqrt{\frac{N}{n} \sum \left[ \left( \ln \frac{O_i}{C_{i-1}} \right)^2 + \frac{1}{2} \left( \ln \frac{H_i}{L_i} \right)^2 - (2 \ln 2 - 1) \left( \ln \frac{C_i}{O_i} \right)^2 \right]}$$

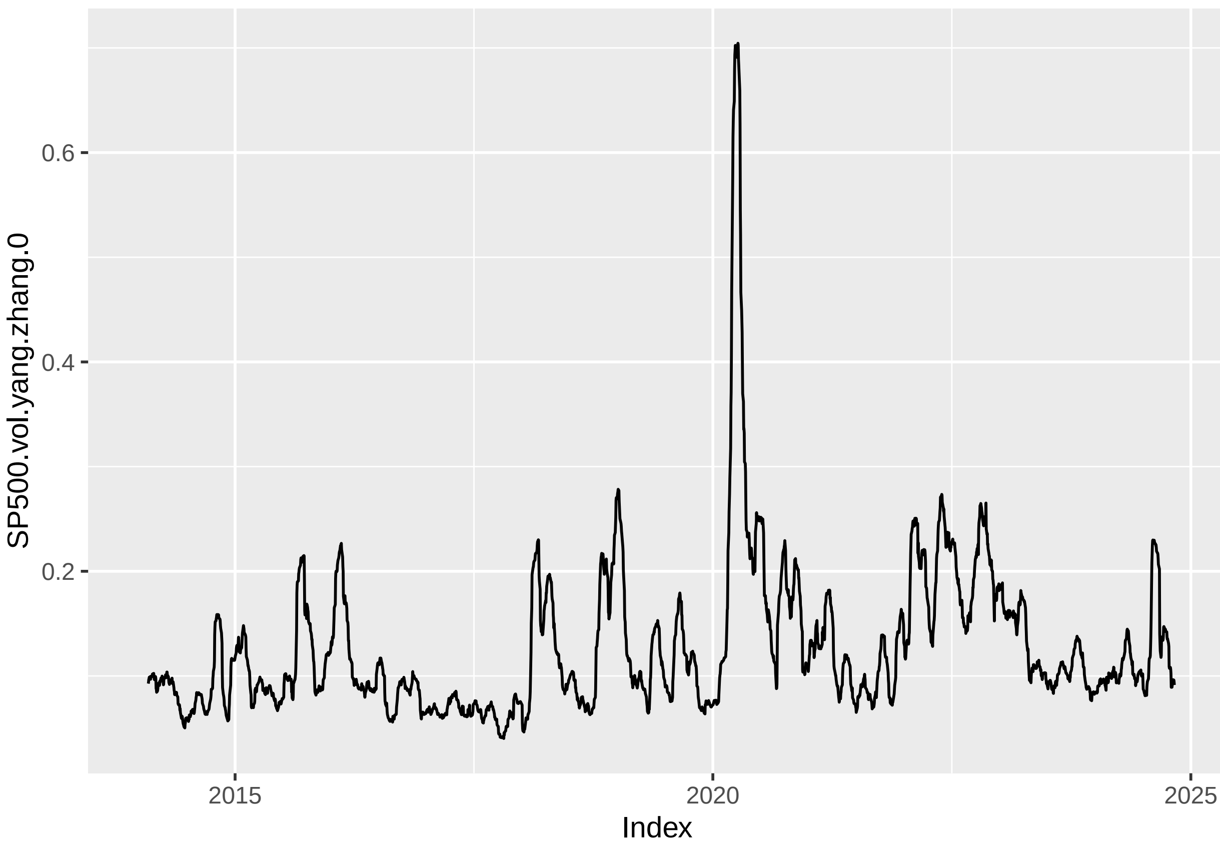


## 1.6 Yang and Zhang Volatility Estimator

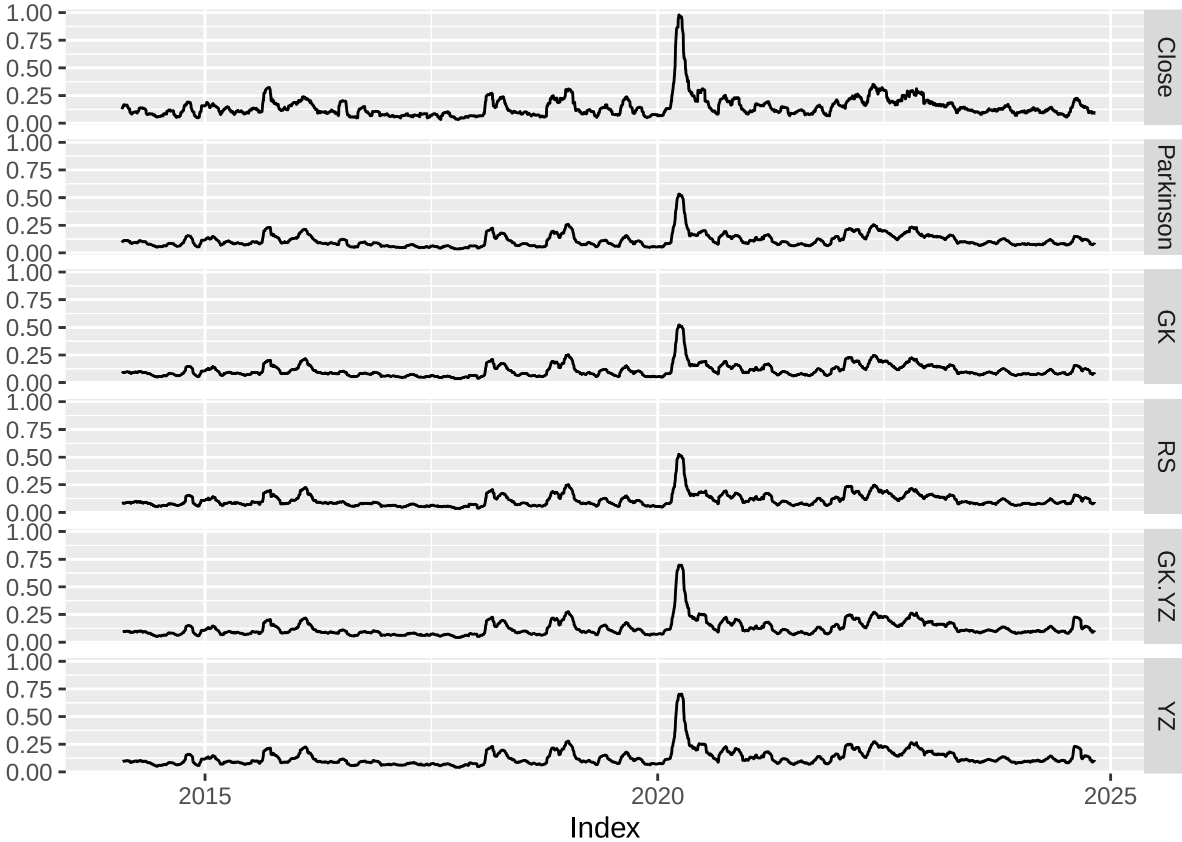
The Yang and Zhang historical volatility estimator has minimum estimation error, and is independent of drift and opening gaps. It can be interpreted as a weighted average of the Rogers and Satchell estimator, the close-open volatility, and the open-close volatility.

When using the volatility() function of the R package TTR (in tidyquant), users may override the default values of  (1.34 by default) or $k$ used in the calculation by specifying alpha or $k$ in the following expressions, respectively. Specifying $k$ will cause alpha to be ignored, if both are provided.

$$\begin{aligned}
s &= sqrt(s2o + k * s2c + (1-k) * (s2rs^2)) \\
s2o &= N * runVar(log(Op/lag(Cl,1)), n = n) \\
s2c &= N * runVar(log(Cl/Op), n = n) \\
s2rs &= volatility(OHLC, n, "rogers.satchell", N, ...) \\
k &= (alpha - 1)/(alpha + (n+1)/(n-1))
\end{aligned}$$



## 1.7 Panel time series plot of all volatility measures



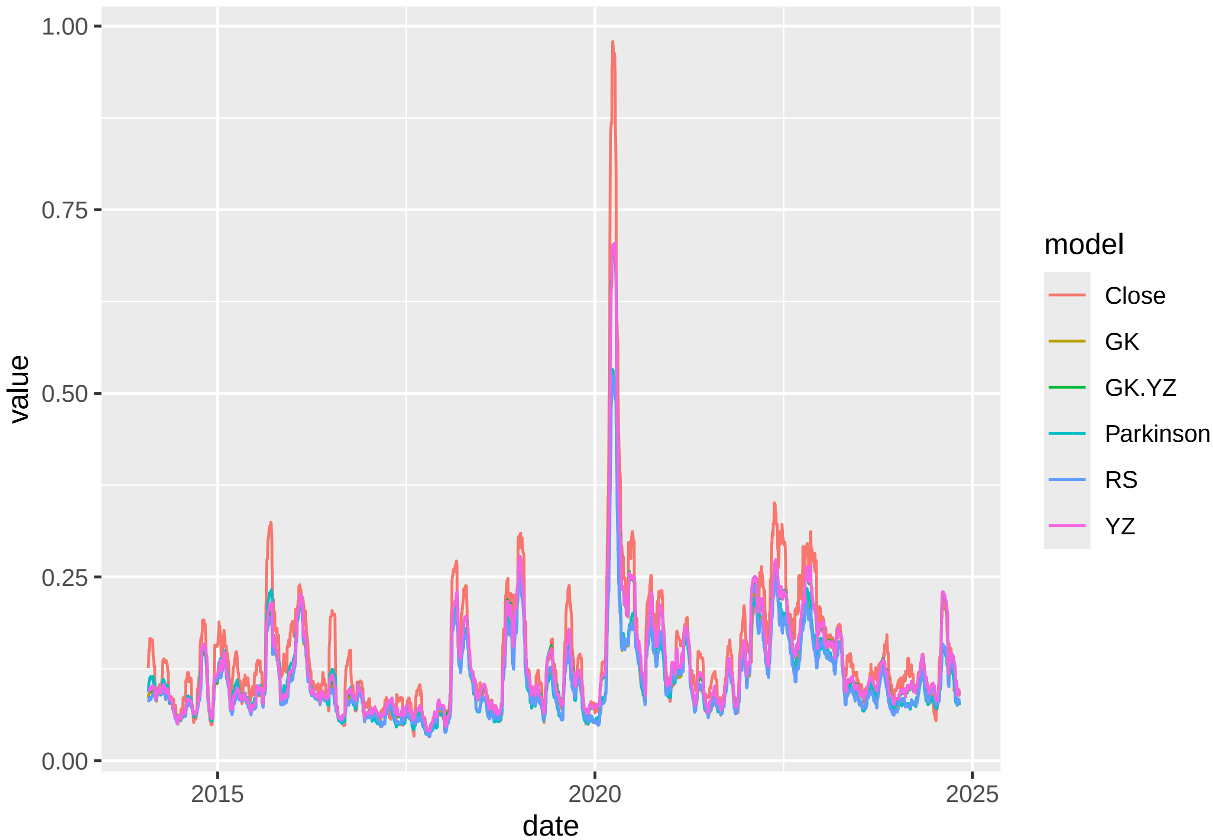

## 1.8 Time series plot of all volatility measures

In first version of this program the following code worked

```
autoplot(SP500.vols, facets =FALSE)
```

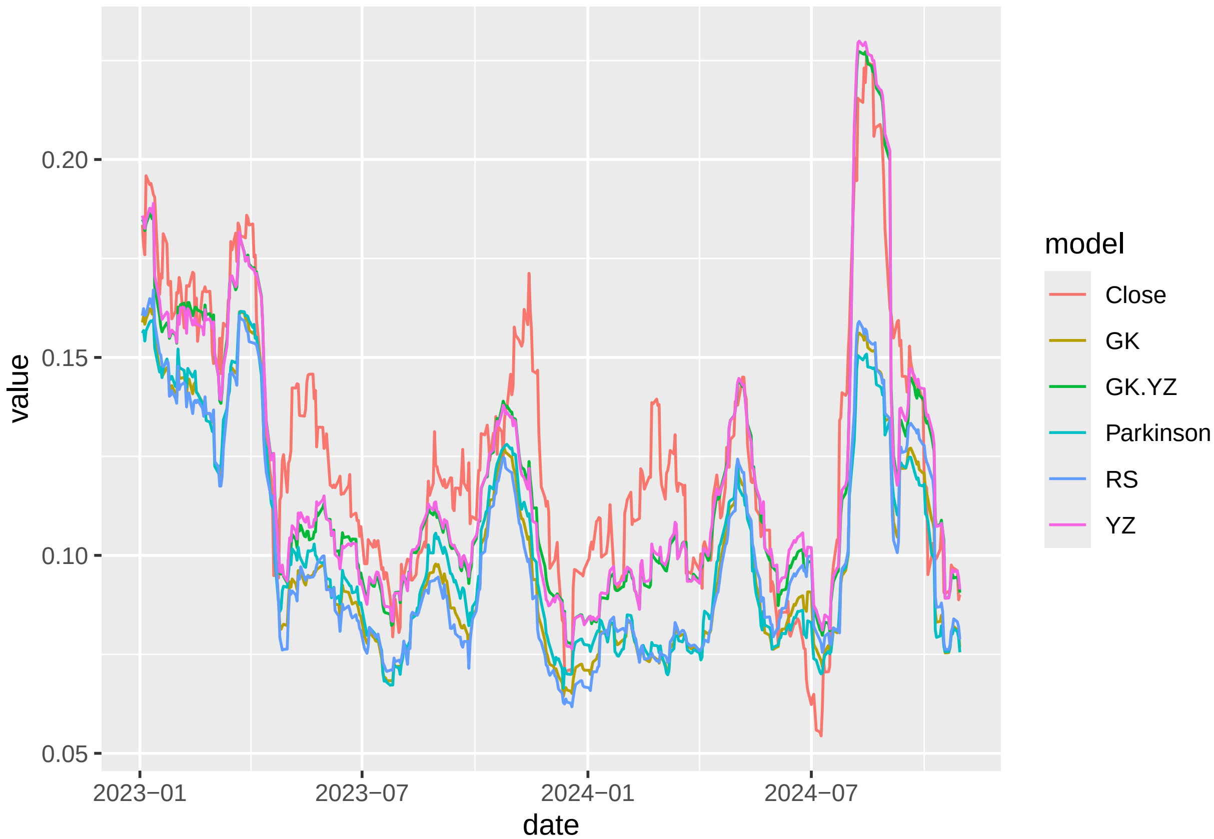
In RStudio cloud it did not. We now construct the figure directly.

```
## [1] "Close"     "Parkinson" "GK"        "RS"        "GK.YZ"     "YZ"
```

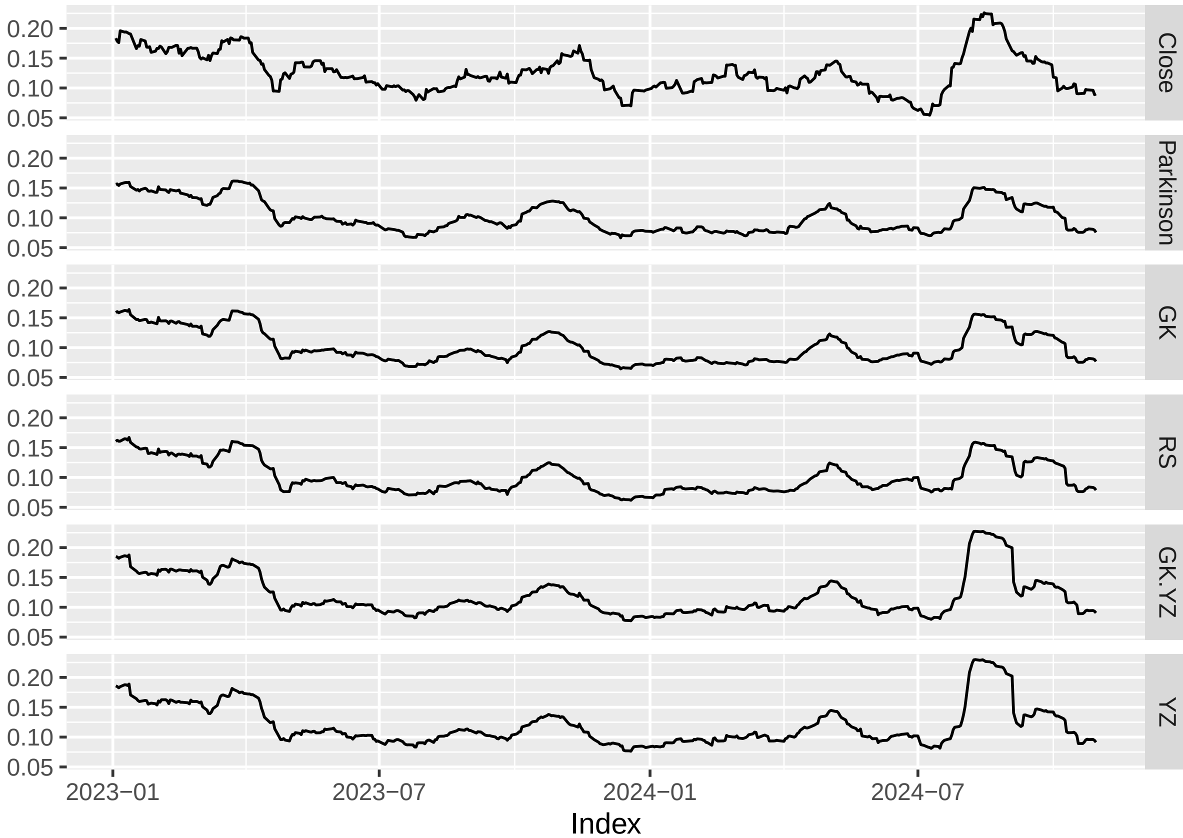


1.00
0.75
0.50
0.25
0.00
value
2015
2020
2025
date
model
Close
GK
GK.YZ
Parkinson
RS
YZ

‘1.9 All volatility measures for 2023-2024

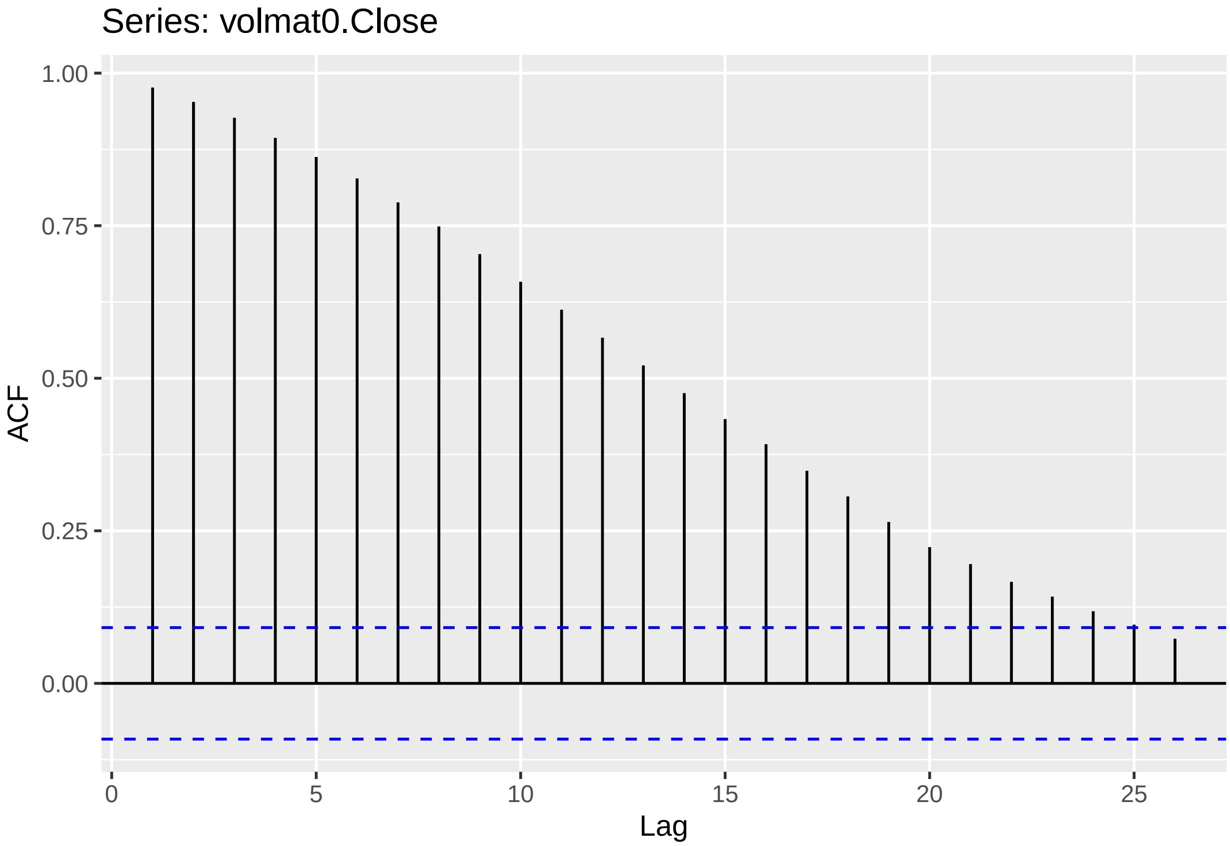


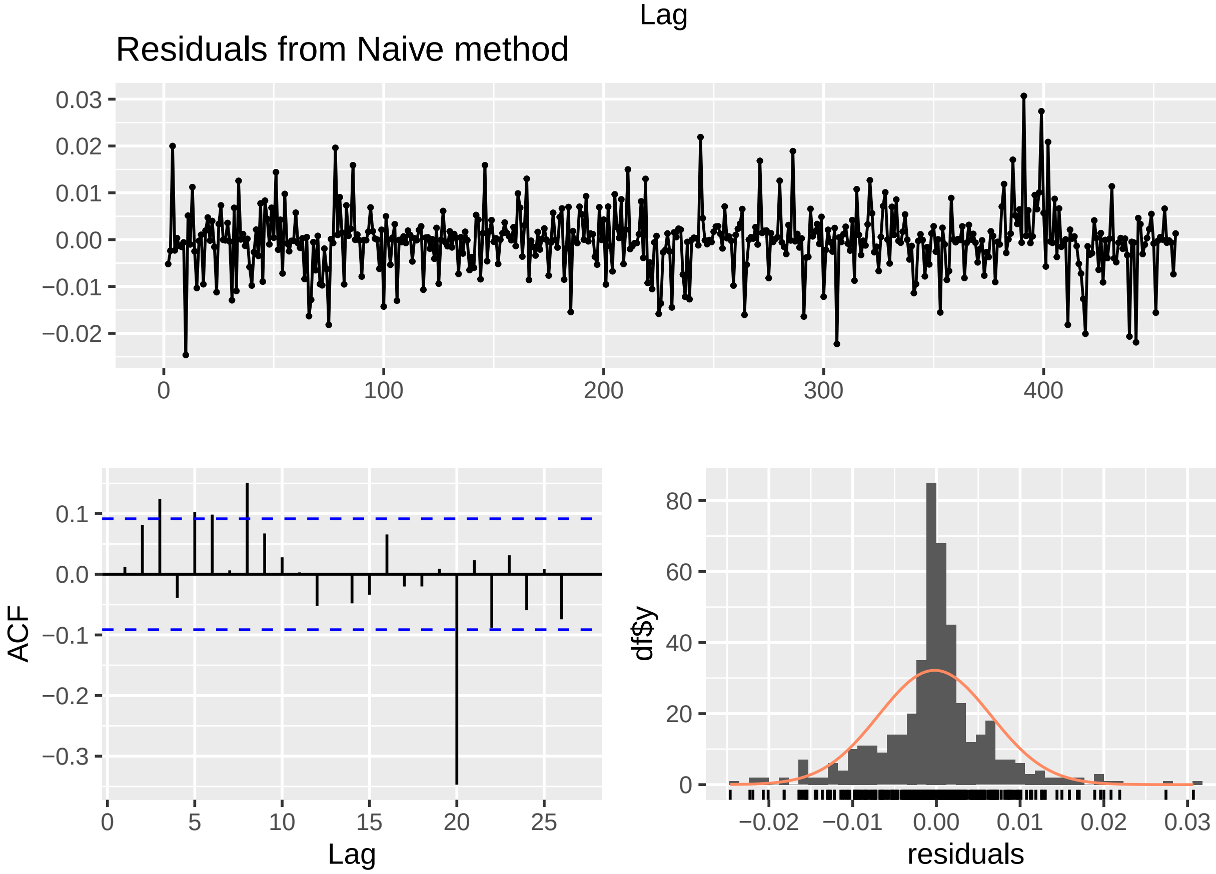
## 2. Modeling 2023-2024 Volatilities



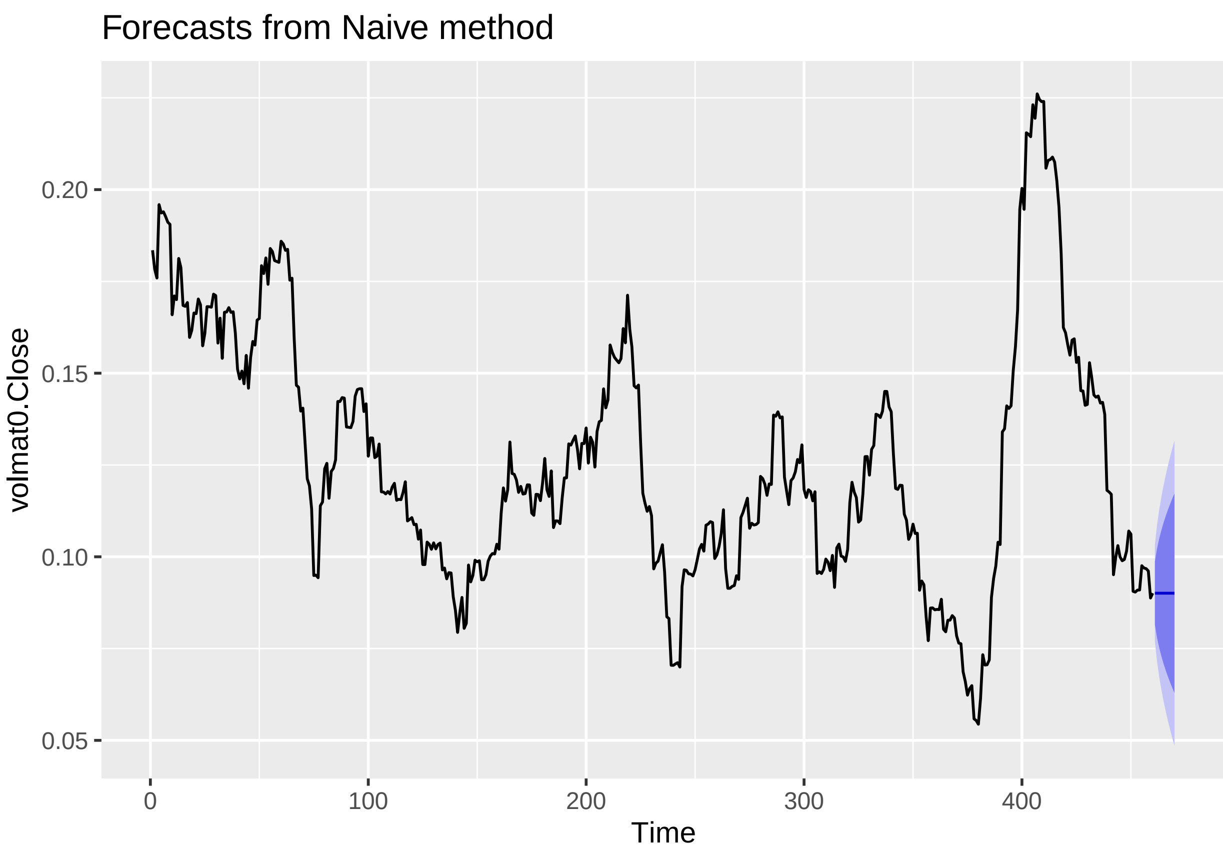
## 2.1 Modeling Close-to-Close Volatility

### 2.1.1 Naive model (Random Walk) of Close-to-Close Volatility



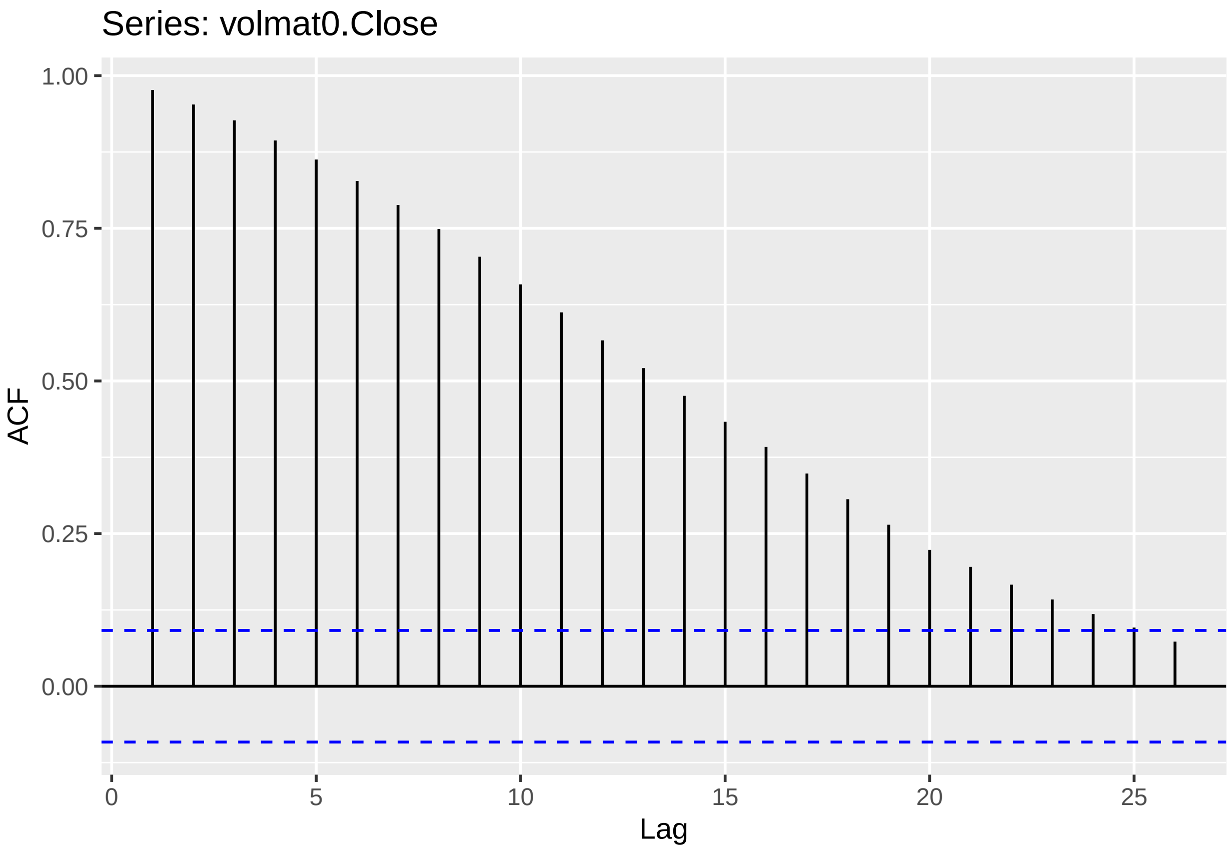



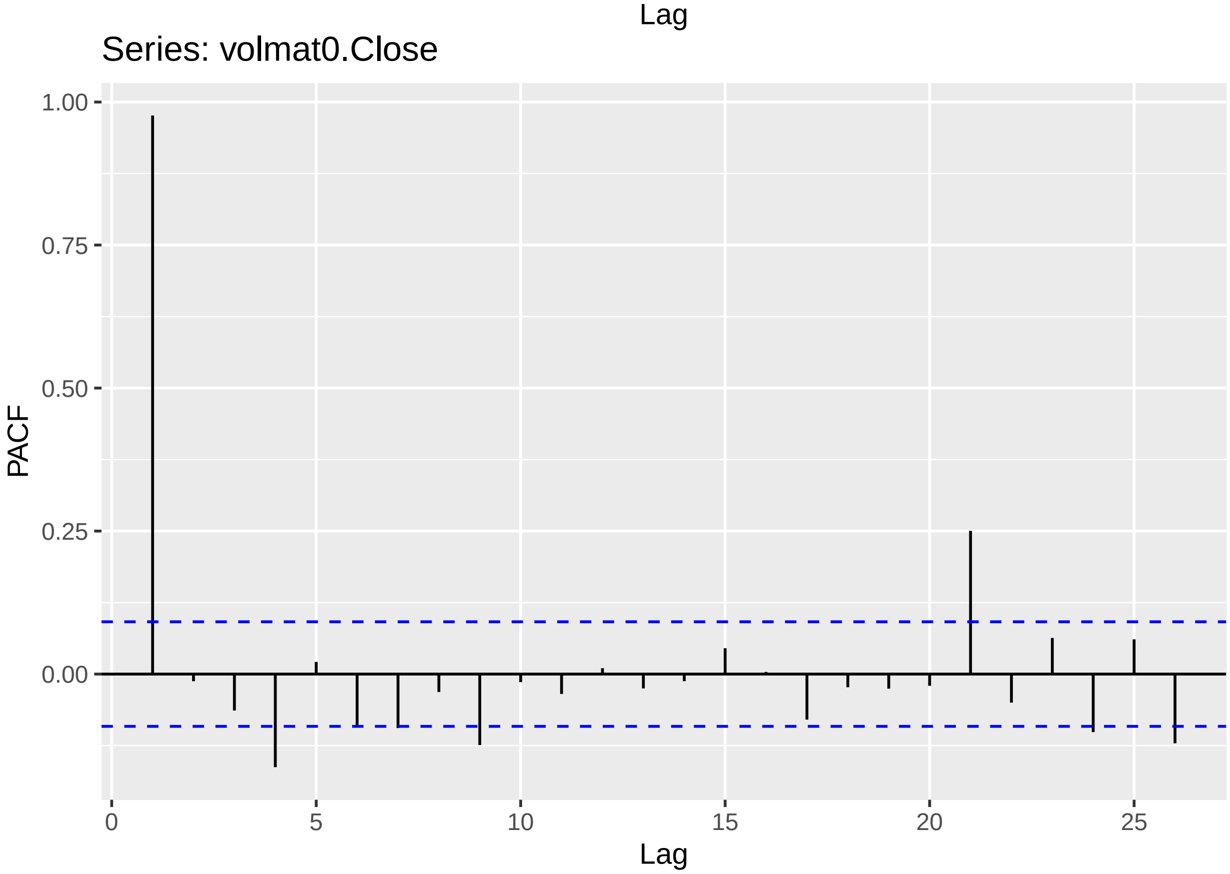
```
##
##  Ljung-Box test
##
## data:  Residuals from Naive method
## Q* = 33.608, df = 10, p-value = 0.0002151
##
## Model df: 0.   Total lags used: 10
```



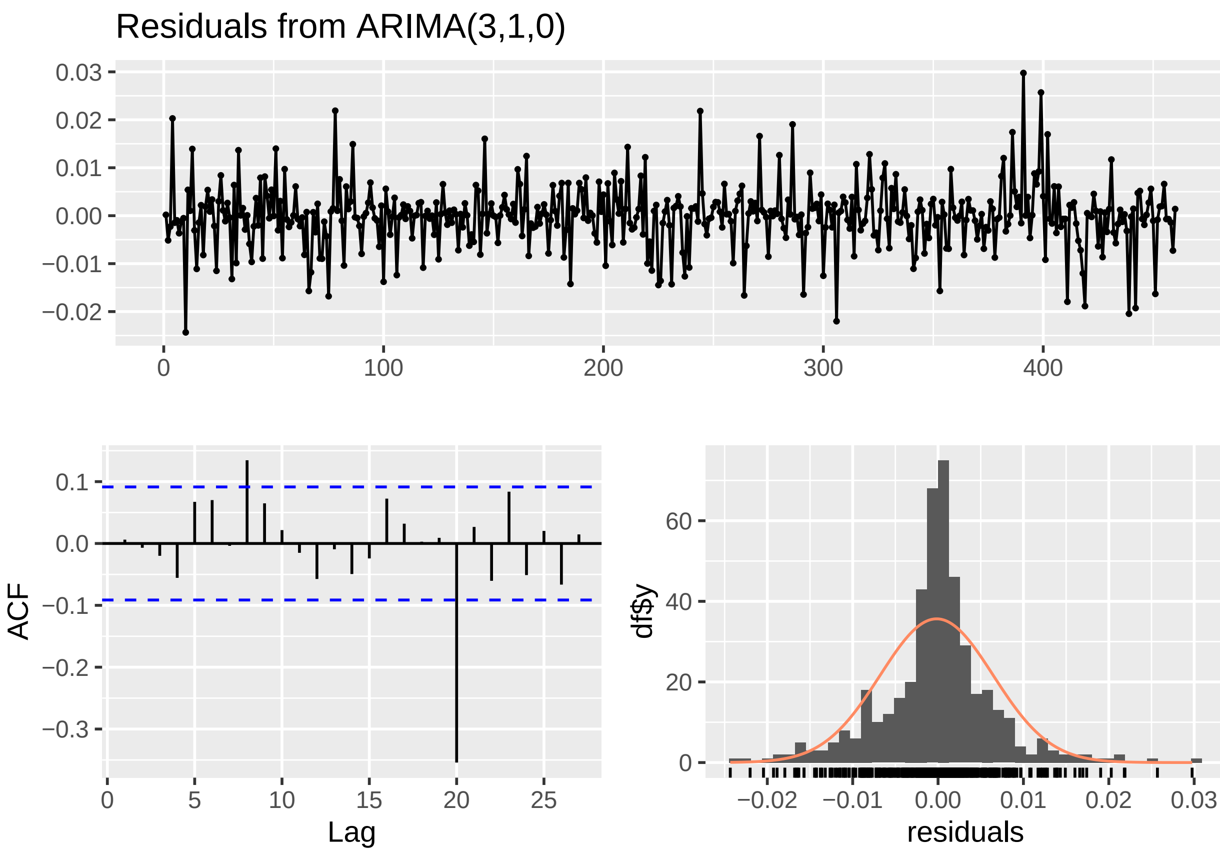

```
##                          ME        RMSE        MAE       MPE     MAPE MASE
## Training set -0.0002035058 0.006707503 0.00437037 -0.314298 3.625842    1
##                   ACF1
## Training set 0.0119396
```

#### 2.1.2 Arima model of Close-to-Close Volatility

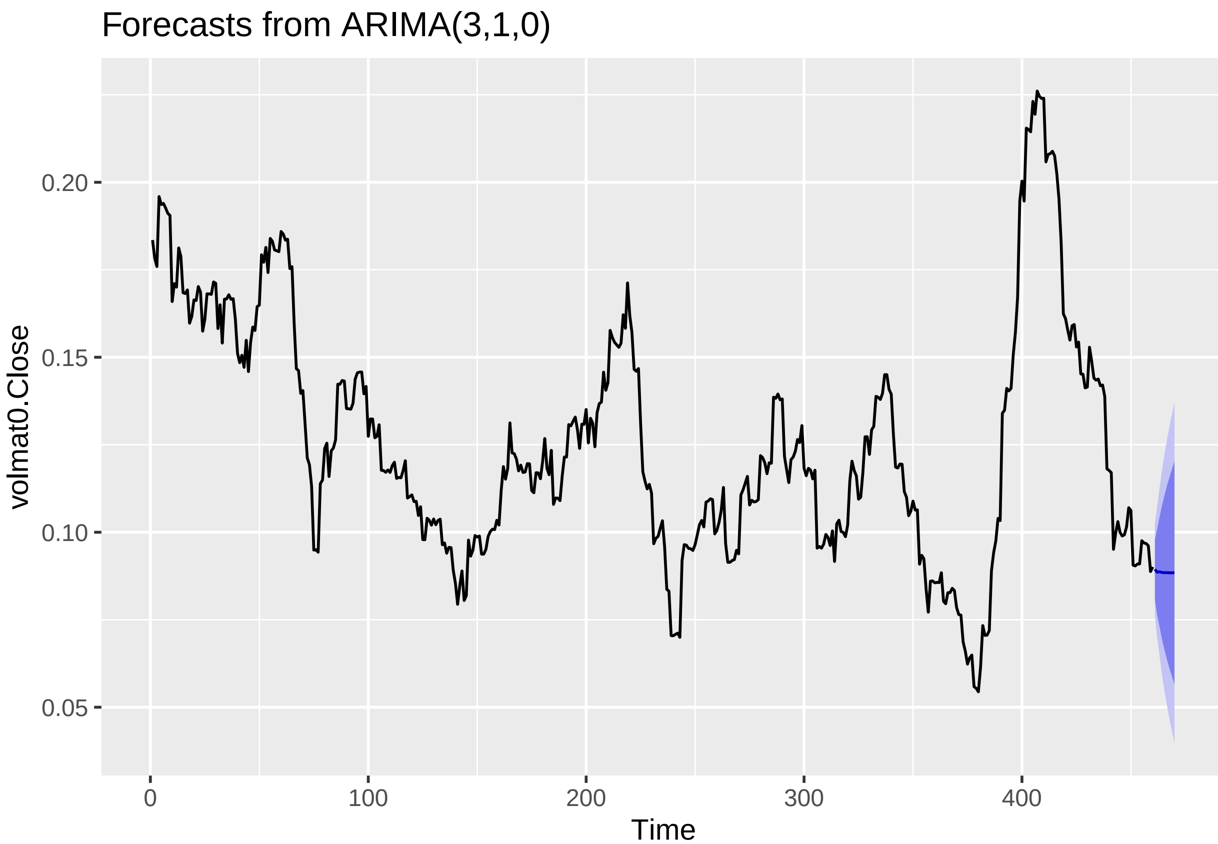


```
## Series: volmat0.Close
## ARIMA(3,1,0)
##
## Coefficients:
##          ar1     ar2     ar3
##       0.0008  0.0799  0.1259
## s.e.  0.0463  0.0461  0.0467
##
## sigma^2 = 4.427e-05:  log likelihood = 1650.95
## AIC=-3293.91   AICc=-3293.82   BIC=-3277.39
```



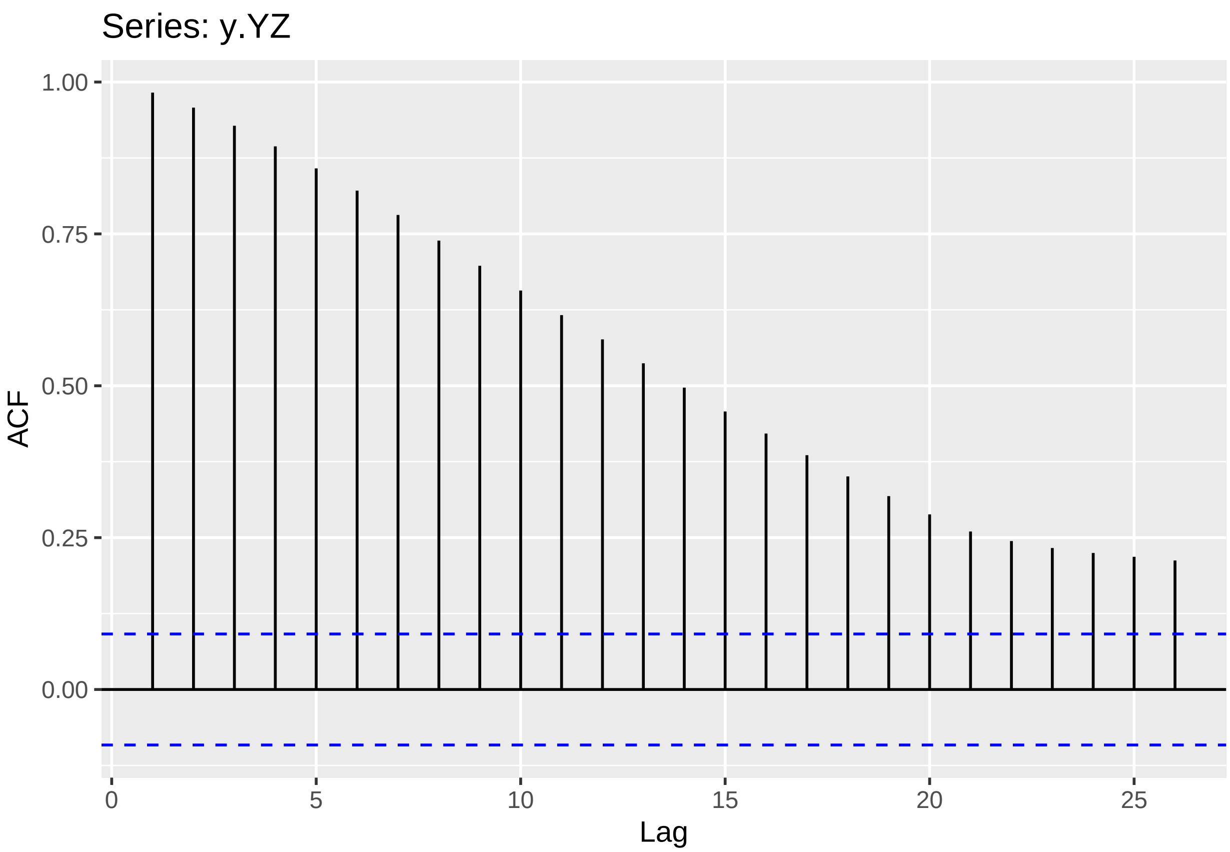

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(3,1,0)
## Q* = 16.795, df = 7, p-value = 0.01876
##
## Model df: 3.   Total lags used: 10
```

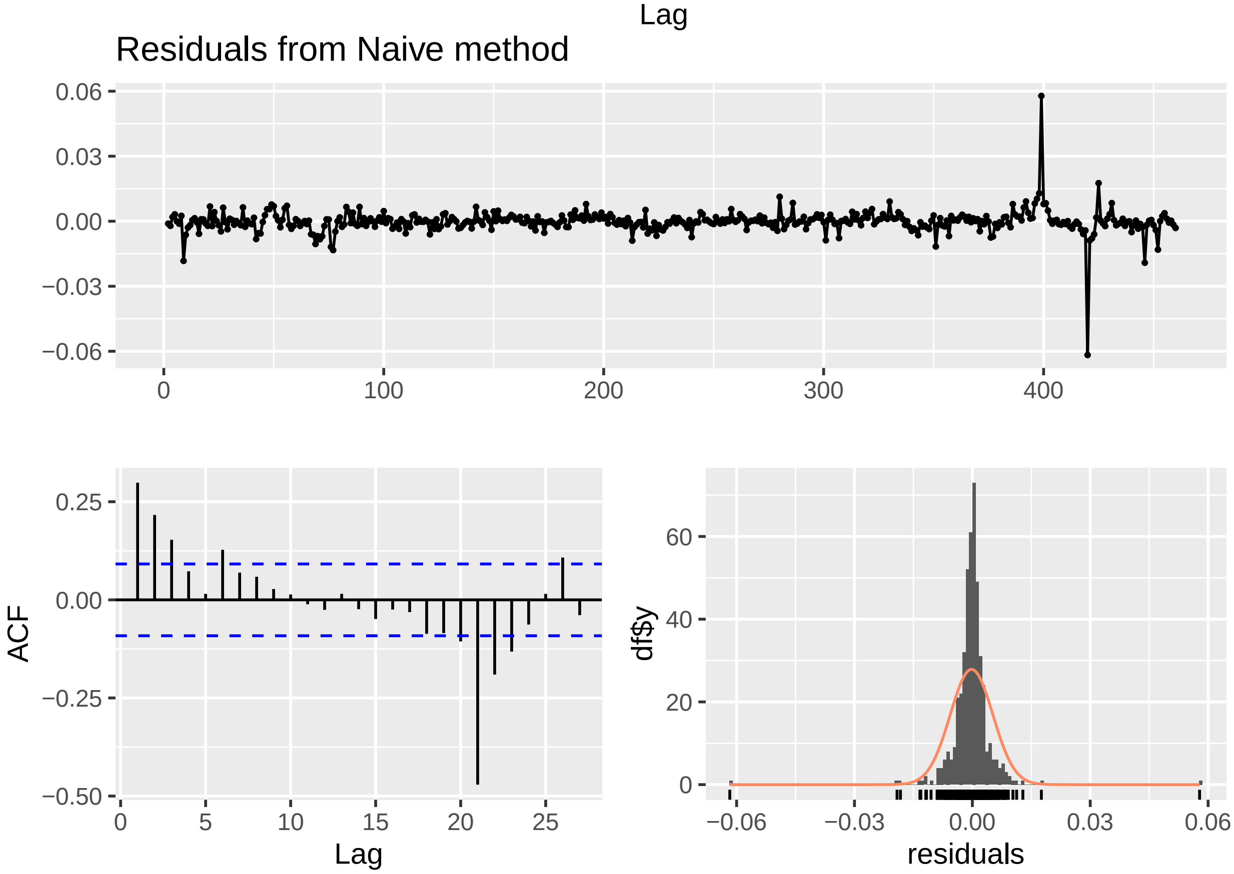



```
##                          ME        RMSE         MAE        MPE    MAPE     MASE
## Training set -0.0001635794 0.006624756 0.004461834 -0.2432893 3.71425 1.020928
##                     ACF1
## Training set 0.006175674
```

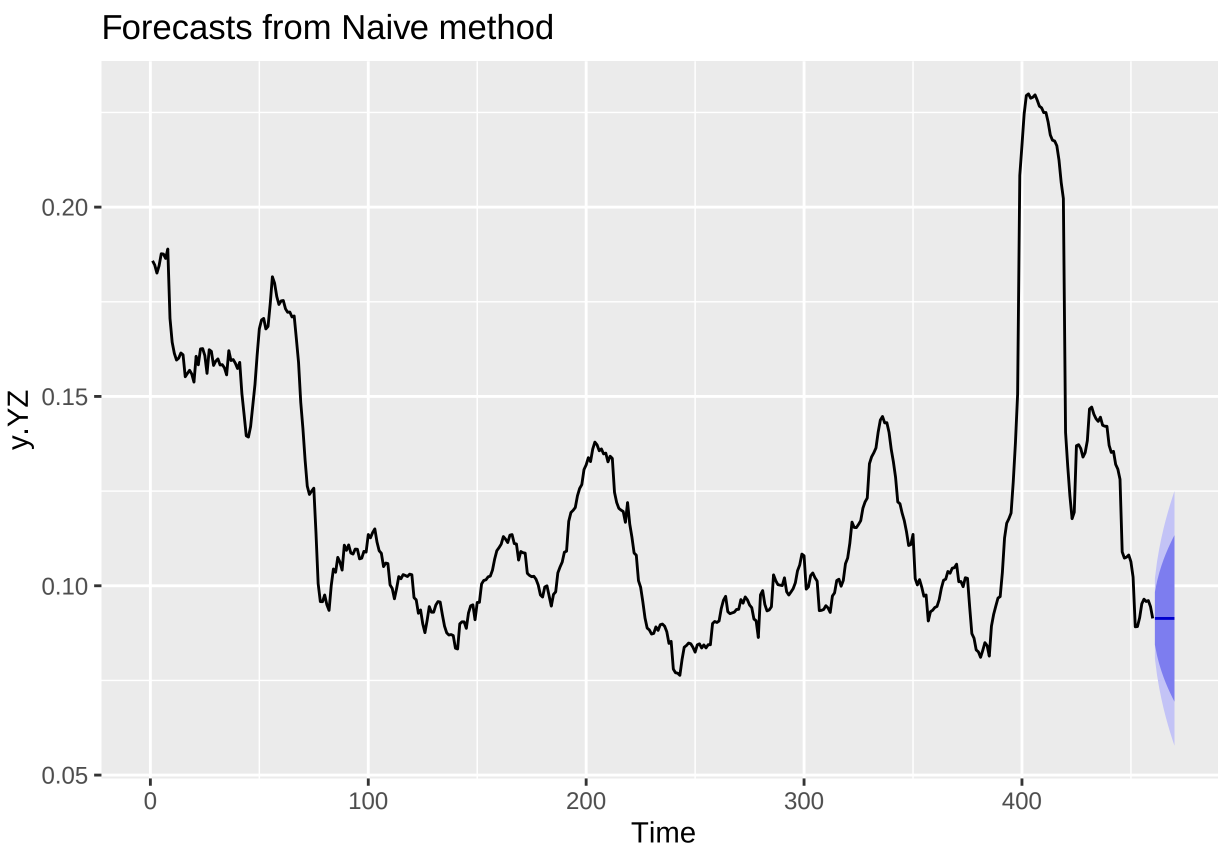
## 2.2 Modeling Yang Zhang OHLC Volatility

### 2.2.1 Naive model (Random Walk) of Yang Zhang OHLC Volatility



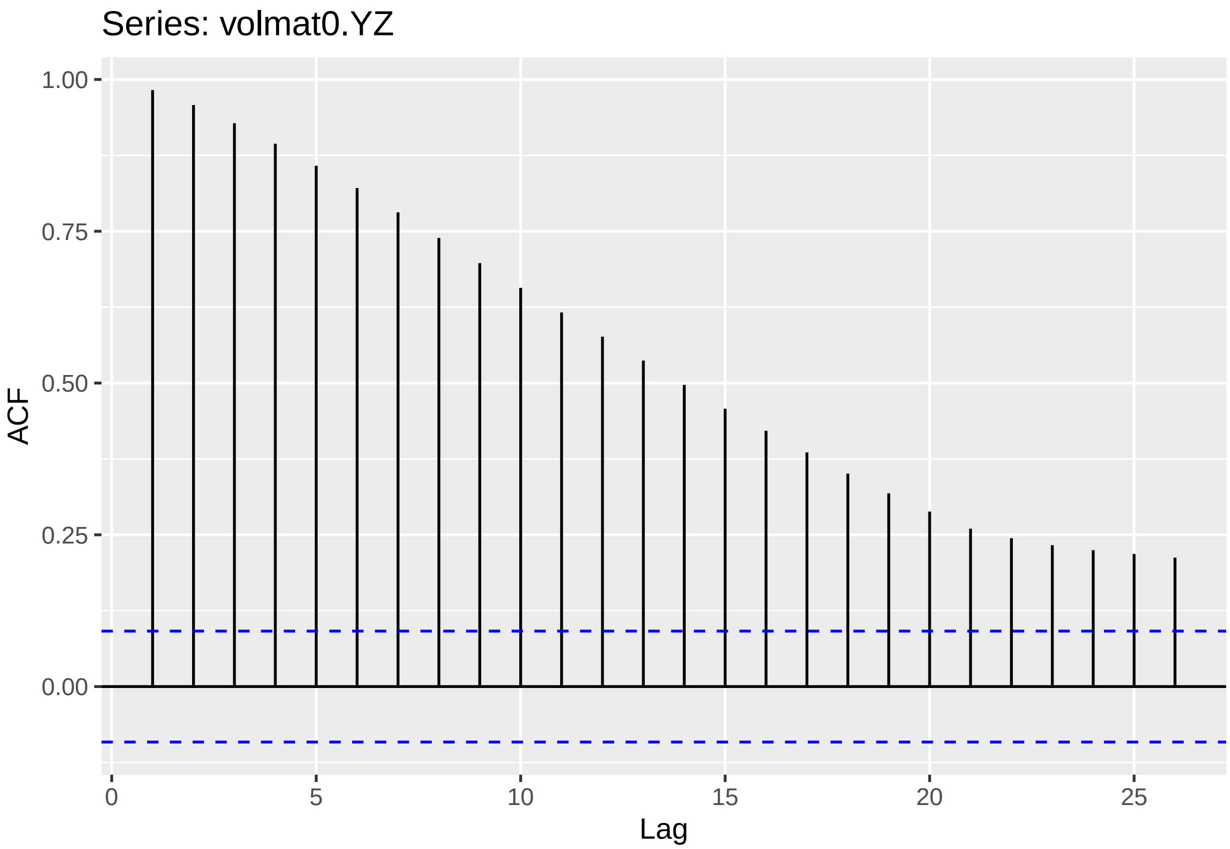



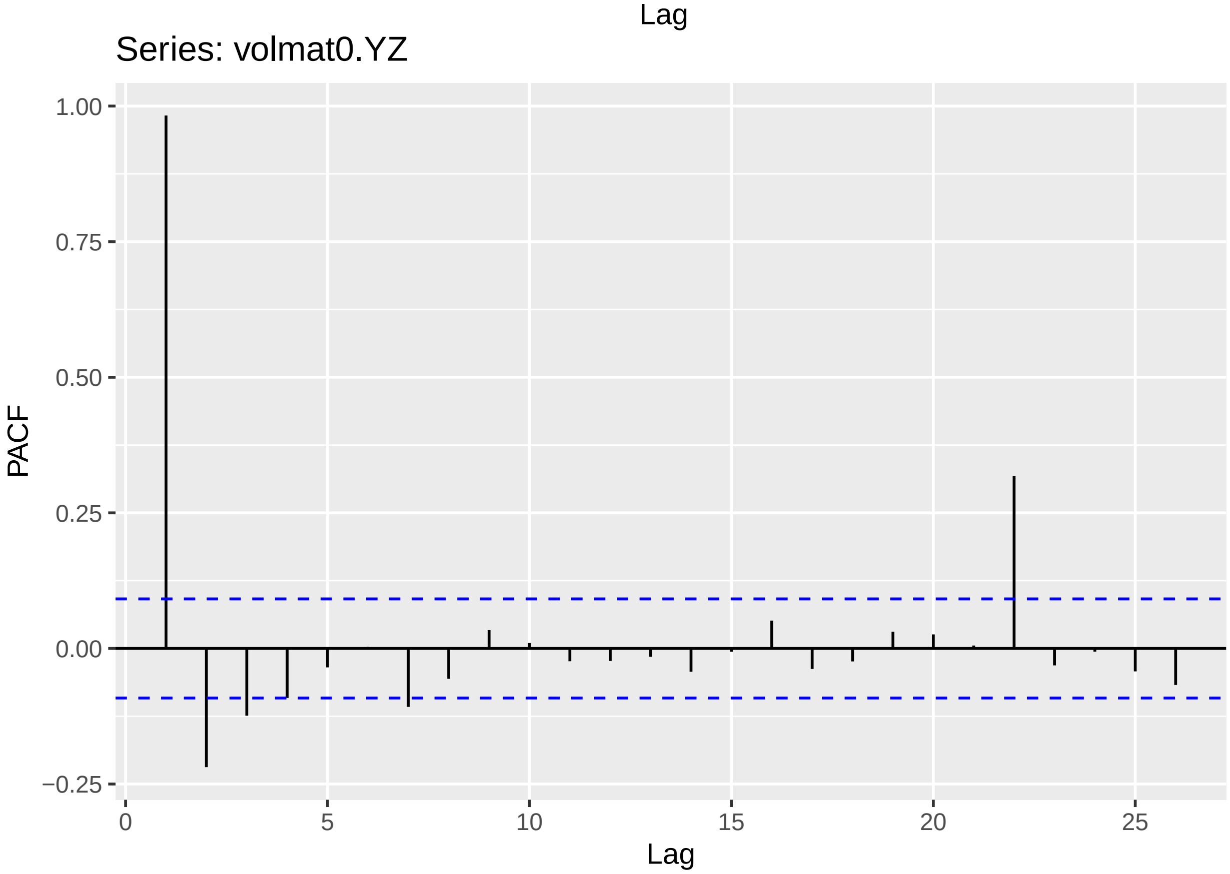
```
##
##  Ljung-Box test
##
## data:  Residuals from Naive method
## Q* = 88.241, df = 10, p-value = 1.199e-14
##
## Model df: 0.   Total lags used: 10
```



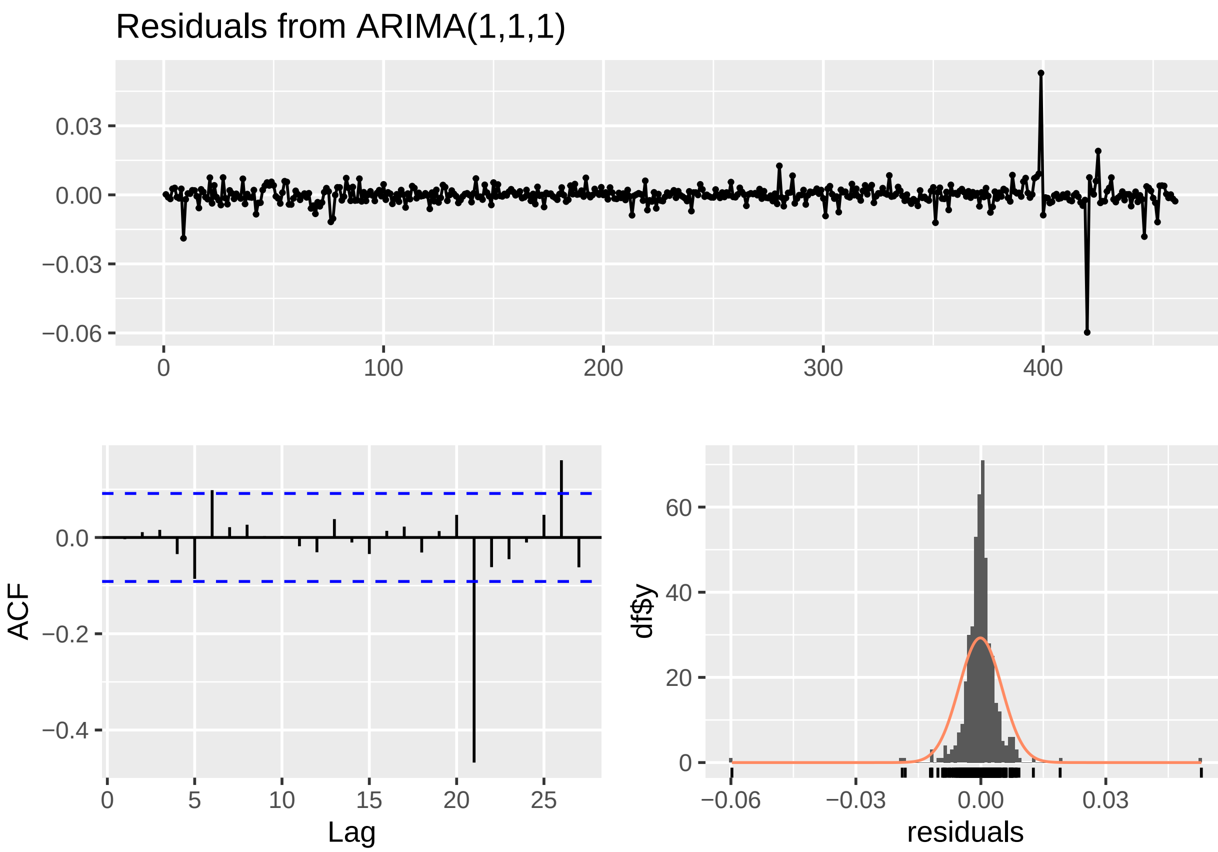

```
##                          ME        RMSE         MAE        MPE     MAPE MASE
## Training set -0.0002057623 0.005429542 0.002755094 -0.2322268 2.307454    1
##                   ACF1
## Training set 0.2985387
```

#### 2.2.2 Arima model of Yang Zhang OHLC Volatility



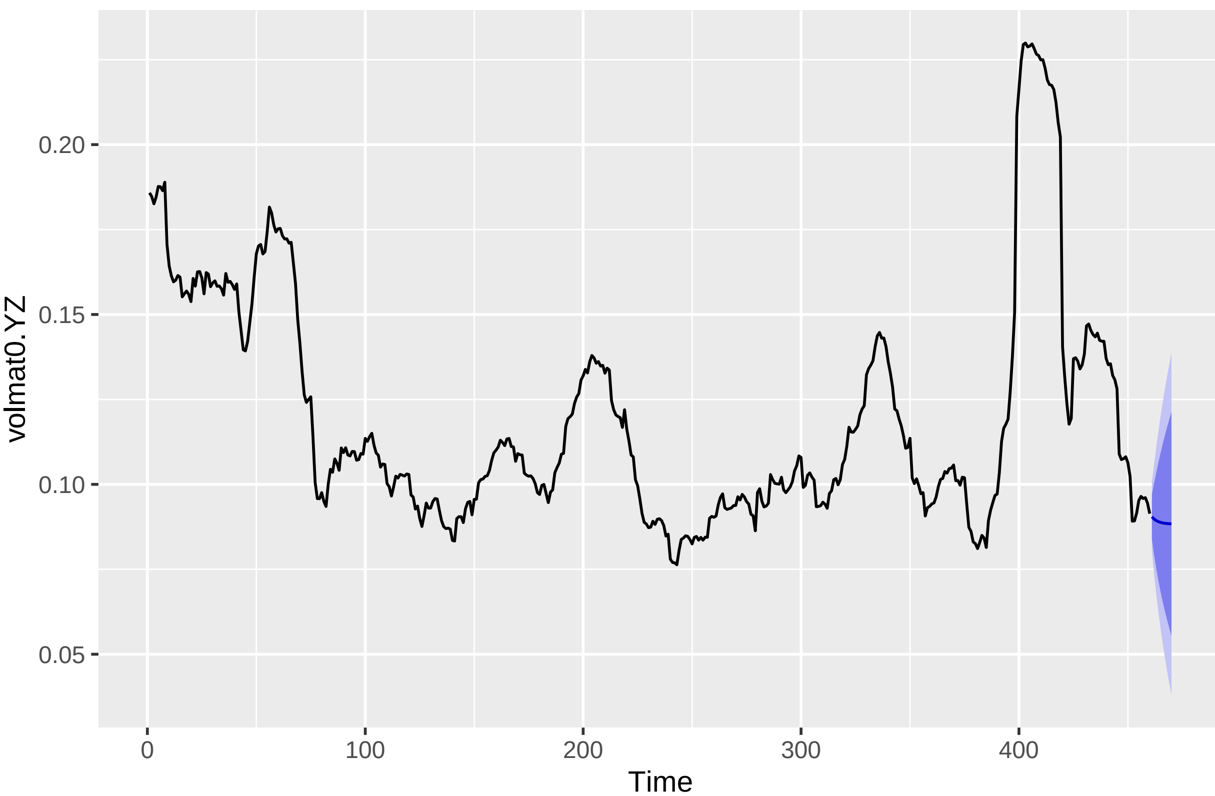



```
## Series: volmat0.YZ
## ARIMA(1,1,1)
##
## Coefficients:
##          ar1      ma1
##       0.6958  -0.4437
## s.e.  0.0926   0.1160
##
## sigma^2 = 2.634e-05:  log likelihood = 1769.57
## AIC=-3533.14   AICc=-3533.09   BIC=-3520.76
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,1,1)
## Q* = 9.2545, df = 8, p-value = 0.3213
##
## Model df: 2.   Total lags used: 10
```

Forecasts from ARIMA(1,1,1)

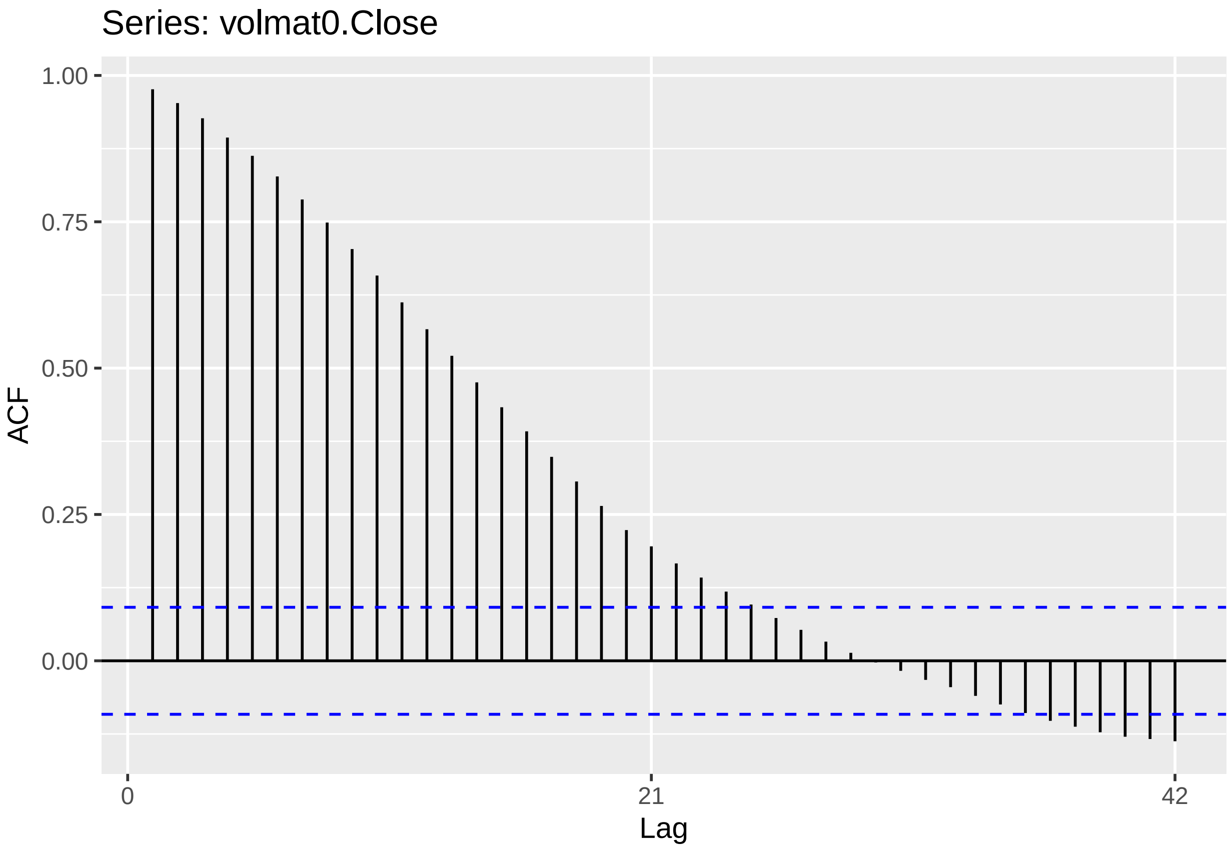


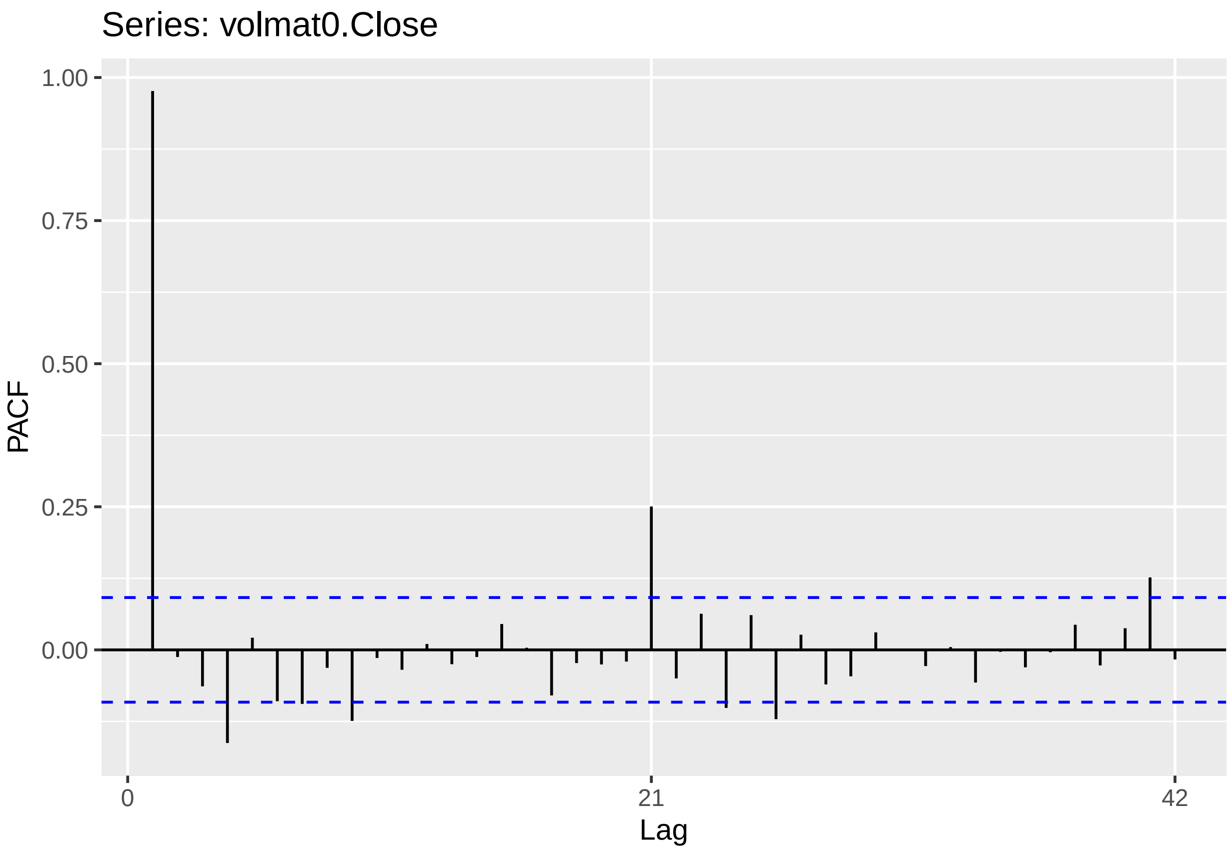

```
##                           ME        RMSE         MAE        MPE     MAPE
## Training set -0.0001150971 0.005115615 0.002625843 -0.1021231 2.220704
##                   MASE         ACF1
## Training set 0.9530867 -0.003515841
```

# 3. Fitting Seasonal Arima Models

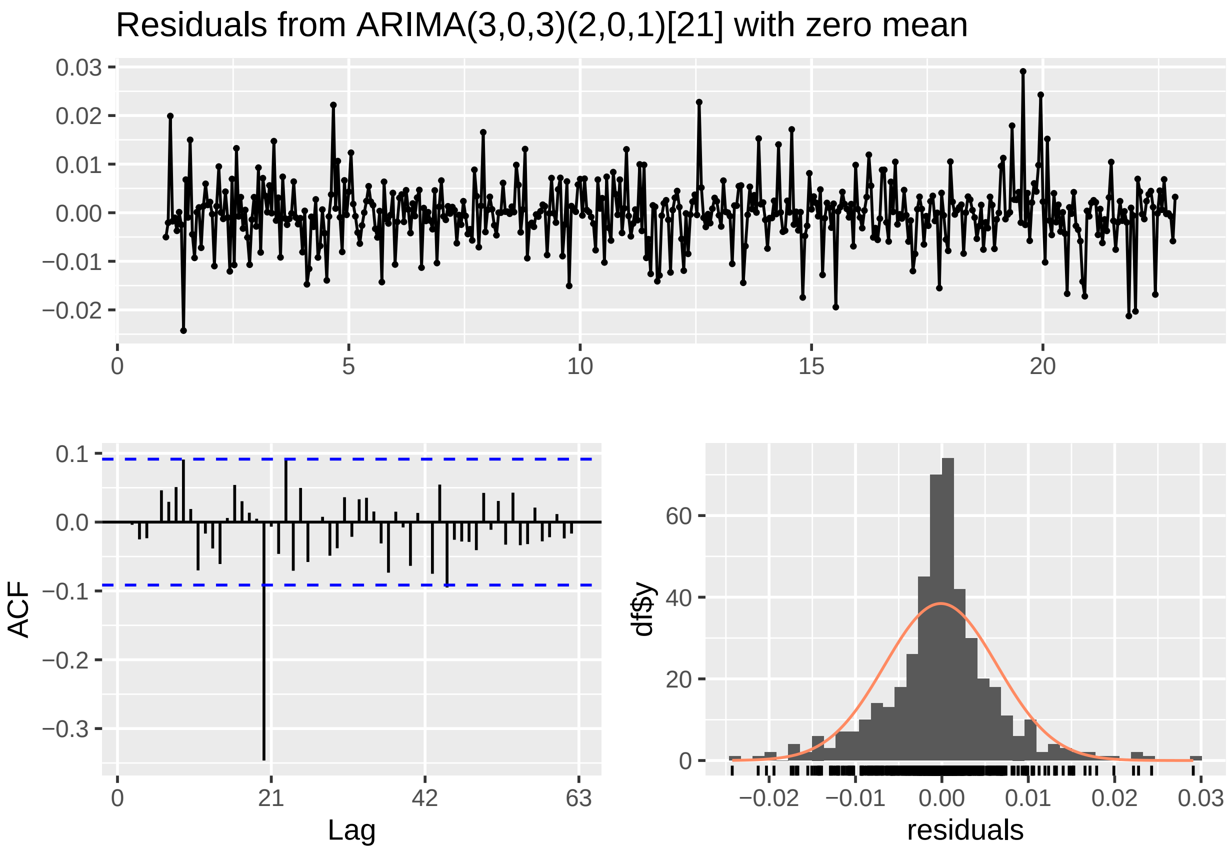
## 3.1 Seasonal Arima Model of Close-to-Close Volatility

The ACF of the residuals to the naive model for close-to-close volatility had high autocorrelations at lag 21, equal to the time period for computing the rolling volatility.

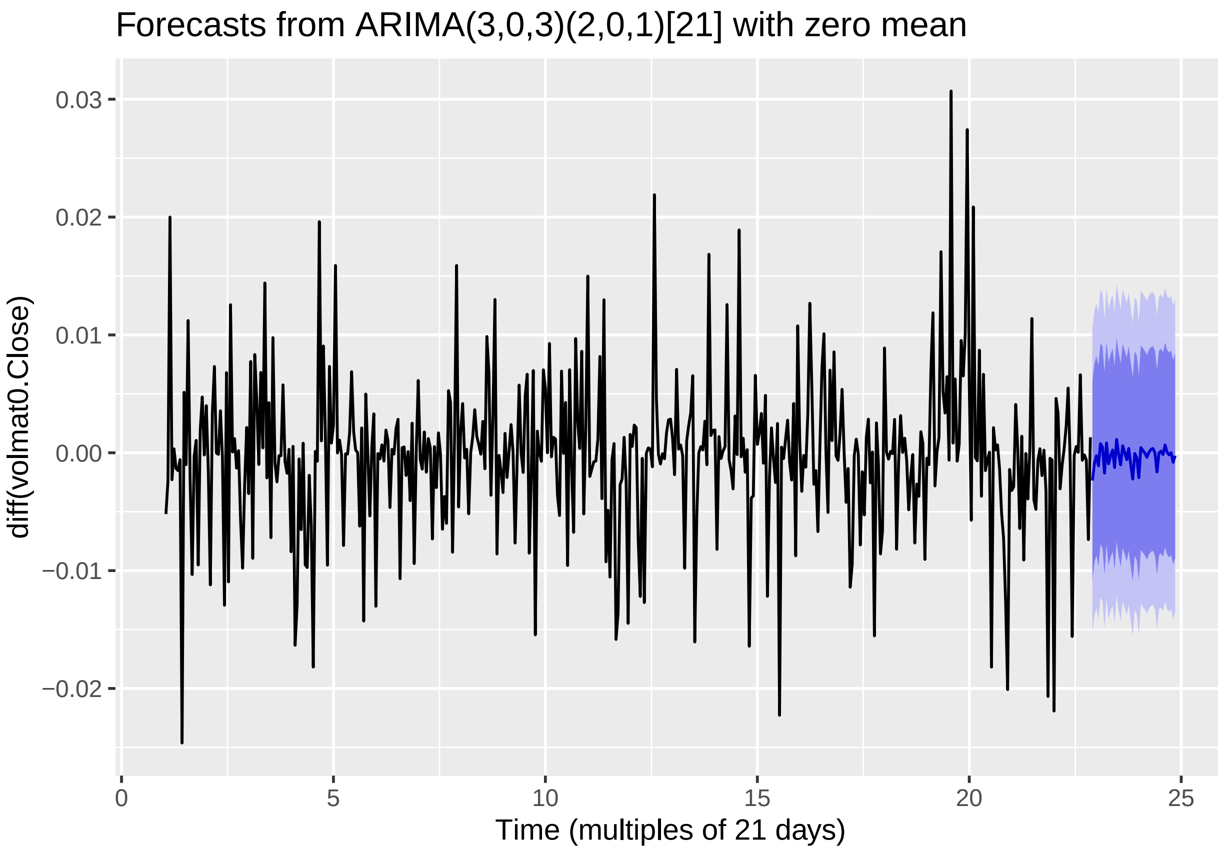





```
## Series: diff(volmat0.Close)
## ARIMA(3,0,3)(2,0,1)[21] with zero mean
##
## Coefficients:
##           ar1     ar2     ar3     ma1      ma2      ma3    sar1    sar2
##       -0.4468  0.2509  0.7583  0.4848  -0.1656  -0.6018  0.1336  0.0994
## s.e.   0.0950  0.1307  0.0885  0.1167   0.1620   0.1037  0.3550  0.0558
##          sma1
##       -0.0849
## s.e.   0.3573
##
## sigma^2 = 4.323e-05:  log likelihood = 1659.08
## AIC=-3298.17   AICc=-3297.68   BIC=-3256.88
```



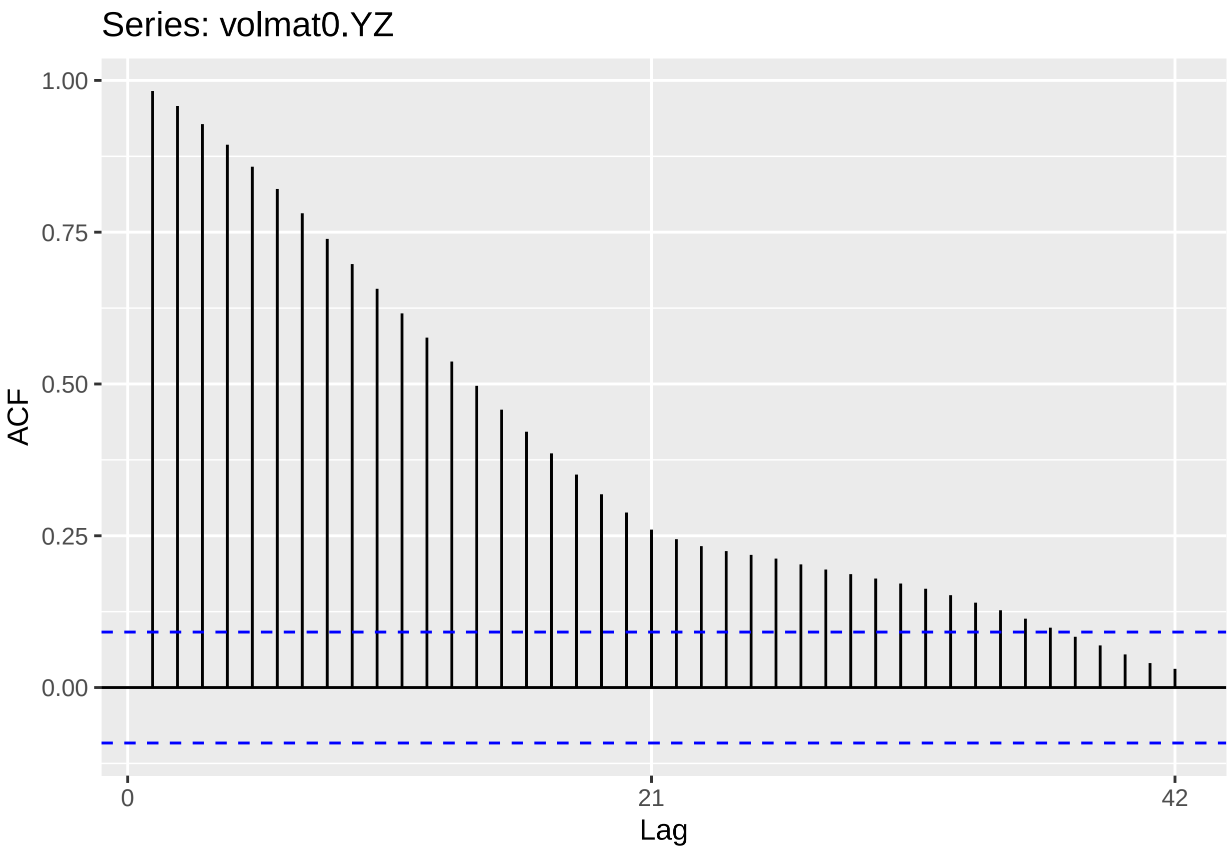

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(3,0,3)(2,0,1)[21] with zero mean
## Q* = 91.934, df = 33, p-value = 1.826e-07
##
## Model df: 9.   Total lags used: 42
```

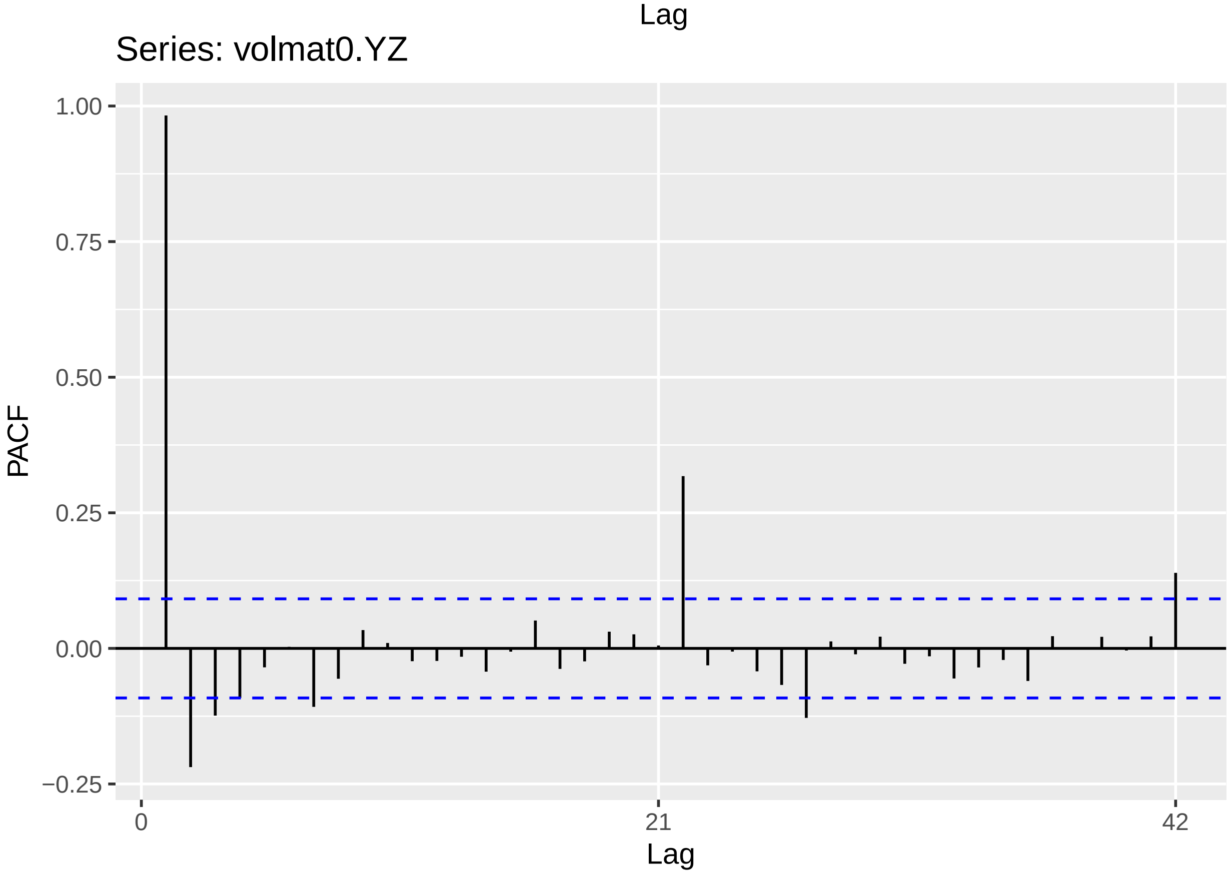



```
##                         ME        RMSE         MAE       MPE     MAPE      MASE
## Training set -0.0001222467 0.006510296 0.004432713 -146.9053 627.1043 0.6578914
##                        ACF1
## Training set -0.0005554823
```

### 3.2 Seasonal Arima Model of Yang Zhang OHLC Volatility

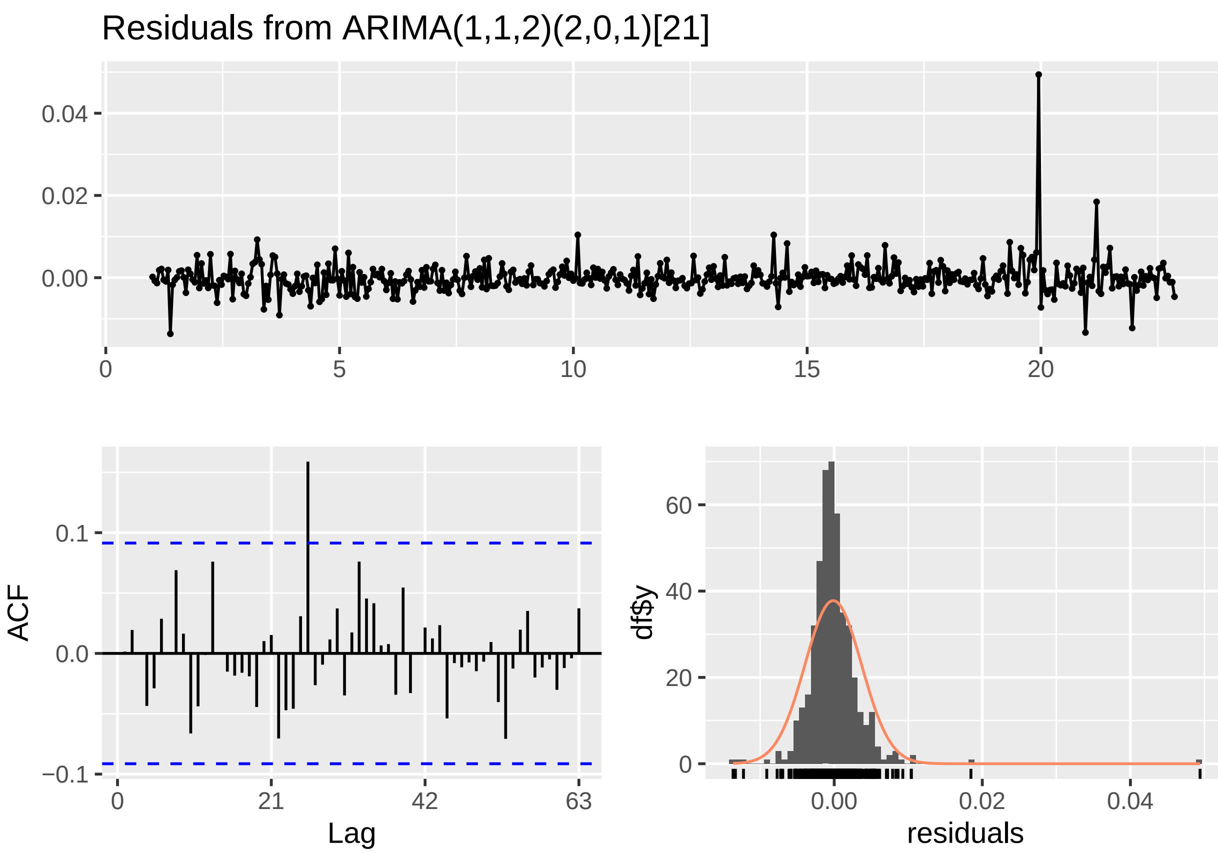
With the Yang Zhang Volatility computed on price series of length 21 days, fit seasonal models with frequency (number of seasons per 'year') equal to 21.



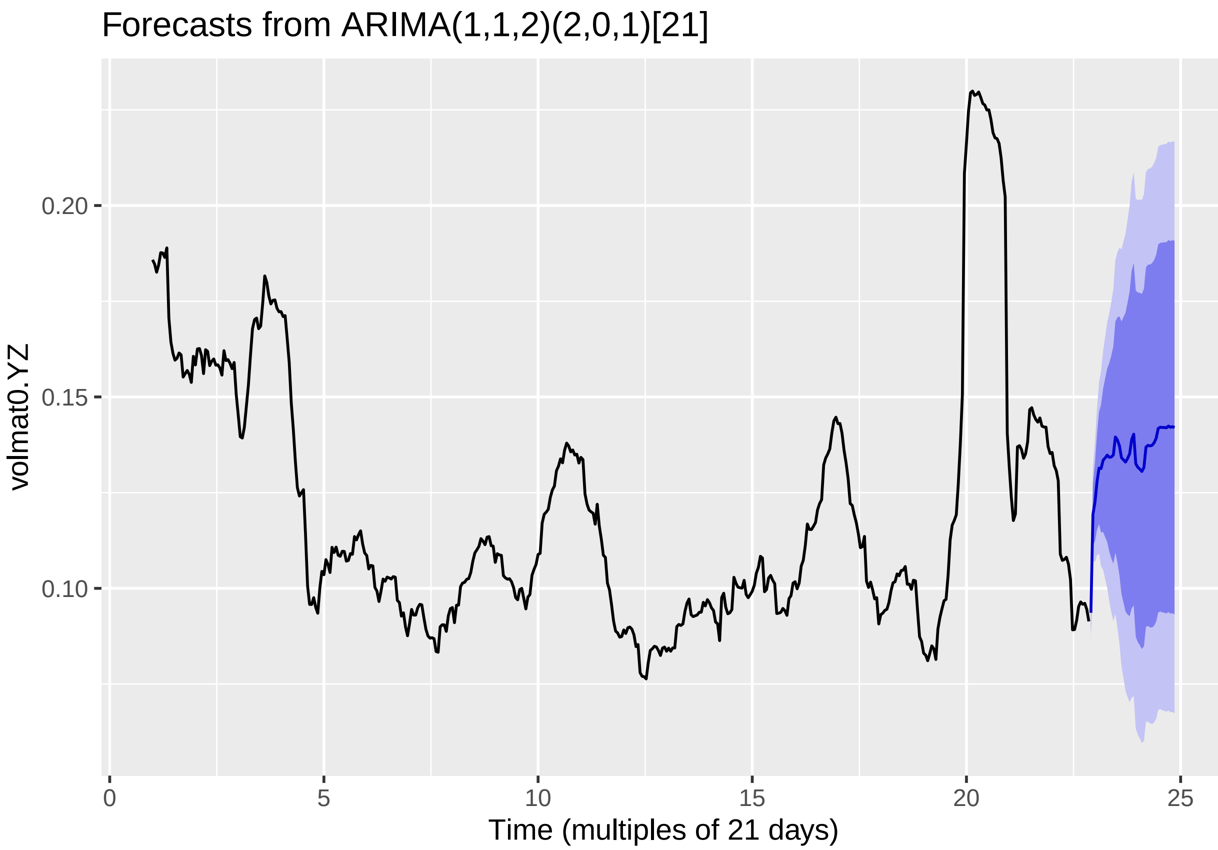




```
## Series: volmat0.YZ
```

```
## ARIMA(1,1,2)(2,0,1)[21]
##
## Coefficients:
##          ar1      ma1      ma2     sar1    sar2     sma1
##       0.9305  -0.6998  -0.1038  -0.3333  -0.305  -0.5834
## s.e.  0.0478   0.0685   0.0539   0.0938   0.083   0.0856
##
## sigma^2 = 1.487e-05:  log likelihood = 1890.41
## AIC=-3766.81   AICc=-3766.56   BIC=-3737.91
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,1,2)(2,0,1)[21]
## Q* = 38.526, df = 36, p-value = 0.356
##
## Model df: 6.   Total lags used: 42
```



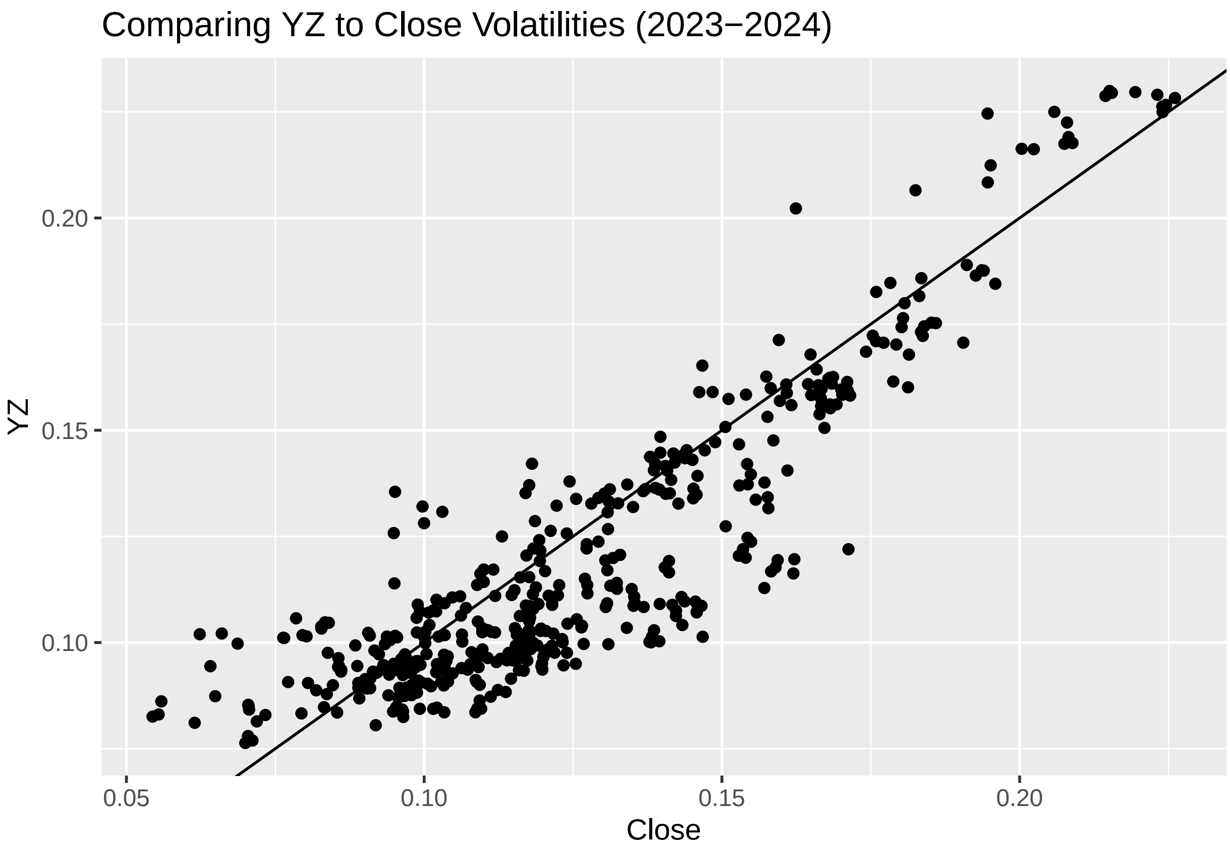

```
##                          ME        RMSE         MAE        MPE     MAPE
## Training set -0.000127886 0.003826245 0.002264091 -0.1398705 1.905408
##                    MASE        ACF1
## Training set 0.08381064 0.001466067
```

*Note:*

- The MAPE of the Seasonal Arima model of the YZ volatility is 1.9054079 versus 3.6258417 for the Naive model of the Close volatility
- Note whether the coefficients of the seasonal Arima model are significant or highly significant in terms of being many multiples of their standard errors.
- In addition to being a more efficient estimate of historical volatility (theoretically), evaluate whether the predictability of the YZ volatility is stronger than that of the Close volatility (compare accuracy statistics for the two models).

## 4. Comparing Close-to-Close and Yang-Zhang Volatilities

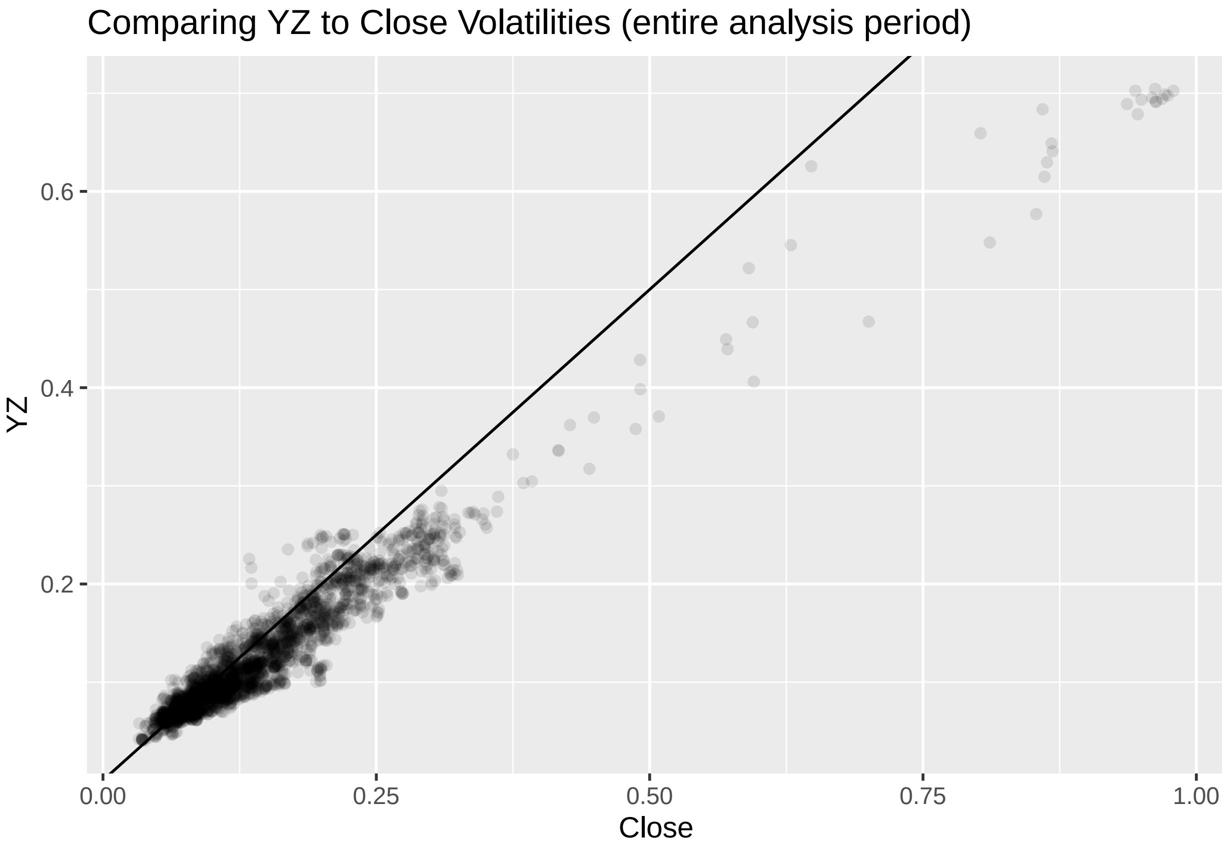
We first compare daily pairs of (Close,YZ) volatilities for 2023 to 2024




*Note:*

- Is the YZ volatility is generally smaller than the Close volatility?
- There are cases when the YZ volatility is much higher than the Close volatility
- Theoretically, the variance of the Close volatility is much higher than the variance of the YZ volatility.

Second, we compare the volatilities for the entire analysis period



# References


- Garman, M. B. and Klass, M. J. (1980). On the estimation of security price volatilities from historical data. Journal of business, pages 67-78
- Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. Journal of business, pages 61-65.
- Rogers, L. C., Satchell, S. E., and Yoon, Y. (1994). Estimating the volatility of stock prices: a comparison of methods that use high and low prices. Applied Financial Economics, 4(3):241-247.
- Rogers, L. C. G. and Satchell, S. E. (1991). Estimating variance from high, low and closing prices. The Annals of Applied Probability, pages 504-512.
- Yang, D. and Zhang, Q. (2000). Drift-independent volatility estimation based on high, low, open, and close prices. The Journal of Business, 73(3):477-492.
- Hyndman, Rob J, and Athanasopoulos, George (2018) Forecasting: Principles and Practice (2nd ed), https://otexts.com/fpp2/