Analysis and forecasting of fishing in the United States

Group 3

Vallabh Sawant

GauravKumar Vishwakarma

Rakshit Karkera

Goal

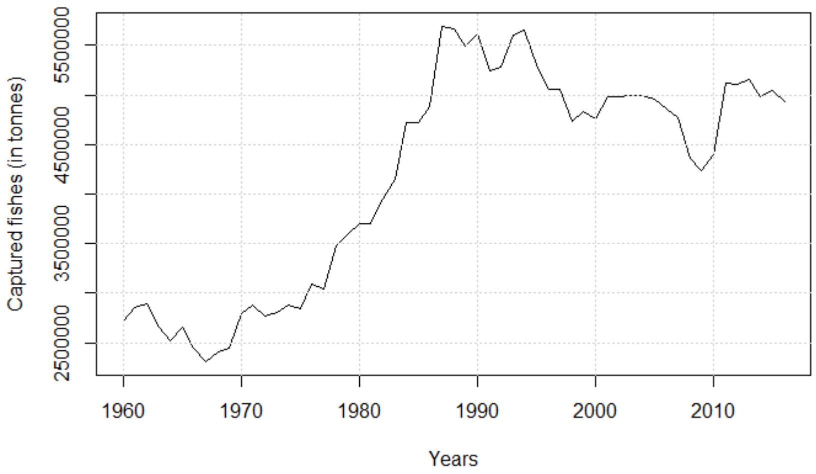
- Our goal is to predict the amount of fish in metric tons that needs to be captured to suffice the needs of population for United States.
- We plan to do a long forecast of 3 years to predict the same.
- We are using various forecasting techniques such as Naïve, Simple Moving Average, Holt-Winters, Regression, Arima, Exponential Smoothing and Random Walk Forest to achieve the goal.
- The best forecasting technique would be decided based on the accuracy measures. We are going to consider MAPE as a good accuracy measure since it is scale independent and can be used to compare different forecast scenarios.

Dataset

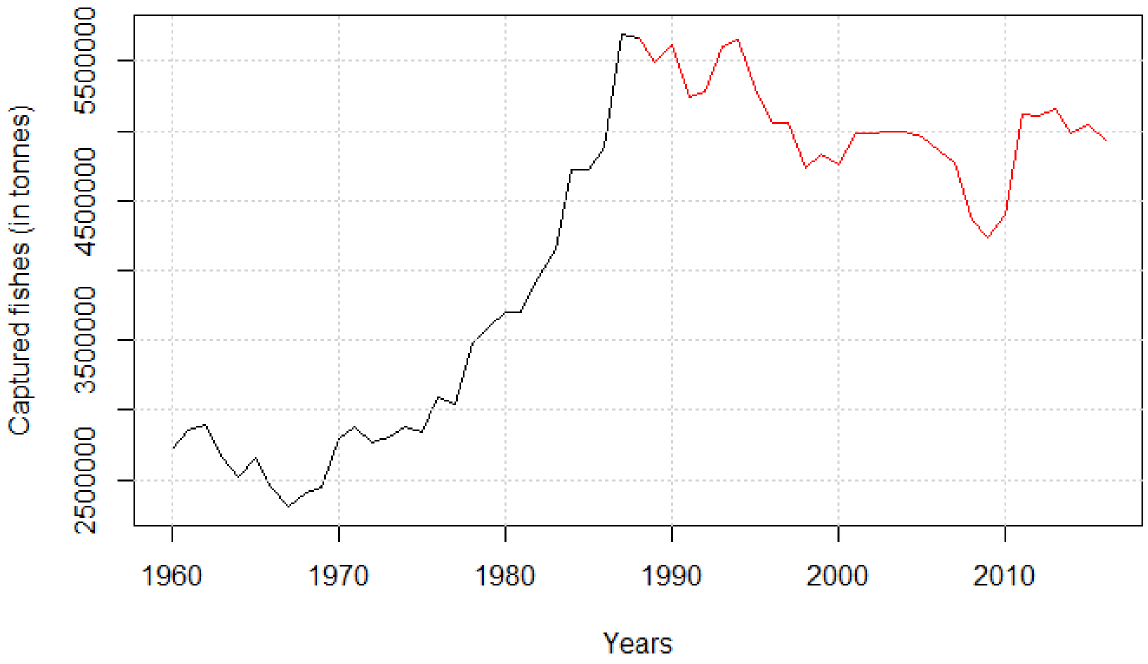
Entity <chr>	Code <chr>	Year <dbl>	Capture fisheries production (metric tons) <dbl>
United States	USA	1960	2714623
United States	USA	1961	2852004
United States	USA	1962	2897963
United States	USA	1963	2655052
United States	USA	1964	2519951
United States	USA	1965	2649980

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2311726 2885967 4721775 4171508 5045443 5694242
```

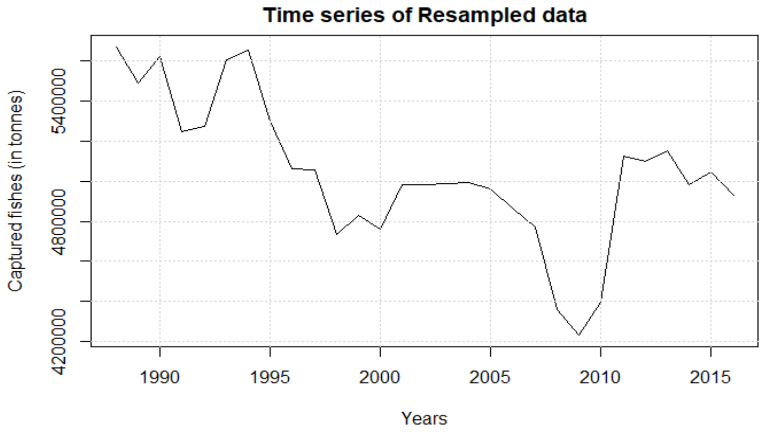
Our data includes fish capture in metric tons for the United States from 1960 to 2016. This data was found on Kaggle after looking for fish captured in the world.



- Sudden rise before the data stabilizes
- Previous data not so relevant for forecasting as there was a rise due to certain factors which don't exist anymore.

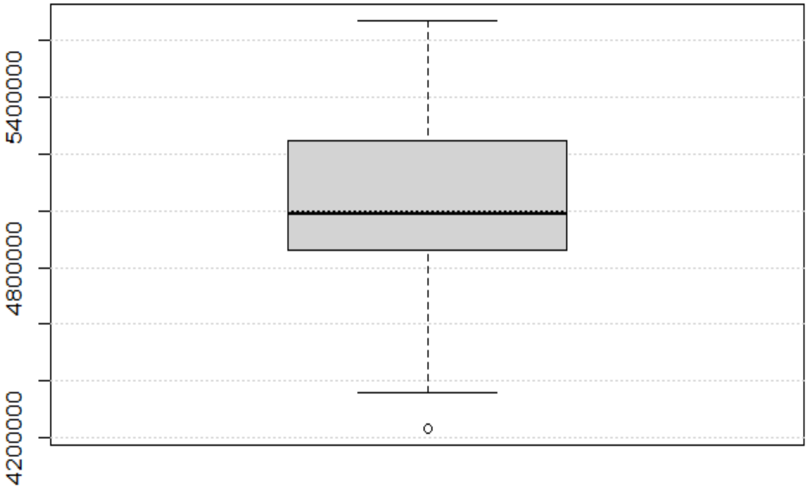

Captured fishes (in tonnes)
2500000
3500000
4500000
5500000
1960
1970
1980
1990
2000
2010
Years

- New data from 1988 to 2016



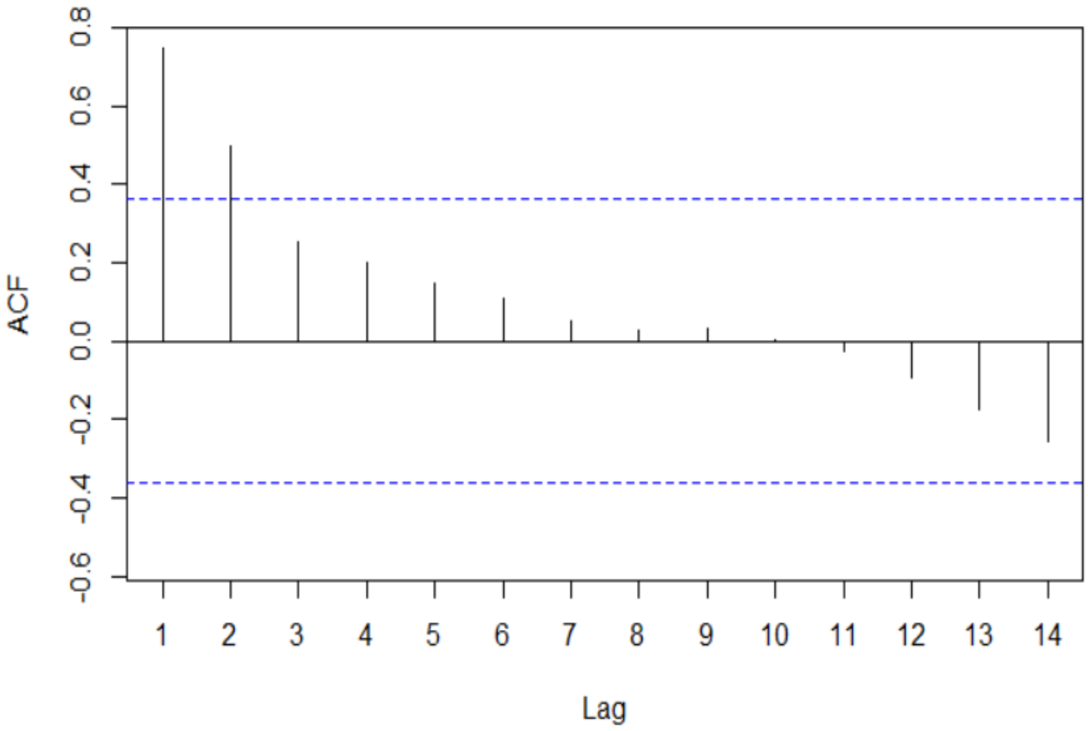
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4233804	4858805	4995418	5040506	5244569	5670666

- The boxplot confirms the mean of the data from 1988 to 2016 is around 5000000 with an outlier.
- The 1st quartile is 4858805 whereas 3rd quartile is at 5244569. The median is at 4995418 and the mean is 5040506.



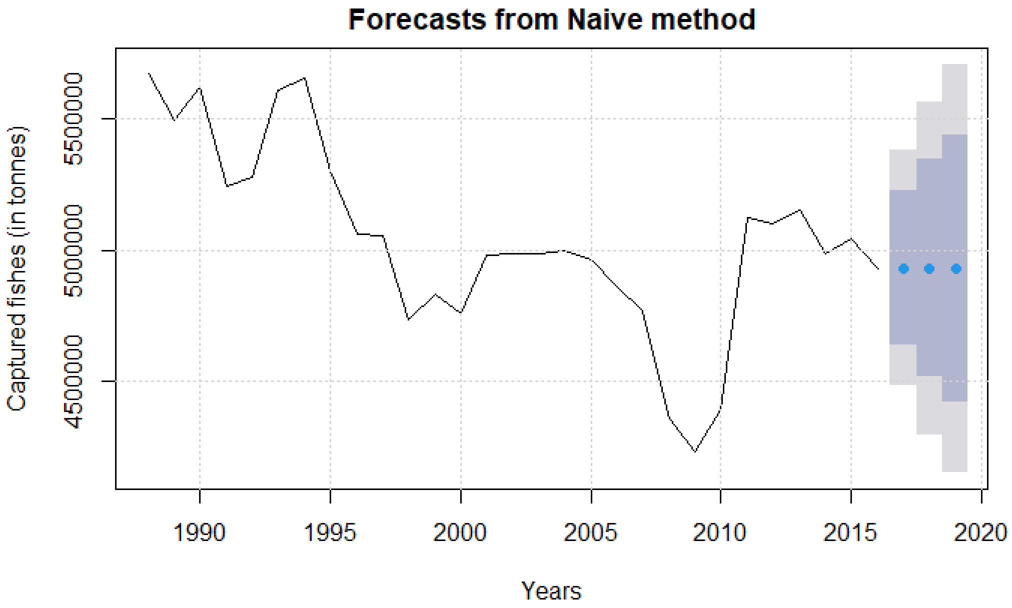
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4233804	4858805	4995418	5040506	5244569	5670666

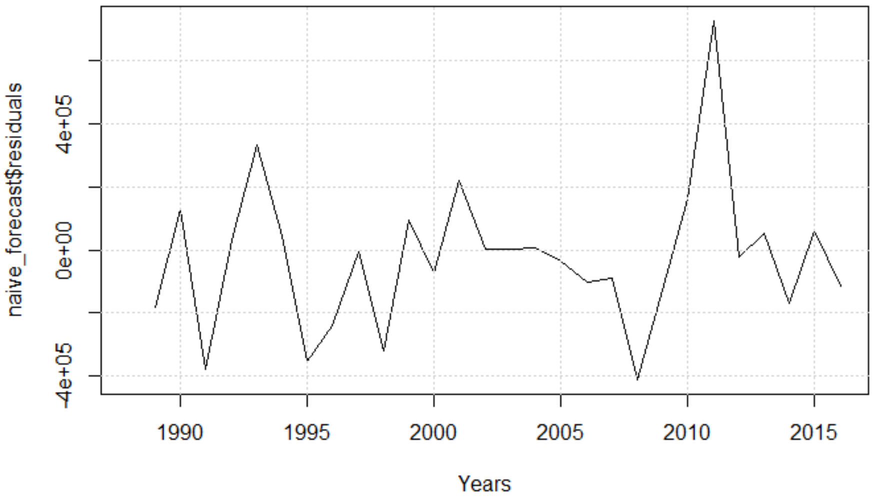
- Acf shows a high correlation with lag1.
- This states the current values are highly dependent on the previous values.



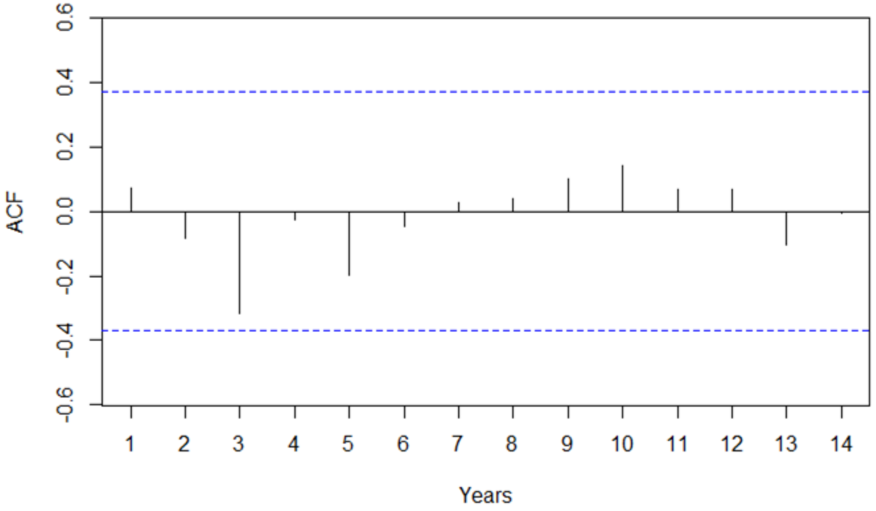
Naive Forecast

- Naïve forecast plot

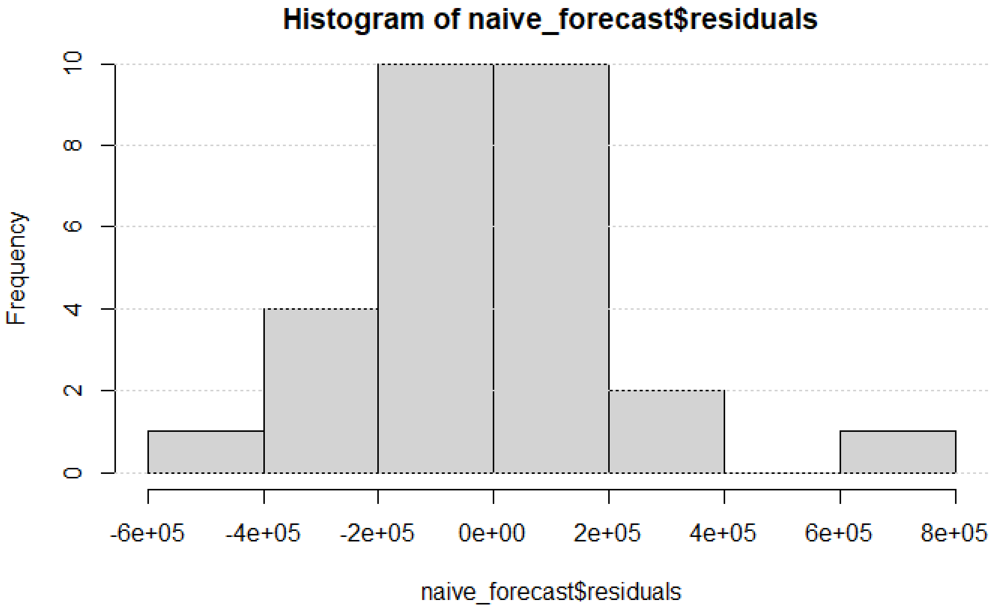


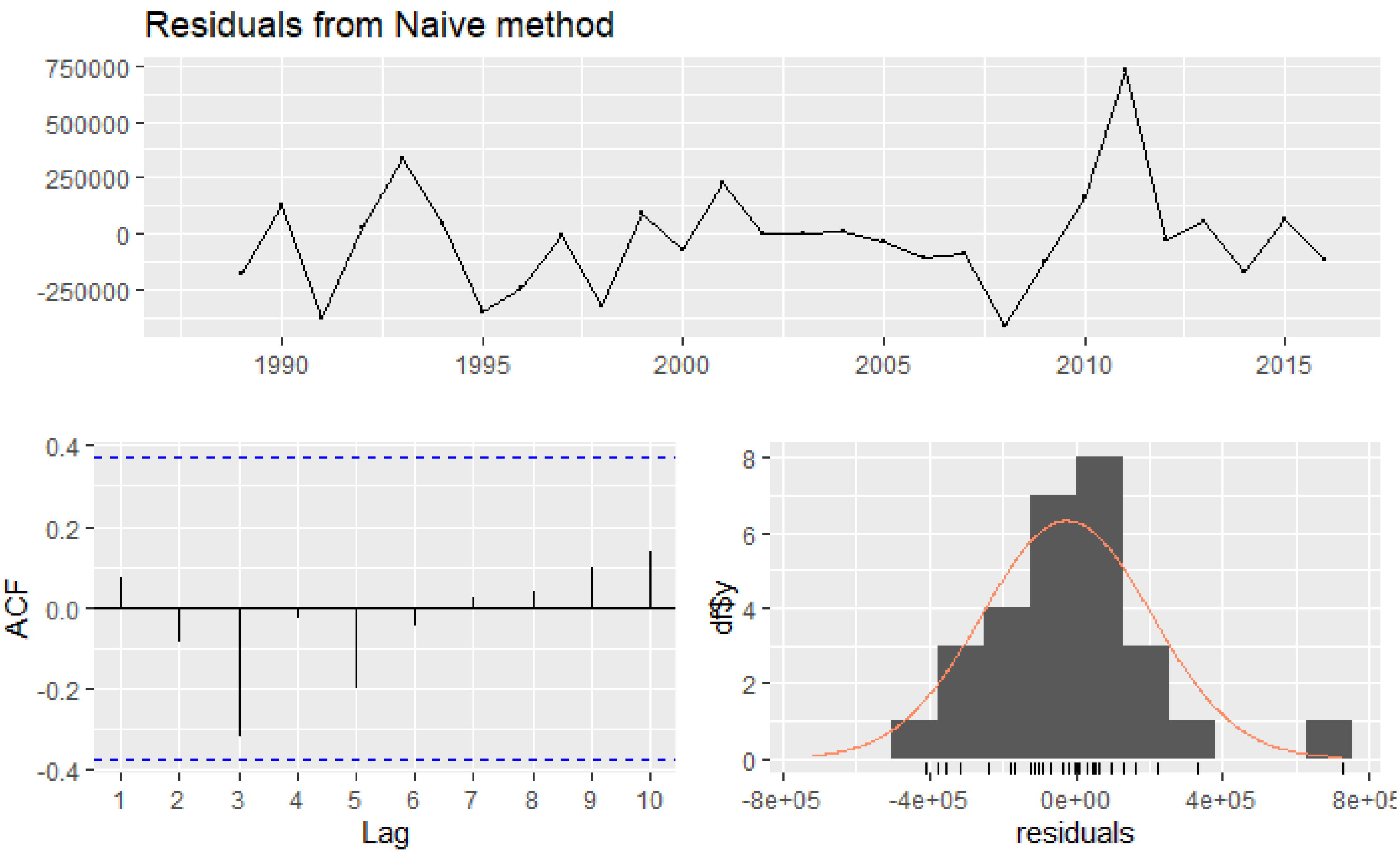


- The residual plot does not show any pattern.
- This means residuals are scattered randomly and they do not contribute much to data fluctuation.
- ACF also shows no correlation between residuals.



- The histogram is somewhat normally distributed with a few outliers.





Residuals from Naive method
750000
500000
250000
0
-250000
1990
1995
2000
2005
2010
2015
0.4
0.2
0.0
-0.2
-0.4
ACF
1
2
3
4
5
6
7
8
9
10
Lag
8
6
4
2
0
df$y
-8e+05
-4e+05
0e+00
4e+05
8e+05
residuals

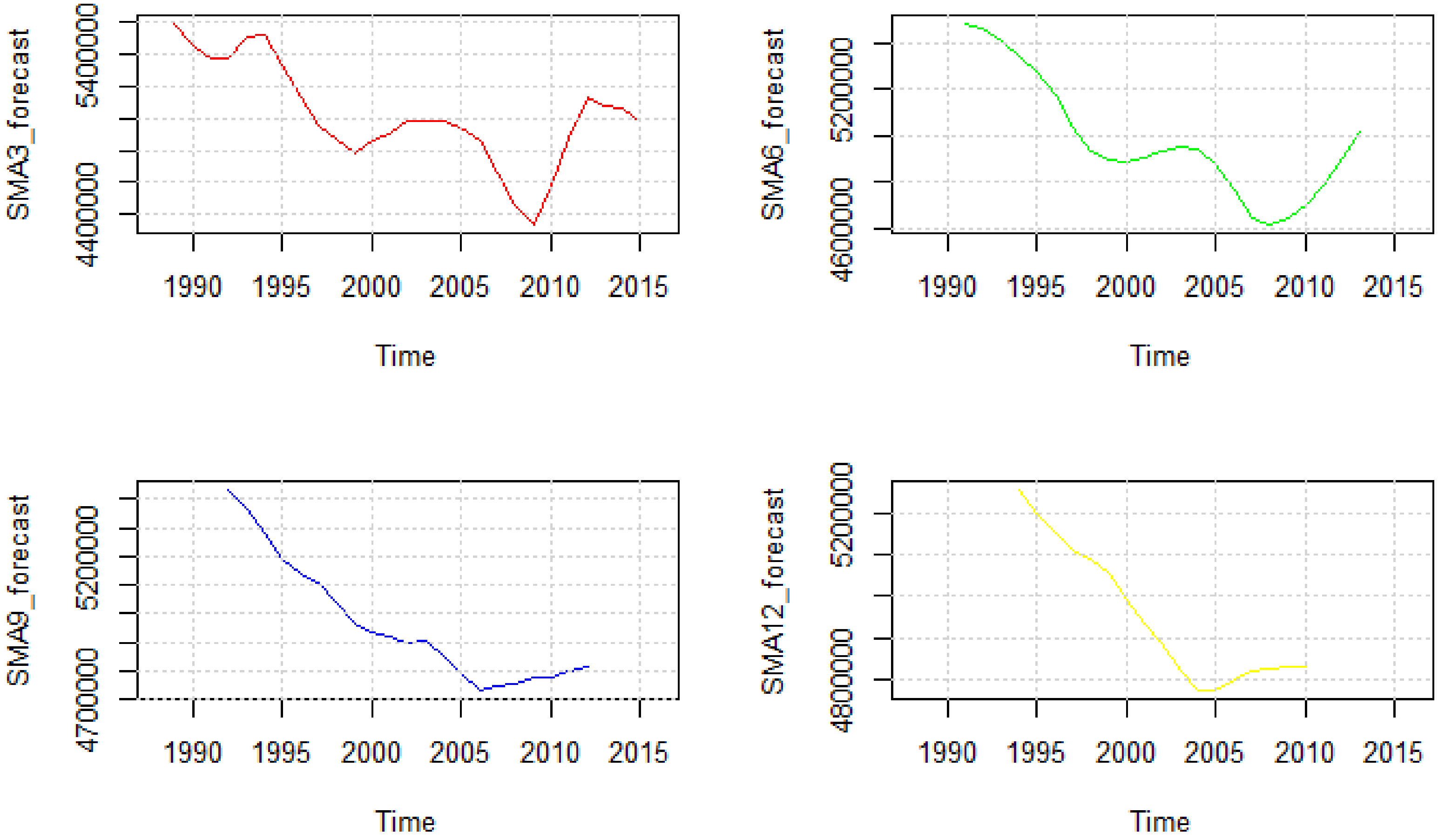
```
forecast(naive_forecast,h=3)
```

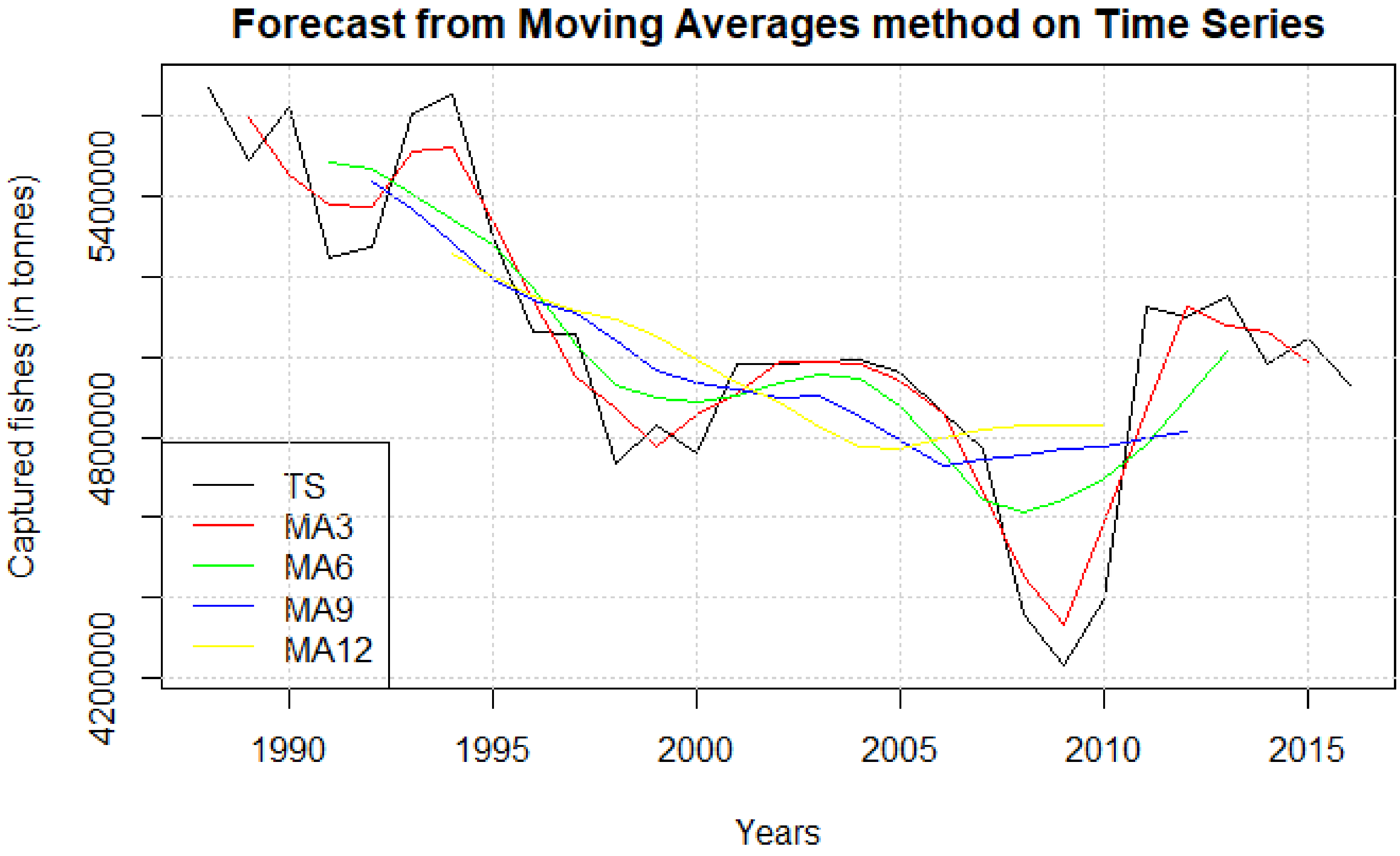
	Point Forecast <dbl>	Lo 80 <dbl>	Hi 80 <dbl>	Lo 95 <dbl>	Hi 95 <dbl>
2017	4931017	4638643	5223391	4483870	5378164
2018	4931017	4517538	5344496	4298656	5563378
2019	4931017	4424611	5437423	4156536	5705498

3 rows

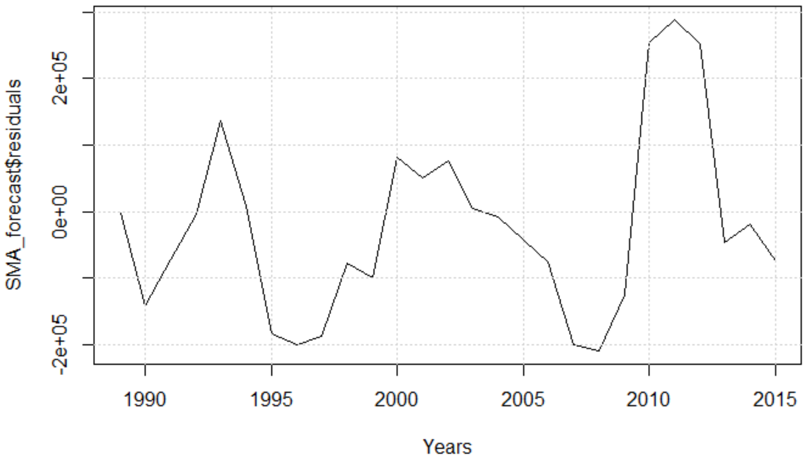
- The above table shows us the forecasted value.
- Naïve model predicted the point forecast to be 4931017 metric tons from 2017 to 2019.

Moving Averages


SMA3_forecast
4400000
5400000
1990
1995
2000
2005
2010
2015
Time
SMA6_forecast
4600000
5200000
1990
1995
2000
2005
2010
2015
Time
SMA9_forecast
4700000
5200000
1990
1995
2000
2005
2010
2015
Time
SMA12_forecast
4800000
5200000
1990
1995
2000
2005
2010
2015
Time

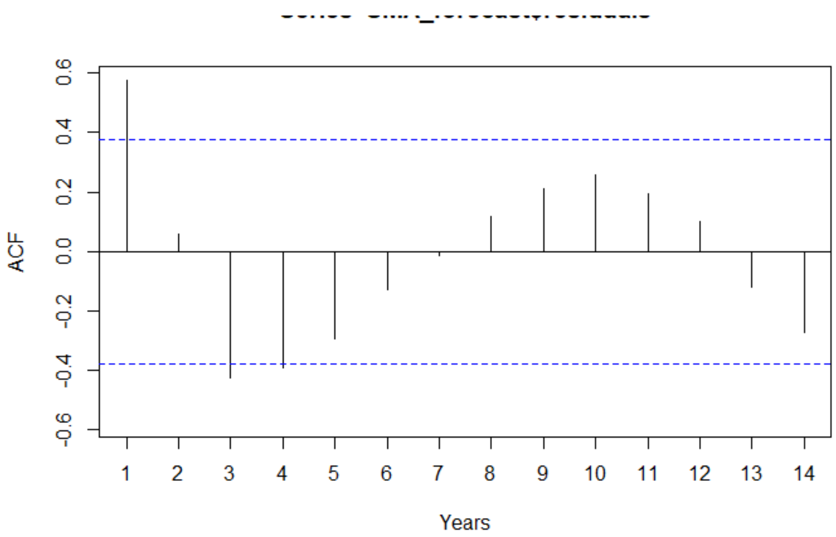


Forecast from Moving Averages method on Time Series
Captured fishes (in tonnes)
42000000
48000000
54000000
1990
1995
2000
2005
2010
2015
Years
TS
MA3
MA6
MA9
MA12

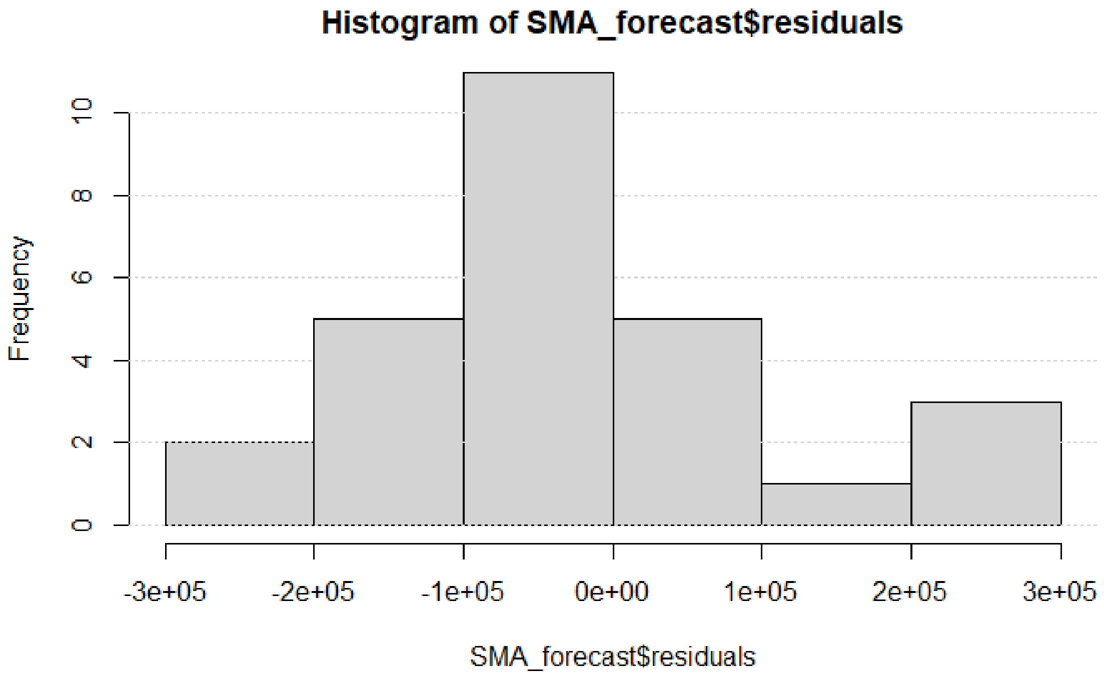


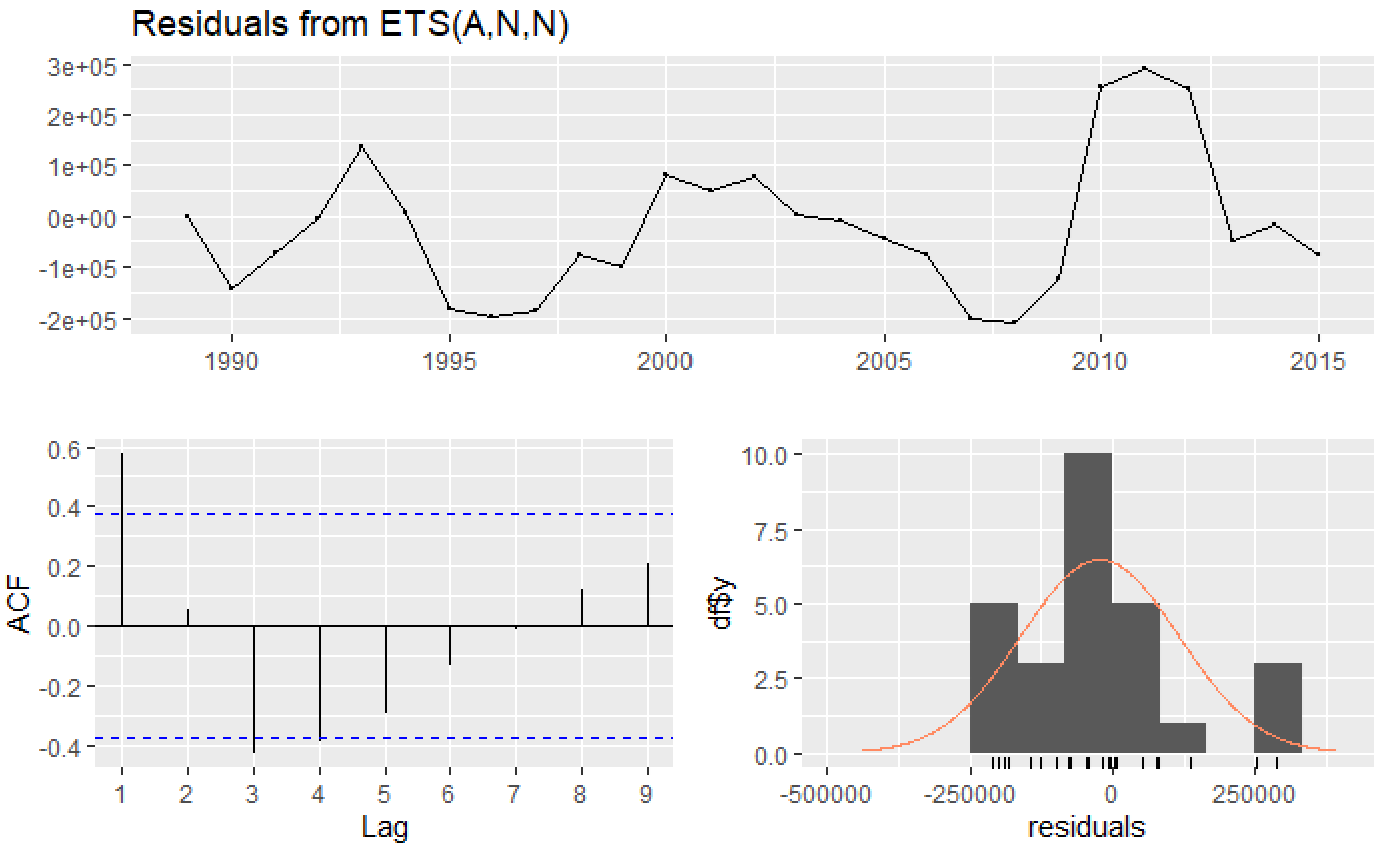
- The residual plot does not show any pattern.
- This means residuals are scattered randomly.

- ACF does shows some significant correlation between residuals. This is not ideal.

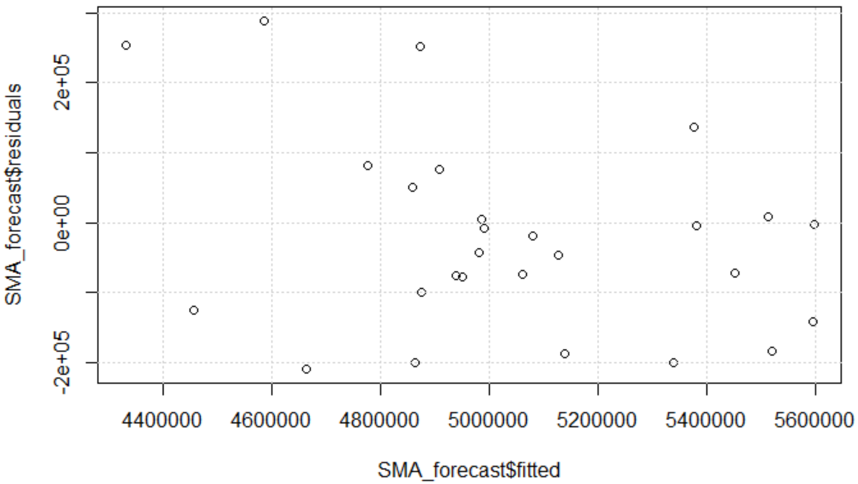


The residuals is not normally distributed. We can say it seems to right skewed very marginally.

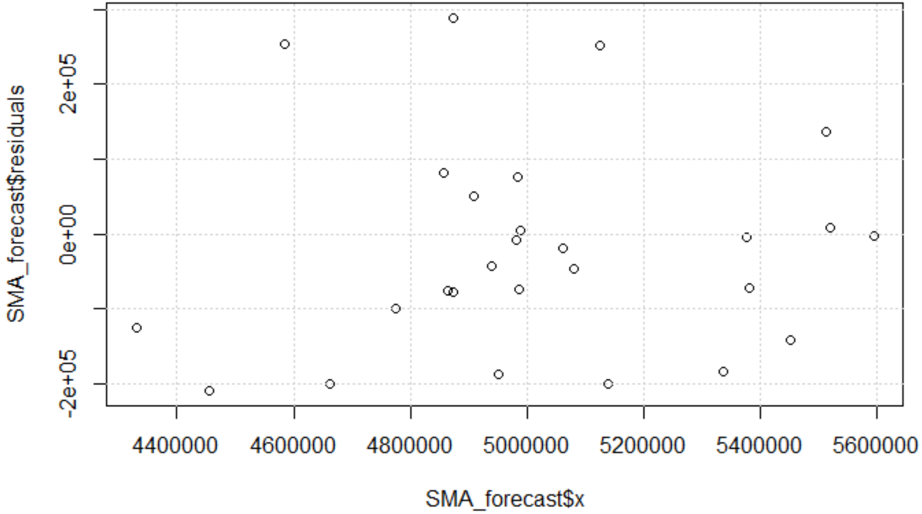




Residuals from ETS(A,N,N)
3e+05
2e+05
1e+05
0e+00
-1e+05
-2e+05
1990
1995
2000
2005
2010
2015
ACF
0.6
0.4
0.2
0.0
-0.2
-0.4
1
2
3
4
5
6
7
8
9
Lag
df$y
10.0
7.5
5.0
2.5
0.0
-500000
-250000
0
250000
residuals



Here we don't see a pattern in the graph. This is a good sign. The residuals are scattered randomly.

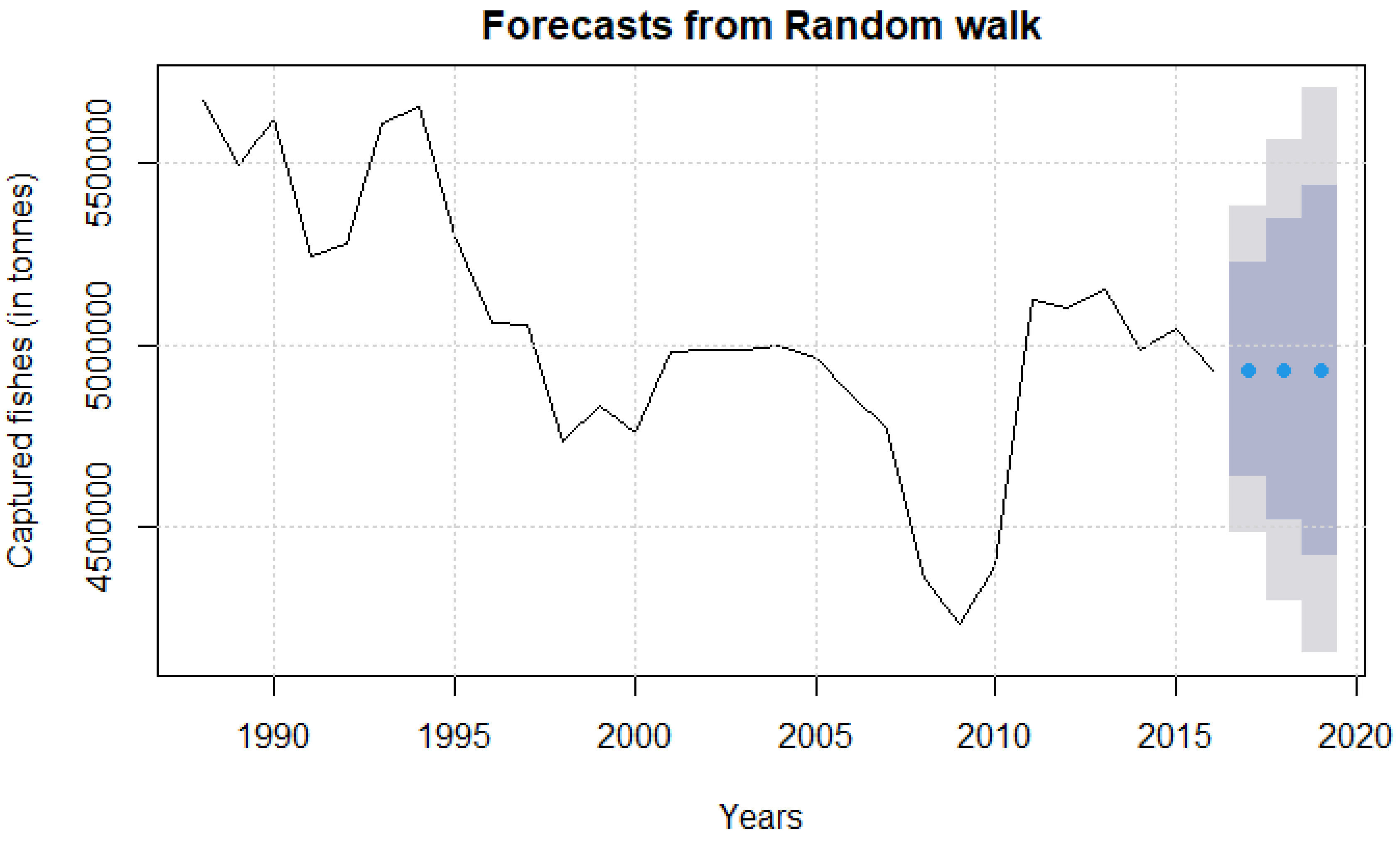


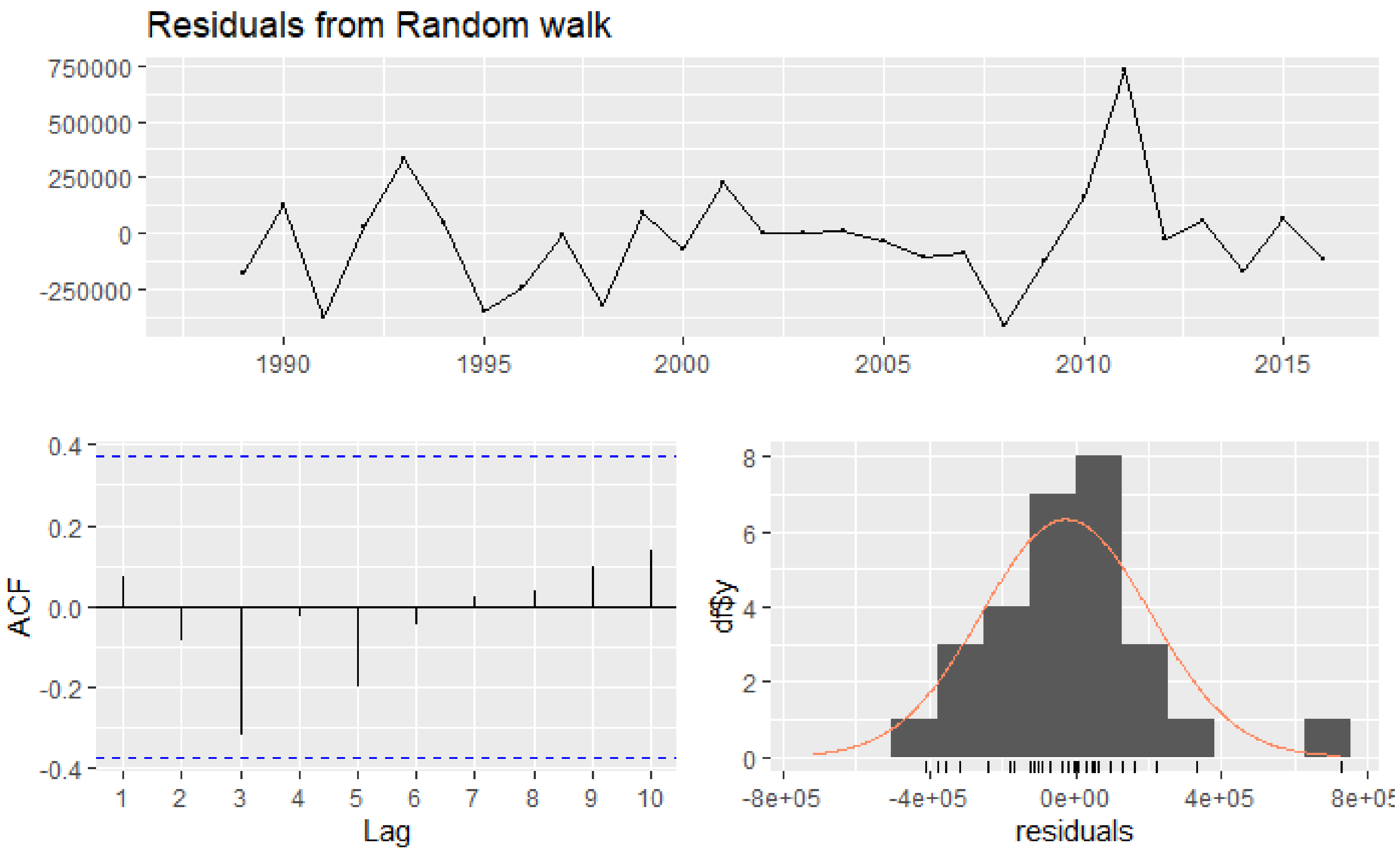
- Using Moving Avg with order 3 as recent data is better than all observations and smaller window provides more weight to recent data points

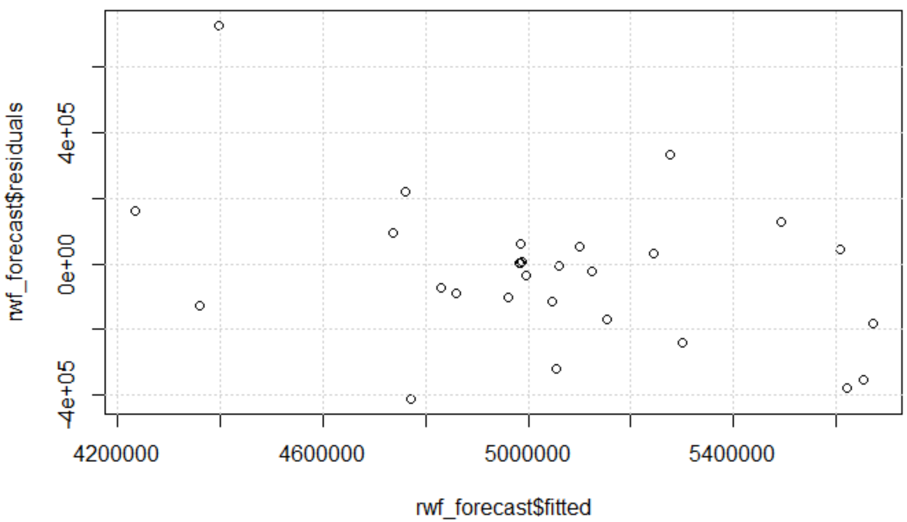
	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2016	4986988	4804071	5169905	4707240	5266735
2017	4986988	4728317	5245659	4591385	5382591
2018	4986988	4670187	5303788	4502483	5471492
2019	4986988	4621181	5352794	4427535	5546441

4 rows

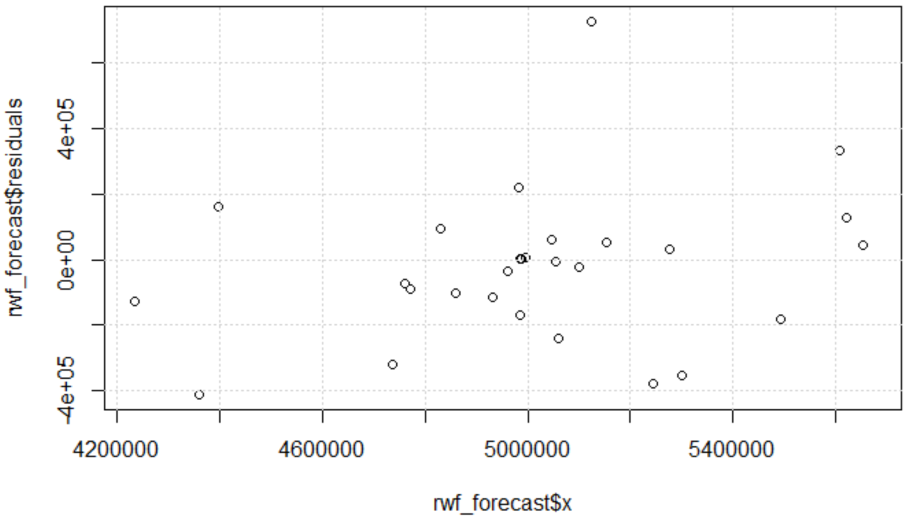
Random Walk Forecast


Forecasts from Random walk
Captured fishes (in tonnes)
4500000
5000000
5500000
1990
1995
2000
2005
2010
2015
2020
Years


Residuals from Random walk
750000
500000
250000
0
-250000
1990
1995
2000
2005
2010
2015
0.4
0.2
0.0
-0.2
-0.4
ACF
1
2
3
4
5
6
7
8
9
10
Lag
8
6
4
2
0
df$y
-8e+05
-4e+05
0e+00
4e+05
8e+05
residuals



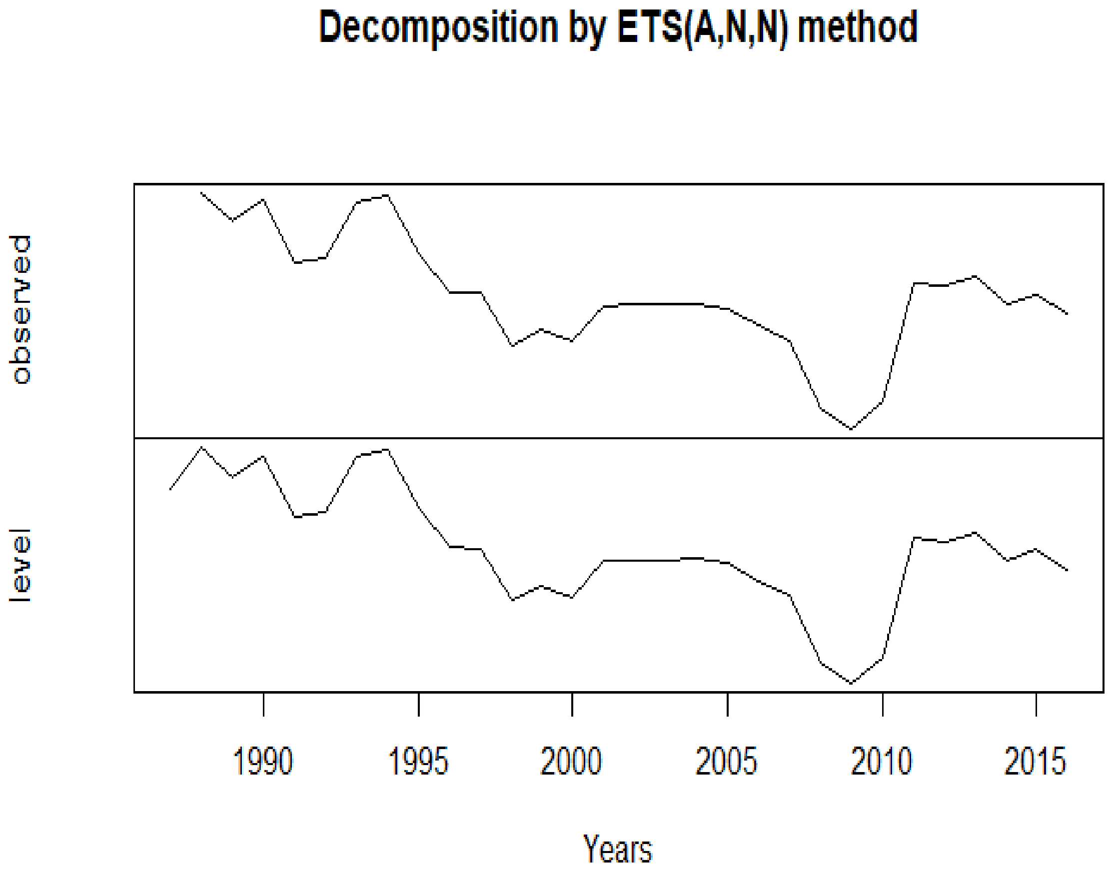
- The plots are scattered evenly.
- The residual is not significant



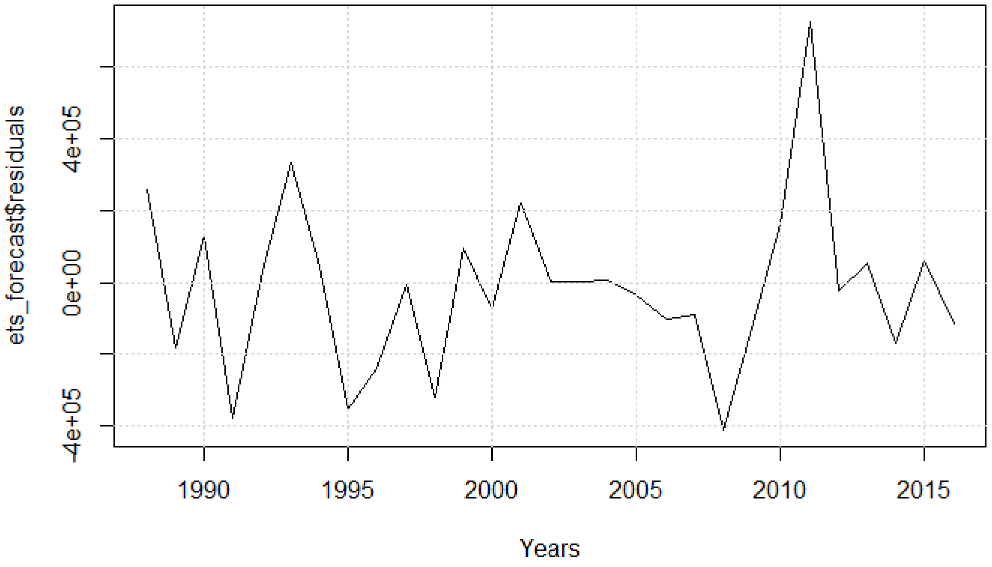
	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2017	4931017	4638643	5223391	4483870	5378164
2018	4931017	4517538	5344496	4298656	5563378
2019	4931017	4424611	5437423	4156536	5705498

3 rows

Exponential Smoothing

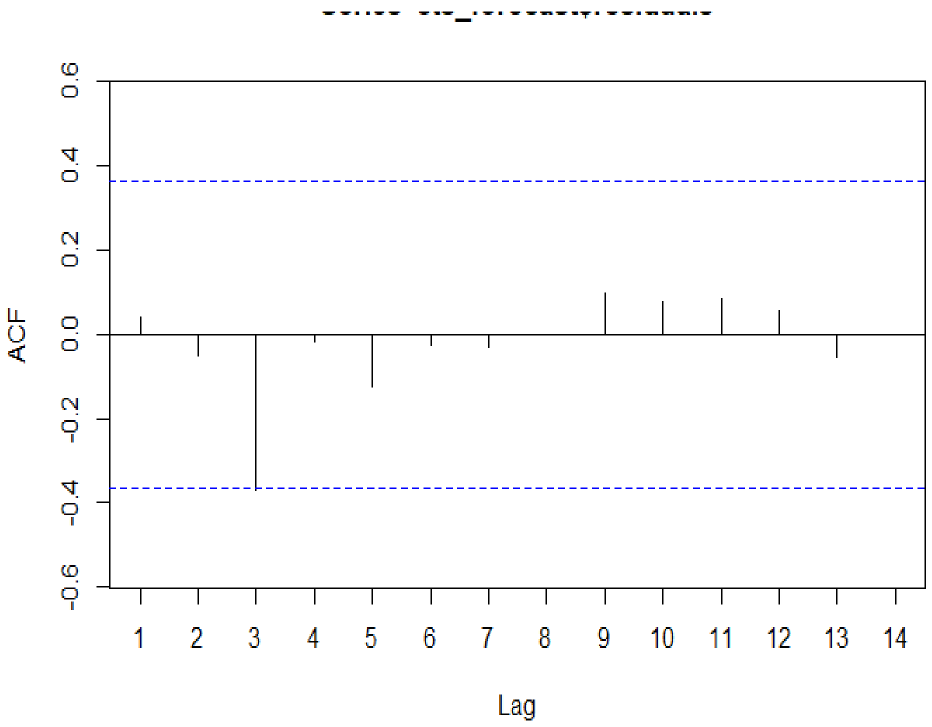


- The above graph is a comparision of the observed and the factors that affect the graph. Here we see the observed and level the same as there is no seasonality to change it.

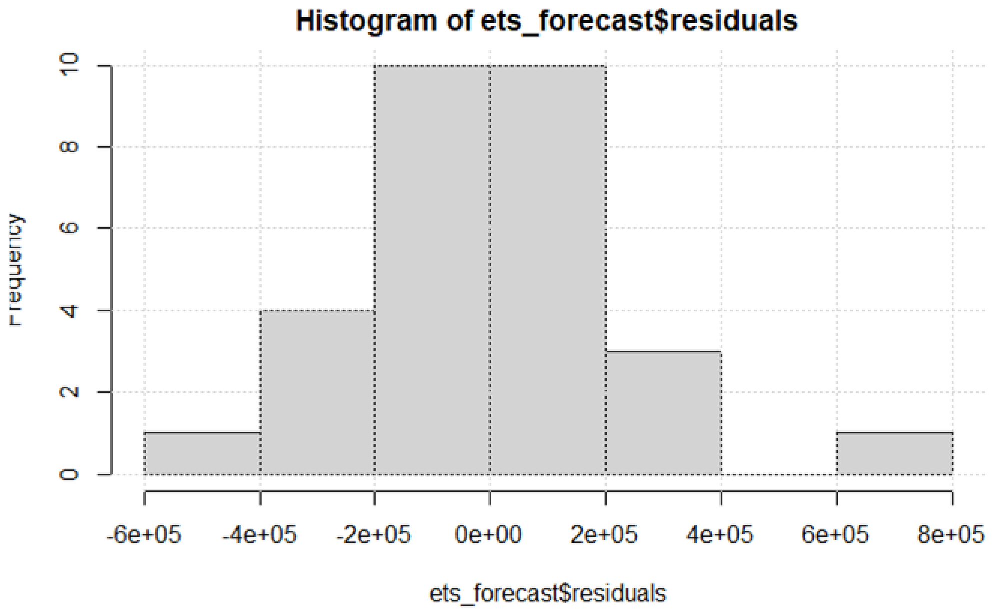


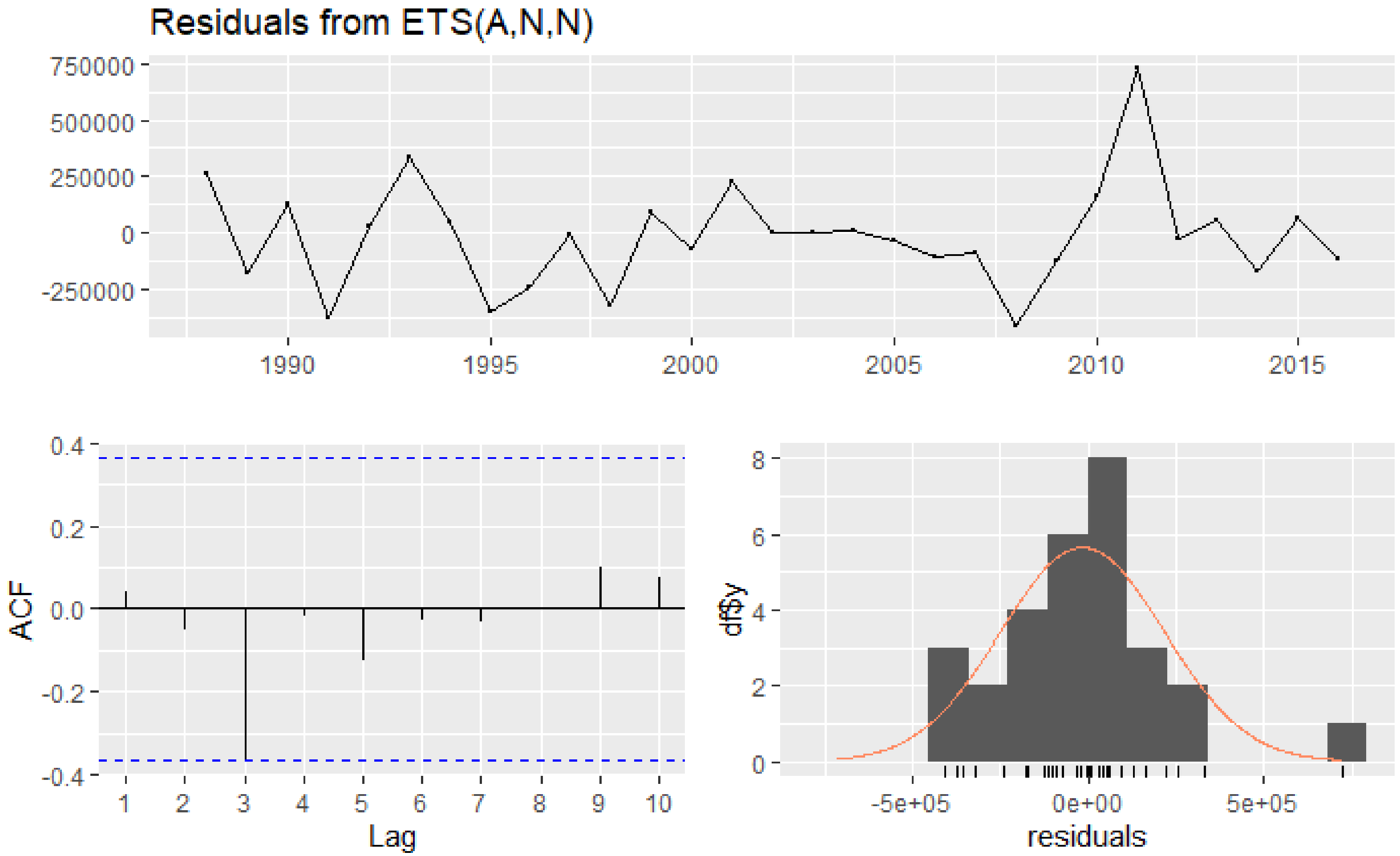
The plot is equally spread residuals around the horizontal line without a distinct pattern. This is a good indication that the residuals do not fluctuate the data.

- Here, in the Acf, we don't see any significant lines which states that there is no correlation between the errors.

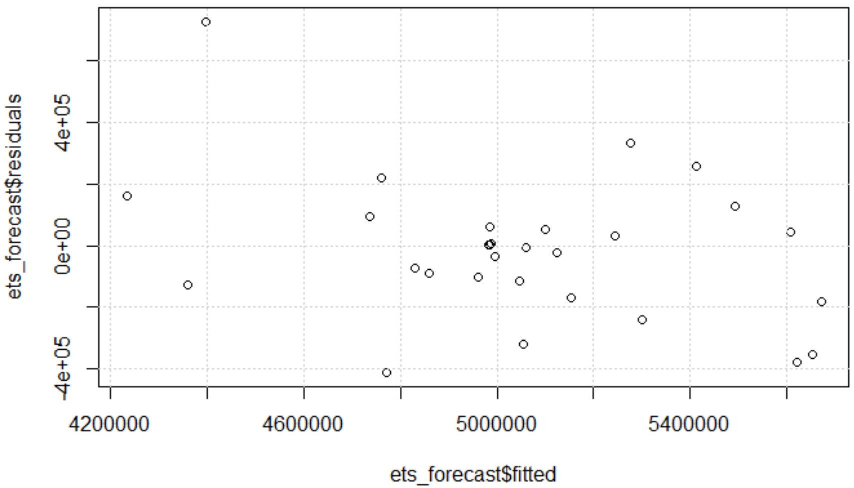


- The histogram is somewhat normally distributed with a few outliers.

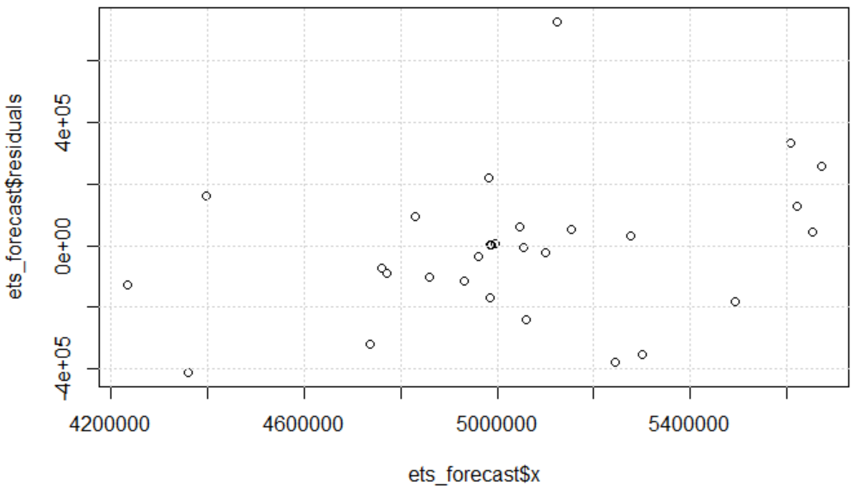




Residuals from ETS(A,N,N)
750000
500000
250000
0
-250000
1990
1995
2000
2005
2010
2015
ACF
0.4
0.2
0.0
-0.2
-0.4
1
2
3
4
5
6
7
8
9
10
Lag
df$y
8
6
4
2
0
-5e+05
0e+00
5e+05
residuals



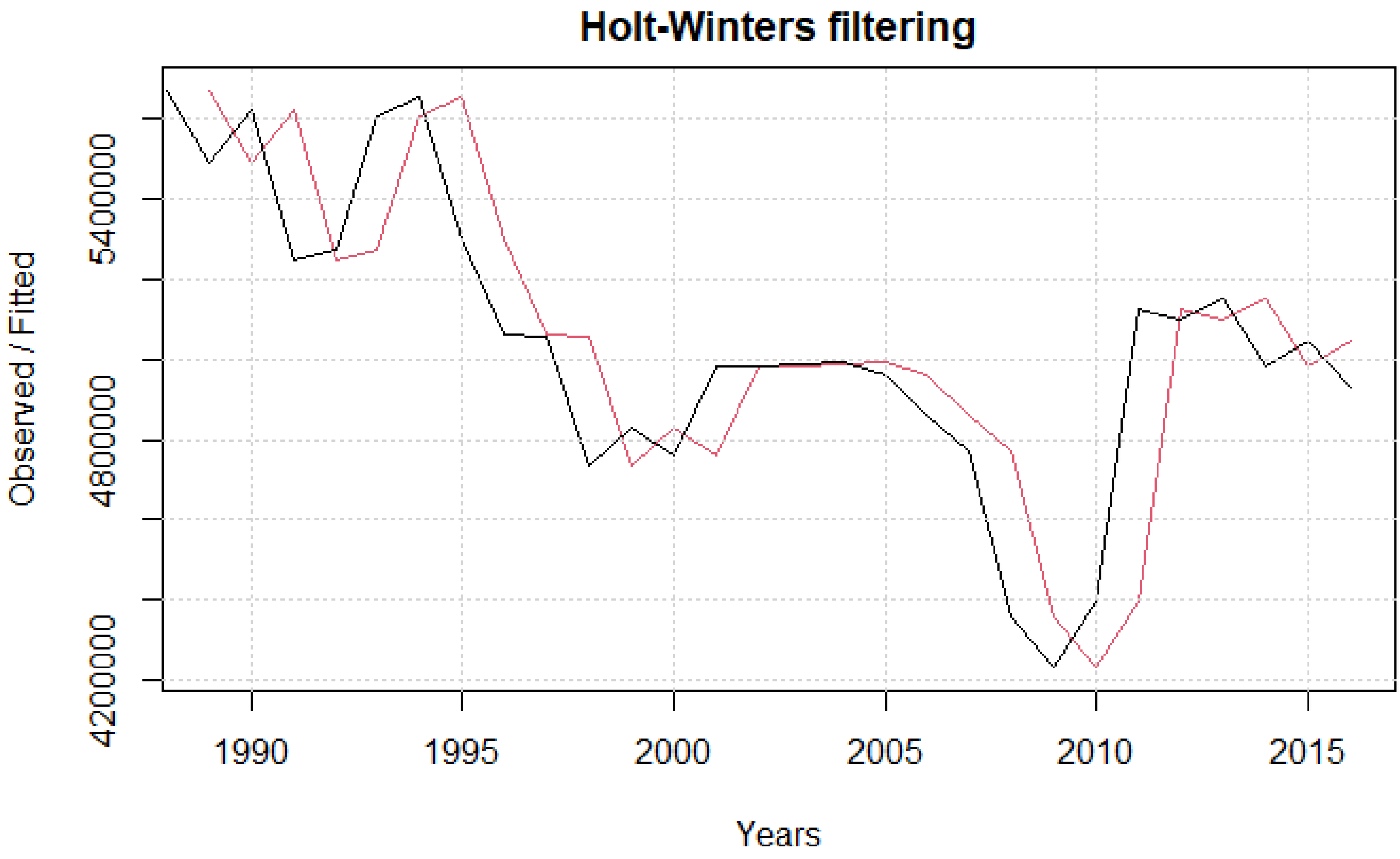
There is no patterns between the fitted vs residual and actual vs residual plot



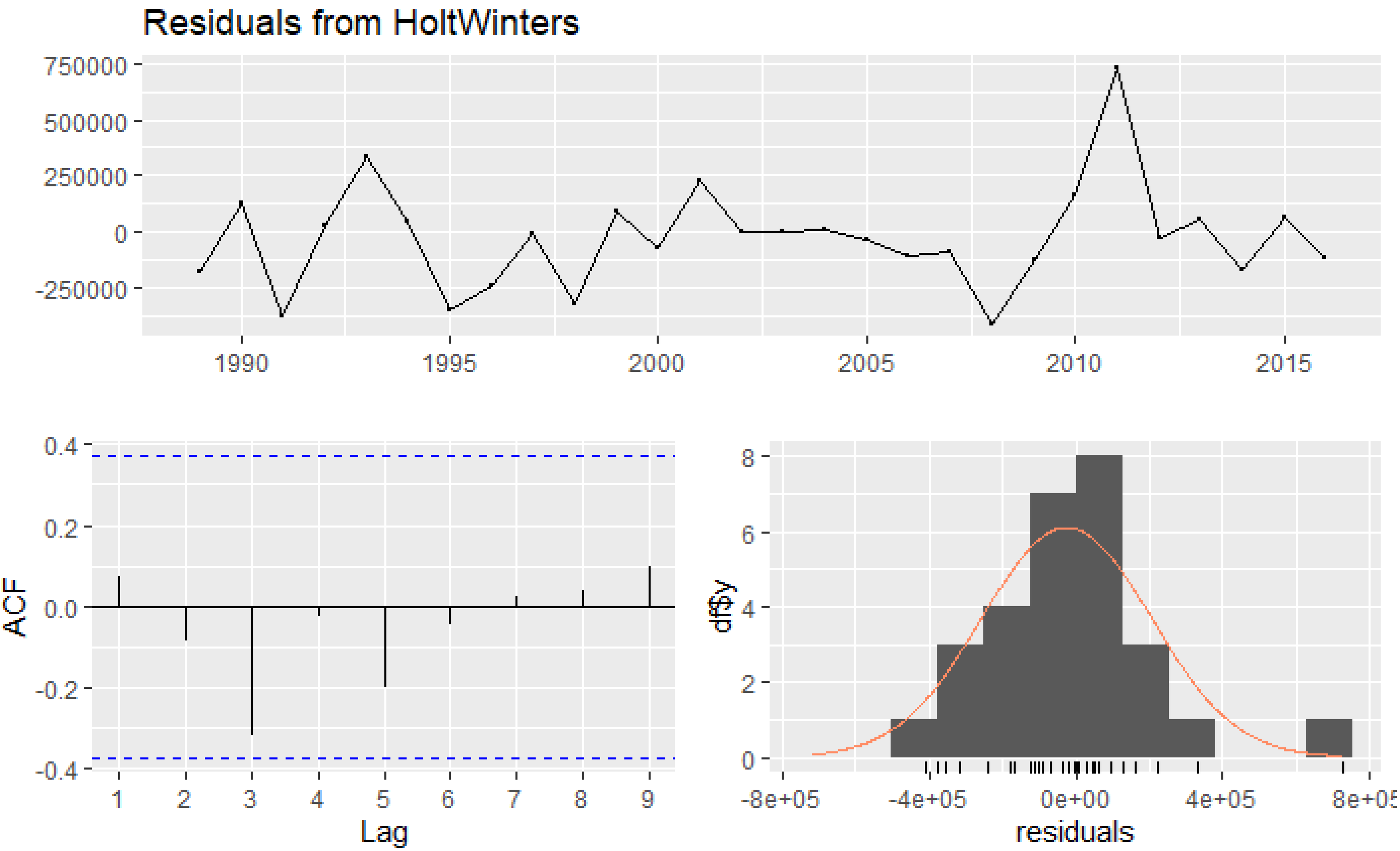
	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2017	4931029	4626568	5235489	4465396	5396661
2018	4931029	4500478	5361579	4272558	5589499
2019	4931029	4403723	5458334	4124584	5737473

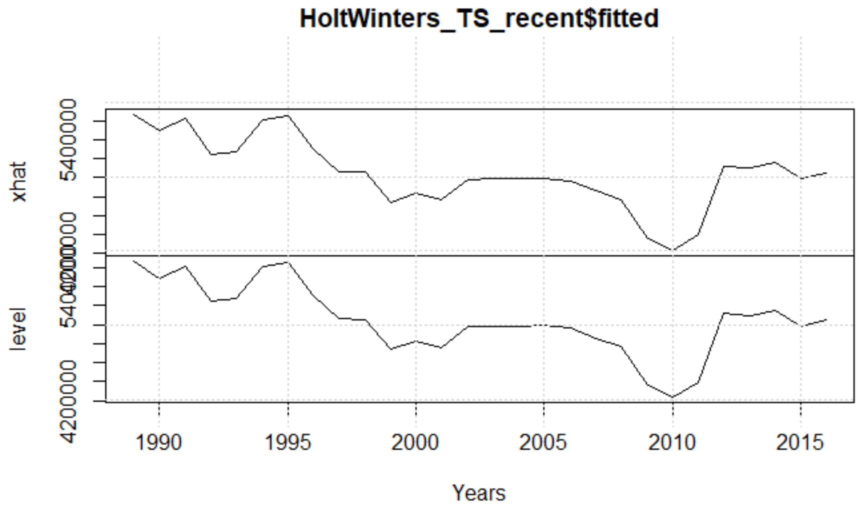
3 rows

Holt-Winter Forecast

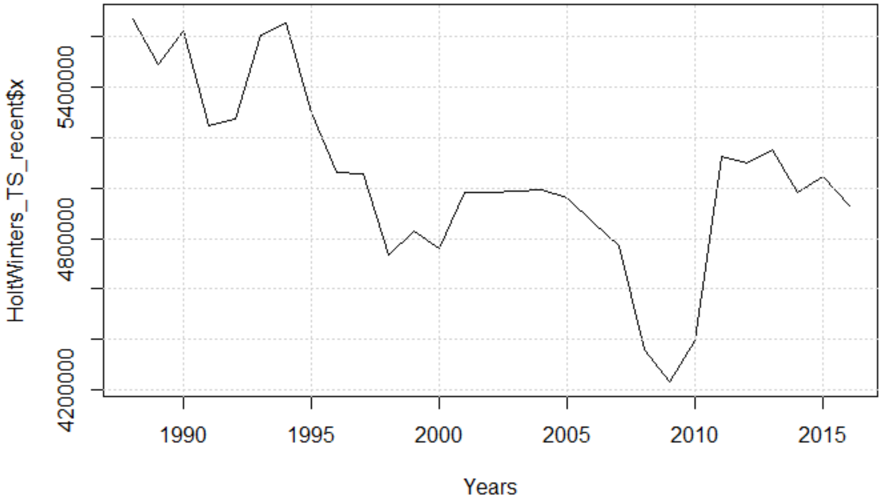


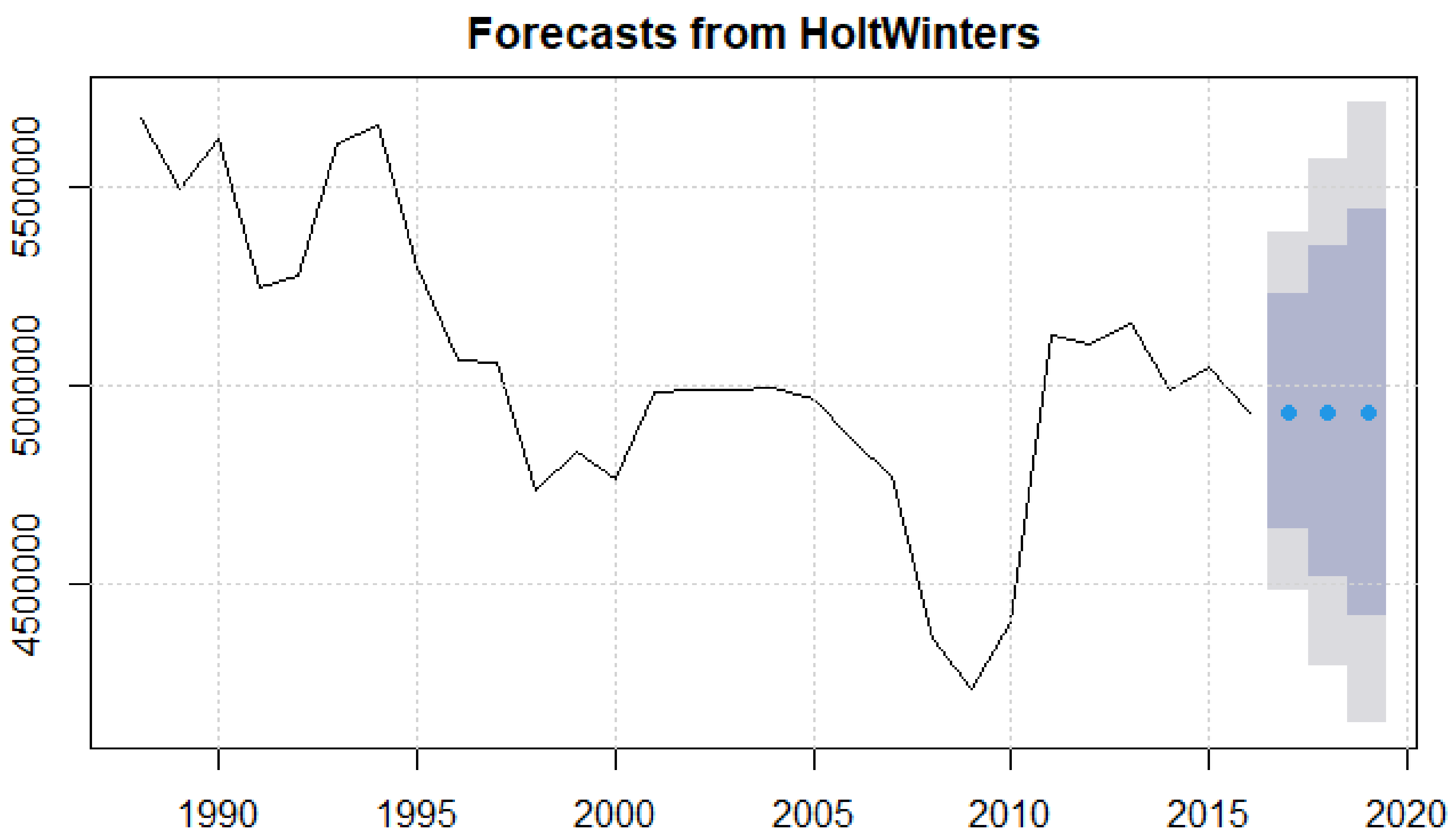
Holt-Winters filtering
Observed / Fitted
42000000
48000000
54000000
1990
1995
2000
2005
2010
2015
Years


Residuals from HoltWinters
750000
500000
250000
0
-250000
1990
1995
2000
2005
2010
2015
0.4
0.2
0.0
-0.2
-0.4
ACF
1
2
3
4
5
6
7
8
9
Lag
8
6
4
2
0
-8e+05
-4e+05
0e+00
4e+05
residuals



- Observed value compared to level
- Since there is no seasonality, both plots are similar




Forecasts from HoltWinters
4500000
5000000
5500000
1990
1995
2000
2005
2010
2015
2020

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2017	4931025	4635287	5226762	4478733	5383316
2018	4931025	4512802	5349247	4291409	5570640
2019	4931025	4418815	5443234	4147667	5714382

3 rows

Regression

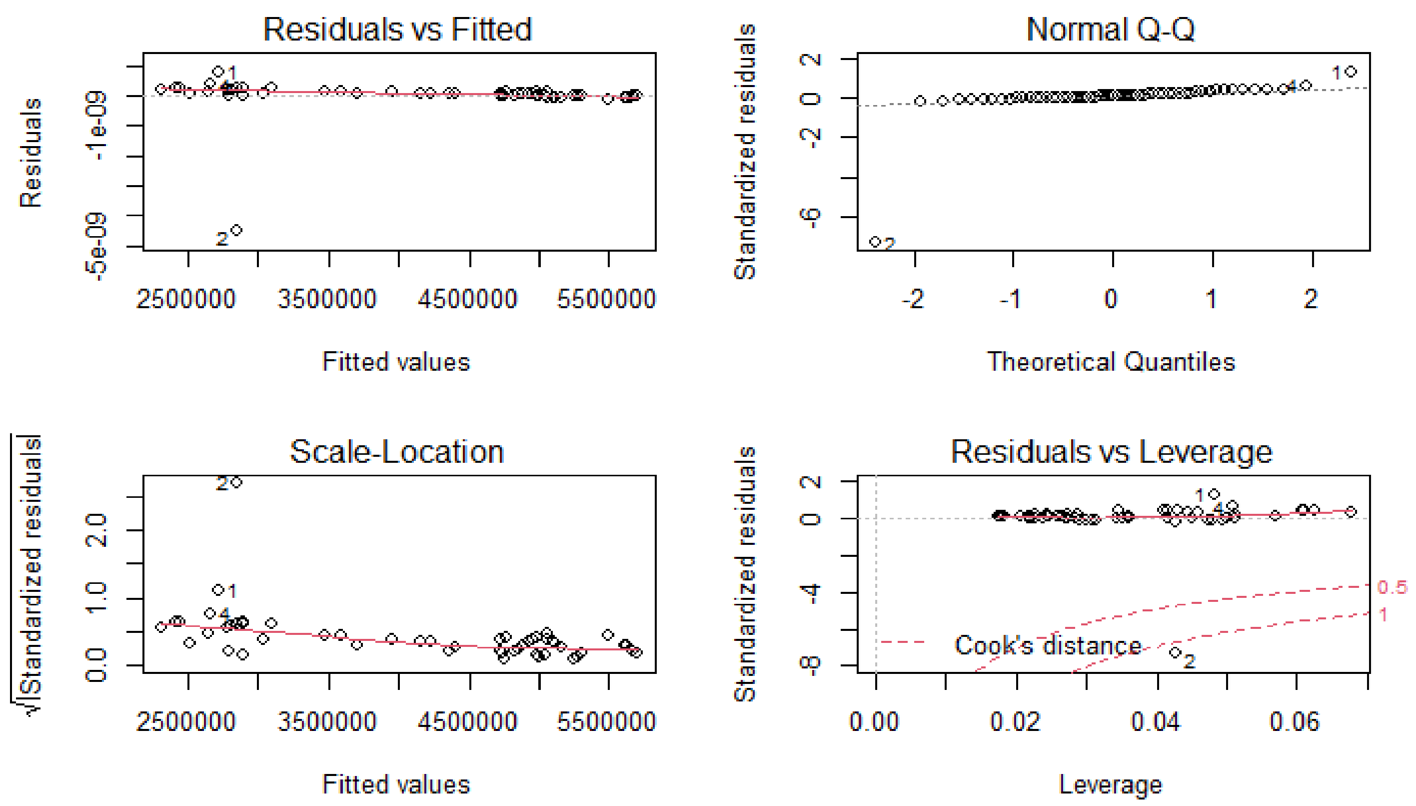
- Recent value is considered without any change, the current value will differ by -1.974e-09. For every increase of one metric ton in lag1, the current value will increase by 1.000e+00.
- The adjusted R-squared value is 1.

```
essentially perfect fit: summary may be unreliable
Call:
lm(formula = FP_US$`Capture fisheries production (metric tons)` ~ 
    lag(FP_US$`Capture fisheries production (metric tons)`, +1), 
    data = FP_US)

Residuals:
       Min         1Q     Median         3Q        Max 
-4.503e-09 -1.580e-11  5.160e-11  1.264e-10  7.671e-10 

Coefficients:
                                                              Estimate Std. Error    t value Pr(>|t|)    
(Intercept)                                                 -1.974e-09  3.268e-10 -6.039e+00 1.39e-07 ***
lag(FP_US$`Capture fisheries production (metric tons)`, +1)  1.000e+00  7.575e-17  1.320e+16  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.29e-10 on 55 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 1.743e+32 on 1 and 55 DF,  p-value: < 2.2e-16
```


Residuals vs Fitted
Residuals
Fitted values
-1e-09
-5e-09
2500000
3500000
4500000
5500000
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance
0.5
1
0.00
0.02
0.04
0.06

ARIMA

- This gives the d value in the Arima function. It tells you the number of differences that you should take to make the time series stationary.

```
ndiffs(FP_US_TS_recent)
```

```
[1] 1
```

- Best model is the one with lowest AIC
- (0,1,0)

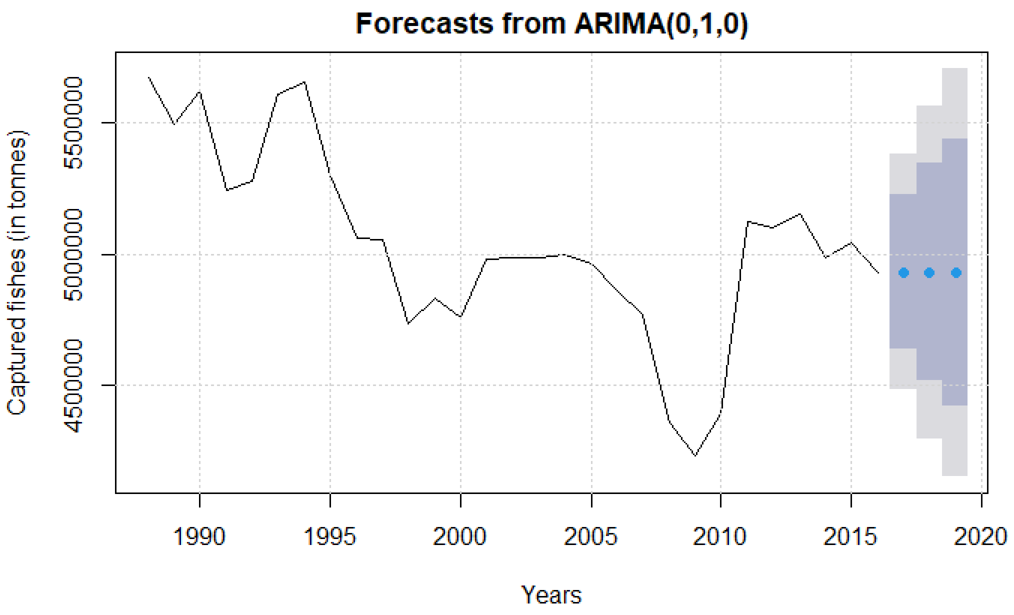
```
ARIMA(0,1,0)                    : 772.5265
ARIMA(0,1,0) with drift         : 774.4747
ARIMA(0,1,1)                    : 774.6563
ARIMA(0,1,1) with drift         : 776.8304
ARIMA(0,1,2)                    : 777.1605
ARIMA(0,1,2) with drift         : 779.5101
ARIMA(0,1,3)                    : 775.9411
ARIMA(0,1,3) with drift         : Inf
ARIMA(0,1,4)                    : 778.0335
ARIMA(0,1,4) with drift         : 780.1718
ARIMA(0,1,5)                    : Inf
ARIMA(0,1,5) with drift         : Inf
```

```
Best model: ARIMA(0,1,0)
```

```
Series: FP_US_TS_recent
ARIMA(0,1,0)

sigma^2 estimated as 5.205e+10:  log likelihood=-385.19
AIC=772.37   AICc=772.53   BIC=773.7
```

- Arima forecast plot



- Naïve and Arima Forecast.

	Point Forecast <dbl>	Lo 80 <dbl>	Hi 80 <dbl>	Lo 95 <dbl>	Hi 95 <dbl>
2017	4931017	4638643	5223391	4483870	5378164
2018	4931017	4517538	5344496	4298656	5563378
2019	4931017	4424611	5437423	4156536	5705498

3 rows

- Holts – Winter Forecast

	Point Forecast <dbl>	Lo 80 <dbl>	Hi 80 <dbl>	Lo 95 <dbl>	Hi 95 <dbl>
2017	4931025	4635287	5226762	4478733	5383316
2018	4931025	4512802	5349247	4291409	5570640
2019	4931025	4418815	5443234	4147667	5714382

3 rows

- Exponential Smoothing Forecast

	Point Forecast <dbl>	Lo 80 <dbl>	Hi 80 <dbl>	Lo 95 <dbl>	Hi 95 <dbl>
2017	4931029	4626568	5235489	4465396	5396661
2018	4931029	4500478	5361579	4272558	5589499
2019	4931029	4403723	5458334	4124584	5737473

3 rows

```
SMA_forecast
```

- Simple Moving Average Forecast

	Point Forecast <dbl>	Lo 80 <dbl>	Hi 80 <dbl>	Lo 95 <dbl>	Hi 95 <dbl>
2016	4986988	4804071	5169905	4707240	5266735
2017	4986988	4728317	5245659	4591385	5382591
2018	4986988	4670187	5303788	4502483	5471492
2019	4986988	4621181	5352794	4427535	5546441

4 rows

- If you notice, we are getting same point forecast value for all three years within the model. This is because our data is not seasonal and cyclic. So, it will simply take avg and print same result for all forecast.

Best Model

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Naïve	-26416.04	228140.4	160037.6	-0.6049683	3.205752	1	0.07203671
Moving Average	-22583.83	137342.9	108031.7	-0.467046	2.197547	0.9638667	0.5752261
Holt-Winters	-26417.51	228141.6	160036.7	-0.6050042	3.205737	0.9999942	0.07209646
Decomp	-16612.05	229233.4	163411.2	-0.4272953	3.252019	1.02108	0.03910814
ARIMA	-25309.6	224174.9	154714.6	-0.5806591	3.098657	0.9667391	0.06850866
Random Walk	-26416.04	228140.4	160037.6	-0.6049683	3.205752	1	0.07203671

- We are using MAPE as a measure of accuracy.
- Since the lowest MAPE value is of Moving average, we consider it as the best forecasting model.
- This was expected as Moving average works best when recent observations are better than all observations.
- Since, we saw correlation between the residuals in Acf for SMA, we choose some other model.
- The second best accuracy is for Arima.
- Final prediction :

```
MAPE <- 5
best_accuracy[1] <- naive_accuracy[MAPE]
best_accuracy[2] <- SMA_accuracy[MAPE]
best_accuracy[3] <- HoltWinters_accuracy[MAPE]
best_accuracy[4] <- ets_accuracy[MAPE]
best_accuracy[5] <- Arima_accuracy[MAPE]
best_accuracy[6] <- rwf_accuracy[MAPE]
```

```
best_accuracy
```

```
[1] 3.205752 2.197547 3.205737 3.252019 3.098657 3.205752
```

```
best_accuracy_MAPE = min(best_accuracy)
best_accuracy_MAPE
```

```
[1] 2.197547
```

	Point Forecast <dbl>	Lo 80 <dbl>	Hi 80 <dbl>	Lo 95 <dbl>	Hi 95 <dbl>
2017	4931017	4638643	5223391	4483870	5378164
2018	4931017	4517538	5344496	4298656	5563378
2019	4931017	4424611	5437423	4156536	5705498

3 rows