Examining and Forecasting Masters Tournament Quality through Time Series Analysis

Maddox Johnston
Math 561: Dr. Jin-Hong Park
College of Charleston

Introduction

Background

The Masters Tournament is an esteemed professional golf championship, first established in 1934. The Masters Tournament, colloquially referred to as "The Masters", is unique in the fact that it always occurs at the same location (Augusta National Golf Club in Augusta, Georgia) on roughly the same dates (Thursday-Sunday on the first full week of April). The golf course the Masters is played on has seen some alterations over the years, namely the lengthening of certain holes to account for golfers increasingly able to drive the ball longer distances (Whitten, 2021). However, despite these alterations, the Total Score[1] of competitors has remained largely unaffected in years following alterations designed to increase difficulty (Whitten, 2021). In fact, it is apparent that both the Total Score of the Tournament Winner and the Average Total Score of the Field[2] have steadily decreased over time.

Motivation

The standardized conditions of The Masters coupled with the breadth of historical data make these general trends meaningful. This project seeks to fit certain data from past Masters Tournament results to time series models to both analyze tournament quality & competitiveness and forecast future tournament results. The statistics this project focuses on are Total Score of the Tournament Winner, Average Total Score of the Field, Strokes Gained[3] of the Tournament Winner, and the Variance of the Total Scores of Competitors surviving the final cut.

Data Preparation

All data was gathered from pgatour.com, which has full results of every Masters Tournament that has ever occurred. However, these were simply individual webpages for each Tournament year, and I was not able to find any existing dataset for what I was interested in. I was forced to scrape each year of data and format it into its own dataset, concatenate these datasets into one "superdataset", and then calculate the statistics of interest for each year before entering them into the final dataset I created, named "MastersData".

	A	B	C	D	E	F	G	H	I
1	YEAR	WINNER_SCORE	AVG_SCORE_Top45	MEDIAN SCORE_Top45	VARIANCE_SCORE_Top45	STROKES_GAINED		*Note: Top 45 players were used due to	
2	1946	282	299.22	300	71.31	17.22		varying numbers of players making the cut	
3	1947	281	293.68	294.5	42.54	12.68		over the years.	
4	1948	279	298.14	297	68.16	19.14			
5	1949	282	297.7	298	43.47	15.7			
6	1950	283	299.88	300.5	46.57	16.88			

Figure 1: The first 5 rows of MastersData

I elected to only include data beginning in the year 1946 since the Tournament did not occur in 1943-1945 due to World War II. As such, the final dataset is 76 rows with a yearly grain, and six columns, consisting of a "Year" column and the rest containing the statistics described in the Introduction: Total Score of the Tournament Winner, Average Total Score of the

[1] Total Score is the stroke total through four rounds. In golf, the winner is that with the lowest Total Score.
[2] The Average Total Score of the Field is the average stroke total of the golfers who play in all four rounds (certain golfers are "cut" in the preceding rounds, and not allowed to participate in the following rounds). Since differing numbers of players compete in the final round over the years, we compute this statistic using only the top 45 finishers.
[3] Strokes Gained is simply computed as: (Total Score) – (Average Total Score of the Field)

Field, Strokes Gained of the Tournament Winner, and the Variance of the Total Scores of Competitors surviving the final cut. The Median Score of the Field statistic was calculated, but not used[4].

It is worth noting that the Sample Variance was used since only data from the Top 45 finishers in the final round was considered due to varying numbers of players making the cut to play in the final round yearly.

Data Analysis

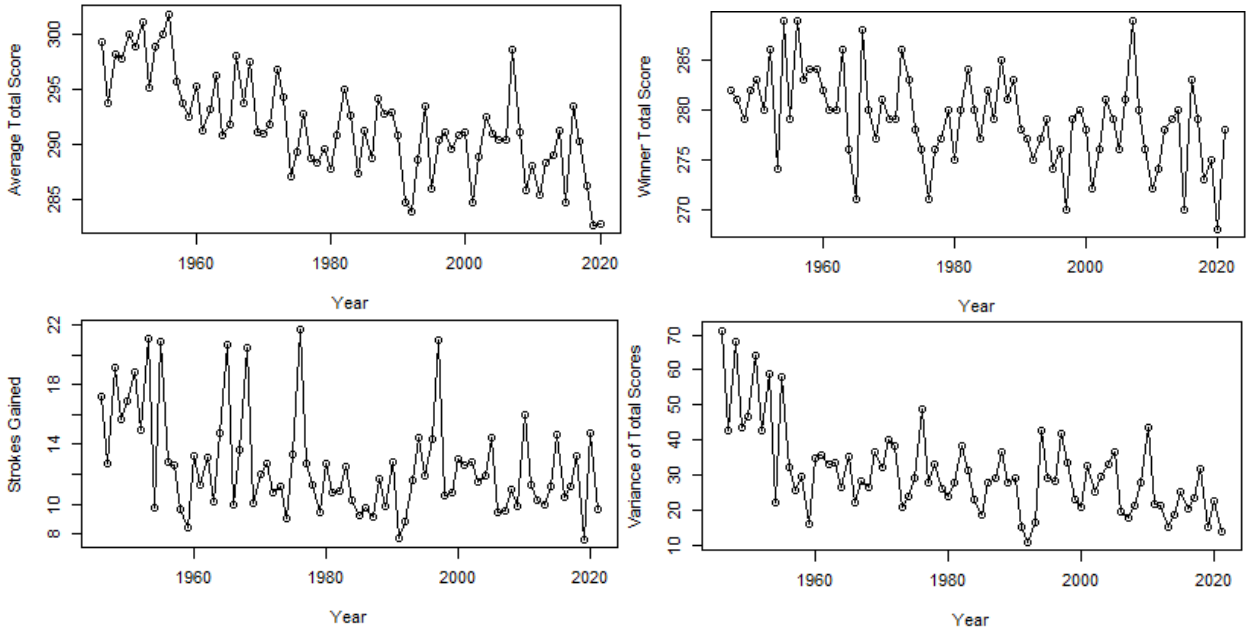


***Figure 2**: (Clockwise, from top left):* (*Top Left*) Yearly Average Total Score, (*Top Right*) Yearly Total Score of the Tournament Winner, (*Bottom Left*) Variance of Total Scores by Year, and (*Bottom Right*) Strokes Gained by Year

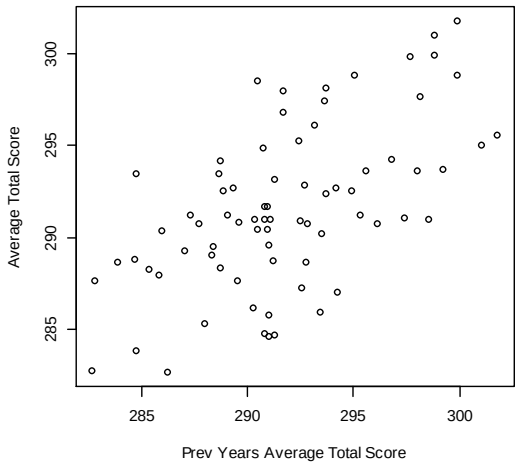
Yearly time series plots of the four statistics of interest are shown above. All four of these time series are presumed nonstationary due to the common characteristic of a decreasing mean over time- a presumption confirmed by the Augmented Dickey-Fuller test[5].This trend is intuitive when we consider the fact that the Masters has become more competitive in recent years than it was in the early 1950's: all golfers are performing better, on average, and the gap between the tournament winner and the field is closer (evident by both the Variance of the Total Scores and Strokes Gained decreasing over time).

[4] The mean was used in favor of the median since the data was not consistently skewed either way, and the two measures exhibited very similar trends.

[5] ADF Test p-values of WinnerTs, AvgTs, StrokesTs, and VarianceTs, respectively: 0.056, 0.088, 0.124, 0.0931. These are not small enough to lead us reject the null hypothesis of nonstationarity.

We also notice that points in our time series tend to "hang together"- there seems to be fairly strong lag 1 autocorrelation. This strong positive lag 1 autocorrelation between the Average Total Score, and the Previous Year's Average Total Score is exemplified in Figure 3 below, a fact that we will keep in mind for model selection.

Figure 3 (*below****)****: Average Total Score vs. Previous Years Average Total Score***



It is evident that transformations are necessary to make our time series stationary. A difference transformation is preferred, since only considering the discrepancy between consecutive years tournaments will eliminate long term historical trends.

First differencing transformations were succesful in making all four of time series stationary, confirmed again by the Dickey-Fuller test[6]. Content with our stationary first differenced models, no more transformations are deemed necessary. Figure 4 below highlights two of the transformed time series, however all four again exhibit similar trends.

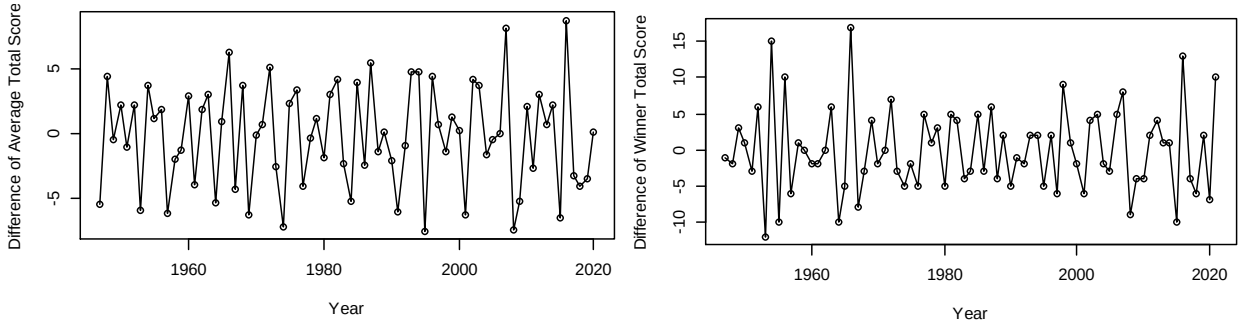
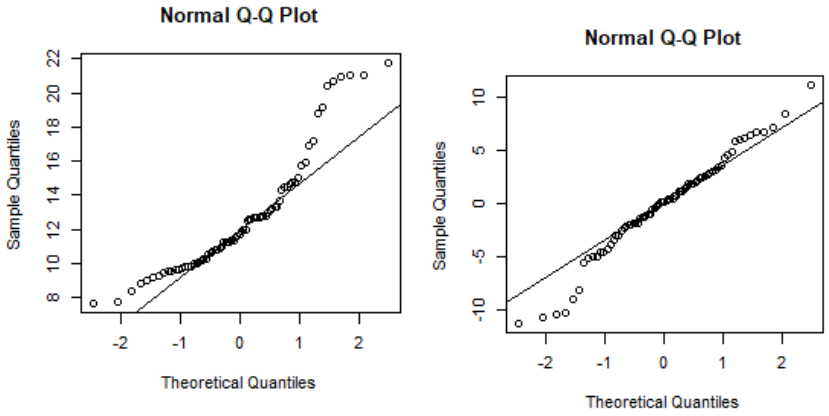


Figure 4*: (Left)* Yearly Difference of Average Total Score, *(Right)* Yearly Difference of Total Score of the Tournament Winner



We also note that the differencing transformation improves the normality of our data dramatically, as seen for Strokes Gained in Figure 5.

Figure 5*: (Left*) Normal Q-Q Plot of Strokes Gained, *(Right*) Normal Q-Q Plot of Difference of Strokes Gained

[6] ADF test results for all four differenced time series yielded p-values <0.01, indicating strong evidence against the null hypothesis of nonstationarity.

Content with our now stationary transformed time series, we now move onto parameter estimation to fit models to our data. We first focus on our transformed "WinnerTs", the first difference of the Yearly Total Scores of the Tournament Winner. We notice in Figure 6 below that the ACF cuts off after lag 1, and the PACF dies off after lag 1. This seems to indicate that a MA(1) model would be appropriate for this differenced data.

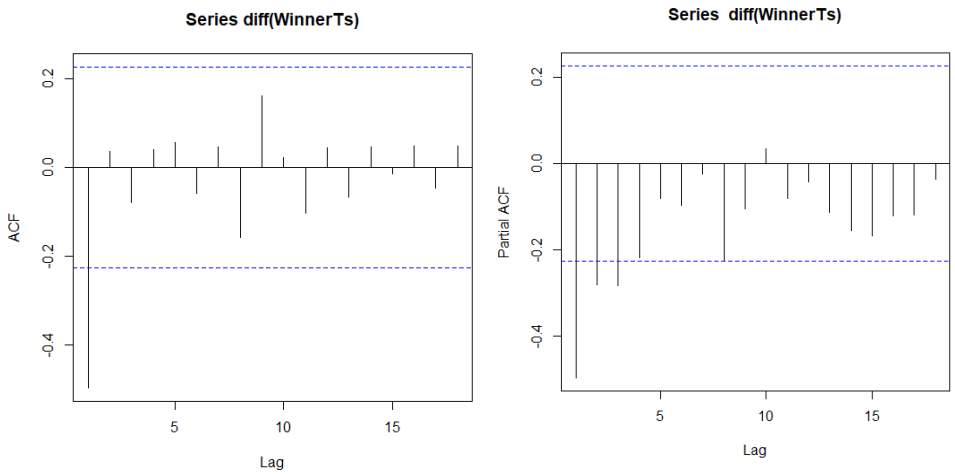


Figure 6: (*Left)* ACF and (*Right)* PACF tables of Differenced WinnerTs.

Figure 7: EACF of Differenced WinnerTs.

AR/MA

	0	1	2	3	4	5	6	7	8	9	10
0	x	o	o	o	o	o	o	o	o	o	o
1	x	x	o	o	o	o	o	o	o	o	o
2	x	x	x	o	o	o	o	o	o	o	o
3	x	o	x	x	o	o	o	o	o	o	o
4	x	x	o	o	x	o	o	o	o	o	o
5	x	x	o	o	x	o	o	o	o	o	o
6	o	x	o	o	x	o	o	o	o	o	o
7	o	x	o	x	x	o	o	o	o	o	o

A look at the EACF table shown in Figure 7 to the left supports evidence toward an MA(1) model, as well as suggesting possibly looking into an ARMA(3,1) model.

Now we shift our attention to our transformed "StrokesTs", first difference of Yearly Strokes Gained by the Tournament Winner. The ACF and PACF are showin in Figure 8 below. We notice how the ACF cuts off after lag 1, but has a significant autocorrelation at lag 4. The PACF seems to tail off after lag 1, and could also resemble a dampened sinusoidal, but is not clear cut. It seems we should entertain a MA(q) or ARMA(p,q) model.

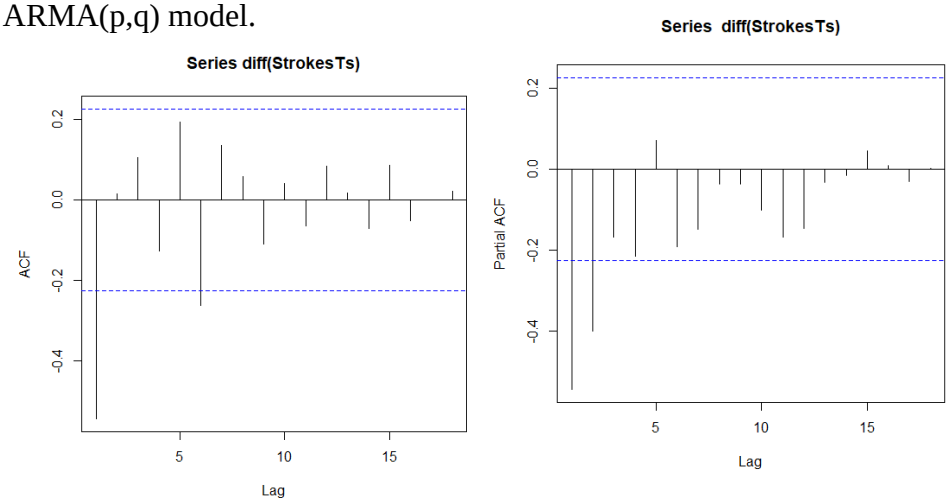


Figure 8: *(Left)* ACF and *(Right)* PACF of Differenced StrokesTs.

Similar trends arise from examining the ACF and PACF of the first differenced Average Score and Variance of Total Score time series. We now

attempt to fit a variety of different models to the transformed time series in light of this information. Results of fitting various models to our differenced "WinnerTs" and "StrokesTs" are shown below.

Various Models for Differenced WinnerTs							
		MA1					
MA(1)	***Coefficients***	-0.8932				***AIC: 436.81***	
	Standard Error	0.052					
		MA1	MA2				
MA(2)	***Coefficients***	-0.8749	-0.0204			***AIC: 438.77***	
	Standard Error	0.1118	0.1094				
		AR1	AR2	AR3	MA1		
ARMA(3,1)	***Coefficients***	0.0253	0.0356	-0.0358	-0.8986	***AIC: 442.58***	
	Standard Error	0.1299	0.1287	0.1260	0.0656		

Figure 9: The parameter estimates, and AIC scores for various models of our differenced WinnerTs

Various Models for Differenced StrokesTs						
		MA1				
MA(1)	***Coefficients***	-0.8724				***AIC: 394.81***
	Standard Error	0.0661				
		MA1	MA2			
MA(2)	***Coefficients***	-0.9196	0.0643			***AIC: 396.5***
	Standard Error	0.1093	0.1142			
		AR1	MA1			
ARMA(1,1)	***Coefficients***	-0.0855	-0.8416			***AIC: 396.45***
	Standard Error	0.1436	0.0999			
		AR1				
AR(1)	***Coefficients***	-0.5539				***AIC: 413.47***
	Standard Error	0.0966				
		AR1	AR2			
AR(2)	***Coefficients***	-0.762	-0.396			***AIC: 403.12***
	Standard Error	0.1055	0.1077			
		AR1	AR2	AR3		
AR(3)	***Coefficients***	-0.8263	-0.5154	-0.1622		***AIC: 403.18***
	Standard Error	0.1138	0.1357	0.1156		

		AR1	AR2	AR3	AR4	
AR(4)	***Coefficients***	-0.8610	0.6285	-0.3334	-0.2157	***AIC: 401.67***
	Standard Error	0.1125	0.1456	0.1444	0.1135	

Figure 10: The parameter estimates, and AIC scores for various models of our differenced WinnerTs

Here we note that a MA(1) seems to be our best model for our differenced WinnerTs, due to having the lowest AIC score and the other two overfit models having insignicant parameters. However, for our differenced "StrokesTs", the best choice is not as obvious. The MA(1) model has the lowest AIC score, and we note MA(2) and ARMA(1,1) do not seem like good choices do to the presence of insignicant parameters. However, an AR(p) model seems worth further consideration as well; in particular, an AR(4) model has a reasonable AIC score, all the coefficients are signicant, and would capture the lag 4 autocorrelation we noticed in Figure 8. Further, by looking at the ACF of the residuals (see Figure 11 below) and the L-B test results[7], despite a common outlier at lag 6, both MA(1) and AR(4) models look good.

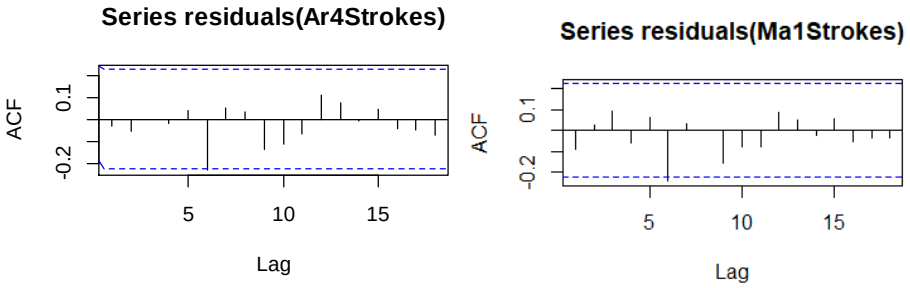
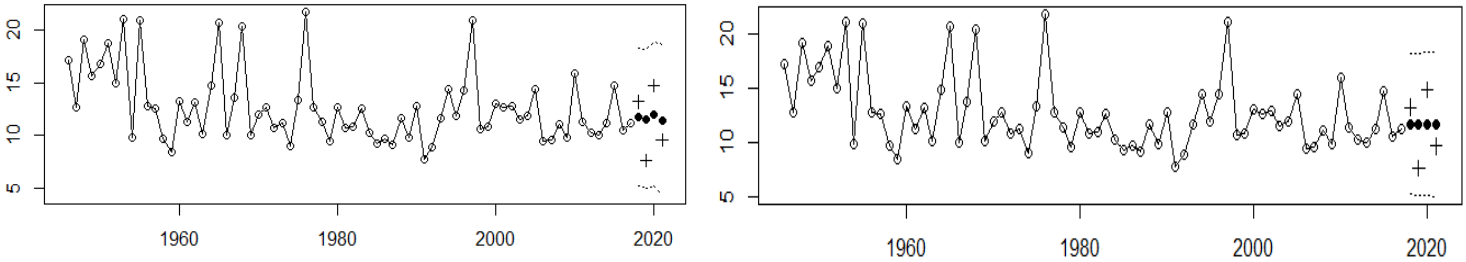


Figure 11: ACF of residuals of AR(4) and MA(1) models for differenced StrokesTs

We will now attempt to use both models to predict in sample data and deduce the best. Figure 12 below shows the results of excluding the last 4 years of data, fitting AR(4) (left) and MA1(right) models to these truncated series, and then attempting to predict the actual values, shown by the '+' symbol. We note that the AR(4) model accurately characterizes the stochastic trend of the actual values. However, a RMSE of 2.66 indicates that this is not the best model. Also, we know that an MA(1) model will only give one predicted value that is not equal to the series mean, so the nature of that plot is not surprising.

Figure 11: Four years of predicted vs. actual in-sample data using AR(4) (*Left*) and MA(1) (*Right*) models for differenced StrokesTs



[7] The L-B test using the AR(4) model for differenced "StrokesTs" yielded a p-value of 0.215, while the MA(1) model yielded a p-value of 0.457. In both cases, we conclude there is no autocorrelation in the residuals.

Now, we repeat the same procedure, except only excluding one year of data (2021). Both the AR(4) model and MA(1) model correctly predict a decrease in Strokes Gained the next year although the MA(1) model gets a little closer, as shown in Figure 13 below.

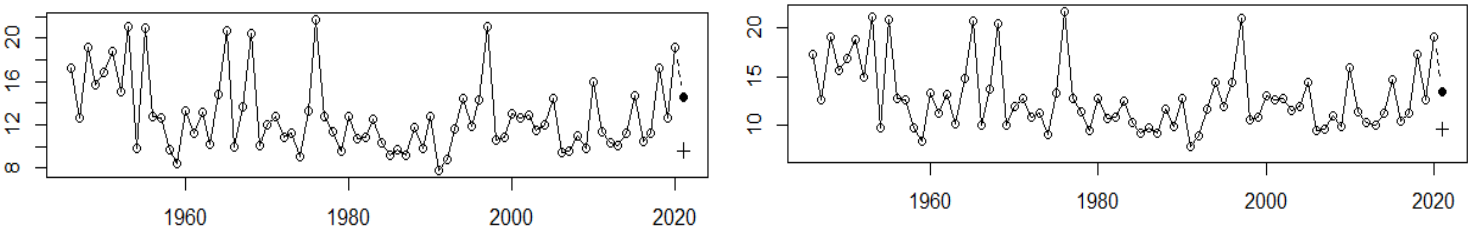


Figure 12: One year of predicted vs. actual in-sample data using AR(4) (*Left*) and MA(1) (*Right*) models for differenced StrokesTs

We conclude that out of the models we tried, the MA(1) model fits our differenced 'StrokesTs' the best, although it does not offer much predictability outside of one year ahead. However, turning our attention back to the differenced "WinnerTs", we note that by excluding data from 2021, our MA(1) model comes extremely close to predicting the true score of the 2021 Masters Winner, as shown in Figure 13 below (The MA(1) model predicted 276, and the 2021 Tournament Winner scored 278).

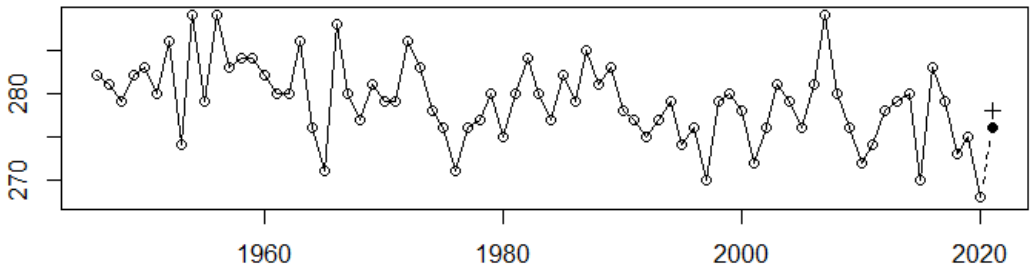


Figure 13: Predicted vs. Actual Total Score using and MA(1) model for differenced WinnerTs

Finally, we forecast the Winner Total Score, Average Total Score, Strokes Gained, and Variance of Scores of Competitors surviving the final cut for the Masters Tournament in 2022 using our fit models. Below, Figure 14 summarizes these results, and Figure 15 show plots highlighting the predicted values for the Tournament Winner (left), and the Variance of Scores of Competitors (right) in 2022. We note how our predicted values seem in accordance with the trends.

2022 Masters Tounament Predicted Values			
Total Score	Average Total Score	Strokes Gained	Variance of Scores
276.3	286.58	10.3	20.41

Figure 14: Summary of forecasted statistics for the 2022 Masters. Predicted using a differenced MA(1) model.

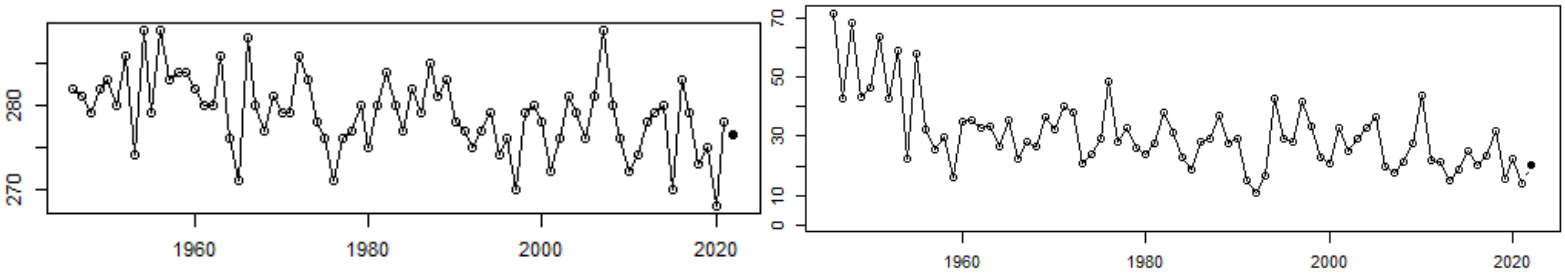


Figure 15: *Left*: Yearly total scores for the Masters tournament winner, with a forecasted value for 2022 (denoted by a solid point). *Right*: Yearly variance of scores of competitors surviving the cut, with a forecasted value for 202.2

Conclusion

Our data points toward a common conjecture around the golf world: golfers are getting better, and the Masters is becoming more competitive. We notice decreasing mean over time in all the statistics we are interested in, and our forecasted values for the 2022 Masters are in line with this tendency. In particular, we can look to the decreasing Total Score of the winner and average Total Score of golfers surviving the cut to see how golfers are performing better despite the course becoming more difficult. Similarly, decreasing Strokes Gained of the winner and variance of the Total Scores of competitors surviving the final cut highlights how The Masters is becoming more competitive.

Another interesting trend worth commenting on is the strong lag 1 autocorrelation shown in Figure 3 (see page 4) between the average total score, and previous year's average total score. This seems intuitive; The Masters often has the same competitors year to year, and the course only changes sporadically. It seems there is predictive value here.

Although it was difficult to model this data in a way conducive to forecasting more than one year out, after transforming the various time series, the MA(1) models seems to capture the essence of this trend present in the data. Future work could go into finding a better predictive model for this data I have created.

Appendix: R Code

```
> MastersData <- read.csv(file.choose())
> attach(MastersData)
> AvgTs <- ts(AVG_SCORE_Top45, start=c(1946), end=c(2021), frequency=1)
> WinnerTs <- ts(WINNER_SCORE, start=c(1946), end=c(2021), frequency=1)
> StrokesTs <- ts(STROKES_GAINED, start=c(1946), end=c(2021), frequency=1)
> VarianceTs <- ts(VARIANCE_SCORE_Top45, start=c(1946), end=c(2021), frequency=1)
> adf.test(AvgTs); adf.test(WinnerTs) ; adf.test(StrokesTs); adf.test(VarianceTs);
> win.graph(width=4.5,height=2.8,pointsize=8)
> plot(WinnerTs,type='o',ylab='Winner Total Score',xlab='Year')
> plot(AvgTs, type='o',ylab='Average Total Score',xlab='Year')
> plot(StrokesTs, type='o',ylab='Strokes Gained',xlab='Year')
> plot(VarianceTs, type='o',ylab='Variance of Total Scores',xlab='Year')
>win.graph(width=4,height=4,pointsize=8)
>  plot(x=zlag(AVG_SCORE_Top45),y=AVG_SCORE_Top45,xlab="Prev Years Average Total Score",ylab="Average Total
Score")> plot(diff(WinnerTs),type='o',ylab='First Difference of Winner Total Score',xlab='Year')
> plot(diff(AvgTs),type='o',ylab='First Difference of Average Total Score',xlab='Year')
> acf(diff(WinnerTs)); pacf(diff(WinnerTs))
> eacf(diff(AvgTs))
> adf.test(diff(AvgTs)); adf.test(diff(WinnerTs)) ; adf.test(diff(StrokesTs)); adf.test(diff(VarianceTs));
> arima(WinnerTs,c(0,1,1)); arima(WinnerTs,c(0,1,2)); arima(WinnerTs,c(3,1,1)); arima(StrokesTs, c(0,1,1)); arima(StrokesTs,
c(0,1,2));  arima(StrokesTs, c(1,1,1)); arima(StrokesTs, c(1,1,0)); arima(StrokesTs, c(2,1,0)); arima(StrokesTs, c(3,1,0));
arima(StrokesTs, c(4 ,1,0))
> Ar4Strokes=arima(StrokesTs, c(4,1,1)); Ma1Strokes=arima(StrokesTs, c(0,1,1))
> acf(residuals(Ar4Strokes)); acf(residuals(Ma1Strokes))
> actualAR=window(STROKES_GAINED,start=73) ;  seriesAR=window(STROKES_GAINED,end=72)
>ARStrokesTs <- ts(seriesAR, start=c(1946), end=c(2017), frequency=1)
>MAlag4StrokesTs <- ts(seriesAR, start=c(1946), end=c(2017), frequency=1)
>ARStrokesModel=arima(ARStrokesTs,c(4,1,0))
> MAlag4StrokesModel=arima(ARStrokesTs,c(0,1,1))
>  resultAR=plot(ARStrokesModel,n.ahead=4,ylab='Series, Forecasts, & Actual Values', ,pch=19)
> points(x=2018:2021,y=actualAR,pch=3)
>resultMAlag4=plot(MAlag4StrokesModel,n.ahead=4,ylab='Series, Forecasts, & Actual Values', ,pch=19)
> points(x=2018:2021,y=actualAR,pch=3)
>  actualMA=window(STROKES_GAINED,start=76); seriesMA=window(STROKES_GAINED,end=75)
> MAStrokesTs <- ts(seriesMA, start=c(1946), end=c(2020), frequency=1)
> MAStrokesModel=arima(MAStrokesTs,c(0,1,1))
>resultMA=plot(MAStrokesModel,n.ahead=1,type='b',ylab='Series ,Forecast, & Actual Value',pch=19)
> points(x=2021,y=actualMA,pch=3)
>actualAverageMA=window(AVERGAGE_SCORE_Top45,start=76)
> seriesAverageMA=window(AVERAGE_SCORE_Top45,end=75)
>MAAverageTs <- ts(seriesAverageMA, start=c(1946), end=c(2020), frequency=1)
> MAAverageModel=arima(MAAverageTs,c(0,1,1))
> resultMAAverage=plot(MAAverageModel,n.ahead=1,type='b',ylab='Series ,Forecast, & Actual Value',pch=19)
> points(x=2021,y=actualAverageMA,pch=3)
>actualWinnerMA=window(WINNER_SCORE,start=76); seriesWinnerMA=window(WINNER_SCORE,end=75)
>MAWinnerTs <- ts(seriesWinnerMA, start=c(1946), end=c(2020), frequency=1)
> MAWinnerModel=arima(MAWinnerTs,c(0,1,1))
> resultMAWinner=plot(MAWinnerModel,n.ahead=1,type='b',ylab='Series ,Forecast, & Actual Value',pch=19)
> points(x=2021,y=actualWinnerMA,pch=3)
>actualVarianceMA=window(VARIANCE_SCORE_Top45,start=76)
> seriesVarianceMA=window(VARIANCE_SCORE_Top45,end=75)
>MAVarianceTs <- ts(seriesVarianceMA, start=c(1946), end=c(2020), frequency=1)
> MAVarianceModel=arima(MAVarianceTs,c(0,1,1))
> resultMAVariance=plot(MAVarianceModel,n.ahead=1,type='b',ylab='Series ,Forecast, & Actual Value',pch=19)
> points(x=2021,y=actualVarianceMA,pch=3)
>resultMAWinner$pred ; resultMA$pred ; resultAR$pred; resultMAWinner$pred ; resultMAAverage$pred;
resultMAVariance$pred
```

References

Whitten, R. (2021, April 09). A Comprehensive History Of Every Change Made To Augusta National Golf Club. Retrieved from https://www.golfdigest.com/story/complete-changes-to-augusta-national

“Masters Tournament: Past Results.” PGATour.com, PGA Tour, www.pgatour.com/tournaments/masters-tournament/past-results.html.