FIFA Project-Report

Try To Forecasting Around The World

DTSC422: Time series and statistical forecasting
Ali Al Hadi Ayache

EPFL

FIFA Project-Report

Try To Forecasting
Around The World

by

Ali Al Hadi Ayache

Preface

This report for project called FIFA done by ali al hadi ayache student in USAL university for course name Time series and statistical forecasting, may be this report contain some mistakes in writing or in result the accurcy of the mistakes about 5% ,we using R-Studio to extract the results the instructor Dr ali Ezzidine

Ali Al Hadi Ayache
January 2023

Summary

The main of this project to making forcasting for number of goals in year 2022 for Uruguay national team for this we making many steps:

- Read the datasets and Give a brief explanation.

- Specify team and create a new data frame that contains the total number of goals score by every country during every year.

- Plot the number of goals of your team in each year

- Calculate mean and standard deviation of the total number of goals scored.

- Convert your data to time series object.

- Plot autocorrelation and partial autocorrelation function.

- Making differencing to make it stationary.

- Plot autocorrelation and partial autocorrelation function and Identify a couple of ARIMA models that might be useful in describing the time series. Which of your models is the best according to their AIC values.

- Use the ARIMA function to estimate and fit the identified models.

- Forecast the number of goals in year 2022 using your best fitted model.

Contents

1

Introduction

Read the datasets using read.csv() function use two variables to store the read datasets: df/results for results.csv and df/goal.scorers for the goalscorers.csv df/results contains information about every match, including the final score, the tournament and location of the match. Specify your team and filter df/results to show all the matches where "Your team" is the home team and was the winner.

	date	home_team	away_team	home_score	away_score	tournament	city	country	neutral
1	1872-11-30	Scotland	England	0	0	Friendly	Glasgow	Scotland	FALSE
2	1873-03-08	England	Scotland	4	2	Friendly	London	England	FALSE
3	1874-03-07	Scotland	England	2	1	Friendly	Glasgow	Scotland	FALSE
4	1875-03-06	England	Scotland	2	2	Friendly	London	England	FALSE
5	1876-03-04	Scotland	England	3	0	Friendly	Glasgow	Scotland	FALSE
6	1876-03-25	Scotland	Wales	4	0	Friendly	Glasgow	Scotland	FALSE
7	1877-03-03	England	Scotland	1	3	Friendly	London	England	FALSE
8	1877-03-05	Wales	Scotland	0	2	Friendly	Wrexham	Wales	FALSE
9	1878-03-02	Scotland	England	7	2	Friendly	Glasgow	Scotland	FALSE
10	1878-03-23	Scotland	Wales	9	0	Friendly	Glasgow	Scotland	FALSE
11	1879-01-18	England	Wales	2	1	Friendly	London	England	FALSE
12	1879-04-05	England	Scotland	5	4	Friendly	London	England	FALSE
13	1879-04-07	Wales	Scotland	0	3	Friendly	Wrexham	Wales	FALSE
14	1880-03-13	Scotland	England	5	4	Friendly	Glasgow	Scotland	FALSE
15	1880-03-15	Wales	England	2	3	Friendly	Wrexham	Wales	FALSE
16	1880-03-27	Scotland	Wales	5	1	Friendly	Glasgow	Scotland	FALSE

we can see all infomation about the results.csv and what conatin now we need to looking into another dataset.

	date	home_team	away_team	home_score	away_score	tournament	city	country	neutral
1	1872-11-30	Scotland	England	0	0	Friendly	Glasgow	Scotland	FALSE
2	1873-03-08	England	Scotland	4	2	Friendly	London	England	FALSE
3	1874-03-07	Scotland	England	2	1	Friendly	Glasgow	Scotland	FALSE
4	1875-03-06	England	Scotland	2	2	Friendly	London	England	FALSE
5	1876-03-04	Scotland	England	3	0	Friendly	Glasgow	Scotland	FALSE
6	1876-03-25	Scotland	Wales	4	0	Friendly	Glasgow	Scotland	FALSE
7	1877-03-03	England	Scotland	1	3	Friendly	London	England	FALSE
8	1877-03-05	Wales	Scotland	0	2	Friendly	Wrexham	Wales	FALSE
9	1878-03-02	Scotland	England	7	2	Friendly	Glasgow	Scotland	FALSE
10	1878-03-23	Scotland	Wales	9	0	Friendly	Glasgow	Scotland	FALSE
11	1879-01-18	England	Wales	2	1	Friendly	London	England	FALSE
12	1879-04-05	England	Scotland	5	4	Friendly	London	England	FALSE
13	1879-04-07	Wales	Scotland	0	3	Friendly	Wrexham	Wales	FALSE
14	1880-03-13	Scotland	England	5	4	Friendly	Glasgow	Scotland	FALSE
15	1880-03-15	Wales	England	2	3	Friendly	Wrexham	Wales	FALSE
16	1880-03-27	Scotland	Wales	5	1	Friendly	Glasgow	Scotland	FALSE

Based on the previous result and the df/goal.scorers dataframe. the result of the match of your team that took place at 2022-12-02 match in World cup .

	date	home_team	away_team	home_score	away_score	tournament	city	country	neutral
1	2022-12-02	Ghana	Uruguay	0	2	FIFA World Cup	Al Wakrah	Qatar	TRUE

top 5 tournaments:

	tournament	n
1	FIFA World Cup qualification	7788
2	Friendly	5231
3	UEFA Euro qualification	2566
4	FIFA World Cup	950
5	African Cup of Nations qualification	866

Now, we want to create a new data frame that contains the total number of goals score by every country during every year. The table will have the columns team tournament/year goals/total. We will mainly work with df/results.

	year	home_team	sum_home_score
1	1872	Scotland	0
2	1873	England	4
3	1874	Scotland	2
4	1875	England	2
5	1876	Scotland	7
6	1877	England	1
7	1877	Wales	0
8	1878	Scotland	16
9	1879	England	7
10	1879	Wales	0
11	1880	Scotland	10
12	1880	Wales	2
13	1881	England	1
14	1881	Wales	1
15	1882	Northern Ireland	0
16	1882	Scotland	10

merge the two frames based on the country and year then add new column that contain the summation of the sum/away/score and sum/home/score.

Calculate mean and standard deviation of the total number of goals scored.

	year	away_team	sum_away_score
1	1872	England	0
2	1873	Scotland	2
3	1874	England	1
4	1875	Scotland	2
5	1876	England	0
6	1876	Wales	0
7	1877	Scotland	5
8	1878	England	2
9	1878	Wales	0
10	1879	Scotland	7
11	1879	Wales	1
12	1880	England	7
13	1880	Wales	1
14	1881	Scotland	11
15	1881	Wales	1
16	1882	England	17

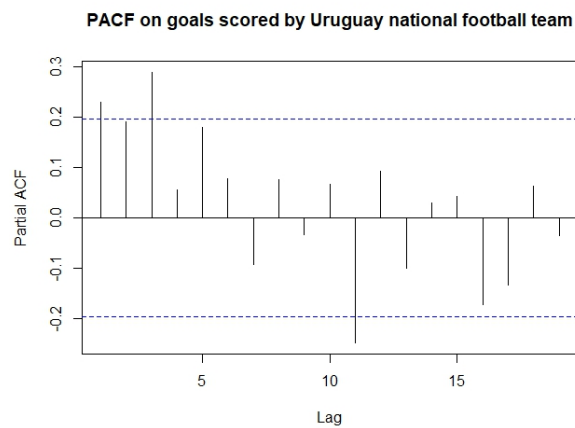|	year	Team	Goals_total
1	1906	Uruguay	1
2	1908	Uruguay	4
3	1909	Uruguay	4
4	1910	Uruguay	14
5	1911	Uruguay	8
6	1912	Uruguay	10
7	1913	Uruguay	4
8	1914	Uruguay	4
9	1915	Uruguay	5
10	1916	Uruguay	17
11	1917	Uruguay	13
12	1918	Uruguay	6
13	1919	Uruguay	20
14	1920	Uruguay	14
15	1921	Uruguay	7
16	1922	Uruguay	7

Uruguay_goals_mean	12.8811881188119
Uruguay_goals_vari…	9.10722606438171

2

R-For-casting

Convert your data to time series object and Plot autocorrelation and partial autocorrelation function.

ACF on goals scored by Uruguay national football team

PACF on goals scored by Uruguay national football team
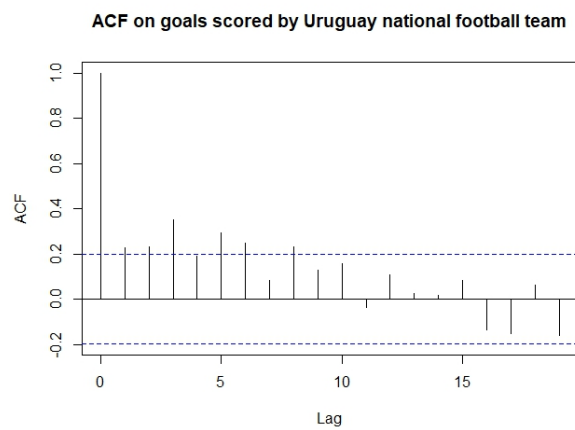
This series are not stationary perform the corresponding transformations and/or differencing to make it stationary

**Uruguay national football team's total goals per year
after applying log and differencing of first order**

**PACF on goals scored by Uruguay national football team
after applying log and differencing of first order**

**ACF on goals scored by Uruguay national football team
after applying log and differencing of first order**
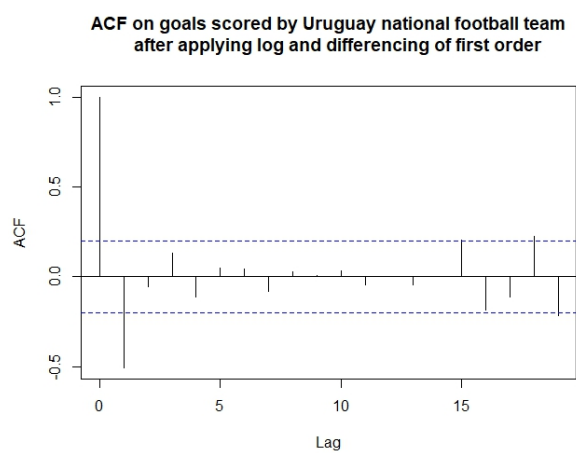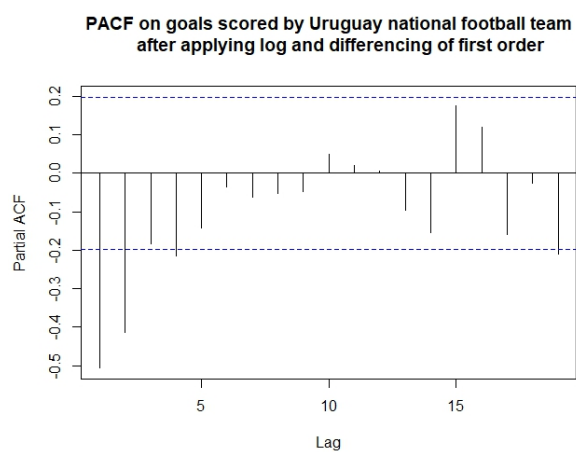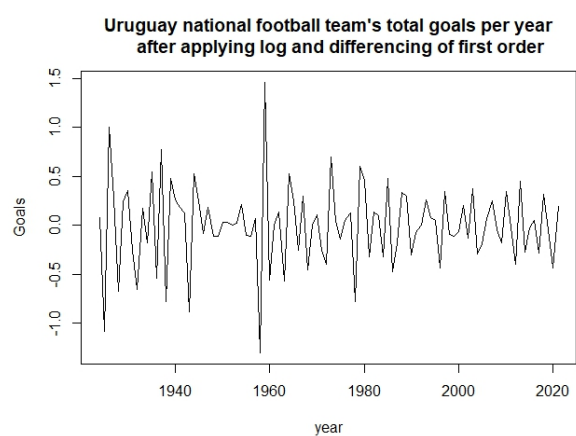
Identify a couple of ARIMA models that might be useful in describing the time series and Use the ARIMA function to estimate and fit the identified models. the models of arima be in last page in R code but the best one ARIMA(0,1,1)

Residuals from ARIMA(0,1,1)
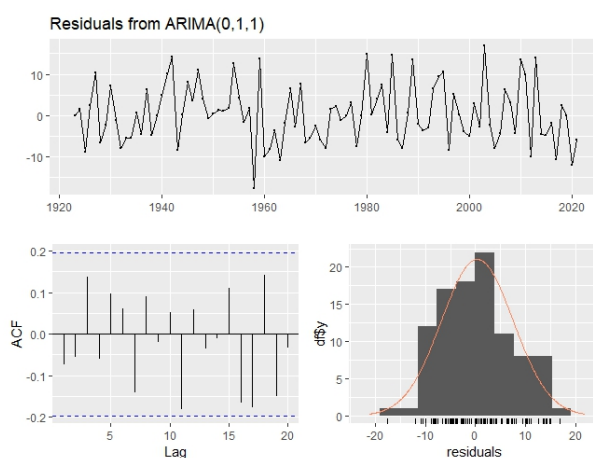
After looking on graphs residuals resemble white noise and Normally distributed.
The results will be in this figure in 2022 about 16 goal while in true Uruguay scoring 20 goals may be the result be not bad.

Uruguay national football team's total goals per year in addition to 2022 prediction.
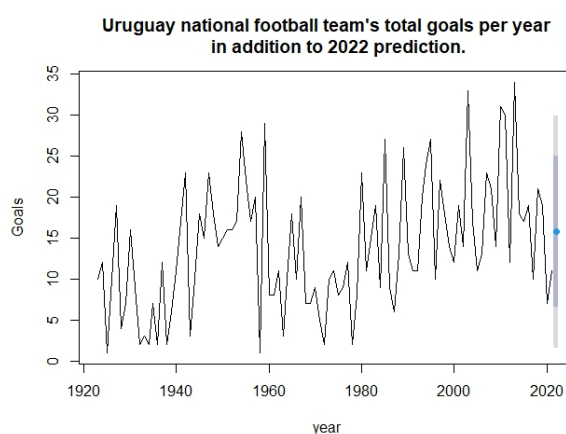
	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2022	15.75645	6.513843	24.99906	1.621101	29.8918

3
Conclusion

As conclusion we using R to extarct the results and making forcasting ,may be the result not 100% but in this dataset and this model it be good , we can make the result more accure but there is no more time.

A
Source Code
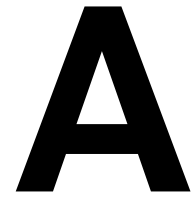
```r
1   #Ex-1
2   #part-a
3   df_results <- read.csv('C:/Users/Hadi Ayache/Desktop/TSPROJECT/PROJECT/results.csv')
4   df_goal.scorers <- read.csv('C:/Users/HadiAyache/Desktop/TSPROJECT/PROJECT/goalscorers.csv')
5   #part-b
6   summary(df_results)
7   summary(df_goal.scorers)
8   #part-c
9   library(dplyr)
10  my_team <- filter(df_results,
11              home_team == 'Uruguay' | away_team == 'Uruguay')
12  my_team_home_wins <-
13      filter(my_team,home_team == 'Uruguay' & home_score > away_score)
14  View(my_team_home_wins)
15  #part-d
16  my_team_1981_12_21 <- my_team %>% filter(date == '2022-11-23')
17  View(my_team_1981_12_21)
18  #part-e
19  match_1981_12_21_scorers <- df_goal.scorers %>%
20      filter(date == '2022-11-23' & home_team == 'Uruguay') %>%
21      .$scorer %>%
22      table() %>%
23      sort() %>%
24      names() %>%
25      tail(1)
26  #part-f
27  #In case question is correct
28  tournaments_in_df_goal.scorers <- df_results %>%
29      filter(.$date %in% df_goal.scorers$date) %>%
30      group_by(tournament) %>%
31      tally(sort = T)
32  top_tournaments_in_df_goal.scorers <-
33      head(tournaments_in_df_goal.scorers, 5)
34  #In case question df_results instead of df_goal.scorers
35  tournaments_in_df_results <- df_results %>%
36      group_by(tournament) %>%
37      tally(sort = T)
38  top_tournaments_in_df_results <-
39      head(tournaments_in_df_results, 5)
40  #####
41  #Ex-2
42  #part-a
43
44  library(lubridate)
45  df_fifa_results <- df_results %>%
46      mutate(year=year(ymd(df_results$date)))
47  df_Uruguay_goals_home_2022 <- df_fifa_results %>%
48      dplyr::filter(home_team == 'Uruguay' & year == 2022) %>%
49      select(home_score) %>%
50      sum()
```

```r
51  df_Uruguay_goals_away_2022 <- df_fifa_results %>%
52      dplyr::filter(away_team == 'Uruguay' & year == 2022) %>%
53      select(away_score) %>%
54      sum()
55  Uruguay_goals_2022 <- df_Uruguay_goals_home_2022 +
56      df_Uruguay_goals_away_2022
57  #part-b
58  df_fifa_results <- df_fifa_results %>%
59      filter(year < 2022)
60  #part-c
61  df_fifa_home_results_grouped <-df_fifa_results %>%
62      group_by(year, home_team) %>%
63      summarise(sum_home_score = sum(home_score), .groups = 'drop') %>%
64      select(year, home_team, sum_home_score)
65
66
67  df_fifa_away_results_grouped <-df_fifa_results %>%
68      group_by(year, away_team) %>%
69      summarise(sum_away_score = sum(away_score), .groups = 'drop') %>%
70      select(year, away_team, sum_away_score)
71
72
73
74  df_fifa_goals <- df_fifa_home_results_grouped %>%
75      inner_join(df_fifa_away_results_grouped,
76                  by=c('year'='year',
77                       'home_team'='away_team'))
78  df_fifa_goals <- df_fifa_goals %>%
79      rename('Team' = 'home_team')
80  df_fifa_goals <- df_fifa_goals %>%
81      mutate(Goals_total =
82                  rowSums(df_fifa_goals [ ,c("sum_home_score", "sum_away_score")])) %>%
83      select(year, Team, Goals_total)
84  df_fifa_goals_Uruguay <- df_fifa_goals %>% filter (Team == 'Uruguay')
85  View(df_fifa_goals_Uruguay)
86  #####
87  #Ex-3
88  fifa_goals_mean <- df_fifa_goals %>% .$Goals_total %>% mean()
89  fifa_goals_variance <- df_fifa_goals %>% .$Goals_total %>% sd()
90  Uruguay_goals_mean <- df_fifa_goals_Uruguay %>%
91      .$Goals_total %>%
92      mean()
93  Uruguay_goals_variance <- df_fifa_goals %>%
94      .$Goals_total %>%
95      sd()
96  df_fifa_goals_Uruguay <- df_fifa_goals %>% filter (Team == 'Uruguay')
97  total_years <- 1923:2021
98  avg <- 0
99  j <- 0
100 temp_df <- data.frame(0,'',0)
101 names(temp_df) <- names(df_fifa_goals_Uruguay)
102 class(temp_df$year) = 'double'
103 class(temp_df$Goals_total) = 'double'
104 for (i in total_years){
105     j <- j + 1
106     if(i %in% df_fifa_goals_Uruguay$year){
107         temp_df[j,] = df_fifa_goals_Uruguay[
108             which(df_fifa_goals_Uruguay$year==i),
109         ]
110     }
111     else{
112         temp_df[j,] = (c(i,'Uruguay',NA))
113     }
114 }
115 class(temp_df$year) = 'double'
116 class(temp_df$Goals_total) = 'double'
117 library(zoo)
118 approx_missing_goals <- na.approx(temp_df$Goals_total)
119 df_approx_missing_goals <- data.frame(year = 1923:2021,
120                                       Goals_total_apx =
121                                           approx_missing_goals)
```

```r
122 df_approx_missing_goals$Goals_total_apx <- round(
123     df_approx_missing_goals$Goals_total_apx)
124 temp_df <- inner_join(temp_df,df_approx_missing_goals) %>%
125     dplyr::select(year, Team, Goals_total_apx)
126 df_fifa_goals_Uruguay <- temp_df
127 df_fifa_goals_Uruguay <- df_fifa_goals_Uruguay %>%
128     rename('Goals_total' = 'Goals_total_apx')
129 View(df_fifa_goals_Uruguay)
130 #####
131 #ARIMA
132 #part-1
133 ts_fifa_Uruguay_goals <- ts(df_fifa_goals_Uruguay$Goals_total,
134                        start = 1923)
135 plot(ts_fifa_Uruguay_goals, xlab='year',ylab='Goals',
136     main = "Uruguay national football team's total goals per year")
137 #part-2
138 acf(ts_fifa_Uruguay_goals,
139     main = 'ACF on goals scored by Uruguay national football team')
140 pacf(ts_fifa_Uruguay_goals,
141     main = 'PACF on goals scored by Uruguay national football team')
142 #part-3
143 ts_fifa_Uruguay_goals_diff <- diff(ts_fifa_Uruguay_goals)
144 plot(ts_fifa_Uruguay_goals_diff, xlab='year',ylab='Goals',
145     main = "Uruguay national football team's total goals per year
146     after applying differencing of first order")
147 ts_fifa_Uruguay_goals_log <- log10(ts_fifa_Uruguay_goals)
148 ts_fifa_Uruguay_goals_log_diff <- diff(ts_fifa_Uruguay_goals_log)
149 plot(ts_fifa_Uruguay_goals_log_diff, xlab='year',ylab='Goals',
150     main = "Uruguay national football team's total goals per year
151     after applying log and differencing of first order")
152 #part-4
153 acf(ts_fifa_Uruguay_goals_log_diff,
154     main = 'ACF on goals scored by Uruguay national football team
155     after applying log and differencing of first order')
156 pacf(ts_fifa_Uruguay_goals_log_diff,
157     main = 'PACF on goals scored by Uruguay national football team
158     after applying log and differencing of first order')
159 #part-5
160 library(forecast)
161 candidate_ar1 <- Arima(ts_fifa_Uruguay_goals_log_diff,
162                        order = c(1,0,0))
163 candidate_ar2 <- Arima(ts_fifa_Uruguay_goals_log_diff,
164                        order = c(2,0,0))
165 candidate_ma1 <- Arima(ts_fifa_Uruguay_goals_log_diff,
166                        order = c(0,0,1))
167 candidate_ma2 <- Arima(ts_fifa_Uruguay_goals_log_diff,
168                        order = c(0,0,1))
169 candidate_models_name <- c('ARIMA(1,0,0)',
170                 'ARIMA(2,0,0)',
171                 'ARIMA(0,0,1)',
172                 'ARIMA(0,0,2)')
173 candidate_models_aicc <- c(candidate_ar1$aicc,
174                         candidate_ar2$aicc,
175                         candidate_ma1$aicc,
176                         candidate_ma2$aicc)
177 df_candidate_models <- data.frame(
178     model = candidate_models_name,
179     aicc = candidate_models_aicc
180 )
181 df_best_candidate_model <- df_candidate_models$model[
182     which.min(df_candidate_models$aicc)]
183 string1 <- 'The best model from our candidates, {ARIMA(1,0,0), ARIMA(2,0,0), ARIMA(0,0,1), ARIMA(0,0,2)
        } according the AICc value is'
184 result <- paste(string1,
185                 df_best_candidate_model)
186 print(result)
187 #part-6
188 bestFit <- auto.arima(ts_fifa_Uruguay_goals,
189                        trace = TRUE,
190                        approximation = FALSE,
191                        lambda = TRUE
```

```r
192                             )
193 #part-7
194 string2 <- 'The best model suggested by auto.arima() is'
195 paste(string2,
196       bestFit)
197 View(bestFit)
198 tsdiag(bestFit)
199 checkresiduals(bestFit)
200 #part-8
201 predicted_goals_2022 <- forecast(bestFit, 1)
202 string2 <- 'The best model suggested that Uruguay will score'
203 string3 <- 'in 2022.'
204 paste(string2,
205       round(tail(predicted_goals_2022$fitted,1)[1]),
206       string3)
207 plot(predicted_goals_2022, xlab='year',ylab='Goals',
208      main = "Uruguay national football team's total goals per year
209      in addition to 2022 prediction.")
```