Soccer championship analysis using Monte Carlo simulation

Isa Castro

University of Neuchâtel, Switzerland

May 2019

Introduction

- Statistics in Sport => Soccer
- Soccer is the most famous sport specially in Brazil.
- In soccer championship, there are two stages:
  a. Classificatory stage
  b. Playoffs = 8 best teams were classified while last 4 teams were relegated

Goal of the article

- Simulate a model to generate an entire championship (by MC simulation) and try to obtain estimators to find the final ranking that will determine which teams will be classified for the playoffs or relegated to a lower rank competition.

Assumptions

1. Equality among the teams : no team has more probability than another to win the game
2. Independence: each game is independent to another one
3. The probability that a game ends up in a draw is the same for all games.

General ideas for the simulation

1. Simulate randomly the number of points for each game :
   a. if a team win = the winner obtains 3 points.
   b. if the team lose = the loser obtain no points.
   c. If there is a draw = each team obtain 1 point.
2. Calculate the cumulative number of points for each team
3. We will rank them

Monte Carlo simulation

1. Determine the number of teams.
2. Generate random number between 0 and 1 (probabilities) according to the number of teams –> we will store it in the matrix.
3. We assume only a part of our matrix = when (i>j) -> to know how many points each team will obtain. Remark : The result of a game, will determine the result of his opponent
4. Dont take in account the diag (team1 x team1) and we will assign for that 0 points.
5. Then, we fill another matrix if the number of point for each team per game according to the probability that the match ends up in a draw.
6. Finally, we will sum each column of the matrix for each team.
7. Rank them

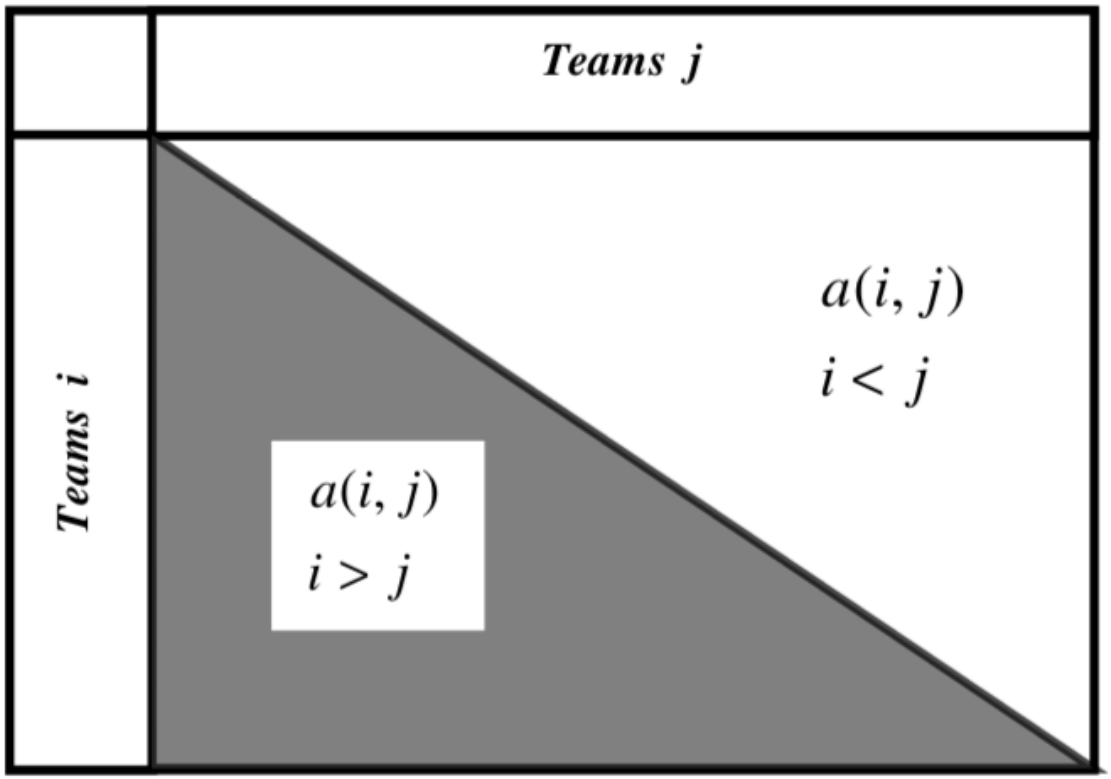


Figure 1: Representation of the Results Matrix, A(M,N)

1

[1] Article:

Validation and Exploring the results

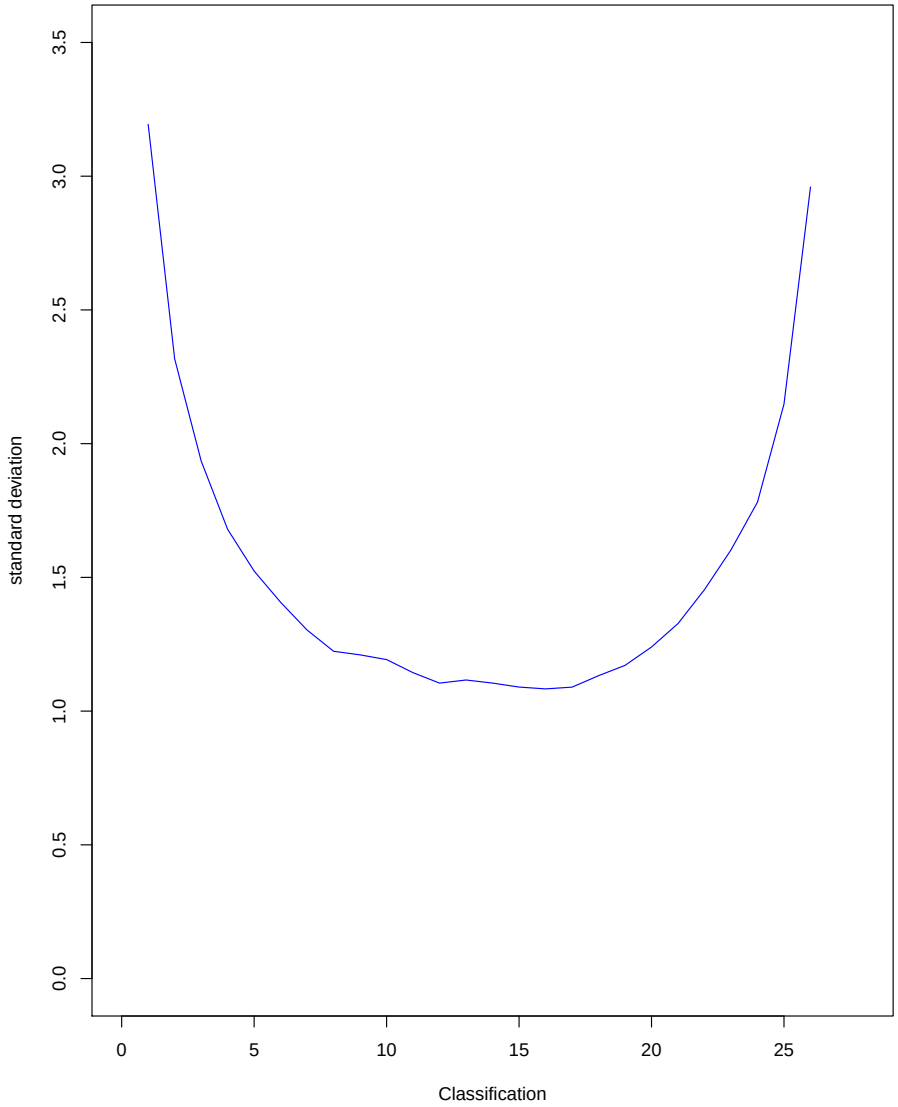
- We will compare the simulation model with observations (using the number of probability of draws based in previous championships -> historical data)

1. First validation: The accuracy of the model and the number of observation per run.We will choose 26 teams,the probability of draw will be generate by a triangular distribution with parameters (0.20, 0.24,0.30) and we will compare the mean score and the mean standard error for the position 8th and position 22th. We will do the simulation for a different numbers of observation: 200, 500, 1000, 2000, based each time in a 10 simulations runs. Here the single observation is the entire championship => single observation is a matrix

Number of obs per 10 runs	Mean 8th	MSE 8th	Mean 22th	MSE 22th
200	**38.1295**	**0.0006377**	**28.0630**	**0.0007237**
500	**38.1266**	**0.0002449**	**28.0846**	**0.0002861**
1000	**38.0923**	**0.0001259**	**28.1353**	**0.0001413**
2000	**38.1141**	**0.00006340**	**28.1256**	**0.00007129**

Validation and Exploring the results

2. Second validation: Look at the relationship between the variability of the scores estimators (measured by the standard deviation) and the teams positions.


standard deviation
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
5
10
15
20
25
Classification

Validation and Exploring the results

3. Third validation and comparison: Look at the percentiles from the simulation results of necessary scores to achieve classification and avoid relegation according to the number of the teams per championship

Percentile	22 teams		24 teams		25 teams		26 teams		28 teams	
	Class	Rel	Class	Rel	Class	Rel	Class	Rel	Class	Rel
5th	31	15	34	17	33	18	37	19	41	21
10th	31	16	35	18	34	19	38	20	41	22
15th	32	17	35	19	35	20	38	21	42	23
20th	32	18	35	20	35	21	39	22	42	24
25th	32	19	36	21	35	22	39	23	42	25
30th	33	19	37	21	36	22	39	23	43	26
35th	33	20	37	22	36	23	40	24	43	26
40th	34	20	37	22	37	23	40	24	44	26
45th	34	20	38	23	37	24	41	25	44	27
50th	34	21	38	23	38	24	41	25	44	27
55th	35	21	39	23	38	24	42	25	45	28
60th	35	22	39	24	39	25	42	26	45	28
65th	36	22	39	24	39	25	43	26	46	28
70th	36	22	40	24	40	25	43	26	46	29
75th	37	22	40	25	41	26	44	27	47	29
80th	38	23	41	25	41	26	45	27	48	29
85th	38	23	42	25	42	26	46	27	49	30
90th	40	24	43	26	44	27	47	28	50	30
95th	41	24	45	26	46	28	49	29	52	31

Discussion

- My results for the simulations for the accuracy of the model according to the numbers of observations are quite similar to the authors results for first and second part.
  1. For the mean of the score for 8th and 22th position = similar values
  2. For the MSE, I obtain smaller values => my model seems to be more accurate. But in general, as for the authors analysis, the standard errors tend to decrease with the increase of the numbers of observations per runs, so the variability decrease and the model becomes more accurate.
- For the plot, we can see that the variability of its score is bigger for the extremes of the ranking (between 1.5 to 3.5)
- Comparing our second table with the table of the article => I have more variability for the percentiles.
- Comparing with the real data, in 2001, the number of teams were 28 and the classification points observed were 45 and 29 to avoid the relegation. I obtain 45 for the 50th percentile and 29 for the 80th percentile. My results are not really similar to the real data at the 95th percentile but there are not completely different.

Conclusions

- The assumptions that we made at the beginning to simplify the model are too strong:

1. We just took in account the fact that a team plays against another team once. In the reality, a team plays twice against another team : one at home , and another outside so these can have an impact to the accuracy of the model.
2. We didnt take in account correlation between injuries/players values and games, that can have a huge impact for a game.

- The model is to simple but give us good general idea.