Proceedings of the 2002 Winter Simulation Conference
E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, eds.

SOCCER CHAMPIONSHIP ANALYSIS USING MONTE CARLO SIMULATION

Caio Fiuza Silva
Eduardo Saggioro Garcia
Eduardo Saliby

COPPEAD / UFRJ
Universidade Federal do Rio de Janeiro
C. Postal 68514
Rio de Janeiro, RJ, 21949-900, BRAZIL

ABSTRACT

Sports had always fascinated humanity. In this context, soccer was taken as a study source. The objective of this paper is to formulate a simulation model to generate estimators for necessary scores to achieve certain places at the final classification ranking of the Brazilian National Soccer Championship. The main data used are the rules of the championship, the number of competitors and the probability that a match ends up in a draw.

1 INTRODUCTION

Since its development by the English in the nineteenth century to nowadays, soccer has been one of the most known and played sports throughout the world. In many countries, especially in Europe and South America, soccer has long become the most popular sport.

In Brazil, for example, soccer is so popular and widespread that it is almost a religion for the country's 170 million people. The Brazilian National Soccer Championship, its most important soccer competition, brings every year millions to stadiums and to the front of TVs to watch the matches.

The best Brazilian soccer teams annually dispute this championship. Its last edition presented 28 teams and was developed in two stages. The first stage was classificatory, with each team playing against each other, totaling 27 matches per team and 378 matches for the entire stage. At the end of this stage, the 8 best teams were classified for the playoffs stage, while the last 4 teams were relegated to a lower rank competition.

Despite its strong popularity, Brazilian soccer has been going through a crisis. Political, financial and administrative difficulties, corruption and lack of organization in the institutions responsible for soccer management submit teams to calendars that often compel them to dispute 2 or 3 championships at the same season. This situation affects the players' behavior and performance, since the time between two successive matches is often not enough for their physical rest and recovery. The consequence of that is the raise in the number of injuries, impairing teams, which usually have to give priority to some championships and to forsake others. To illustrate the situation, the most important Brazilian teams played approximately 100 matches last year. Considering the fact that the athletes have 30 days on vacation, the teams played, on average, a match every 3.3 days.

In this manner, estimating the necessary score to reach a desired position in a championship can support teams on the formulation of their strategies, helping them to optimize the use of resources and to plan efficiently the athletes' physical preparation.

The purpose of this paper is therefore to present a simple simulation model that represents the Brazilian National Soccer Championship, providing the estimation, with a certain level of confidence, of the necessary scores to achieve classification to the playoffs and to avoid relegation.

2 THE SIMULATION MODEL

The main idea of the proposed model is to generate randomly, through Monte Carlo sampling (Hammersley and Handscomb, 1964), the number of points obtained by each team in each match. The entire championship is simulated and the teams are then classified in accordance with their cumulative number of points, which is their final score.

Until 1994, in the Brazilian National Soccer Championship, the winner team obtained 2 points in a match, while the loser obtained 0 points. If the result were a draw, each team would obtain 1 point. Nevertheless, the rules were changed. Now, when a team wins a match it obtains 3 points. The loser still receives no points and a draw still gives each team 1 point.

By simplification, it is assumed in the proposed model that the probability that a match ends up in a draw is the same for all the matches in a simulated championship. Other assumptions are:

- Equality among all teams – The chance of winning in a match is the same for every team, i.e., it is not considered that any team has advantages or is stronger than the others;
- Independence between results – It is assumed that the result of a match is totally independent from previous matches.

Considering the aforementioned assumptions and rules, the first step to elaborate the model is to create the matrix A(M,N), with M = N = number of competitors in the championship. This matrix, illustrated in Figure 1, is used to store the matches' results. The rows and the columns show the teams enrolled in the championship. Each element a(i,j), with i different from j, represents the number of points obtained by team i in the match between team i and team j. The value of a(i,j) belongs to the set R = {0,1,3}, i.e., 0 for loss, 1 for draw and 3 for victory of team i against team j.

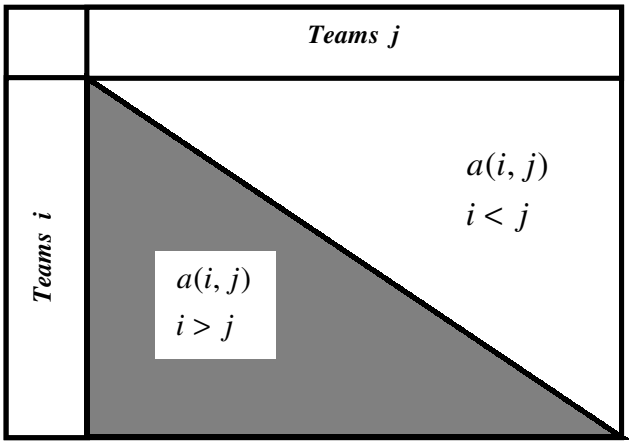

Figure 1: Representation of the Results Matrix, A(M,N)

The process of generating the matches' results is done in two stages. In the first stage, random numbers between 0 and 1 are generated and stored in the matrix B(M,N). These random numbers are assigned to each element b(i,j), with i bigger than j, and will determine the result in a match, considering that the match is a non-zero-sum and non-collaborative game (Luce, Dunkin, and Raiffa 1957), in which the result of a player determines the result of his opponent.

The second stage of the process is to fill the results matrix A(M,N). The elements a(i,j), with i bigger than j, are defined by their association with the values of each b(i,j) and P(D), the probability that a match ends up in a draw.

After filling the entire matrix A(i,j), the sum of each column gives the total score of each team in the championship. In this way, it is possible to classify them, finding the score of each position in the classification ranking.

The logic used, developed in Visual Basic Language is shown in the following algorithm:

```
For j=1 to N
For I=1 to M
    While I>j
    If b(I,j)<=P(D) then
    a(I,j)=1 else
        if b(I,j)<=[P(D)+(1-P(D))/2] then
        a(I,j)=3 else
            a(I,j)=0
        end if
    end if
    wend
    while I<j
    if a(I,j)=1 then
    a(j,i)=1 else
        if a(I,j)=3 then
        a(j,I)=0 else
            a(j,I)=3
        end if
    end if
    wend
next
next
```

3 VALIDATION

Validation consists in comparing the simulation model results with observations in real systems and, through analysis of differences and deviations, improving the model until acceptable results are achieved.

In the case of the proposed model for the Brazilian National Soccer Championship, the validation process was developed in the following stages:

- Historical analysis of the percentage of draws;
- Design of experiments;
- Verification of model's applicability;
- Comparison between simulation results and real systems observations.

Next, each of the stages will be described in more details.

3.1 Historical Analysis of the Percentage of Draws

Data for the last 6 Brazilian National Soccer Championships were collected in order to investigate the behavior of the aggregate percentage of draws.

It can be seen that the number of matches that ended up in a draw oscillated between 21.7% and 30.1%, with

an average of 24.2%. Figure 2 shows the percentage of draws through the development of the championships analyzed.

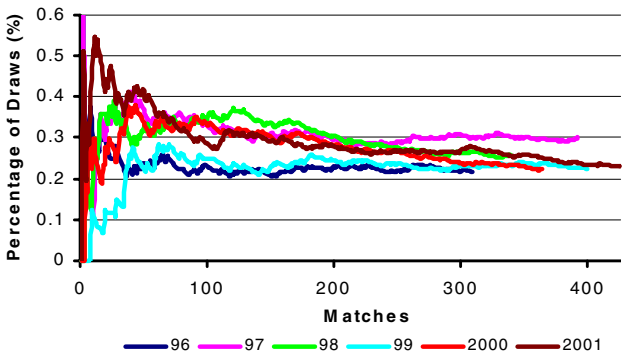

Figure 2: Historical Data of Draws in the Brazilian National Soccer Championship from 1996 to 2001

The probability that a match ends up in a draw, a basic parameter for the simulation, can be modeled based on the historical data presented.

3.2 Design of Experiments

According to Banks (1995), in every simulation project, it is needed to make decisions related to size of warm up process, number of simulation runs and number of observations in each run. In this paper, an observation is the simulation of one entire championship, while a run is a set of championships.

Based on the assumption that the observations are totally independent one to another, the results of one championship cannot be affected by previous championship results. Thus, it is not necessary to proceed with the warm up process in the proposed model.

The definition of the number of simulation runs is important when it is required to analyze statistics among results of different runs. A number of runs bigger than 1 is useful, for example, in determining the variability among the results of sets of championships. If no comparison among sets of championships is needed, the number of simulation runs can arbitrarily be set to 1.

The number of observations in each simulation run must be determined more carefully. Tests with different numbers of observations per run should be done in order to determine which values are appropriate to the proposed model.

In this manner, an experiment was elaborated to test the results of simulations with different numbers of observations per run. A championship with 26 teams was considered and the probability of draws was modeled as a triangular distribution with parameters (0.20, 0.24, 0.30). The outputs of the simulation are the mean and the standard error of the scores of the 8th and 22nd positions in the championship, based on the results of 10 runs of n observations. The number of observations per runs tested was 200, 500, 1000 and 2000. The results are shown in Table 1:

Table 1: Comparison of the Simulation Results with Different Number of Observations per Run

	Results based on 10 Simulation Runs			
	Minimum Score to achieve Classification (8th)		Minimum Score to avoid Relegation (22nd)	
Number of Observations per Run	Mean	Mean Standard Error	Mean	Mean Standard Error
200	38,144	0.101	28,099	0.079
500	38,102	0.057	28,115	0.059
1000	38,117	0.055	28,130	0.052
2000	38,108	0.023	28,120	0.026

An analysis of Table 1 shows that the averages of the outputs do not change significantly when the number of observations per run is increased. As expected, with the increase on the number of observations per run, the standard errors tend to decrease, i.e., the variability of the results is diminished and the model becomes more accurate.

3.3 Verification of Model's Applicability

Once analyzed the relation between the accuracy of the model and the number of observations per run, a single run simulation with 2000 observations of a championship with 26 teams was done. Estimators for the scores of each position at the classification ranking were collected and their standard deviations were calculated. The results are shown in Figure 3.

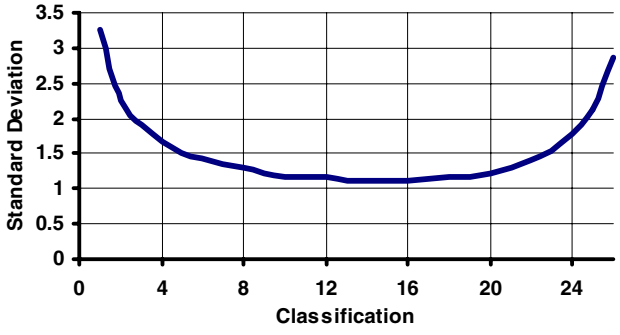

Figure 3: Relationship Between the Standard Deviation of the Scores Estimators and the Teams Position at the Ranking

The variability of a team's score, measured by its standard deviation, is a function of its position in the classification ranking. The more a team is positioned at the extremes of the ranking, the bigger is the variability of its score, as seen in Figure 3.

It can be noticed that between the 4th position and the 24th position the standard deviation ranges from 1.0 to 1.6 points. The estimators of the scores for the first and last places, however, presented standard deviations around 3.0 points.

Since the most interesting estimators are the necessary scores to avoid relegation and to achieve classification to the playoffs, which are intermediate positions at the classification ranking, it is possible to conclude that the use of the proposed model should be adequate to the Brazilian National Soccer Championship.

3.4 Comparison between Simulation Results and Real Observations

The final step to complete the validation process is to check the consistency between values generated by the simulation model and values observed in real systems.

In this paper, real systems are the Brazilian National Soccer Championship editions from 1996 to 2001.

Differently from European soccer championships, in which rules are the same every season, the rules for the Brazilian National Soccer Championship changes every year. Therefore, the simulation parameters have to be changed too. Table 2 shows the parameters observed in each championship edition.

Table 2: Simulation Parameters for the Real Systems

Year	NT	Classified	Relegated	D%
1996	24	8	4	21.7
1997	26	8	4	30.1
1998	24	8	4	25.7
1999	22	8	0	22.5
2000	25	12	0	22.3
2001	28	8	4	23.0

where:

NT = Number of teams enrolled in the championship;
Classified = Number of teams that were classified to the playoffs;
Relegated = Number of teams that were relegated to the lower division;
D% = Percentage of draws observed in each championship.

Once defined the real systems and their parameters, an application of the proposed model with one simulation run of 2000 observations was done. The probability of draws was modeled as a triangular probability distribution with parameters (0.217, 0.242, 0.301).

The results of the simulation, given in Table 3, show the percentiles of the empirical distribution of the estima

Table 3: Percentiles, Obtained from the Simulation Model Results, of Necessary Scores to Achieve Classification and to Avoid Relegation, as a Function of the Number of Teams Competing in the Championship.

Percentile	22 teams		24 teams		25 teams		26 teams		28 teams	
	Class.	Releg.	Class.	Releg.	Class.	Releg.	Class.	Releg.	Class.	Releg.
5th	30	22	33	24	34	25	36	26	39	28
10th	30	22	33	24	35	25	37	26	40	28
15th	30	22	34	24	35	26	37	27	40	29
20th	30	23	34	25	35	26	37	27	40	29
25th	31	23	34	25	36	26	37	27	41	29
30th	31	23	34	25	36	26	37	27	41	30
35th	31	23	34	25	36	27	38	28	41	30
40th	31	23	34	26	36	27	38	28	41	30
45th	31	24	35	26	36	27	38	28	41	30
50th	31	24	35	26	36	27	38	28	41	30
55th	31	24	35	26	36	27	38	28	42	31
60th	32	24	35	26	37	27	38	28	42	31
65th	32	24	35	26	37	28	38	29	42	31
70th	32	24	35	27	37	28	39	29	42	31
75th	32	25	36	27	37	28	39	29	42	31
80th	32	25	36	27	37	28	39	29	42	32
85th	32	25	36	27	38	28	39	30	43	32
90th	33	25	36	28	38	29	40	30	43	32
95th	33	26	37	28	38	29	40	30	44	33

tors for the necessary scores to achieve classification and to avoid relegation.

According to the data presented in Table 3, if a team, which is enrolled in a championship of 22 teams, finishes the championship with a score of 32 points, it has a chance between 85% and 90% of being classified to the playoffs. In another way, if a team obtains 22 points at the same championship, its probability of avoiding relegation lies between 15% and 20%.

Table 4 presents data from the real systems analyzed. For the 6 Brazilian National Soccer Championships editions, the observed necessary scores to achieve classification and to avoid relegation were collected in order to be compared with the simulation results presented in Table 3.

Table 4: Observed Scores in Real Systems

	Number of Teams	Classification	Relegation
1996	24	36	27
1997	26	37	26
1998	24	36	24
1999	22	33	-
2000	25	36	-
2001	28	45	29

The combined analysis of Tables 3 and 4 shows that the simulation model is quite consistent and adherent to the real systems, since almost all real observed scores lay between the 5th and 95th percentiles of the empirical distribution obtained from the simulation model. The only exception is the score obtained by the 8th team in the classification ranking of the 2001 championship. Its real score was 45 points, while the 95th percentile from the simulation indicates 44 points.

4 MODEL'S APPLICATION

The proposed simulation model represents an a priori analysis of a soccer championship, considering exclusively initial conditions, general rules and simplified assumptions for the whole competition, such as equality among teams. In this kind of analysis, it is possible to study only the number of points conditioned to a specific event, like the necessary score to avoid relegation, but it is impossible to predict anything about the chances of a particular team.

A posteriori analysis of the championship can be done in order to resolve this limitation of the model. In this analysis, after the end of a championship round, the table of results is updated with real values, reducing the effects of uncertainties in the next simulation.

To make easier the utilization of the proposed model, SIMUL – SOCCER 2002, a software developed in Visual Basic 6.0, was created.

The main screen allows the user to define the number of competitors, the team's names and the probability of draws. Analysts can insert matches results after they occur and simulate only the remaining matches. This procedure generates the individual probabilities of classification to the playoffs and of relegation, besides making the results more accurate.

The great benefit of simulating using SIMUL – SOCCER 2002 is the dynamic analysis of chances for each team through the competition.

5 SUMMARY AND CONCLUSION

The main objective of this paper was to formulate a simple and easy of utilization simulation model, which is able to generate reliable estimators for the necessary score to achieve a certain position at the classification ranking of a soccer championship.

The results proved that the proposed method is quite representative of the real systems investigated. Until now, the most used models were based on heavy databases containing historical statistics.

It is also possible to extend the application horizon of the model. Other soccer championships around the world, such as the French, Italian, English, German and Spanish soccer leagues can be successfully simulated. Moreover, it is also reasonable the application of the model in other sports in which the possible results for the teams are draw, victory and loss.

6 REFERENCES

Banks, Jerry. 1999. Discrete-event system simulation. Prentice-Hall, Inc. Upper Saddle River, New Jersey

Hammersley , J.M., and D.C. Handscomb. 1964. Monte Carlo Methods. Chapman and Hall, London.

Luce, R. Duncan and H. Raiffa.1957. Games and Decisions: introduction and critical survey. John Wiley & Sons, Inc.

AUTHORS BIOGRAPHIES

CAIO FIUZA SILVA is a student of industrial engineering at Federal University of Rio de Janeiro. He has a scholarship from ANP, Brazilian National Agency for regulation of the Petroleum Sector, to research special topics of industrial engineering related to the oil industry. His research interests are in business games development, statistical forecasting and simulation applied to logistics and supply chain management. He is a member of SOBRAPO (Brazilian O. R. Society). His email address is <caiofiuz@coppead.ufrj.br>

EDUARDO SAGGIORO GARCIA is a student of industrial engineering at Federal University of Rio de Ja-

neiro. He has a scholarship from ANP, Brazilian National Agency for regulation of the Petroleum Sector, to research special topics of industrial engineering related to the oil industry. His research interests are in inventory management, statistical forecasting, simulation modeling and operations research applied to logistics and supply chain management. He is a member of SOBRAPO (Brazilian O. R. Society). His email address is <edsg@ufrj.br>

EDUARDO SALIBY is a full professor at COPPEAD/ UFRJ, Graduate Business School, Federal University of Rio de Janeiro, Brazil. He received a B.S. degree in industrial engineering from the University of Sao Paulo, a M.S. degree in industrial engineering and operations research from the Federal University of Rio de Janeiro, and a Ph.D. in Operations Research from The University of Lancaster, UK, in 1980. His research interests are simulation methodology, with special emphasis on simulation sampling, simulation practice and simulation software. He is a member of SOBRAPO (Brazilian O. R. Society), ORS (UK) and SCS. His email address is <saliby@coppead.ufrj.br>