A Sequential Markov Chain Model of FIFA World Cup Winners

Nehemiah A. Ikoba
Department of Statistics, University of Ilorin, Ilorin, Nigeria

In this paper, a sequential Markov chain conceptualization of the winners of the FIFA World Cup is presented. The aim was to capture the dynamics of the World Cup and predict the future winner via Markov chain analysis. A sequentially incremented state-space Markov chain is used to approximate the process of winning the FIFA World Cup. The corresponding Markov chains at every epoch where the state space increases were computed. The result of the analysis showed a close predictive ability of the model to predict the previous World Cup winners. It is predicted that a new winner may emerge in the 2022 World Cup in Qatar. However, if a new winner does not emerge, on the basis of both the sequential Markov chain and the first passage matrix of the conceptualized model, then Brazil is the most probable winner of the 2022 World Cup, followed by Italy and Germany. The sequential Markov chain approach can be applied to other sporting events and scenarios in which there is only a small probability that the number of observed states may increase from a small set of states.

Keywords: stochastic processes, sports analytics; transition probability matrix, mean first passage time, mean recurrence time

1. Introduction

Soccer is one of the most popular team sports in the world and the Federation Internationale de Football Federation (FIFA) World Cup has blossomed in terms of followership and economic benefits over the years. Since the first edition of the FIFA World Cup in 1930, the number of participating teams has increased to the current 32 teams, with over 200 national teams participating in the qualifying matches every cycle (Paul and Mitra, 2008). The World Cup is a global brand with massive sponsorship and advertisement revenues accruing to FIFA.

The 21 World Cup tournaments have been won by only 8 national teams (three from South America – Uruguay, Brazil and Argentina; and five from Europe – Italy, Germany, England, France and Spain). Brazil leads the winners chart with five wins and is the only team to have featured in all 21 tournaments (Wikipedia).

Sixteen countries have hosted the World Cup, five of which have hosted it twice. South America accounts for 9 victories compared to Europe's 12 till date (FIFA).

It is of interest to study the dynamics of the FIFA World Cup in terms of the small group of winning nations. Only eight countries have won the World Cup spread across only two continents: South America and Europe. Furthermore, since 1966 when England won the competition for the first and only time, only three other countries – Argentina, France and Spain have joined the group of World Cup winners. No team from Africa has gotten to the semi-finals of the World Cup, even though Nigeria and Ghana are global powerhouses in the age-grade competitions. In fact, Nigeria has won the Under-17 World Cup a total of five times, more than any other country in the world.

Several researches have been done in relation to analysis of match outcomes and prediction of winners of football competitions like the World Cup, European Championships and the various club competitions in various continents.

Kuonen *et al.* (1997) applied logistic regression using seed positions to model the win/ loss result of knockout games of European club football competitions. The aim was to predict the probability of winning the tournament and the probability of each team reaching a particular stage of the competition. The analysis showed a predictive strength of the model was 64.5%. Since the model did not take into consideration the scores of matches, the authors proposed the incorporation of goal for and against teams in the conceptualization of the seeding coefficient, as a possible way of improving the predictive capability of the model.

Dyte and Clarke (2000) suggested a method for predicting the distribution of scorers in international soccer matches, treating each team's goals scored as independent Poisson variables dependent on the FIFA ranking of each team, and the venue of the match. Poisson regression estimates of model parameters were used to simulate matches played during the 1998 World Cup, and the model was found to perform plausibly well. However, there was a subjective adjustment of the FIFA ranking points of countries from non-traditional soccer playing regions of Europe and South America, thereby introducing some form of bias into the model analysis.

Hoffmann *et. al* (2002), in studying the socioeconomic determinants of international soccer performance, used a multiple regression model to analyze the performance of a sample of countries on the basis of their FIFA ranking points. Economic, demographic, cultural and climatic factors were adduced as the factors that determine the performance of teams in football international tournaments. The coefficient of multiple determination showed that only about 32% of the variation in the response variable could be explained by the independent variables. This pointed to a very poor regression model and the need to include more relevant variables in the model, which was not done by the authors. The regression model thus has several shortcomings and is not a fair predictor of performance at the FIFA World Cup.

Szymanski (2003) built a multiple linear regression model showing that wealthier countries tend to be sportier, using predictor variables like gross domestic product (GDP) per head, participation rate of the country, and popularity of football in the country. The analysis provided a low value for the coefficient of multiple determination ($R^2 = 0.40$). A major finding was that much of what determines success is beyond the immediate control of football administrators.

Torgler (2004) claimed that for the 2002 World Cup, there was an element of uncertainty, as teams did not perform according to their FIFA ranking.

The benchmark for evaluation was the last FIFA ranking (May) before the commencement of the competition. Performance metrics used included the number of games won and the number of goals scored in comparison with the team's FIFA ranking. Two econometric base models were proposed: the probit model and the ordinary least squares model with the goal difference as the dependent variable, while the independent variable was the rank difference between the teams. It was claimed that a team with higher number of yellow cards and red cards was more likely to win the World Cup. A number of seemingly spurious regression conclusions were presented, which do not have any relationship with the actual winning or losing of the World Cup.

In the context of application of Markov chain methodology to other sports, Jones (2004) modelled the overtime period of the National Football League (NFL) using an absorbing Markov chain model. Two types of overtime – sudden death and the first-to-six rule were examined to determine the optimal overtime rule. The models were validated by data from the 2002 NFL season. It was shown that the first-to-six rule was better than the sudden death rule. However, a tradeoff exists as the number of ties in matches increases marginally using the first-to-six overtime rule.

Baker and Scarf (2006) analyzed data from 20 sporting competitions and some simple models were proposed to forecast match outcomes. The main focus was on annual sporting contests in which the same two teams strive to overcome each other year after year. The motivation was an attempt to predict the winner of the Oxford-Cambridge boat race. The runs test for randomness was used to analyze the sequence of wins, losses or draws of these competitions. Significance tests for studying the effect of 'match covariates' were introduced, and the effect of these covariates was found to be quite large in many instances. Several potential application areas of the conceptualized models were provided.

Lago (2007) explained the effect of performance and chance in the results of the teams in the 2006 FIFA World Cup matches. The analyses were based on multiple linear regression, mean comparison test and logit multinomial models. It was concluded that the performance was important during the first round of matches, but during the knock-out phase, the role of performance becomes less important.

Paul and Mitra (2008) carried out an analysis of 4 World Cup winners via the monthly FIFA ranking. It was stated that the top-seeded team never won the World Cup except in 1994. The researchers examined the strength of the element of uncertainty in the World Cup using two empirical models. A higher ranked team always won the World Cup.

Volf (2009) modelled the score in a football match as two interacting time dependent random point processes. This was captured via a semi-parametric multiplicative regression model of intensity. The model was validated by analyzing the performance of the quarter-finalists of the 2006 FIFA World Cup.

Suzuki *et al.* (2010) proposed a Bayesian methodology to predicting match outcomes of the 2006 FIFA World Cup. The FIFA ranking and specialists' opinions were used as prior information to calculate the win, draw and loss probabilities of each match. The whole competition was also simulated in order to estimate the classification probabilities in the group stage and winning tournament chances for each participating team. The De Finetti measure and the percentage of correct

forecasts were used to assess the predictive capability of the model. The approach of Suzuki *et al.* (2010) is similar to Dyte and Clarke (2000), but done within a Bayesian framework. The advantage of this approach was that the incorporation of specialists' knowledge seeks to overcome the challenge of little knowledge about teams and updates the team's strength as the competition progresses. The choice of specialists' opinion could be hugely subjective, impacting on the predictive capability of the model. The model was unable to predict correctly the eventual finalists and winner of the 2006 World Cup.

Cattelan, *et al.* (2013) provided a dynamic version of the Bradley-Terry model for paired comparison data that was applied to determine the outcomes of sporting contests. The model allowed for time-varying team abilities which depend on past results through exponentially weighted moving average processes. Model validation was done through data elicited from the 2009/2010 regular season of the National Basketball Association (NBA) tournament and the 2008/2009 Italian Serie A football season.

Scoppa (2013) underscored the role of fatigue in sports performance from an economic standpoint. Drawing data from the FIFA World Cup and UEFA European Championship, a regression model of team performance against several possible indicators of fatigue was used. Team performance was found to be correlated both with past outcomes and the world ranking of the countries, while the host country also enjoys considerable advantage. All the variables of performance in the regression model were defined in terms of differences, and the impact of rest on the team's performance was shown to be significant. Also, stronger teams and the home team were found to have a higher probability of enjoying more rest between matches, hence subject to less fatigue.

Kaufmann (2014) proposed that the major factors contributing to national teams winning the World Cup are good governance in the country and a strong fan base. The basis of the analysis was quite fuzzy, as some of the strongest democracies in the world are yet to win the World Cup.

Some conceptualized statistical models of performance at the World Cup also incorporate the market value of the national team (the aggregate market value of the individual players of the team). Other predictor variables in a multiple linear regression model of outcomes include the size of a country's economy, the market value of the Head Coach, whether the Coach is a foreigner or not, the resource wealth level of the country, the population of the country and home advantage, among several other factors (Kaufmann, 2014).

An extensive review of within-game and between-game data analysis strategies for various sports is provided in Percy (2015). Various stochastic processes were identified and their relevant applications in different sports highlighted. Furthermore, a Bayesian framework was presented for sequentially updating parameters estimates on the basis of within-game dynamic learning Markov processes in Percy (2015). However, the use of within-game data limits the utility of such researches, as the model needs to be continually updated.

Rumpf *et al.* (2017) studied the technical and physical performance parameters that distinguish between teams winning and losing matches in the 2014 FIFA World Cup. Twelve physical and 21 technical parameters were analyzed for each team. The winning teams, as should be expected, scored significantly greater number of goals, as well as other goal-related metrics, and received significantly

lower yellow cards. Binary logistic regression analysis showed that shot accuracy was the best predictor for success. It was concluded that scoring efficacy from open play and from set-pieces are crucial to win matches in the World Cup. However, from the history of the World Cup, these on-field physical and technical parameters may not adequately predict the winner of the World Cup.

Seth (2018) examined the determinants of FIFA World Cup performances of nations via a multiple linear regression model. The variables that were significant in the regression model were the seeding status of a country in the tournament, the presence of star players in the team, in addition to whether the country is a host of the World Cup and being a member nation of FIFA before 1924. None of the socioeconomic or demographic variables incorporated into the model were statistically significant in determining performance in the competition. The justification for choosing 1924 as the cut-off date for the variable FIFA membership was not sufficiently established.

In section 2, the methodology which provides a description of the adapted Markov chain model and its theoretical analysis are presented. The results from the analysis of the past World Cups are captured in section 3. Finally, the conclusions are expressed in section 4.

2. Materials and Methods

Markov chains are discrete-time, discrete-space stochastic processes (Grinstead and Snell, 1997) that provide useful information on the dynamics of the phenomenon being studied. A set of random variables (X_n, n = 0,1, …) forms a Markov chain if the conditional probability of the next state X_{n+1} depends only on the immediate past state, X_n, and not on previous states.

The process through which winners emerge at the FIFA World Cup is conceptualized as a Markov chain. A Markov Chain categorization of the World Cup is possible with the state of the process, X_n being defined as the winner of the World Cup in the n^{th} edition.

Let the stochastic process $\{Xn,\ n = 1,2, \ldots\}$ represent the winners of the World Cup from its inception in 1930. This is a discrete state Markov chain with a large state space representing all member countries of FIFA. However, the corresponding transition probability matrix may not be estimable from data, as the corresponding columns and rows of countries yet to win the World Cup will have zero entries. This is due to the fact that only 8 countries have won the World Cup so far. A conventional Markov chain conceptualization of the World Cup winners will be impossible because the process has only 21 observations spanning a 90-year period. Thus, in order to observe the process sufficiently and have up to 60 observations, say, we may have to wait for over 150 years. An alternative approach is thus needed to overcome the challenge of an extremely large state space of the Markov chain.

The nature of the FIFA World Cup has been such that there is a sequential increase in the number of winners of the competition. The winners come from a small group of countries while occasionally, new winners join this group. A sequential Markov chain is thus a feasible approximation of the process of winning

the World Cup. The Markov chain is such that there is a sequential updating of the state space at certain epochs. This implies that the corresponding transition probability matrix changes whenever there is an update of the state space.

At time t=0, corresponding to the time prior to the commencement of the World Cup competition in 1930, it is assumed that there are no winners, that is, $X_0 = 0$. At time t=1, it is assumed that the Markov chain has data corresponding to only two states and $X_1 = 1$. That is, the process is in state 1 after the first step, representing the first winner of the competition. This is the starting point for the sequential Markov chain model. This two-state chain, whose transition probability matrix represents the transitions between the first and second winner of the competition, is observed until the instant when a new winner emerges in the competition, increasing the state space to 3. The process is continued until we have the existing 8-state Markov chain that reflects the 8 winners of the competition till date.

This abridged state space is sequentially increased and comprises all the winners of the World Cup: $S=\{1,2,\ldots,8\}$ representing Uruguay (1), Italy (2), Germany (3), Brazil (4), England (5), Argentina (6), France (7), and Spain (8). It is noted that not every edition of the World Cup results in a new winner emerging, hence the state space only increases on those editions of the competition where a new winner emerges.

The process can be viewed as a *compound Markov chain* whose state space evolution is governed by another process that sequentially increases the size of the state space.

The Markov property is given as

$$\Pr(X_n = i_n | X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, \cdots, X_1 = i_1) = \Pr(X_n = i_n | X_{n-1} = i_{n-1}) = p_{ij}$$

and represents the one-step transition probability of moving from state i to state j in exactly one step.

The one-step transition probabilities (p_{ij}) are the ij^{th} entries of the stochastic matrix P, called the *one-step transition probability matrix.*

It is possible to incorporate the dynamics of the state space process into the Markov chain whenever a new winner emerges. The state space can only increase at those epochs where a new winner emerges and the TPM will thereafter be updated to reflect the new situation. If a process governs the evolution of the state space, then that process also influences the nature of the corresponding TPM. Incorporating a dynamic mechanism to take care of the increment in the state space will yield the appropriate TPM of the process. The increment process is viewed as a Poisson process which is attached to the corresponding Markov chain.

Assume that the process through which new winners emerge in the World Cup is Poisson with rate $\lambda = 1/\mu$ where μ is the mean waiting time till a new winner emerges, that is, the interarrival time for new winners.

The assumption is also made that no two winners will emerge at the same time. Once there is an arrival in the Poisson process, the state space of the Markov chain is updated by 1 and the corresponding transition probability matrix is computed.

The Markov chain goes sequentially from a 2-state model up to a k-state model, incremented after each tournament that produces a new winner. The reasoning behind this sequential Markov chain is that the phenomenon being studied has very limited observations even though it has been observed for a long time, the process revolves round a very small set of possible states which could be incremented occasionally.

The transition probabilities p_{ij} are estimated from data. The maximum likelihood estimates of the transition probabilities are obtained by maximizing the likelihood function (or joint probability density) subject to the constraint that $\Sigma_j p_{ij} = 1$.

The log likelihood function of the model is given as

$$\log_e L = \sum_{ij} n_{ij} \ln(p_{ij}) \tag{1}$$

The maximum likelihood estimate of the transition probabilities are then obtained as (Tan and Yilmaz, 2002)

$$\hat{p}_{ij} = \frac{number\ of\ transitions\ from\ i\ to\ j}{total\ number\ of\ i\ ocurrences} = \frac{n_{ij}}{n_{i.}} \tag{2}$$

where n_{ij} is the number of transitions from state i to state j, while $n_{i.}$ is the total number of transitions from state i, i,j=1,2,...,k.

A compact representation of equation (2) could be obtained. Let **N** denote the matrix whose entries are the number of transitions from state i to state j. Let **c** be a column vector of length k with all entries equal to 1. Then the row sums are obtained as **Nc**. Furthermore, let **B** be a diagonal matrix whose diagonal entries are the entries of **Nc**. That is, $\boldsymbol{B} = diag(\boldsymbol{Nc})$. Then, the corresponding transition probability matrix **P** is obtained as

$$\boldsymbol{P} = (\boldsymbol{N}'\boldsymbol{B}^{-1})' = \boldsymbol{B}^{-1}\boldsymbol{N} \tag{3}$$

The maximum likelihood estimates of the transition probability matrix of the conceptualized Markov chain is obtained using historical data, which also provide information on the estimates of the probability of a new entrant joining the group of winners and the mean interarrival time for the embedded Poisson process governing the state space evolution of the Markov chain.

The sequence of transition probability matrices, $\{\boldsymbol{P}_1, \boldsymbol{P}_2, \cdots, \boldsymbol{P}_r\}$ are TPMs representing an evolving Markov chain. A particular $\boldsymbol{P}_i$ is valid until the chain is updated to the next higher state, when the underlying Poisson process has a new arrival, at which point the next transition probability matrix $\boldsymbol{P}_{i+1}$ becomes valid. The initial probability vector for each TPM $\boldsymbol{P}_i$ is such that all entries of the vector are zero except that of the last state, which is fixed at 1. This represents the new winner in the sequentially updated Markov chain. This is true except in the case of $\boldsymbol{P}_i$ where the initial probability vector is (1 0) representing the very first winner of the competition.

The TPMs $\{\boldsymbol{P}_1, \boldsymbol{P}_2, \cdots, \boldsymbol{P}_r\}$ specify generally *ergodic* Markov chains, as there is a possibility of movement between all the states. In an ergodic Markov chain, it is possible to move from every state to every other state, not necessarily in a single step.

Given a transition probability matrix $\boldsymbol{P}$ of an ergodic Markov chain, the stationary or long-run distribution of the chain satisfies the relationship:

$$\boldsymbol{\Pi P} = \boldsymbol{\Pi}$$

where $\boldsymbol{\Pi}$ is the unique fixed probability vector of the Markov chain.

The stationary probability distribution, **W** is a square matrix which has the same dimension as **P** and each row is the fixed probability vector $\boldsymbol{\Pi}$.

The stationary probability distribution is also obtained as the limiting form of the one-step transition probability matrix **P**, that is

$$\boldsymbol{W} = \lim_{\mathrm{n}\to\infty} \mathbf{P}^{\mathrm{n}}$$

Thus, higher powers of $\boldsymbol{P}$ approach the stationary probability distribution $\boldsymbol{W}$ (Grinstead and Snell, 1997).

It is of interest in the analysis of Markov chains to determine the first time a state is reached, as well as the time of first return to any state. These concepts are captured by the mean first passage time and the mean recurrence time.

The *mean first passage time* specifies the expected waiting time for any of the previous winner to win the competition, given that a particular team has won it.

On the other hand, the *mean recurrence time* is the expected number of steps it takes to return to a particular state, having started from that state. It is the expected length of time of waiting between wins for each of the World Cup winners.

The concept of fundamental matrices in absorbing Markov chains could be extended to ergodic Markov chains (Grinstead and Snell, 1997). Once the fundamental matrix of the ergodic chain is obtained, then the corresponding mean first passage time matrix can easily be determined.

Two important theorems and a relevant proposition from Grinstead and Snell (1997) are presented without proof:

Theorem 1 For an ergodic Markov chain, the mean recurrence time for state i is $\boldsymbol{r_i} = 1/\pi_i$ where π_i is the $\boldsymbol{i^{th}}$ component of the fixed probability vector for the transition matrix.

Proposition 1 Let **P** be the transition matrix of an ergodic chain, and let W be the matrix all of whose rows are the fixed probability row vector for **P**. Then the matrix

$$\boldsymbol{I - P + W}$$

has an inverse, called the Fundamental Matrix, **Z** of the ergodic chain. That is

$$\boldsymbol{Z=(I - P + W)^{-1}}$$

Theorem 2 The mean first passage matrix $\boldsymbol{M}$ for an ergodic chain is determined from the fundamental matrix **Z** and the fixed row probability vector by

$$m_{ij} = \frac{z_{jj} - z_{ij}}{\pi_j} \qquad (4)$$

The mean first passage to state j, starting from state i, m_{ij} can be expressed in a more compact form via the mean first passage matrix $\boldsymbol{M}$. This is presented, as well as the proof, in Theorem 3 below.

Theorem 3 The mean first passage matrix $\boldsymbol{M}$ for an ergodic Markov chain is given by

$$\boldsymbol{M = U{-}Z)D} \tag{5}$$

where $\boldsymbol{D}$ is a diagonal matrix whose entries are the mean recurrence times $\boldsymbol{r_i} = 1/\pi_i$, and $\boldsymbol{U}$ a matrix whose rows are the diagonal entries of the fundamental matrix $\boldsymbol{Z}$.

Proof:

The mean first passage to state j, starting from state i, is given as

$$m_{ij} = \frac{z_{jj} - z_{ij}}{\pi_j}$$

The matrix $\boldsymbol{U}$ is defined as

$$\boldsymbol{U} = \begin{pmatrix} diag(Z) \\ \vdots \\ diag(Z) \end{pmatrix} = \begin{pmatrix} z_{11} & \cdots & z_{kk} \\ \vdots & \vdots & \vdots \\ z_{11} & \cdots & z_{kk} \end{pmatrix}$$

The matrix $\boldsymbol{D}$ is a diagonal matrix whose entries are the mean recurrence times $\boldsymbol{r_i} = 1/\pi_i$, and is given as

$$\boldsymbol{D} = \begin{pmatrix} r_{11} & 0 & 0 & 0 \\ 0 & r_{22} & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & r_{kk} \end{pmatrix} = \begin{pmatrix} 1/\pi_1 & 0 & 0 & 0 \\ 0 & 1/\pi_2 & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & 1/\pi_k \end{pmatrix}$$

$$\boldsymbol{U - Z} = \begin{pmatrix} z_{11} & \cdots & z_{kk} \\ \vdots & \vdots & \vdots \\ z_{11} & \cdots & z_{kk} \end{pmatrix} - \begin{pmatrix} z_{11} & \cdots & z_{1k} \\ \vdots & \vdots & \vdots \\ z_{k1} & \cdots & z_{kk} \end{pmatrix} = \begin{pmatrix} 0 & z_{22} - z_{12} & \cdots & z_{kk} - z_{1k} \\ z_{11} - z_{21} & 0 & \cdots & z_{kk} - z_{2k} \\ \vdots & \vdots & & \vdots \ \vdots \\ z_{11} - z_{k1} & z_{22} - z_{k2} & \cdots & 0 \end{pmatrix}$$

The matrix $\boldsymbol{U{-}Z}$ is thus a matrix with zero in all its diagonal cells and $z_{jj}{-}z_{ij}$ in the off-diagonal cells. Since $\boldsymbol{D}$ is a diagonal matrix, then multiplying $\boldsymbol{U{-}Z}$ by $\boldsymbol{D}$ will yield

$$(\boldsymbol{U - Z})\boldsymbol{D} = \begin{pmatrix} 0 & z_{22} - z_{12} & \cdots & z_{kk} - z_{1k} \\ z_{11} - z_{21} & 0 & \cdots & z_{kk} - z_{2k} \\ \vdots & \vdots & & \vdots \ \vdots \\ z_{11} - z_{k1} & z_{22} - z_{k2} & \cdots & 0 \end{pmatrix} \times \begin{pmatrix} 1/\pi_1 & 0 & 0 & 0 \\ 0 & 1/\pi_2 & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & 1/\pi_k \end{pmatrix}$$

Thus

$$(\boldsymbol{U - Z})\boldsymbol{D} = \begin{pmatrix} 0 & \frac{z_{22} - z_{12}}{\pi_2} & \cdots & \frac{z_{kk} - z_{1k}}{\pi_k} \\ \frac{z_{11} - z_{21}}{\pi_1} & 0 & \cdots & \frac{z_{kk} - z_{2k}}{\pi_k} \\ \vdots & \vdots & & \vdots \ \vdots \\ \frac{z_{11} - z_{k1}}{\pi_1} & \frac{z_{22} - z_{k2}}{\pi_2} & \cdots & 0 \end{pmatrix} = (m_{ij}) = \boldsymbol{M}$$

Hence

$$M = (\boldsymbol{U - Z})\boldsymbol{D}$$

Therefore, given an ergodic markov chain, the mean first passage matrix can be computed by the formula provided in Equation (5). This provides a straightforward and computationally less tedious way of obtaining the mean first passage matrix instead of carrying out individual computations of m_{ij}.

3. Results and Discussion

Data of the winners of the FIFA World Cup from 1930 to 2018 were obtained from the FIFA website www.fifa.com. The list of World Cup winners is presented in table 1, while some important statistics are presented in tables 2 and 3. Table 4 summarizes the analysis of the conceptualized sequential Markov chain model for the estimation of the winning probabilities at each edition of the World Cup, as well as a projection of the probable winner of the 2022 World Cup, slated for Qatar. The 8-state transition probability matrix of the current chain is thereafter presented, as well as its mean recurrence times and mean first passage matrix.

Table 1. FIFA World Cup winners from 1930 to 2018, showing years when new winners emerged

Year	Winner	State Space Length	Year	Winner	State Space Length
1930	Uruguay	2	1982	Italy	6
1934	Italy (new)	2	1986	Argentina	6
1938	Italy	2	1990	Germany	6
1950	Uruguay	2	1994	Brazil	6
1954	Germany (new)	3	1998	France (new)	7
1958	Brazil (new)	4	2002	Brazil	7
1962	Brazil	4	2006	Italy	7
1966	England (new)	5	2010	Spain (new)	8
1970	Brazil	5	2014	Germany	8
1974	Germany	5	2018	France	8
1978	Argentina (new)	6			

Source: www.fifa.com

It is seen from table 2 that, on the average, a new winner emerges in the FIFA World Cup every three editions, the same as the variance. The minimum time taken to get a new winner is one competition (in 1958) and the longest wait before a new winner emerged is 5 competitions from 1978 when Argentina won for the first time to 1998 when France joined the class of World Cup winners. As shown from the value of the variance, there is little variability around the mean number of competitions before a new winner emerges.

Table 2. Summary statistics of the number of competitions before a new winner emerges in the FIFA World Cup

Minimum	Median	Mean	Maximum	Variance
1	3	2.57	5	1.95

Table 3. Summary of the number of wins, proportion of wins and estimates of the mean time until the next win for winners of the FIFA World Cup.

S/N	Country	Number of wins	Proportion of wins	Mean time until the next win
1	Uruguay	2	0.095	3
2	Italy	4	0.190	5
3	Germany	4	0.190	5
4	Brazil	5	0.238	3
5	England	1	0.048	Not applicable
6	Argentina	2	0.095	2
7	France	2	0.095	5
8	Spain	1	0.048	Not applicable

Table 4. Prediction of winners of the FIFA World Cup using the Sequential Markov chain model and the appropriate initial probability vector

Year	Winner	TPM	Predicted probability vector	Comments
1930	Uruguay	–	–	New winner.
1934	Italy	P_1	(0 1)	Winner correctly predicted
1938	Italy	P_1^2	(0.5 0.5)	Equal chances for both teams
1950	Uruguay	P_1^3	(0.25 0.75)	Winner not correctly predicted
1954	Germany	–	–	New winner
1958	Brazil	–	–	New winner.
1962	Brazil	P_3	(0 0 0 1)	Winner correctly predicted.
1966	England	–	–	New winner.
1970	Brazil	P_4	(0 0 0 1 0)	Winner correctly predicted.
1974	Germany	P_4^2	(0 0 0.33 0.33 0.33)	Equal chances for Germany, Brazil and England.
1978	Argentina	–	–	New winner
1982	Italy	P_5	(0 0.5 0.5 0 0 0)	Equal chances for both Italy and Germany
1986	Argentina	P_5^2	(0.17 0.17 0 0.33 0 0.33)	Equal chances for both Brazil and Argentina.
1990	Germany	P_5^3	(0.06 0.30 0.36 0.11 0.11 0.06)	Winner correctly predicted.
1994	Brazil	P_5^4	(0.10 0.16 0.09 0.39 0.04 0.22)	Winner correctly predicted.
1998	France	–	–	New winner.
2002	Brazil	P_6	(0 0 0 1 0 0 0)	Winner correctly predicted.
2006	Italy	P_6^2	(0 0.2 0.2 0.2 0.2 0 0.2)	Equal chances for Italy, Germany, Brazil, England and France.
2010	Spain	–	–	New winner.
2014	Germany	P_7	(0 0 1 0 0 0 0 0)	Winner correctly predicted.
2018	France	P_7^2	(0 0 0 0.5 0 0.25 0.25 0)	Winner not correctly predicted.
*2022		P_7^3	(0 0.23 0.23 0.35 0.1 0 0.1 0)	Brazil is predicted to win, followed by Italy or Germany.

The sequential Markov chain model for the FIFA World Cup winners is updated when a new winner emerges, and as such, the corresponding transition probability matrix for those editions where a new winner emerges cannot be computed. From the information available in table 4, when those competitions where a new winner emerges are left out, the model correctly predicted the winner of the World Cup 50% of the time (7 Competitions), predicted equal chances for two or more countries (inclusive of the true winner) about 36% of the time (5 Competitions) and only incorrectly predicted the winner 14% of the time (2 Competitions). If the final transition probability matrix P_7 had been used instead of the sequential approach, then it would have been impossible for the model to predict the winners of the competition from inception. In fact, the estimated probability vectors would have incorrectly predicted winners in most of the competitions and by the seventh edition (1962), the process would have significantly approached its stationary distribution, which would give the country with the most wins (Brazil) the highest probability of winning every subsequent competition.

For the sequential transition probability matrix P_i to be estimable, there must be at least one more competition where a new winner does not emerge, otherwise, there will be insufficient data, hence the TPM for the 3-state Markov chain P_2 could not be estimated because Brazil became a new winner in the 1958 competition after Germany emerged winner in the previous edition (1954).

A projection is made for the forthcoming 2022 World Cup on the basis of the sequential model which gives a high probability that Brazil will emerge victorious in the competition. The top three countries projected to win the 2022 World Cup are Brazil, Italy and Germany. These three countries are ordinarily the frontrunners in any World Cup competition and hold a chunk of the overall victories in the competition. However, on the basis of the process governing the sequential increment of the group of World Cup winners, it is predicted that a new winner may emerge in the 2022 competition, thus increasing the size of the state space from 8 to 9.

The transition probability matrix for the final 8-state Markov chain, which has been sequentially updated from the two-state model, is presented below. It shows the transitions between all the countries that have won the FIFA World Cup from inception to 2018.

$$P_7 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \begin{pmatrix} 0 & ½ & ½ & 0 & 0 & 0 & 0 & 0 \\ 1/4 & ¼ & 0 & 0 & 0 & 1/4 & 0 & ¼ \\ 0 & 0 & 0 & 2/4 & 0 & 1/4 & 1/4 & 0 \\ 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & ½ & ½ & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

The stationary or long-run probabilities (the row entries of matrix $\boldsymbol{Z}$) is given as

$$\boldsymbol{\pi}_7 = (0.04 \quad 0.17 \quad 0.17 \quad 0.32 \quad 0.06 \quad 0.09 \quad 0.11 \quad 0.04)$$

The corresponding mean recurrence times are given as

$$\mathbf{r}_7 = (23.5 \quad 5.9 \quad 5.9 \quad 3.1 \quad 15.7 \quad 11.8 \quad 9.4 \quad 23.5)$$

The mean recurrence times specifies the mean waiting time until a country wins the World Cup, having won it previously. It is seen that Brazil has the lowest waiting time, which is about 3 competitions between wins, followed by Italy and Germany with about 6 competitions each.

The mean first passage matrix of the 8-state Markov chain is given below:

$$\boldsymbol{M} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \begin{pmatrix} 0 & 5.1 & 6.5 & 5.3 & 20.0 & 11.8 & 10.6 & 19.4 \\ 18.4 & 0 & 5.0 & 6.1 & 20.8 & 9.6 & 11.4 & 14.3 \\ 26.6 & 8.3 & 0 & 2.6 & 17.3 & 11.9 & 7.9 & 22.5 \\ 26.0 & 7.6 & 6.0 & 0 & 14.7 & 14.3 & 8.4 & 21.9 \\ 27.0 & 8.6 & 7.0 & 1.0 & 0 & 15.3 & 9.4 & 22.9 \\ 23.5 & 5.1 & 3.5 & 5.3 & 20.0 & 0 & 10.6 & 19.4 \\ 27.0 & 8.6 & 7.0 & 1.0 & 16.0 & 15.3 & 0 & 22.9 \\ 27.6 & 9.3 & 1.0 & 3.6 & 18.3 & 12.9 & 8.9 & 0 \end{pmatrix} \end{array}$$

The mean first passage matrix contains the expected number of competitions a country will wait before winning the World Cup, given that a particular country won the immediate past edition. From the mean first passage matrix above, some observations are pertinent. The entries with 1, namely $\boldsymbol{M}_{54}$, $\boldsymbol{M}_{74}$ and $\boldsymbol{M}_{83}$ imply that there is little information from countries who had only won the completion once (England and Spain) or any current winner who had won it the second time (France). Since France is the current holder of the World Cup, on the basis of the mean first passage time matrix, it is expected that Brazil will probably win the 2022 competition.

While on the average, it takes 3 competitions before a new winner emerges from the historical data, the last new winner was Spain in 2010, hence it could be expected that a new winner may emerge in the 2022 World Cup. A possible new winner is Belgium, as the country has been exhibiting consistent performances over the years.

4. Conclusion

A sequential Markov chain has been used to model the process of winners emerging in the FIFA World Cup. The sequential transition probability matrices were estimated from the data of World Cup winners. Using an appropriate initial probability vector, it was shown that the sequential Markov chain model performed reasonably well in predicting winners of the competition, with a very few wrong predictions. On the basis of the projections for the 2022 World Cup, Brazil has the

highest probability of winning the competition, followed by Italy and Germany. However, there is also a possibility that a new winner may emerge in the 2022 World Cup, thus increasing the state space of the sequential Markov chain.

The sequential Markov chain model has the advantage that it can be updated after every competition and used to make subsequent prediction for the next competition. This is achieved by updating the transition probability matrix as appropriate depending on whether a new winner emerged or not.

Possible wider application of the model are scenarios in which the process takes only a small set of values after it has been observed for a long period and the other set of possible values have only a negligible chance of occurring. The state space process can then be viewed as a time-varying system.

Any sports competition that only a few countries or clubs had won over the lifetime of the competition could also be modelled via the Markov chain conceptualization of the FIFA World Cup winners.

References

BAKER, R. AND SCARF, P. (2006). Predicting the outcomes of annual sporting contests. Applied Statistics, 55 (2), 225 – 239.

CATTELAN, M.; VARIN, C. AND FIRTH, D. (2013). Dynamic Bradley-Terry Modelling of Sports Tournaments. Journal of The Royal Statistical Society, Series C (Applied Statistics), 62 (1), 135 – 150.

DYTE, D. AND CLARKE, S. R. (2000). A Ratings Based Poisson Model for World Cup Soccer Simulation. The Journal of the Operational Research Society, 51 (8), 993 – 998.

FEDERATION INTERNATIONALE DE FOOTBALL FEDERATION (FIFA). URL:www.fifa.com

GRINSTEAD, C. M. AND SNELL, J. L. (1997). Introduction To Probability. American Mathematical Society.

HOFFMANN, R.; GING, L. C. AND RAMASAMY, B. (2002). The Socio-Economic Determinants of International Soccer Performance. Journal of Applied Economics, 5 (2), 253 – 272.

JONES, M. A. (2004). Win, Lose, or Draw: A Markov Chain Analysis of Overtime in the National Football League. The College Mathematics Journal, 35 (5), 330 – 336.

KUONEN, M.; MORGENTHALER, S. AND CHAVEZ, E. (1997). Statistical Models for Knock-Out Soccer Tournaments. Ecole Polytechnic, Lausanne, Switzerland.

LAGO, C. (2007). Are Winners Different from Losers? Performance and Chance in the FIFA World Cup Germany 2006. International Journal of Performance Analysis in Sport, 7 (2), 36 – 47.

PAUL, S. AND MITRA, R. (2008). How Predictable are the FIFA World Cup Football Outcomes? An Empirical Analysis. Applied Economics Letters, 15, 1171 – 1176.

PERCY, D. F. (2015). Strategy Selection and Outcome Prediction in Sport Using Dynamic Learning for Stochastic Processes. Journal of the Operational Research Society, 66, 1840–1849.

RUMPF, M. C.; SILVA, J. R.; HERZOG, M.; FAROOQ, A. AND NASSIS, G. (2017). Technical and Physical Analysis of the 2014 FIFA World Cup Brazil: Winners Vs Losers. Journal of Sports Medicine and Physical Fitness, 57 (10), 1338 – 1343.

SCOPPA, V. (2013). Fatigue and Team Performance in Soccer: Evidence from the FIFA World Cup and the UEFA European Championship. IZA Discussion Paper No. 7519.

SETH, S. (2018). FIFA World Cup: Factors that explain the performances of National Football Teams. Claremont Colleges Thesis. http://scholarship.claremont.edu/cmc_theses/1919

SUZUKI, A. K.; SALASAR, L. E. B.; LEIT, J. G. AND LOUZADA-NETO, F. (2010). A Bayesian Approach for Predicting Match Outcomes: The 2006 (Association) Football World Cup. The Journal of the Operational Research Society, 61 (10), 1530 – 1539.

SZYMANSKI, S. (2003). The Economic Design of Sporting Contests. Journal of Economic Literature, 41 (4), 1137 – 1187.

TAN, B. AND YILMAZ, K. (2002). Markov Chain Test for Time Dependence and Homogeneity: An Analytical and Empirical Evaluation. Eur. Jour Op. Res., 137, 524 – 543.

TORGLER, B. (2004). The Economics of the FIFA Football World Cup, KYKLOS, 57, 287–300.

VOLF, P. (2009). A random point process model for the score in sport matches. IMA Journal of Management Mathematics 20(2): 121–131.

WIKIPEDIA. www.wikipedia.org/FIFA_World_Cup