Journal of Advance Research in Mathematics and Statistics (ISSN: 2208-2409)

DOI: https://doi.org/10.53555/nnms.v3i7.549 **Publication URL: http://nnpub.org/index.php/MS/article/view/549**

PREDICTING THE WINNER OF GAMES IN WORLD CUP SOCCER MATCHES

Mohammed Sylla[1] and Rhonda Magel[2]*

*[1,2]*Department of Statistics North Dakota State University Fargo, ND 58108 Email: Rhonda.magel@ndsu.edu*

Corresponding Author: -
Email: Rhonda.magel@ndsu.edu

Abstract:

In this research, statistical models were developed that can be used to predict the outcomes of World Cup soccer matches. Least squares regression and logistic regression techniques were used in the development of the models' using data from the 2006 World Cup Matches. The models were tested using data from the 2010 World Cup Matches. Predictions were made for the 2014 World Cup Championship assuming no results were known ahead of time.

Keywords: *Least Squares Regression; Logistic Regression; Point Spread Models; Win Probability Models*

I.

1. INTRODUCTION

World Cup soccer is a heavily viewed sporting event. It is estimated that it is broadcasted to 204 countries with over 715 million people watching [1]. Preliminary games are played among teams on all seven continents and from this, there are 32 teams selected to that are qualified for the World Cup, which occurs every four years [1]. A random drawing is conducted as explained in [2] for the 2014 World Cup in which the 32 teams are placed in eight groups of four teams each. The first round in the World Cup is the Round Robin. Each team in the group of four plays each other in this round and hence plays 3 games. A team is given 3 points for winning a game in this Round Robin; 1 point for a draw, and 0 points for a lost game. The two teams in each group of four that have the highest number of points at the end of the Round Robin make it to Round 2 also referred to as Round 16 (Knock out stage). Round 2 is followed by the quarter-final (Round 3), semi-final (Round 4) and then the final round (Round 5). If a team loses a game in any round after the Round Robin, they are out of the World Cup [1].

In this research, models were developed that estimate the number of points that each team scores in the Round Robin. Models were developed that estimate the goal margin of each game for Round 2 and then Rounds 3-5. Models were also developed for Round 2 and then Rounds 3-5 to estimate the probability that a particular team will win the game given the two teams playing. The models were developed by using data collected from the 2006 World Cup. They were validated by using the data from the 2010 World Cup and then used to predict the results for the 2014 World Cup.

Past research of forecasting soccer outcomes has involved both direct and indirect approaches. Koning [3] developed a logistic regression model to estimate the probability of a team having a win, draw, or loss based on some measure of the quality of the team, and whether or not the team was playing at home. Moroney [4] and Karlis and Ntzourfras [5] used an indirect approach and modeled the goals scored by each team playing in the match using a bivariate Poisson model. The model actually underestimated the number of draws in the Round Robin [5]. Reep and Benjamin [6] also used an indirect approach but modeled the number of goals scored by each team using a negative binomial distribution instead of using a bivariate Poisson distribution. Reep and Benjamin [6] found that 80% of the goals scored occurred after a sequence of 3 passes or less, giving evidence to the fact that goals scored are associated with the number of passes between players on the same team. This was based on data gathered from 3,213 soccer matches between 1953 and 1968. Croucher [7] studied the effect of a tiebreaker factor. McGarry and Schultz [8] investigated whether or not it was better for a team to be randomly placed in one of the 8 groups in the Round Robin over another. This is based on how the top 7 seeded teams are placed into the groups along with the host country. Magel and MeInykov [9] studied factors that were significant in predicting the outcomes of European soccer games. They found that differences between goals scored and goals against based on k previous games of both teams were significant. They also found the differences in cards received by both teams and their opponents during the last k games were also significant.

In this paper, we extended the findings of [9] to predict the results of the World Cup. This research focused on considering Goals Scored, Goals Against, and number of cards received in the last k games as well as a team's winning probability prior to entering the competition.

2. Model Development

For Round 1 (Round Robin), we developed a model using data from the 2006 World Cup to estimate the number of points that each team in the Round Robin would obtain for the three games that they would play. We predicted that the two teams with the highest number of estimated points in their groups would advance to Round 2 (or Round 16). Ordinary least squares regression was used to develop the point model for the Round Robin. Four variables were considered for entry into the model using the stepwise regression technique with an entry level of 0.25 and an exit level of 0.20. The four variables considered were the following: Average Goals Scored per game by the team before the 2006 World Cup (AvgGS_Game), the Average Goals Scored Against the team per game before the 2006 World Cup (AvgGA_Game), the Average number of disciplinary Cards received by the team per game before the 2006 World Cup (Ave_Cards), and the winning probability of the team before the 2006 World Cup (WinP). All of the variables were found to be significant at alpha equal to 0.05. The R-Square for the model was 0.92 with the adjusted R-Square equaling 0.91. The intercept term was set equal to 0. The data was taken from [1] and included games played between August 18, 2004 and November 16, 2005 which is when the preliminaries took place for the 2006 World Cup.

The equation for estimating the number of points that a team will receive in the Round Robin of the World Cup is given by the following:

$$\hat{y}\ (\text{estimated number of points}) = 3.5105 \times \text{AvgGS} - 2.0834\,\text{AvgGA} + 0.4582 * \text{Ave Cards} + 2.4101 * \text{WinP} \quad \textbf{(1)}$$

In order to predict the teams that made it past Round 1, we also tried developing a goal margin model to predict the difference between goals made of the two teams playing in each game of the Round Robin. Stepwise regression was used with an alpha entry of 0.25 and an alpha stay of 0.20. Four variables were considered for entry into the model with those four variables being the differences between each of the four variables used in the point model development between the two teams. One team in each game was randomly designated as "Team A" and the other team as "Team B". The differences were all taken in the order of "Team A" minus "Team B" and the estimated point spread was found in the same order. Only two of the four variables were found to be significant and they both had p-values less than 0.01. These variables included the differences in the average goals scored per game by each team and the average goals per game scored against each team. The R^2 for the goal margin model for Round 1 was 0.63 with the adjusted R^2 being 0.59. This model was not further used because it did have a lower R^2 value than the point model and it did not directly predict the

two teams with the highest number of points, but it predicted individual winners for each game and these had to be combined together further to predict the teams in each group with the highest number of points.

Round 2 (Round 16) is a knock-out stage. If a team loses a game in this round, they are out of the tournament. If a team wins the game, they go to Round 3. Two models were developed for this round to determine which teams would go on to the next round. The first model was a goal margin model which was developed to estimate the difference in goals scored between the two teams playing in a game. There were eight variables considered for entry into the model. These eight variables all involve differences between two teams in the order of "Team A" minus "Team B" with the goal margin being predicted in the same order. The intercept term was set to zero and one of the two teams playing in each game of this round was randomly selected to be "Team A" with the other team being "Team B". The eight variables considered for the model were the following with only data on teams making it to this round in World Cup 2006 being used:

- Difference in Average Goals Scored between two teams before World Cup 2006
- Difference in Average Goals Scored Against two teams before World Cup 2006
- Difference in Average disciplinary Cards given by a referee before World Cup 2006
- Difference in Average winning percentage between two teams before World Cup 2006
- Difference in Average Goals Scored during Round Robin of the World Cup 2006
- Difference in Average Goal Scored against during Round Robin of the World Cup 2006
- Difference in Average disciplinary cards given during Round Robin of the World Cup 2006
- Difference in Average Number of wins in the Round Robin World Cup 2006.

Three variables were found to be significant using the stepwise regression technique with the same entry and exit levels used as before: Difference in Average Goals Scored during Round Robin of 2006 World Cup (ADiffGS); Difference in Average Goals Scored against during Round Robin of 2006 World Cup (ADiffGA), Difference in Average disciplinary cards given during Round Robin of 2006 World Cup (ADiffCard**s)**. The model had an adjusted R^2 of 0.67 and is given in equation (2).

$$\hat{y}\text{ (estimated point spread)} = 2.0226\text{xADiffGS} - 0.9351\text{xADiffGA} - 1.1514\text{xADiffCards} \quad \textbf{(2)}$$

The second model developed for Round 2 was a logistic regression model designed to estimate the probability of "Team A" winning the game. The same set of eight variables considered in the development of the goal margin model for Round 2 were also considered for this model. Stepwise regression was used to help in the development of the model. The same three variables were found to be significant as in the goal margin model for this round. Hosmer-Lemeshow (HL) was used to assess the goodness of fit, where the null hypothesis indicates that our current model fits well and the alternative hypothesis indicates the model does not fit well. The p-value for HL was 0.544 meaning that we do no reject the null hypothesis and we conclude that the model fit is fine. An ROC curve was graphed for this model and the area under the curve was 0.78. An ROC of 0.78 implies this model is acceptable for determining which team won the game [10].

The model estimated probability of a win by Team A using this model is then given by

$$p(win) = \frac{1}{1+e^{-\hat{y}}} \quad \text{where } \hat{y} = 1.2135\text{xADiffGS} - 2.5469\text{xADiffGA} - 0.6724*\text{ADiffCards} \quad \textbf{(3)}$$

Two models were developed that encompassed Rounds 3-5 together. The models were again a goal margin model and a logistic regression model. Past World Cup 2006 data were used to develop a model [1]. Again, the stepwise technique was used with a select entry of 0.25 and select stay of 0.20. The following variables were considered for entry into the model:

- Difference in average Goals Scored between two teams up to this present round in the World Cup
- Difference in average Goals Scored against between two teams up to this present round in the World Cup
- Difference in average Cards received between two teams up to this present round in the World Cup The variables found to be significant were the following: Difference in average Goals Scored between two teams up to this present round in the World Cup(AdiffGS) and Difference in average Goals Scored against between two teams up to this present round in the World Cup (AdiffGA). Our adjusted R-square value was 0.62 and the model is given below:

$$\text{where } \hat{y}\text{ (estimated point spread)} = 1.0067\text{xADiffGS} - 0.7044\text{xAdDffGA} \quad \textbf{(4)}$$

All of the three variables considered were found to be significant for the logistic model with the difference in average cards received between the two teams being denoted by ADiffCards. The Hosmer-Lemeshow test [10] was used to assess the goodness of fit for the logistic regression model. The p-value for HL was found to equal 0.29 meaning that we do no reject the null hypothesis and we concluded that the model is a good fit. An ROC curve was graphed for this model and the area under the curve was found to equal 0.86 which indicates this model is excellent in determining which team won the game [10].

The estimated probability of a win by "Team A" using this model is given by

$$p(win) = \frac{1}{1+e^{-\hat{y}}} \quad \text{where } \hat{y} = 0.7813xADiffGS-1.5953xADiffGA-0.4062*ADiffCards \qquad \textbf{(5)}$$

3. Model Validation

In order to validate the point model for the Round Robin, we used the 32 teams that qualified for World Cup 2010 and the data associated with these teams. The 32 teams were divided into 8 groups of 4 teams each with the groups being labeled with letters from A to H. The point model was used to predict the number of points that each team in a group would receive during the Round Robin based on the four variables found to be significant in the point model: Average Goals scored by the team per game before World Cup 2010; Average Goals scored against the team per game before World Cup 2010; Average number of cards per game received by the team before World Cup 2010; and the win probability of the team before World Cup 2010. As an example, as to how the model was applied, we will consider the teams in Group G. Data for the four significant variables was collected for each of the teams and is given in Table 1. The data was then placed in equation **(1)** for each game to predict the number of points the team would acquire during the Round Robin.

- Predicted number of points (Brazil) = 3.5105 x (1.83)-2.0834x (0.61) +0.4582x (6.33) +2.4011x (0.5) = 9.26
- Predicted number of points (Portugal) = 3.5105x(1.7)- 2.0834x(0.5) +0.4582x (4.33) +2.4011x (0.5) = 8.12
- Predicted number of points (Korea DPR) = 3.5105 x (0.875)- 2.0834x(0.625) +0.4582x (8.33) +2.4011x (0.3) = 6.31
- Predicted number of points (Ivory Coast) = 3.5105 x (1.16)-2.0834x (0.66) +0.4582x (4.66) +2.4011x (0.8) = 6.76

In group G, our model predicted both Brazil and Portugal to qualify having the most points with 9 and 8 points, respectively. Both teams did actually qualify with Brazil receiving 7 points and Portugal receiving 5 points. The results for all the groups are given in Table 1. The model was correct on 23 of the 32 teams for an accuracy of 71.9%.

We next went on to validate both models developed for Round 2. As an example, as to how the goal margin model for Round 2 was validated, we considered one game played in Round 2 of World Cup 2010 between Uruguay and South Korea. The data is given in Table 2 and placed into equation **(2).**

- Goal Margin estimate (Uruguay vs South Korea) = 2.0226x(-0.33)-1.1514x (2)-0.9351x (2.0226) = 2.26 ≈2 (Uruguay)

The Goal Margin model estimates number of goals scored by "Team A" (first team listed) minus number of goals scored by "Team B"(second team listed). When the Goal Margin estimate is positive it is predicted "Team A" wins. When the Goal Margin estimate is negative, it is predicted "Team B" wins. Our model predicted that Uruguay will win by 2 goals. Uruguay did win, but won with a difference of 1 goal. All of the results for this round are given in Table 2. Out of eight games our model correctly predicted six of teams which would win the game. In this case, the correct prediction percentage was 75%.The logistic regression model for Round 2 was also validated. To illustrate this validation process, we considered the 2010 World Cup game between Uruguay and South Korea. Data from this game was collected and given in Table 3. The data was placed in equation **(3).** The results were the following:

- $$P(\text{Uruguay}) = \frac{}{1+\exp(-1.235x-0.33+2.5469x-0.67+0.6724x-2)}$$

$$=0.93$$

Since the estimated probability that Uruguay will win the game is greater than 0.5, our model is predicting Uruguay to be the winner and Uruguay did win the game. All of the results for this round are given in Table 3.
There were 5 games out of 8 that this model predicted correctly for an accuracy of 62.5%.

Table 1: Validation Results from Point Model Round Robin 2010

Teams 2010	Predicted number of points	Actual number of points	Predicted to Qualified Y/N	Actually Qualified Y/N	AV_GS	AV_GA	AV_CARDS	Winp
Group A								
South Africa	4.36	4	N	N	1	1	1.66	0.9
Mexico	6.64	4	N	Y	1.8	1.2	3	0.6
Uruguay	6.56	7	Y	Y	1.55	1.11	4.33	0.6
France	6.96	1	Y	N	1.8	0.9	2.33	0.6
Group B								
Argentina	6.15	9	Y	Y	1.27	1.11	4	0.9
Nigeria	6.01	1	N	N	1.5	0.66	2	0.5
Korea Republic	6.35	4	N	Y	1.5	0.5	2	0.5
Greece	7.14	3	Y	N	2	1	1.66	0.6
Group C								
England	14.08	5	Y	Y	3.4	0.6	2.667	0.9
USA	9.22	5	Y	Y	1.9	1.3	8.33	0.6
Algeria	10.21	1	N	N	1.5	0.66	10.33	0.66
Slovenia	10.20	4	Y	N	2.2	1	6.27	0.7
Group D								
Germany	11.69	6	Y	Y	2.6	0.5	3.66	0.8
Australia	8.23	4	N	N	1.5	0.125	3.36	0.7
Serbia	8.73	3	N	N	2.2	0.8	3.2	0.5
Ghana	8.72	4	Y	Y	1.5	0.5	6.66	0.6
Group E								
Netherlands	10.24	9	Y	Y	2.125	0.25	3	0.8
Denmark	8.16	3	N	N	1.6	0.5	4.66	0.6
Japan	6.60	6	N	Y	1.375	0.75	4.66	0.5
Cameroon	8.77	0	Y	N	1.5	0.33	6	0.6
Group F								
Italy	9.60	2	N	N	1.8	0.7	6.66	0.7
Paraguay	8.68	5	N	Y	1.33	0.833	8.33	0.8
New Zealand	10.51	3	Y	N	2.33	0.833	4.66	0.8
Slovakia	9.77	3	Y	Y	2	1	6.33	0.8
Group G								
Brazil	9.26	7	Y	Y	1.83	0.61	6.33	0.5
Korea DPR	6.31	0	N	N	0.875	0.625	8.33	0.3
Ivory Coast	6.76	4	N	N	1.16	0.66	4.66	0.8
Portugal	8.12	5	Y	Y	1.7	0.5	4.33	0.5
Group H								
Spain	12.94	6	Y	Y	2.8	0.5	4.33	0.9
Switzerland	8.54	4	N	N	1.8	0.8	5.33	0.6
Honduras	7.33	1	N	N	0.7	1.1	13	0.5
Chile	8.67	6	Y	Y	2	1	5	0.6

Table 2: 2010 Validation Results from the Goal Margin Model-Round 2

Team A vs B 2010	Goal A	Goal B	Actual Goal Margin(A -B)	Estimate Goal Margin	Predicted Team to Win	Actual Team which Won	ADiffCards	ADiffGA	ADiffGS
Uruguay vs South Korea	2	1	1	2.26	Uruguay	Uruguay	-2	-0.67	-0.33
United States vs Ghana	1	2	-1	0.94	United States	Ghana	0.34	0.02	0.67
Nethelands vs Slovania	2	1	1	1.38	Netherlands	Netherlands	0.33	-1.165	0.33
Brazil vs Chile	3	0	3	1.60	Brazil	Brazil	0.17	-0.49	0.66
Argentina vs Mexico	3	1	2	2.91	Argentina	Argentina	0.69	-1.0833	1.33
Germany vs England	4	1	3	3.04	Germany	Germany	-0.9	0	0.99
Paraguay vs Japan	5	3	2	0.30	Paraguay	Paraguay	-0.327	0.8	0.33
Spain vs Portugal	1	0	1	-3.01	Portugal	Spain	0.66	0.247	-1

Table 3: 2010 Validation Results from the Logistic Regression- Round 2

Team A vs B 2010	ADiffCards	ADiffGA	ADiffGS	Estimated Probability of Winning Team A	Estimated Probability of Team B	Predicted to Advanced	Actual Team which Won
Uruguay vs South Korea	-2	-0.67	-0.33	0.93	0.06	Uruguay	Uruguay
United States vs Ghana	0.34	0.02	0.67	0.63	0.37	United States	Ghana
Nethelands vs Slovania	0.33	-1.165	0.33	0.96	0.04	Netherlands	Netherlands
Brazil vs Chile	0.17	-0.49	0.66	0.87	0.13	Brazil	Brazil
Argentina vs Mexico	0.69	-1.0833	1.33	0.98	0.01	Argentina	Argentina
Germany vs England	-0.9	0	0.99	0.86	0.14	Germany	Germany
Paraguay vs Japan	-0.327	0.8	0.33	0.19	0.80	Japan	Paraguay
Spain vs Portugal	0.66	0.247	-1	0.09	0.91	Portugal	Spain

We used the past data of the World Cup 2010 [1] to validate our Goal Margin model for Rounds 3-5. We correctly predicted 5 out of 7 games for a 71% correct prediction rate. Table 4 gives values of the significant variables needed for the equation between each of the two teams playing. As an example, consider the game between Uruguay and Ghana and use equation **(4)** with data given in Table 4.

- Estimated Goal Margin (Uruguay vs Ghana) = 1.0067x(ADiffGS=0.5)-0.7044x (AdiffGA=-0.5)=0.86 (Uruguay)

It is predicted the Uruguay will win since the goal margin is positive. Uruguay did win the game. All of the results are given in Table 4 for the goal margin model for Rounds 3-5.

We validated the logistic model for use with Rounds 3-5 by using past data of the 2010 World Cup [1]. The values of the variables found to be significant in the model given by equation **(5)** were found between the two teams playing. As an example of how the model was used, we will consider the 2010 World Cup game between Netherlands and Brazil. The values of the variables are given in Table 5 for this game as well as all of the games. Since the estimated probability of the Netherlands winning was greater than 0.5, we predicted a win for the Netherlands and they did win. We correctly predicted 5 out of 7 games for 71 % correct prediction rate. All of the results are found in Table 5.

- $P(\text{Netherlands}) = \dfrac{1}{1+\exp(-0.7813x-0.25+1.5953x-0.25+0.406x-1)} = 0.65$

Table 4: Validation from Goal Margin –Rounds 3-5

Rounds	Team A-2010	Average Goal For A	Teams B-2010	Average Goal for B	Difference In Average GS between Team A and B	Difference In Average GA between Team A and B	Estimated Goal Margin	d Team to win	Teams who won
3	Uruguay	1.5	Ghana	1	0.5	-0.5	0.86	Uruguay	Uruguay
3	Netherlands	2.25	Brazil	2	0.25	0	0.25	Netherlands	Netherlands
3	Argentina	2.5	Germany	2.75	-0.25	0	-0.25	Germany	Germany
3	Paraguay	0.75	Spain	1.25	-0.5	0.5	-0.86	Spain	Spain
4	Uruguay	1.5	Netherlands	2.25	-0.75	0.25	-0.93	Netherlands	Netherlands
4	Germany	2.75	Spain	1.25	1.5	0	1.51	Germany	Spain
5	Netherlands	2.25	Spain	1.25	1	0	1.01	Netherlands	Spain

Table 5: Validation for Logistic Regressionfor Rounds 3-5

Rounds	Teams A vs B 2010	AdiffGS	AdiffGA	AdiffCards	Estimated probability of Winning Team A	Estimated probability of Winning Team B	Predicted to advanced	Actual Team who won
3	Uruguay vs Ghana	0.5	-0.5	-0.5	0.80	0.20	Uruguay	Uruguay
3	Netherlands vs Brazil	0.25	0	-1	0.65	0.35	Netherlands	Netherlands
3	Argentina vs Germany	-0.25	0	-2.25	0.67	0.33	Argentina	Germany
3	Paraguay vs Spain	-0.5	0.5	0.25	0.21	0.78	Spain	Spain
4	Uruguay vs Netherlands	-0.75	0.25	0.5	0.23	0.73	Netherlands	Netherlands
4	Germany vs Spain	1.5	0	3	0.49	0.51	Spain	Spain
5	Netherlands vs Spain	1	0	0.25	0.66	0.33	Netherlands	Spain

4. Model Prediction

Our last step was using the models to predict the winner of the World Cup from the beginning before any games began. We predicted the winner of the 2014 World Cup starting with the teams in 2014 who qualified for the World Cup match and with the knowledge of which group they were assigned to for the Round Robin. We estimated the number of points each team would acquire using equation (1) and the values of the significant variables of all teams using the 2014 data. From the point model, we predicted which of the two teams in each group would go on to Round 2. We used the goal margin model to predict which teams would make it past Round 2. We then used the goal margin model developed for Rounds 3-5 to determine the winner. The goal margin models were used instead of the logistic regression models since the goal margin model for Round 2 did slightly better than the logistic regression model in the validation phase and both models for Rounds 3-5 had the same correct percentage in the validation phase. Using this method, data on teams predicted to make it to Round 2 based on the point model and the predicted matches were placed into the goal margin model for Round 2 (equation (2)). The data on teams predicted to win in Round 2 using the model in equation (2) and play against each other in Round 3 were placed into equation (4). In Round 2, the matches are decided by A1vs B2; C1vs D2; E1 vs F2; G1 vs H2; B1 vs A2; D1 vs C2; F1 vs E2; and H1 vs G2 where A1 represents the team in group A with the greatest number of points acquired during the Round Robin and B2 represents the team with the 2nd highest number of points in Group B during the Round Robin [11]. The others are defined similarly. Data on teams predicted to win in Round 3 and play each other in Round 4 was placed into equation (4) and then data on the two teams predicted to play each other in Round 5 was placed into equation **(4)** and a winner predicted. The predicted results for the Round Robin are given in Table 6. The predicted results for Round 2 using the Goal Margin Model are given in Table 7. Table 8 gives the predicted results for Rounds 3-5. Germany was predicted to win the World Cup and they did win.

Table 6: Predicted Results for Round Robin 2014 using Point Model

Teams 2014	Predicted number of points	Actual number of points	Predicted to Qualify Y/N	Actually Qualified Y/N	AV_GS	AV_GA	AV_CARDS	Winp
Group A								
Brazil	12.60	7	Y	Y	3	0.4	1.6	0.9
Croatia	4.73	3	N	N	1.2	0.9	2.6	0.5
Mexico	10.73	7	Y	Y	2.5	0.66	2	1
Cameroon	6.21	0	N	N	1.33	0.5	2.16	0.66
Group B								
Spain	7.54	3	Y	N	1.75	0.375	0.8	0.75
Netherlands	13.89	9	Y	Y	3.4	0.5	1.8	0.9
Chile	5.45	6	N	Y	1.81	1.56	2.18	0.56
Australia	5.87	0	N	N	1.5	0.875	3.33	0.375
Group C								
Colombia	10.16	9	Y	Y	2.5	0.83	3.33	0.66
Greece	8.09	4	Y	Y	2	0.625	1.5	0.7
Ivory Coast	6.63	3	N	N	1.68	0.81	2.33	0.56
Japan	8.09	1	N	N	2	0.625	1.5	0.7
Group D								
Uruguay	4.46	6	N	Y	1.56	1.56	2.62	0.43
Costa Rica	6.85	7	Y	Y	1.3	0.7	5.53	0.5
England	12.14	1	Y	N	3.1	0.4	1.4	0.6
Italy	6.84	3	N	N	1.9	0.9	1.3	0.6
Group E								
Switzerland	7.32	6	Y	Y	1.7	0.6	2	0.7
Ecuador	4.34	4	N	N	1.25	1	2.1875	0.43
France	7.39	7	Y	Y	1.875	0.75	1.7	0.66
Honduras	4.70	0	N	N	1.3	1.2	3.66	0.4
Group F								
Argentina	7.97	9	Y	Y	2.18	0.93	1.973	0.56
Bosnia	11.62	3	Y	N	3	0.6	0.9	0.8
Iran	4.65	1	N	N	1	0.25	0.33	0.625
Nigeria	4.92	4	N	Y	1.16	0.5	1.5	0.5
Group G								
Germany	13.46	7	Y	Y	3.6	1	1.6	0.9
Portugal	7.51	4	N	N	2	0.9	2	0.6
Ghana	12.57	1	Y	N	3	0.5	2.33	0.833
USA	7.81	4	N	Y	1.83	0.83	4.167	0.5
Group H								
Belgium	7.96	9	Y	Y	1.8	0.4	1.2	0.8
Algeria	9.31	4	Y	Y	2.16	0.66	2.4	0.83
Russia	8.17	2	N	N	2	0.5	1.1	0.7
Korea Rep.	6.53	1	N	N	1.625	0.875	3.16	0.5

Table 7: Predicted 2014 Results from the Goal Margin Model – Round 2

Team 2014 Round 2 Results	AdiffGS	AdiffCards	AdiffGA	Estimate Goal Margin	Predicted Team to win	Actual Results
Brazil vs Spain	1	-1.67	1	2.43	Brazil	Brazil
Netherlands vs Mexico	2	0.67	0.5	2.84	Netherlands	Netherlands
Colombia vs Costa Rica	0	1.34	2	-3.56	Costa Rica	Costa Rica
England vs Greece	0	-0.67	-2.3	3.27	England	X
France vs Argentina	0.33	-0.33	1.67	-0.95	Argentina	Argentina
Bosnia vs Switzerland	-1	0.67	-1.4	-1.04	Switzerland	X
Germany vs Russia	0.95	-1.7	-0.50	4.35	Germany	Germany
Ghana vs Algeria	1.1	-1.90	1	3.48	Ghana	X

Table 8: Predicted Results from Goal Margin Model Rounds 3-5 2014

Round	Team A 2014	AGS Team A	Team B 2014	AGS Team B	ADiffGS	AGA Team A	AGA Team B	ADiffGA	Estimated Goal Margin	Team Predicted to Win	Team who Won
3	Netherlands	3	Costa Rica	1.25	1.75	1	0.5	0.5	1.41	Netherlands	Netherlands
3	Argentina	1.75	Ghana	1.1	0.65	0.75	0.95	-0.2	0.80	Argentina	Argentina
3	Germany	3.25	Switzerland	1.75	1.5	0.75	1.5	-0.75	1.51	Germany	Germany
3	Brazil	2	England	0.5	1.5	0.75	1	-0.25	1.51	Brazil	Brazil
4	Netherlands	3	Argentina	1.75	1.25	1	0.75	0.25	1.26	Netherlands	Argentina
4	Germany	3.25	Brazil	2	1.25	0.75	0.75	0	1.26	Germany	Germany
5	Germany	3.25	Netherlands	3	0.25	0.75	0.75	0	0.25	Germany	Germany

5. Conclusion

A method was developed that could be used to predict future World Cup soccer matches and the Champion. The method used a point model in the Round Robin to predict which teams would make it to Round 2 and play each other in that round. Based on the predicted matches for Round 2, another model was used to estimate the point spread of a game played between the two teams in a predicted match in this round and from this, it was predicted which teams would make it to Round 3. Another model was used for Round 3, and then Rounds 4-5, to estimate what the point spread would be in a game between the two teams predicted to play each other in one of these rounds. Each of the models did well in the validation stage and the overall method did predict Germany as the winner in the prediction stage.

6. References

[1].*The Official Website of the FIFA World Cup™ –FIFA.com*. Retrieved March 7, 2014, from http://www.fifa.com.
[2].https://en.wikipedia.org/wiki/2014_FIFA_World_Cup_seeding. Retrieved May 26, 2016.
[3].Koning, R. H. (2000). Balance in competition in Dutch soccer. *Journal of the Royal Statistical Society: Series D (The Statistician), 49*, 419-431. doi: 10.1111/1467-9884.00244
[4].Moroney, M. J. (1956). *Facts from figures* (3rd ed.). London: Penguin.
[5].Karlis, D., & Ntzoufras, J. (2003). Analysis of sports data using bivariate poisson models. *The Statistician, 52*, 381-393.
[6].Reep, C., & Benjamin, B. (1968). Skill and chance in association football. *Journal of the Royal Statistical Society, 134*, 581-585.
[7].Croucher, J. S. (1984). The effect of changing competition in the English football league. *Teaching Statistics, 6*, 39-42.

[8].McGAarry, T., & Schutz, R. W. (1994). Analysis of the 1986 and 1994 World Cup Soccer tournament. In *ASA Proceedings from the 1994 Joint Statistical Meeting in Toronto, Statistics in Sports*, pp. 61-65.
[9].Magel, R., & Melnykov, Y. (2014, June). Examining influential factors and predicting outcomes in European soccer g*ames. International Journal of Sports Science*, *4*(3), 91-96.
[10]. Hosmer, David and Lemeshow, Stanley (2000). Applied Logistic Regression – Second Edition, John Wiley & Sons: New York, New York.
[11]. Team playing orders. Retrieved fromhttp://worldsoccertalk.com/wp-content/uploads/2014/07/world-cup-bracket-new.pdf