

Predictive Power of Elo Rating Systems and Markovian System on Association Football

Xiaochen Yang

1. Introduction

This independent study is about the predictive power of Elo rating systems and other ranking procedures in sports. Rating system has been a topic of interest. One reason is that a reasonable rating system could provide appropriate seeding, grouping players of similar skill levels. Besides, many rating systems also yield prediction on a specific match, which could be used in betting and pricing.

The Elo rating system is a widely accepted rating system that is used to quantify the skill levels of game players, and it was initially created by Arpad Elo, a Hungarian-born American physics professor. The Elo system was initially applied to rank the chess players, however, it gradually was applied to many other sports and games, including baseball, basketball, and association football. The Elo rating system can also be used to generate an expected winning probability of a team, once given the team's Elo rating and the opponent's rating. This is one reason why we are interested in the Elo system - it could be extended to generate prediction of a match outcome.

We decide to study the performance of Elo rating system on the association football dataset. One reason is that FIFA (Fédération Internationale de Football Association) maintains monthly world ranking for men's association football and quarterly world ranking for women's association football. FIFA has different ranking procedures for men's and women's football, however, both ranking procedures could be considered as Elo variants. In addition to the Elo system, we are also curious about other rating systems. One system being studied this time is the Markovian rating system, which utilizes the equilibrium distribution of winning probabilities. Details about both systems will be explained in Section 4.

2. Data Collection and Cleaning

Data were collected from various sources. FIFA does have all past ratings for men's football, however, the ratings website has missing data problem. Therefore all the FIFA men's official rating releases are collected from a personal website. In addition, the exact date of the monthly release is necessary, since the Elo rating would be updated on a daily basis. These dates are collected from FIFA website.

Ratings cannot be computed without actual match records. Initially, data of international matches involving European teams were collected from the online European Football Database, [however, later, another more comprehensive online database was found at <http://www.eloratings.net/>. The initial data inspection was done using the European Football Database, however, the study of ranking procedure was done using the [eloratings.net](http://www.eloratings.net/) database.

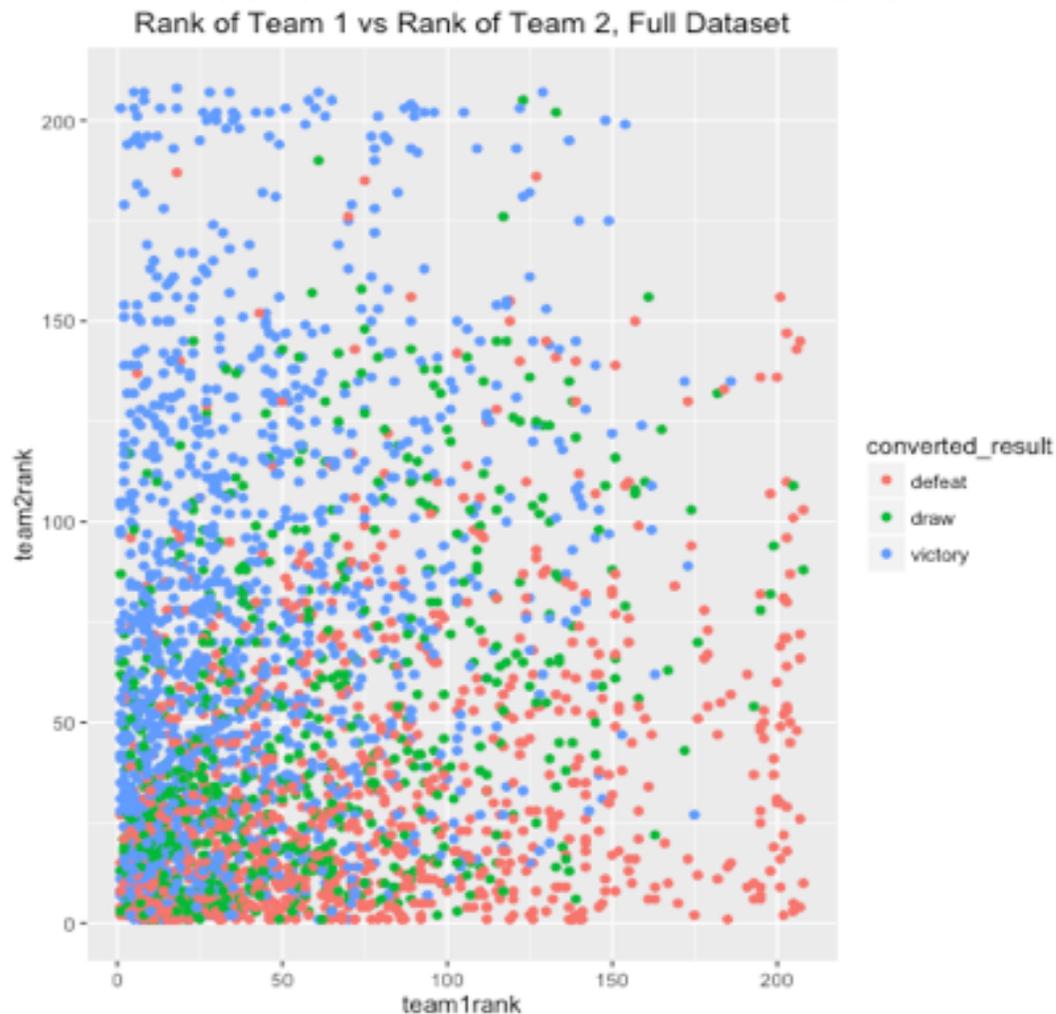
The dataset from the [eloratings.net](http://www.eloratings.net/) database consists of all international matches documented since June 12, 2006, when FIFA last updated the men's world ranking procedure. The matches include World Cup, intercontinental confederations, qualifiers of former two, other regional tournaments and friendlies. In addition to the match outcomes, the goals of each match are also documented and would be used in some versions of Elo rating systems.

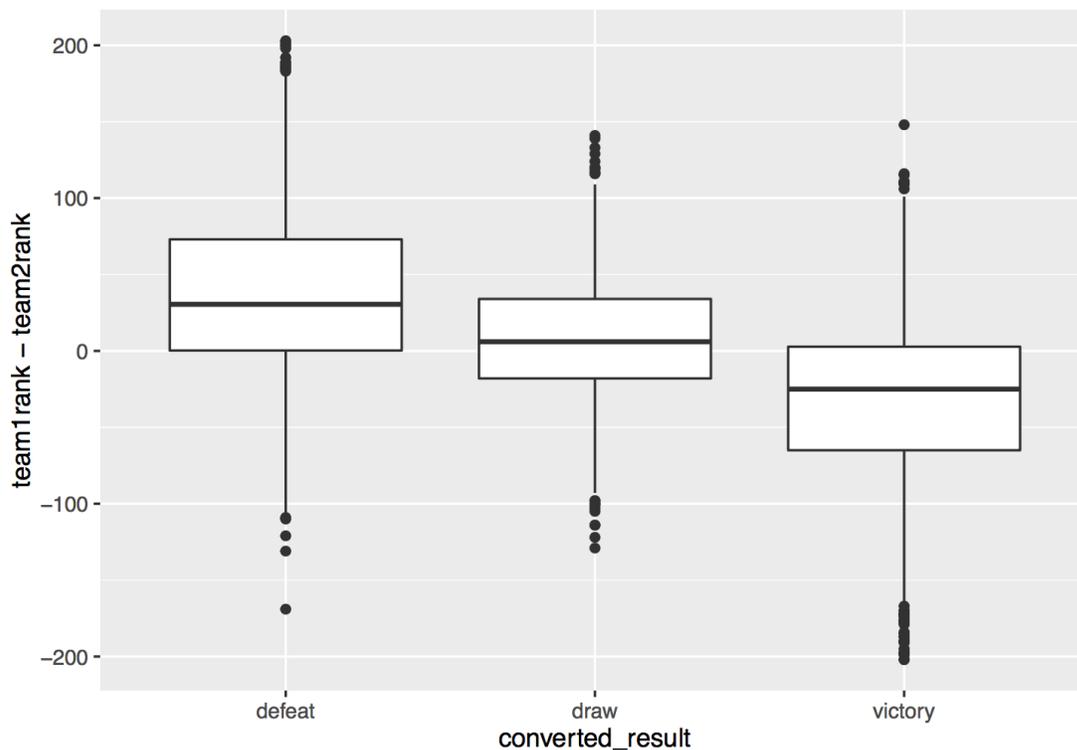
Inspired by various articles discussing home advantage in sports, including football, basketball and baseball, also a result of initial data inspection, we decide to incorporate home-away information in the dataset, believing this would add more information. Often home team

would have a slight advantage over the away team. [6 any paper] The elratings.net database contains the location of the match, and the home/away team information is retrieved according to the country of the match location. There are matches played on neutral grounds, meaning that neither of the participating teams is playing on home ground. We examined the performance of the Elo system incorporating home advantage and the one without home advantage.

3. Initial Inspection

Initially we attempted to look for association between team rating and the match result. After looking at the scatterplot of the dataset obtained from the European Football Database, we decided to try fitting linear models to the dataset, since there may be a pattern in the plot - a team with higher ranking (having higher Elo ratings than many players) tend to win when the opponent team has a lower ranking, and when a match ended with a draw, the two participating teams tend have similar ratings. The boxplot of difference in team ranking grouped by match result also indicates there may be a pattern about team ranking and game outcome, as when a team wins (victory group in the boxplot), it tends to have a ranking higher than its opponent.





Several linear models were used, including ordinary least squares, binary logistic regression, and ordinal logistic regression. Binary logistic regression does not handle the draw scenario, which is not uncommon in football matches. Here are some model summaries:

Binary logistic regression with two covariates:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.118659	0.085468	1.388	0.165
team1rank	-0.023724	0.001334	-17.783	<2e-16 ***
team2rank	0.028572	0.001536	18.597	<2e-16 ***

Binary logistic regression with interaction terms:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.042213	0.121296	0.348	0.728
team1rank	-0.024286	0.001764	-13.766	<2e-16 ***
team2rank	0.030189	0.002069	14.588	<2e-16 ***
converted_comp1	0.160945	0.171034	0.941	0.347
team1rank:converted_comp1	0.001995	0.002739	0.729	0.466
team2rank:converted_comp1	-0.004144	0.003143	-1.318	0.187

Ordinal logistic regression with two covariates:

	Value	Std. Error	t value	p value
team1rank	-0.01893907	0.0009105965	-20.798531	4.462712e-96
team2rank	0.02249665	0.0010198149	22.059541	7.736409e-108
defeat draw	-0.72099907	0.0681731896	-10.575991	3.850831e-26
draw victory	0.46929807	0.0672573364	6.977649	3.001611e-12

Ordinal logistic regression with interaction terms:

	Value	Std. Error	t value	p value
team1rank	-0.020067546	0.001202775	-16.684368	1.702999e-62
team2rank	0.025592679	0.001415140	18.084907	4.190749e-73
converted_comp1	0.200006046	0.128963693	1.550871	1.209326e-01
team1rank:converted_comp1	0.003610334	0.001830415	1.972413	4.856249e-02
team2rank:converted_comp1	-0.007353560	0.002039727	-3.605168	3.119507e-04
defeat draw	-0.624655617	0.093315643	-6.694008	2.171402e-11
draw victory	0.567483886	0.093184540	6.089893	1.129861e-09

Indeed some coefficient estimates are statistically significant, however, they are not very informative and do not exactly match the possible pattern indicated by the scatterplot and boxplot. Therefore I attempted to fit models to a subset of the full dataset. This subset includes only matches where both participating teams are from Europe. I chose to use such a dataset in the hope to reduce the heterogeneity in the dataset. However, the fitted models have similar problems. That is, the coefficients have small absolute value and the interpretability is limited.

Thanks to the advice from Professor David Aldous, I noticed that the focus should be the predictive power of a rating system, and the attribution of a certain covariate may not be of the most importance. Next section covers details about different rating systems and their performances on the association football dataset.

4. Rating System Comparison

4.1 Overview of Elo rating system

Elo rating system is an ‘earned rating’ system, in the sense that there is an exchange of rating points when two players play against each other. The winner obtains a number of rating points according to its previous rating and the opponent’s rating. The loser loses the same amount of rating points. Before a match begins, the expected winning probability, also known as the expected score, of a player can be computed once the other player’s rating point is known. A formula that is widely used is:

$$P_H = \frac{1}{1 + c^{-(a+R_H-R_A)/d}}$$

c is a constant of choice, a is a constant representing the amount of home advantage, R_H is the rating of the home team, and R_A is the rating of the away team. Both ratings are ratings before the match. If the match happens on neutral ground, then a is set to 0. d is a scaling factor, which usually takes value 400. Common choices of c include e , Euler’s constant, and 10. All the Elo rating systems below are implemented using $c = 10$. The expected score for the away team is 1 minus the expected score for the home team.

$$P_A = 1 - P_H$$

The rating for a player is updated after the match, dependent on the match outcome. The rating update formula is

$$R'_H = R_H + K * (S_H - P_H)$$

The R'_H is the updated rating of the home team, R_H is the rating of the home team before the match. K is a factor controlling the amount of rating point exchange, also affecting the speed of a team's rating change. S_A is actual outcome of the match and takes the value of 0, 0.5, or 1, which corresponds to loss, draw, or victory, respectively. P_A is the expected score computed based on the ratings of both teams.

4.1.1 The Classic Elo System

The first Elo rating system we tested is the most “basic” one, in the sense that the K-factor remains constant for all matches. In some Elo rating system variants, the K-factor for one specific match is adjusted according to the competition type and the goal difference on that match. Some of the variants would be discussed later - the elratings.net version, FIFA men's rating system and FIFA women's rating system. Home advantage is included for all Elo rating systems, and we choose the conventional value 100 for the home advantage constant for all Elo systems. Following is the expected score formula and the update formula for this classic Elo system.

$$P_H = \frac{1}{1 + 10^{-(100 + R_H - R_A)/400}}$$

$$R'_H = R_H + 20 * (S_H - P_H)$$

4.1.2. The Eloratings.net System

The second Elo rating system we studied is the one adopted by elratings.net system. There are two differences between this system and the classic one. The first one is that the K-factor changes according to the competition type. The second one is that there is one extra match importance multiplier G in the rating update formula.

The K-factor is adjusted in this way: for a friendly match, it takes value 20; for a minor tournament, it takes value 30; for World Cup and continental qualifiers, and major tournaments, it takes value 40; for continental championship finals and major intercontinental tournaments, it takes value 50; for FIFA World Cup finals, it takes value 60. The match importance multiplier G takes value 1 if the goal difference N is no greater than 1, 1.5 if the goal difference N is exactly 2, and $(N+11)/8$ if the goal difference N is no less than 3. And the rating update formula is:

$$R'_H = R_H + K * G * (S_H - P_H)$$

This Elo rating is updated on a daily basis, too.

4.1.3. The FIFA Men's World Ranking System

The third Elo-variant system we studied is the FIFA Men's World Ranking System. The rating points P being exchanged after a match is dependent on the outcome of the match (M), the importance of the game (I), the rating of the opponent (T), and the average of confederation strengths of participating teams (C). This is the formula:

$$P = M * I * T * C$$

We did not update the rating on a daily basis. Instead, we directly used the monthly-updated official release as the input when generating prediction of match outcome.

4.1.4. The FIFA Women's World Ranking System

FIFA has a different ranking procedure for the Women's Football, and this system is more closely related to the classic Elo system than the one for Men's Football. However, this system also differs from the classic one in 3 aspects.

Firstly, the K -factor is set to a different constant. K equals 15 for any match in this system.

Secondly, there is a different game importance multiplier I . It takes value 1 for friendlies and small competitions/tournaments, 2 for confederation-level qualifiers, 3 for FIFA Confederation Cup, FIFA World Cup qualifiers and Confederation-level finals, 4 for FIFA World Cup finals.

Lastly, the actual game result S_A does not always take values in $\{0, 0.5, 1\}$ in this system. Instead, the value of S_A depends on the goal difference. Following is the table designating value of S_A :

	0	1	2	3	4	5	6+
0	0.5	0.15	0.08	0.04	0.03	0.02	0.01
1	0.5	0.16	0.089	0.048	0.037	0.026	0.015
2	0.5	0.17	0.098	0.056	0.044	0.032	0.02
3	0.5	0.18	0.107	0.064	0.051	0.038	0.025
4	0.5	0.19	0.116	0.072	0.058	0.044	0.03
5+	0.5	0.20	0.125	0.080	0.065	0.05	0.035

The row names represent the goal difference, and the column names represent the goals scored by the losing team. This choice of S_A is meant to reward the losing team if they have made efforts and scored many goals. The expected score is computed using the same formula as the classic one. The following is the rating update formula for FIFA Women's Ranking System:

$$R'_A = R_A + K * I * (S_A - P_A)$$

4.2. Markovian System

The Markovian rating system utilizes the equilibrium distribution of a discrete Markov chain. Consider the transition probability P_{ij} , the probability that team i would win the match against team j . This probability can be estimated using the following formula, which involves the number of victories of team i over team j in the past:

$$\hat{p}_{ij} = \frac{V_j + 1}{V_i + V_j + 2}$$

Here, V_j denotes the number of victories of team j over team i , vice versa. The constant 1 in the nominator and 2 in the denominator is added in order to avoid zero division. If team i and team j have never met before, than the transition probability cannot be directly estimated. However, by computing all estimable transition probabilities, we can normalize the transition matrix P by row, and the inestimable probabilities are assigned the same value. Using the formula for the equilibrium distribution π :

$$\pi = \pi * P$$

Home advantage is not considered in this rating system.

4.3. Prediction and Errors

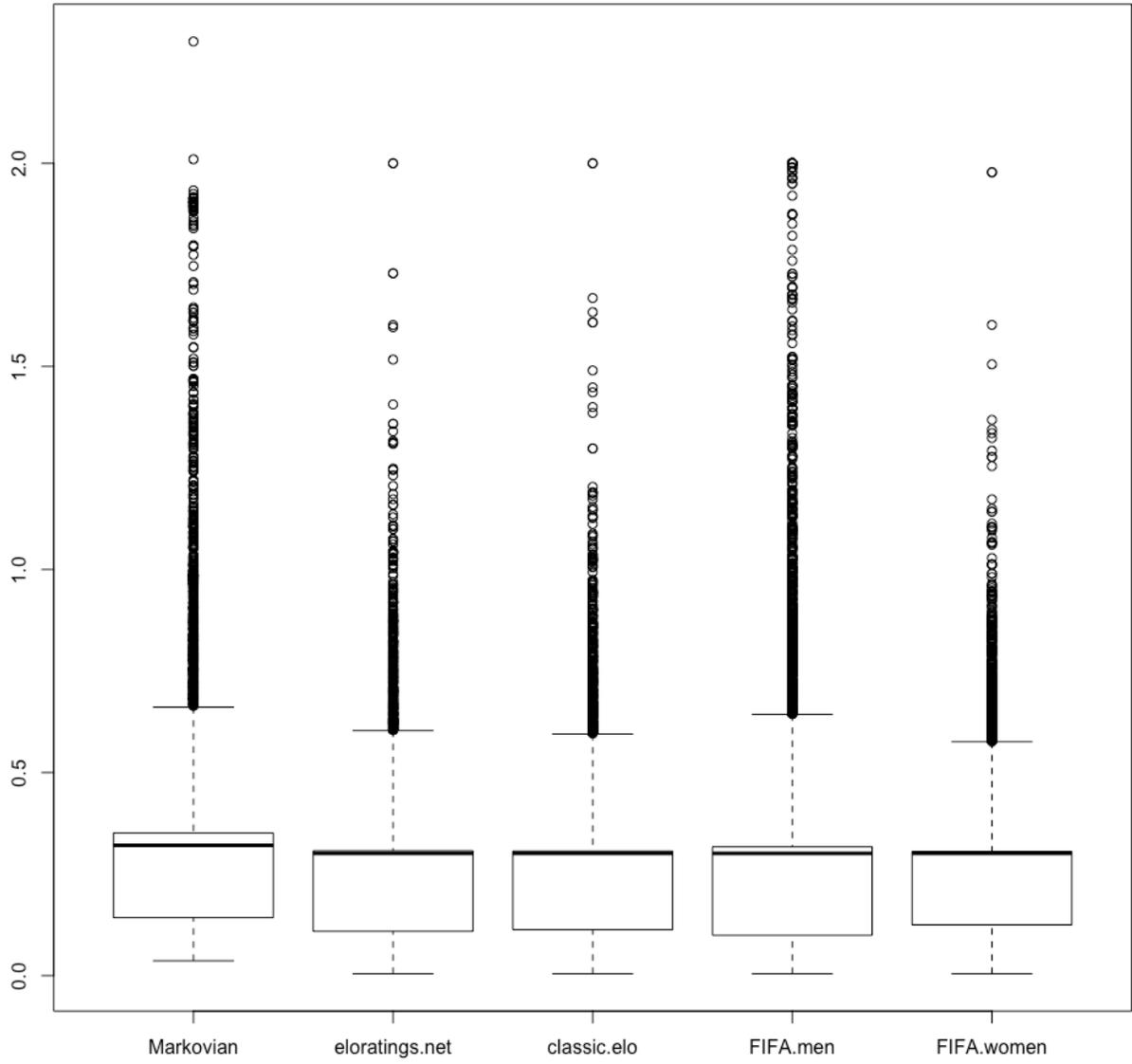
With the five rating systems introduced above, we generated the prediction of match result based on the daily-updated rating (except for the FIFA Men's World Rankin) and used a sliding-window approach to compute the prediction error, meaning that we use the rating updated on day t for the match outcome prediction on day $t+1$. We used both binomial deviance and squared errors as the error measure, and the results are below:

	Mean of Binomial Deviance	SD of Binomial Deviance	Mean of Squared Error	SD of Square Error
Classic Elo	0.2567951	0.1720152	0.1466897	0.1594267
Eloratings.net	0.2579407	0.1773721	0.1475498	0.1622564
FIFA Men	0.2831777	0.2532942	0.1605018	0.1904928
FIFA Women	0.2563387	0.1602452	0.146411	0.1536974
Markovian	0.3139173	0.2378511	0.1694812	0.1779247

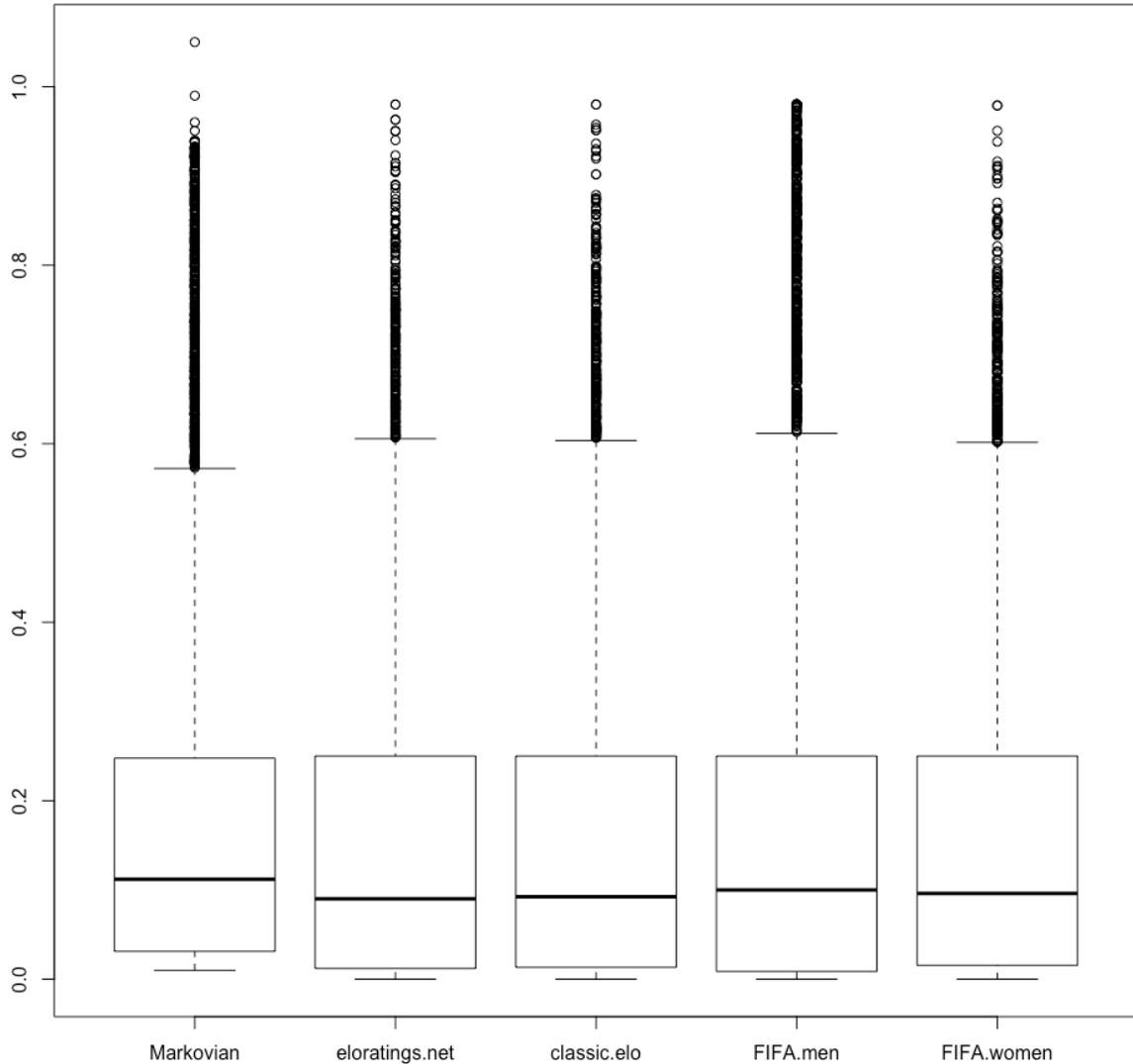
It is surprising that the classic Elo system outperforms the FIFA Men's World Ranking Procedure, which, supposedly, should be accurate and helpful in prediction, since it is designed for association football. The classic Elo system also achieves similar accuracy as the Eloratings.net system, though the latter is more complex. FIFA Women's World Ranking Procedure also outperforms the other system designed by FIFA. Besides, the Markovian system does relatively poorly, perhaps because a large number of transition probabilities are estimated from a relatively small sample. The sample size of a particular team against another particular team is usually small.

Below are boxplots for the binomial deviance and squared error measure comparison.

Five Ranking Procedures, Binomial Deviance



Five Ranking Procedures, Squared Error



5. Future Work

In this independent study, the Markovian system being studied does not take home advantage into account. However, there actually are works on Markovian rating system with home advantage for other sports, for example, NCAA College Basketball. J. Sokol proposed a Markovian rating system for NCAA that considers home advantage, however, some modification may be necessary for it to be applied to association football. One reason is that NCAA is a tournament, and most of the teams are guaranteed to have matches against all other teams, meaning that the transition probability is always estimable, though the small sample size may still be a problem. One possible modification is to compute the long-term transition probability using other estimated transition probabilities, i.e. estimate the m -step transition probability $P^n(x,y)$ and $p_{x,y}$, the probability that a Markov chain starting at x will be in state y at some positive time. However, sample size probably would still be a problem, and more work is needed.

References

- Clarke, Stephen R., and John M. Norman. "Home Ground Advantage of Individual Clubs in English Soccer." *The Statistician*44.4 (1995): 509. Web.
- Dixon, Mark J., and Stuart G. Coles. "Modelling Association Football Scores and Inefficiencies in the Football Betting Market." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*46.2 (1997): 265-80. Web.
- Goddard, John. "Regression models for forecasting goals and match results in association football." *International Journal of Forecasting*21.2 (2005): 331-40. Web.
- "World Football Elo Ratings: Rating System." *World Football Elo Ratings: Rating System*. N.p., n.d. Web. 16 Dec. 2016.
- Hvattum, Lars Magnus, and Halvard Arntzen. "Using ELO ratings for match result prediction in association football." *International Journal of Forecasting*26.3 (2010): 460-70. Web.
- "Live Scores, Sports News, Scores, Results, Fixtures, Odds, Database and more - 7M Sports." *Live Scores, Sports News, Scores, Results, Fixtures, Odds, Database and more - 7M Sports*. N.p., n.d. Web. 16 Dec. 2016.
- "National football teams of Europe, international matches history." *National football teams of Europe, international matches history*. N.p., n.d. Web. 16 Dec. 2016.
- @fifacom. "The FIFA/Coca-Cola World Ranking - FIFA.com." *FIFA.com*. N.p., n.d. Web. 16 Dec. 2016.
- @fifacom. "The FIFA/Coca-Cola World Ranking - Ranking Procedures - FIFA.com." *FIFA.com*. N.p., n.d. Web. 16 Dec. 2016.
- "The predictive power of ranking systems in association ..." N.p., n.d. Web. 16 Dec. 2016.
- "A logistic regression/Markov chain model for NCAA basketball." N.p., n.d. Web. 16 Dec. 2016.