doi.org/10.26398/IJAS.0030-007

RANKING AND FORECASTING SERIE A 1934-2016 SOCCER OUTCOMES THROUGH BIVARIATE POISSON REGRESSION

Roberto Benedetti,[1] Alessandro Pandimiglio

Department of Economics, Gabriele d'Annunzio University of Chieti-Pescara

Federica Piersimoni

Istat, Directorate for Methodology and Statistical Process Design, Rome, Italy

Marco Spallone

Department of Economics, Gabriele d'Annunzio University of Chieti-Pescara

Abstract. *Soccer rankings, based on previous performance of each team, have several important roles with a crucial impact on betting such as to provide an objective indication of the strength of each team (McHale and Davies, 2007). Typically tournaments are scheduled so to avoid the pairing of the best teams from the early stages. They also provide an efficient tool to predict outcomes in order to properly fix the betting odds through an objective criterion. The football betting market is based on fixed odds that generally remain unchanged in relation to bettor demand (Goddard, 2005; Goddard and Asimakopoulos, 2004). The efficiency of the estimates of the bookmakers could add a risk of exposure that may generate ample opportunities to uncover inefficiencies in the market. In the previous literature, the approach of modeling the goals scored and conceded by each team showed to be more flexible than directly modeling win-draw-lose match results (Dobson and Goddard, 2003; Goddard, 2005; Lasek et al., 2013). Bivariate Poisson regression (Dixon and Coles, 1997) is used to estimate ranking models on the Italian Serie A historical data from 1934 to 2016. Promising and encouraging forecasting performance is achieved tuning appropriately the reference period of the data used to estimate the model. Such a model is flexible to the introduction of additional team specific covariates that can improve its predictive capabilities.*

Keywords: *Betting odds, Poisson distribution, likelihood function, predictive capabilities, time varying parameters.*

1. INTRODUCTION

Sports betting has always existed in Italy, but since 2000, it experienced a strong growth due to market liberalization. However, it is only in the last 15 years that sports betting has become relevant in terms of business volumes: in fact, market

[1] Corresponding author: Roberto Benedetti, email: benedett@unich.it

turnover increased from €1.75 billions in 2004 to €5.5 billions in 2015 (Figure 1), 90% of which are soccer bets; moreover, from 2012 on, the payout rate hasalways been above 80%, due to an increase of market competition. In 2015 winnings were 86% of turnover.

The contribution to tax revenues of this industry is significant: from 2012 to 2015, tax revenues have been on average €178 millions (Figure 2). Until 2015, due taxation was about 4% of turnover; from 2016 on, tax base became gross margin. The new tax rules should increase market competition: in fact, bookmakers may find convenient to raise the payout ratio (that is, marginal costs) in order to reduce tax incidence. Therefore, since consumers face lower prices (that is, amount at stake less expected payout), they bet more, hence increasing both turnover and fiscal revenues.

However, market efficiency not only depends on the tax system, but also on the accuracy of the predictions about the outcomes of a sport event.

In competitive markets, efficiency requires that the odds issued by bookmakers are as closest as possible to the best predicted odds: in fact, market efficiency requires that there exists no strategy that yields positive expected returns to betters. In non-competitive markets, instead, accuracy of predictions about outcomes reduces the volatility of expected profits of book-makers and allows for positive returns on market side.

There is an extensive academic literature on modeling the probability of outcomes of sports events, and on testing market efficiency. Nevertheless, our research is restricted to the recent literature on football betting market: in particular, our contribution is aimed at predicting probabilities about outcomes of Italian Serie A football matches.

As for probability modeling, the study of Moroney (1956) was the first that suggested to make use of a negative binomial distribution and, as a second choice, of the Poisson distribution to predict the various outcomes of football matches. Maher (1982) obtained reasonably accurate estimation of football scores and derived more sophisticated predictions, by making use of a bivariate Poisson model in which the scores of the home and the away teams are assumed to be independent Poisson distributions, based on the previous performances of each team.

As for testing market efficiency, bivariate Poisson models seem the most commonly used technique: for example, Dixon and Coles (1997); Dixon and Robinson (1998); Goddard (2005); Goddard and Asimakopoulos (2004) for the English Leagues; Dyte and Clarke (2000) for the 1998 Fifa World Cup.

Others made use of more sophisticated techniques: Rue and Salvesen (2000) predicted English Premier League results with a Bayesian dynamic generalized linear model using Markov chain Monte Carlo iterative simulation technique;

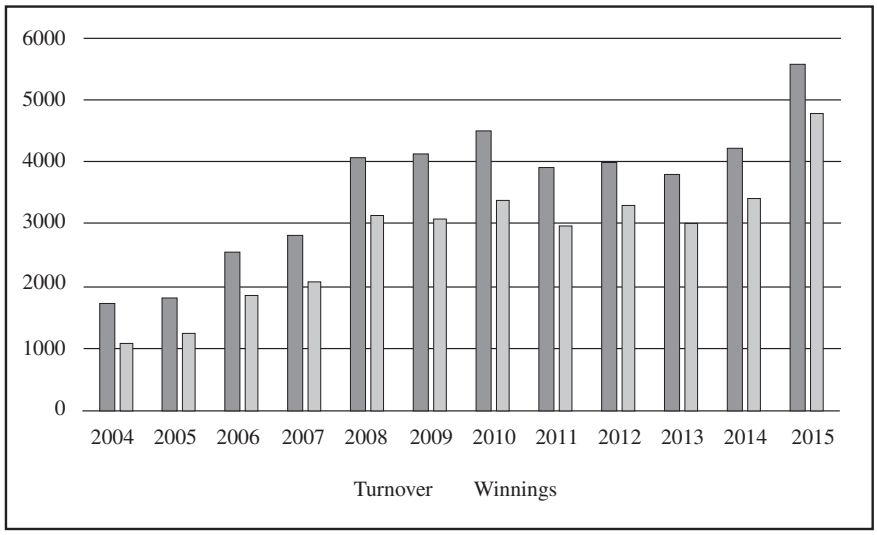

Fig. 1: Italian sport betting market: turnover and winnings (mil euro)

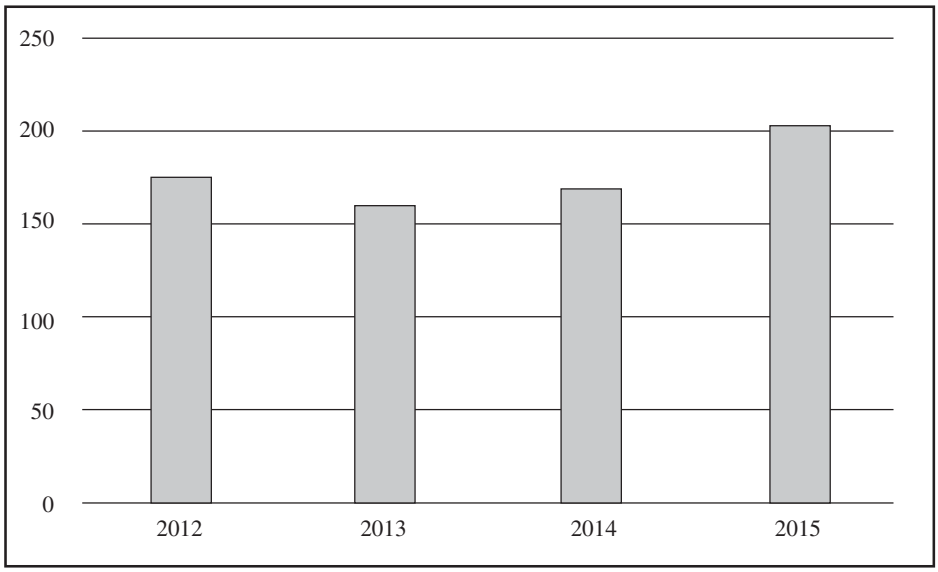

Fig. 2: Italian sport betting market: fiscal revenue (mil euro)

Forrest and Simmons (2008) estimated odds, by using a clustered probit regression for Spain; Demir et al. (2012) implemented a Fibonacci strategy to predict odds for 8 European soccer leagues.

To our knowledge, there are no estimates of soccer fixed odds for Italian Serie A league. Following Following Dixon and Coles (1997) and Dixon and Robinson (1998), this paper applies a bivariate Poisson regression in order to predict probabilities of outcomes of Italian Serie A football league, by using historical data from 1934 to 2016.

Usually, football betting is based on fixed odds: book-makers issue odds on out comes, such as full time or half time results, exact scores, number of goals (under/over), etc. The aim of this paper is to predict the probabilities of these outcomes without considering market efficiency issues. In other words, the paper focuses on predictive techniques and does not deal with the implementation of optimal betting strategies.

In the last few years book-makers also issue live betting odds on many events, such as the first team to score, next player to score, etc. We believe that the predictive model suggested by this paper, together with an adequate database, could provide a better understanding of live betting.

Four sections follow this introduction. Section 2 describes available data. Section 3 briefly sketches the model (built on the structure of Dixon and Coles (1997) and Dixon and Robinson (1998)). Section 4 shows the main results. Finally section 5 concludes and suggests refinements and further improvements.

2. DATA DESCRIPTION

The data comprise the final results of matches played in Italian Serie A from 1934 to 2016. The results are collected from the site:
https://github.com/jalapic/engsoccerdata
which provides only the final score for all the matched played in the period 1934-2016. Each game has three possible outcomes: home-side win, draw or away-side win. Since the database provides the final scores of each match, it is possible to get information about the attack and defense ability of every team, distinguishing between home and away matches; moreover, it is possible to get the relative frequency of goals in home and away games.

Tables 1-3 show the relative frequency of home and away goals in the whole sample (Table 1), in the last five years (Table 2) and in the last season (Table 3). For the whole sample and for the other two sub-samples the most likely outcome is 1-1, while the second most likely outcome is a 1-0 home win. Moreover, it can be

noted that in the last few years the number of goals scored by the away teams is increasing: in fact, in 19.8% of matches away teams scored 2 goals and in 8.4% of matches they scored 3 goals.

Tab. 1: Relative frequency of games in 1934-2016 by home goals (rows) and away goals (columns)

	0	1	2	3	4	5	6	7	8	Tot
0	11.297	6.109	3.401	1.169	0.398	0.102	0.031	0.016	0.008	22.532
1	12.443	13.175	4.468	1.952	0.543	0.201	0.024	0.016	0.000	32.822
2	8.089	8.955	4.940	1.193	0.417	0.102	0.051	0.004	0.004	23.756
3	4.287	4.629	2.181	0.807	0.173	0.071	0.020	0.008	0.000	12.175
4	1.941	2.043	0.996	0.390	0.087	0.028	0.000	0.004	0.000	5.487
5	0.744	0.807	0.362	0.142	0.031	0.004	0.000	0.000	0.000	2.090
6	0.228	0.268	0.150	0.075	0.016	0.008	0.000	0.000	0.000	0.744
7	0.083	0.106	0.047	0.016	0.000	0.000	0.000	0.000	0.000	0.252
8	0.047	0.016	0.016	0.000	0.000	0.000	0.000	0.000	0.000	0.079
9	0.016	0.031	0.008	0.004	0.000	0.000	0.000	0.000	0.000	0.059
10	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004
Tot	39.179	36.140	16.568	5.747	1.665	0.516	0.126	0.047	0.012	100.000

Tab. 2: Relative frequency of games in 2011-2016 by home goals (rows) and away goals (columns)

	0	1	2	3	4	5	6	7	Tot
0	8.526	7.684	5.105	1.895	0.632	0.053	0.000	0.053	23.947
1	9.421	11.579	5.526	3.053	0.842	0.211	0.053	0.000	30.684
2	9.316	8.895	5.368	1.421	0.842	0.158	0.000	0.000	26.000
3	4.211	4.158	2.368	1.368	0.211	0.105	0.000	0.000	12.421
4	1.474	1.947	1.158	0.474	0.105	0.053	0.000	0.000	5.211
5	0.421	0.526	0.211	0.105	0.053	0.000	0.000	0.000	1.316
6	0.158	0.053	0.053	0.053	0.000	0.000	0.000	0.000	0.316
7	0.105	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.105
Tot	33.632	34.842	19.789	8.368	2.684	0.579	0.053	0.053	100.000

3. THE MODEL

This paper uses a statistical prediction model based on Dixon and Coles (1997) and Dixon and Robinson (1998) methodology that allows to take into account sev- eral factors which can influence the final score of the match such as the ranking, the ability to attack, the ability to defend of both teams, the home effect, the last n performances of both teams, and so on.

The basic idea is that the number of goals scored by the home and the away teams are independent Poisson variables, whose means are the ability to attack and de- fend of both teams. In particular, teams are indexed i and j and Xi, j and Yi, j are the number of goals scored by the home and away team respectively. The model can predict full time results and is represented as follows:

$$X_{i,j} \sim Poisson(\alpha_i \beta_j \gamma_h), \tag{1}$$

$$Y_{i,j} \sim Poisson(\alpha_j \beta_i), \tag{2}$$

where $X_{i,j}$ and $Y_{i,j}$ are independent and α_i, $\beta_i > 0 \; \forall i$. The attack ability is measured by α_i, while β_i and $\gamma_h > 0$ measure the defence ability and the home effect respectively.

From equations (1) and (2) follows that with n teams, $(\alpha_1, ..., \alpha_n)$ attack parameters, $(\beta_1, ..., \beta_n)$ defence parameters and the home effect parameter γ_h, have to be estimated. Since the model will have too many parameters, the following constraint must be imposed:

$$n^{-1} \sum_{i=1}^{n} \alpha_i = 1, \tag{3}$$

The likelihood function is the basic tool for inference. With matches indexed $k = 1, ..., N$ and corresponding scores (x_k, y_k) the likelihood function takes the form:

$$L(\alpha_i, \beta_i, \gamma_i; i = 1, ..., n) = \prod_{k=1}^{N} exp(-\lambda_k) \lambda_k^{Xk} exp(-\mu_k) \mu_k^{Yk}, \tag{4}$$

where:

$$\lambda_k = \alpha_{i(k)} \beta_{j(k)} \gamma_h,$$
$$\mu_k = \alpha_{j(k)} \beta_{j(k)},$$

while $i(k)$ and $j(k)$ measure the indices of home and away team paling in match k.

In the next section the main results of attack and defense scores of the majorItalian soccer teams will be presented.

4. RESULTS

The Poisson regressions are obtained using the following package:
https://CRAN.R-project.org/package=fbRanks

If the database contains unconnected clusters (that is, if it contains teams that have never played each other or haven't played each other very often, due to promotion and relegation), each cluster is ranked separately relative to the median team strength in the cluster. The package contains functions for predicting and simulating tournaments and leagues from estimated models and allows efficient fitting of a very large numbers of teams. The fitting algorithm analyzes match data and determines which teams form a cluster, hence fitting each cluster separately.

The complete set of parameters is obtained by maximizing equation (4). In this way it is possible to obtain attack and defense abilities for each team, that will be employed to estimate the probabilities of match outcomes.

Although the predictions on the final result of a match should be based on a short sample since strength or weakness of teams can change significantly over time, we provide estimates for the whole sample (1934-2016), for the last five years (2011-2016) and for the last available season 2015-2016. So only the last two estimates give an idea of the real strength or weakness of the Italian Serie A teams. Figure 3 shows the mean attack and defense parameters for Italian serie A teams in the whole sample 1934-2016, in the sub-sample 2011-2016 and in the last season 2015-2016. As expected, the teams' performance rates, as determined by α_i and β_i, are changing over time: this means that attack and defense scores tend to be dynamic, varying from one season to another. Therefore also the probabilities of match outcomes could vary a lot according to the chosen sample. This evidence is even clearer, if the total score (attack and defense) of the top four Italian teams in the sample 1934-2016, namely Juventus, Inter, Milan and Roma, are compared: they are described in Figure 4, where the total score is measured on the basis of both five-year data and one-year data.

As expected, the total score is more volatile when measured on an annual basis: this suggests that the time interval matters in making correct predictions about the outcome of matches.

Tables 4 and 5 confirm this evidence. Table 4 measures the attack, defense, and total score of the 35 teams with higher ranks in the three different samples (1934-2016, 2011-2016 and 2015-2016). Even though Juventus is the first Italian team in all samples, rankings of other teams changes a lot over time; therefore, also the probability of the final outcomes of matches varies over time. Table 5 provides the predicted probabilities for home wins, ties and away wins, by using different season data range.

To sum up, we believe that accurate predictions of match outcomes should be based on a short time basis (i.e., the performances of each team in the last two or three years); moreover, it is crucial to take into account some other information, like teams' league positions, unavailable players, and many external factors that are not available in our database.

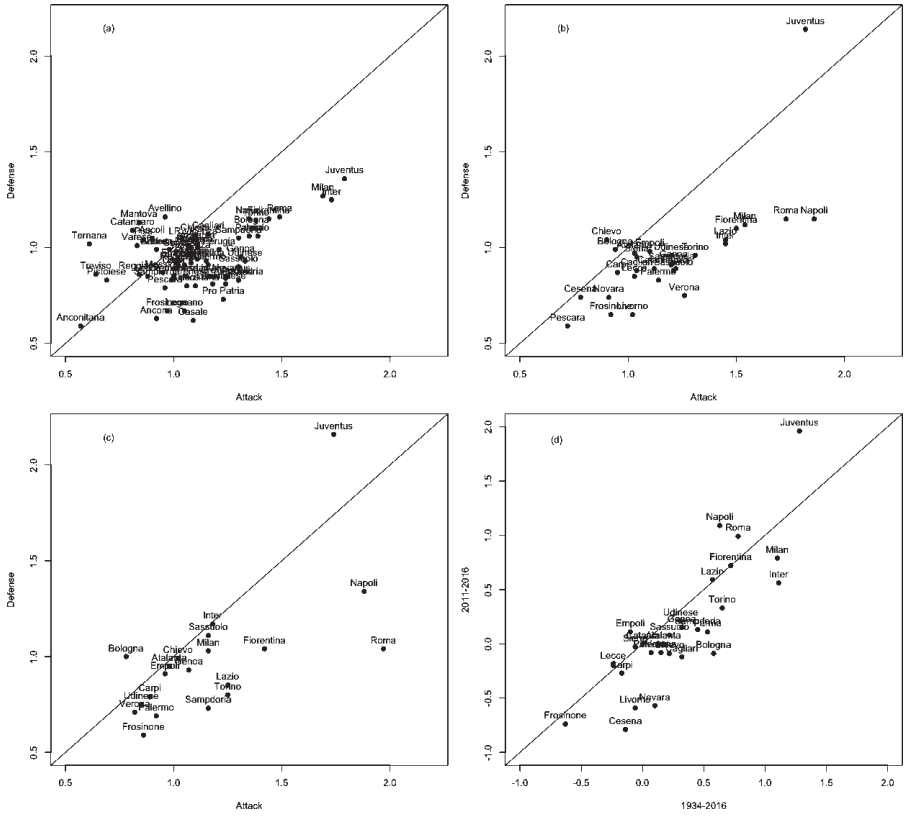

Fig. 3: (a) Serie A 1934-2016: teams attack and defence scores, (b) Serie A 2011-2016: teams attack and defence scores, (c) Serie A 2015-2016: teams attack and defence scores, (d) Total Scores: 2011-2016 vs 1934-2016

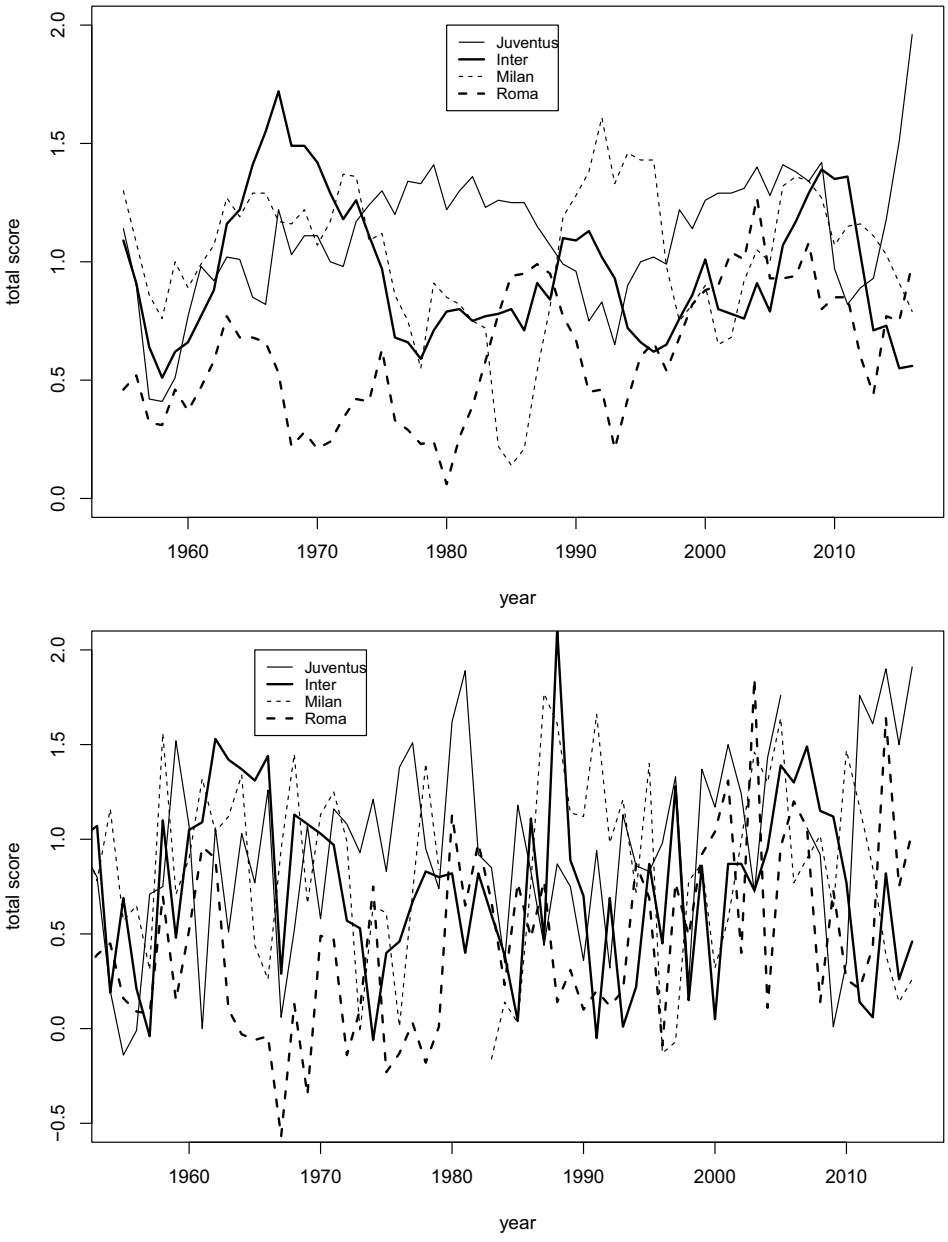

Fig. 4: Juventus, Inter, Milan and Roma Total Scores (top) 5 year data, (bottom) 1 year data

Tab. 4: Scores by using different season data range. Only the 35 teams with higher rank are shown

		Data 1934-2016				Data 2011-2016				Data 2015-2016			
Rank	**Team Name**	**score**	**att.**	**def.**	**games**	**score**	**att.**	**def.**	**games**	**score**	**att.**	**def.**	**games**
1	Juventus	1.28	1.79	1.36	2808	1.96	1.82	2.14	190	1.91	1.74	2.16	38
2	Inter	1.11	1.73	1.25	2847	0.56	1.45	1.02	190	0.46	1.18	1.17	38
3	Milan	1.10	1.69	1.27	2786	0.79	1.54	1.12	190	0.26	1.16	1.03	38
4	Roma	0.78	1.49	1.16	2802	0.99	1.73	1.15	190	1.04	1.97	1.04	38
5	Fiorentina	0.72	1.44	1.15	2626	0.72	1.50	1.10	190	0.57	1.42	1.04	38
6	Torino	0.65	1.38	1.14	2419	0.33	1.31	0.96	152	0.01	1.25	0.80	38
7	Napoli	0.63	1.35	1.15	2348	1.09	1.86	1.15	190	1.33	1.88	1.34	38
8	Bologna	0.58	1.36	1.10	2330	-0.09	0.94	0.99	152	-0.35	0.78	1.00	38
9	Lazio	0.57	1.39	1.06	2489	0.59	1.45	1.04	190	0.08	1.25	0.85	38
10	Parma	0.53	1.35	1.06	856	0.11	1.22	0.89	152	-	-	-	-
11	Sampdoria	0.45	1.30	1.05	2008	0.13	1.20	0.91	152	-0.24	1.16	0.73	38
12	Cagliari	0.32	1.16	1.07	1216	-0.12	1.04	0.88	152	-	-	-	-
13	Genoa	0.32	1.31	0.95	1690	0.15	1.21	0.92	190	-0.01	1.07	0.93	38
14	Udinese	0.31	1.33	0.93	1486	0.21	1.20	0.96	190	-0.66	0.85	0.75	38
15	Perugia	0.26	1.21	0.99	418	-	-	-	-	-	-	-	-
16	Chievo	0.22	1.10	1.06	520	-0.09	0.90	1.04	190	0.02	1.02	0.99	38
17	Sassuolo	0.22	1.30	0.90	114	0.08	1.21	0.88	114	0.36	1.16	1.11	38
18	LR Vicenza	0.18	1.09	1.04	1008	-	-	-	-	-	-	-	-
19	Atalanta	0.17	1.11	1.02	1906	0.00	1.03	0.97	190	-0.11	0.98	0.95	38
20	Avellino	0.16	0.96	1.16	300	-	-	-	-	-	-	-	-
21	Verona	0.15	1.11	1.00	866	-0.08	1.26	0.75	114	-0.76	0.82	0.71	38
22	Padova	0.12	1.30	0.83	560	-	-	-	-	-	-	-	-
23	SPAL	0.12	1.07	1.01	544	-	-	-	-	-	-	-	-
24	Alessandria	0.11	1.30	0.83	437	-	-	-	-	-	-	-	-
25	Foggia	0.10	1.08	1.00	358	-	-	-	-	-	-	-	-
26	Novara	0.10	1.25	0.85	446	-0.57	0.91	0.74	38	-	-	-	-
27	Triestina	0.10	1.15	0.93	896	-	-	-	-	-	-	-	-
28	Palermo	0.07	1.16	0.91	1012	-0.08	1.14	0.83	152	-0.65	0.92	0.69	38
29	Piacenza	0.07	1.08	0.97	272	-	-	-	-	-	-	-	-
30	Modena	0.05	1.11	0.94	466	-	-	-	-	-	-	-	-
31	Pro Vercelli	0.05	1.24	0.84	200	-	-	-	-	-	-	-	-
32	Brescia	0.01	1.08	0.94	762	-	-	-	-	-	-	-	-
33	Lucchese	0.01	1.24	0.81	278	-	-	-	-	-	-	-	-
34	Catania	0.00	1.08	0.92	602	0.00	1.12	0.89	114	-	-	-	-
35	Como	-0.01	1.01	0.98	426	-	-	-	-	-	-	-	-

Tab. 5: Predicted probabilities for Home Wins - Tie - Away Wins by using different season data range

Data 1934-2016

	Juventus	Inter	Milan	Roma	Fiorentina	Torino
Juventus	———	41.01-25.71-33.28	41.02-25.84-33.14	47.83-25.47-26.71	48.89-25.34-25.78	50.07-25.33-24.61
Inter	33.21-25.57-41.22	———	37.51-25.82-36.67	44.01-25.61-30.38	45.41-25.55-29.04	46.78-25.62-27.60
Milan	33.45-25.89-40.66	36.99-25.63-37.38	———	43.33-26.10-30.57	44.96-25.64-29.40	46.62-25.99-27.40
Roma	26.89-25.52-47.59	30.21-25.50-44.30	30.39-26.01-43.60	———	37.90-26.72-35.38	39.34-26.95-33.71
Fiorentina	25.97-25.25-48.78	29.17-25.80-45.03	29.31-26.01-44.69	35.37-26.64-37.99	———	37.52-27.37-35.11
Torino	24.62-25.36-50.02	27.61-25.64-46.75	27.98-25.67-46.35	33.85-27.01-39.14	34.55-27.28-38.16	———

Data 2011-2016

	Juventus	Inter	Milan	Roma	Fiorentina	Torino
Juventus	———	64.10-22.34-13.57	58.81-24.76-16.43	55.53-25.44-19.03	60.14-24.09-15.78	68.03-20.88-11.09
Inter	13.67-22.53-63.80	———	32.42-25.18-42.40	28.27-24.15-47.58	33.71-25.42-40.87	42.48-25.09-32.44
Milan	16.45-24.83-58.73	42.55-25.20-32.25	———	33.16-24.86-41.98	39.08-25.22-35.70	47.73-24.77-27.50
Roma	19.27-25.09-55.64	47.77-23.80-28.43	42.35-24.81-32.84	———	43.49-24.75-31.76	52.84-23.30-23.86
Fiorentina	15.45-24.23-60.32	40.96-25.20-33.84	35.57-25.71-38.72	31.74-24.79-43.47	———	45.73-25.13-29.14
Torino	11.08-20.79-68.13	32.53-25.20-42.27	27.75-24.86-47.39	24.09-22.99-52.92	29.01-24.90-46.08	———

Data 2015-2016

	Juventus	Inter	Milan	Roma	Fiorentina	Torino
Juventus	———	60.37-26.20-13.43	65.46-23.14-11.40	55.13-24.14-20.73	61.66-23.80-14.54	73.98-17.36- 8.66
Inter	13.23-25.92-60.85	———	38.90-29.64-31.46	25.48-24.03-50.48	33.80-28.02-38.18	46.11-26.09-27.80
Milan	11.43-23.04-65.53	31.63-29.36-39.00	———	21.95-22.21-55.84	30.38-26.66-42.96	42.40-25.76-31.84
Roma	20.30-24.50-55.21	50.38-23.95-25.68	56.19-22.01-21.80	———	49.78-22.30-27.91	64.93-17.96-17.10
Fiorentina	14.43-23.75-61.83	38.24-28.01-33.75	43.46-26.54-30.00	27.98-22.42-49.59	———	50.69-23.39-25.92
Torino	8.52-17.22-74.26	27.73-26.16-46.11	31.50-26.10-42.40	17.09-17.99-64.93	25.67-23.46-50.87	———

5. CONCLUDING REMARKS

In this study we estimated a ranking model on the Italian Serie A with historical data from 1934 to 2016, by using a bivariate Poisson regression. Studies on foot- ball betting markets are relatively scarce and, at least to our knowledge, this is the first attempt to estimate the Italian soccer betting odds, by using such a long historical dataset. Following Dixon and Coles (1997) and Dixon and Robinson (1998), we made use of a bivariate Poisson distribution to estimate the number of goals scored by each team and the probability of the final outcome of Italian Se- rie A football matches. Basically, we estimated the attack and defense scores, by exploiting the information about past performances of each team. Although esti- mations are accurate in many respects, we noted that teams' performance rates are changing over time; this implies that attack and defense scores are dynamic, vary- ing from one season to another. So, the probabilities of match outcomes could vary a lot according to the chosen sample. Generally speaking, teams' perfor- mance rates are more volatile when measured on shorter time intervals.

Our model is based on information about teams' past performances (the final score of Italian Serie A matches); therefore, our estimations are based on the history of each team. Our recommendations for future research are that further refinements are needed to improve the accuracy of the estimations: in particular, big data availability could provide more information about the teams, such as unavailable players, timing of the goals, goal scorers, newly signed players, weather conditions, and many other external factors that are are not currently at our disposal.

REFERENCES

Demir, E., Hakan, D. and Rigoni, U. (2012). Is the soccer bet market efficient? a cross country investigation using the fibonacci strategy. In *The Journal of Gambling Business and Economics*. 6(2): 29-49.

Dixon, M.J. and Coles, S.C. (1997). Modelling association football scores and inefficiencies in the football betting market. In *Journal of the Royal Statistical Society: Series C (Applied Statistics).* 46: 265-280.

Dixon, M.J. and Robinson, M.E. (1998). A birth process model for association football matches. In *Journal of the Royal Statistical Society: Series D (The Statistician).* 47(3): 523-538.

Dobson, S. and Goddard, J. (2003). Persistence in sequences of football match results: A montecarlo analysis. In *Journal of Operational Research.* 2: 247-256.

Dyte, D. and Clarke, S.R. (2000). A ratings based poisson model for world cup soccer simulation. In *Journal of the Operational Research Society.* 51(8): 993-998.

Forrest, D. and Simmons, R. (2008). Sentiment in the betting market on spanish football. In *Applied Economics.* 40: 119-126.

Goddard, J. (2005). Regression models for forecasting goals and match results in association football. In *International Journal of Forecasting.* 21: 331-340.

Goddard, J. and Asimakopoulos, I. (2004). Forecasting football match results and the efficiency of fixed-odds betting. In *Journal of Forecasting.* 23: 51-66.

Lasek, J., Szlàvik, Z. and Bhulai, S. (2013). The predictive power of ranking systems in association football. In *International Journal of Applied Pattern Recognition.* 1(1): 27-46.

Maher, M.J. (1982). Modelling association football scores. In *Statistica Neerlandica.* 43: 309-315.

McHale, I. and Davies, S. (2007). Statistical analysis of effectiveness of the FIFA world rankings. In J. Albert and R.H. Koning, eds., *Statistical Thinking in Sports.* Chapman and Hall/CRC, Boca Raton: 77-90.

Moroney, M.J. (1956). *Fact from figures.* Penguin, London, 3rd edn.

Rue, H. and Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. In *Journal of the Royal Statistical Society: Series D (The Statistician).* 49: 399-418.