PARX model for football matches predictions*

Giovanni Angelini and Luca De Angelis

Department of Statistical Sciences & School of Economics, Management and Statistics

University of Bologna

March 2, 2017

Abstract

We propose an innovative approach to model and predict the outcome of football matches based on the Poisson AutoRegression with eXogenous covariates (PARX) model recently proposed by Agosto, Cavaliere, Kristensen and Rahbek (2016). We show that this methodology is particularly suited to model the goals distribution of a football team and provides a good forecast performance that can be exploited to develop a profitable betting strategy. This paper improves the strand of literature on Poisson-based models, by proposing a specification able to capture the main characteristics of goals distribution. The betting strategy is based on the idea that the odds proposed by the market do not reflect the true probability of the match because they may also incorporate the betting volumes or strategic price settings in order to exploit bettors' biases. The out-of-sample performance of the PARX model is better than the reference approach by Dixon and Coles (1997). We also evaluate our approach in a simple betting strategy which is applied to English football Premier League data for the 2013/2014, 2014/2015, and 2015/2016 seasons. The results show that the return from the betting strategy is larger than 30% in most of the cases considered and may even exceed 100% if we consider an alternative strategy based on a predetermined threshold which makes it possible to exploit the inefficiency of the betting market.

Keywords: Sports forecasting, Density forecasts, Count data, Poisson autoregression, Betting market.

J.E.L. Classification: C22; C25; C51; C53; L83.

*The authors are grateful to the Editor, Terence Mills, and two anonymous referees for their helpful comments on the earlier version of this article. Correspondence to: Luca De Angelis, Department of Statistical Sciences, University of Bologna, via Belle Arti 41, 40126 Bologna, Italy. E-mail: l.deangelis@unibo.it

1 Introduction

Over the last few years, the football betting market has experienced the fastest growth in gambling markets (Constantinou et al., 2013). Not surprisingly, many different methodologies have been developed to construct a profitable betting strategy which is able to capture the mispricing of odds. Starting with the pioneering works of Maher (1982) and Dixon and Coles (1997), many econometric methods have been proposed to predict football match results.

The purpose of our paper is twofold: (i) to develop an approach able to compute a set of probabilities associated with each possible result, and (ii) to use these probabilities to profit from the potential mispricing of the odds offered on the betting market. The odds proposed by the bookmakers may be influenced by betting volumes and, therefore might not always reflect the true probability of the match outcomes. Indeed, one of the aims of bookmakers is to encourage bettors to subdivide their wagers on each odd (Vlastakis et al., 2009). In doing so, they minimise the risk and gain from the unfairness of the proposed odds. Moreover, bookmakers may systematically set odds in order to take advantage of bettors' biases, such as the well-known preference for favourites and local teams, in order to increase profits (Levitt, 2004). Therefore, the comparison between true probabilities and odds can be exploited to define a profitable betting strategy.

The mainstream econometric approach to predict the probabilities associated with the outcomes of a football match is to model the number of goals scored and conceded by the two teams based on Poisson distributions. These probabilities are then aggregated to obtain the expectations of different match results (Maher, 1982; Dixon and Coles, 1997; Lee, 1997; Karlis and Ntzoufras, 2003). Another approach is to model these probabilities directly using discrete choice regression models such as ordered probit regression models (Kuypers, 2000; Goddard and Asimakopoulos, 2004; Forrest et al., 2005; Goddard, 2005). An interesting overview of different forecasting methods is proposed by Spann and Skiera (2009). Our proposal lines up with the first approach. In particular, our purpose is to compute the home and away team goals probability distributions based on Poisson models. This approach is more flexible than the one based on discrete choice regression because, once the distribution associated with the goals scored by the two team is computed, it is possible to derive the probability for each possible match result. Dixon and Coles (1997) propose a model based on the product of two univariate Poisson distributions which generates probabilities for home and away team goals under the assumption of independence between the goal distributions of two opposite teams. With respect to the original work of Maher (1982), Dixon and Coles (1997) also take into account an additional parameter which governs the dependence between home and away goals distributions for the results 0-0, 0-1, 1-0 and 1-1, which are found to be statistically dependent in their sample. Conversely, Karlis and Ntzoufras (2003) use a bivariate Poisson distribution arguing that the dependence parameter, albeit small, leads to a more accurate prediction of the number of draws. Koopman and Lit (2015) extend this idea by proposing a dynamic bivariate Poisson distribution

to jointly model the distribution of home and away team goals, thus allowing for a framework where the attack and defence strengths of teams can slowly vary stochastically over time.

A further strand of the literature on football matches prediction focuses on machine learning-based methods. Among others, Tsakonas et al. (2002) show the better forecasting performance of a genetic programming-based technique compared to methods based on fuzzy models and neural network. Rotshtein et al. (2005) propose a combination of a genetic and neural optimisation techniques. More recently, various authors underscore the importance of expert judgements in the evaluation of football matches and their contribution to Bayesian networks (see, e.g., Joseph et al., 2006). Within this class of methods, Baio and Blangiardo (2010) propose a hierarchical Bayesian model without subjective judgements while Constantinou et al. (2012) consider a Bayesian network where subjective variables assume a central role in the prediction of football matches.

Our paper focuses on football matches prediction using the Poisson Autoregression with eXogenous covariates (PARX) model introduced by Agosto et al. (2016) which extends the Poisson Autoregression (PAR) model originally proposed by Fokianos et al. (2009) to include covariates in its specification. This model has been successfully used to predict corporate defaults and, within the framework of football betting, it is particularly useful, since the intensity of the goals scored by a team is characterized by an autoregressive persistence that the PARX model is able to account for. Therefore, these probabilities can be compared with those implied by odds in order to detect potential mispricings in the betting market. Our approach is slightly different from those proposed in the literature for football match predictions, because the PARX model captures the dynamics of the conditional intensity of goal distribution. More specifically, the adopted specification is able to model the autoregressive persistence in the propensity of scoring goals by a football team. With respect to other Poisson-based methods proposed in the literature, the dynamic specification of the PARX model allows to capture the form of a team and its recent performance. To the best of our knowledge, the only paper which takes into account a dynamic specification for the intensity has been proposed by Koopman and Lit (2015). The main difference of our method as compared to Koopman and Lit (2015) is that we exploit additional information by including exogenous covariates in the model specification which can greatly improve forecasting performances. The inclusion of covariates such as proxies for the attack or defence abilities of the opposing teams can be particularly useful for predicting football match results, as the information on the strength and/or form of the teams is taken into account.

The one-step ahead forecast accuracy of the PARX-based approach is compared to that of Dixon and Coles (1997), which is one of the main references in this context, as well as to that obtained on the basis of a pure PAR model. According to the mean-squared forecasting error, our proposal outperforms the one by Dixon and Coles (1997) in predicting the number of goals of the away team and provides an overall better forecasting performance than the PAR-based approach.

We finally propose a suitable betting strategy based on a set of different bookmaker odds to further evaluate the out-of-sample forecasting performance of our model. Our betting strategy is based on the comparison between the probabilities computed by the PARX model and the corresponding odds proposed in the betting market. The results obtained when applying our PARX-based betting strategy to the 2013/2014, 2014/2015, and 2015/2016 seasons of the English Premier League show that our approach is profitable and is able to detect the mispricing of the betting market by spotting the most underpriced odds, i.e., payouts higher than expected. We also compare our betting strategy performances with those of three 'naive' strategies: (i) always bet on the home team, (ii) always bet on the favourite team, and (iii) the 'longshot' approach as described in Forrest and Simmons (2002).

The rest of the paper is organized as follows. The estimation of the PARX model and its model selection are outlined in section 2. In section 3 we discuss how to predict the number of goals using the PARX model and the forecasting evaluation, and we propose a simple and profitable betting strategy. In section 4 we apply the PARX model to the English Premier League data. Section 5 concludes the paper.

2 Modelling football goals with PARX

In this section we propose an innovative approach to derive the probabilities associated with each possible outcome of a football match by taking into account the main features of the goal distributions. In particular, PAR and PARX denote a class of models which are characterized by a linear autoregressive intensity and allow to fit data that show serial dependence, a typical characteristic of football goals distributions.

These models also capture the phenomenon of goal clustering which, analogously to the well-known volatility clustering in financial literature, identifies periods during which football teams tend to score more goals than during other periods. A further advantage of PAR and PARX models is that they account for overdispersion, a feature observed in numerous count data, including goals scored by a football team. The difference between PAR and PARX models is that the latter makes it possible to include exogenous covariates in the model specification. This model extension is particularly suitable in our framework as it enables us to incorporate additional information about the team's strength, ability, and/or form, with the aim of improving the forecast accuracy.

Let y_t denote the number of goals scored by a football team at time t, where $t = 1, ..., T$. The PARX model of intensity λ_t can be specified as

$$y_t|\mathcal{F}_{t-1} \sim Pois(\lambda_t), \quad t = 1, ..., T \tag{1}$$

$$\lambda_t = \omega + \sum_{j=1}^{p} \alpha_j \lambda_{t-j} + \sum_{j=1}^{q} \beta_j y_{t-j} + \gamma \mathbf{x}_{t-1} \tag{2}$$

where $\mathbf{x}_{t-1}$ denotes a vector of m exogenous (non-negative) covariates and $\mathcal{F}_{t-1}$ is the information set available at time $t-1$, i.e., $\mathcal{F}_{t-1} = \{y_{t-m},\ \mathbf{x}_{t-m} : m \geq 1\}$. The parameters $\omega > 0$ and $\alpha_1, ..., \alpha_p, \beta_1, ..., \beta_q, \gamma \geq 0$ are time-invariant and, when $\gamma = 0$, the PARX reduces to the PAR model. As for ARMA-type processes, the system in (1)-(2) is labelled as PARX(p, q). One particular feature of the model in (1)-(2) is that, in the case of a single covariate, $\mathbf{x}_{t-1} = x_{t-1}$, the expected value of the number of goals is given by

$$E\left[y_t\right] = E\left[\lambda_t\right] = \frac{\omega + E\left[x_{t-1}\right]}{1 - \sum_{j=1}^{\max(p,q)} \left(\alpha_j + \beta_j\right)} \tag{3}$$

and $Var\left[y_t\right] \geq E\left[y_t\right]$, that is the model is able to capture overdispersion in the marginal distribution. The reader is referred to Agosto et al. (2016) for more details and properties of the PARX model.[1]

2.1 Estimation, model selection and specification tests

Following the formalization in Agosto et al. (2016) the conditional log-likelihood of the model in (1)-(2) for the parameter vector $\theta = \left(\omega, \alpha_1, ..., \alpha_p, \beta_1, ..., \beta_q, \gamma\right)'$ is given by

$$\ell_T\left(\theta\right) = \sum_{t=1}^{T} l_t\left(\theta\right), \qquad l_t\left(\theta\right) := y_t \log \lambda_t\left(\theta\right) - \lambda_t\left(\theta\right).$$

The maximum likelihood estimator of θ is given by

$$\hat{\theta} = \arg\max_{\theta} \ell_T\left(\theta\right). \tag{4}$$

The maximization problem in (4) is subject to the restrictions $\omega > 0$, $\alpha_1, ..., \alpha_p, \beta_1, ..., \beta_q, \gamma \geq 0$, and $\sum_{j=1}^{\max(p,q)} \left(\alpha_j + \beta_j\right) < 1$. The first set of conditions are required to ensure that $\lambda_t > 0$ while the latter is used to ensure the stability of the process. The innovations are assumed to be i.i.d. and jointly i.i.d. with y_t over time. The latter assumption does not imply that innovations and the current number of goals are independent. On the contrary, simultaneous dependence between y_t and innovations to the exogenous variables is allowed. These conditions imply that the PARX model admits a stationary and weakly dependent solution. The above restrictions mimic the ones used in GARCHX(p, q) models (see Han and Kristensen, 2014) and are discussed more in detail in Agosto et al. (2016). Model selection, i.e., the selection of the lags of λ_t and y_t in (2) (p and q, respectively), can be performed according to an information criterion. One of the commonly used criteria is the Akaike Information Criterion (AIC; Akaike, 1974):

$$AIC(\hat{\theta}) = -2\ell_T(\hat{\theta}) + 2k$$

[1] The specification of (3) in Agosto et al. (2016) consider $E\left[f\left(x_{t-1}\right)\right]$ instead of $E\left[x_{t-1}\right]$ to ensure that the covariates x_{t-1} are positive.

where k denotes the number of model parameters. In spite of its well known tendency to overparametrise models also in large samples, AIC is particularly useful in a forecasting context as it is asymptotically equivalent to cross-validation (Stone, 1977).

To evaluate the goodness of fit of the specified PARX models, we consider two different tools designed for time series count data.[2]

The first diagnostic tool is the analysis of the standardised Pearson residuals given by

$$\epsilon_t = \frac{y_t - E(y_t|\mathcal{F}_{t-1})}{\sqrt{Var(y_t|\mathcal{F}_{t-1})}} \tag{5}$$

where, since $y_t|\mathcal{F}_{t-1} \sim Pois(\lambda_t)$, $E(y_t|\mathcal{F}_{t-1}) = Var(y_t|\mathcal{F}_{t-1}) = \lambda_t$. If the model is correctly specified, the estimated counterparts of the standardised Pearson residuals in (5) should be a white noise process.

The second tool is the (randomised) probability integral transform (PIT) introduced by Brockwell (2007) and generally adopted to evaluate the misspecification of Poisson autoregressive models as in Davis and Liu (2015), Agosto et al. (2016) and Liboschik et al. (2016). For any t, the PIT is defined as

$$\tilde{u}_t = F_t(y_t - 1) + \upsilon_t[F_t(y_t) - F_t(y_t - 1)] \tag{6}$$

where υ_t is a sequence of i.i.d. uniform (0,1) random variables and $F_t(\cdot)$ is the predictive cumulative distribution, which in our case is the CDF of a Poisson with parameter λ_t. If the model is correctly specified, then $\tilde{u}_t$ is an i.i.d. sequence of uniform (0,1) random variables.[3]

Recently, Kheifets (2015) argues that applying Kolmogorov-Smirnov type tests to check the functional form of a marginal distribution and whether it is i.i.d. is a mistake, since this test is designed to verify the marginal distribution of i.i.d. random variables and should not be used with generalised residuals as in (6). In particular, the null hypothesis requires that $\tilde{u}_t$ is simultaneously uniform and independent (and not uniform under independence). Therefore, standard Kolmogorov-Smirnov tests may miss important deviations from the null as does not control the dynamics in $\tilde{u}_t$. The new test proposed by Kheifets (2015) allows to measure how far $\tilde{u}_t$ is from being independent and uniform. Specifically, under the null for $r = (r_1, r_2) \in [0,1]^2$ and for the l-th lag ($l = 1, 2, ...$), we test for pairwise independence, i.e., $P(\tilde{u}_t \leq r_1, \tilde{u}_{t-l} \leq r_2) = r_1 r_2$. Following Kheifets (2015), we define the process

$$V_{2T,l}(r) = \frac{1}{\sqrt{T-l}} \sum_{t=l+1}^{T} (I(\tilde{u}_t \leq r_1) I(\tilde{u}_{t-l} \leq r_2) - r_1 r_2),$$

[2]For more details on these methods, see Davis et al. (2003), Jung and Tremayne (2011) and Davis and Liu (2015).

[3]Note that the asymptotic distribution of the Kolmogorov-Smirnov tests is invalid due to the estimation of the model parameters (Kheifets, 2015). Therefore, we also consider an i.i.d. bootstrap approximation of the test critical values.

where $I(\cdot)$ is the indicator function, and the test statistics

$$D_{2T,l}(r) = \max_{[0,1]^2} \left| V_{2T,l}(r) \right|. \tag{7}$$

Critical values for (7) can be obtained by i.i.d. bootstrap approximation as in Kheifets (2015).

3 Forecasting football matches with PARX

In this section we outline how the PARX model can be applied to forecast football match outcomes. We also discuss forecasting accuracy and how forecasts can be used to develop a profitable betting strategy.

First, we show that the Poisson distribution is not the suitable probability distribution for analysing and forecasting football match outcomes and that PARX models are useful in this context. In particular, Figure 1 shows some preliminary analysis on the goal distribution of four teams which played the last seasons of the English Premier League (from season 2005/2006 to season 2015/2016). In Panel A (B) of Figure 1 we depict the empirical home (away) goals distribution and the corresponding reference theoretical Poisson distribution with parameter λ (black lines in Figure 1), where λ is the mean of goals scored in the sample considered. The comparison between the (marginal) empirical distributions and the corresponding Poisson distributions shows the presence of overdispersion as stressed by the results in Panel C of Figure 1 where, for all the teams considered, the mean of goals is smaller than the variance.[4]

All these results underscore the fact that the (marginal) empirical goals distribution is not a Poisson, and this is mainly due to overdispersion. As discussed in Agosto et al. (2016; section 3), PARX models are able to capture overdispersion in marginal distributions and therefore are potentially suitable for analysing and forecasting the dynamics of the goals distribution.

[Figure 1 about here]

The analysis and forecast of the outcomes of football matches is obtained as follows. For each match, we estimate two PARX models, one for the home team and one for the away team. In particular, we define two different (conditional) Poisson distributions for each match. Let H denote the home team and A the away team in a specific match, so that $y_t^H|\mathcal{F}_{t-1} \sim Pois(\lambda_t^H)$ is the distribution of the goals scored by the home team when it plays at home, and $y_t^A|\mathcal{F}_{t-1} \sim Pois(\lambda_t^A)$ is the number of goals scored by the away team when it plays away. Once the two (conditional) Poisson distributions for home and away

[4]Overdispersion has been formally tested using the equidispersion test by Cameron and Trivedi (1990) where the null hypothesis of mean-variance equality $H_0 : Var(y_t) = E(y_t) = \lambda_t$ is checked against the alternative $H_1 : Var(y_t) = \lambda_t + \alpha g(\lambda_t)$, where $g(\cdot)$ is a specific function that maps from R^+ to R^+. The results for the four teams considered in Figure 1 show significant overdispersion for the empirical distributions of Aston Villa's home goals and Manchester City's home and away goals, for different choices of the specific function $g(\cdot)$. However, it should be noted that Cameron and Trivedi (1990)'s test must be employed with caution in our context as it assumes independence across observations.

goals are estimated using the PARX model in (1)-(2), it is easy to derive the two forecast distributions. As we consider two different models, one for the home team and one for the away team, there is no need to take into account any additional parameters for home advantage as, e.g., in Dixon and Coles (1997). Indeed, the sample considered for estimating the PARX model for team H (team A) consists of the matches played by H at home (A away) only. The two different and independent PARX models for H and A teams allow to capture the home advantage and takes into account the fact that the means (and the variances) of the number of goals scored at home and away by a team are usually different. The use of the PARX model for forecasting is discussed in Agosto et al. (2016; section 4.3) and it is shown to be very similar to forecasting with GARCH models (Hansen et al., 2012). In our case, we are interested in one-step ahead forecasts, i.e., $y^i_{T+1}|\mathcal{F}_T$, for $i = H, A$, which denotes the number of goals scored by a team in the next match conditional on the information available at time T. The conditional distribution of $y^i_{T+1}|\mathcal{F}_T$ is a Poisson of parameter λ^i_{T+1}, so it is necessary to compute λ^i_{T+1} to obtain a forecast of the number of goals in the next match. Given the information available at time T and the vector of parameters θ, the value $\lambda^i_{T+1|T}$ is given by the process

$$\lambda^i_{T+1|T}(\theta) = \omega + \sum_{j=1}^{p} \alpha_j \lambda^i_{T+1-j}(\theta) + \sum_{j=1}^{q} \beta_j y^i_{T+1-j} + \gamma \mathbf{x}_T, \quad i = H, A.$$

Once we have computed a point forecast of the underlying intensity, $\lambda^i_{T+1|T}(\theta)$, it is straightforward to forecast the distribution of y^i_{T+1} as

$$\hat{P}\left(y^i_{T+1} = y^i|\mathcal{F}_T\right) = Pois\left(y^i|\lambda^i_{T+1|T}(\theta)\right), \qquad y \in \{0, 1, 2, ...\}, \tag{8}$$

where $Pois(y|\lambda) = \lambda^y \exp(-\lambda)/y!$. For our purposes, we first derive the forecast distribution in (8) for the home and away teams, $\hat{P}\left(y^H_{T+1} = y^H|\mathcal{F}_T\right)$ and $\hat{P}\left(y^A_{T+1} = y^A|\mathcal{F}_T\right)$, respectively, and then, assuming the independence of the two distributions, we can derive the joint forecast distribution[5]

$$\hat{P}\left(y^H_{T+1} = y^H, y^A_{T+1} = y^A|\mathcal{F}_T\right) = \hat{P}\left(y^H_{T+1} = y^H|\mathcal{F}_T\right) \cdot \hat{P}\left(y^A_{T+1} = y^A|\mathcal{F}_T\right), \tag{9}$$

for $y^i \in \{0, 1, 2, ...\}$. In other words, the estimation of the intensities $\lambda^H_{T+1|T}$ and $\lambda^A_{T+1|T}$ allows to derive the probability associated with any possible match outcome. For instance, given $\hat{\lambda}^H_{T+1|T}$ and $\hat{\lambda}^A_{T+1|T}$, the probability that the home team wins is given by $\hat{P}(y^H_{T+1} > y^A_{T+1}|\mathcal{F}_T)$, the probability of a draw is $\hat{P}(y^H_t = y^A_t|\mathcal{F}_T)$, and the probability of an away win is $\hat{P}(y^H_t < y^A_t|\mathcal{F}_T)$. On the basis of the estimated

[5]In their paper, Dixon and Coles (1997) find that the assumption of independence between scores is reasonable except for the results 0-0, 1-0, 0-1 and 1-1. However, by replicating their analysis based on the Pearson's chi-squared test for independence considering the matches played in the Premier League seasons from 2005/2006 to 2015/2016, we find that the test statistic is 46.18 ($DF = 36$, p-value = 0.1190). Therefore, we do not reject the null hypothesis of independence between home and away goals.

joint probabilities, we can also compute the aggregate probability for other popular bets such as the total number of goals and under/over, e.g., the probability that the number of total goals in the match will be equal to 1 is $\hat{P}(y^H_{T+1} + y^A_{T+1} = 1|\mathcal{F}_T)$, while the probability of over 2.5 goals is $\hat{P}(y^H_{T+1} + y^A_{T+1} \geq 3|\mathcal{F}_T)$.

In our analysis, we consider the mean of the goals conceded by the opponent team as the only covariate, so that $\mathbf{x}_{t-1} = x_{t-1}$ in (2). This covariate plays a decisive role in our framework, as the mean of the goals conceded by the opponent can be interpreted as a proxy of the defensive ability of the team. Indeed, a team with a good defence (a small value of x_{t-1}) tends to concede fewer goals than a team with a poor defence (a high value of x_{t-1}). We have also tried other covariates as proxies of team ability and/or form. However, we empirically find that the covariate of the mean of goals conceded by the opponent team provides the best results in terms of return of the proposed betting strategy.[6]

As an example, consider the match played between Aston Villa and Chelsea on March 15^{th}, 2014. The match ended with the (true) result of 1-0. The data we consider for home (Aston Villa) and away (Chelsea) teams are summarised in Table 1.

[Table 1 about here]

The value of $x_{t-1} = 1.09$ reported in the fourth column and last row of Table 1 indicates that Chelsea conceded, on average, 1.09 goals in the last three seasons[7] (prior to March 15^{th}, 2014) when it played away. Analogously, the value of $x_{t-1} = 1.32$ is the mean of the goals conceded by Aston Villa in the last three seasons, when it played at home. Figure 2 shows the time series dynamics of the goals scored at home by Aston Villa (first panel) and away by Chelsea (second panel) and the respective autocorrelation functions. From this figure, we can observe that there are periods in which the teams tend to score more goals than in other periods thanks to, for instance, an improvement of the physical (or tactical) condition. This phenomenon can be interpreted as 'goal clustering', analogous to the well-known volatility clustering in finance literature which is usually modelled by GARCH models and is well captured by PARX models. Indeed, 90% confidence bands represented by blue lines in Figure 2 show statistically significant autocorrelations, in particular lag 3 and lag 2 for Aston Villa and Chelsea, respectively.

[Figure 2 about here]

The model selection approach based on the AIC described in section 2.1 select a PARX(0,2) to model the goals scored by Aston Villa while a PARX(3,0) model is suggested to model the goals scored by Chelsea. The estimated parameters for the two models are reported in Table 2. The results in Table 2

[6]The results for other covariates are available upon request.

[7]In our analysis we consider the last three seasons before the date of the match. This choice is a compromise between the need to have a sufficient number of observations for estimating purposes and the issue of the presence of structural breaks which obviously may occur among different seasons.

show that the estimated covariate coefficients γ are both highly statistically significant as well as β_3 and α_2 for Aston Villa's PARX(0,3) and Chelsea's PARX(2,0), respectively. The significance of the autoregressive coefficients are in line with the autocorrelation functions shown in Figure 2 and the significance of γ coefficients strengthens the use of the PARX specification over the simple PAR model.[8]

[Table 2 about here]

The forecast distribution of the goals scored by Aston Villa (home team) against Chelsea (away team) is $\hat{P}\left(y^H_{T+1} = y^H|\mathcal{F}_T\right) = Pois\left(y^H|\hat{\lambda}^H_{T+1|T} = 1.8104\right)$, whereas the forecast distribution of goals scored by Chelsea (away team) against Aston Villa is $\hat{P}\left(y^A_{T+1} = y^A|\mathcal{F}_T\right) = Pois\left(y^A|\hat{\lambda}^A_{T+1|T} = 1.5814\right)$. Therefore, since the expected value of a Poisson distribution equals the intensity parameter, the expected number of goals scored by Aston Villa versus Chelsea is 1.8104. On the other hand, Chelsea is expected to score 1.5814 goals against Aston Villa.

[Table 3 about here]

The joint probability distribution summarised in Table 3 allows the computation of any possible match result and, thus, it is extremely useful for pursuing a profitable betting strategy. For instance, the results in Table 3 allow us to compute the probability that either Aston Villa or Chelsea wins and the probability of a draw; i.e., $\hat{P}\left(y^H_{T+1} > y^A_{T+1}|\mathcal{F}_T\right) = 0.4349$, $\hat{P}\left(y^H_{T+1} < y^A_{T+1}|\mathcal{F}_T\right) = 0.3398$, and $\hat{P}\left(y^H_{T+1} = y^A_{T+1}|\mathcal{F}_T\right) = 0.2253$, respectively, which are obtained as the sum of the probabilities above the main diagonal, below the main diagonal and those on the main diagonal, respectively. The results in Table 3 show that the most likely match result is 1-1 with probability 0.0963, followed by 2-1 (0.0872). The probability associated with the true result (1-0) is 0.0609.

Using the specification tests discussed in section 2.1 we check the model specifications for the example considered. In Panel A of Table 4 we report the Kolmogorov-Smirnov tests to check if the PIT in (6) is a uniform (0,1) distribution. Both asymptotic and bootstrap (in brackets in Table 4) p-values confirm that the PITs for Aston Villa and Chelsea are two i.i.d. sequences of uniform (0,1) distributions. To complete our specification analysis we also perform the test proposed by Kheifets (2015) and outlined in 7 to jointly evaluate uniformity and l-th lag ($l = 1, ..., 5$) pairwise independence of the PIT. The results reported in Panel B of Table 4 confirm the uniform distribution and serial independence of the PIT for the first five lags. These two diagnostic tools ensure that the two PARX models are correctly specified and can thus be used for forecasting purposes.

[Table 4 about here]

[8] All the analyses were performed using MATLAB. The programs are available on request. Alternatively, the recent R package 'tscount' by Liboschik, Fokianos and Fried (2016) can be used to estimate PAR and PARX models, as well as other generalised linear models.

The forecasting accuracy is evaluated using the Mean-Squared Forecasting Error (MSFE)

$$MSFE^i = \frac{1}{N}\sum_{s=1}^{N}\left(y_s^i - \hat{\lambda}_s^i\right)^2, \quad i = H, A \tag{10}$$

where N is the number of matches analysed, y_s^i is the true number of goals scored in match s and $\hat{\lambda}_s^i$ is the one-step ahead forecast of the intensity, that is the expected value of the number of goals for team i in match s.

3.1 Betting strategy

We now define a simple and profitable betting strategy, which is similar in spirit to the one adopted by Dixon and Coles (1997) and Koopman and Lit (2015), which exploits the predictions obtained by the PARX models.

In particular, we use the joint distribution in (9) to derive the probability associated with any possible outcome. Table 3 shows the results of the example analysed in the previous section (Aston Villa vs. Chelsea). Once a table like Table 3 is computed for each match, it is easy to develop a betting strategy for the results 1 (home win), X (draw), and 2 (away win), which is one of the most popular betting choices offered by the market. To each of the results 1, X and 2 is associated an odd. An odd is how much the bet is paid. For instance, the odd associated with result 1 for Aston Villa vs. Chelsea match (played on March 15^{th}, 2014) is 6.80, this means that if someone bets £1 on Aston Villa he wins £6.80 (i.e., the net profit is $6.80-1 =$ £5.80). In our analysis, we consider a set of 26 international bookmakers.[9] The different odds are averaged and the mean odd is considered in the analysis. The data are taken from the website `www.betexplorer.com`, a large database containing the results and the odds for different football championships, as well as other sports.

The betting strategy we propose is based on two conditions:

1. Select the result associated with the highest probability;
2. Evaluate whether this probability is appealing with respect to the offered odd.[10]

More formally, let $P_1 = \hat{P}\left(y_{T+1}^H > y_{T+1}^A|\mathcal{F}_T\right)$, $P_X = \hat{P}\left(y_{T+1}^H = y_{T+1}^A|\mathcal{F}_T\right)$, $P_2 = \hat{P}\left(y_{T+1}^H < y_{T+1}^A|\mathcal{F}_T\right)$ be the probabilities and O_1, O_X, O_2 the odds associated with results 1, X and 2, respectively. The first step of the betting strategy is to select the most likely result, i.e. $\arg\max_{b=1,X,2} P_b$, which, in the case of Aston Villa vs. Chelsea, is the home win with $P_1 = 0.4349$. The second step consists of deciding whether betting on

[9]The 26 bookmakers considered in the analysis are: 10Bet, 12BET, 188BET, 888sport, bet-at-home, bet365, Betclic, Betfair, Betsafe, Betsson, BetVictor, Betway, bwin, ComeOn, Expekt, Interwetten, Ladbrokes, mybet, Paddy Power, Pinnacle, SBOBET, Sportingbet, Tipico, Unibet, William Hill, youwin.

[10]The proposed betting strategy is chosen in order to maximise the probability of winning. Indeed, condition 1 implies that the probability of selecting the correct match outcome is higher than one third.

this result is profitable. Let P_b^o be the implicit probability defined as the inverse of the odds associated with result b, for $b = 1, X, 2$. In the previous example, $P_1^o = 6.80^{-1} = 0.1471$. Therefore, according to the bookmakers, the probability of an Aston Villa's win is less than 15%, against 43.5% predicted by the PARX model, hence the payout proposed by the bookmaker's odd is higher than expected. The expected value of the bet for result 1 (home win), say B_1, is then given by $E[B_1] = \frac{P_1}{P_1^o} - 1$. We bet on home win only if $E[B_1] > 0$, i.e., only if the probability estimated by the PARX model is higher than the implicit probability (the inverse of the odd) proposed by the market ($P_1 > P_1^o$). In the case of the match between Aston Villa and Chelsea, $P_1 = 0.4349 > P_1^o = 0.1471$, hence, according to our betting strategy, it is rational to place a bet on an Aston Villa win. In this way we develop a betting strategy which detects and exploits the most profitable (underpriced) odds in the market.

Following the idea proposed by Dixon and Coles (1997) and Koopman and Lin (2015), we also propose an alternative strategy. In particular, we consider picking only the matches whose $E[B_b] > \tau$, i.e., only if $P_b > P_b^o(1 + \tau)$, where $\tau > 0$ and $b = 1, X, 2$. Therefore, we only bet on the match outcomes whose profitability is higher than a specific threshold τ. In the example considered, we bet on a Aston Villa's win in its home match against Chelsea because $P_1 > P_1^o$ (case of $\tau = 0$). However, by adopting this alternative strategy, we bet only if $0 < \tau < E[B_1] = \frac{P_1}{P_1^o} - 1 = 1.957$. Thus, it is still convenient to bet on an Aston Villa win in this match as long as we select a threshold $\tau < 1.957$.

4 Empirical analysis of the English Premier League

In this section we evaluate the forecasting accuracy discussed in section 3 and the performance of the betting strategy described in section 3.1 in predicting the outcomes of one of the most popular and betted football championships in Europe: the English Premier League. In particular, we analyse the matches played in the Premier League in 2013/2014, 2014/2015 and 2015/2016 seasons. We exclude from the analysis the matches (i) where at least one team played fewer than 15 matches in the past 3 years in the Premier League; (ii) played in the first 18 rounds of each season; (iii) played in the last month of each season. These conditions are found to increase the profitability of the betting strategy. In particular, point (i) is necessary to guarantee a sufficient number of observations for the model estimation. Moreover, as pointed out in (ii), the first 18 rounds of each season are used as burn-in with the purpose to capture the information on the current season by using the first part of the season. Finally, point (iii) is considered in order to avoid biases due to the different levels of motivation of teams when playing end of season matches.[11] Indeed, we find that the last month of each season always leads to negative returns for the

[11] It is indeed well-known that, at the end of the season, there may be teams with no possibility of improving their standing position, while others are in desperate need of points (e.g., teams seeking to achieve a prestigious standing position or teams closed to relegation). It is likely that the latter will make an incredible effort to achieve their objective, while the former do not have the same level of motivation.

proposed betting strategy.

For each match played, we estimate the PARX model in (1)-(2), the joint probabilities in (9) and then we apply the betting strategies proposed in section 3.1.

To evaluate its forecasting performance, we compare the predictions of the PARX model with those obtained by the popular approach proposed by Dixon and Coles (1997) [D&C, hereafter]. In addition, with the purpose of evaluating the role of the exogenous variable in forecasting the number of goals, we also compare the PARX forecasting performance with that of the pure PAR model. In Table 5 we show the ratios between the MSFEs of the PARX and D&C models (Panel A) and the MSFE ratios between the PARX and the PAR models (Panel B). The results in Panel A of Table 5 show that the prediction of the number of goals of the away team by the PARX model is significantly better than the one provided by the D&C model, while there is no significant improvement in predicting the number of goals scored by the home team. Specifically, for all the seasons considered (a total of $N = 330$ matches), we obtain a ratio of 0.8119 for the away team, which, according to the Diebold-Mariano test (see the asterisks reported in Table 5), is statistically significant. As for the comparison between PARX and PAR models, the results for 'All seasons' reported in the last row of Panel B in Table 5 show that the PARX model outperforms in terms of forecasting performance the pure PAR model in predicting the number of goals for both home and away teams. In particular, according to the Diebold-Mariano tests, the MSFE ratios are significant at 1% and 10% levels for the home and away teams, respectively. These results show that the inclusion of the exogenous covariate in the model specification significantly improves the one-step ahead forecasting accuracy.

Before focusing on the betting strategy, we first consider the percentages for results 1, X and 2 for all the matches in the 2013/2014, 2014/2015, and 2015/2016 seasons which satisfy the three conditions outlined above. Result 1 is observed 46.76% of the time, 23.53% of the time matches ended with a draw, while the percentage of result 2 is 29.71%. The corresponding percentages obtained by the PARX model are very close to the actual ones: 46.00% for result 1, 23.02% for X and 30.98% for result 2. It is therefore interesting to note that, unlike other Poisson-based models proposed in the literature, our approach does not seem to underestimate draws.

[Table 5 about here]

We now summarise the main results of the betting strategy which are reported in Table 6 and Figure 3. In particular, we consider the percentage and absolute returns for different values of τ, namely $\tau \in \{0, 0.1, 0.2, 0.3\}$. The results reported in the first column of Table 6 show that this strategy performs reasonably well, even with $\tau = 0$, leading to an absolute return of 24.23 for the 2013/2014 season, 21.58 for the 2014/2015 season and 8.84 for the 2015/2016 season. These absolute returns correspond to a percentage return of 43.27%, 44.96% and 12.63% for the 2013/2014, 2014/2015, and 2015/2016

seasons, respectively. The number of bets placed for these three seasons are 56, 48, and 70, with a winning percentage of 62.50%, 48.00%, and 42.86% respectively. Overall, our betting strategy leads to an aggregate percentage return of 31.41% and an absolute return of 54.65 for all three seasons.

[Table 6 and Figure 3 about here]

In line with Koopman and Lin (2015), the return of the betting strategy improves for higher τ values. This is somewhat expected as consequence of the fact that the higher the value of τ the higher the underpricing, then we are reasonably confident that the bookmaker odds are inaccurate. Given this increased confidence at higher τ values, we expect profitability to increase. Obviously, the number of bets decreases as the value of τ increases. In fact, as reported in the lower panel of Table 6, the number of bets for all three seasons is 174 when $\tau = 0$ and decreases to 115, 80 and 50 when $\tau = 0.1$, 0.2 and 0.3.[12] The results in Table 6 and the upper-left panel of Figure 3 show that $\tau = 0.3$ leads to the best performance in terms of percentage return, that is 76.36%, 124.08%, and 75.29% for the 2013/2014, 2014/2015, and 2015/2016 seasons, respectively, corresponding to an aggregate percentage return of 87.30% for all three seasons. The results in Table 6 and the lower-left panel in Figure 3 show the percentage of winning bets. The interesting feature is that this percentage does not decrease with the value of τ, but remains rather constant. Conversely, the mean of the odds associated with winning bets increases as τ increases (see the results in Table 6 and the lower-right panel of Figure 3). This result is particularly interesting because it provides a clear evidence that our approach is able to detect the mispricing of the odds offered by the betting market, without any loss in the forecasting ability. Indeed, the higher the value of τ, the higher the underpricing of the odd. Nevertheless, our betting strategy delivers similar performances in terms of winning percentage for all the values of τ considered, but an increasing expected profitability.

Finally, we compare the PARX-based betting strategy with three popular 'naive' strategies, namely Home, Longshot, and Favourite (see the caption in Table 6 for more details). The results in Table 6 show that the betting strategy based on the PARX model outperforms the three naive strategies. As a further comparison, we also consider the performance of the same betting strategy as outlined in section 3.1 but based on a simple PAR model. Results show that the inclusion of the exogenous covariate in the model specification improves the performance of the betting strategy for the all seasons and τ values considered.[13] This result can be explained by the fact that, overall in our analysis, the coefficient of the exogenous covariate is statistically significant in 80% of PARX estimations.

[12] When $\tau \geq 0.3$ the number of bets is very small and therefore it becomes somewhat worthless to consider values of τ higher than 0.3.

[13] The results for the PAR-based betting strategies are not reported for reasons of space, but are available from the authors upon request.

5 Conclusions

In this paper we have proposed a novel approach to predict football match outcomes. The analysis is based on the PARX model introduced by Agosto et al. (2016) which allows to model and forecast the goals distribution of a football team by including exogenous variables in the Poisson autoregressive model specification. The role of the covariates is crucial in capturing the key features of the performance of a football team such as attack and defence abilities and form. This method is able to model the autoregressive intensity of the goal scored distributions and the goal clustering phenomenon. With this approach we determine the joint probability distribution of all possible match outcomes. We can then define a suitable betting strategy comparing the probabilities estimated by PARX models and the odds proposed by the bookmakers. The main idea of our betting strategy is to bet only on matches where the probability estimated on the basis of the PARX model is larger than the implicit probability provided by odds, thus identifying the potential mispricing of the odds offered on the betting market. As shown in the empirical analysis based on the matches played in the 2013/2014, 2014/2015 and 2015/2016 English Premier League seasons, the PARX model outperforms the popular Dixon and Coles (1997) approach and the simple PAR model in terms of accuracy in forecasting the number of goals. Moreover, the proposed PARX-based betting strategy leads to a return of 43.27%, 44.96% and 12.63% for the 2013/2014, 2014/2015 and 2015/2016 seasons, respectively. Interestingly, by selecting a threshold $\tau = 0.3$ we achieve a return larger than 87% for these three Premier League seasons.

References

Agosto, A., Cavaliere, G., Kristensen, D. and Rahbek, A. (2016). Modeling corporate defaults: Poisson autoregressions with exogenous covariates (PARX). *Journal of Empirical Finance*, forthcoming.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.

Amisano, G. and Giacomini, R. (2007). Comparing Density Forecasts via Weighted Likelihood Ratio Tests. *Journal of Business and Economic Statistics*, 25, 177–190.

Baio, G. and Blangiardo, M. (2010). Bayesian hierchical model for the prediction of football results. *Journal of Applied Statistics*, 37(2), 253–264.

Brockwell, A.E. (2007). Universal residuals: A multivariate transformation. *Statistics and Probability Letters*, 77(14), 1473-1478.

Cameron, A. C. and Trivedi, P. K. (1990). Regression-based Test for Overdispersion in the Poisson Model. *Journal of Econometrics*, 46: 347–364.

Constantinou, A. C., Fenton, N. E. and Neil, M. (2013). Profiting from an inefficient Association Football gambling market: Prediction, Risk and Uncertainty using Bayesian networks. *Knowledge-Based Systems*, 50: 60–86.

Constantinou, A. C., Fenton, N. E. and Neil, M. (2012). pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Systems*, 36, 322–339.

Davis, R., Dunsmuir, W. and Streett, S. (2003). Observation-driven models for Poisson counts. *Biometrika*, 90, 777-790.

Davis, R. and Liu, H. (2015). Theory and inference for a class of nonlinear models with application to time series of counts. *Statistica Sinica*, forthcoming.

Dixon, M.J. and Coles, S.G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Applied Statistics*, 46(2), 265–280.

Fokianos, K., Rahbek, A., and Tjostheim, D. (2009). Poisson autoregression. *Journal of the American Statistical Association*, 104, 1430–1439.

Forrest, D., Goddard, J. and Simmons, R. (2005). Odds-setters as forecasters: the case of English football. *International Journal of Forecasting*, 21, 551–564.

Forrest, D., and Simmons, R. (2002). Outcome uncertainty and attendance demand in sport: the case of English soccer. *The Statistician*, 2, 241-291.

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

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

Han, H., and Kristensen, D. (2014). Asymptotic theory for the QMLE in GARCH-X models with stationary and non-stationary covariates. *Journal of Business and Economic Statistics*, 32, 416-429.

Joseph, A., Fenton, N. and Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based System*, 7, 544–553.

Jung, R. and Tremayne, A. (2011). Useful models for time series of counts or simply wrong ones?. *AStA Advances in Statistical Analysis*, 95, 59-91.

Karlis, D. and I. Ntzoufras (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3), 381–393.

Kheifets, I.L. (2015). Specification tests for nonlinear dynamic models. *Econometrics Journal*, 18, 67–94.

Koopman, S.J. and Lit, R. (2015). A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society: Series A*, 178(1), 167–186.

Kuypers, T. (2000). Information and efficiency: an empirical study of a fixed odds betting market. *Applied Economics*, 32, 1353–1363.

Lee, A.J. (1997). Modeling scores in the Premier League: is Manchester United really the best?. *Chance*, 10, 15–19.

Levitt, S.D. (2004). Why are gambling markets organised so differently from financial markets?. *Economic Journal*, 114, 223–246.

Liboschik, T., Fokianos, K. and Fried, R. (2016). tscount: An R package for analysis of count time series following generalized linear models. Vignette of R package tscount version 1.3.0.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36(3), 109–11.

Rotshtein, A., Posner, M. and Rakytyanska, A. (2005). Football predictions based on a fuzzy model with genetic and neural tuning. *Cybernetics and Systems Analysis*, 41(4), 619630.

Spann, M. and Skiera, N. (2009). Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, 28(1), 55–72.

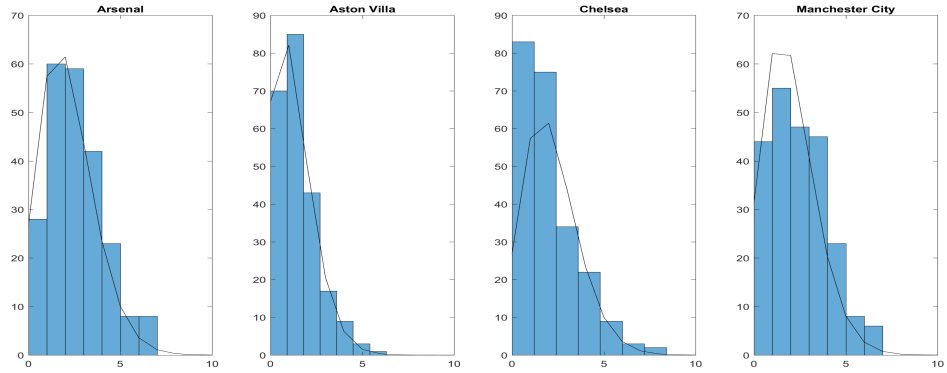
Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B*, 39(1), 44-47.

Tsakonas, A., Dounias, G., Shtovba, S. and Vivdyuk, V. (2002). Soft computing-based result prediction of football games, in: The First International Conference on Inductive Modelling (ICIM2002), Lviv, Ukraine, 2002.

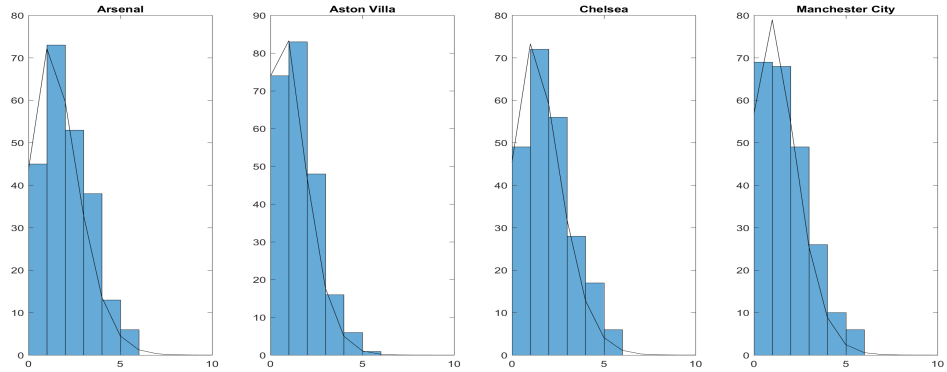
Vlastakis, N., Dotsis, G. and Markellos, R.N. (2009). How Efficient is the European Football Betting Market? Evidence from Arbitrage and Trading Strategies. *Journal of Forecasting*, 28, 426-444.

Woodland, L. and Woodland, B. (1994). Market efficiency and the favourite-longshot bias: the baseball betting market. *The Journal of Finance*, 49, 269–279

Panel A: Home goals distribution



Panel B: Away goals distribution



Panel C: Mean and variance of goals distribution

	Arsenal	Aston Villa	Chelsea	Man. City
Mean (variance) Home	2.14 (2.37)	1.22 (1.42)	2.14 (2.33)	1.99 (2.40)
Mean (variance) Away	1.63 (1.69)	1.12 (1.15)	1.61 (1.75)	1.39 (1.76)

Figure 1: Preliminary analysis of the goals distributions for Arsenal, Aston Villa, Chelsea and Manchester City. Panels A and B depict the home and away goals distribution, respectively, where the blue bars represent the empirical distribution and the black line is the corresponding theoretical Poisson distribution. Panel C reports the means and variances of the goals distributions.

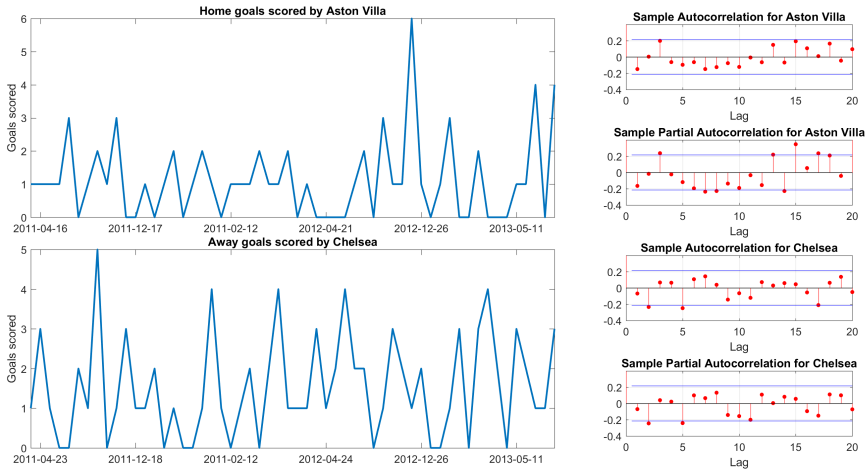


Figure 2: Time series of the goals scored at home by Aston Villa and away by Chelsea in the three years previous to March 15^{th}, 2014. The right panel of the figure shows the autocorrelation and the partial autocorrelation functions (ACFs). Blue lines in the ACFs indicate the 90% confidence bands.

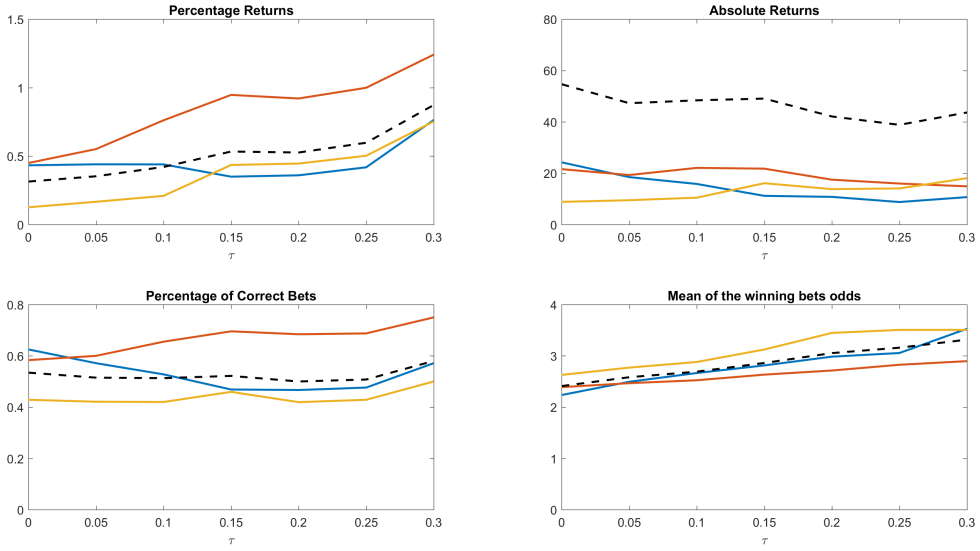


Figure 3: Evaluation of the performance of the PARX-based betting strategy described in section 3.1 for different values of τ, for the 2013/2014 (blue line), 2014/2015 (red line), and 2015/2016 (yellow line) seasons as well as the aggregate results for all three seasons (black dashed line). The upper-left panel shows the percentage returns of the betting strategy, the upper-right panel shows the absolute returns, the lower-left panel displays the percentage of correct bets, and the lower-right panel shows the mean of odds for the winning bets.

Aston Villa				**Chelsea**			
Date	Opponent Team	y_t^H	x_{t-1}	Date	Opponent Team	y_t^A	x_{t-1}
2011-04-10	Newcastle	1	1.53	2011-04-02	Stoke City	1	1.00
2011-04-23	Stoke City	1	1.62	2011-04-16	West Brom.	3	1.63
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2014-02-08	West Ham	0	1.68	2014-02-11	West Brom.	1	1.32
2014-03-02	Norwich	4	2	2014-03-01	Fuhlam	3	1.53
2014-03-15	Chelsea	?	1.09	2014-03-15	Aston Villa	?	1.32

Table 1: Dataset example of the match between Aston Villa (home team) and Chelsea (away team) played on March 15^{th}, 2014. The first (fifth) column reports the home (away) match dates for Aston Villa (Chelsea) in the three years before 2014-03-15, the second (sixth) column reports the opponent teams, the third (seventh) column reports the number of goals scored by Aston Villa (Chelsea) and the fourth (eighth) column indicates the mean of the goals conceded by the opponent team when it played away (at home) in the three years before the date of the match.

	Aston Villa	**Chelsea**
Parameters	PARX(0,3)	PARX(2,0)
ω	0.0001 (0.0001)	0.0001 (0.0005)
α_1	-	0.0001 (0.0005)
α_2	-	0.7205 (8.1657)
β_1	0.0001 (0.0009)	-
β_2	0.0001 (0.0008)	-
β_3	0.3116 (15.471)	-
γ	0.5129 (9.4535)	0.3943 (5.3485)

Table 2: Results for the two PARX models estimated for the example discussed in section 3 of the match between Aston Villa and Chelsea played on March 15^{th}, 2014; t-statistics are reported in brackets.

y_t^H \ y_t^A	$P(y_t^A) = 0$	1	2	3	4	5	6	7	8
$P(y_t^H) = 0$	0.034	0.053	0.042	0.022	0.090	0.002	0.000	0.000	0.000
1	**0.061**	0.096	0.076	0.040	0.016	0.005	0.001	0.000	0.000
2	0.055	0.087	0.069	0.036	0.014	0.004	0.001	0.000	0.000
3	0.033	0.052	0.042	0.022	0.009	0.003	0.000	0.000	0.000
4	0.015	0.024	0.019	0.010	0.004	0.001	0.000	0.000	0.000
5	0.005	0.009	0.007	0.004	0.001	0.000	0.000	0.000	0.000
6	0.002	0.003	0.002	0.001	0.000	0.000	0.000	0.000	0.000
7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 3: Estimated joint probability goal distribution for the match between Aston Villa (y_t^H) and Chelsea (y_t^A) played on March 15^{th}, 2014. In bold, the probability associated with the true result.

Panel A: Kolmogorov-Smirnov test	
$H_0 : \tilde{u}_t \sim U(0,1)$	
Aston Villa	Chelsea
p-value = 0.7282 (0.8788)	p-value = 0.5287 (0.6768)
Panel B: Test for uniformity and independence (Kheifets, 2015)	
$H_0 : P(\tilde{u}_t \leq r_1, \tilde{u}_{t-l} \leq r_2) = r_1 r_2$	
Aston Villa	Chelsea
$l = 1$, p-value = 0.5915	$l = 1$, p-value = 0.6466
$l = 2$, p-value = 0.5414	$l = 2$, p-value = 0.7218
$l = 3$, p-value = 0.3860	$l = 3$, p-value = 0.6165
$l = 4$, p-value = 0.4962	$l = 4$, p-value = 0.5639
$l = 5$, p-value = 0.3684	$l = 5$, p-value = 0.7544

Table 4: Specification tests for the fitted PARX models for the match between Aston Villa and Chelsea played on March 15^{th}, 2014. Panel A: Kolmogorov-Smirnov tests to check whether the PIT in (6) is distributed as a uniform (0,1) distribution; bootstrap p-values are reported in brackets. Panel B: Kheifets' (2015) test to check the PIT l-lag ($l = 1, ..., 5$) pairwise independence and uniformity.

Forecasting performance comparison (MSFE Ratio)		
Panel A: Dixon and Coles (1997)		
Season	Home Team	Away Team
2013/2014	0.9701	0.7511**
2014/2015	1.0005	0.8220*
2015/2016	0.9874	0.8184**
All seasons	0.9743	0.8119***
Panel B: PAR model		
Season	Home Team	Away Team
2013/2014	0.9315***	0.9603
2014/2015	0.9554	0.9615*
2015/2016	0.9599	0.9827
All seasons	0.9467***	0.9683*

Table 5: Forecasting performance comparison between PARX model and Dixon and Coles (1997, D&C) approach (Panel A), and between PARX and PAR models (Panel B). The values reported are the ratio between the MSFEs of the PARX-based approach and the D&C-based (Panel A) or PAR-based (Panel B) approaches. Values lower than one indicate that the PARX provides a better forecasting performance than the competing approach. *, ** and *** denote significance at 10%, 5%, and 1% levels, respectively.

	PARX model				PAR model	Naive betting strategies		
	$\tau = 0$	0.1	0.2	0.3	$\tau = 0$	Home	Favourite	Longshot
2013/2014 season								
Percentage return	43.27	43.94	35.97	76.36	32.19	3.89	3.06	-11.70
Absolute return	24.23	15.82	10.79	10.69	15.45	6.49	1.90	-10.30
Percentage of correct bets	62.50	52.78	46.67	57.14	54.17	49.10	74.19	12.50
Mean of the winning bets odds	2.29	2.72	2.91	3.08	2.44	2.11	1.39	7.06
Number of bets	56	36	30	14	48	167	62	88
2014/2015 season								
Percentage return	44.96	76.14	92.05	124.08	42.14	-0.05	13.89	-35.02
Absolute return	21.58	22.08	17.49	14.89	17.70	-0.07	6.25	-30.12
Percentage of correct bets	58.33	65.52	68.42	75.00	54.76	45.64	80.00	12.79
Mean of the winning bets odds	2.48	2.68	2.80	2.98	2.59	2.19	1.42	5.08
Number of bets	48	29	19	12	42	149	45	86
2015/2016 season								
Percentage return	12.63	20.96	44.55	75.29	-0.21	-2.68	-19.07	32.35
Absolute return	8.84	10.48	13.81	18.07	-0.12	-4.37	-7.82	26.85
Percentage of correct bets	42.86	42.00	41.94	50.00	38.60	41.72	56.10	22.89
Mean of the winning bets odds	2.62	2.88	3.44	3.50	2.58	2.33	1.44	5.78
Number of bets	70	50	31	24	57	163	41	83
All seasons								
Percentage return	31.41	42.07	52.61	87.30	22.47	0.43	0.22	-5.28
Absolute return	54.65	48.38	42.09	43.65	33.03	2.05	0.33	-13.57
Percentage of correct bets	53.45	51.31	50.00	58.00	46.09	45.51	70.95	15.95
Mean of the winning bets odds	2.34	2.64	2.93	3.14	2.54	2.20	1.41	5.93
Number of bets	174	115	80	50	147	479	148	257

Table 6: Performance of the PARX and PAR-based betting strategies and three naive betting strategies. The results for the PARX-based betting strategy are reported for different values of τ. The results for the PAR-based betting strategy are reported for $\tau = 0$ (results for other values of τ are available upon request). 'Home' identifies a betting strategy where the bettor always wagers on the home team (irrespectively of the odd); 'Favourite' denotes a betting strategy where the bettor always wagers on the favourite team (corresponding to odds lower than 1.67, i.e., an implied probability of at least 60%); 'Longshot' identifies a betting strategy where the bettor always wagers on underdogs (corresponding to odds higher than or equal to 4, i.e., an implied probability of 25% at most).