Giovanni Angelini, Luca De Angelis

PARX model for football matches predictions

Quaderni di Dipartimento

Serie Ricerche 2016, n. 2
ISSN 1973-9346

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dipartimento di Scienze Statistiche “Paolo Fortunati”

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

August 31, 2016

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. 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 incorporate also 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 the English football Premier League data for the 2013/2014 and 2014/2015 seasons. The results show that the return from the betting strategy is larger than 35% in all the cases considered and may even exceed 100% if we consider an alternative strategy based on a predetermined threshold which allows to exploit the inefficiency of the betting market.

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

1 Introduction

In the last few years, the football betting market has experienced the fastest growth among the gambling markets (Constantinou and Norman, 2012). Not surprisingly, many different methodologies have been developed to construct a profitable betting strategy which is able to capture the mispricing of the odds.

*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

Starting with the pioneering works of Maher (1982) and Dixon and Coles (1997), many econometric methods have been proposed to predict football matches.

The main purpose of our paper is to develop an approach able to compute a set of probabilities associated with each possible result and use these probabilities to profit from the potential mispricing of the odds offered on the betting market. The odds proposed by the bookmakers often reflect the betting volumes rather than the true probability of the match outcomes. Indeed, the aim of the 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. Moreveor, bookmakers may systematically setting odds in a manner that takes advantage of bettor's biases, such as the well-known preference for favorites and local teams, in order to increase profits (Levitt, 2004). Therefore, the comparison between the true probabilities and the 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 the different match results (Maher, 1982; Dixon and Coles, 1997; Lee, 1997; Karlis and Ntzoufras, 2003). Another approach is to model directly these probabilities 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 the 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 consider 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 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, allowing for a framework where the strengths of attack and defence of the teams can slowly vary stochastically over time.

Our paper focuses on football matches prediction using the Poisson Autoregression with eXogenous covariates (PARX) model introduced by Agosto, Cavaliere, Kristensen and Rahbek (2016) which extends the Poisson Autoregression (PAR) model originally proposed by Fokianos, Rahbek and Tjostheim

(2009) to include covariates in its specification. This model has been successfully used to predict corporate defaults and, in 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. These probabilities can thus be compared with those implied by the odds in order to detect potential mispricing on the betting market. Our approach is slightly different from those proposed in the literature for football matches predictions because the PARX model captures the dynamics of the intensity of the goals distribution. To the best of our knowledge, the only paper which considers a dynamic specification for the intensity has been proposed by Koopman and Lit (2015). The main difference of our method with respect to Koopman and Lit (2015) is that we exploit additional information by including exogenous covariates in the model specification which can greatly improve the forecasting performances. The inclusion of covariates such as proxies for attack or defense ability of the opposing teams can be particularly useful for predicting football matches results as we take into account information related to the strength and/or form of the teams.

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. 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.

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 applying our PARX-based betting strategy to the 2013/2014 and 2014/2015 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.

The rest of the paper is organized as follow. 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, 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 goals 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 goals clustering that, analogously to the well-known volatility clustering in the finance literature, identifies periods in which football teams tend to score more goals than other periods. A further advantage of PAR and PARX models is that they account for the phenomenon of overdispersion, a feature observed in many count data, including goals scored by a football team (see Panel C of Figure 1). The main difference between PAR and PARX models is that the latter allows for the inclusion of exogenous covariates in the model specification. This model extension is particularly suitable in our framework as it allows us to incorporate additional information about the team 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, $\mathcal{F}_{t-1}$ 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

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

and $Var[y_t] \geq E[y_t]$, 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 and model selection

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

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

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

The maximum likelihood estimator of θ is given by

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

and is obtained as the solution of $S_T(\theta) = 0$, where the score $S_T(\theta)$ is defined as

$$S_T(\theta) = \sum_{t=1}^{T} \left(\frac{y_t}{\lambda_t(\theta)} - 1 \right) \frac{\partial \lambda_t(\theta)}{\partial \theta}.$$

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 guarantee that $\lambda_t > 0$ while the latter is used to ensure the stability of the process. 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). Theorem 2 in Agosto et al. (2016) shows that

$$\sqrt{T}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, H^{-1}(\theta_0)\right), \qquad H(\theta) = -E\left[\frac{\partial^2 l_t(\theta)}{\partial \theta \partial \theta'}\right],$$

where the Hessian matrix $H(\theta)$ can be consistently estimated by

$$H_T(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \frac{1}{\lambda_t(\theta)} \left(\frac{\partial \lambda_t(\theta)}{\partial \theta} \right) \left(\frac{\partial \lambda_t(\theta)}{\partial \theta} \right)'.$$

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 criterion is the Akaike Information Criterion (AIC; Akaike, 1974):

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

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

2.2 Specification tests

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

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

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) and Agosto 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 $\tilde{u}_t$ are 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

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

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]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.

3 Forecasting football matches with PARX

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

First, we show that the Poisson distribution is not the suitable probability distribution for analysing and forecasting the football match outcomes and that PARX models are useful in this context. In particular, Figure 1 shows some preliminary analysis on the goals distribution of four teams which played the last ten seasons of the English Premier League (from season 2005/2006 to season 2014/2015). 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 show 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.

All these results underline 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 the marginal distributions and are thus potentially suitable for analysing and forecasting the dynamics of the goals distribution.

[Figure 1 about here]

The analysis and the 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 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 the home (away) team consists of the matches played at home (away) only. The use of 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., $y_{T+1}^i|\mathcal{F}_T$, 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 $y_{T+1}^i|\mathcal{F}_T$ is a Poisson of parameter λ_{T+1}^i, so it is necessary to compute λ_{T+1}^i 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 independence of the two distributions, we can derive the joint forecast distribution[4]

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

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 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 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 opposite team can be interpreted as a proxy of the defence ability of the team. Indeed, a team with a good defence (a small value of x_{t-1}) tends to concede a less number of 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.[5]

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

[4] 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 2014/2015, 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.

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

(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[6] (prior to March 15^{th}, 2014) when it played away. Analogously, the value of $x_{t-1} = 1.32$ is the mean of goals conceded by Aston Villa in the last three seasons, when it played at home. Figure 2 shows the time series dynamics of goals scored at home by Aston Villa (first panel) and by Chelsea (second panel). From this figure, we can observe that there are periods in which the teams tend to score more goals than in other periods. This goals clustering is analogous to the well-known volatility clustering in the finance literature which is usually modelled by (G)ARCH models and is well captured by PARX models.

[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 show that the estimated covariate coefficients γ are both highly significant as well as β_3 and α_2 for Aston Villa's PARX(0,3) and Chelsea's PARX(2,0), respectively.

[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 to pursue a profitable betting strategy. For instance, the results in Table 3 allow us to compute the probability that Aston Villa will win, the one for a Chelsea's win, 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

[6] 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.

$\hat{P}\left(y^H_{T+1} = y^A_{T+1}|\mathcal{F}_T\right) = 0.2253$, 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 2.2 we check the model specifications for the example considered. In Panel A of Figure 3 we show the autocorrelation functions (ACF) of the standardised Pearson residuals in (5) for Aston Villa and Chelsea. The ACFs show no significant lags and hence the residuals follow a white noise process. In Panel B of Figure 3 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 Figure 3) 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 C of Figure 3 confirm the uniform distribution and the serial independence of the PIT for the first five lags. These three diagnostic tools ensure that the two PARX models are correctly specified and can thus be used for forecasting purposes.

[Figure 3 about here]

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

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

where N is the number of matches analysed, y^i_s is the true number of goals scored in match s and $\hat{\lambda}^i_s$ 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), that 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. For the example analysed in the previous section (Aston Villa vs. Chelsea) the results are reported in Table 3. 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 are one of the most popular betting choices offered by the market. To each of the results 1, X and 2 is associated an odd.[7] An odd

[7]In our analysis, we consider a set of bookmakers. The different odds are averaged and the mean odd is then used.

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, that means that if one bets 1£ on Aston Villa he wins 6.80£ (i.e., the net profit is $6.80 - 1 = 5.80$£).

The betting strategy we propose is based on two conditions:

1. Select the result associated with the highest probability;

2. Evaluate if this probability is appealing with respect to the offered odd.

More formally, let $P_1 = \hat{P}\left(y^H_{T+1} > y^A_{T+1}|\mathcal{F}_T\right)$, $P_X = \hat{P}\left(y^H_{T+1} = y^A_{T+1}|\mathcal{F}_T\right)$, $P_2 = \hat{P}\left(y^H_{T+1} < y^A_{T+1}|\mathcal{F}_T\right)$ 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, which, in the case of Aston Villa vs. Chelsea, is the home win with $P_1 = 0.4349$. The second step consists in deciding whether betting on this result is profitable. Let P^o_b be the implicit probability defined as the inverse of the odds associated with result b, for $b = 1, X, 2$. In the previous example, $P^o_1 = 6.80^{-1} = 0.1471$. 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 $E[B_1] = \dfrac{P_1}{P^o_1} - 1$. 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 ($P_1 > P^o_1$). In the case of the match between Aston Villa and Chelsea, $P_1 = 0.4349 > P^o_1 = 0.1471$, hence, according to our betting strategy, it is rational to place a bet on a Aston Villa's 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 $P_b > P^o_b(1 + \tau)$, 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 $P_1 > P^o_1$ (case of $\tau = 0$). However, by adopting this alternative strategy, we bet only if $0 < \tau < E[B_1] = \dfrac{P_1}{P^o_1} - 1 = 1.957$. Hence, it is still convenient to bet on a Aston Villa's 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 and 2014/2015 seasons.[8] 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.

In order to evaluate the forecasting performance of the PARX model we compare it with the one obtained on the basis of the popular model proposed by Dixon and Coles (1997). In particular, we compute the MSFE described in section 3 for both approaches and in Table 4 we report the ratios between the MSFEs of the PARX and Dixon and Coles (1997) models. The results in Table 4 show that the PARX-based approach produces better forecasts for the away goals distribution. Specifically, for all the seasons considered (a total of $N = 263$ matches), we obtain a ratio of 0.7830. According to the Diebold-Mariano tests (see the asterisks reported in Table 4), the prediction of the number of goals of the away team for the PARX model is significantly better than the one for the Dixon and Coles (1997) model, while there is no significant improvement in predicting the number of goals scored by the home team.

[Table 4 about here]

We now summarise the main results of the betting strategy which are reported in Table 5 and in Figures 4 and 5. 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 5 show that this strategy performs reasonably well, even with $\tau = 0$, leading to an absolute return of 24.23 (21.58) which corresponds to a percentage return of 43.27% (44.96%) for the 2013/2014 (2014/2015) season. The number of placed bets for the 2013/2014 (2014/2015) season is 56 (48) with a winning percentage of 62.50% (48.00%).

[Table 5 about here]

In line with Koopman and Lin (2015) and as one might expect, the return of the betting strategy improves for higher values of τ. Obviously, the number of bets decreases as the value of τ increases. In fact, the number of bets decreases to 36, 30 and 14 (29, 19 and 12) when $\tau = 0.1$, 0.2 and 0.3, respectively, for the 2013/2014 (2014/2015) season. 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. The results in Table 5 show that $\tau = 0.3$ leads to the best performance in terms of percentage return, that is 76.36% and 124.08% for 2013/2014 and 2014/2015 seasons, respectively.

[Figures 4 and 5 about here]

[8] From the analysis we exclude: (i) the teams that played less than 15 matches in the last 3 years of the Premier League; (ii) the first 18 rounds of each season; (iii) the last month of each season. Points (i) and (ii) are required to include a sufficient number of observations in the analysis. Point (iii) is considered since we found that the last month of each season often leads to a negative performance.

The first block in Figures 4 and 5 shows the percentage of winning bets. The interesting feature is that this percentage does not decrease with the value of τ but it remains rather constant. This result is particularly interesting because it is 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.

5 Conclusion

In this paper we have proposed a novel approach to predict football matches 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 to capture 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 goals 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 the PARX models and the odds proposed by the bookmakers. The main idea of our betting strategy is to bet only on the matches where the probability estimated on the basis of the PARX model is larger than the implicit probability provided by the odds, thus identifying the potential mispricing of the odds provided by the bookmakers. As shown in the empirical analysis based on the matches played in the 2013/2014 and 2014/2015 English Premier League seasons, the PARX model outperforms the popular Dixon and Coles (1997) model in terms of accuracy in forecasting the number of goals. Moreover, the proposed betting strategy leads to a return of 43.27% and 44.96% for the 2013/2014 and 2014/2015 seasons, respectively. As shown in the empirical analysis, by selecting a threshold $\tau = 0.3$ we achieve a return larger than 100% for the 2014/2015 Premier League season.

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, 177190.

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

Constantinou, A., Fenton, C., Norman, 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.

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.

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.

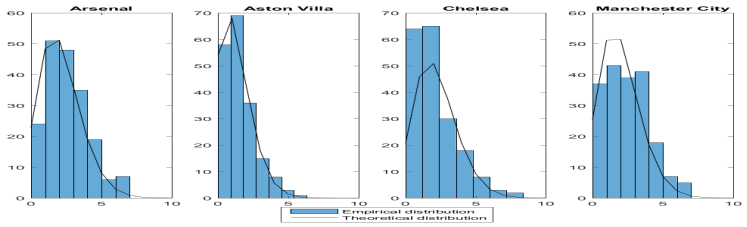
Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36(3), 109–11.

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.

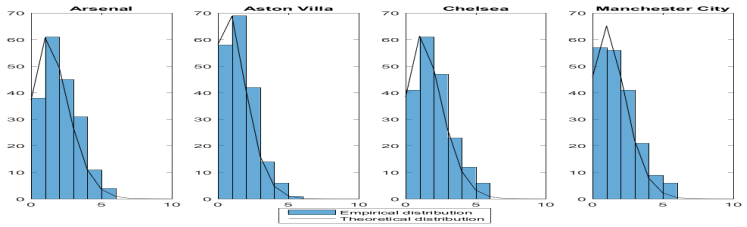
Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B*, 39(1), 44-47.

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.

Panel A: Home goals distribution



Panel B: Away goals distribution



Panel C: Mean and variance goals distribution

	Arsenal	Aston Villa	Chelsea	Man. City
mean (variance) Home	2.12 (2.37)	1.26 (1.52)	2.22 (2.46)	2.01 (2.41)
mean (variance) Away	1.63 (1.65)	1.18 (1.21)	1.60 (1.76)	1.42 (1.86)

Figure 1: Preliminary analysis of the goals distribution. Panel A: home goals distribution. Panel B: away goals distribution. Panel C: mean and variance of the goals distributions.

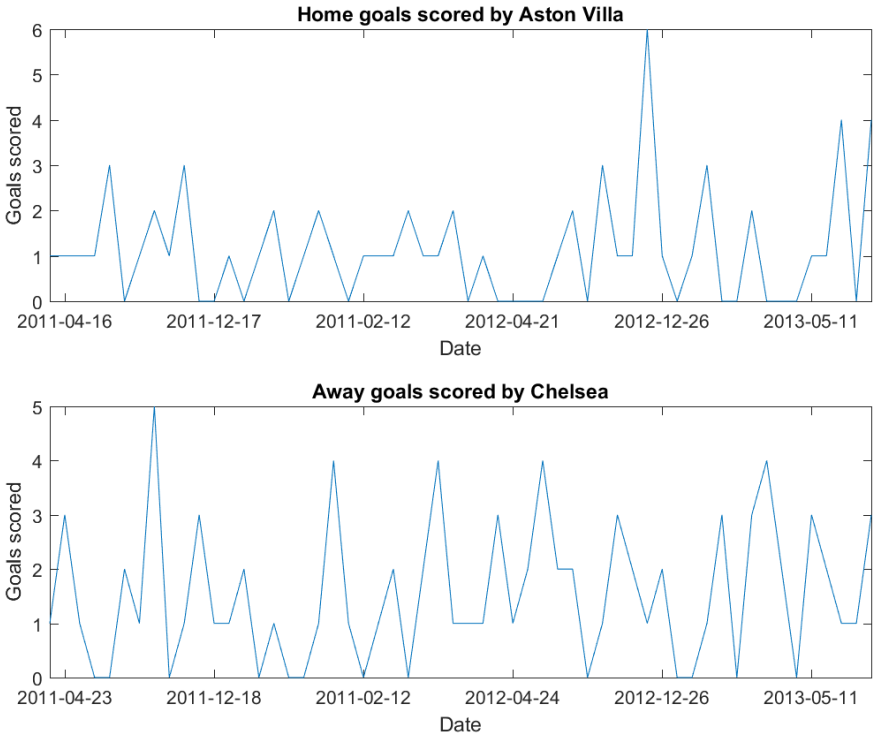
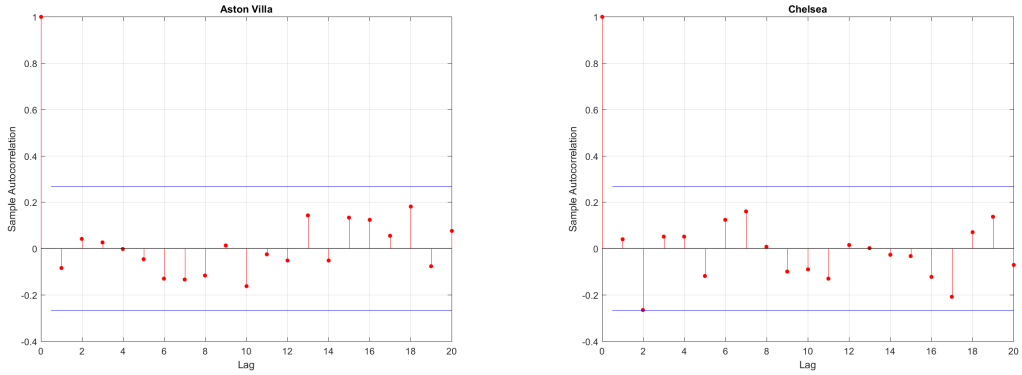


Figure 2: Time series of goals scored at home by Aston Villa and away by Chelsea in the last three years previous March 15^{th}, 2014.

Panel A: Autocorrelation functions



Panel B: 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 C: 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
Test for $l = 1$, p-value = 0.5915	Test for $l = 1$, p-value = 0.6466
Test for $l = 2$, p-value = 0.5414	Test for $l = 2$, p-value = 0.7218
Test for $l = 3$, p-value = 0.3860	Test for $l = 3$, p-value = 0.6165
Test for $l = 4$, p-value = 0.4962	Test for $l = 4$, p-value = 0.5639
Test for $l = 5$, p-value = 0.3684	Test for $l = 5$, p-value = 0.7544

Figure 3: Specification tests for the estimated PARX models for Aston Villa and Chelsea. Panel A: autocorrelation functions of the standardised Pearson residuals for Aston Villa (left) and Chelsea (right). Panel B: Kolmogorov-Smirnov tests to check whether the PIT is distributed as a uniform (0,1) distribution, in brackets the bootstrap p-values. Panel C: Kheifest (2015)'s to test the PIT l-lag ($l = 1, ..., 5$) pairwise independence and uniformity.

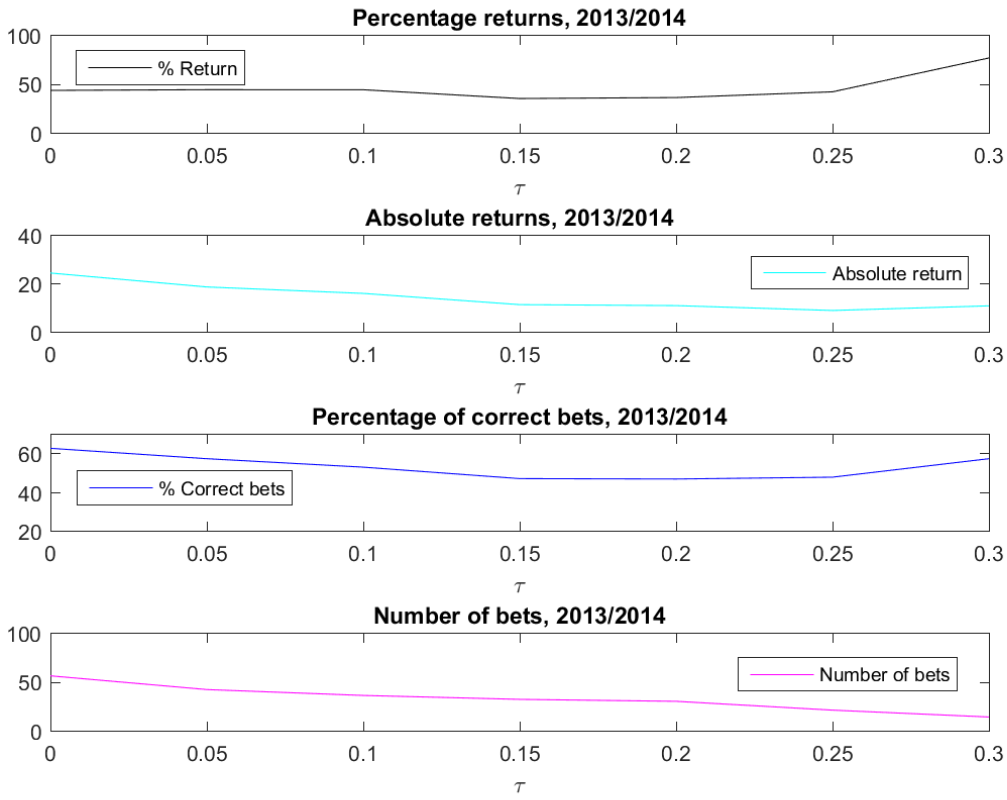


Figure 4: Evaluation of the forecasting performance of the betting strategy for different values of τ - 2013/2014 season.

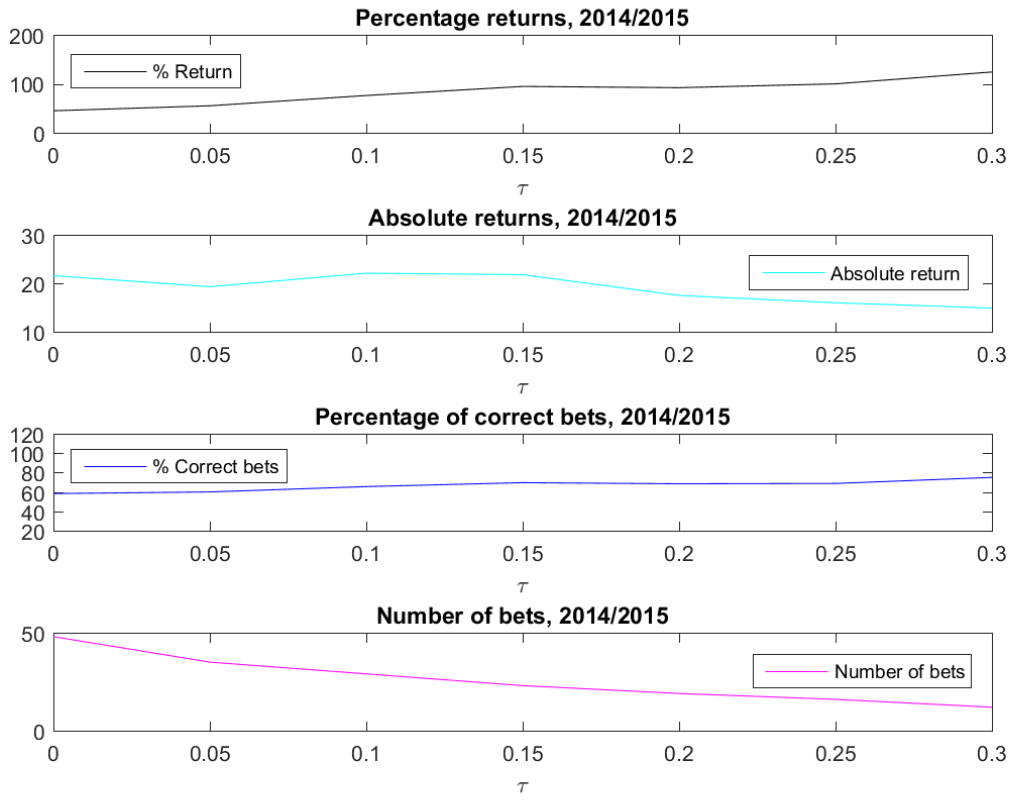


Figure 5: Evaluation of the forecasting performance of the betting strategy for different values of τ - 2014/2015 season.

Aston Villa				**Chelsea**			
Date	Opposite Team	y_t^H	x_{t-1}	Date	Opposite 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 for the match between Aston Villa and Chelsea played on 2014-03-15.

	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: Estimations example for Aston Villa (home team) and Chelsea (away team). In brackets the t-statistics.

y_t^H \ y_t^A	$P(y_t^A) = 0$	1	2	3	4	5	6	7	8	9	10
$P(y_t^H) = 0$	0.034	0.053	0.042	0.022	0.090	0.002	0.000	0.000	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	0.000	0.000
2	0.055	0.087	0.069	0.036	0.014	0.004	0.001	0.000	0.000	0.000	0.000
3	0.033	0.052	0.042	0.022	0.009	0.003	0.000	0.000	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	0.000	0.000
5	0.005	0.009	0.007	0.004	0.001	0.000	0.000	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	0.000	0.000
7	0.000	0.000	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	0.000	0.000
9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 3: Joint probability goals distribution for Aston Villa (y_t^H) and Chelsea (y_t^A). In bold the probability associated with the true result.

	MSFE	
Season	Home Team	Away Team
2013/2014	0.9701	0.7511**
2014/2015	1.0005	0.8220*
All seasons	0.9847	0.7830**

Table 4: Forecasting performance comparison between PARX model and Dixon and Coles (1997, D&C) approach. The values reported are the ratio between the MSFEs of the PARX-based and D&C-based approaches. Values lower than one indicate that PARX provides a better forecasting performance than D&C. ** and * denote significance at 1% and 5% levels, respectively.

	$\tau = 0$	0.1	0.2	0.3
2013/2014 Season				
Percentage return	43.27	43.94	35.97	76.36
Absolute return	24.23	15.82	10.79	10.69
Percentage of correct bets	62.50	52.78	46.67	57.14
Number of bets	56	36	30	14
2014/2015 Season				
Percentage return	44.96	76.14	92.05	124.08
Absolute return	21.58	22.08	17.49	14.89
Percentage of correct bets	58.33	65.52	68.42	75.00
Number of bets	48	29	19	12

Table 5: Forecasting performances of the PARX model for different values of τ.