Proceedings Paper:

Hargreaves, Jessica orcid.org/0000-0002-7173-7902 and Powell, Ben orcid.org/0000-0002-0247-7713 (2022) Modelling bookings in association football. In: Proceedings of MathSport International 2022 Conference. 9th International Conference on Mathematics in Sport, 11-13 Jul 2022, University of Reading. , GBR , pp. 34-44.

Modelling bookings in association football

J. Hargreaves* and B. Powell**

*Department of Mathematics, University of York, York, UK + email address: jessica.hargreaves@york.ac.uk
** Department of Mathematics, University of York, York, UK + email address: ben.powell@york.ac.uk

Abstract

In this report, we describe preliminary investigations into, and opinions on, the subject of statistical analysis of bookings in association football. More specifically, we consider whether more developed methodology for modelling goals can be re-purposed to predict the number of bookings in a given match, and, if so, how. A more thorough description of the topic and a more rigorous analysis of our modelling experiments will appear in forthcoming work.

1 Introduction

1.1 Executive summary

Preliminary modelling suggests that the process by which bookings are made is intrinsically more noisy and less predictable than that for goals. With a low signal-to-noise ratio and a potentially large number of team-specific parameters to infer, over-fitting is a major problem. This leads to poor out-of-sample predictions, which undermine the practical utility of our models. A natural recourse in such situations is to apply shrinkage/variance-reduction methods and/or to constrain models using additional prior knowledge. Our working conclusion is that out-of-the-box shrinkage methods, namely the L_1 coefficient-penalizing Lasso, do help alleviate the over-fitting problem but that further improvements require a more thoughtful consideration of the factors driving bookings.

We note that the models we employ in these investigations are penalized generalized linear models (GLMs), generally not including interaction effects. However, it is possible that this important, and generally well understood, class of models is not the best tool to use for the problem under consideration. Next-generation machine learning models such as artificial neural networks or random forests may allow us to identify and exploit covariate interaction effects that the GLMs are blind to. While potentially valuable, we suspect that these flexible, highly parameterized models will be even more susceptible to over-fitting. Further work in this direction is deferred for future studies.

1.2 Background and Motivation

Since the seminal paper by [Maher, 1982], a substantial statistical literature has developed on the topic of models for association football (soccer) matches and, in particular, probabilistic models for match scores. Maher's model assumes that the numbers of goals scored by each team in a football match follow independent Poisson processes and that the rates at which the teams expect to score goals are functions of the ability of the two teams to attack and defend. Subsequent papers, most notably [Dixon and Coles, 1997], developed the Maher model in a variety of directions. For example: additional important work on an alternative bivariate Poisson model is presented in [Karlis and Ntzoufras, 2003]; dynamic models are presented in [Owen, 2011] and [Koopman and Lit, 2015]; and [Boshnakov et al., 2017] demonstrate improvements in model fit with the use of a Weibull based count model.

The Dixon-Coles model, whose influence on the literature has been especially strong, is essentially a Poisson regression model with one important structural difference. While marginally Poisson, the number of goals scored by each team within a match are, for a certain range of values, modelled as being probabilistically dependent. The dependence is motivated on both theoretical and empirical grounds, which the authors explain in their paper. Covariates for the model include dummy variables encoding offensive and defensive strengths for each team and another encoding whether a team plays at their home ground. The regression coefficients are implicitly time varying, with their values being estimated using data that is weighted according to an envelope/kernel function applied to match dates. In Section 3 we present first attempts at applying modified versions of the Dixon-Coles model to bookings for each team in a match.

2 Exploratory data analysis

Our principle data set for this study, which is publicly available at `football-data.co.uk`, consists of UK Premier League matches between 2011 and 2019. These data inform the majority of the findings below, and the modelling exercise in Section 3. The period has been intentionally chosen to avoid the effects of the COVID–19 pandemic, whose destabilizing effects are partially explored in Section 2.1.

2.1 COVID–19 pandemic effects

Due to social restrictions during the COVID–19 pandemic, many professional sports leagues around the world experienced varying levels of spectator reductions during 2020–2021. Several studies (for example, [Endrich and Gesche, 2020, Tilp and Thaller, 2020, Fischer and Haucap, 2021, McCarrick et al., 2021, Bryson et al., 2021]) have since investigated the effects of these reductions on home performance and referee decisions. The studies reach a general consensus that lower spectator numbers affected both home team performance and referee decisions. In particular, as pointed out by [Endrich and Gesche, 2020], home teams appeared to be treated less favorably during this period due to the reduction in crowd pressure on referees. Additionally, [Bryson et al., 2021] identify a (statistically significant) reduction in the number of bookings for away teams, thus reducing the ‘home advantage’ further.

These findings have the potential to adversely affect our current work modelling bookings. To assess this potential more formally we investigated goal and booking counts in the UK Premier League during the COVID period using a set of varying coefficient models. These models, introduced by [Hastie and Tibshirani, 1993] and developed more recently by [Fan and Gijbels, 2018], allow us to estimate smoothly time-varying covariate effects. Preliminary analyses (illustrated in Figures 1 and 2) indicate distinct and atypical changes in base-rates for goal and booking counts as well as relationships between the counts and match locations, and the counts and bookmaker’s odds. These observations align broadly with the aforementioned studies and motivate the exclusion of the COVID-affected seasons from our main modelling exercises.

2.2 Team interaction effects

The title of this subsection euphemistically refers to the antagonistic relationships and rivalries between specific pairs of teams, often, but not exclusively, based on geographical proximity. In Figure 3 we present means of counts of bookings for matches for different team pairings. Immediate conclusions are not obvious from the plot. In principle, very large values in a row or column with mostly small values would indicate that an interaction effect which, if not accounted for, could skew team-specific parameters inappropriately. We suggest that further analysis of the significance and persistence over time of the apparent interaction effects is required before we attempt to incorporate them into models. For now, however, we choose not to do so.

2.3 Team strength disparity effects

Next, we take a closer look at the relationships between a team's strength relative to its opponent in a match and the number of goals scored and bookings incurred in that match. We quantify a team's relative strength as the logarithm of the ratio of its win and loss probabilities as implied by bookmaker's odds. In Figure 4 we illustrate the relationships found in the data. Obviously, we see a strong positive correlation between a team's relative strength and the number of goals it scores. We also see a less strong, but still statistically significant, negative correlation between relative strength and the number of bookings incurred. This finding is in line with the idea that weaker teams resort to foul play more often when faced with a more technically able opponent.

2.4 Temporal effects

Our analyses appear to show no significant systematic variation in marginal goal counts over time. We do, however, detect temporal effects on the booking count. To be more precise, these effects have been quantified by fitting to the counts Poisson regression models whose covariates consist of a linear trend and a pair of sinusoids with a period of one year. While these models are clearly only able to capture very simple dynamics, they generally support arguments against including temporal or seasonal effects in models for goals but for including them in models for bookings. The latter argument is further strengthened by anecdotal reports of teams being more cautious about incurring bookings as they accumulate over a season.

2.5 Home/away effects

The main factors impacting home advantage are believed to be: crowd support (through *both* the encouragement of home players' performance and biasing referee's decisions (e.g. penalties and bookings) in favor of the home team); familiarity with the stadium; and travel fatigue. In-depth discussion of these topics can be found in [Ponzo and Scoppa, 2018, Tilp and Thaller, 2020].

In the current work, the home advantage manifests itself numerically in significant coefficient estimates in the models of Section 3 and graphically in Figure 6 where we see goal counts and booking counts skewed in favour of the home team. The significance of both effects is also detectable from simple paired t-tests, for example, which show the average home team scoring approximately 0.35 more goals and incurring 0.24 fewer bookings than the away team.

2.6 Gender effects

Quantifying differences between goal and booking counts for women's and men's football is undermined by the relative scarcity of data relating to the latter. Nevertheless, data collected from the RSSSF `http://www.rsssf.com/intland-women.html`, for example, appear to show qualitatively similar phenomena for both genders. The home advantage effect, for instance, appears as a statistically significant difference in home and away goal counts of approximately 0.68 for international women's matches. We note that data to inform additional inferences is often available but is generally collated less systematically and distributed less widely than the data for the men's game. We anticipate that this is likely to change as the women's game increases in popularity.

3 Predictive modelling

We test the ability of three varieties of model to predict bookings counts and, for comparison purposes, goals counts. The first two models are Dixon-Coles Poisson and logistic regression models with the characteristic adjustments for intra-match dependence that effectively inflate simultaneous zero counts for both teams. The third model is a 'hurdle model' that combines the first two. More specifically, it involves modelling the non-zero status of the counts, then modelling the counts given that they are greater than zero. All the models use the same set of covariates in order to accommodate:

1. team-specific offensive and defensive effects as in the original Maher and Dixon-Coles models,
2. a home advantage effect,
3. referee-specific effects,
4. a team-disparity effect informed by pre-match bookmaker's odds,
5. 3 parametric temporal effects (linear trend and a pair of seasonal sinusoids).

All models lead to predictive distributions for goal and booking counts. A specific cumulative probability from these distributions will eventually be used to assess them in Section 3.4.

3.1 Poisson regression

The response variables for these models are the goal and booking counts. Regression coefficients are subject to L_1 penalization whose strength is calculated to minimize a cross-validated error estimate.

3.2 Logistic regression

The response variables for these models are values in $\{0,1\}$, indicating whether the counts are less than or equal to the relevant population median. As will become clear in Section 3.4, these models directly target the probability according to which they will be assessed. The model coefficients are penalized in the same way as the Poisson regression models.

3.3 Logistic/Poisson hurdle models

These models, introduced by [Cragg, 1971] and recently applied to football goals by [Owen, 2017], provide an alternative method for accommodating an over- or under-abundance of zero counts in the data. They do not include the intra-match dependence adjustment that characterizes the Dixon-Coles models, which partly serves to accommodate the same phenomenon. The hurdle models essentially describe the distribution of a count via a Bernoulli distribution for a its non-zero status and a truncated Poisson distribution for its value given that it is non-zero. For fitting these models we employ the *pscl* package for R, which contains code developed by [Zeileis et al., 2008]. This code currently does not penalize fitted coefficients.

3.4 Model comparison

Our comparisons for predictive performance are based on whether we can predict if a team's goal or booking count in a given match will fall above or below a specific threshold value, k. The idea is informed by the practical importance of such predictions for under/over-type gambles in betting markets. The specific threshold values in question are chosen to be the marginal median counts across all matches in the data set.

Our model-based predictions are informed by a training subset of the data consisting of matches during the first five of the seven seasons under consideration. The remaining n_{test} matches are held back for evaluation. Specifically, the probability assigned by our models to the outcome (a count being less than or equal to, or greater than the threshold) that did occur is computed for matches in the test set. Geometric averages of these probabilities, corresponding to exponentiated scaled log-likelihoods, are presented in Table 1. Indexing the probabilities for each match in the test set by i and denoting them $\hat{p}_i$, the scores are computed according to the formula

$$\begin{aligned} GAPS =& \text{Geometric Average Probability Score} && (1)\\ =& \left(\prod_{i=1}^{n_{\text{test}}} \hat{p}_i^{1(\text{count } i \text{ is} \leq k)}(1-\hat{p}_i)^{1(\text{count } i \text{ is} > k)}\right)^{1/n_{\text{test}}}. && (2) \end{aligned}$$

This quantity is an exponentiated average log-score for the predictions $\hat{p}_i$. The benchmark against which we suggest measuring the regression models is the GAPS achieved by a forecaster who specifies for every match the same probability, which is computed as the marginal proportion of all goals or bookings in the training set that are less than or equal to the relevant threshold k.

To reiterate, the GAPSs are (geometric) averages of probabilities assigned by models to outcomes that do occur. Good models will therefore produce high GAPSs. The nature of the under/over-outcomes has been selected so that it is relatively easy for a naive model, which allocates the same outcome probabilities for each event, to achieve a score of around 0.5.

Our results are distinctly underwhelming. For the goal counts, both Poisson and logistic regression models allocate to outcomes that do occur probabilities that are significantly, but only modestly, greater than the naive marginal method. For the bookings, the Poisson and logistic regression models allocate probabilities that are, on average, not significantly different from those of the naive marginal method. The hurdle models perform significantly worse than the naive marginal method for both goals and bookings. In each of the preceding three sentences the word 'significantly' is used in its technical sense and is informed by paired Wilcoxon signed-rank tests comparing pairs of logged probabilities from different models for the same under/over events.

4 Remarks

Summary statistics from the fitted Dixon-Coles models for bookings suggest that they are picking up on and quantifying real effects. For example, the causal effects discussed in Section 2 are all reflected in correspond-

	Marginal	Poisson reg.	Logistic reg.	Hurdle model
Goals	0.5115	0.5527	0.5519	0.4447
Bookings	0.4998	0.4995	0.4920	0.4844

Table 1: Geometric averages of probabilities assigned to outcomes in the test data.

ingly large fitted model coefficients. Despite this, the predictive skill of the models on test booking data is not significantly different from those of the most naive methods. We tentatively conclude that the model is also (over-)fitting to illusory trends in the training data that are not present in the test data. Shrinking all coefficients towards zero by a degree calibrated to minimize cross-validated prediction errors improves the models' out-of-sample predictive skill but the gain over the naive methods remains very small. We conjecture that further improvements, if possible, will require more nuanced, contextually motivated selection or shrinkage of covariate effects or a reformulation of the prediction problem that targets more predictable quantities.

A particularly promising direction for further investigation, also identified by [Kharrat et al., 2020], involves the effects of individual players on a match. Since individuals are arguably more capable of skewing booking counts than goal counts, we ought to expect them to play a greater role in predictive modelling of bookings. However, as discussed above, without very careful treatment, adding player effects is likely to exacerbate over-fitting problems.

We also anticipate opportunities for model improvement in the form of 'match importance' indices to be used as extra covariates in predictive models. These might, for example, be based on relative league position(s) or on a 'derby' indicator that acts as a proxy for reputational standing. The relevance of these indices, as anticipated by [Buraimo et al., 2010] for example, is based on the premise that teams perform differently when the consequences of their performance are higher or lower. Computing such measures of importance is likely to benefit from a combination of expert background knowledge and numerical experimentation.

Acknowledgements

The work above is motivated and informed by an undergraduate student project in collaboration with Sky Betting Gaming. Accordingly, the authors would like to acknowledge valuable contributions from Thomas Hemery to the original project. They would also like to thank the team at Sky Betting Gaming (including Donough Regan, Jon Carter, Mitch Bond, Will Cook and Fredrik Bjorkeroth) for helpful conversations and insights.

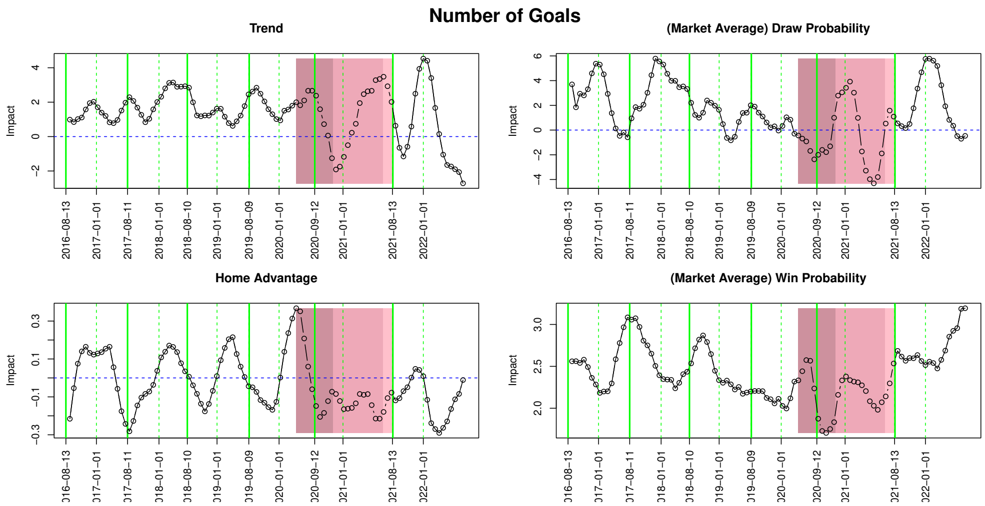


Figure 1: The estimated coefficient functions of the varying coefficient models (Section 2.1): Goals. Vertical green lines indicate: the start of a season (solid); New Year's Day (dashed). Red background indicates COVID restrictions in place (dark red: zero spectators permitted; dark pink: "tiered" restrictions in England (i.e. some grounds permitted restricted spectator numbers); light pink: all grounds permitted restricted numbers of spectators).

A Figures

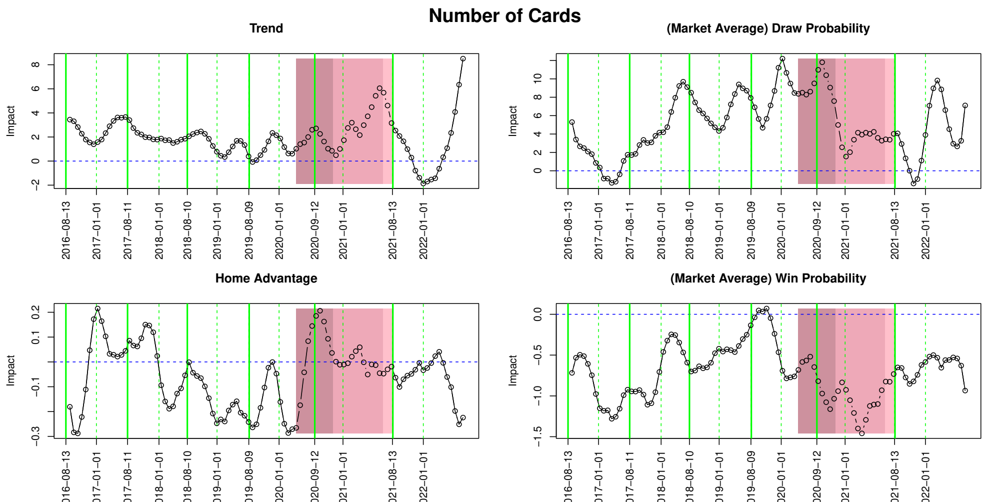


Figure 2: The estimated coefficient functions of the varying coefficient models (Section 2.1): Cards. Vertical green lines indicate: the start of a season (solid); New Year's Day (dashed). Red background indicates COVID restrictions in place (dark red: zero spectators permitted; dark pink: "tiered" restrictions in England (i.e. some grounds permitted restricted spectator numbers); light pink: all grounds permitted restricted numbers of spectators).

References

[Boshnakov et al., 2017] Boshnakov, G., Kharrat, T., and McHale, I. G. (2017). A bivariate weibull count model for forecasting association football scores. *International Journal of Forecasting*, 33(2):458–466.

[Bryson et al., 2021] Bryson, A., Dolton, P., Reade, J. J., Schreyer, D., and Singleton, C. (2021). Causal effects of an absent crowd on performances and refereeing decisions during covid-19. *Economics Letters*, 198:109664.

[Buraimo et al., 2010] Buraimo, B., Forrest, D., and Simmons, R. (2010). The 12th man?: refereeing bias in english and german soccer. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2):431–449.

[Cragg, 1971] Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: Journal of the Econometric Society*, pages 829–844.

[Dixon and Coles, 1997] Dixon, M. J. and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280.

[Endrich and Gesche, 2020] Endrich, M. and Gesche, T. (2020). Home-bias in referee decisions: Evidence from "ghost matches" during the covid19-pandemic. *Economics Letters*, 197:109621.

[Fan and Gijbels, 2018] Fan, J. and Gijbels, I. (2018). *Local polynomial modelling and its applications*. Routledge.

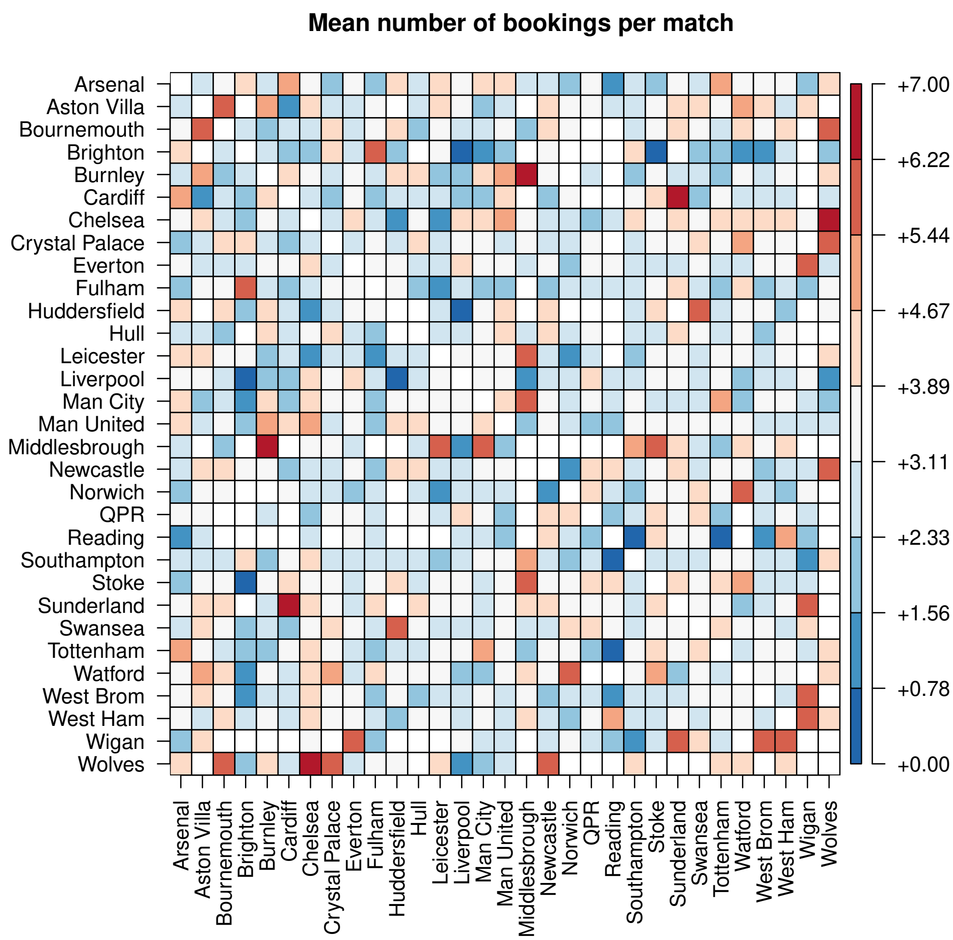
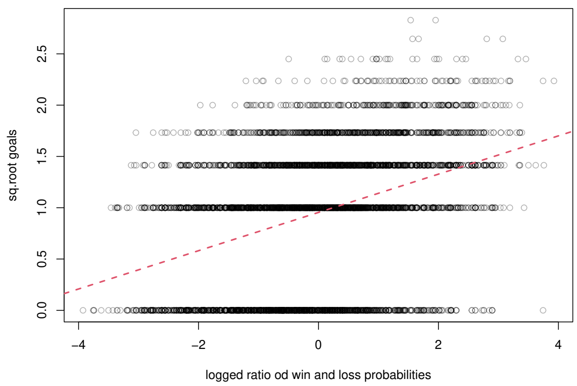
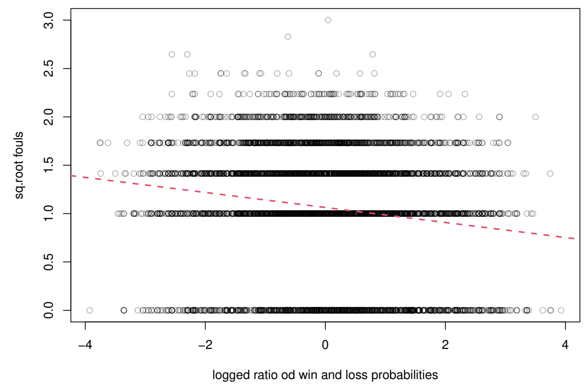


Figure 3: Average booking counts for matches between specific teams. The few pairings that do not occur in the data set are imputed with the mean booking count.

Figure 4: Per-match goal and booking counts for teams with varying win probabilities.

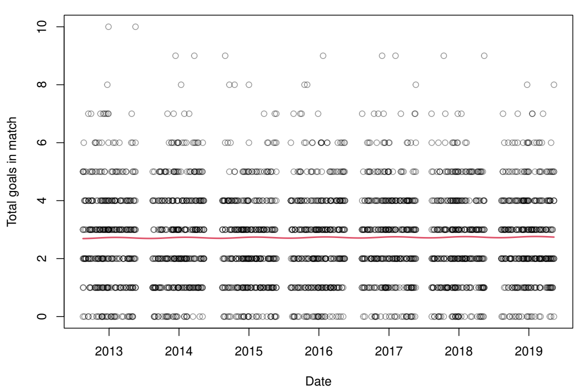


(a) Goals.

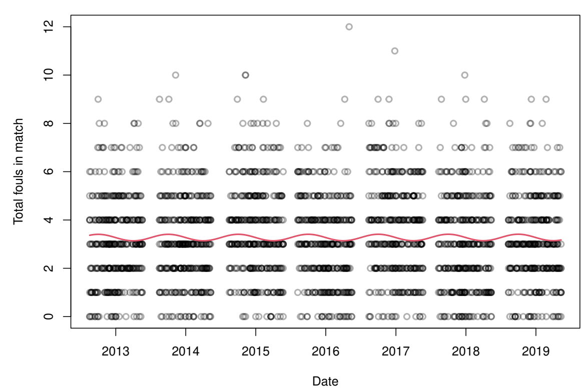


(b) Cards.

Figure 5: Per-match goal and booking counts for both teams over time.

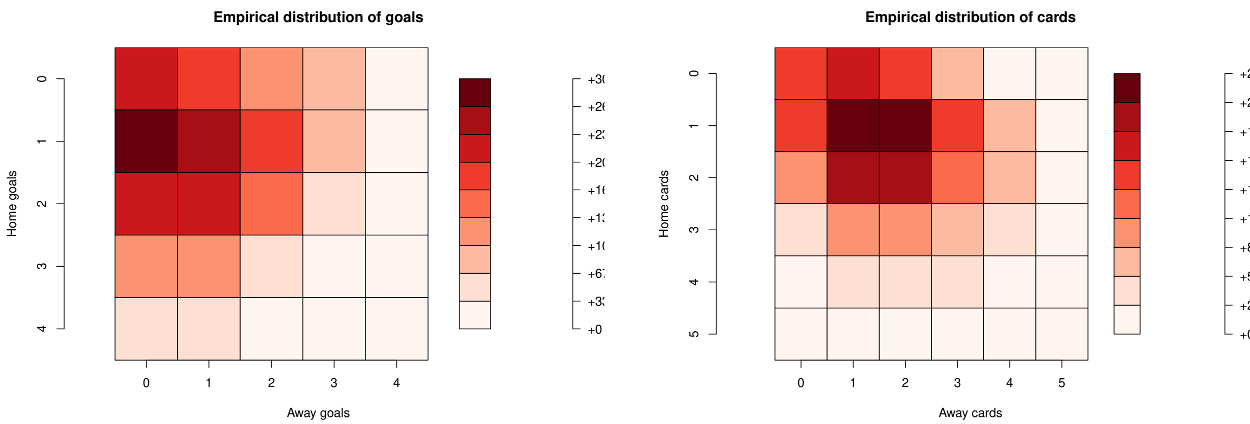


(a) Goals.



(b) Cards.

Figure 6: Counts for matches in which home/away goal and booking counts take different values.



(a) Goals. (b) Cards.

[Fischer and Haucap, 2021] Fischer, K. and Haucap, J. (2021). Does crowd support drive the home advantage in professional football? evidence from german ghost games during the covid-19 pandemic. *Journal of Sports Economics*, 22(8):982–1008.

[Hastie and Tibshirani, 1993] Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779.

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

[Kharrat et al., 2020] Kharrat, T., McHale, I. G., and Peña, J. L. (2020). Plus–minus player ratings for soccer. *European Journal of Operational Research*, 283(2):726–736.

[Koopman and Lit, 2015] 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 (Statistics in Society)*, 178(1):167–186.

[Maher, 1982] Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118.

[McCarrick et al., 2021] McCarrick, D., Bilalic, M., Neave, N., and Wolfson, S. (2021). Home advantage during the covid-19 pandemic: Analyses of european football leagues. *Psychology of sport and exercise*, 56:102013.

[Owen, 2011] Owen, A. (2011). Dynamic bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA Journal of Management Mathematics*, 22(2):99–113.

[Owen, 2017] Owen, A. (2017). The application of hurdle models to accurately model 0-0 draws in predictive models of football match outcomes. In *Proceedings of MathSport International 2017 Conference*, page 295.

[Ponzo and Scoppa, 2018] Ponzo, M. and Scoppa, V. (2018). Does the home advantage depend on crowd support? evidence from same-stadium derbies. *Journal of Sports Economics*, 19(4):562–582.

[Tilp and Thaller, 2020] Tilp, M. and Thaller, S. (2020). Covid-19 has turned home advantage into home disadvantage in the german soccer bundesliga. *Frontiers in sports and active living*, 2:593499.

[Zeileis et al., 2008] Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, 27(8).