Forecasting number of corner kicks taken in association football using compound Poisson distribution

Stan Yip[a,*], Yinghong Zou[a], Ronald Tsz Hin Hung[b], Ka Fai Cedric Yiu[a]

[a] *Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong*
[b] *Department of Economics, Hong Kong Baptist University, Hong Kong*

Abstract

This article presents a holistic compound Poisson regression model framework to forecast number of corner kicks taken in association football. Corner kick taken events are often decisive in the match outcome and inherently arrive in batch with serial clustering pattern. Providing parameter estimates with intuitive interpretation, a class of compound Poisson regression including a Bayesian implementation of geometric-Poisson distribution is introduced. With a varying shape parameter, the corner counts serial correlation between matches is handled naturally within the Bayesian model. In this study, information elicited from cross-market betting odds was used to improve the model predictability. Margin application methods to adjust market inefficiency in raw odds is also discussed.

Keywords: Bayesian hierarchical models, compound Poisson distribution, corner kick, geometric-Poisson distribution, football, negative binomial distribution

1. Introduction

Association football is the most popular sport in the world with billions of followers. The sport drives significant contribution to the economy from the

*The Hong Kong Polytechnic University, Hung Hom
Email addresses: gstanyip@connect.polyu.hk (Stan Yip), yinghong.zou@connect.polyu.hk (Yinghong Zou), 19481217@life.hkbu.edu.hk (Ronald Tsz Hin Hung), cedric.yiu@polyu.edu.hk (Ka Fai Cedric Yiu)

global television revenues (Roberts et al., 2016), tourism income from extensive travels (Allmers & Maennig, 2009) and sponsorship from betting operators (Deutscher et al., 2019). Thanks to the rapid development of broadcasting and big data technology, a considerable amount of information from different sources allows forecasting the sport outcomes more accessible.

Forecasting the likelihood of football match outcomes is a timely research topic within the statistics and econometrics community in the past decades. There are at least 40,000 articles in JSTOR and 3,700 entries in Econ Lit that refer to sports events modelling and related factors (Stekler et al., 2010). Quantifying uncertainty of the association football outcomes first attracts the attention from statisticians. The seminal paper by Maher (1982) utilises a bivariate Poisson distribution to capture teams' inherent attacking and defensive strengths factors. Dixon & Coles (1997) elaborate Maher's work to exploit the betting market inefficiency. Koopman & Lit (2019) provide a dynamic forecasting model with time-varying coefficients that generates a significant positive return over the bookmaker's odds. Angelini & De Angelis (2017) use a Poisson autoregression with exogenous covariates (PARX) developed by econometricians to model the football matches outcomes. Ley et al. (2019) provide 10 different methods to rank football teams based on their historical performance. Baboota & Kaur (2019) adopt a gradient boosting to determine English Premier League results. A binary dynamic time-series model is introduced in Mattera (2021) for forecasting two-way football outcomes such as red cards occurrence, total goals over/under (O/U) and goal/no goal events. Hierarchical Bayes framework is often used in modelling football which can naturally handle the relations between unobserved variables and componentise a model into different parts. Rue & Salvesen (2000) perform a Bayesian predictive and retrospective study in the evolution of football team strengths by incorporating dynamic attacking and defensive factors. Baio & Blangiardo (2010) develop a Bayesian model to formulate home advantage and latent team effects. In Karlis & Ntzoufras (2009), goal supremacy is adequately described with a Bayesian Skellam regression model.

Several attempts have been made to extract information from the pre-match betting market and expert pundits. Clarke et al. (2017) study the forecasting ability after adjusting bookmaker's odds to allow for overround. Cain et al. (2000) study a high odds overvaluation phenomenon termed favourite-longshot bias in the UK football betting market. Deschamps & Gergaud (2007) find a negative favourite-longshot bias on draw odds in the home-away-draw (HAD) market. Pundits and experts information are often considered by the general public. However, Forrest et al. (2005) discover that subjective forecasting is inferior to model-based forecast. Forecasts from statistical model often contain the information that the betting market not taken into account (Pope & Peel, 1989; Goddard & Asimakopoulos, 2004; Dixon & Pope, 2004). Sung & Johnson (2007) illustrate that combining probabilities from statistical models and instant market odds with a simple statistical model improves the accuracy probabilistic forecast.

Apart from the simple HAD prediction, corners kicks are also of both sport analytics and betting interest. Corner kicks are considered to be a reasonable goal-scoring opportunity. Pulling et al. (2013) study the tactical behaviour when defending corner kicks. Casal et al. (2015) provide a thorough analysis on the factors to increase the likelihood of shot-on-goal from corner kicks. Fitt et al. (2006) give a valuation framework on the most commonly traded football spread bets including the total number of corners betting option. Corner counts market is one of the most popular sidebet markets offered by sports bookmakers. A more advanced understanding on the statistical properties of the corner counts distribution helps the bookmakers to manage risk through setting the odds more adequately. Number of corners taken is also an important match statistics that never available prior to the match. Along with other match statistics such as number of shots and number of interceptions throughout the game, if these match statistics can be forecasted prior to the match start, it contributes more information for predicting match outcomes (Wheatcroft, 2021).

Data provenance is a crucial component of an accurate forecasting system. Data collected in football has lately become a concern due to novel sensor modalities, with potentially high commercial and research interest (Stein et al., 2017). Hutchins (2016) discuss that the infrastructure and resources necessary to generate real-time data are contributing to rising disparities between top 'data-rich' sports and comparatively poor 'data-poor' sports.

The remainder of this article is organised as follows. Section 2 discusses the data and the information extraction issues. Section 3 introduces the family of compound Poisson distributions and model development. Section 4 implements the algorithm and performs betting simulation. A detailed discussion of the model implications and future work is provided in Section 5.

2. Data acquisition and information extraction

2.1. Data acquisition

Football match information were collected from the Hong Kong Jockey Club (HKJC) website (www.hkjc.com) which includes 20,190 matches with corner over-under market offered from 1st August 2016 to 10 June 2021 across 90 leagues and cup games. Pre-match HAD, total goals scored over-under and total corner over-under market odds are recorded. On the other hand, post-match information including scorelines and number of corners are also obtained. The HKJC set the market odds at least a day before the match starts. All the prices collected are the odds just before the match start.

The integrated dataset used in this study (Table 1) also consists of an alternative data source with extra match information from seven elite leagues (i.e.: English Premier League, Scottish Premier League, German Bundesliga, Italian Serie A, Spanish La Liga, French Ligue 1 and Dutch Eredivisie) including number of shots, number of shot-on-goals, number of fouls and number of cards given and number of the corner kicks taken by each team. These can be downloaded from `www.football-data.co.uk`. A mild correlation (9.68%/17.37%) between the number of shots/shot-on-goals and the number of corner counts is found.

In order to estimate parameters more accurately and construct useful covariates from bookmaker odds in the model, utilising a proper overdispersed distribution due to clustered counts (not due to excess of zeros and attacking/defensive tactics) is crucial in the information extraction process which will be discussed in the following subsections.

Table 1: Top 15 competitions with total corner over-under market odds offered by HKJC from 1st August 2016 to 10 June 2021

Competitions	Number of corner markets offered	Extra match statistics available
Italian Serie A	1863	Yes
English Premier League	1849	Yes
Spanish La Liga	1784	Yes
USA Major League Soccer	1766	No
German 2. Bundesliga	1517	No
German Bundesliga	1503	Yes
Japanese J-League	1494	No
French Ligue 1	1066	Yes
Europa Cup	942	No
Australian A-League	722	No
European Champions League	675	No
Russian Premier League	475	No
Japanese J-League 2	305	No
Asian Champions League	300	No
Korean K-league	282	No
Other 75 leagues and competitions	3647	No
Total	**20190**	

2.2. Overdispersed behaviour of the total goals scored and corner kicks taken

Overdispersion in Poisson models occurs when the response variance exceeds the mean (Hilbe, 2011). Apart from attributing to the excess of zero counts in the statistical distribution (Lambert, 1992) or the dependency to another variable (Kocherlakota, 1988; Kocherlakota & Kocherlakota, 2001; i Morata, 2009;

Sengupta et al., 2016), it also arises from serial correlation between events or counts of event are clustered that violates the distributional assumption of Poisson random variable. Weak overdispersion occurs in association football total goals scored (Rodríguez-Avi & Olmo-Jiménez, 2017) whilst this statistical property can also be found in many other sports such as basketball, baseball, hockey (Higgs & Stavness, 2021) in varying degrees. Unlike some American sports that the scores depending on numbers of factor brought out extra variation (Pollard, 1973), the nature of the overdispersion in football total goals scored is inherently different. Maher (1982) asserts that Poisson distribution is a better model than the overdispersed negative binomial distribution due to its close-to-one variance-to-mean ratio. Dixon & Coles (1997) introduce a low scoring modification on a bivariate Poisson to handle teams' changing tactics when two teams are one goal apart. Contrast to the overdispersed behaviour of the total goals scored, modelling overdispersed corner kicks taken counts is a relatively under-explored area that no known literature has ever discussed it.

In the HKJC dataset recorded matches from 2016-2021, the sample variance-to-mean ratio for the number of total goals scored has a value of 1.0396 which agrees with the suggestion in Maher (1982) that the scoreline exhibits a Poisson distribution. On the other hand, the ratio for the number of corner kicks taken is 1.1856 which indicates a fairly significant overdispersion. The corner kick clusters are arisen from "parent" corners or offensive duels which produces one or more "offsprings" through multiple clearances off the goal line that last touched a player of the defending team (for schematic illustration, see Fig. 1). A cluster of corner kicks is understood as a few corners occurring in a short time span. In other words, corners from open play might be a Poisson, but the number of total corners kicks taken rather follows an overdispersed distribution. Failing to recognise this type of temporal clustering leads to an underestimation of number of corner kicks taken.

2.3. Bookmaker odds as forecasters and margin removal methods

Betting odds issued by bookmakers reflect implicit probabilities about sporting outcomes (Stekler et al., 2010; Leitner et al., 2010; Kovalchik, 2016; Clarke et al., 2017). Some empirical evidences show that the implied probabilities from betting odds provide moderate accuracy and outperform tipsters prediction (Spann & Skiera, 2009; Stekler et al., 2010; Reade, 2014). The reciprocals of the bookmaker odds are understood as the bookmaker's probabilistic belief. These values sum to more than one. The sum of these probabilities π is known as the margin of the book which determines how large the edge toward to the bookmaker. Its inverse $1/\pi$ is the payout rate which indicates the expected value from the bettor point of view. The method to normalise this probabilistic belief is called margin removal method.

It is found that the market often overvalues low probability events. Cain et al. (2000) examine that there is a favourite-longshot bias in the football fixed-odds betting market. One explanation is that there are well-informed insiders in the market and the bookmakers maximises their profits through applying less percentage of margin to a particular selection in the book (Shin, 1992, 1993). Another popular interpretation is that the bias come from amateur bettors' risk-loving behaviour who overvalue longshot selections.

2.3.1. Multiplicative margin removal

The multiplicative margin removal method, also known as basic method, normalise the inverse odds proportionally and divided by its booksum π. Most studies has widely adopt this method due to simplicity (Štrumbelj, 2014). Let $\mathbf{o} = (o_1, o_2, \ldots, o_l)$ be the quoted decimal odds for a football match with $l \geq 2$ possible outcomes and the inverse odds are $\mathbf{\Pi} = (\pi_1, \pi_2, \ldots, \pi_l)$. Four popular margin removal methods are considered in this article. The implied probability derived from the multiplicative margin removal method is:

$$p_i = \pi_i/\pi,$$

where $\pi = \sum_i^l \pi_i$ is the booksum.

2.3.2. Odds ratio margin removal

Implemented in the R package `implied` (Lindstrøm, 2021), the odds ratio margin removal method fixes the ratio between the implied probability and the raw probability with margin π on all possible outcomes. The odds ratio (OR) is defined as:

$$\mathrm{OR} = \frac{p_i(1-\pi_i)}{\pi_i(1-p_i)},$$

where OR is selected so that $\sum_i^l p_i = 1$. The implied probability can be calculated with the following equation:

$$p_i = \frac{\pi_i}{\mathrm{OR} + \pi_i - (\mathrm{OR} \times \pi_i)}.$$

In spite of its mathematical elegance and the ability to capture the favourite-longshot bias, this method is relatively unknown to academic community.

2.3.3. Shin's method

Shin's method handles the information asymmetry for bookmakers as less-informed price-setters who face a group of insiders with superior information. Štrumbelj (2014) finds that the implied probabilities derived from the betting odds using Shin's method are more accurate than those derived from other simple methods. The formula of taking out margin from the raw probability is given in Jullien & Salanié (1994):

$$p_i = \frac{\sqrt{z^2 + 4\,(1-z)\,\frac{\pi_i^2}{\pi}} - z}{2\,(1-z)},$$

where z can be numerically estimated by an iterative method starting at $z_0 = 0$:

$$z = \frac{\sum_{i=1}^l \sqrt{z^2 + 4(1-z)\frac{\pi_i^2}{\pi}} - 2}{l-2}.$$

In the special case of $l = 2$, Clarke et al. (2017) show that it is equivalence to a rarely used additive method that does not guarantee a positive probability for $l > 2$ and has a tractable analytic solution:

$$z = \frac{(\pi_+ - 1)(\pi_-^2 - \pi_+)}{\pi_+(\pi_-^2 - 1)},$$

where $\pi_+ = \pi_1 + \pi_2$ and $\pi_- = \pi_1 - \pi_2$.

2.3.4. Power margin removal

The power method is originally described in Vovk & Zhdanov (2009) and further developed in Clarke (2016) and Clarke et al. (2017). The power parameter k scales the inverse odds to the implied probabilities:

$$p_i = \pi_i^{1/k},$$

where the parameter k is chosen that $\sum_i^l p_i = 1$. The power method outperforms other marginal removal methods in the ATP men's singles tennis two-way market. It is thought that the method is commonly used by bookmakers through empirical evidence (Clarke et al., 2017).

2.4. Cross-market information

Market odds from HAD and O/U provides information of how the market view the intensities of goals in a football match. It is generally understood that the expected number of total goals scored and the relative strength between home and away teams play a strong role in explaining number of corners. In order to incorporate such an information, implied expected number of goal scores by both teams can be extracted from the double independent Poisson model proposed by (Maher, 1982):

$$f_{X_1,X_2}(x_1, x_2|\lambda_1, \lambda_2) = \frac{\lambda_1^{x_1} e^{-\lambda_1}}{x_1!} \times \frac{\lambda_2^{x_2} e^{-\lambda_2}}{x_2!}.$$

The implied expected total goals scored $\mathbf{TG}_i$ and home team's relative goal supremacy $\mathbf{SUP}_i$ of match i can be reparameterised from the implied values of $\hat{\lambda}_1$ and $\hat{\lambda}_2$:

$$\begin{aligned} \mathbf{TG}_i &= \hat{\lambda}_1 + \hat{\lambda}_2, \\ \mathbf{SUP}_i &= \hat{\lambda}_1 - \hat{\lambda}_2. \end{aligned}$$

A variant of the limited memory Broyden–Fletcher–Goldfarb–Shanno algorithm (L-BFGS-B; Bryd et al., 1995) is used to optimise the square loss function of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ for each match with implied probabilities from HAD and total goals scored O/U odds after removing bias. The implied parameters $\mathbf{TG}_i$ and $\mathbf{SUP}_i$ from the market odds are used to reflect how the market account for the match expected total goals scored and goal supremacy between two teams. Therefore, the L-BFGS-B algorithm is used to minimise the following loss function:

$$(p_H - p'_H)^2 + (p_D - p'_D)^2 + (p_L - p'_L)^2 ,$$

where p_H, p_D, p_L are the implied probabilities from the match HAD home, HAD draw, O/U under odds respectively. The model HAD home, HAD draw and O/U under probabilities p'_H, p'_D, p'_L are derived from the Equation (2.4).

2.5. Competition-dependent factor and team-level historical records

Over 90 leagues or cup competitions being offered the total number of corners taken O/U market by the HKJC during the period 2016-2021. Table 2 shows the average corner counts in some competitions are higher and inherently more volatile with a higher variance-to-mean ratio which is estimated by maximising the log-likelihood of a negative binomial generalised linear model. Target encoding techniques transform categorical variables to a numerical covariate by a statistic computed using the predictors. The following equation generalises the target encoding technique proposed by Micci-Barreca (2001) to replace the categorical variable by a numerical value $\hat{x}_k^i$ of match i that belong to competition k:

$$\hat{x}_k^i = \frac{n_k^i \cdot \hat{\theta_k} + \hat{\theta} m}{n_k^i + m},$$

where n_k^i is the number of matches in competition k, $\hat{\theta}$ is the statistic of interest such as logarithm of average corner counts and maximum likelihood estimator of the shape parameter, $\hat{\theta_k}$ is the statistic of interest on competition k and m is a hyperparameter that set to a reasonable sample size that makes an estimator. Although target leakage problem is a common cause of overfitting in many machine learning algorithms (Zhang et al., 2013; Prokhorenkova et al., 2017), this problem is negligible in this study since the covariates encoded $\hat{x}_k^i$ via the target encoding method is obtained from a pre-training period of 2016-2018 only which is not a part of the data in the fitted model.

Table 2: Average number of corner kicks taken and estimated variance-to-mean ratio D by competition using the `glm.nb` function in R for the leagues with sample size greater than 500 during the study period of 2016-2021

Competitions	Average number of corners	$\hat{D}$
Australian A-League	11.04	1.20
Italian Serie A	10.43	1.25
English Premier League	10.38	1.14
USA Major League Soccer	10.18	1.12
German 2. Bundesliga	10.00	1.16
UEFA Champions League	9.73	1.30
Japanese J-League	9.70	1.11
French Ligue 1	9.65	1.13
German Bundesliga	9.58	1.19
UEFA Europa League	9.40	1.23
Spanish La Liga	9.39	1.10

3. Model development

3.1. Discrete compound Poisson distribution

The discrete compound Poisson (DCP) distribution is a broad family of distributions widely used in modelling contagious disease attacks, repeated accidents (Greenwood & Yule, 1920; Feller, 1943), bacteria spawn (Neyman, 1939) and batch arrival counts (Adelson, 1966). It is a reasonable model to consider a Poisson process for each cluster of counts and it also captures the intensity varied from cluster to cluster. With Poisson distribution as its special case, it enjoys a great flexibility for overdispersed count data. Suppose the number of clusters N whose distribution is given as:

$$N \sim \mathbf{Poisson}(\lambda).$$

A DCP random variable Y is defined by the form:

$$Y = \sum_{i=1}^{N} X_i, \tag{1}$$

where $X_1, X_2, \ldots, X_N$ are independent identically distributed discrete random variables. A random variable Y is said to be a DCP random variable (Wimmer & Altmann, 1996) when its probability generating function (pgf) $G(s)$ has a form:

$$G_Y(s) = \exp\left\{\sum_{i=1}^{\infty} \alpha_i \lambda \left(s^i - 1\right)\right\}, \quad s \in \mathbb{R}, \tag{2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots) \in \mathbb{R}^\infty$ is the probability set describing a discrete random variable X_i with the sum $\sum_{i=1}^{\infty} \alpha_i = 1$. Thus, the DCP random variable Y is denoted by:

$$Y \sim \text{DCP}(\lambda, \boldsymbol{\alpha}).$$

The pgf in the Equation (2) is used to study the characteristics of DCP distribution that includes many widely-used models such as Poisson, Hermite (Kemp & Kemp, 1965), Neyman Type A (Neyman, 1939), negative binomial and geometric-Poisson (Pólya, 1930). The following subsections provide the key characteristics of a few popular DCP distributions including Poisson, negative binomial and geometric-Poisson distributions. These three distributions refer to three plausible cases of “no clustering”, “logarithmic distributed clustering”, “geometric distributed clustering” respectively.

3.1.1. Poisson distribution

Poisson distribution is the trivial case of DCP distribution where the random variables X_i in the Equation (1) is a degenerated distribution with probability mass function (pmf):

$$p_{X_i}(x_i) = \begin{cases} 1, & x_i = 1 \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, the parameter $\boldsymbol{\alpha}$ is set to a unit vector of $\alpha_1 = 1$. Thus, the pgf of the Poisson random variable Y is

$$G_Y(s) = \exp\left(\lambda(s-1)\right).$$

3.1.2. Negative binomial distribution

Negative binomial distribution is proven to be a DCP distribution with a logarithmically distributed counts of each cluster (Quenouille, 1949). The pmf of the logarithmic distribution is given by:

$$p_X(x) = \frac{-1}{\log(1-p)}\frac{p^x}{x}, \quad x = 1, 2, \dots,$$

where $0 < p < 1$. The pgf of the negative binomial distribution can be derived by considering a composition function of the pgf of Poisson distribution and the pgf of logarithmic distribution (Johnson et al., 2005). The composition function of the $G_N(s) = \exp(\lambda(s-1))$ and $G_X(s) = \log(1-ps)/\log(1-p)$ is given by:

$$G_Y(s) = G_N(G_X(s)) = \left(\frac{p}{1-(1-p)s}\right)^{-\frac{\lambda}{\log p}}.$$

A parameterisation used by Gelman et al. (1995) is adopted in this paper with a pmf:

$$p_Y(y) = \binom{y+\kappa-1}{y}\left(\frac{\lambda}{\lambda+\kappa}\right)^y\left(\frac{\kappa}{\lambda+\kappa}\right)^\kappa.$$

The expectation, variance and variance-to-mean ratio are $E[Y] = \lambda$, $Var[Y] = \lambda + \lambda^2/\kappa$ and $D = 1 + \lambda/\kappa$ respectively. The pgf is given by:

$$\left(\frac{\kappa}{\lambda(1-s)+\kappa}\right)^\kappa.$$

The distribution also includes Poisson model as a limiting case, i.e.:

$$\lim_{\kappa\to\infty} G_Y(s) = \lim_{\kappa\to\infty}\left(1 + \frac{\lambda(1-s)}{\kappa}\right)^{-\kappa} = \exp\{\lambda(s-1)\}.$$

3.1.3. Geometric-Poisson distribution

An alternative approach to account for clustered counts in the Poisson model is to place a geometric distribution on each cluster count. A geometric-Poisson random variable is defined on a DCP random variable with a zero-truncated geometric random variable within clusters. Özel & Inal (2010) derive the mass function as:

$$p_Y(y) = \begin{cases} e^{-\lambda}, & y = 0 \\ e^{-\lambda}\sum_{k=1}^{y}\frac{\lambda^k}{k!}\binom{y-1}{k-1}\theta^k(1-\theta)^{y-k}, & y = 1, 2, 3, \dots \end{cases}$$

where $\lambda > 0$, $0 < \theta < 1$. The expectation, variance and variance-to-mean ratio are $E[Y] = \lambda/\theta$, $Var[Y] = \lambda(2-\theta)/\theta^2$ and $D = (2-\theta)/\theta$ respectively. The pgf is given by

$$G_Y(s) = \exp\left\{-\lambda\left[1 - \frac{(1-\theta)s}{1-\theta s}\right]\right\},$$

which is clearly a pgf of the DCP distribution in the Equation (2).

3.2. Compound Poisson regression

Consider a regression model with the following covariates:

$$\begin{aligned}\log\lambda_i \quad = \quad & \beta_0 + \beta_1\log(\text{TG}_i) + \beta_2\log(|\text{SUP}|_i + 0.01) + \beta_3\text{TCTarget}_i + \\ & \beta_4\log(\text{HomeAvg3}_i + 0.01) + \beta_5\log(\text{AwayAvg3}_i + 0.01) + \\ & \beta_6\log(\text{HomeShotOnGoalAvg3}_i + 0.01) + \\ & \beta_7\log(\text{AwayShotOnGoalAvg3}_i + 0.01),\end{aligned}$$

where HomeAvg3_i and AwayAvg3_i are the home and away team's average counts in their previous 3 matches imputed by the overall league average if lack of the previous match information. The covariate TCTarget_i generated from the target encoding method in the Equation (2.5) handles the adjusted mean corner count. $\text{HomeShotOnGoalAvg3}_i$ and $\text{AwayShotonGoalAvg3}_i$ are the home and away team's number of shots in their previous 3 matches. If the team has not competed previously, the counts are imputed by the league-dependent value encoded in the Equation (2.5) and by the grand mean if no league information is available. Adding a small positive constant to a covariate is a popular approach to tackle the log transformation of values that include zero (Bellego et al., 2021).

3.3. Regression on the shape parameters

On the grounds that the flexibility of Bayesian hierarchical model, an additional regression structure can be added to the negative binomial and geometric-Poisson shape parameters κ_i and θ_i. The regression components on the shape parameter are defined as:

$$\begin{aligned}\text{negative binomial regression:} \quad \log(\kappa_i) \quad &= \quad \alpha_0 + \alpha_1\log(|\text{SUP}_i| + 0.01) \\ \text{geometric-Poisson regression:} \quad \text{logit}(\theta_i) \quad &= \quad \alpha_0 + \alpha_1\log(|\text{SUP}_i| + 0.01).\end{aligned}$$

Apart from using the implied goal supremacy considered above as regression covariate, target encoding method can be incorporated in the shape component. The maximum likelihood estimates of the shape parameters for each competition according to the target encoding method in Equation (2.5) can be used as the covariate. For estimating the dispersion parameter κ of negative binomial distribution, Levin & Reeds (1977) prove the existence of unique solution provided that $s^2 > \bar{y}$. The same treatment is also applied to the geometric-Poisson model. However, some of the league-dependent shape parameter estimates give large standard errors. The covariate $\log(|\text{SUP}_i| + 0.01)$ is a reasonable choice for capturing the heterogeneity of corner counts between matches.

4. Model performance, diagnostics and betting simulation

The existence of favourite-longshot bias in the corner O/U, HAD and number of total goals scored O/U markets are examined by a negative mean logarithmic scoring rule (Gneiting & Raftery, 2007). The R package `implied` (Lindstrøm, 2021) offers an implementation of popular margin removal methods to calculate implied probabilities of betting odds namely multiplicative method, odds ratio method, power method and Shin method (Shin, 1993; Clarke et al., 2017). Except the multiplicative method allocates margin proportionally to the odds, all other methods penalise longshot and apply less margin on favourite selections. The corner O/U bet-type is defined as the total number of corners taken being higher or lower than the number specified. Table 3 shows that the multiplicative method provides optimal predictability on corner O/U and the number of total goals scored O/U markets whilst Shin's method offers the best prediction by discounting longshot probabilities. Therefore, the Shin method and the multiplicative method are used for handling **TG** and **SUP** covariates.

Five candidate models are considered here, namely naïve Poisson, negative binomial and geometric-Poisson regression models; negative binomial model and geometric Poisson models with varying shape parameters. The result of the preliminary fitted models shows that only TG, TCTarget, HomeAvg3 and

Table 3: Performance comparison of margin removal methods

	Negative mean logarithmic scoring rule		
Method	Corner O/U	HAD	Total goals O/U
Multiplicative	**14187.07**	19618.98	**14465.44**
Odds ratio	14209.67	19614.28	14522.77
Power	14219.82	19633.16	14548.68
Shin	14207.74	**19612.15**	14517.71

AwayAvg3 are significantly associated with the number of the total corner kicks taken. Although non-significant covariates SUP, HomeShotOnGoalAvg3 and AwayShotOnGoalAvg3 with 95% credible interval of covariate estimates covering zero do not provide a good insight to the mean value of corner count. They are able to capture the variation of the distribution's higher moment hence the predictability. For example, a high absolute value of goal supremacy can be translated to a more unpredictable corner count. The models are fitted by Markov chain Monte Carlo (MCMC) algorithms via No U-turn sampler (Hoffman et al., 2014) implemented in Stan language (Carpenter et al., 2017) with 4000 iterations of each. The overall fit of each model is summarised in Table 4. To check whether all candidate models are well-specified, the Pareto K is examined that all estimates are less than 0.5. It shows all fitted models have high reliability and exhibit convergence. Leave-one-out expected log pointwise predictive density (elpd_loo) is used for model comparison (Vehtari et al., 2017). The geometric-Poisson regression model has the highest elpd_loo which fits the data better than other models. However, an additional regression component on the shape parameter does not improve the goodness-of-fit of the geometric Poisson regression model since it has an even smaller elpd_loo whilst the negative binomial regression model with varying shape parameter has a significantly higher elpd_loo value (the se_diff between two models is 5.7 which is not shown in the Table 4 than its fixed shape parameter counterpart.

Table 5 shows that implied number of total goals scored has a positive effect

Table 4: Model performance by the expected log-pointwise predictive density (elpd_loo) and LOO information criterion (looic) of all candidate models, where p_loo is the estimated effective number of parameters of the model, se_diff is the standard error of the difference in the elpd estimate to the one for geometric-Poisson model with varying shape parameter

Model	elpd_loo (criterion)	p_loo (penalty)	se_diff
(A) Geometric-Poisson	−30856.4	10.0	0.0
(B) Geometric-Poisson + shape reg	−30856.8	11.5	1.4
(C) Negative binomial + shape reg	−30858.4	11.6	1.7
(D) Negative binomial	−30871.4	9.1	5.3
(E) Poisson	−30877.2	10.1	6.9

on the number of corner kicks. For each of the expected goal, the expected corner count is increased by 16.70% for the model A. Previous corner kicks taken and conceded by the same home and away teams also play a crucial role. The league average corner counts exerts a negative effect on the number of corner kicks taken. It can be seen as an adjustment to a stronger effect of previous corner kicks taken and conceded. The median of the posterior distribution of model A's shape parameters θ is 0.9577. It indicates that a single corner induces additional 0.0423 corners and also implies that the variance-to-mean ratio has an estimate of 1.0883.

The betting simulation is performed for 2057 matches offered by the HKJC in 2021 from 1st January to 10th June in 2021 (Figure 2). Although average overall margin taken by the operator is 7.61% (payout rate 92.38%). However, the payout rate on the "over" selection is 85.71% whilst the "under" selection one is 98.24%. It indicates that, empirically, the market has a bias in favour of the "under" selection.

A bet of 100 dollars is placed on each of the selection when the model expected value is positive. With the five candidate models considered in Table 6 along with a blind betting strategy on all "under" outcomes, cumulative return is calculated for the simulated period. Sharpe ratio (Sharpe, 1966) is calculated from the ratio of aggregated daily profit and its sample standard deviation an-

Table 5: Parameter estimates of all candidate models A-E

Parameter	Model		
	(A) Geometric-Poisson	(B) GP + shape	(C) NB + shape
β_0	$0.711(0.0219, 1.4154)$	$0.788(0.0938, 1.4654)$	$0.7782(0.0562, 1.5142)$
β_1	$0.1545(0.1046, 0.2069)$	$0.1688(0.1216, 0.2231)$	$0.1573(0.099, 0.2109)$
β_2	$-0.001(-0.007, 0.0049)$	$-0.0018(-0.0072, 0.0037)$	$-0.001(-0.0068, 0.0047)$
β_3	$-0.6168(-0.9374, -0.2813)$	$-0.6588(-1.0025, -0.3218)$	$-0.6206(-0.9676, -0.2577)$
β_4	$0.5957(0.53, 0.6625)$	$0.6092(0.5436, 0.6725)$	$0.5957(0.5326, 0.6628)$
β_5	$0.6052(0.5405, 0.6652)$	$0.609(0.5417, 0.6754)$	$0.5996(0.537, 0.6659)$
β_6	$-0.0024(-0.0084, 0.0033)$	$-0.0029(-0.0086, 0.0033)$	$-0.0022(-0.0084, 0.0037)$
β_7	$-0.0019(-0.0088, 0.0043)$	$-0.0025(-0.0088, 0.0042)$	$-0.002(-0.009, 0.0039)$
θ	$0.9577(0.9441, 0.9705)$		
α_0		$3.7470(2.9766, 4.8491)$	$4.7198(4.3913, 5.1212)$
α_1		$-0.1978(-0.7523, 0.1141)$	$36.3012(-12.0218, 91.3719)$
	Model		
	(D) Negative binomial	(E) Poisson	
β_0	$0.7575(0.0924, 1.4591)$	$0.7438(0.0774, 1.4203)$	
β_1	$0.1559(0.1045, 0.2028)$	$0.156(0.1032, 0.2041)$	
β_2	$-0.001(-0.0066, 0.0049)$	$-0.0009(-0.0061, 0.0045)$	
β_3	$-0.607(-0.9541, -0.2698)$	$-0.6048(-0.9345, -0.2794)$	
β_4	$0.5931(0.5273, 0.6607)$	$0.5959(0.532, 0.6686)$	
β_5	$0.5974(0.5343, 0.6664)$	$0.6043(0.5434, 0.6659)$	
β_6	$-0.0024(-0.0082, 0.0043)$	$-0.0023(-0.0081, 0.004)$	
β_7	$-0.0019(-0.0089, 0.0044)$	$-0.0021(-0.0088, 0.0039)$	
κ	$59.8998(52.2233, 68.2215)$		

nualised by the square root of the total 364 trade days (Sharpe, 1966; Lo, 2002). The negative binomial regression model with match varying shape parameter outperforms other candidates and gives the best final profit of $6578 and Sharpe ratio 3.065. The Table (5) shows that the location level regression parameter estimates ($\beta_0 - \beta_7$) of the negative binomial model with match varying parameter (model C) are almost identical to those in model (A). The estimate of the parameter α_1 indicates that the variance-to-mean ratio D has a weak but notable relationship with the magnitude of the implied goal supremacy $\mathbf{SUP}_i$.

Table 6: Out-of-sample betting simulation for the period between 1st Jan 2021 to 10th June 2021 ranked by its Sharpe ratio

Model	# of bets	profit \$	profit %	Sharpe ratio
(C) Negative binomial + shape reg	1137	6578	5.785%	3.065
(B) Geometric-Poisson + shape reg	1125	6256	5.561%	2.922
(A) Geometric-Poisson	1123	5308	4.727%	2.427
(D) Negative binomial	1085	4805	4.429%	2.263
(E) Poisson	1156	4283	3.705%	1.519
Blind bet on under	2056	−3530	-1.717%	−1.267

5. Discussion

This article presents a Bayesian overdispersed Poisson regression framework to forecast future corner counts leading to a profitable forecasting model. The model incorporates multiple sources of match and market information including previous team's corner counts, cross-market information and competition-level statistics. This work could be also extended to accommodate underdispersed count through the Convey-Maxwell-Poisson (CMP) distribution (Conway & Maxwell, 1962). The probability mass function is given by:

$$p_Y(y) = \frac{\lambda^y}{(y!)^\nu} \frac{1}{Z(\lambda, \nu)},$$

where $Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^i}{(j!)^\nu}$. When $\nu = 1$, the CMP distribution becomes Poisson distribution; when $\nu = 0$, it becomes geometric distribution. A recent result in Geng & Xia (2022) shows that a CMP random variable is infinitely divisible if it is Poisson or geometric. This theoretical result corroborates that the CMP distribution does not fall into the DCP distribution family. Exploiting the property of exponential family, a generalised linear model (GLM) for CMP distribution with two link functions on the location and shape parameters can be applied (e.g.: Guikema & Goffelt 2008), the interpretation is different from the framework introduced in this paper.

Our study shows that the favourite long-shot bias in the HAD odds is captured adequately by the Shin's method. Except the HAD market, the implied probability of the total goals scored and corner O/U markets can be easily calculated through handling the margin weights proportional to the original odds with the multiplicative method. The market efficiency of the corner O/U markets is also biased in favour of the "under" selection. The asymmetry toward the "under" selection is understood as the market overvalues the "over" selection. The margin removal methods utilised in this study is unable to capture this kind of asymmetry. Along with the favourite-longshot bias studied previously, the O/U asymmetry is an area one should work on.

Our compound Poisson framework allows a complex model structure with multiple regression equations in the same model. With the application of MCMC algorithms in Stan language, the model development process can be seen as iterative and continuous under the hierarchical Bayesian modelling paradigm. Its flexibility is already proven in other research areas such as epidemiology (for example, Yip et al., 2022). A possible extension of this model is to expand the regression component with a multi-level hierarchical structure to adjust the estimates of data-rich competitions:

$$\begin{aligned} \log \lambda_i &= \log \lambda_{2i} + \mathbf{X}_{1i}' \boldsymbol{\alpha} \\ \log \lambda_{2i} &= \mathbf{X}_{2i}' \boldsymbol{\beta}, \end{aligned}$$

where the term $X_{2i}\boldsymbol{\beta}$ is the regression component on covariates of all competitions and $X_{1i}\boldsymbol{\alpha}$ is the regression component of data-rich competitions such as inclusion of zonal and player-based statistics. These information can be obtained from OPTA sportsdata (Liu et al., 2013) takes into account of on-pitch events and player actions with spatial locations. From the betting simulation, some competitions with a large fan base such as English Premiere League and

European Champions League are recorded negative profits. Further modelling on the data-rich elite competitions will need to be developed and experimented.

Acknowledgements

We thank Jonas Christoffer Lindstrøm to provide helpful supports in using the `implied` package in R.

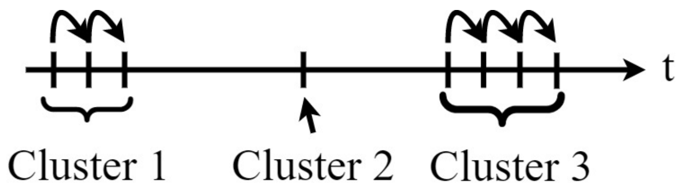

Figure 1: Schematic diagram to illustrate how corner events spawned from its "parent" event.

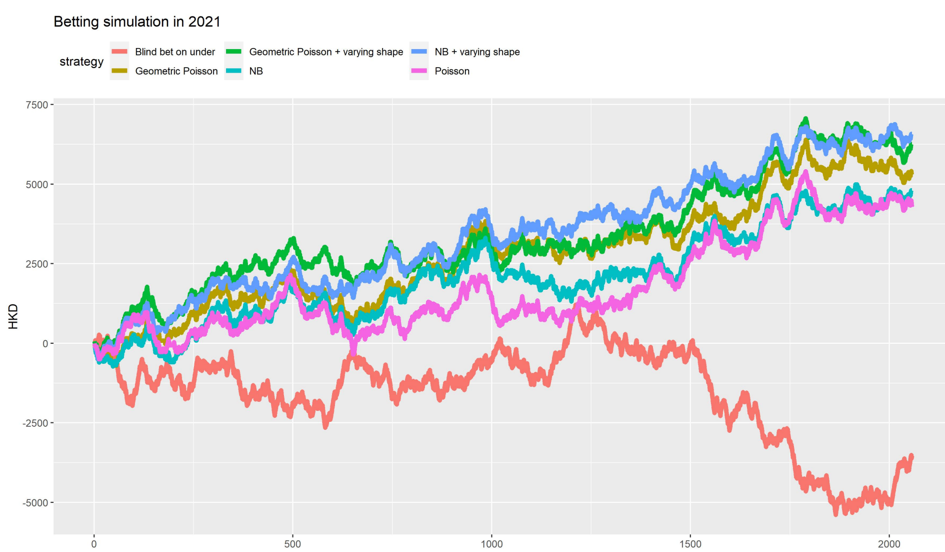

Figure 2: Cumulative profits for 2057 matches offered by the HKJC in 2021 from 1st January to 10th June in 2021.

References

Adelson, R. (1966). Compound Poisson distributions. *Journal of the Operational Research Society*, *17*, 73–75.

Allmers, S., & Maennig, W. (2009). Economic impacts of the FIFA soccer world cups in France 1998, Germany 2006, and outlook for South Africa 2010. *Eastern Economic Journal*, *35*, 500–519.

Angelini, G., & De Angelis, L. (2017). Parx model for football match predictions. *Journal of Forecasting*, *36*, 795–807.

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, *35*, 741–755.

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

Bellego, C., Benatia, D., & Pape, L.-D. (2021). Dealing with logs and zeros in regression models. *CREST-Série des Documents de Travail*, .

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*, 1190–1208.

Cain, M., Law, D., & Peel, D. (2000). The favourite-longshot bias and market efficiency in UK football betting. *Scottish Journal of Political Economy*, *47*, 25–36.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*, 1–32.

Casal, C. A., Maneiro, R., Ardá, T., Losada, J. L., & Rial, A. (2015). Analysis of corner kick success in elite football. *International Journal of Performance Analysis in Sport*, *15*, 430–451.

Clarke, S., Kovalchik, S., & Ingram, M. (2017). Adjusting bookmaker's odds to allow for overround. *American Journal of Sports Science*, *5*, 45–49.

Clarke, S. R. (2016). Adjusting true odds to allow for vigorish. In *Proceedings of the 13th Australasian Conference on Mathematics and Computers in Sport. R. Stefani and A. Shembri, Eds* (pp. 111–115).

Conway, R. W., & Maxwell, W. L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, *12*, 132–136.

Deschamps, B., & Gergaud, O. (2007). Efficiency in betting markets: evidence from English football. *The Journal of Prediction Markets*, *1*, 61–73.

Deutscher, C., Ötting, M., Schneemann, S., & Scholten, H. (2019). The demand for English premier league soccer betting. *Journal of Sports Economics*, *20*, 556–579.

Dixon, M. J., & 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*, 265–280.

Dixon, M. J., & Pope, P. F. (2004). The value of statistical forecasts in the UK association football betting market. *International Journal of Forecasting*, *20*, 697–711.

Feller, W. (1943). On a general class of "contagious" distributions. *The Annals of Mathematical Statistics*, *14*, 389 – 400. URL: `https://doi.org/10.1214/aoms/1177731359`. doi:`10.1214/aoms/1177731359`.

Fitt, A., Howls, C., & Kabelka, M. (2006). Valuation of soccer spread bets. *Journal of the Operational Research Society*, *57*, 975–985.

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

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Geng, X., & Xia, A. (2022). When is the Conway–Maxwell–Poisson distribution infinitely divisible? *Statistics & Probability Letters*, *181*, 109264.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*, 359–378.

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

Greenwood, M., & Yule, G. U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society*, *83*, 255–279.

Guikema, S. D., & Goffelt, J. P. (2008). A flexible count data regression model for risk analysis. *Risk Analysis: An International Journal*, *28*, 213–223.

Higgs, N., & Stavness, I. (2021). Bayesian analysis of home advantage in North American professional sports before and during COVID-19. *Scientific Reports*, *11*, 1–11.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hoffman, M. D., Gelman, A. et al. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, *15*, 1593–1623.

Hutchins, B. (2016). Tales of the digital sublime: Tracing the relationship between big data and professional sport. *Convergence*, *22*, 494–509.

Johnson, N. L., Kotz, S., & Kemp, A. W. (2005). *Univariate discrete distributions*. John Wiley & Sons.

Jullien, B., & Salanié, B. (1994). Measuring the incidence of insider trading: A comment on Shin. *The Economic Journal*, *104*, 1418–1419.

Karlis, D., & Ntzoufras, I. (2009). Bayesian modelling of football outcomes: using the skellam's distribution for the goal difference. *IMA Journal of Management Mathematics*, *20*, 133–145.

Kemp, C., & Kemp, A. W. (1965). Some properties of the 'Hermite' distribution. *Biometrika*, *52*, 381–394.

Kocherlakota, S. (1988). On the compounded bivariate poisson distribution: A unified treatment. *Annals of the institute of Statistical Mathematics*, *40*, 61–76.

Kocherlakota, S., & Kocherlakota, K. (2001). Regression in the bivariate poisson distribution, .

Koopman, S. J., & Lit, R. (2019). Forecasting football match results in national league competitions using score-driven time series models. *International Journal of Forecasting*, *35*, 797–809.

Kovalchik, S. A. (2016). Searching for the goat of tennis win prediction. *Journal of Quantitative Analysis in Sports*, *12*, 127–138.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*, 1–14.

Leitner, C., Zeileis, A., & Hornik, K. (2010). Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the EURO 2008. *International Journal of Forecasting*, *26*, 471–481.

Levin, B., & Reeds, J. (1977). Compound multinomial likelihood functions are unimodal: Proof of a conjecture of IJ Good. *The Annals of Statistics*, (pp. 79–87).

Ley, C., Wiele, T. V. d., & Eetvelde, H. V. (2019). Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches. *Statistical Modelling*, *19*, 55–73.

Lindstrøm, J. C. (2021). *implied: Convert Between Bookmaker Odds and Probabilities*. URL: `https://CRAN.R-project.org/package=implied` r package version 0.4.0.

Liu, H., Hopkins, W., Gómez, A. M., & Molinuevo, S. J. (2013). Inter-operator reliability of live football match statistics from OPTA Sportsdata. *International Journal of Performance Analysis in Sport*, *13*, 803–821.

Lo, A. W. (2002). The statistics of Sharpe ratios. *Financial Analysts Journal*, *58*, 36–52.

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

Mattera, R. (2021). Forecasting binary outcomes in soccer. *Annals of Operations Research*, (pp. 1–20).

Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter*, *3*, 27–32.

i Morata, L. B. (2009). A priori ratemaking using bivariate poisson regression models. *Insurance: Mathematics and Economics*, *44*, 135–141.

Neyman, J. (1939). On a new class of "contagious" distributions, applicable in entomology and bacteriology. *The Annals of Mathematical Statistics*, *10*, 35–57.

Özel, G., & Inal, C. (2010). The probability function of a geometric Poisson distribution. *Journal of Statistical Computation and Simulation*, *80*, 479–487.

Pollard, R. (1973). Collegiate football scores and the negative binomial distribution. *Journal of the American Statistical Association*, *68*, 351–352.

Pólya, G. (1930). Sur quelques points de la théorie des probabilités. *Annales de l'institut Henri Poincaré*, *1*, 117–161.

Pope, P. F., & Peel, D. A. (1989). Information, prices and efficiency in a fixed-odds betting market. *Economica*, (pp. 323–341).

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2017). Catboost: unbiased boosting with categorical features. *arXiv preprint arXiv:1706.09516*, .

Pulling, C., Robins, M., & Rixon, T. (2013). Defending corner kicks: analysis from the English Premier League. *International Journal of Performance Analysis in Sport*, *13*, 135–148.

Quenouille, M. H. (1949). A relation between the logarithmic, Poisson, and negative binomial series. *Biometrics*, *5*, 162–164.

Reade, J. (2014). Information and predictability: Bookmakers, prediction markets and tipsters as forecasters. *Journal of Prediction Markets*, *8*, 43–76.

Roberts, A., Roche, N., Jones, C., & Munday, M. (2016). What is the value of a Premier League football club to a regional economy? *European Sport Management Quarterly*, *16*, 575–591.

Rodríguez-Avi, J., & Olmo-Jiménez, M. J. (2017). A regression model for overdispersed data without too many zeros. *Statistical Papers*, *58*, 749–773.

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

Sengupta, P., Chaganty, N. R., & Sabo, R. T. (2016). Bivariate doubly inflated poisson models with applications. *Journal of Statistical Theory and Practice*, *10*, 202–215.

Sharpe, W. F. (1966). Mutual fund performance. *The Journal of business*, *39*, 119–138.

Shin, H. S. (1992). Prices of state contingent claims with insider traders, and the favourite-longshot bias. *The Economic Journal*, *102*, 426–435. doi:10.2307/2234526.

Shin, H. S. (1993). Measuring the incidence of insider trading in a market for state-contingent claims. *The Economic Journal*, *103*, 1141–1153.

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

Stein, M., Janetzko, H., Seebacher, D., Jäger, A., Nagel, M., Hölsch, J., Kosub, S., Schreck, T., Keim, D. A., & Grossniklaus, M. (2017). How to make sense of team sport data: From acquisition to data modeling and research aspects. *Data*, *2*, 2.

Stekler, H. O., Sendor, D., & Verlander, R. (2010). Issues in sports forecasting. *International Journal of Forecasting*, *26*, 606–621.

Štrumbelj, E. (2014). On determining probability forecasts from betting odds. *International Journal of Forecasting*, *30*, 934–943.

Sung, M., & Johnson, J. E. (2007). Comparing the effectiveness of one-and two-step conditional logit models for predicting outcomes in a speculative market. *The Journal of Prediction Markets*, *1*, 43–59.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432.

Vovk, V., & Zhdanov, F. (2009). Prediction with expert advice for the Brier game. *Journal of Machine Learning Research*, *10*.

Wheatcroft, E. (2021). Forecasting football matches by predicting match statistics. *Journal of Sports Analytics*, *7*, 77–97.

Wimmer, G., & Altmann, G. (1996). The multiple Poisson distribution, its characteristics and a variety of forms. *Biometrical Journal*, *38*, 995–1011.

Yip, S., Him, N. C., Jamil, N. I., He, D., & Sahu, S. K. (2022). Spatio-temporal detection for dengue outbreaks in the central region of malaysia using climatic drivers at mesoscale and synoptic scale. *Climate Risk Management*, *36*, 100429.

Zhang, K., Schölkopf, B., Muandet, K., & Wang, Z. (2013). Domain adaptation under target and conditional shift. In *International Conference on Machine Learning* (pp. 819–827). PMLR.