Does accounting for an artificial turf advantage in Dutch football increase predictive accuracy of probabilistic models?

Gertjan S Verhoeven (gertjan.verhoeven@gmail.com)

20 juli, 2018

Summary[1]

Currently, one in three matches in Dutch professional football ("Eredivisie") is played on an artificial turf surface. Recently, statistical evidence was reported that suggest an increased home advantage for Dutch teams playing on artificial turf against an away team that plays its home games on natural grass (van Ours, 2017). Here we investigate if accounting for this effect increases out-of-sample predictive accuracy of match outcomes. To do so, we implemented existing probabilistic models to make one-step-ahead forecasts, with and without the additional artificial turf predictor. The ranked probability score (RPS) is used to assess the accuracy of the forecasts and compare between models. We find that including the artificial turf home advantage as additional predictor does not improve the accuracy of the forecasts. We conclude that the evidence for a large artificial turf advantage in the Eredivisie is not strong.

Keywords: forecasting, soccer, artificial turf

Introduction

In the past few years, several professional football clubs in the Netherlands have switched to playing their home field games on an artificial pitch, apparantly due to financial reasons. Currently (since season 2014/2015), one in three matches (6 of 18 teams) in Dutch professional football (Eredivisie) is played on an artificial turf surface. This has led to a heated debate in the Netherlands which in 2017 resulted in the team captains of the twelve natural grass teams to call for a ban on artificial turf in the Eredivisie. Later that year, this was followed by a manifesto "Quit using artificial turf" signed by over a hundred coaches, former players, trainers etc.

[1]The author would like to thank Rutger Lit, Ramsis Croes, Misja Mikkers and two anonymous reviewers for useful comments and suggestions.

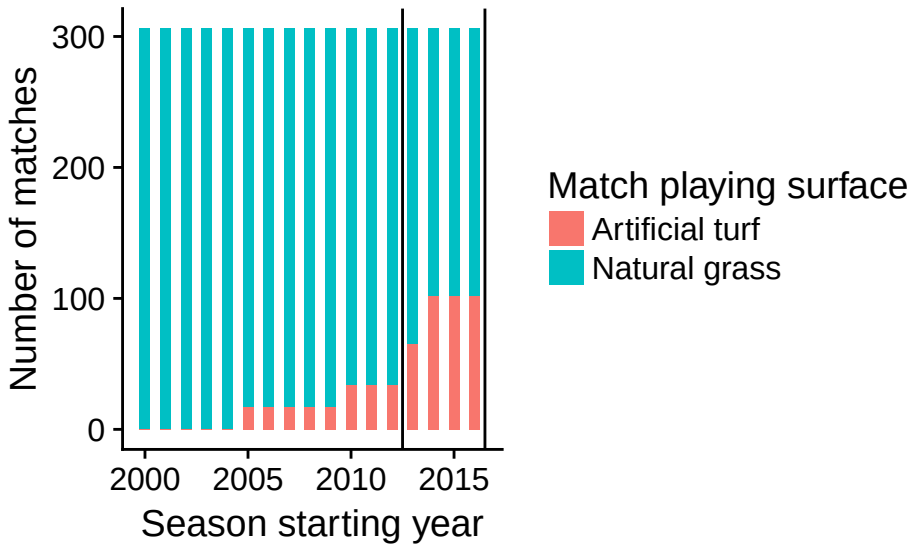

Figure 1: Number of matches played on artificial turf and natural grass in the Dutch Eredivisie by season starting year. Vertical lines indicate the matches analyzed in this paper.

One of the arguments against allowing artificial turf in league competitions (there are several more, see (Hvattum, 2015) for a discussion with references) is that a team playing on artificial turf would have an unfair advantage when playing against a team that plays its home matches on natural grass. Research on four teams with artificial pitches competing in the English Football League during the 1980's (Barnett & Hilditch, 1993) concluded that this advantage existed and was of sufficient scale to be cause of concern. However, since then, artificial turf technology has evolved and this conclusion might not hold true anymore for current artificial playing surfaces. There are now three recent studies with results on the effect of artificial turf on match outcomes. Trombley (Trombley, 2016) concludes that artifical turf does not affect the competitive balance in Major League Soccer in the US for the years 2011-2014. Hvattum (Hvattum, 2015) finds that team strength on average increases after switching from natural grass to artificial turf for Norwegian soccer teams, and using a model based on ELO rating differences, finds a statistically significant artificial turf advantage in-sample. However, out-of-sample predictive performance is virtually unchanged when including the effect as a predictor.

Finally, van Ours (van Ours, 2017) finds a large in-sample effect using an analysis based on group differences. The statistical analysis was based on aggregate differences between four types of matches. A substantial point-estimate (+0.5 additional goals per match holding all else equal) was reported, together with a model to assess the statistical significance (the estimate was statistically significant non-zero with p-value 0.02). It was concluded that artificial pitches lead to unfair advantages that could lead to undeserved relegation of lower ranking teams.

Given the large point estimate (even larger than typical reported values for the regular, well documented home advantage), as well as the large percentage of affected matches (25% of all matches in the last three completed seasons), one would expect that incorporating the artificial turf home advantage in forecasting models would substantially increase their predictive accuracy.

To test this hypothesis, a predictive model is needed that uses a delineated information set as a basis for its predictions, to which information on the playing surface can be added. As the Dutch artificial

turf advantage finding (van Ours, 2017) was based on analysis of historical match result data, we restrict our analysis to this dataset as well (available on e.g. http://www.football-data.co.uk). This rules out more refined predictive models based on e.g. within match shot attempts (expected goals) or models that incorporate information on individual players (abilities, injuries etc). This also rules out using Bookmaker odds as predictions, since we do not known what information has been incorporated in them.

We surveyed the literature on probabilistic modeling of football outcomes to identify predictive models that appear to perform well on historic match outcomes. We setttle on two variants that model the goal difference of a match using latent parameters for team strength. The outcome of a match in terms of Win, Draw or Loss can be directly derived from the predicted goal difference. Both variants are dynamic, i.e. they allow the team strengths to vary over time. They differ in the distribution that is assumed for the goal differences: The first variant assumes a Poisson difference distribution, also known as the Skellam distribution (Skellam, 1946, Karlis & Ntzoufras (2009), Lit (2016)). The second variant assumes a t-distribution, that allows for long tails and includes the normal distribution as a special case (Kharratzadeh, 2017). Note that the Skellam distribution is the difference of two independent Poisson models and is therefore a discrete distribution, whereas the t-distribution is a continuous distribution.

We subsequently expand these two probabilistic model variants with an additional dummy variable that equals one if the match satifies the two conditions for the additional home advantage due to artificial turf. This requires a) that the match is played on artificial turf, and b) that the away team plays its home matches on natural grass.

The main result of this paper is that we find no evidence of improved forecasts when accounting for the additional home advantage due to artificial turf. In contrast, leaving out the regular home advantage gives a strong decrease in predictive accuracy. Apart from this, we find that differences in accuracy between the various predictive model variants are relatively small. No model outperforms forecasts derived from the betting odds of two well-known bookmakers, although the models do come surprisingly close.

The paper is organized as follows. After motivating and describing our model variants, we then describe our data as well as how we perform inference. Then we describe our model comparison approach, which is based on one-step-ahead forecasts that are evaluated with the Ranked Probability Score (RPS) (Epstein, 1969, Constantinou & Fenton (2012)). Finally we present and discuss the results, including a section where we perform model checking (also known as posterior predictive checking).

What model to use?

Using statistical models to predict the result of football matches has a long tradition going back to Maher (Maher, 1982). Maher used two static independent Poisson models for the home and away team respectively, to predict the number goals scored for each team. A landmark paper by Dixon and Coles (Dixon & Coles, 1997) improved on Maher's model in several ways, most notably by introducing time-dependence for the team abilities. Another important work is Rue and Salvesen (Rue & Salvesen, 2000) who present a Bayesian dynamic generalized linear model, with team abilities modeled as a random walk in time (Brownian motion). There are many more papers in this literature, for an extensive review we refer the reader to (Constantinou & Fenton,

2013, Koopman & Lit (2017)).

After reviewing the literature, we conclude that must-haves for a competitive parametric football model are:

1. Inclusion of the regular home advantage (Pollard, 2008)
2. Allowance for time-varying team abilities (Lit, 2016)
3. Adressing the correlation between goals made by the home and the away team (Koopman & Lit, 2015)
4. Partial pooling of the variance of the team ability time evolution (Lit, 2016, Kharratzadeh (2017))

The first two points are described in more detail in the next section, where we present the model variants used in this paper.

Regarding the third point, this has been typically adressed in the literature in one of two main ways. The first is modelling the match score as a joint distribution of two independent Poisson processes, and adding a correlation parameter γ. This is the bivariate Poisson model (Karlis & Ntzoufras, 2003, Koopman & Lit (2015)).

An alternative approach, and the one we adopt here, is to take the difference of the pair of goals for each match, and model the score difference. For the distribution of score differences, we consider both a Poisson difference (or Skellam) distribution (Karlis & Ntzoufras, 2009, Koopman & Lit (2015))[2], as well as a t-distribution. It has been shown by (Koopman & Lit, 2017) that both approaches (modeling the score or the score difference) give very similar predictive accurracy for the outcome in terms of Win, Draw, Loss.

Regarding the fourth and last point: Team strength is expected to change even within a single season for various reasons (e.g. players come and go, get injured, learning effects). In addition, because of the relatively low amount of goals made in each match, a single match outcome in terms of goal difference does not provide us with a strong signal on relative team ability. Estimating team-specific time dynamics (No pooling) might lead to overly noisy, volatile team abilities. Estimating equal time dynamics for all teams (complete pooling) is one way to cope with this situation. However, (Lit, 2016) found that higher ranking teams having more stable abilities compared to lower ranking teams, and that this distinction improves model fit noticably. An attractive alternative is to use partial pooling (i.e. multilevel / hierarchical modelling) for the time dynamics, which estimates the amount of pooling from the data (Greenland, 2000). We therefore adopt the multilevel variance model of Milad Kharratzadeh (Kharratzadeh, 2017)[3] for both model variants.

[2] Both the difference of two independent Poisson distributions, as well as the difference between two Bivariate Poisson counts with correlation γ is distributed as a Skellam distribution, because the correlation term cancels out by differencing (Koopman & Lit 2014)

[3]The paper of Milad Kharratzadeh has been peer reviewed and is fully reproducible available at https://github.com/stan-dev/stancon_talks/tree/master/2017/Contributed-Talks/02_kharratzadeh. For the Skellam model, The Stan code of Ben Torvaney of the Karlis & Ntzoufras static Skellam model on GitHub (https://github.com/Torvaney/soccerstan) provided a starting point. Combining these two sources led to a dynamic multilevel Skellam model, that shares many similarities (apart from the multilevel aspect) with the dynamic Skellam model of Koopman & Lit.

Description of the model variants

We adopt a Bayesian approach in our modelling. This requires defining prior distributions for all our parameters, as well as a likelihood of observing the data given specific values for all parameters. We then use Bayes' rule to compute the posterior probabilities, conditional on the data and our model. As mentioned earlier, we model the goal difference Y of a football match. From the goal difference, the outcome of the match in terms of win, draw, or loss can be derived directly.

A natural way to model the goal difference Y_{ijt} for the match between home team i and away team j at time t is as a difference of two Poisson distributions, each with its own rate parameter λ_{it} and λ_{jt} respectively. The unit of time t is week number, since each team plays at most once a week.

$$Y_{ijt} \sim Skellam(\lambda_{it}, \lambda_{jt}) \tag{1}$$

The rate parameters λ ("scoring intensities") are modeled as follows:

$$\lambda_{it} = \exp(\mu + \delta + \kappa d_{ijt} + \alpha_{it} - \beta_{jt}) \tag{2}$$

$$\lambda_{jt} = \exp(\mu + \alpha_{jt} - \beta_{it}) \tag{3}$$

In equation (2), δ captures the regular home advantage, whereas d_{ijt} is a dummy predictor capturing the additional home advantage κ due to artificial turf. d_{ijt} equals one when at time t, the game between home team i and away team j satisfies the two required conditions: a) the match is played on artificial turf, and b) the away team plays its home matches on natural grass. The parameter μ, together with the home advantage, determines the scoring intensities when two teams with equal strength play each other. Because we expect Poisson rates of order of magnitude 1 [4], and since we have an exponential link function (to constrain the rate λ to be always positive), this translates to a value around zero for the exponents (2) and (3). We therefore choose for μ, δ and κ normally distributed priors around zero:

$$\mu, \delta \sim Normal(0, 1) \tag{4}$$

$$\kappa \sim Normal(0, 5) \tag{5}$$

We choose the scale for the artificial turf advantage as largely uninformative for the expected range of plausible values (plus or minus 1) to reflect that we have only very little prior information for this parameter.

The parameters α_{it} and β_{it} model the attack and defense ability of team i at time t. The time evolution of these parameters is modeled as a random walk [5]: The ability of a team at time t is

[4] The mean of a Poisson distribution is equal to its rate parameter λ. In the last four complete seasons, on average 1.7 and 1.3 goals were scored by the home and away team respectively.

[5] A random walk is not a stationary process and can drift off to infinity. The process of conditioning on the data however prevents this from happening. In addition, Koopman & Lit (2015) model the latent ability α as an first order autoregressive process ($\alpha_{it} = \phi\alpha_{i,t-1}$) and estimate a value for the persistence parameter ϕ of 0.996, i.e. very close to 1, the value of a random walk.

assumed to be equal to the ability at time $t-1$ plus a noise term that is normally distributed around zero with a team-specific variance σ_i:

$$\alpha_{it} = \alpha_{i,t-1} + \eta_{it} \tag{6}$$

$$\eta_{it} \sim Normal(0, \sigma_i) \tag{7}$$

We initialize the time series using data on previous season's performance for each team. We follow the approach of (Kharratzadeh, 2017): The previous performance scores z_i are calculated by the final sum of league points of each team in the previous season (with 3 points for a win, 1 point for a draw and zero points for a loss), transformed to a range between (-1, +1) using linear scaling between the lowest and highest points. For teams that got promoted from the second highest league ("Jupiler league") we choose $z = -1$ as most likely ability given that they were relegated in the past from the Eredivisie.

The first week's abilities are then modeled as a weighted average of the z_i, as well as latent variables η_{i0} generated from a normal distribution with a variance σ_0 that is estimated from the data.

$$\alpha_{i0} = wz_i + \eta_{i0} \tag{8}$$

$$\eta_{i0} \sim Normal(0, \sigma_0) \tag{9}$$

The weight parameter w is a free parameter that can give more weight to the supplied team historical performance z_i if this fits the data well.

Finally, we model the variances σ_i of the separate time series (two for each team, for the Attack and defense ability parameters) as coming from a half-normal distribution with hyperparameter τ:

$$\sigma_i \sim HalfNormal(0, \tau) \tag{10}$$

This has the effect of shrinking the variances somewhat towards zero, preventing overfitting by smoothing the latent ability parameters estimates (Kharratzadeh, 2017). See also the discussion on partial pooling in the previous paragraph. For the defense abilities β_{it} a similar set of equations (with equal priors) is used, with the parameters σ_0, σ_i and τ assumed equal for both attack and defense parameters.

For the Skellam model, a more parsimonious model variant can be obtained by using a single team ability parameter θ_{it} ("Strength") instead of α_{it} and β_{it}.

If we replace the Skellam distribution by a t-distribution and model the mean of this distribution as the difference of two "scoring intensities", we obtain the model of (Kharratzadeh, 2017). A t-distribution has, in addition to a location (Here defined as $(\lambda_{it} - \lambda_{jt})$) and a scale parameter σ_Y, a third parameter ν (also called "degrees of freedom") that determines the shape of the tails. For large values of ν, a t-distribution approaches a normal distribution.

This modeling approach does not allow diffentiating team strength into two separate team ability parameters α_{it} and β_{it} ("Attack" and "Defense"). To see that such a model would not be identified,

we rearrange the terms in (2) and (3) (omitting the exponent) as $\lambda_{it} - \lambda_{jt} = \delta + d_{ijt}\kappa + (\alpha_{it} + \beta_{it}) - (\alpha_{jt} + \beta_{jt}) = \delta + d_{ijt}\kappa + \phi_{it} - \phi_{jt}$. This leaves us with a single ability parameter ϕ_{it} for each team. Including a constant μ for each scoring intensity λ would also leave the model unidentified.

We end up with the following model equations:

$$Y_{ijt} \sim t(\lambda_{it} - \lambda_{jt}, \sigma_Y, \nu) \tag{11}$$

$$\lambda_{it} = \delta + d_{ijt}\kappa + \phi_{it} \tag{12}$$

$$\lambda_{jt} = \phi_{jt} \tag{13}$$

$$\phi_{it} = \phi_{i,t-1} + \eta_{it} \tag{14}$$

Note that we no longer need the exponential function to constrain the λ's in the Skellam distribution to the range $[0, \infty]$.

Finally, following (Karlis & Ntzoufras, 2009, Lit (2016)), the Skellam models contain a zero inflation component. This results in a mixed model, where with a probability p the goal difference is 0, and with a probability $1 - p$ the goal difference is generated by the Skellam model. One model variant leaves out this zero inflation component to learn the effect of including it. A uniform distribution between 0 and 1 is used as a prior for p.

Inference from data

All our models are coded in Stan, a probabilistic programming language for Bayesian modelling (Carpenter *et al.*, 2016). Stan programs are compiled and during execution generate MCMC samples from the posterior distribution. Stan uses Hamiltonian Monte Carlo sampling, in particular the the No-U-Turn Sampler (Hoffman & Gelman, 2014), which allows for very efficient exploration of high-dimensional parameter spaces.

Match level data for the dutch Eredivisie was obtained from http://www.football-data.co.uk). This dataset also contained Win/draw/loss odds for each match from various bookmakers, including William-Hill and Bet365, that we use as a benchmark. To convert the bookmaker odds to probabilities we use standard normalization, where the sum of the raw probabilities obtained by inverting the odds are scaled to equal 100% (Štrumbelj & Šikonja, 2010) [6].

Information on playing surface (natural or artificial) for each team was obtained from various sources, and verified by comparing with news media postings online.

Each model variant is fitted separately for each out-of-sample week (2 seasons x 34 playing weeks = 68 weeks in total). Fitting a model consists of running six MCMC chains with 1000 sample

[6] For a match with outcomes Win/Draw/Lose with given bookmakers odds $o_i = o_W, o_D, o_L$, the inverse odds $\pi_i = 1/o_i$ do not represent probabilities because they do not sum to one. Dividing by the sum $\Pi = \sum_{i=1}^{3} \pi_i$, we obtain $p_i = \pi_i/\Pi$. p_i are called standard (or basic) normalized outcome probabilities.

draws each (sampled in parallel on six cores) with the first 500 samples of each chain discarded as warmup, retaining the second 500 draws. This gives a total of $M = 3000$ posterior draws for each model fit. A few model variants initially gave warnings during samples (divergent iterations or maximum treedepth reached), after increasing adapt_delta to 0.99 and max_treedepth to 15 all chains sampled without any warnings. To check convergence and diagnose potential problems during sampling, Stan reports the potential scale reduction factor (PSRF) statistic ($\hat{R}$) as well as the effective sample size n_{eff} for each parameter. PSRF values greater than 1.05 and ratios $n_{eff}/N < 0.1$ can heuristically be interpretated as worrisome.

For all fitted models, for all parameters $\hat{R} < 1.02$ except for a few (N = 35, mostly σ_0 and τ) parameters with $1.02 < \hat{R} < 1.04$, and a single parameter with $\hat{R} = 1.06$. Also a few (N = 25) parameters had $n_{eff}/N < 0.1$, of which 15 had also elevated PSRF values (> 1.02). Our subjective judgement is that all MCMC chains have converged succesfully.

The set of 68 out-of-sample model fits results in 68 x 9 = 612 - 4 = 608 forecasts[7]. To check whether $M = 3000$ draws gave sufficiently stable forecasts, we ran variants of the $t-$distribution model (with and without artificial turf advantage) twice, generating for each model two vectors of 608 ranked probability scores. The correlation coefficient between these vectors was 0.998, for both model variants. Both runs had the same average RPS. This was deemed to be sufficiently stable for model comparison. We verified that the models were coded properly by simulating parameters and data from the generative model and verifying that the parameters are properly recovered.

All scripts for this paper, including the paper itself will become available on a Github repository.

Forecasting approach

Our goal is to compare various models on out-of-sample predictive accuracy (forecasting). The motivation for our approach is that out-of-sample evidence for the artificial turf advantage is more convincing than in-sample evidence. For this approach, we face a trade-off in the choice which data we use to estimate the parameters ("in-sample"), and which data to use for out-of-sample prediction.

The artificial turf advantage effect was reported using data from the three most recent complete seasons (2014/2015 up to 2016/2017) (van Ours, 2017). Ideally, we would like to forecast match outcomes for these three seasons. However, we also need data to estimate the parameters, and to estimate the artficial turf advantage, this requires sufficient in-sample matches that satisfy the ATA conditions. Therefore, we decided to use the last four seasons of data (see also Figure 1), and decide (somewhat arbitrarily) to use two seasons worth of data (612 matches) to estimate the parameters, and use an expanding window to obtain one-step-ahead forecasts for the remaining two seasons (612 matches).

Since the team abilities change over time, increasing the estimation window size by going further back in time is not expected to increase predictive accuracy by much. Only for the time-independent parameters do we expect an impact. Whether the impact is beneficial or detrimental will depend on time horizon, as the constant regular home advantage is an assumption that decreases in plausibility as we increase the period over which we average (in fact it will likely vary in time and across teams).

[7]Four matches could not be predicted because a newly promoted team (not present in the training data up to that point) was playing for the first time at home or away.

To compare model predictive performance we use a proper scoring rule (Gneiting & Raftery, 2007). Scoring rules are used to compare the quality of forecasts. A scoring rule is proper when it incentivizes a forecaster to use his true beliefs when making forecasts that are ranked by applying the scoring rule (Gneiting & Raftery, 2007). For ordered discrete outcomes (Win, draw, loss) with r categories (here $r = 3$), the Ranked Probability Score (RPS) (Epstein, 1969) is a proper scoring rule (Constantinou & Fenton, 2012):

$$RPS = \frac{1}{r-1} \sum_{i=1}^{r-1} (\sum_{j=1}^{i} (p_j - e_j))^2 \tag{15}$$

Here p_j is the cumulative probability density function (CDF) of the model forecast, and e_j is the cumulative outcome vector, that, while stepping through the ordered categories {Win, Draw, Loss} changes from 0 to 1 when the realized outcome occurs. For example, when the actual outcome is a Draw, and the forecast probabilities are p(Win) = 0.2, P(draw) = 0.35, and P(Loss) = 0.45, the RPS is calculated as $[(0.2 - 0)^2 + (0.55 - 1)^2 + (1 - 1)^2]/2 = 0.1212$. Lower RPS values mean more accurate forecasts.

To calculate the probabilities needed for the RPS, we use the posterior predictive distribution $p(y^{rep}|y)$ for observing a goal difference y^{rep}, conditional on the observed data y and the model. This gives for a particular match, for each of the potential outcomes (Win, draw, loss for the home team) the probability of occuring, that can plugged into Equation (15) to obtain the RPS for that match.

$$p(y^{rep}|y) = \int_{\Theta} p(y^{rep}|\theta) p(\theta|y) d\theta \tag{16}$$

Here $p(\theta|y)$ is the posterior distribution of the parameters after conditioning on the data y. $p(y^{rep}|\theta)$ is the likelihood of observing a particular new value of y, given particular values of the parameters of the model θ. The integral can be seen as a weighted average of the likelihood over all possible sets of values for all parameters, each in proportion with their posterior probability. It combines both our uncertainty about the parameters of the model (lack of knowledge), as well as the uncertainty caused by random variability in the data-generated process.

The posterior predictive distribution for new data can be conveniently calculated during MCMC sampling. For each draw θ^m of a total of M draws from the posterior distribution $p(\theta|y)$, we can plug θ^m into the likelihood function. This turns the likelihood function in a sampling distribution. We can directly sample from this distribution using forward simulation. The M samples (one for each draw from the posterior) together form the posterior predictive distribution (16). For example, for the Skellam model variant, a single predicted goal difference is a sample from a Skellam distribution with parameters fixed at θ^m.

The resulting posterior predictive distribution over goal differences y^{rep} forms the basis of our probabilistic forecast. For the Skellam model, the probabilities p(win), p(draw) and p(loss) follow directly from the MCMC sample $y_m^{rep} = y_1^{rep}, y_2^{rep}, y_3^{rep}, \ldots$ with p(win) the percentage of y^{rep} values greater than zero, p(draw) the percentage of y^{rep} values equal to zero and p(loss) the percentage of y^{rep} values lower than zero.

For the t-distribution models, we first discretize the predicted continuous values y^{rep} for the goal differences by defining a draw as a predicted goal difference in the range [-0.5, +0.5]. We then apply

the same procedure as described for the Skellam models.

Bookmaker odds as benchmark

We present two extremes to compare the model-based predictions with: the Online bookmaker' odds implied probabilities as the target "to beat", and a random forecasting strategy of equal probability for each of the three outcomes as uninformed reference prediction. Probabilities derived from bookmaker odds seem to provide very accurate forecasts that are hard to beat with parametric statistical models that use objective data as inputs (Constantinou *et al.*, 2012). Only after incorporating detailed subjective information did their model achieve comparable forecasting performance (measured by RPS) to that of bookmakers. To our knowledge, no parametric football model using match level data is currently able to outperform the predictive accuracy of published betting odds. We note here that predictive accuracy is not the same as profitability of betting strategies based on parametric models.

Average RPS scores of the betting market (including the Dutch Eredivisie) are reported in (Constantinou & Fenton, 2013). For the Dutch Eredivisie the average RPS for the period 2005-2012 is 0.19. However, caution is needed when comparing RPS values, as there is considerable variability in average RPS values over seasons, much larger than the variation in average RPS between bookmakers for the same season. We therefore calculate the average RPS on the exact same data that we let the models predict for.

We chose somewhat arbitrarily the Bet365 odds since it is one of the largest online betting sites, and the William Hill odds since these odds were available for a longer time period (at least since 2000/2001 season) and is a very large and well known bookmaker as well. It is likely that by pooling the various bookmakers even better forecasts can be obtained, but as our goal here is get some intuition in how well our models are performing, this lead was not pursued further.

The William-Hill and Bet365 odds-based forecasts have an average RPS of 0.19 and 0.189 respectively, and correlate strongly (Pearson's correlation coefficient of 0.994) for the 2015/2016 and 2016/2017 Eredivisie seasons combined. The random forecasting strategy of equal probabilities has an average RPS of 0.237.

Model checking

We have performed various checks on the models. A separate document is available that contains all checks. Here we present a few results that illustrate the working of the models.

Figure 2 shows the distribution of goal differences, both of the actual data, as well as predicted by the dynamic Skellam model (This figure is adapted from (Karlis & Ntzoufras, 2009)). This figure is constructed by pooling all posterior predictive distributions for the goal difference (one distribution of size M draws for each of N matches) and creating a histogram that is normalized by M. This gives a visual impression of model calibration.

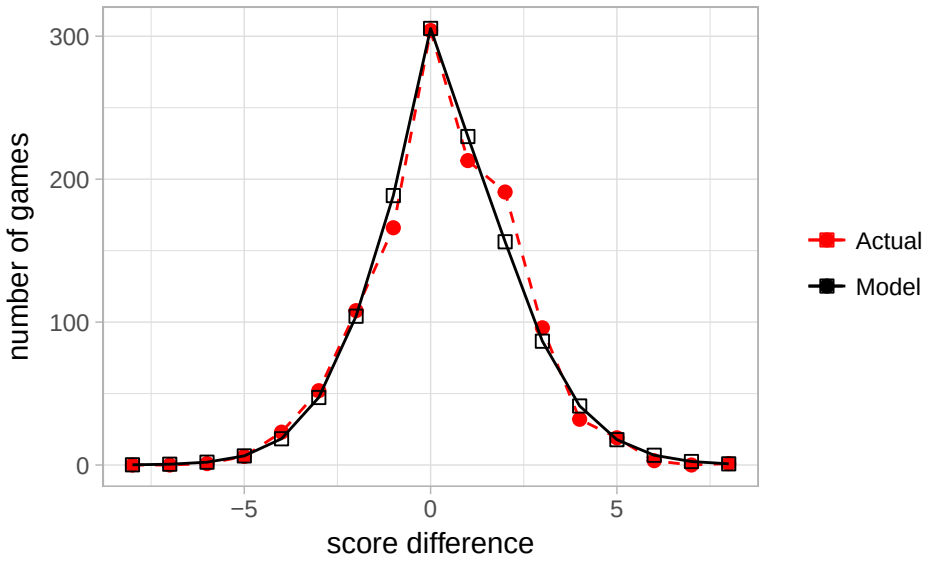

Figure 2: Predicted distribution of goal differences versus actual distribution. Shown are the predictions for the Skellam model.

Figure 3 shows for the t-distribution model the estimated latent team abilities over time for Ajax and Feyenoord. Panel A shows the abilities for the no pooling model (team-specific random walk variance), Panel B with partial pooling (multilevel random walk variance). As expected, the multilevel model leads to smoothed ability estimates.

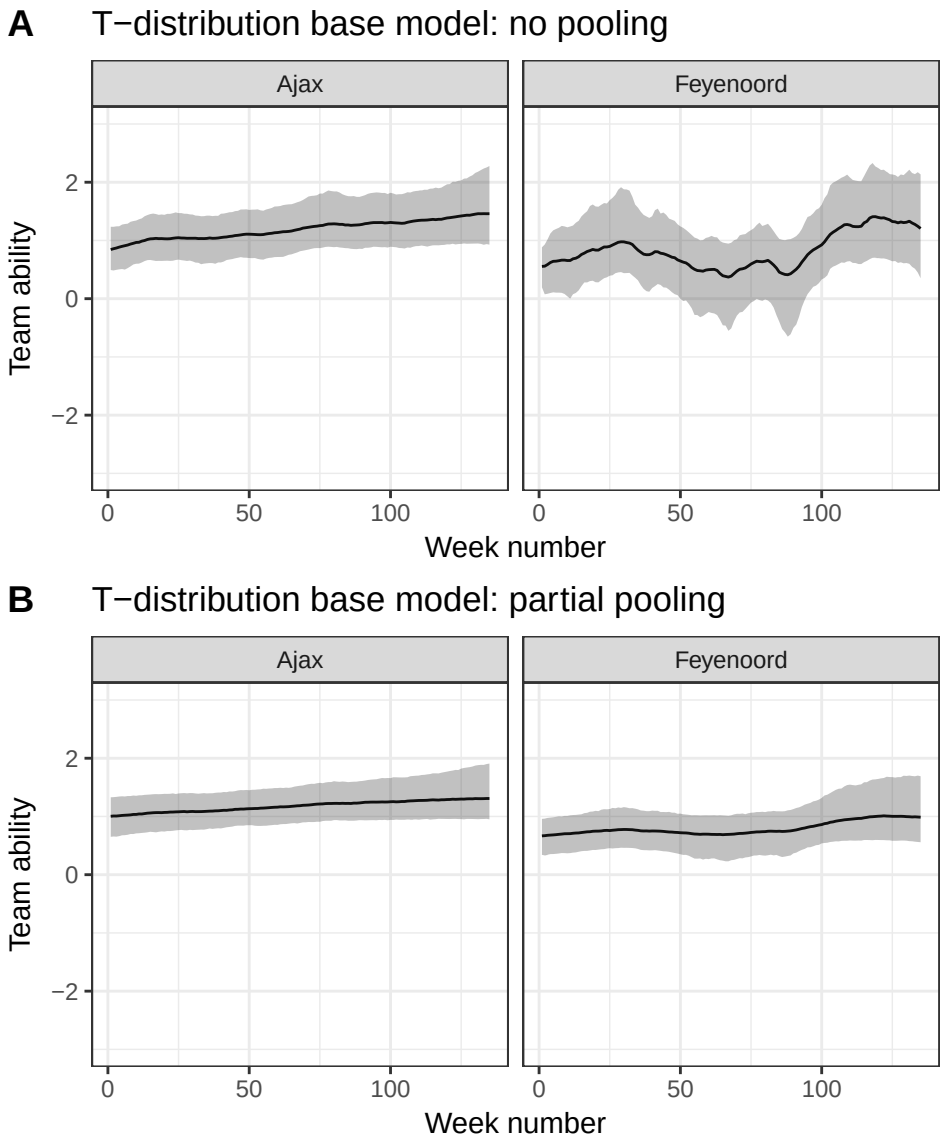

Figure 3: Estimated team abilities: effect of partial pooling of the team-specific random walk variance. The model was fitted on the dataset used to predict the final round of the 4th season (2016/2017).

We also compare the quality of our model predictions against the bookmaker odds. For this we use the Ranked probability score for each match, and aggregate by week to compare over time (Figure 4), as well as by match type (Figure 5).

For the comparison over time, we compare the best model (Skellam model without zero inflation) with the Bet365 odds. Figure 4A shows that the average quality of the forecasts between statistical model and bookmaker odds correlate strongly, and Figure 4B shows that there is no time / seasonal trend or structural break visible over time. The vertical line marks the first week of the second out-of-sample season. Figure 4B contains a local polynomial regression fit to smooth the time series (blue line).

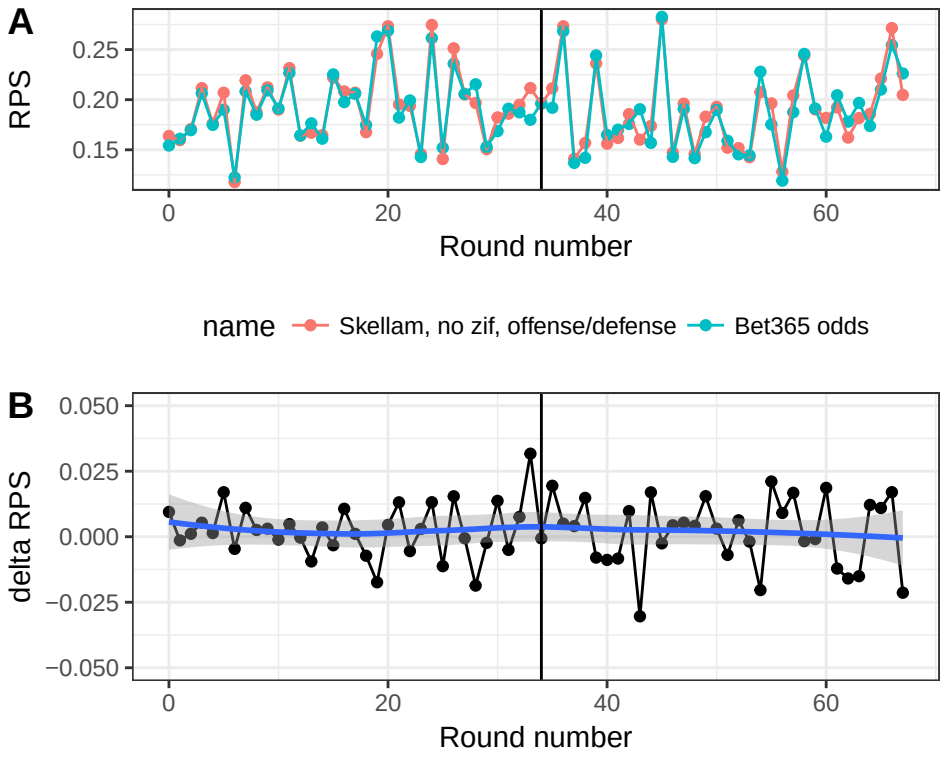

Figure 4: Bet365 vs best Skellam model: A) weekly averaged Ranked Probability Scores over time, for all playing rounds in seasons 2015/2016 and 2016/2017. B) Difference between the values plotted in A) over time, with a polynomial smoothing fit to visualize possible time trends. The vertical line marks the first week of season 2016/2017.

We expect that matches where the team abilities are similar are more difficult to predict. Based on a two-season league table, we label each team as relatively "strong" when it ranks in the upper halve of the league table, and as relatively "weak" when it ends up in the lower halve of the ranking. With these labels, we then classify each match as a match between two strong teams, a match between a weak and a strong team, or a match between two weak team. For each of the three match types we calculate the average RPS for each model. We show results for the best Skellam model (no zero inflation) and for the best t-distribution model, and compare against the Bet365 odds as well as the equal probability reference forecasts.

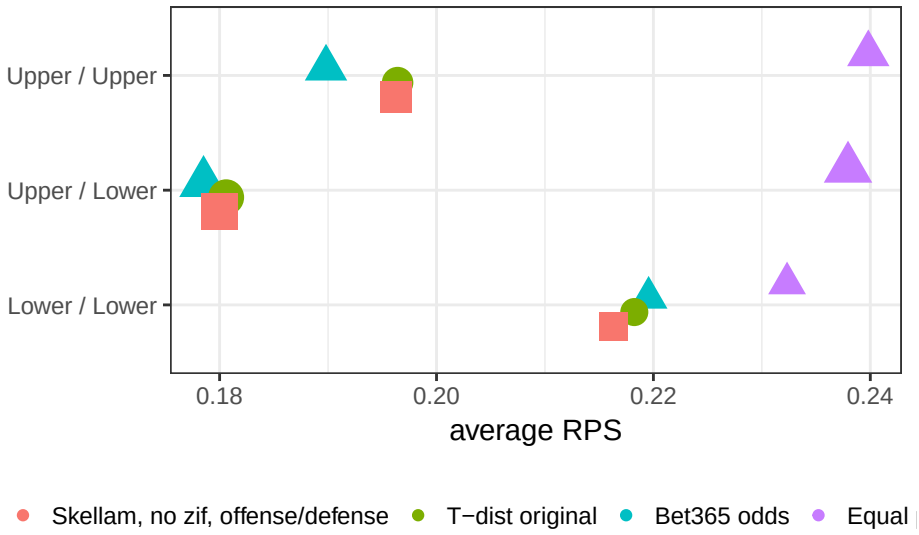

Figure 5: Average Ranked probability scores: models vs. bookmakers, stratified by relative team strength.

Results: In-sample parameter estimates of the home advantage and artifical turf advantage parameters

Figure 6 shows the median posterior value, as well as the 90% plausible intervals of both the regular home advantage as well as the artificial turf advantage. Figure 6A displays the t-distribution based models, Figure 6B displays the Skellam based models.

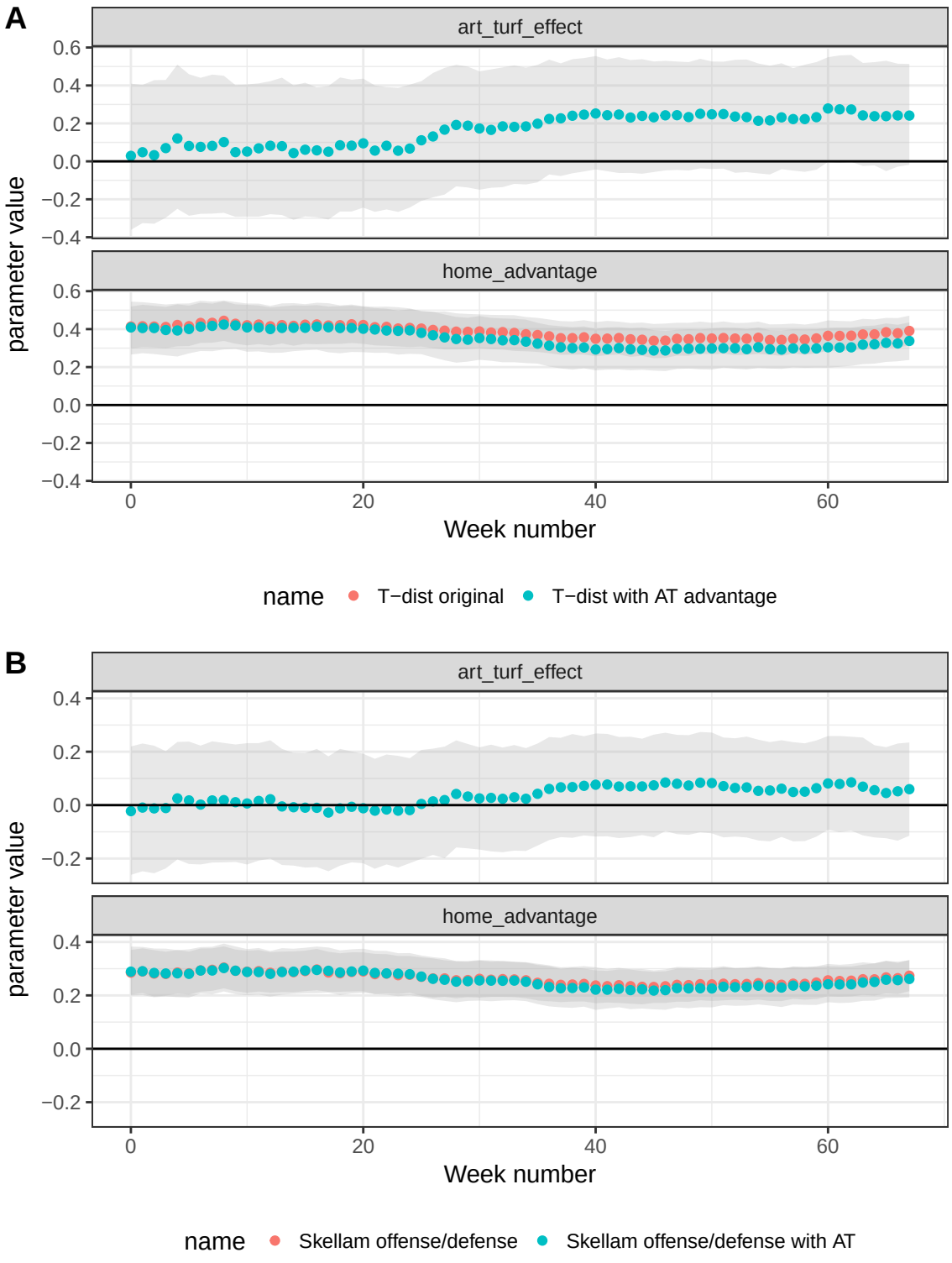

Figure 6: Home advantage & AT advantage during window expansion, including data up to week t.

In both model variants the median posterior value of the artificial turf (AT) advantage parameter starts to increase after including the 20th playing round of the first out-of-sample season (2015/2016)

in the expanding window. After round 36 (The beginning of the second out-of-sample season) for the t-distribution model, the median value of the artificial turf advantage stabilizes at around 0.2, with the 90% credible intervals displaying a lot of uncertainty, including a value of zero. The increase in the AT advantage value correlates with a smaller decrease in the home advantage value. The Skellam model displays the same patterns over time, but the median posterior value remains close to zero, with a lot of uncertainty.

Note that our models do not allow the artificial turf advantage to vary over time, here we estimate an average effect over the full in-sample period, that changes as we expand the window of data included in the estimation. We also like to point out that the AT effect values cannot be directly compared between both model variants, because of different marginal effects of the dummy variable on the goal difference Y. The Skellam model has an exponential link function that shifts the expected value of the Poisson distribution, whereas the t-distribution model has a linear link function that directly shifts the center of the t-distribution.

Results: out-of-sample predictive accuracy

Table 1 shows the main result of the paper, the averaged RPS for each model over two seasons of Eredivisie match outcomes (N = 608 matches). Four matches could not be predicted because a newly promoted team (not present in the training data up to that point) was playing for the first time at home or away. We use the Diebold-Mariano test statistic to test the null hypothesis that two models have equal predictive accuracy. For this we assume that the difference d in RPS value for match i between models m and m', $d_{i,m,m'} = RPS_{i,m} - RPS_{i,m'}$, is normally distributed around zero. The DM-test statistic is then simply the t-statistic for the intercept of a regression of the RPS values of model m on the RPS values of model m'. We use as reference model the model with the lowest average RPS, which are the Bet365 betting odds. The test-statistics and p-values are displayed in Table 1:

Model	distribution	aRPS	DM statistic	DM p-value
Bet365 odds	Benchmark	0.1893	NA	NA
William_hill odds	Benchmark	0.1902	-1.5	0.140
Skellam, no zif, offense/defense	Skellam	0.1914	-1.3	0.200
Skellam offense/defense with AT	Skellam	0.1917	-1.4	0.150
Skellam offense/defense	Skellam	0.1917	-1.4	0.150
Skellam single ability	Skellam	0.1920	-1.7	0.095
T-dist original	T-dist	0.1921	-1.7	0.085
T-dist with AT advantage	T-dist	0.1923	-1.7	0.087
T-dist no pooling	T-dist	0.1957	-3.0	0.003
T-dist no HA	T-dist	0.1981	-2.9	0.004
Equal probability odds	Benchmark	0.2375	-8.4	0.000

It is clear that on average, no model has outperformed the bookmakers predictions. As expected, the reference model of fixed equal probabilities has the worst performance. Two elements clearly matter for the quality of the forecasts: Including the regular home advantage, and including the multilevel model for the time evolution of the team abilities. Apart from these observations, we find that for the remaining models we cannot reject the null hypothesis of equal predictive accuracy.

Surprisingly, a relatively simple model (t-distribution with partial pooling and team strength as a single ability) gives a similar performance compared to a more complex Poisson-based model with attack- and defense parameters. The zero-inflation component does not noticably improve the model accuracy, as was previously found by (Karlis & Ntzoufras, 2009, Lit (2016)). In retrospect, this is not surprising as it adds a fixed amount of probability density to the Draw outcome, irrespective of which teams are playing.

Conclusions

This paper analyses the possibility of an artificial turf advantage in the Dutch Eredivisie, where a team that plays its home matches on artificial turf has an additional advantage against the away team that plays its home matches on natural grass. We report both in-sample model estimates as well as out-of-sample predictive accuracy to learn about the effect. Given the recently reported large point estimate of +0.5 extra goals and the large number of matches that satisfy the artificial turf advantage conditions, we had expected to find a substantial beneficial effect in predictive accuracy of statistical forecasts. However, we find that only the regular home advantage, as well as a multilevel modeling approach have a strong effect on the predictive accuracy. In-sample, for models that are fitted on data consisting of the four most recent seasons, we find for the t-distribution model variant a point estimate of around +0.24 for the artificial turf advantage, with a 90% credible interval that includes 0. For the Skellam model variant, we find a point estimate of +0.06 with a 90% credible interval that includes 0. We conclude therefore that a strong, convincing signal is absent from the data and thus we remain uncertain about the magnitude and degree of variability of a artificial turf advantage in recent Dutch Eredivisie league play.

License

Cite using Zenodo: DOI

References

Barnett, V. & Hilditch, S. (1993) The Effect of an Artificial Pitch Surface on Home Team Performance in Football (Soccer). *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **156**, 39–50.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M.A., Guo, J., Li, P. & Riddell, A. (2016) Stan: A probabilistic programming language. *Journal of Statistical Software*, **20**, 1–37.

Constantinou, A.C. & Fenton, N.E. (2013) Profiting from arbitrage and odds biases of the European football gambling market. *The Journal of Gambling Business and Economics*, **7**, 41–70.

Constantinou, A.C. & Fenton, N.E. (2012) Solving the problem of inadequate scoring rules for

assessing probabilistic football forecast models. *Journal of Quantitative Analysis in Sports*, **8**.

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

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.

Epstein, E.S. (1969) A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, **8**, 985–987.

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

Greenland, S. (2000) Principles of multilevel modelling. *International journal of epidemiology*, **29**, 158–167.

Hoffman, M.D. & Gelman, A. (2014) The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, **15**, 1593–1623.

Hvattum, L.M. (2015) Playing on artificial turf may be an advantage for Norwegian soccer teams. *Journal of Quantitative Analysis in Sports*, **11**, 183–192.

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

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.

Kharratzadeh, M. (2017) Hierarchical Bayesian Modeling of the English Premier League. *Proceedings of the First Stan Conference, StanCon*.

Koopman, S.J. & 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**, 167–186.

Koopman, S.J.(.J.). & Lit, R. (2017) *Forecasting Football Match Results in National League Competitions Using Score-Driven Time Series Models*, Tinbergen Institute.

Lit, R. (2016) Time-Varying Parameter Models for Discrete Valued Time Series.

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

Pollard, R. (2008) Home Advantage in Football: A Current Review of an Unsolved Puzzle. *The Open Sports Sciences Journal*, **1**.

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.

Skellam, J.G. (1946) The frequency distribution of the difference between two Poisson variates belonging to different populations. *Journal of the Royal Statistical Society. Series A (General)*, **109**, 296.

Štrumbelj, E. & Šikonja, M.R. (2010) Online bookmakers' odds as forecasts: The case of European

soccer leagues. *International Journal of Forecasting*, **26**, 482–488.

Trombley, M.J. (2016) Does artificial grass affect the competitive balance in major league soccer? *Journal of Sports Analytics*, **2**, 73–87.

van Ours, J. (2017) *Artificial Pitches and Unfair Home Advantage in Professional Football*, Social Science Research Network, Rochester, NY.