Kicking the Odds: A Bayesian Framework for Football Match Outcome Prediction in Allsvenskan

William Johansson Joaquín Barrutia

October 31, 2023

Uppsala University

Introduction

Motivation

- On 2022-10-03, the Swedish betting site SvenskaSpel introduced a new betting game called **Fullträff.**
- No one has ever been able to predict all 13 games.
- Can a Bayesian Framework help in predicting Allsvenskan results?

Literature Review

- Statistics has extensive literature dedicated to forecasting soccer match outcomes identifying profitable betting opportunities.
- Early works that gained popularity employed:
  - Poisson models to predict soccer scores (Maher, 1982; Dixon and Coles, 1997).
- Authors have recently moved to address this prediction problem within a Bayesian framework (Baio and Blangiardo, 2010; Robberechts et al., 2021)
- Our paper leverages the fantastic tools developed by Egidi (2022) to:
  - Predict score differences before the commencement of matches within a Bayesian framework.
  - We focus on an under-explored league, such as the Allsvenskan.

Data

- We constructed a dataset using information from two distinct sources:
  1. football-data.co.uk website, which has updated match results for leagues around the world.
  2. a https://www.kaggle.com/ dataset with ratings for every team featured in the EA Sports FIFA 22 Video Game.
- Our final dataset includes each of the games of the 16 teams in Allsvenskan in 2022 Season ($240 = 2\binom{16}{2}$)
- For each match we have the following variables:
  - full-time goals scored by both the home and away teams
  - match date
  - overall average of the FIFA 22 teams' rating, computed using attack, midfield and defense ratings

Models

Models

- We used two different kinds of models Static and Dynamic
- The proposed models assume that the data followed two distributions: Bivariate Poisson and Skellam.
- We will describe the easiest base model in detail which is the Static Bivariate Poisson
    - it is really easy to modify the base model to have a Skellam distribution or to change it from Static to Dynamic.

Static Bivariate Poisson

The model is described as

$$y_n^h, y_n^a | \lambda_{1n}, \lambda_{2n}, \lambda_{3n} \sim BivPoisson(\lambda_{1n}, \lambda_{2n}, \lambda_{3n}) \tag{1}$$

$$log(\lambda_{1n}) = \mu + home + att_{h_n} + def_{a_n} + \frac{\gamma}{2}(rank_{h_n} - rank_{a_n}) \tag{2}$$

$$log(\lambda_{2n}) = \mu + att_{a_n} + def_{h_n} - \frac{\gamma}{2}(rank_{h_n} - rank_{a_n}) \tag{3}$$

$$log(\lambda_{3n}) = \beta_0 + \gamma_1 \beta_{h_n} + \gamma_2 \beta_{a_n} + \gamma_3 \beta w_n \tag{4}$$

where (y_n^h, y_n^a) denote the number of goals scored by home and away team in n-th game, $\lambda_{1n}, \lambda_{2n}$ represent home and away teams' scoring rates. Each λ_{in}, μ represent a constant intercept, *home* is the benefit from playing at your stadium, att_{h_n} and def_{h_n} represent the attack and defense ability of home team h playing in game n. $(rank_{h_n} - rank_{a_n})$ are FIFA ranking incorporated as a predictor.

Static Bivariate Poisson

The priors for the parameters of each team t in the model, are:

$$att_t \sim N(\mu_{att}, \sigma_{att}) \tag{5}$$

$$def_t \sim N(\mu_{def}, \sigma_{def}) \tag{6}$$

$$\sigma_{att}, \sigma_{def} \sim Cauchy^{+}(0, 5) \tag{7}$$

$$\gamma \sim N(0, 1) \tag{8}$$

For our sensitivity analysis we changed these priors.

Results

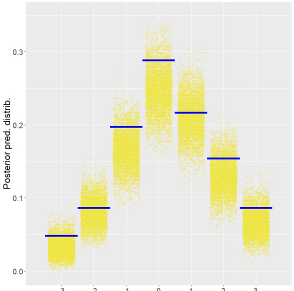
Model Comparisons

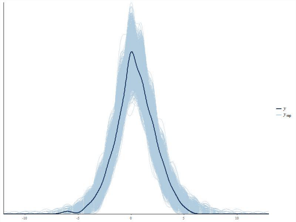
- We approximated the posterior distributions for the parameters of the models employing the Markov Chain Monte Carlo (MCMC).
- Pointwise out-of-sample prediction accuracy

Model	Priors	ELPD$_{loo}$	p_{eff}	LOOIC
BivPoiss Dyn	original	-597.39	17.75	1,194.79
		(11.75)	(1.39)	(23.51)
BivPoiss Stat	original	-596.14	19.61	1,192.28
		(11.29)	(1.35)	(22.59)
Skellam Dyn	original	-399.5	25	799.01
		(11.06)	(2.54)	(22.12)
Skellam Stat	original	-395.86	17.61	791.72
		(11.01)	(1.83)	(22.03)
BivPoiss Dyn	modified	-596.57	19.98	1,193.13
		(11.58)	(1.53)	(23.17)
BivPoiss Stat	modified	-596.26	19.19	1,192.53
		(11.36)	(1.33)	(22.73)
Skellam Dyn	modified	-399.67	26.18	799.33
		(10.98)	(2.81)	(21.96)
Skellam Stat	modified	-395.99	17.59	791.98
		(11.10)	(1.83)	(22.196)

Model Checking

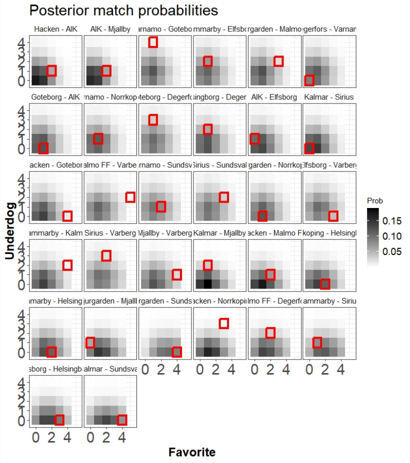
- The score difference frequencies seem to be appropriately captured by the Static Skellam model's replications $(y^h - y^a)^{rep}$ (Figure 1).
- Figure 2: overlap between observed goal difference density and replicated densities,the Static Skellam model reasonably captures the goal difference.





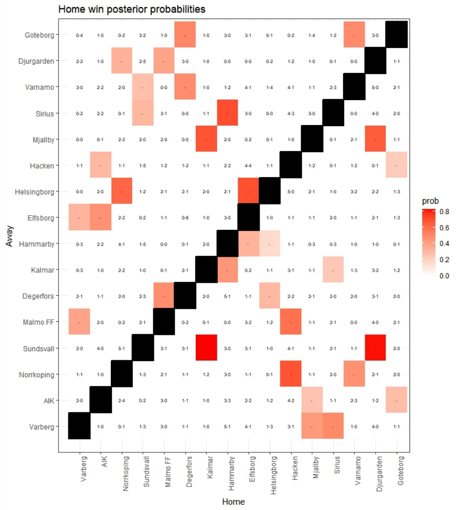
Predictive Accuracy

We can compute the probabilities of the posterior results for all matches used in test set –last 2 weeks of 22 Season ($y^h - y^a$). The red square represents the observed result.



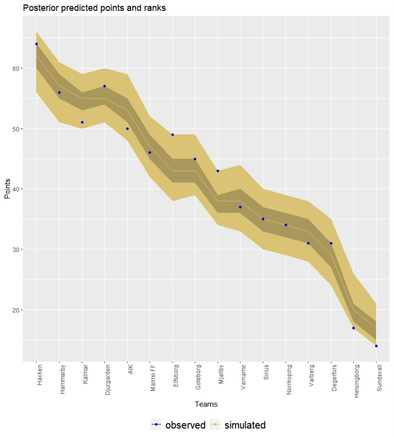
Predictive Accuracy

We can also show the computed probabilities for a home win for the 32 test matches. The red cells denote more likely home wins.



Predictive Accuracy

Finally, we can reconstruct a final rank league table, predicting position and total amount of points at the end of the season using the in-sample replications $(y^h - y^a)^{rep}$ to compute credible intervals.



Conclusions

Conclusions

- The model accurately predicted 18 out of 32 games.
- Simulating the simplest betting strategy
  - betting 1 unit on the most probable outcome according to the model in each of the 32 test-games.
- The profit of betting 100 SEK using this strategy would be 569 SEK.
- The model can be developed even further, with more advanced measures of team rankings or abilities
- The betting strategy can also be improved.

G. Baio and M. Blangiardo. Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, 37 (2):253–264, Feb. 2010. ISSN 0266-4763. doi: 10.1080/02664760802684177. URL https://doi.org/10.1080/02664760802684177. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/02664760802684177.

M. J. Dixon and S. G. Coles. Modelling Association Football Scores and Inefficiencies in the Football Betting Market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46 (2):265–280, 1997. ISSN 1467-9876. doi: 10.1111/1467-9876.00065. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9876.00065. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9876.00065.

L. Egidi. Fitting football models and visualizing predictions with the footBayes package. 2022. URL https://www.researchgate.net/profile/Leonardo-Egidi/publication/362155489_footBayes_fitting_football_models_and_visualizing_predictions_with_the_footBayes_package/data/62d91fb79dd86c7a09204902/footbayes-egidi.pdf.

M. J. Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982. ISSN 1467-9574. doi: 10.1111/j.1467-9574.1982.tb00782.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.1982.tb00782.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9574.1982.tb00782.x.

P. Robberechts, J. Van Haaren, and J. Davis. A Bayesian Approach to In-Game Win Probability in Soccer. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pages 3512–3521, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8332-5. doi: 10.1145/3447548.3467194. URL https://doi.org/10.1145/3447548.3467194.