Analysis of Home Court advantage in Lithuanian Basketball League

Aurimas Račas

April 25, 2022

1 Motivation

Home court advantage is a well-known phenomenon in various sports, and is particularly associated with team sports where teams typically have a home-base arena they play in [1]. Factors that underlie this phenomenon are typically considered to be team's familiarity with the venue, unfavorable impact to the visiting team related to pre-game travel, and the favourable spectator influence (sometimes referred to as the "sixth" player in basketball, for example). Research has shown that favourable spectator influence appears to be the dominant factor explaining this phenomenon. [2].

In basketball in particular, researchers have quantified the home-court advantage (vs. a neutral court) in USA college basketball to be $4-5$ points [4], with limited differences between teams. This course project uses Bayesian techniques builds on the previous research and aims to analyse home-court advantage in the Lithuanian Basketball League. It borrows concepts and ideas from [3] with specific focus on estimating the size of the home court advantage. All associated code and data is available at Github repository.

2 Data

I have collected game scores, including home- and away- team information, indicator whether it is a game in a regular season or a cup tournament from the Lithuanian Basketball League website. The data was scraped on April 18, 2022, and contains all games played in this league since 1993 (refer to the code repository for the scraping code and the dataset itself). The full dataset comprises 5,133 games between 34 unique teams in 36 seasons (the annual King Mindaugas Cup (KMT), played since 2016 is recorded as separate seasons in the data). Importantly, the data structure allows identifying home- and away- teams, as well as track clubs over time.

Unlike in the US, all games are played on a club's court and thus there is no need to separately identify any games played on neutral courts. While I am sure that there may have been some instances of such games occurring due to technical reasons (e.g. an area being temporarily closed for repairs), I did not collect further information required to identify such games.

Basic exploratory analysis confirms the presence of the the home-court advantage in the data. The average difference between home-team score and away-team score across all games is 3.79 points, not too far from home court advantage estimated in the previous research. The presents of it is observed every year with exception of 2012 (see figure below), with the latest seasons exhibiting slightly lower levels of the home-court advantage than previously as indicated by the rolling 3-season average line.

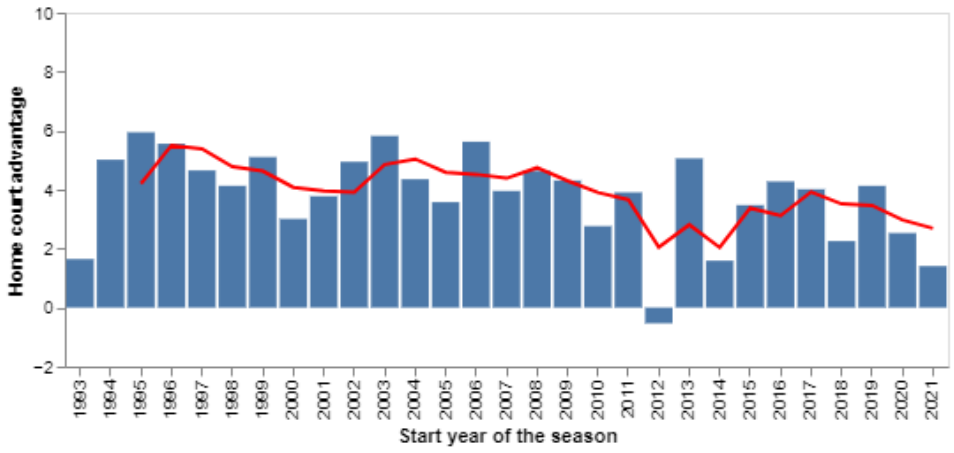

Figure 1: Average difference between home and away score by year

3 Modelling approach

I approach the estimation of home-court advantage by borrowing ideas from Brown and Sokol (2010) work on NCAA basketball predictions [3]. In particular, I model game score outcomes as pair-wise team strength differentials that are impacted by home-court advantage. Such an approach allows me to avoid modelling the individual latent team strengths which then would need to be converted to game scores requiring further assumptions.

I consider two main Bayesian model specifications: home-court advantage as an overall phenomenon and home-court advantage as a team-specific phenomenon. Previous research has found that team-to-team differences of the estimated home-court advantage are limited and thus the former model may be sufficient; however, I include the team-specific model to specifically test whether team-to-team differences are indeed limited.

The following notation is used throughout the document:

- $D^k_{i,j}$ - game k outcome between team i and team j
- $U_{i,j}$ - the unobservable strength differential between team i and j
- $h_{[i]}$ - home court advantage (of team i)
- c - incremental home-court advantage because the game is played in the KMT cup (vs. regular season)
- $\hat{\mu_{i,j}}$ - prior of strength differential between team i and j
- $\hat{h}$ - prior of home court advantage (of team i)

- X_i^k - indicator variable that is equal to 1 if game k is played at team i home arena.
- $Y_{i,j}^k$ - indicator variable that is equal to 1 if game k is played at KMT cup.

The overall logic of the below described models can be summarized in the following equation.

$$D_{i,j}^k = U_{i,j} + h_{[i]}X_i^k - h_{[j]}X_j^k + cX_i^kY_{i,j}^k - cX_j^kY_{i,j}^k$$

In a nutshell, I expect that the observed score differential has three underlying components: the real team strength differential, plus the home-court advantage of one of the playing teams (note that one of the X_i^k and X_j^k is always zero), plus an incremental home-court advantage in case the game is played in the KMT cup. The last variable is motivated by the fact that the KMT cup games are of higher stakes, and thus I would like to understand if the home-court advantage is larger in those situations.

3.1 Overall home-court advantage model

The overall home-court advantage model is defined as follows:

$$\alpha_{i,j} \sim HC^{1}(5) \quad \text{(hyper-prior for pair-wise difference variance)}$$

$$U_{i,j} \sim N(\hat{\mu_{i,j}}, \alpha_{i,j}) \quad \text{(pair-wise team strenghts)}$$

$$\beta \sim HC(5) \quad \text{(hyper-prior for home advantage variance)}$$

$$h = N(\hat{h}, \beta) \quad \text{(home court advantage)}$$

$$\gamma \sim HC(2) \quad \text{(hyper-prior for additional cup advantage variance)}$$

$$c \sim N(0, \gamma) \quad \text{(incremental home court advantage when played in the cup)}$$

$$S_{i,j}^k = U_{i,j} + hX_i^k - hX_j^k + cX_j^kY_{i,j}^k - cX_j^kY_{i,j}^k$$

$$\epsilon \sim HC(5) \quad \text{(hyper-prior for residual variance)}$$

$$D_{i,j}^k \sim N(S_{i,j}^k, \epsilon) \quad \text{(likelihood)}$$

For computational purposes, I also experiment with a model where β is fixed at value of 10 instead of being drawn from half-Cauchy distribution.

[1] Half-Cauchy distribution; I encountered computational issues with using Gamma distributions and switched to Half-Cauchy distributions as priors to variance across the models.

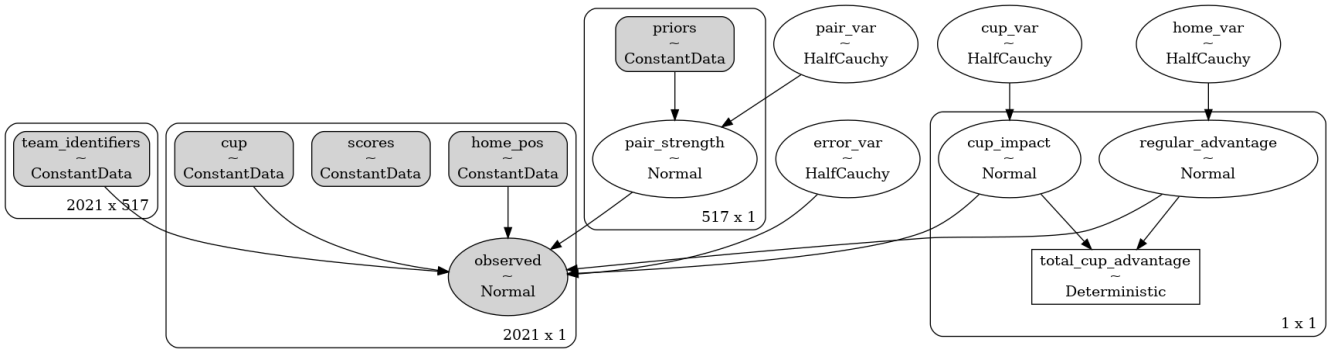

Figure 2: Single home-advantage model graph

3.2 Team-specific home-court advantage model

The team specific model is designed similarly to the previous model. Instead, of relying on a single home-court advantage, however, it uses it as a hyper-prior to draw team-specific advantages which are then incorporated into the scores:

$$\alpha_{i,j} \sim HC(5) \quad \text{(hyper-prior for pair-wise difference variance)}$$

$$U_{i,j} \sim N(\hat{\mu_{i,j}}, \alpha_{i,j}) \quad \text{(pair-wise team strenghts)}$$

$$\beta \sim HC(2) \quad \text{(hyper-prior for overall home advantage variance)}$$

$$h = N(\hat{h}, \beta) \quad \text{(overall home court advantage)}$$

$$\delta \sim HC(2) \quad \text{(hyper-prior for team-specific home advantage variance)}$$

$$h_{i,j} = N(h, \delta) \quad \text{(team-specific home court advantage)}$$

$$\gamma \sim HC(2) \quad \text{(hyper-prior for additional cup advantage variance)}$$

$$c \sim N(0, \gamma) \quad \text{(incremental home court advantage when played in the cup)}$$

$$S^k_{i,j} = U_{i,j} + h_{i,j}X^k_i - h_{i,j}X^k_j + cX^k_j Y^k_{i,j} - cX^k_j Y^k_{i,j}$$

$$\epsilon \sim HC(5) \quad \text{(hyper-prior for residual variance)}$$

$$D^k_{i,j} \sim N(S^k_{i,j}, \epsilon) \quad \text{(likelihood)}$$

Similarly to the previous model, I also experiment with a model where β is fixed at value of 10 instead of being drawn from half-Cauchy distribution.

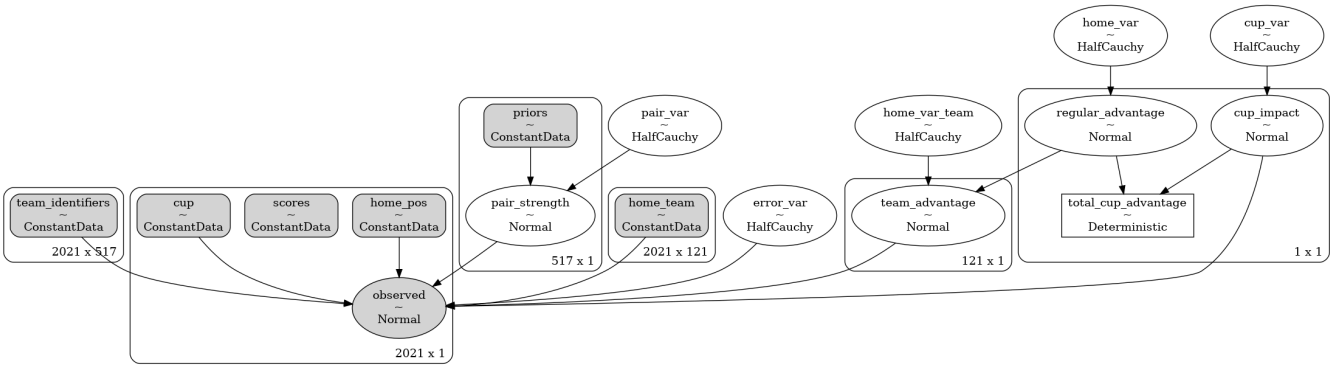

Figure 3: Team-specific home-advantage model graph

3.3 Prior selection

I experiment with uninformative priors for home-court advantage $\hat{h} = 0$ as well as $\hat{h} = 4$, given the range of home-court advantage found by previous research. Similarly, I experiment with uninformative priors for pair-wise team strengths $\hat{\mu}_{i,j} = 0$, and empirically informed priors, where $\hat{\mu}_{i,j}$ is estimated as the average score differential in the 2 seasons prior the season whose data is used for estimation.

3.4 Dataset sub-samples

Performing Bayesian analysis on the entire dataset at once was not computationally feasible. Instead, I chose two specific time periods: the latest complete season (starting in year 2020) that comprised 230 observations and 51 unique team pairs, and a longer time period spanning from 2012 to 2021 with 2021 observations in total with 517 unique team pairs.

Note that in the latter dataset, a team was treated as unique every season to avoid misrepresenting the team as the same entity across a long period of time during which its composition would have changed nearly 100%.

4 Results

4.1 Model selection

Previously described model selection, prior selection, and sub-sample variations resulted in fitting 24 models in total. First, I performed model comparison checks to identify any models that appears to be best fit to the data. I used a Leave-One-Out cross-validation approach with deviance as the selection criteria. The results are plotted below.

I observe that the differences between the models are minimal; in case of 10-season dataset, the best performing model used team-specific specification, with non-informative priors for the home advantage, however all other team-specific models perform worse than the simpler models. Among the models fit on 2020 season only, the best performing model was a single-advantage model that used empirical priors for pair-wise team strengths, non-informative priors for home-court advantage (and no-hyper prior for its variance). Importantly, I find that models with empirical priors for pair-wise strengths all performed better than any of the models that used non-informative priors.

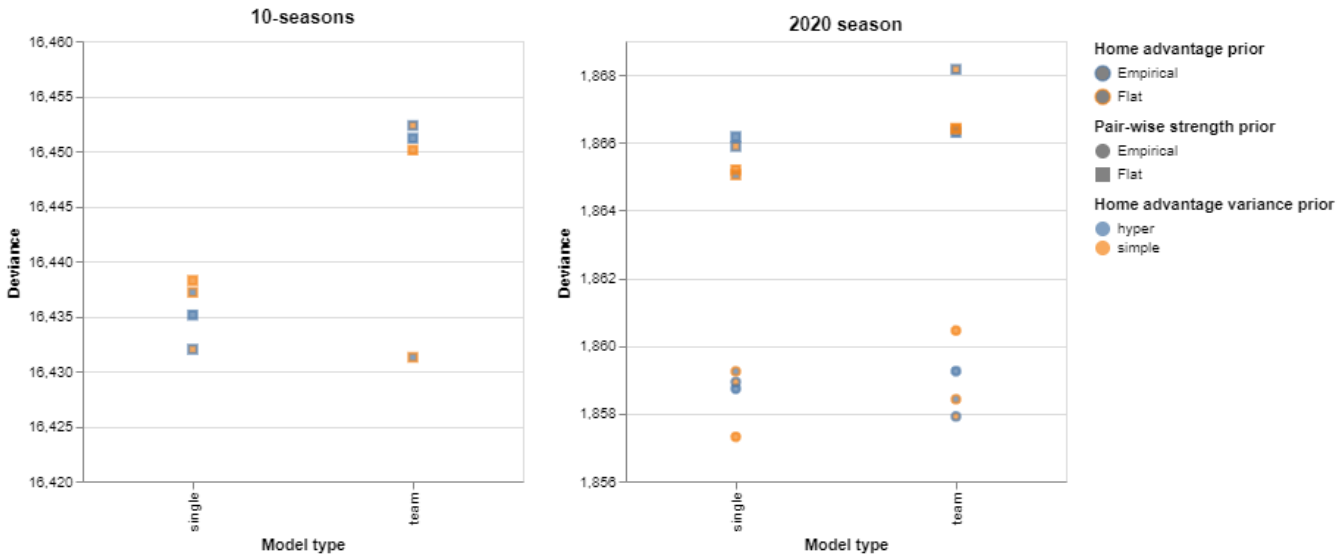

Figure 4: Model comparison using leave-one-out cross-validation based on deviance criteria

4.2 Model diagnostics

In Appendix A, I include posterior distribution plots and sampling chain plots of the two best performing models discussed above. While not ideal, the plots indicate no major issues with

the posterior distributions. During the fitting process, I observed some chain divergences as reported by pymc [5], a Python library used for modelling purposes.

4.3 Estimators of home-court advantage

Given limited differences in deviance across the models, it is unsurprising that they produced similar estimators, too. Models fit on the 10-year dataset arrived at the overall home-court advantage Bayes estimators with a mean ranging from 2.62 to 2.92, with 90% HDI credible sets not including zero in all cases. The model with the lowest deviance produced an estimator with a posterior mean of 2.626 and 90% HDI credible set $[1.76; 3.28]$.

Models fit on 2020 season only produced similar results, too. The overall home-court advantage posterior mean ranged from 1.705 to 3.149 (with models that used an informative home-court advantage prior tending towards higher values). All but one model specifications led to 90% HDI credible sets not including zero. The model with the lowest deviance produced and estimator with a posterior mean of 2.3 and 90% HDI credible set of $[0.945; 3.641]$.

Further, I find that playing in KMT cup does not confer an additional home-court advantage. None of the models produced an 90% HDI credible set outside zero. Interestingly, the models fit on the 10-year dataset all resulted in posterior mean of cup impact estimator being negative. To ensure that this did not negate the home-court advantage in these games, I also estimated the probability of the total home-court advantage (incl. cup impact) in these models. The posterior probability was at least 0.994 across all models, confirming that home-court advantage remains in KMT games, but may actually be smaller than in the regular season.

4.4 Team-specific home-court advantages

Finally, I explore whether team-specific home-court advantage estimators provide additional insights. The below analysis is based on the the best-performing model (based on deviance criteria) for the 10-year dataset. As discussed above, this model arrived at overall home-court advantage estimator with a posterior mean of 2.626. The posterior means of team-specific home-court advantages spanned from -2.7 points to 8.33 points, with an average of 2.66. Out of 121 unique teams captured in the dataset, 34 teams had team-specific home-court advantages that did not include 0 in the 90% HDI credible set, and all those advantages were positive (with a minimum posterior mean advantage of 3.09).

5 Discussion and further work

In this project, I applied Bayesian methods to estimate home-court advantage in the Lithuanian Basketball League. In line with prior research, I find that home-court advantage exists, and its overall magnitude is somewhat smaller than what has been identified in the research focused on US basketball (c. 2.3 - 2.6 points vs. above 4 points identified in the US). Intuitively, this makes sense given smaller arena sizes in Lithuania and thus, arguably, smaller spectator impact, as well smaller distances that the opponent teams need to travel compared to the US.

I also explored whether team-specific models perform better, with inconclusive results. It appears that the models that allow for team-specific home-court advantages perform similarly to the ones that treat home-court advantage as a uniform phenomenon. At the same time, such models do not necessarily perform worse, and provide further insights into team-level dynamics, where it appears that only a fraction of the teams benefit from a consistent home-court advantage.

It would be interesting to extend this analysis further. Specifically, it would be interesting to model home-court advantage over time and confirm whether it is actually diminishing as the exploratory analysis indicated. Further, the models built on the 10-year dataset currently treated the same club as different entity every season to account for varying team strength. Arguably, a more flexible model should allow for varying team strength while keeping the home-court advantage fixed to the club - this may be a better specification given that the home-court advantage is associated with factors that are linked to the club itself and not its particular team composition in a given year.

References

[1] John Berdell, James Ciecka, and Anthony C Krautmann. "There's no place like home: home court advantage in North American sports leagues". In: *International Journal of Sport Management and Marketing* 14.1-4 (2013), pp. 133–145.

[2] Christopher J Boudreaux, Shane D Sanders, and Bhavneet Walia. "A natural experiment to determine the crowd effect upon home court advantage". In: *Journal of Sports Economics* 18.7 (2017), pp. 737–749.

[3] Mark Brown and Joel Sokol. "An improved LRMC method for NCAA basketball prediction". In: *Journal of Quantitative Analysis in Sports* 6.3 (2010).

[4] David A Harville and Michael H Smith. "The home-court advantage: How large is it, and does it vary from team to team?" In: *The American Statistician* 48.1 (1994), pp. 22–28.

[5] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. "Probabilistic programming in Python using PyMC3". In: *PeerJ Computer Science* 2 (2016), e55.

6 Appendix A

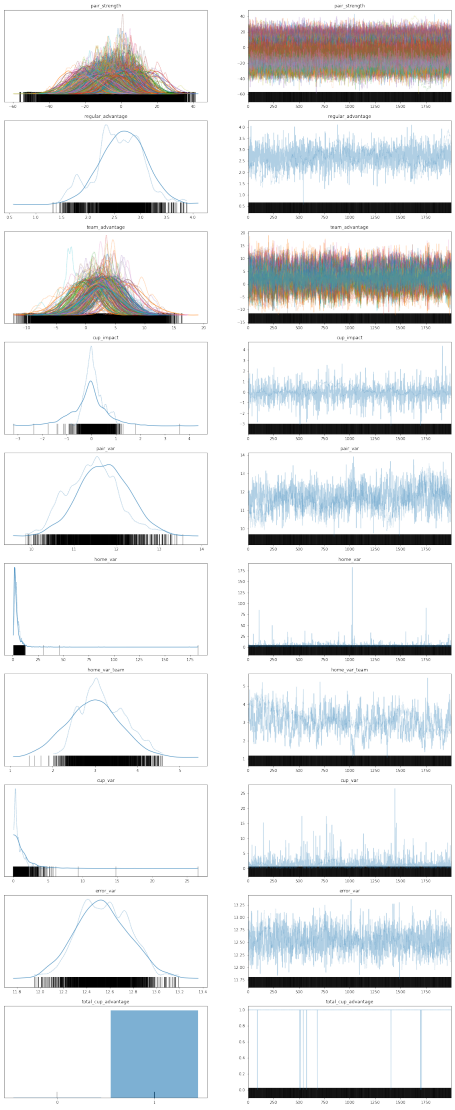

Figure 5: Trace diagnostics - 10-year model

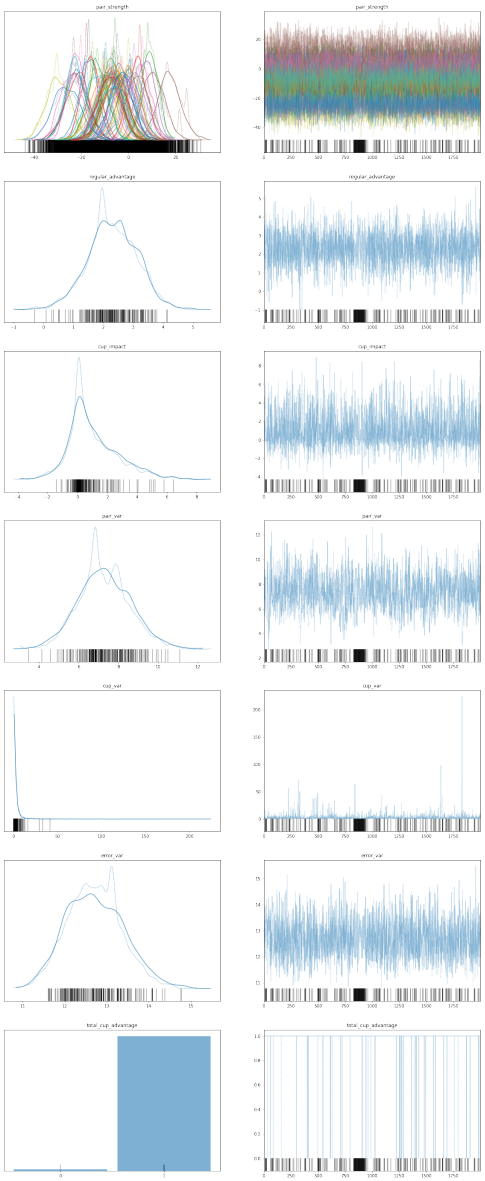

Figure 6: Trace diagnostics - 1-year model