Hierarchical Bradley-Terry Models: Extension to Incorporate Neutral Venue Scenario

Thesis Report - MSc Statistics

Varun Ranganath

Supervisor: Alan Huang

June 2023

Contents

List of Figures

List of Tables

Chapter 1

Literature Review

1.1 Introduction

In 1929, Zermelo [36] proposed to solve the problem of ranking players in chess tournaments. In instances where players might have matched up with weaker (or stronger) opponents initially, and the tournaments are canceled, Zermelo approached the situation by using relative player strengths as probabilities. In 1952, this idea was popularized by Ralph Allen Bradley and Milton E. Terry [9], giving the name to the model.

The Bradley-Terry model has been popular, facilitating pairwise comparisons, not only in sports like Zermelo initially used it for, but also in various other fields. This review of literature traces the journey of these models from their inception in 1952 until today, where programming software is being used to build these models with all the new extensions that have been added to the initial model proposed by Bradley and Terry.

As mentioned, although Zermelo initially used it to facilitate player rankings in chess tournaments, the model has been used in various fields, where pairwise comparison is necessary. Matthews and Morris have used an extension of the Bradley-Terry models to the measurement of pain [25]. Dittrich et al have used these models to the preference of universities by using rankings [30]. Stuart-Fox et al have used a version of these models to design and

analyze animal contests [34]. A closer application to our paper is the example given in the book by Alan Agresti [1], exploring home-ground advantage in Major League Baseball(MLB).

There have been many attempts to improve the initial Bradley-Terry model proposed by Bradley and Terry, and some of these attempts are successful. Given our paper is modeling sporting outcomes, there are extensions such as hierarchy and home-ground advantage, which are akin to the sporting world. As we move ahead, we will look to review these extensions that have been explored in the past and explore their relevance to our current paper.

1.2 Inception

In the 1950s, Bradley and Terry were authors of a 3-part paper that involved rank analysis through pairwise comparison. The first paper was the work of Bradley and Terry, but the second and third were published by Bradley alone, extending the work Terry and himself had done.

The model described in the 1952 article [9] is as follows:
Consider t treatments in an experiment involving paired comparisons. We shall consider that these treatments have ratings (or preferences) $\pi_1, \pi_2, ...\pi_t$ which are considered throughout an experiment. We can pose restrictions on these π_i values by saying that every $\pi_i \geq 0$ and $\Sigma\pi_i = 1$

The model proposes that the probability for a treatment i to obtain a better rating than treatment j is given by

$$\mathbb{P}(i > j) = \frac{\pi_i}{\pi_i + \pi_j}$$

A new variable r_{ijk} is created which indicates the rank of the i^{th} treatment on k^{th} repetition against the j^{th} treatment. When the i^{th} treatment wins

against j^{th} treatment, we get $r_{ijk} = 1$ and $r_{jik} = 3 - r_{ijk} = 2$

As a result, the likelihood function looks as follows:

$$L(\pi_i, \pi_j | r_{ijk}) = (\frac{\pi_i}{\pi_i + \pi_j})^{2-r_{ijk}} \ (\frac{\pi_j}{\pi_i + \pi_j})^{2-r_{jik}}$$

After the likelihood function is generalized for all expressions, it is seen

$$L(\pi|r) = \prod_i \pi_i^{2n(t-1)-\sum_{j\neq i}\sum_k r_{ijk}} \prod_{i<j} (\pi_i + \pi_j)^{-n}$$

Bradley, in the year 1954, in his paper [7] proposes various kinds of hypothesis tests to the model built in his previous paper. These hypothesis tests are surrounded by a certain statistic called the 'B' statistic, along with a likelihood ratio test measure. These 'B' statistics are primarily used to test treatment equality among paired comparisons and are also used to test the appropriateness of the model as a whole. In the third installment of the papers on rank analysis [8], Bradley gives an account of examining the tests of significance he had developed in his previous papers. He also examines the variance and covariance of the estimators as well as the power of test conducted using the model.

As years have gone by, the notation of the initial Bradley-Terry models has taken different forms. Jones, in an honors thesis [23], notes that the usage of the variable r_{ijk} has slowly disappeared from the literature over the years. The variable has been replaced by w_{ij}, which indicates the number of times i has been preferred over j, which leads to a very simple likelihood function

$$L(\pi|w) = \prod_i \prod_{j<i} (\frac{\pi_i}{\pi_i + \pi_j})^{w_{ij}} (\frac{\pi_j}{\pi_i + \pi_j})^{w_{ji}}$$

As we move ahead with the paper, we will be using a very similar representation of the likelihood function.

1.3 Modelling Ties

As the central focus of this paper is to model sporting outcomes, ties are almost a regular result in most sports. Although there are tie-breakers, there are some sports like chess, soccer, and others where a draw(tie) is a valid result for a game. Although Bradley and Terry developed their models based on Zermelo's attempt to model chess outcomes and rankings, their final model did not have a provision to estimate and predict ties as an outcome. As a development, many researchers tried to extend the Bradley-Terry models to be able to model ties. Some of them were Glenn and David [22], Rao and Kupper [31], Davidson [15], and Kousgaard [24].

Glenn and David [22] in the year 1960, tried to improve the Thurstone-Mosteller model by trying to incorporate ties. They introduce a new threshold value τ and state that if the difference between the two responses lies in an interval between $-\tau$ and τ, the judge will declare a tie. This was further improved by Rao and Kupper [31] in the year 1967, by the introduction of another threshold parameter η, which can be considered analogous to the τ parameter introduced by Glenn and David.

Rao and Kupper state that if a parameter η exists such that the differences in the logs of levels of two treatments are less than η, the judge will not distinguish between the two treatments and declare a tie. Representing this in an equation, Rao and Kupper mention that if we consider $\theta = exp(\eta)$ and true treatment ratings $\pi_1, \pi_2, ...\pi_t$, we have:

$$\mathbb{P}(i > j) = \frac{\pi_i}{\pi_i + \theta\pi_j}$$

$$\mathbb{P}(i < j) = \frac{\pi_i}{\theta\pi_i + \pi_j}$$

$$\mathbb{P}(i = j) = \frac{\pi_i\pi_j(\theta^2 - 1)}{(\pi_i + \theta\pi_j)(\theta\pi_i + \pi_j)}$$

This model becomes the Bradley-Terry model when $\eta = 0$, implying $\theta = 1$. After obtaining the equation, the paper continues with methods for obtaining maximum likelihood estimators for the parameters of the modified model.

Rao and Kupper's proposition was further extended by Davidson [15] in the year 1970, where he proposed that the probability of no preference is proportional to the geometric mean of the probabilities of preferring the treatments, which looks like:

$$\mathbb{P}(i = j) = \nu\sqrt{\mathbb{P}(i < j)\mathbb{P}(i > j)}$$

where $\nu \geq 0$ is a constant of proportionality, which does not depend on the treatments. The extended model looks like:

$$\mathbb{P}(i > j) = \frac{\pi_i}{\pi_i + \pi_j + \nu\sqrt{\pi_i\pi_j}}$$

$$\mathbb{P}(i = j) = \frac{\nu\sqrt{\pi_i\pi_j}}{\pi_i + \pi_j + \nu\sqrt{\pi_i\pi_j}}$$

This model turns into a classic Bradley-Terry model when $\nu = 0$

Apart from these extensions, one could also look at implementing ties as a very simple instance of half a win, which will reduce our power to predict ties but will be able to use draws as an outcome in the process of building the model. As introduced earlier, for two treatments i and j, we could record $w_{ij} = \frac{1}{2}$. In our current research, we have used this extension of recording a draw as half a win for both teams.

1.4 Modelling Order Effects

Gymnastics, the sport, involves a set of judges evaluating the performance of gymnasts based on a set of criteria. The judges will have to evaluate individual performances as the gymnasts perform in a pre-determined order. Given the judges have not seen all of the performances, the scores they provide

might depend on the order in which they see the individuals perform. In sporting competitions like diving and gymnastics where judges are involved, the order in which athletes perform will have an effect on their probability of winning. Aside from sports, a cooking show judge might provide a different score to a participant based on the order in which the participants present their dishes. Bradley-Terry models have not provided any provisions for such orders.

'Home Ground Advantage' is highly talked about in sports. In Major League Baseball, the home team will always have to bat in the bottom half of the inning. There is a preconceived notion that a team is expected to perform better in home conditions than away, due to factors like the familiarity of the venue, the fans, and others. It would be interesting to see whether home ground has an influence on the team's strengths, which would make the measurement of their true skill level more comprehensive.

One of the most important papers to extend the Bradley-Terry model to include order effects was written by Davidson and Beaver [16]. A multiplicative order effect was suggested to be incorporated in the models, unlike the additive order effect, suggested by Beaver and Gokhale [5].

In equation form, Beaver and Gokhale's model looks like this:

$$\mathbb{P}(i > j) = \frac{\pi_i + \delta_{ij}}{\pi_i + \pi_j}$$

$$\mathbb{P}(i < j) = \frac{\pi_i - \delta_{ij}}{\pi_i + \pi_j}$$

where δ_{ij} is a parameter associated with the pair (i, j) and has the restriction $|\delta_{ij}| \leq min(\pi_i, \pi_j)$. This model is a Bradley-Terry model if $\delta_{ij} = 0$

Davidson and Beaver write that the values ln π_1 ..... ln π_i can be represented on a linear scale. Thus, the logarithm of the worths could also be assumed to be affected additively by the order of presentation. By this, they have also assumed that they can see a multiplicative within-pair order effect γ_{ij} when objects i and j appear together in a pair.

In equation form, Davidson and Beaver's model looks like this:

$$\mathbb{P}(i > j) = \frac{\pi_i}{\pi_i + \gamma_{ij}\pi_j}$$

$$\mathbb{P}(i < j) = \frac{\gamma_{ij}\pi_j}{\pi_i + \gamma_{ij}\pi_j}$$

where $\gamma_{ij} \geq 0$. When $\gamma_{ij} = 1$, there is no order effect, turning into a Bradley-Terry model. When $\gamma_{ij} > 1$, object j, being presented second, will have its worth greater than the object presented first, i. When $\gamma_{ij} < 1$, object j, being presented second, will have its worth lesser than the object presented first, i. The paper also mentions that the case of $\gamma_{ij} = \gamma$ for all (i, j) is important as it significantly reduces the number of parameters in the model. Davidson and Beaver in [16] also assume that $\gamma_{ij} = \gamma_{ji}$.

In a sporting context, we do not have to agree with this assumption as this might not be true in several cases. Paine in 2013 [29] mentioned in his article that Utah and Denver NBA franchises might have a slightly higher home advantage than other franchises due to their higher altitude home venues.

This multiplicative order effect will be used throughout the paper to model home-ground advantage across different sporting competitions.

1.5 Shift to Log-Linear Parameterization

A shift to a log-linear parameterization of the Bradley-Terry model was one of the most important advancements of the model yet. Jones [23] mentions that this shift has enabled us to apply the findings from logistic regression to this model, enabling us to use a more general and well-researched framework.

From the initial stages of the development of the Bradley-Terry model in 1952, there are mentions in the first paper [9] about comparing all the $\log\pi_i$s on a linear scale. Bradley and Terry believed that this would facilitate overall comparisons of the experimental treatments. Despite this, there were very

few attempts to provide a complete log-linear parameterization of the model.

There are mentions of log-linear parameterization in the early 1970s by Cox [13], Atkinson [3], Maxwell [26], Fienberg and Larntz [18], and many others. Atkinson describes the Bradley-Terry, in a very similar manner as Cox [13]. Atkinson states that the Bradley-Terry model assumes that the treatments can be ranked in one dimension with a parameter ρ_i and treatment rating π_{ij} depends on the value of $\rho_i - \rho_j$. The model looks like this:

$$\lambda_{ij} = log(\frac{\pi_{ij}}{1-\pi_{ij}}) \ = \rho_i - \rho_j$$

A similar version of this model will be used in our paper, except with a small change in the notation where θ is used instead of ρ.

Jones [23] mentions that with log-linear parameterization, some of the introduced concepts like order effects can be easily extended onto the model as linear terms. For example, when considering the case of $\gamma_{ij} = \gamma$ as a common home ground advantage, we can just add on the term $\alpha = ln\gamma$ in the model, which comes out as:

$$log(\frac{p_{ij}}{1-p_{ij}}) \ = \theta_i - \theta_j + \alpha$$

where $\theta_i = ln\pi_i$ and p_{ij} indicates the probability that team i beats j at team i's home

1.6 Modelling External Covariates

In 1991, Critchlow and Fligner discussed the topic of external covariates in Bradley-Terry models in their paper [14]. In their paper, Bradley-Terry models are represented analogously to the conventional linear model using the GLM representation of the model. We have

$$\theta_i = \sum_{j=1}^{p} c_{ij}\beta_j \qquad for\ i = 1, 2.....t$$

where $\beta_1, ..., \beta_p, p \leq t$ are unknown parameters and c_{ij} is the value of the j^{th} covariate. They also mention that c_{ij} may take the role of an indicator variable. In the paper, these models are called 'constrained models' with covariates.

In this paper, Critchlow and Fligner mention using the GLIM statistical package to fit these models. GLIM is a statistical package developed by the Royal Statistical Society's Working Party on Statistical Computing [2]. In the paper, while introducing the package, they mention that maximum likelihood estimates are obtained by GLIM using an iteratively reweighted least squares approach, and the software also offers likelihood ratio test statistics for relevant hypotheses.

There have been instances of usage of covariates in Bradley-Terry models since Critchlow and Fligner. In 1998, Dittrich et al used Bradley-Terry models to model the effect of subject-specific covariates with an application to university rankings [30]. When choosing preferences, the subjects in the study might have factors influencing their preference for a university like their gender or their ability to speak foreign languages. These covariates are added to the log-linear version of the Bradley-Terry model as categorical covariates, as above in the case of the Critchlow and Fligner paper.

In 2006, Giambona et al used the Bradley-Terry models to assess the factors of attractiveness across territorial areas in Italy [21]. In this paper, Giambona et al try to explore territorial factors and university factors as independent covariates to assess attractiveness towards territorial areas.

In a sporting context, in a sport like Chess, where all genders are entering the same competition, we could explore gender as an external covariate. Chabris and Glickman conducted a study on the sex difference between chess players by comparing the rating of men and women over a period of 13 years [12]. Analyzing this in a Bradley-Terry model context with gender as an external covariate could be something that could be explored in the future.

Cricket is a sport played almost exclusively in the summer in both England and Australia. The reason behind that is that cricket games can be interrupted by rain as rain will alter the condition of the pitch and the ground, making the conditions unfavorable for the game to continue. In case of a

long rain interruption during a limited overs[1] game, the number of overs to be bowled will be reduced, and the number of runs scored will be appropriately scaled. The method used to execute the scaling is called the Duckworth-Lewis-Stern method [33]. In the future, this could be considered an external categorical covariate, where the Duckworth-Lewis-Stern method being used or not could be a covariate.

1.7 Current Software

We aim to create a package on R to suit our needs with regard to Bradley-Terry models. Before we proceed further, it is important to explore and understand the current state of software and what the software will be able to obtain according to our needs. There were two packages on R to model data in line with the Bradley-Terry models:

- BradleyTerry by David Firth in 2005 [20].
- BradleyTerry2 by David Firth and Heather Turner in 2012 [35].

The BradleyTerry package, built in 2005 [20], formulates the Bradley-Terry models by setting the first log-skill level $\theta_1 = 0$. The other θ_is are estimated by using maximum likelihood estimation. The reference θ_i can be changed accordingly. The package is able to model linear predictors. As mentioned in 1.4, this package can add the order effect to the Bradley-Terry model. In a sporting context, the order effect can be seen as a home-ground advantage (or disadvantage) indicated by 1 (or -1) in a pairwise comparison. The package produces outputs of the log-skill levels along with important statistics like standard errors, Z scores, p-value, and residuals. This package can obtain two kinds of residuals: contest-wise and player residuals. Apart from the default MLE[2] fitting, the package can also fit a bias-reduced maximum likelihood, as mentioned by Firth [19].

[1]Limited overs games in cricket are games where the number of overs to be bowled is restricted. Currently, two internationally conducted formats of limited overs are 50 overs and 20 overs.

[2]Maximum Likelihood Estimation

The BradleyTerry2 package, built in 2012 [35], was an upgrade to the previous BradleyTerry package. The package allows for the fitting of Bradley-Terry models using the logit, cauchit, and probit formulation, an improvement from only logit in the previous package. Inclusion of the prediction error U_i looks like $\lambda_i = \sum_{r=1}^{p} \beta_r X_{ir} + U_i$ where $U_i \sim N(0, \sigma^2)$ is another feature of the new package. Contest-specific and player-specific predictors are also added to the package, with a few new contest-specific predictors capable of adding factors to the model. With player-specific predictors, the Bradley-Terry model becomes a GLM[3], which makes the package fit the model using a quasi-likelihood algorithm by Breslow and Clayton [10]

1.8 Research

The discussion of order effects as in 1.4 is our primary research of interest. The previous packages of R were able to measure order effects (For example Home-ground advantage) for only one level. We would like to extend this across multiple levels and explore this in different sporting competitions. The two competitions of interest are the NBA[4] and T20 Blast. The two cases mentioned in Davidson and Beaver [16] of $\gamma_{ij} = \gamma$ for all (i, j) and $\gamma_{ij} = \gamma_{ji}$ need not hold true and we would like to look at this using match data from these competitions.

With CoViD-19 rocking the world over the past two years, sporting tournaments across the world were affected by the deadly virus. Many tournaments were interrupted halfway, putting many teams in a situation similar to the scenario Zermelo explained in his paper [36]. The tournament committees decided to build bio-secure bubbles in a different country or a state and conduct all the games in these bio-secure bubbles. These bio-secure bubbles led to teams having no particular home-ground advantage as almost all games were played in neutral venues, snatching the home advantage (or disadvantage). Incorporating the neutral venue effect in modeling the strengths of

[3]Generalized Linear Model

[4]National Basketball Association

teams with order effects is another focal point of the thesis.

One of the special features of North American sports leagues is the hierarchical structure involved in the competition. The MLB[5] has 30 teams equally split into two leagues, 'American League' and 'National League'. These leagues are further split into 3 divisions with 5 teams in each division. The NFL[6] has 32 teams equally split into two conferences, 'National Football', and 'American Football' Conference. These conferences are further split into 4 divisions with 4 teams in each division. The creation of the conferences and leagues in both MLB and NFL are based on some small changes in rules and groupings that have been historically the same within the competition. While the American League in MLB has a 'Designated Hitter', the National League requires the pitcher to also hit during games [27]. NFL groupings have been historical and have followed the same structure since 1970.

The NBA has 30 teams equally split into two conferences, 'Eastern' and 'Western' Conference. These conferences are further split into 3 divisions of 5 teams each. The NHL[7] has 32 teams equally split into two conferences, 'Eastern' and 'Western' Conference. These conferences are further split into 2 divisions of 8 teams each. As seen above, both the NBA and NHL have groups based on their geographic location. Moving onto another sporting league, T20 Blast is a T20[8] cricket league in the United Kingdom. This league has 18 county cricket teams, split into two divisions, 'North' and 'South' based on their geographic location.

Since NBA and T20 Blast are our chosen competitions, an example to understand their hierarchy could make things easier. The Golden State Warriors, the current NBA champions (2021-22 season) are a team in the Western Conference, playing in the Pacific division. Hampshire Hawks, the current T20 Blast champions (2022 season) are a team in the South Group.

We would like to understand if these structures in the leagues might bear a difference in the degree of home-ground advantage teams might have when

[5]Major League Baseball
[6]National Football League
[7]National Hockey League
[8]A cricket competition where each team gets to bat for a maximum of 20 overs

facing each other. In the NBA, we would like to explore the level of home-ground advantage when team i faces team j at home when

- Advantage Team i has if Team j is from the same division.
- Advantage Team i has if Team j is from the same conference but a different division.
- Advantage Team i has if Team j is from a different conference

In the T20 Blast, we would like to explore the level of home-ground advantage when team i faces team j at home when

- Advantage Team i has if Team j is from the same group.
- Advantage Team i has if Team j is from a different group.

Along with these, we would also like to explore other levels of home-ground advantages, like the ones mentioned in Davidson and Beaver [16], where $\gamma_{ij} = \gamma$ and the ones where this does not hold true. We would also like to perform hypothesis tests for all the models built and check whether a more complex model with more order effects is fitting the data better or not. We use Likelihood Ratio tests to perform hypothesis testing.

We use the best-fit model for the NBA data and run simulations for the ongoing NBA playoffs (2023 season) and obtain probabilities to predict the championship winner. As there are no ties in NBA[9] and no ties in most competitions in cricket[10], we will not predict ties as an outcome.

[9] 5 minutes of additional time - known as overtime will be played until a result is obtained

[10] A super over i.e. one additional over to each team decides the winner of the game. In the case of a tied super-over, the team that hit the highest number of boundaries in the regular game was declared the winner. But, since 2020, in the tied scenario, an unlimited number of super-overs are available to teams to decide the winner.

Chapter 2

Model Building

After reviewing all the literature on the Bradley-Terry model from its inception until now, we have found that order effects, explored as 'Home Ground Advantage' in a sporting context, would require a further extension. In this chapter, we will introduce all the models that have been built to explore the order effects. Firstly, we introduce all the important notation that is going to be used in the model. Secondly, we would provide basic information about all the models individually, showing the log-likelihood function and its derivative. This derivative, as we all know, will help in minimizing the log-likelihood function with lesser computation time. Finally, we introduce the concept of hypothesis testing of all these models and some theory relevant to hypothesis testing.

2.1 Notation

Some common notation that has been used in the models are in the list

- H - The matrix which depicts the number of home games of the teams, where H_{ij} shows the number of times team i has played a home game against team j.
- W - The matrix which depicts the number of home wins of the teams,

where W_{ij} shows the number of times team i has won a home game against team j.

- L - The matrix which depicts the number of home losses of the teams, where L_{ij} shows the number of times team i has lost a home game against team j.
- NW - The matrix which depicts the number of neutral venue wins of the teams, where NW_{ij} shows the number of times team i has won a neutral venue game against team j.
- NL - The matrix which depicts the number of neutral venue losses of the teams, where NL_{ij} shows the number of times team i has lost a neutral venue game against team j.
- N - Number of teams in the tournament
- R - The matrix which depicts the relationship between two teams, where R_{ij} shows the level of the relationship between team i and team j, in the hierarchy of the tournament.
- n - The number of levels in the hierarchy of a tournament
- $(i > j)$ - The event that team i beats j.
- $(i > j)_i$ - The event that team i beats j at home.
- λ_i - The 'skill level' of team i
- θ_i - The 'log-skill level' of team i, obtained by $\theta_i = \ln \lambda_i$
- γ_k - The home-ground advantage of teams. For all the models that are going to be implemented in the paper, the home-ground advantage is different across different models
- α_k - The log home-ground advantage of teams, obtained by $\alpha_k = \ln \gamma_k$

2.2 Models

We have formulated and implemented a wide variety of extensions of Bradley-Terry models, in line with expanding the order effects. In this section, we will show all the models we have implemented in a format like below:

- Number of Parameters
- Representation of $\mathbb{P}$(i > j)
- Likelihood Function
- Log-Likelihood Function
- Gradient to the log-likelihood function

2.2.1 Vanilla Bradley-Terry Model

- **Number of Parameters**: The model has N parameters, where we try to obtain the parameter value for each team.
- **Representation of $\mathbb{P}$(i > j)**: The initial equation of the model looks like this:

$$\mathbb{P}(i > j) = \frac{\lambda_i}{\lambda_i + \lambda_j}$$

- **Likelihood Function**: The likelihood function of the model is:

$$L(\lambda) = \prod_{i=1} \prod_{j=1} (\frac{\lambda_i}{\lambda_i + \lambda_j})^{W_{ij}} \ (\frac{\lambda_i}{\lambda_i + \lambda_j})^{L_{ji}} \ (\frac{\lambda_i}{\lambda_i + \lambda_j})^{NW_{ij}}$$

- **Log-likelihood Function**: Instead of optimizing the likelihood function, the log-likelihood function is used, making the process more efficient. Thus, the log-likelihood function looks as follows:

$$l(\theta) = \sum_{i=1}^{N}\sum_{j=1}^{N}(W_{ij} + L_{ji} + NW_{ij})(\theta_i - log(e^{\theta_i} + e^{\theta_j}))$$

- **Gradient to the log-likelihood Function**: Considering the partial derivative with respect to θ_i for the gradient, the result is in the form of

$$\frac{\partial l}{\partial \theta_i} = \sum_{j=1}^{N}(W_{ij} + L_{ji} + NW_{ij}) - (H_{ij} + H_{ji} + NW_{ij} + NW_{ji})(\frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}})$$

2.2.2 Common Home-ground Advantage Model

- **Number of Parameters**: The model has $N + 1$ parameters, where we try to obtain the parameter value for each team and the common home-ground advantage parameter γ. We assume that the γ parameter remains constant throughout the period chosen for analysis. This represents the average home-ground advantage of teams.
- **Representation of $\mathbb{P}(i > j)$**: The initial equation of the model looks like this:

$$\mathbb{P}(i > j)_i = \frac{\lambda_i\gamma}{\lambda_i\gamma + \lambda_j}$$

$$\mathbb{P}(i < j)_i = \frac{\lambda_j}{\lambda_i\gamma + \lambda_j}$$

- **Likelihood Function**: The likelihood function of the model is:

$$L(\lambda, \gamma) = \prod_{i=1}\prod_{j=1}(\frac{\lambda_i\gamma}{\lambda_i\gamma + \lambda_j})^{W_{ij}}\ (\frac{\lambda_i}{\lambda_i + \lambda_j\gamma})^{L_{ji}}\ (\frac{\lambda_i}{\lambda_i + \lambda_j})^{NW_{ij}}$$

- **Log-likelihood Function**: Instead of optimizing the likelihood function, the log-likelihood function is used, making the process more efficient. Thus, the log-likelihood function looks as follows:

$$l(\theta,\alpha) = \sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij}(\theta_i+\alpha-log(e^{\theta_i+\alpha}+e^{\theta_j}))+L_{ji}(\theta_i-log(e^{\theta_i}+e^{\theta_j+\alpha})) + NW_{ij}(\theta_i - log(e^{\theta_i}+e^{\theta_j}))$$

- **Gradient to the log-likelihood Function**: Considering the partial derivative with respect to θ_i for the gradient, the result is in the form of

$$\frac{\partial l}{\partial \theta_i} = \sum_{j=1}^{N}(W_{ij}+L_{ji}+NW_{ij}) - H_{ij}\frac{e^{\theta_i+\alpha}}{e^{\theta_i+\alpha}+e^{\theta_j}} - H_{ji}\frac{e^{\theta_i}}{e^{\theta_i}+e^{\theta_j+\alpha}} - (NW_{ij}+NW_{ji})\frac{e^{\theta_i}}{e^{\theta_i}+e^{\theta_j}}$$

and the partial derivative with respect to α is in the form

$$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij} - H_{ij}\frac{e^{\theta_i+\alpha}}{e^{\theta_i+\alpha}+e^{\theta_j}}$$

2.2.3 Common Hierarchical Home-ground Advantage Model

- **Number of Parameters**: The model has N+n parameters, where we try to obtain the parameter λs value for each team and the n home-ground advantage parameter γs. γ_n represents the home-ground advantage at each level of the hierarchy. As mentioned above, R_{ij} represents the level of relationship in the hierarchy between team i

and team j, which brings us to $\gamma_{R_{ij}}$ as the home-ground advantage of team i against team j, based on their hierarchical relationship. Here, $(i, j) = 1, 2, ...N$. We assume that the $\gamma_{R_{ij}}$ parameter remains constant throughout the period chosen for analysis. This represents the average common hierarchical home-ground advantage of teams.

- **Representation of $\mathbb{P}(\mathbf{i > j})$**: The initial equation of the model looks like this:

$$\mathbb{P}(i > j)_i = \frac{\lambda_i \gamma_{R_{ij}}}{\lambda_i \gamma_{R_{ij}} + \lambda_j}$$

$$\mathbb{P}(i < j)_i = \frac{\lambda_j}{\lambda_i \gamma_{R_{ij}} + \lambda_j}$$

- **Likelihood Function**: The likelihood function of the model is:

$$L(\lambda, \gamma) \;=\; \prod_{i=1}\prod_{j=1} \left(\frac{\lambda_i \gamma_{R_{ij}}}{\lambda_i \gamma_{R_{ij}} + \lambda_j}\right)^{W_{ij}} \;\; \left(\frac{\lambda_i}{\lambda_i + \lambda_j \gamma_{R_{ij}}}\right)^{L_{ji}} \left(\frac{\lambda_i}{\lambda_i + \lambda_j}\right)^{NW_{ij}}$$

- **Log-likelihood Function**: Instead of optimizing the likelihood function, the log-likelihood function is used, making the process more efficient. Thus, the log-likelihood function looks as follows:

$$l(\theta, \alpha) = \sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij}(\theta_i + \alpha_{R_{ij}} - log(e^{\theta_i + \alpha_{R_{ij}}} + e^{\theta_j})) + L_{ji}(\theta_i - log(e^{\theta_i} + e^{\theta_j + \alpha_{R_{ij}}})) + NW_{ij}(\theta_i - log(e^{\theta_i} + e^{\theta_j}))$$

- **Gradient to the log-likelihood Function**: Considering the partial derivative with respect to θ_i for the gradient, the result is in the form of

$$\frac{\partial l}{\partial \theta_i} = \sum_{j=1}^{N}(W_{ij} + L_{ji} + NW_{ij}) - H_{ij}\frac{e^{\theta_i+\alpha_{R_{ij}}}}{e^{\theta_i+\alpha_{R_{ij}}} + e^{\theta_j}} - H_{ji}\frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j+\alpha_{R_{ij}}}}$$
$$- (NW_{ij} + NW_{ji})\frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}}$$

and the partial derivative with respect to α_n is in the form

$$\frac{\partial l}{\partial \alpha_n} = \sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij} - H_{ij}\frac{e^{\theta_i+\alpha_{R_{ij}}}}{e^{\theta_i+\alpha_{R_{ij}}} + e^{\theta_j}} \qquad where\, n = R_{ij}$$

2.2.4 Team-specific Home-ground Advantage Model

- **Number of Parameters**: The model has $2N$ parameters, where we try to obtain the parameter λs value for each team and the individual home-ground advantage parameter γ_is. Here, $i = 1, 2, ...N$. We assume that the γ_i parameter remains constant throughout the period chosen for analysis. This represents the average team-specific home-ground advantage of teams.
- **Representation of $\mathbb{P}(\mathbf{i > j})$**: The initial equation of the model looks like this:

$$\mathbb{P}(i > j)_i = \frac{\lambda_i\gamma_i}{\lambda_i\gamma_i + \lambda_j}$$

$$\mathbb{P}(i < j)_i = \frac{\lambda_j}{\lambda_i\gamma_i + \lambda_j}$$

- **Likelihood Function**: The likelihood function of the model is:

$$L(\lambda, \gamma) \quad = \quad \prod_{i=1}\prod_{j=1}(\frac{\lambda_i\gamma_i}{\lambda_i\gamma_i + \lambda_j})^{W_{ij}}(\frac{\lambda_i}{\lambda_i + \lambda_j\gamma_j})^{L_{ji}}(\frac{\lambda_i}{\lambda_i + \lambda_j})^{NW_{ij}}$$

- **Log-likelihood Function**: Instead of optimizing the likelihood function, the log-likelihood function is used, making the process more efficient. Thus, the log-likelihood function looks as follows:

$$l(\theta,\alpha) = \sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij}((\theta_i+\alpha_i) - log(e^{\theta_i+\alpha_i}+e^{\theta_j})) + L_{ji}(\theta_i - log(e^{\theta_i}+e^{\theta_j+\alpha_j})) + NW_{ij}(\theta_i - log(e^{\theta_i}+e^{\theta_j}))$$

  Usage of α_j is for notation purposes to indicate the home-ground advantage of team j.

- **Gradient to the log-likelihood Function**: Considering the partial derivative with respect to θ_i for the gradient, the result is in the form of

$$\frac{\partial l}{\partial \theta_i} = \sum_{j=1}^{N}(W_{ij} + L_{ji} + NW_{ij}) - H_{ij}\frac{e^{\theta_i+\alpha_i}}{e^{\theta_i+\alpha_i}+e^{\theta_j}} - H_{ji}\frac{e^{\theta_i}}{e^{\theta_i}+e^{\theta_j+\alpha_j}} - (NW_{ij}+NW_{ji})\frac{e^{\theta_i}}{e^{\theta_i}+e^{\theta_j}}$$

  and the partial derivative with respect to α_i is in the form

$$\frac{\partial l}{\partial \alpha_i} = \sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij} - H_{ij}\frac{e^{\theta_i+\alpha_i}}{e^{\theta_i+\alpha_i}+e^{\theta_j}}$$

2.2.5 Hierarchical Home-ground Advantage Model

- **Number of Parameters**: The model has N$(n+1)$ parameters, where we try to obtain the parameter λs value for each team and the $n*N$ home-ground advantage parameter $\gamma_{i,n}$s for each team and the level of hierarchy. n denotes the total levels of hierarchy in the tournament.

As mentioned above, R_{ij} represents the level of relationship in the hierarchy between team i and team j, which brings us to $\gamma_{i,R_{ij}}$ as the individual home-ground advantage of team i against team j, based on their hierarchical relationship. Here, $(i, j) = 1, 2, ...N$. We assume that the $\gamma_{i,n}$ parameter remains constant throughout the period chosen for analysis. This represents the average hierarchical home-ground advantage of teams.

- **Representation of $\mathbb{P}(i > j)$**: The initial equation of the model looks like this:

$$\mathbb{P}(i > j)_i = \frac{\lambda_i \gamma_{i,R_{ij}}}{\lambda_i \gamma_{i,R_{ij}} + \lambda_j}$$

$$\mathbb{P}(i < j)_i = \frac{\lambda_j}{\lambda_i \gamma_{i,R_{ij}} + \lambda_j}$$

- **Likelihood Function**: The likelihood function of the model is:

$$L(\lambda, \gamma) \; = \; \prod_{i=1} \prod_{j=1} (\frac{\lambda_i \gamma_{i,R_{ij}}}{\lambda_i \gamma_{i,R_{ij}} + \lambda_j})^{W_{ij}} (\frac{\lambda_i}{\lambda_i + \lambda_j \gamma_{j,R_{ij}}})^{L_{ji}} (\frac{\lambda_i}{\lambda_i + \lambda_j})^{NW_{ij}}$$

- **Log-likelihood Function**: Instead of optimizing the likelihood function, the log-likelihood function is used, making the process more efficient. Thus, the log-likelihood function looks as follows:

$$l(\theta, \alpha) = \sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij}(\theta_i + \alpha_{i,R_{ij}} - log(e^{\theta_i + \alpha_{i,R_{ij}}} + e^{\theta_j})) + L_{ji}(\theta_i - log(e^{\theta_i} + e^{\theta_j + \alpha_{j,R_{ij}}})) + NW_{ij}(\theta_i - log(e^{\theta_i} + e^{\theta_j}))$$

- **Gradient to the log-likelihood Function**: Considering the partial derivative with respect to θ_i for the gradient, the result is in the form of

$$\frac{\partial l}{\partial \theta_i} = \sum_{j=1}^{N}(W_{ij}+L_{ji}+NW_{ij}) - H_{ij}\frac{e^{\theta_i+\alpha_{i,R_{ij}}}}{e^{\theta_i+\alpha_{i,R_{ij}}}+e^{\theta_j}} - H_{ji}\frac{e^{\theta_i}}{e^{\theta_i}+e^{\theta_j+\alpha_{j,R_{ij}}}} - (NW_{ij}+NW_{ji})\frac{e^{\theta_i}}{e^{\theta_i}+e^{\theta_j}}$$

and the partial derivative with respect to $\alpha_{i,n}$ is in the form

$$\frac{\partial l}{\partial \alpha_{i,n}} = \sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij} - H_{ij}\frac{e^{\theta_i+\alpha_{i,R_{ij}}}}{e^{\theta_i+\alpha_{i,R_{ij}}}+e^{\theta_j}} \qquad where\ n = R_{ij}$$

2.2.6 Pairwise Home-ground Advantage Model

- **Number of Parameters**: The model has $N(N+1)$ parameters, where we try to obtain the parameter λs value for each team and the individual pairwise home-ground advantage parameter $\gamma_{i,j}$s where $(i,j) = 1, 2, ...N$. We assume that the $\gamma_{i,j}$ parameter remains constant throughout this period. This represents the average pairwise home-ground advantage of teams.

- **Representation of $\mathbb{P}(\mathbf{i} > \mathbf{j})$**: The initial equation of the model looks like this:

$$\mathbb{P}(i > j)_i = \frac{\lambda_i\gamma_{i,j}}{\lambda_i\gamma_{i,j}+\lambda_j}$$

$$\mathbb{P}(i < j)_i = \frac{\lambda_j}{\lambda_i\gamma_{i,j}+\lambda_j}$$

- **Likelihood Function**: The likelihood function of the model is:

$$L(\lambda,\gamma) = \prod_{i=1}\prod_{j=1}(\frac{\lambda_i\gamma_{i,j}}{\lambda_i\gamma_{i,j}+\lambda_j})^{W_{ij}} \quad (\frac{\lambda_i}{\lambda_i+\lambda_j\gamma_{j,i}})^{L_{ji}}(\frac{\lambda_i}{\lambda_i+\lambda_j})^{NW_{ij}}$$

- **Log-likelihood Function**: Instead of optimizing the likelihood function, the log-likelihood function is used, making the process more efficient. Thus, the log-likelihood function looks as follows:

$$l(\theta,\alpha) = \sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij}((\theta_i+\alpha_{i,j})-log(e^{\theta_i+\alpha_{i,j}}+e^{\theta_j}))+L_{ji}(\theta_i-log(e^{\theta_i}+e^{\theta_j+\alpha_{j,i}})) + NW_{ij}(\theta_i - log(e^{\theta_i}+e^{\theta_j}))$$

- **Gradient to the log-likelihood Function**: Considering the partial derivative with respect to θ_i for the gradient, the result is in the form of

$$\frac{\partial l}{\partial \theta_i} = \sum_{j=1}^{N}(W_{ij}+L_{ji}+NW_{ij}) - H_{ij}\frac{e^{\theta_i+\alpha_{i,j}}}{e^{\theta_i+\alpha_{i,j}}+e^{\theta_j}} - H_{ji}\frac{e^{\theta_i}}{e^{\theta_i}+e^{\theta_j+\alpha_{j,i}}} - (NW_{ij}+NW_{ji})\frac{e^{\theta_i}}{e^{\theta_i}+e^{\theta_j}}$$

and the partial derivative with respect to $\alpha_{i,j}$ is in the form

$$\frac{\partial l}{\partial \alpha_{i,j}} = \sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij} - H_{ij}\frac{e^{\theta_i+\alpha_{i,j}}}{e^{\theta_i+\alpha_{i,j}}+e^{\theta_j}} \quad where(i,j) = 1,2....N$$

2.3 Hypothesis Testing

After we have obtained all the models, we would like to be able to perform some hypothesis testing on these models to check whether a higher-order or complex model fits our data better. As we are using Maximum Likelihood Estimation, we turn to the Likelihood Ratio Test statistic for hypothesis testing. The reason behind using LRT[1] is that the test is valid only if the models

[1]Likelihood Ratio Test

are hierarchically nested within each other, fitting our scenario perfectly.

To put our test in equation form, let us consider two models M_1 and M_0, where M_1 is the higher-order model. After performing the log-likelihood estimation, we obtain log-likelihood values l_1 and l_0 for our models respectively. Then, LRT is given by:

$$-2(l_0 - l_1) \sim \chi^2_{df}$$

Here, the degrees of freedom is calculated by subtracting the number of parameters of the lower-order model from the number of parameters of the higher-order model. For example: If we are to conduct an LRT between the Common Home-ground Advantage model and the Team-specific Home-ground Advantage Model, the degrees of freedom will be:

$$2N - (N + 1) = N - 1$$

where N is the number of teams

Chapter 3

Implementation

We have introduced all the models we would like to implement as an extension of the Bradley-Terry models. Our primary objective is to explore order effects by implementing these models on computer software and obtaining the appropriate results. We use R programming software to create our functions and obtain the parameters. Firstly, we show our function that helps us fit the desired Bradley-Terry models with all the necessary inputs. After introducing the function, we show how we obtain all the inputs in the desired format for the main function to be implemented. We also include another function that performs the hypothesis testing of the nested models based on the user's requirement.

3.1 Process

We have represented the Log-likelihood functions and derivatives as `Log_Likex` and `Log_Like_derivx` where 'x' ranges from 1 to 6, with each number representing a Bradley-Terry model[1]. The model fitting is primarily done using the Log-Likelihood functions in `optim()` in R programming software. The `optim()` function performs general-purpose optimizations for the initial pa-

[1]The numbers represent the Bradley-Terry in the same order they are presented in the paper. For example, the Vanilla Bradley-Terry model is denoted by 1, while Team-specific Home-ground Advantage Model is denoted by 4

rameters. This process' computation time is reduced significantly by using the gradient for the Log-likelihood function (first derivative). For all the models, the number of parameters is obtained using the `ntheta()` function. For the initial guess of the parameters, all the theta values are set to 0 and optimized from thereon. The code that has been used to implement the models and obtain all the results has been uploaded onto GitHub[2]

3.2 Main Functions

We are going to use two functions for fitting the Bradley-Terry model according to our choice and perform hypothesis testing for the models. We are going to list the functions and explain how we go about obtaining the inputs in the required form for these functions.

3.2.1 Fitting the Model

The function `BT_Model()` is used to fit the Bradley-Terry model of the users' choice and returns the most optimal parameters with the initial θ_i parameters scaled to the worst team in the league. The `BT_Model()` has 3 inputs to be provided. They are:

- `matrix` - The list of 5 matrices to represent the win-loss numbers of the teams. The 5 matrices are represented in the following manner:
  - `matrix$w` - Number of home-wins matrix.
  - `matrix$l` - Number of home-losses matrix.
  - `matrix$nw` - Number of neutral-venue wins matrix
  - `matrix$nl` - Number of neutral-venue losses matrix
  - `matrix$h` - Number of home games played matrix
- `Model_Type` - We have discussed 6 models so far in the paper. To input which model is to be fitted, we have used a three-letter indicator that

[2]https://github.com/Varun583/Masters-Thesis

is to be specified. The three-letter indicator matching the model type is listed below:

  - 'VAN' - Vanilla Bradley-Terry Model
  - 'CHA' - Common Home-ground Advantage Model
  - 'CHI' - Common Hierarchical Home-ground Advantage Model
  - 'TSH' - Team-specific Home-ground Advantage Model
  - 'HIE' - Hierarchical Home-ground Advantage Model
  - 'PAI' - Pairwise Home-ground Advantage Model

- `R` - Relationship matrix of the team with respect to the level of hierarchy in the competition[3]. This matrix is a symmetric matrix where $R_{ij} = R_{ji}$

The vital input to execute this function would be the `matrix`. As this information is not readily available on any platform, we perform some pre-processing to get the data in the required form. We start the pre-processing by obtaining all the relevant fixtures that we would like to fit our models into. Since our primary interest is modeling NBA and T20 Blast tournaments, the raw data for both tournaments look quite different, as seen in Figures 3.1 and 3.2. The data is obtained from reliable websites called basketball-reference.com[4] and espncricinfo.com[17]

	Date	Start..ET.	Visitor.Neutral	Points_A	Home.Neutral	Points_B	X	Attend.	Arena	Notes
1	Tue Oct 16 2018	8:00p	Philadelphia 76ers	87	Boston Celtics	105		18624	TD Garden	
2	Tue Oct 16 2018	10:30p	Oklahoma City Thunder	100	Golden State Warriors	108		19596	Oracle Arena	
3	Wed Oct 17 2018	7:00p	Milwaukee Bucks	113	Charlotte Hornets	112		17889	Spectrum Center	
4	Wed Oct 17 2018	7:00p	Brooklyn Nets	100	Detroit Pistons	103		20332	Little Caesars Arena	
5	Wed Oct 17 2018	7:00p	Memphis Grizzlies	83	Indiana Pacers	111		17923	Bankers Life Fieldhouse	
6	Wed Oct 17 2018	7:00p	Miami Heat	101	Orlando Magic	104		19191	Amway Center	
7	Wed Oct 17 2018	7:30p	Atlanta Hawks	107	New York Knicks	126		18249	Madison Square Garden (IV)	
8	Wed Oct 17 2018	7:30p	Cleveland Cavaliers	104	Toronto Raptors	116		19915	Scotiabank Arena	
9	Wed Oct 17 2018	8:00p	New Orleans Pelicans	131	Houston Rockets	112		18055	Toyota Center	
10	Wed Oct 17 2018	8:30p	Minnesota Timberwolves	108	San Antonio Spurs	112		18354	AT&T Center	

Figure 3.1: Raw data imported into R: NBA version

[3]Models 'VAN', 'CHA', 'TSH', and 'PAI' do not need the input of R and can be executed without R

	Team 1	Team 2	Winner	Margin	Ground	Match Date	Scorecard
1	Essex	Sussex	Sussex	36 runs	Chelmsford	Jul 4, 2018	T20
2	Northants	Leics	Leics	4 wickets	Northampton	Jul 4, 2018	T20
3	Notts	Bears	Bears	8 wickets	Nottingham	Jul 4, 2018	T20
4	Middlesex	Surrey	Middlesex	3 wickets	Lord's	Jul 5, 2018	T20
5	Lancashire	WORCS	WORCS	5 wickets	Manchester	Jul 5, 2018	T20
6	Yorkshire	Durham	Yorkshire	44 runs	Leeds	Jul 5, 2018	T20
7	Somerset	Gloucs	Somerset	6 wickets	Taunton	Jul 6, 2018	T20
8	WORCS	Bears	WORCS	4 runs	Worcester	Jul 6, 2018	T20
9	Leics	Durham	Durham	33 runs	Leicester	Jul 6, 2018	T20
10	Northants	Notts	Notts	58 runs	Northampton	Jul 6, 2018	T20

Figure 3.2: Raw data imported into R: T20 Blast version

We would like to format the data into a series of 1s and 0s to make future processes easier[4]. Since data from NBA and T20 Blast are in different forms, we use two different functions, `DesiredFormatNBA()` and `DesiredFormatBlast()`, to convert the data frame of fixtures into the desired form. The formatted data frame looks like Figure 3.3 Given neutral venue scenarios exist in our data, there is no unique way of coding this in our functions. We identify the games and make the necessary changes in the data frame manually. Although this is a cumbersome process, given there are very few neutral venue games[5] in our data, we used a combination of R programming and manual entry to format the neutral venue games appropriately.

The function `Data2Mat()` is the final link in this pre-processing. This function takes the formatted data and converts them into a list of matrices as per our requirement. As mentioned, we would get a list of 5 matrices with all the necessary information for further processes. Although we could represent NBA and T20 Blast Data for the results of the function, given there

[4]Since cricket has both tied games and no results due to rain, we have coded a tie as 0.5 of a win for each team and no result as a 0 of a win for each team

[5]NBA: 214 games out of the 4949 games in the data, Blast: 29 games out of 475 completed games

	Team.A	Team.B	Team.A.win	Team.B.win	Team.A.home	Team.B.home
1	Warwickshire Bears	Nottinghamshire Outlaws	1	0	0	1
2	Leicestershire Foxes	Northamptonshire Steelbacks	1	0	0	1
3	Sussex Sharks	Essex Eagles	1	0	0	1
4	Surrey	Middlesex	0	1	0	1
5	Durham	Yorkshire Vikings	0	1	0	1
6	Worcestershire Rapids	Lancashire Lightning	1	0	0	1
7	Gloucestershire	Somerset	0	1	0	1
8	Warwickshire Bears	Worcestershire Rapids	0	1	0	1
9	Durham	Leicestershire Foxes	1	0	0	1
10	Kent Spitfires	Surrey	1	0	0	1

Figure 3.3: Data formatted into the required form

are 30 and 18 teams respectively in the competition, the representation in this paper would take up a lot of space.

Hence, we take a smaller example of the 2013 season of the Caribbean Premier League[6]. There were 6 teams in the league, with each team playing a total of 7 games, bringing the total games to 21, with 3 additional games being the semifinals and the final. The 6 franchise teams in the league were from Guyana, Jamaica, Barbados, Trinidad and Tobago[7], Antigua, and St Lucia. The matrices in Figures 3.4, 3.5, 3.6 represent the output of the `Data2Mat()` function.

```{r}
matCPL = Data2Mat(CPL)
matCPL$w
```

```
                    Barbados Tridents Guyana Amazon Warriors Antigua Hawksbills St Lucia Zouks Trinbago Red Steel Jamaica Tallawahs
Barbados Tridents                   0                      0                  1              1                 1                 0
Guyana Amazon Warriors              0                      0                  0              0                 1                 1
Antigua Hawksbills                  1                      0                  0              0                 0                 0
St Lucia Zouks                      0                      0                  0              0                 0                 0
Trinbago Red Steel                  0                      1                  1              0                 0                 0
Jamaica Tallawahs                   1                      0                  0              0                 1                 0
```

Figure 3.4: Matrix of Home-Wins: CPL 2013

[6]abbreviated as CPL
[7]referred to as Trinbago

```{r}
matCPL = Data2Mat(CPL)
matCPL$l
```

	Barbados Tridents	Guyana Amazon Warriors	Antigua Hawksbills	St Lucia Zouks	Trinbago Red Steel	Jamaica Tallawahs
Barbados Tridents	0	0	0	0	0	0
Guyana Amazon Warriors	0	0	0	1	0	0
Antigua Hawksbills	0	1	0	1	0	0
St Lucia Zouks	1	0	1	0	0	1
Trinbago Red Steel	0	1	0	0	0	1
Jamaica Tallawahs	0	1	0	0	0	0

Figure 3.5: Matrix of Home-Losses: CPL 2013

```{r}
matCPL = Data2Mat(CPL)
matCPL$nw
```

	Barbados Tridents	Guyana Amazon Warriors	Antigua Hawksbills	St Lucia Zouks	Trinbago Red Steel	Jamaica Tallawahs
Barbados Tridents	0	0	0	0	0	0
Guyana Amazon Warriors	1	0	0	0	0	0
Antigua Hawksbills	0	0	0	0	0	0
St Lucia Zouks	0	0	0	0	0	0
Trinbago Red Steel	0	0	0	1	0	0
Jamaica Tallawahs	1	1	1	0	0	0

Figure 3.6: Matrix of Neutral Venue Wins: CPL 2013

The other vital input is the relationship matrix for the models. We know that each competition has a different level of hierarchy and the hierarchy is established based on different criteria. As we are fitting models for NBA and T20 Blast data, we have been able to draft the relationship matrix for these two competitions alone, in two different functions.

`NBA_R()` takes the input of the formatted data frame of fixtures and uses the unique teams in those fixtures to formulate the relationship matrix of NBA. The relationship level is assigned in this form:

- 3 - Teams are in the same division
- 2 - Teams are in the same conference but in different divisions
- 1 - Teams are in different conferences

`Blast_R()`, just like the `NBA_R()`, formulates the relationship matrix of the T20 Blast competition. The relationship level is assigned in this form:

- 2 - Teams are in the same group
- 1 - Teams are in different groups

3.2.2 Hypothesis Testing of the Models

The function `BT_test()` is used to perform hypothesis testing of the nested models. This function has 4 inputs to be provided. They are:

- `model1` - One of the two models to be compared.
- `model2` - One of the two models to be compared.
- `matrix` - The list of 5 matrices to represent the win-loss numbers of the teams.
- `R` - Relationship matrix of the team with respect to the level of hierarchy in the competition

This function takes the three-letter specification of the models as inputs for `model1` and `model2` and returns the p-value to help the user determine the statistical significance of the two models and the χ^2 statistic with the appropriate degrees of freedom.

3.3 Output

3.3.1 Fitting the model

Figure 3.7 represents the output for the Vanilla Bradley-Terry model. This shows the θ_i values scaled to the worst team in the league, with some information about the worst and best teams in the league.

Figure 3.8 represents the output for the Common Home-ground Advantage model. Just like the vanilla model, this shows the θ_i values scaled to the worst team in the league, along with α common home-ground advantage parameter. This also provides some information about the worst and best teams in the league.

Figure 3.9 represents the output for the Team-specific Home-ground advantage model. While providing the scaled θ_i values, this model also provides

the individual α_i home-ground advantage parameters[8]. In the end, the model shows which team has the highest individual home-ground advantage among all the teams.

```
> BT_Model(matCPL, 'VAN')
[[1]]
                          Theta
Barbados Tridents      0.9821378
Guyana Amazon Warriors 2.1194639
Antigua Hawksbills     0.0000000
St Lucia Zouks         0.0000000
Trinbago Red Steel     0.9644990
Jamaica Tallawahs      2.5583823

$type
[1] "VAN"

$Worst
[1] "Antigua Hawksbills"

$Best
[1] "Jamaica Tallawahs"
```

Figure 3.7: Vanilla Bradley-Terry Model: CPL 2013

```
> BT_Model(matCPL, 'CHA')
[[1]]
                          Theta
Barbados Tridents      0.974507968
Guyana Amazon Warriors 2.139319718
Antigua Hawksbills     0.008073335
St Lucia Zouks         0.000000000
Trinbago Red Steel     0.950817547
Jamaica Tallawahs      2.560118629

$Alpha
[1] 0.1860139

$type
[1] "CHA"

$Worst
[1] "St Lucia Zouks"

$Best
[1] "Jamaica Tallawahs"
```

Figure 3.8: Common Home-ground Advantage Bradley-Terry Model: CPL 2013

[8]In this case, Barbados Tridents have a very high α value as they won all their home games this season. This could be evidence to state that the sample size is very small and more data should be incorporated to make the models more comprehensive.

```
> BT_Model(matCPL, 'TSH')
[[1]]
                            Theta
Barbados Tridents        0.000000
Guyana Amazon Warriors  19.485782
Antigua Hawksbills       7.351507
St Lucia Zouks          19.918481
Trinbago Red Steel      19.223553
Jamaica Tallawahs       20.483679

[[2]]
                             Alpha
Barbados Tridents        31.684712
Guyana Amazon Warriors    1.127545
Antigua Hawksbills        2.498804
St Lucia Zouks          -30.734696
Trinbago Red Steel       -0.138966
Jamaica Tallawahs        -1.129107

$type
[1] "TSH"

$Highest_Home_Advantage
[1] "Barbados Tridents"
```

Figure 3.9: Team-specific Home-ground Advantage Bradley-Terry Model: CPL 2013

3.3.2 Hypothesis testing of the Models

Figure 3.10 represents the LRT for the two models chosen for the CPL 2013 season. With p-value being significant, we can say that a more complex model (Team-specific Home-ground Advantage Model) was a better fit than the nested model (Common Home-ground Advantage Model) for a single season of CPL.

```
```{r}
BT_test('TSH','CHA',matCPL)
```

P-Value is 0.02326319
with a statistic of 13.01236 and 5 degrees of freedom
```

Figure 3.10: LRT for TSH model vs CHA model: CPL 2013

3.4 List of Functions

We have mentioned the functions we have used until now in a very haphazard manner. Hence, here is a list of all the functions we have implemented so far in this paper.

- `Log_Likex()` - The log-likelihood functions introduced in Chapter 2 where 'x' ranges from 1 to 6, with each number representing a model.
- `Log_Like_derivx()` - The derivative of the log-likelihood functions, making the gradient, introduced in Chapter 2 where 'x' ranges from 1 to 6, with each number representing a model.
- `DesiredFormatNBA()` - This function takes the list of NBA fixtures obtained from the website and converts it into a pre-prescribed format.
- `DesiredFormatBlast()` - This function takes the list of T20 Blast fixtures obtained from the website and converts it into a pre-prescribed format.
- `Data2Mat()` - This function converts a standard formatted data frame of fixtures into a list of 5 matrices, with win-loss information of all the teams in the tournament.
- `ntheta()` - The function, when inputted the win/loss matrices and model number from 1 to 6, returns the number of parameters appropriate to the model.
- `NBA_R()` - This function returns a relationship matrix of the NBA, with the appropriate levels of hierarchy.
- `Blast_R()` - This function returns a relationship matrix of the T20 Blast, with the appropriate levels of hierarchy.
- `PCT_Table` - The function, when inputted the list of matrices obtained from `Data2Mat()`, returns a data frame with the total number of wins and losses for the teams, with the final column returning the winning percentage, an important statistic in the NBA.

- `Home_Table()` - The function, when inputting the list of matrices obtained from `Data2Mat()`, returns a data frame with the total number of home wins and losses for the teams, with the final column returning the winning percentage.
- `BT_plots()` - This function, when inputted the model results obtained from `BT_Model()`, returns the plots for the Bradley-Terry model.
- `BT_predict()` - This function returns the probabilities of winning for both the teams involved in a fixture, based on the fitted model chosen by the user.

Chapter 4

Applications

4.1 NBA

4.1.1 Background

The **National Basketball Association**, popularly known as the **NBA**, is a professional basketball league conducted across North America. The National Basketball League (established in 1937) and the Basketball Association of America (created in 1946) were two rival organizations that merged to form the NBA in 1949. Four American Basketball Association (ABA) teams were absorbed into the NBA in 1976. [28]

Since the expansion of the league in 2004-05[1], the league consists of 30 teams, divided into 2 conferences, Eastern and Western conferences based on their geographic location. These conferences are further divided into 3 equal divisions i.e., each division with 5 teams, again based on geographic location. Appendix D provides the exact hierarchy of the league.

As mentioned in Chapter 1, the presence of this hierarchy is one of our major motivations to explore the league using the Bradley-Terry models. With 3

[1]The addition of Charlotte Bobcats (currently known as Charlotte Hornets)

levels of hierarchy i.e., intra-division, intra-conference, and inter-conference, the presence of order effects or advantages would not be a surprise. With the United States being one of the largest countries in the world, geographic locations around the country are not similar in their weather conditions. As we move from south to north of the United States, conditions get colder as we are moving further away from the equator. Jones [23] takes note that the travel time to visit an intra-division team is shorter than when visiting an intra-conference rival, which is further shorter when visiting an inter-conference rival. These changes in conditions might bring out different levels of order effects among teams.

A regular season of the NBA has teams playing 82 games[2] each. The NBA has used the hierarchy to introduce an imbalance in the number of games the teams play against each other. For example, in the 2018-19 season, we see this pattern for a particular team:

- Four games against teams from the same division, 2 at their home court and 2 in away court.
- Four games against 6 of the 10 teams from the same conference but different divisions, 2 at their home court and 2 on the away court.
- Three games against 4 of the 10 teams from the same conference but different divisions, with the home and away games played in a 2:1 ratio, decided randomly by the league.
- Two games against teams from the opposite conference, 1 at their home court and 1 on the away court.

The league uses a 5-year rotation period to decide the scheduling of the matches between intra-conference rivals i.e. whether the team plays a rival three times or four times in the league.

CoViD - 19 and the repercussions

On March 11, 2020, the 2019-20 NBA season was interrupted due to the CoViD-19 pandemic. As a result, most teams could not play their regular

[2] 41 home and 41 away games

season's total of 82 games. With careful consideration of a team's playoff chances, 9 from the Eastern Conference and 13 teams from the Western Conference played 8 games to decide the playoff positions. These 22 teams should have to have played around 72 or 73 games at the end of the regular season, but the Dallas Mavericks played 75 games, while the Los Angeles Lakers and San Antonio Spurs played 71 games. A small play-in tournament helped decide some of the last available playoff spots. [32]

With the 2019-20 season coming to a late conclusion, the 2020-21 season had a late start, which led the NBA to shorten the regular season with 72 games[3]. The league had a structure like this:

- Three games against teams from the same conference, with the home and away games played in a 2:1 ratio, decided randomly by the league.
- Two games against teams from the other conference, 1 at their home court and 1 on the away court.

A bigger play-in tournament was scheduled with 4 teams from each conference fighting for the last two playoff spots. The league returned to a normal 82-game home and away structure from the 2021-22 season.

Until the 2019-20 season, the 8 playoff spots in each conference were decided based on the conference standings of the season. The top-8 teams would automatically gain a qualification to the playoffs. But, as the CoViD-19 interruption created some imbalances in the league, the NBA adopted a play-in tournament, which gave the teams finishing 9th and 10th in the league another shot at qualifying for the playoffs.

Since the 2020-21 season, teams seeded 7th through 10th from both conferences will compete in a play-in tournament, the format of which will be explained in Appendix A. Two teams each from both conferences will advance to the playoffs from this tournament. The playoffs are played separately for both the conferences in a best-of-7 series format, with the first team achieving 4 wins advancing to the next round. The winner of the Eastern Conference will face the winner of the Western Conference in a best-of-7 series to decide the NBA champion. Appendix B depicts the structure of the

[3] 36 home and 36 away games

NBA playoffs.

Bradley-Terry models, which are able to handle such hierarchical structures and fixture imbalances, seem the right fit to model NBA data. In this paper, we would like to obtain data for 4 seasons of the NBA, from the 2018-19 season to the 2021-22 season, and use our functions to get appropriate results. Across all these seasons, we use fixtures of the regular season and the post-season[4] and feed them into our functions.

4.1.2 Results

After fitting the models to the data and performing hypothesis testing, we found that the Common Hierarchical Home-ground Advantage Model was a statistically significant model for the NBA seasons of 2018-19 to 2021-22. While this was the highest-order effect we observed, we should also note that the Common Home-ground Advantage model was also statistically significant. As we were running the pairwise home-ground advantage model, as the amount of data for such comparisons were not enough, the parameters did not converge to the optimal values with the `optim()` function. Hence, we are not including the pairwise advantage model. The hypothesis tests conducted for all the models with some important information about the tests are all provided in Table 4.1

Since there were two models that were statistically significant, we plot the significant models along with the vanilla model. Figure 4.1 depicts the Vanilla Bradley-Terry model. Figure 4.2 depicts the Common Home-ground Advantage model. Figure 4.3 depicts the Common Hierarchical Home-ground Advantage model. The win-loss record along with the winning percentage of the teams are presented in Appendix G. Table 4.2 represents the strengths of the teams across the 3 models with the teams arranged from 1 to 30, with respect to their θ_i values.

The additional specific home-ground advantage Paine [29] suspected in his article, although present, is not statistically significant, according to our models and the data for the mentioned seasons. Although geographical factors

[4]The playoffs, including the play-in tournament

Higher-order Model	Lower-order Model	p-value	χ^2 statistic
Vanilla	-	-	-
Common HG[5]	Vanilla	$p < 0.001$	77.475
Common Hierarchical HG	Common HG	$\mathbf{p = 0.106}$	4.493
Team-specific HG	Common Hierarchical HG	$p = 0.472$	26.848
Hierarchical HG	Team-specific HG	$p = 0.87$	47.886

Table 4.1: LRT for NBA data - 2018-19 to 2021-22

might have some impact, all current NBA games are played indoors with standard court sizes, which might take the stadium or the court advantage out of the picture.

It is also worth noting that α parameters of the intra-division and inter-conference levels of the hierarchy were almost similar. This could be because in both these levels of hierarchy, teams face each other equal number of times in a season, home and away, essentially neutralizing the additional edge teams have on each other. But, with the intra-conference level of the hierarchy, the imbalance of home and away fixtures might have led to an increased home-ground advantage among teams.

The non-significance of the team-specific home-ground model begs the question of the impact of fans on the team's performance at their home. People who follow the NBA might know that the New York Knicks' home, Madison Square Garden is one of the loudest places to watch a game. But, due to CoViD - 19, for almost two seasons, games were held with little to no fans inside the stadium. The non-presence of fans could have been indirectly benefitting the away team, neutralizing home-ground advantage. As fans have started to fill stadiums, we might see different results with the team-specific model.

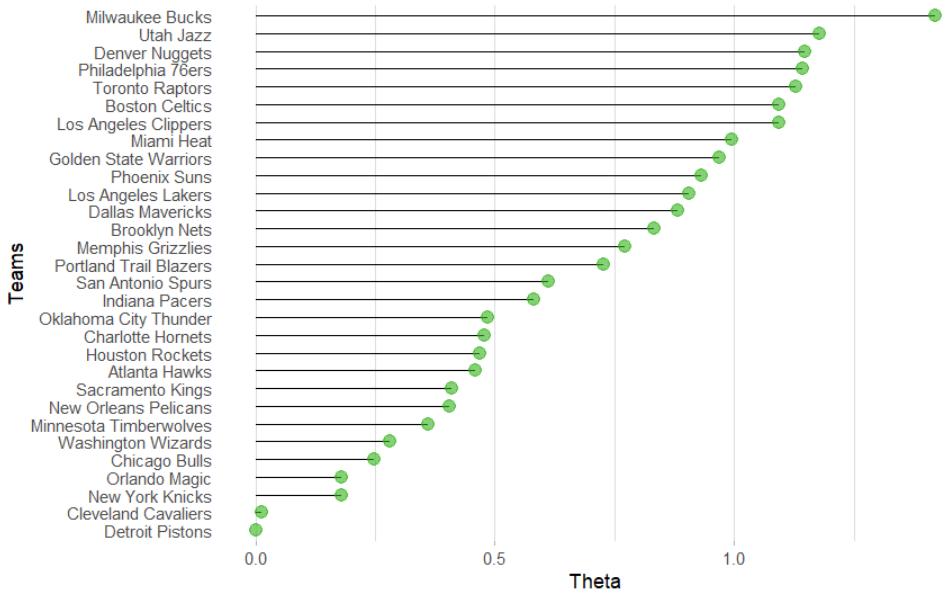


Figure 4.1: Vanilla Bradley-Terry Model: NBA 2018-19 to 2021-22

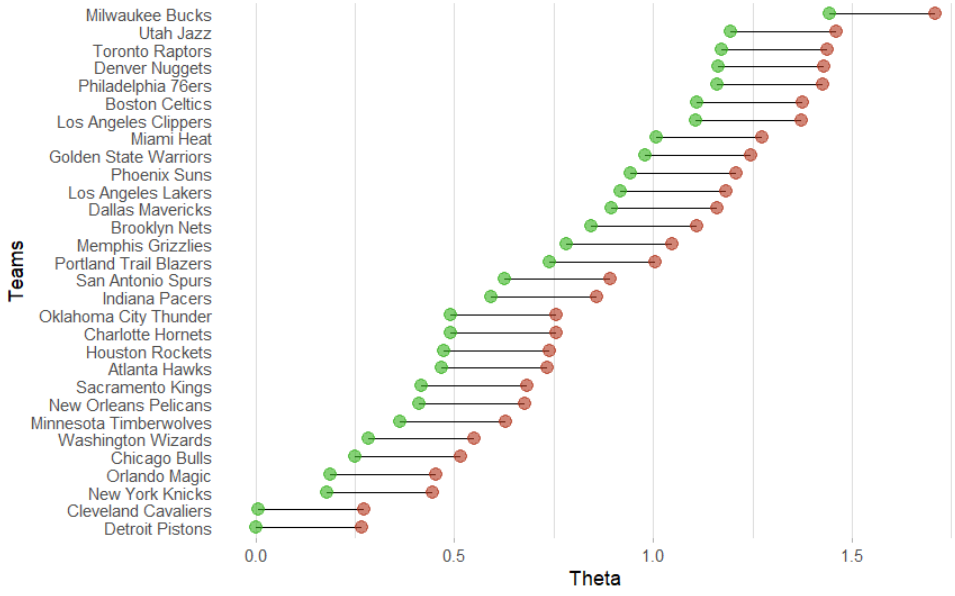


Figure 4.2: Common Home-ground Advantage Model: NBA 2018-19 to 2021-22. Green: θ_i, Red: $\theta_i + \alpha$

Teams	Vanilla	Common HG	Common Hierarchical	PCT[6]
Milwaukee Bucks	1	1	1	1
Utah Jazz	2	2	2	2
Toronto Raptors	5	3	3	5
Denver Nuggets	3	4	4	4
Philadelphia 76ers	4	5	5	3
Boston Celtics	6	6	6	6
Los Angeles Clippers	7	7	7	7
Miami Heat	8	8	8	8
Golden State Warriors	9	9	9	9
Phoenix Suns	10	10	10	10
Los Angeles Lakers	11	11	11	11
Dallas Mavericks	12	12	12	12
Brooklyn Nets	13	13	13	13
Memphis Grizzlies	14	14	14	14
Portland Trail Blazers	15	15	15	15
San Antonio Spurs	16	16	16	17
Indiana Pacers	17	17	17	16
Charlotte Hornets	19	19	18	18
Oklahoma City Thunder	18	18	19	19
Houston Rockets	20	20	20	21
Atlanta Hawks	21	21	21	20
Sacramento Kings	22	22	22	22
New Orleans Pelicans	23	23	23	23
Minnesota Timberwolves	24	24	24	24
Washington Wizards	25	25	25	25
Chicago Bulls	26	26	26	26
Orlando Magic	27	27	27	28
New York Knicks	28	28	28	27
Cleveland Cavaliers	29	29	29	29
Detroit Pistons	30	30	30	30

Table 4.2: Strengths of the teams across 3 models: NBA 2018-19 to 2021-22

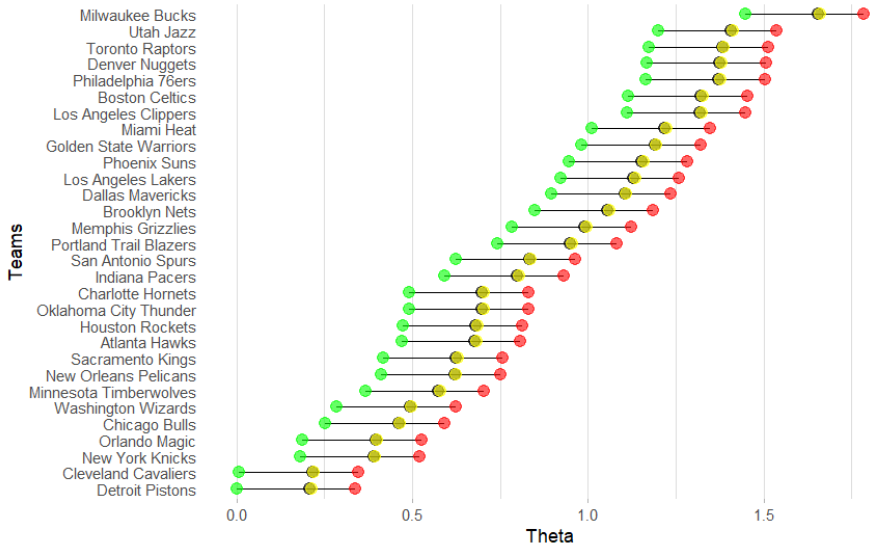


Figure 4.3: Common Hierarchical Home-ground Advantage Model: NBA 2018-19 to 2021-22. Green: θ_i, Black: $\theta_i + \alpha_{R=1}$, Yellow: $\theta_i + \alpha_{R=3}$, Red: $\theta_i + \alpha_{R=2}$

4.2 T20 Blast

4.2.1 Background

T20 Blast, known as **Blast** is a limited-overs professional cricket competition conducted across the United Kingdom. This competition has the status of the first-ever professional T20[7] league, dating back to 2003. This competition was known by several names like the Twenty20 Cup (2003-2009), Friends Provident T20 and Friends Life T20 (2010-2013), and T20 Blast since 2014.

The league, since its inception in 2003, has had 18 teams representing 18 counties of the English cricket system. Over the years, these 18 teams have been divided into groups in different manners based on their geographic location. We represent the league's structure in Table 4.3. The allocation of

[7]20-20, popularly known as T20, is a format in cricket where each team gets to bat for 20 overs and the team who has scored the highest runs wins the game

the teams in these groups is represented in Appendix E and F.

Years	Groups
2003-2009	North, South, Midlands
2010-2011	North, South
2012-2013	North, South, Midlands
2014-2019	North, South
2020	North, South, Central
2021-Present	North, South

Table 4.3: T20 Blast group structure over the years

Since 2014, a regular season of the T20 Blast involves teams playing 14 games each against teams from the same group. There are 9 teams in each group, which creates some imbalance in the number of games of teams against each other. The structure looks like this:

- Two games, against 6 of the 8 teams from the group, 1 played at home ground and 1 on away ground.
- One game, against 2 of the 8 teams from the group, with home and away grounds decided randomly by the league.

With so much history of the game attached, England is one of the homes of cricket. County teams, like Sussex, have been in existence since 1839. Over the years, counties have developed sporting rivalries which make the game reach newer heights with respect to popularity. The league, aware of this, schedules games so the fans get the best out of these on-field rivalries. The league randomly chooses the 6 and 2 teams against which the teams have to play their games. The randomness depends on which two teams bring out the best crowds to the stadiums and increase viewership among the fans.

The well-known rivalries in the county game are Yorkshire-Lancashire[8], Surrey-Middlesex[9]. While these rivalries are ensured two games every regular season, some of the other combinations of teams within groups are not given priority

[8]Popularly known as the Roses rivalry
[9]Popularly known as the London derby

for two games in a season. Some of the fixtures over the last 5 seasons[10], which are played only the minimum number of times are in Table 4.4

Total games	Fixture
5	Derbyshire v Durham Gloucestershire v Surrey Kent v Glamorgan Northamptonshire v Nottinghamshire Somerset v Sussex

Table 4.4: Fixtures: T20 Blast - 2017-2022

CoViD - 19 and the repercussions

As most of the leagues were affected by the virus, T20 Blast had to have a delayed start, leading to a shorter season than regular in 2020. Each team would play 10 games against teams from the same group, with 3 groups in the picture, with a home and away structure. The top 2 teams from each group entered the knockouts. The two best-placed third-ranked teams in the league completed the 8 team knockout lineup. The league reverted back to the 2-group format in 2021, with the regular 14-game structure.

Unlike the NBA, T20 Blast never built a bio-secure bubble for its games but was under strict bio-secure protocols throughout the season. Most of the games followed the regular home-and-away format. But, throughout the cricket season in England, international cricket will run along with domestic competitions like the T20 Blast. As a result, teams might have to move to another venue when an international game is scheduled at their home ground. This leads to neutral venue scenarios, which have been encoded in our fixtures.

Since 2014, excluding 2020, the league has had a knockout structure to decide the winner of the tournament. With 2 groups, the 4 best-placed teams from each group make it to the quarterfinals. Winners of the quarterfinals

[10]ignoring the 2020 season

will move on to the next round, and the team which has won all its games in the knockout stages is the winner of the league. Appendix C explains the format for the knockouts.

Using Bradley-Terry models to understand the hierarchy of the T20 Blast might not give out the best results possible. Although we are considering 4 seasons' worth of data, it is worth mentioning that inter-group matches are very less in number in these fixtures. The reason behind that, as explained, is that the teams do not face teams from the opposite group until the knockout stages. This is just an observation beforehand but we will use our models to quantify this observation.

Bradley-Terry models will help us in handling the imbalance in the T20 Blast fixtures. In this paper, we are considering data from 4 seasons of T20 Blast, starting from the 2018 season to the 2022 season, excluding the 2020 season. The reason behind excluding the 2020 season is the presence of a different hierarchy in the season[11] as opposed to the regular hierarchy[12]. Across all these seasons, we use fixtures of the regular season as well as the knockout stages and feed them into our functions.

4.2.2 Results

After fitting the models to the data and performing hypothesis tests, we found out that none of the models were statistically significant at 90% significance level for the T20 Blast seasons for 2018 to 2022, excluding the 2020 season. We would not be fitting the pairwise home-ground advantage model as we are using only 4 seasons' worth of data, which does not ensure there is at least one game between two teams in the league[13]. Among those that were close to significance, the Team-specific Home-ground Advantage Model was the closest to the 90% significance level. All the hypothesis tests conducted for the models with some important information about the tests are provided in Table 4.5

Since there was only one model that was statistically significant (albeit weak significance), we plot the significant model along with the Vanilla Model.

[11] North, South, and Central

[12] North and South

[13] As teams do not play teams from the opposite group, we have this issue

Higher-order Model	Lower-order Model	p-value	χ^2 statistic
Vanilla	-	-	-
Common HG	Vanilla	p = 0.271	1.21
Common Hierarchical HG	Common HG	p = 0.499	0.457
Team-specific HG	Common Hierarchical HG	**p = 0.152**	21.727
Hierarchical HG	Team-specific HG	p = 0.722	14.113

Table 4.5: LRT for T20 Blast data - 2018 to 2022 (excluding 2020)

Figure 4.4 represents the Vanilla Bradley-Terry model. Figure 4.5 represents the Team-specific Home-ground Advantage model. Just like in the NBA case, we present the home win-loss record for these teams over the course of these seasons in Appendix H. Table 4.6 depicts the strengths of teams, represented by θ_i across these two models and compares them with their respective winning percentages.

If we take a closer look at the results of the models, as expected, with the sample size being small to determine order effects, both the hierarchical models were not statistically significant. With the team-specific home-ground advantage model being significant, a lot of avenues might open up as to what might be causing these different levels of advantage among teams.

Unlike basketball, cricket grounds are not of standard size. There have been instances in the Indian Premier League (IPL)[14] where teams have considered the size of their home ground to construct squads that are suitable for these grounds. Along with the size, the condition of the pitch[15] also plays a huge role in a game of cricket. With so many peculiar aspects in the game, it is unsurprising that each team has a unique home-ground advantage.

An interesting observation with the team-specific advantage model is that out of the 18 teams, about 8 of them have a home-ground disadvantage i.e. they have a better record playing away from their home than at home. This could be evidence of the fact that these teams might not have constructed their squads to maximize their home conditions advantage.

[14]A popular franchise T20 competition based out of India

[15]The central strip of a cricket ground on which the bowler delivers the ball

Lancashire Lightning, the team based out of Manchester, UK, seems to have the highest home-ground advantage. Leicestershire Foxes, the team based out of Leicester, UK, seems to have the highest home-ground disadvantage, with their strength parameter θ reducing significantly when playing at home. This is also backed up by sheer numbers. Lancashire has a 19/4 win/loss record at home during the chosen period, the best win percentage in the league. Leicestershire has a 8/18 win/loss record at home during the same period, the worst win percentage in the league.

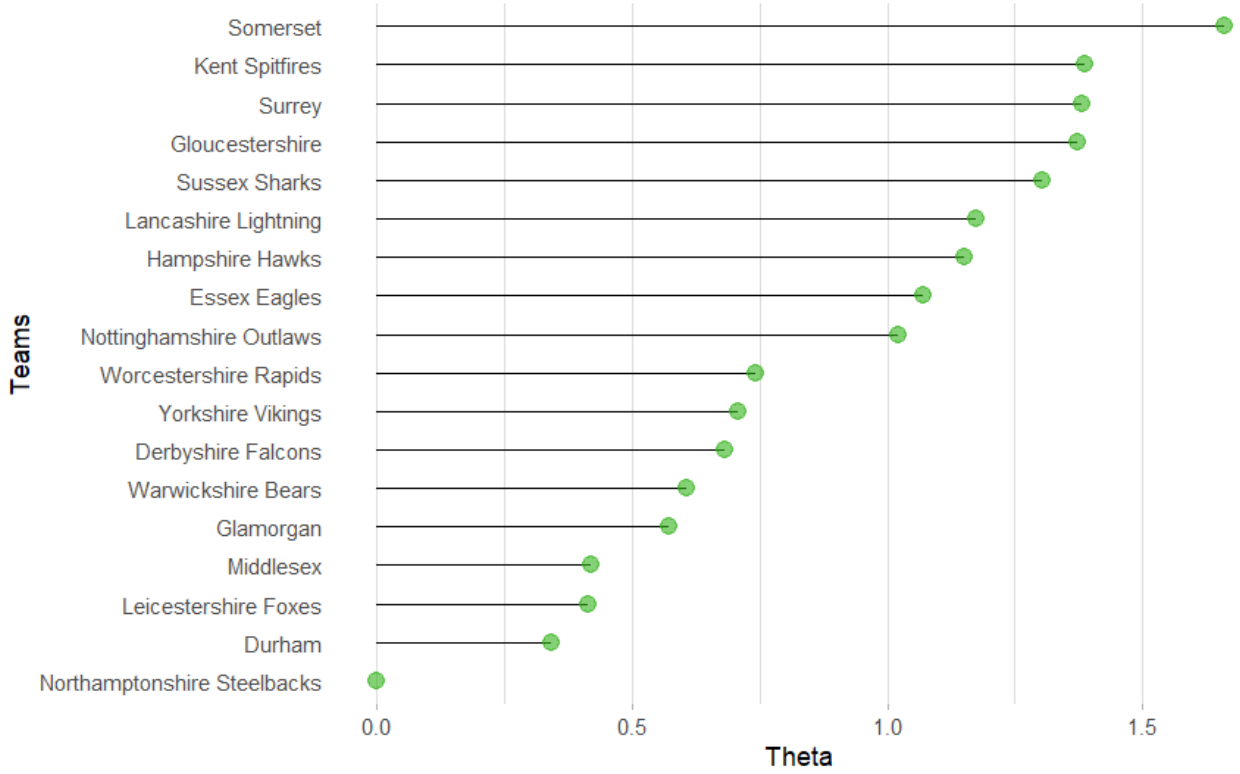


Figure 4.4: Vanilla Bradley-Terry Model: T20 Blast 2018 to 2022 (excluding 2020)

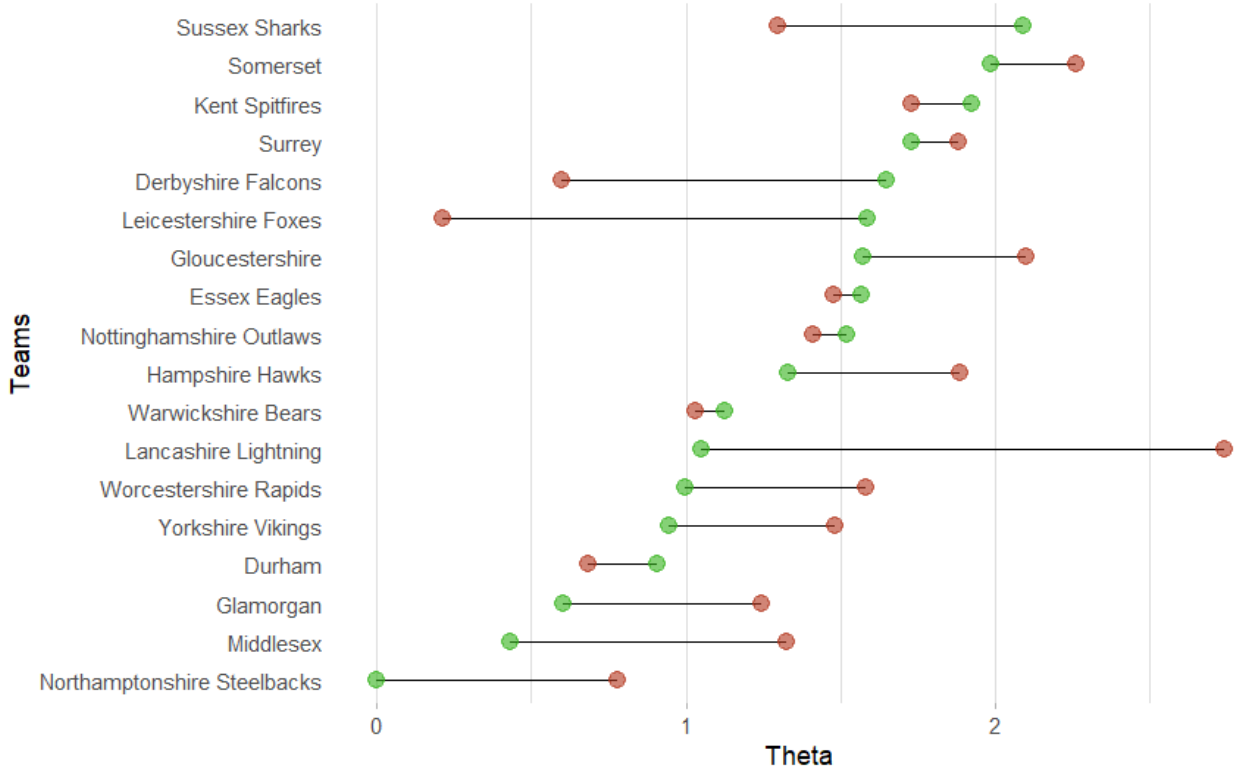


Figure 4.5: Team-specific Home-ground Advantage Model: T20 Blast 2018 to 2022 (excluding 2020). Green: θ_i, Red: $\theta_i + \alpha_i$

Teams	Vanilla	Team-specific	Winning Percentage
Sussex Sharks	5	1	6
Somerset	1	2	1
Kent Spitfires	2	3	7
Surrey	3	4	4
Derbyshire Falcons	12	5	10
Leicestershire Foxes	16	6	14
Gloucestershire	4	7	5
Essex Eagles	8	8	13
Nottinghamshire Outlaws	9	9	3
Hampshire Hawks	7	10	11
Warwickshire Bears	13	11	12
Lancashire Lightning	6	12	2
Worcestershire Rapids	10	13	9
Yorkshire Vikings	11	14	8
Durham	17	15	15
Glamorgan	14	16	16
Middlesex	15	17	18
Northamptonshire Steelbacks	18	18	17

Table 4.6: Strengths of the teams across 2 models: T20 Blast 2018 to 2022 (excluding 2020)

Chapter 5

Simulation and Improvements

After building our models and applying them to our data, we would like to see their predictive power for a set of fixtures. The current NBA season of 2022-23 is almost coming to a conclusion, with Miami Heat and the Denver Nuggets making the championship final. Denver Nuggets, the first seed in the Western Conference have performed exceptionally well and matched up their expectations this season. On the other side, Miami Heat surprised all pre-tournament experts by becoming the first team from the play-in tournament to reach the NBA championship final.

With all the regular season's information, we would like to simulate the playoffs for the current season, starting from the play-in tournament until the NBA finals. Along with the final results, we would like to provide some unique insights about our simulation and some improvements that could be done to the models we have developed so far.

5.1 Simulation Rational

The period from 2014 to 2019 was when the Golden State Warriors built a dynasty and were the undisputed rulers of the NBA. Reaching 5 NBA Finals in all 5 seasons during this period, while winning 3 titles in the process is evidence to that claim. Their win/loss record during this time was 322/88

making their win percentage reach an insurmountable 0.785. This period also included their record-shattering season, with a 73/9 win/loss record, the highest-ever winning percentage achieved by any team in the NBA during a season. But, things changed quickly. Klay Thompson, one of the star players in the side suffered an injury during the 2018-19 NBA finals, which ruled him out for the entire 2019-20 season. Kevin Durant, another star player for the Warriors, requested a trade and moved to Brooklyn Nets. In a remarkable turnaround, the Golden State Warriors went from being the 3rd best team in the league[1] in the year 2019, to the worst team in the league the following season.

To give a more recent example from our data, the Sacramento Kings were one of the lower-ranked sides in the NBA over the last 4 years. When we compare win/loss records from our current data, Sacramento Kings had a record of 131/177 with a PCT[2] of 0.425. In the 4 seasons we have modeled, the Kings stood at the 22nd position in the Common Hierarchical model. Fortunes changed in the current season with a few significant trades, a change in head coach, and previous draft picks repaying the faith leading the side to its first playoff qualification since 2006. The Kings were always an improving side, who reaped huge benefits this last season.

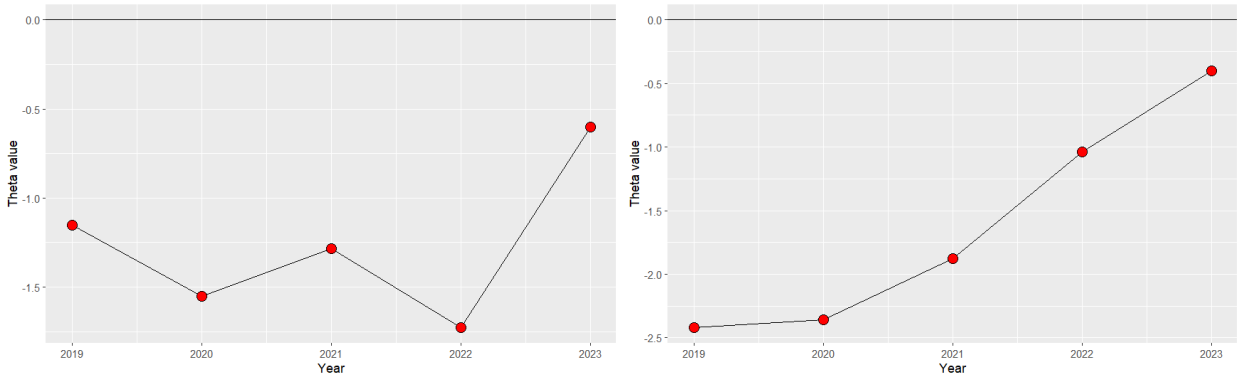
Cleveland Cavaliers were the second-worst side in the Common Hierarchical model built for our data. This was backed up by a win/loss record of 104/199 with a PCT of 0.343. As the new season arrived in 2022, a significant trade in the name of Donovan Mitchell from the Utah Jazz played a huge role in turning the fortunes of the side. The addition of Mitchell to a young and upcoming core helped the side reach the playoffs, which they had failed to do over the last 4 seasons in the NBA.

Figure 5.1 represents the strength parameter θ_i for the Sacramento Kings and Cleveland Cavaliers from the Common Hierarchical Home-ground Advantage model for every year individually from 2018 to 2023. Another unique feature in this figure is that the θ values are scaled to the best-performing team in the league. This means that if a team is closer to 0, their strength is getting similar to the strength of the best team in the league. As we can see

[1]Regular season win PCT

[2]Abbreviation for Winning Percentage: A popular statistic in the NBA

in the 2023 season, both the Kings and Cavaliers have improved significantly and reached the strength level closer to the best team in the league.



Sacramento Kings Cleveland Cavaliers

Figure 5.1: Theta Values - Season by Season

On observation of our results of our models, we have noticed that both Kings and Cavaliers are in the bottom 10 teams of the league with respect to their strengths. But, this season, both teams have successfully qualified for the playoffs, which shows that these sides are not at the strength level depicted by our model. Hence, any simulation by taking the strength level from the 4 seasons' data will not give out appropriate results by undermining the efforts of teams like the Kings and Cavaliers this season. Hence, we take a slightly different approach. We obtain the θ_i values for the team by running the Common Hierarchical Home-ground Advantage model for the 2022-23 regular season. But we obtain the $\alpha_{R(i,j)}$ parameters from the previous 4 seasons' data. Using these θ_i values, we obtain the probabilities of winning for Team A against Team B and vice versa. We will simulate the entire post-season of the NBA, which includes the play-in tournament and the playoffs, 10,000 times and obtain a few notable insights from the same.

5.2 Results and Insights

The play-in tournament in the NBA post-season was started in the 2019-20 season as CoViD - 19 led to interruption and an incomplete regular season. Since the 2020-21 season, teams placed 7th through 10th in the regular season standings from both conferences contest in a small tournament, fighting for the last two playoff spots. The probability of qualifying for the playoffs from this play-in tournament for both conferences this season is given in Tables 5.1 and 5.2

Teams	Playoff Qualifying Probability
Miami Heat	84.5%
Atlanta Hawks	74.6%
Toronto Raptors	24.6%
Chicago Bulls	16.3%

Table 5.1: Playoff Qualifying Probability - Eastern Conference

Teams	Playoff Qualifying Probability
Los Angeles Lakers	84.3%
Minnesota Timberwolves	76%
New Orleans Pelicans	23.7%
Oklahoma City Thunder	16%

Table 5.2: Playoff Qualifying Probability - Western Conference

We also look at the simulations of the teams winning their respective conferences in Figure 5.2 and 5.3. Milwaukee Bucks had the highest chance of winning the Eastern Conference with about a 56% of the simulations being won by them, while the Denver Nuggets had the highest chance of winning the Western Conference with about 55% of the simulations being won by them. Although all the teams in the play-in and the playoffs had at least one simulation of winning the title, teams with a higher chance are presented in the figures.

The NBA Finals are contested between the winner of the Eastern Conference

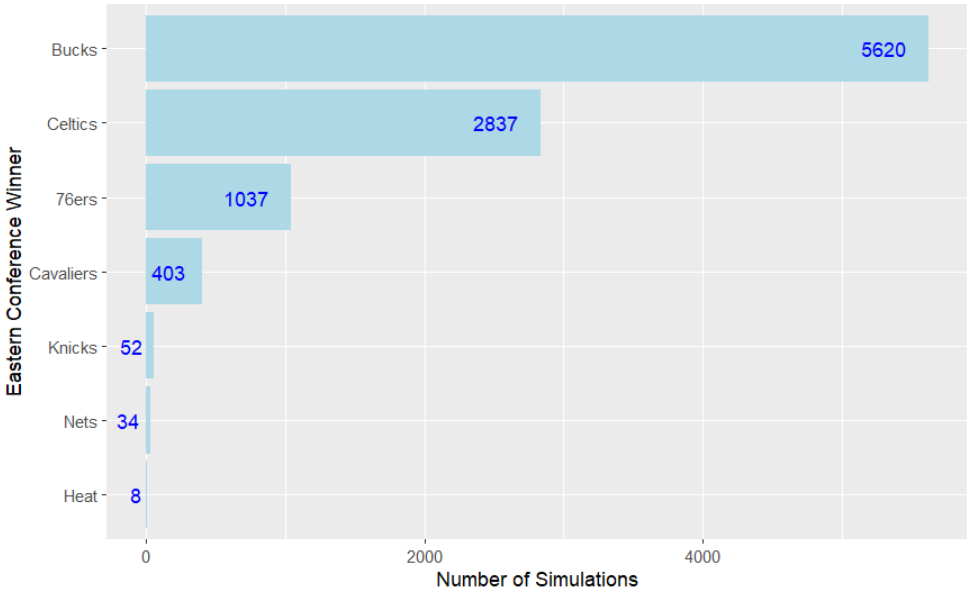


Figure 5.2: Eastern Conference Winners - Simulation

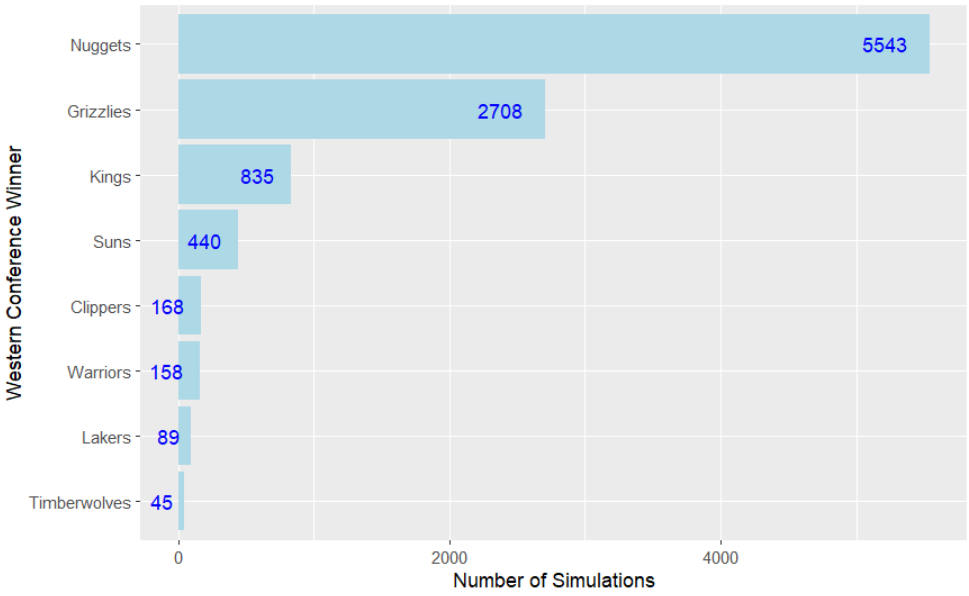


Figure 5.3: Western Conference Winners - Simulation

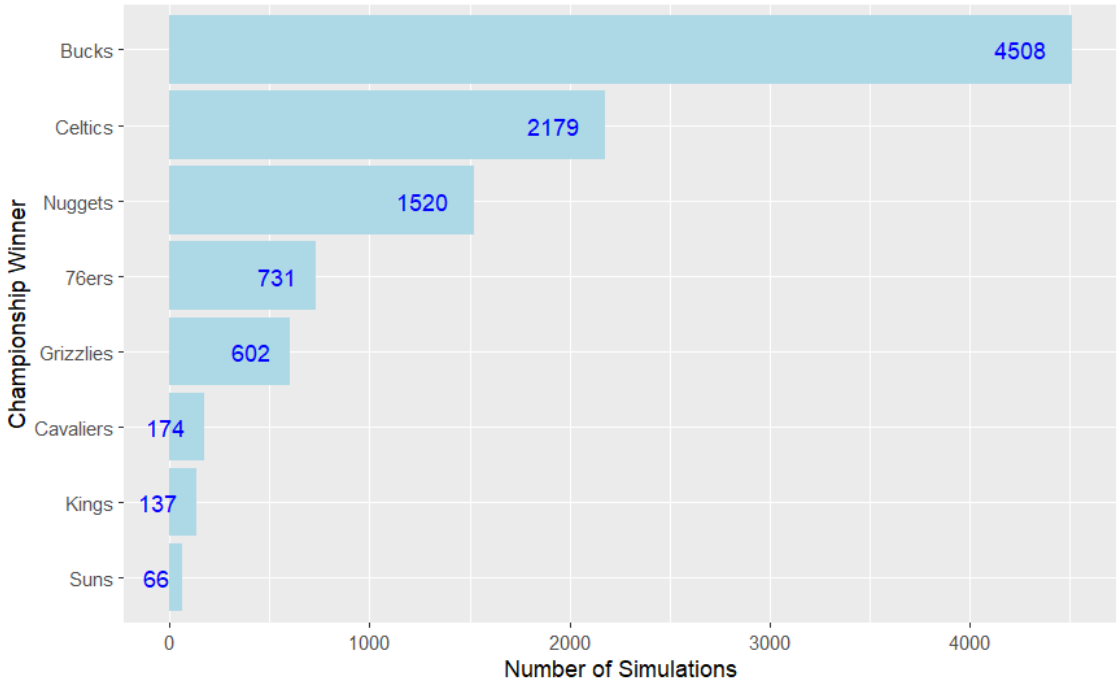


Figure 5.4: Championship Winners - Simulation

and the Western Conference. In a best-of-7 series, the team finishing higher in the overall league standings gets the home-court advantage. Played in a first-to-4 format, the first team to reach 4 wins will be declared the champion. We look at the simulations of the NBA Finals in Figure 5.4. Milwaukee Bucks, the highest-ranked team in the league in the 2022-23 season were the overwhelming favorites to win the league with about 45% of the simulations won by them. They are followed by Boston Celtics (21.8%) and Denver Nuggets (15.2%)

5.2.1 Comparison with the ongoing season

With Milwaukee Bucks being the best side in the regular season, experts all over the world picked the Bucks to win the championship. According to Sports Odds History [6], the Bucks were the favorites to win the league. But, Miami Heat pulled off one of the biggest upsets in NBA history, by knocking out the 2020-21 champions. Our simulation had this exact first-round matchup 2527 times out of the 10,000 simulations. The win probabilities for this matchup look like this:

Teams	Winning Probability
Miami Heat	7.5%
Milwaukee Bucks	92.5%

Table 5.3: Heat v Bucks - First Round Simulation

With the Miami Heat having one of the best playoff runs, our model's results deviated from the actual results. What is more fascinating about this playoff run is that the Miami Heat were one of the teams in the play-in tournament. Among all the 10,000 simulations run, the Miami Heat's championship-winning simulations were at 0.04% (4 out of 10,000 simulations). At the point of submission, the Heat have reached the NBA Finals and have made themselves a solid case to win the championship.

Another team that exceeded expectations compared with our models was the Los Angeles Lakers. With a lackluster beginning to the season, the Lakers had a significant turnaround after a few mid-season trades. Their form picked up significantly and they qualified for the play-in tournament. The

Lakers had an intense matchup against the Memphis Grizzlies. The Grizzlies had the best home record this season with a win/loss of 35/6. The Los Angeles Lakers withstood this challenge and beat the Grizzlies 4-2 and eventually made the Western Conference finals. Our simulation had the Lakers-Grizzlies first-round matchup 5970 times out of the 10,000 simulations. The win probabilities for this matchup looked like this:

Teams	Winning Probability
Los Angeles Lakers	16.7%
Memphis Grizzlies	83.3%

Table 5.4: Lakers v Grizzlies - First Round Simulation

5.3 Improvements

A model's predictive power is thrown off when an upset[3] happens in the league. Although predicting an upset is an arduous task, if one can make the model as comprehensive as possible, that should improve the results churned out by the model. One of the things that we have not covered is the dynamic nature of sporting tournaments. In big tournaments like the NBA, a team can go through a series of ups and downs over the course of a regular season. A team's form at the end of the regular season can affect their playoff run. This could be explored more in detail in the future. Essentially, if we are able to quantify a team's form, we would be able to give out better results ensuring higher predictive power.

Cattelan, Varin, and Firth, in 2013 [11] came up with the idea of modeling home-ground advantage in a dynamic manner. They proposed that the concept of 'home ability' and 'away ability' is assumed to be evolving over time following an Exponentially Weighted Moving Average (EWMA) pro-

[3]Sporting slang when a lower-ranked team beats a higher-ranked team

cess. The dynamic model proposed for the NBA looks like this:

$$\mathbb{P}(Y_i = 1 | Y_{i-1} = y_{i-1} .... Y_1 = y_1) = \frac{\exp(a_{h_i}(t_i) - a_{v_i}(t_i))}{1 + \exp(a_{h_i}(t_i) - a_{v_i}(t_i))}$$

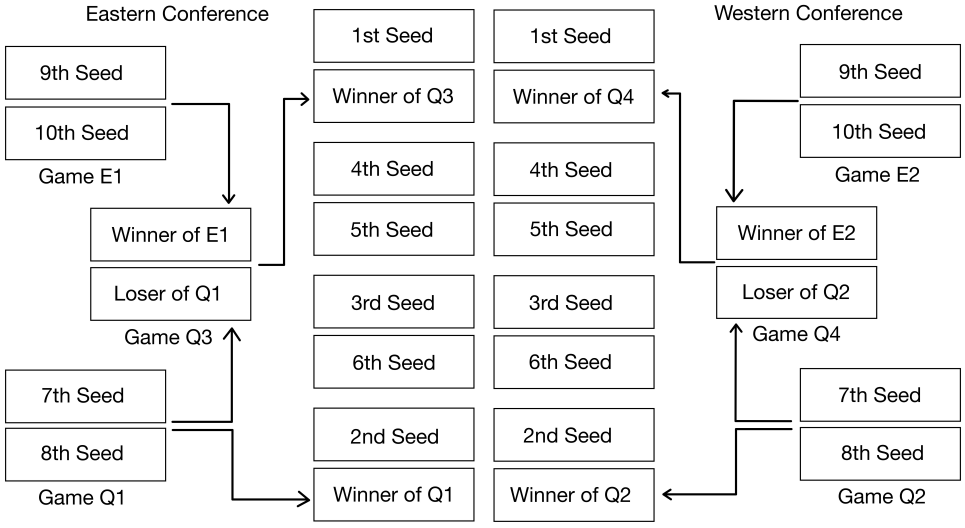
where Y_i denotes the result of the i^{th} matchup, which is binary in the case of NBA. $a_{h_i}(t_i)$ and $a_{v_i}(t_i)$ are the new concepts 'home' and 'away' ability introduced in the paper. Cattelan et al represented these abilities in equation form like this:

$$a_{h_i}(t_i) = \lambda_1 \beta_1 r_{h_i}(t_{i-1}) + (1 - \lambda_1) a_{h_i}(t_{i-1})$$
$$a_{v_i}(t_i) = \lambda_2 \beta_2 r_{v_i}(t_{i-1}) + (1 - \lambda_2) a_{v_i}(t_{i-1})$$

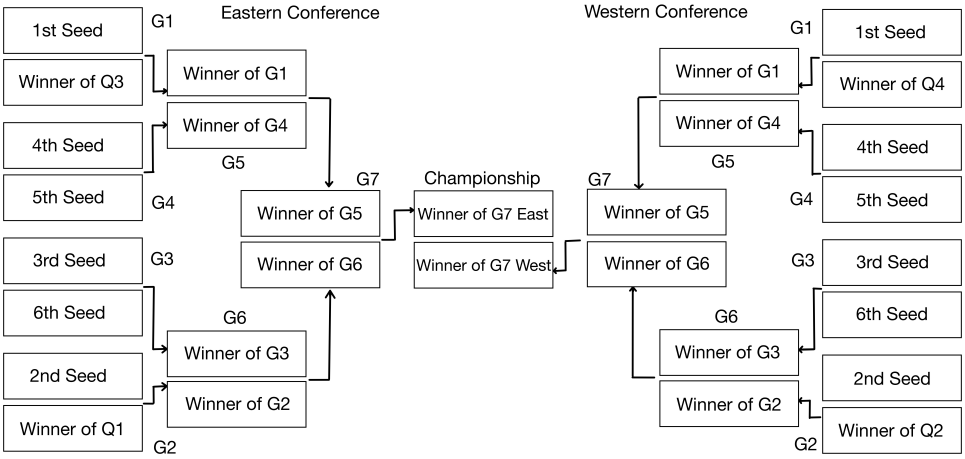
where β_1 and β_2 being home-specific and visitor-specific parameter and λ_1 and λ_2 are the smoothing parameters. If λ_1 is equal to 0, this means that home abilities are constant throughout the period, which is an example of our models. $r_{h_i}(t_{i-1})$ and $r_{v_i}(t_{i-1})$ represent the results of the previous home game, for the home team, and previous away game for the away team respectively. With the initial condition for r_{h_i} and r_{v_i} at the beginning of the season needing to be set, Cattelan et al use the previous season's average home and away results as the initial condition.

With the 2009-2010 NBA season to be modeled, $\bar{r_{h_i}}$ was set to be 0.608, based on the frequency of the victories of home teams in the previous season. With an improved likelihood value compared to the model with $\lambda_1 = \lambda_2 = 0$, there was evidence to the claim that NBA systems are dynamic and form might exist as a result. Further examination of this approach with multiple seasons of data might help us understand the dynamic nature more accurately.

Appendix

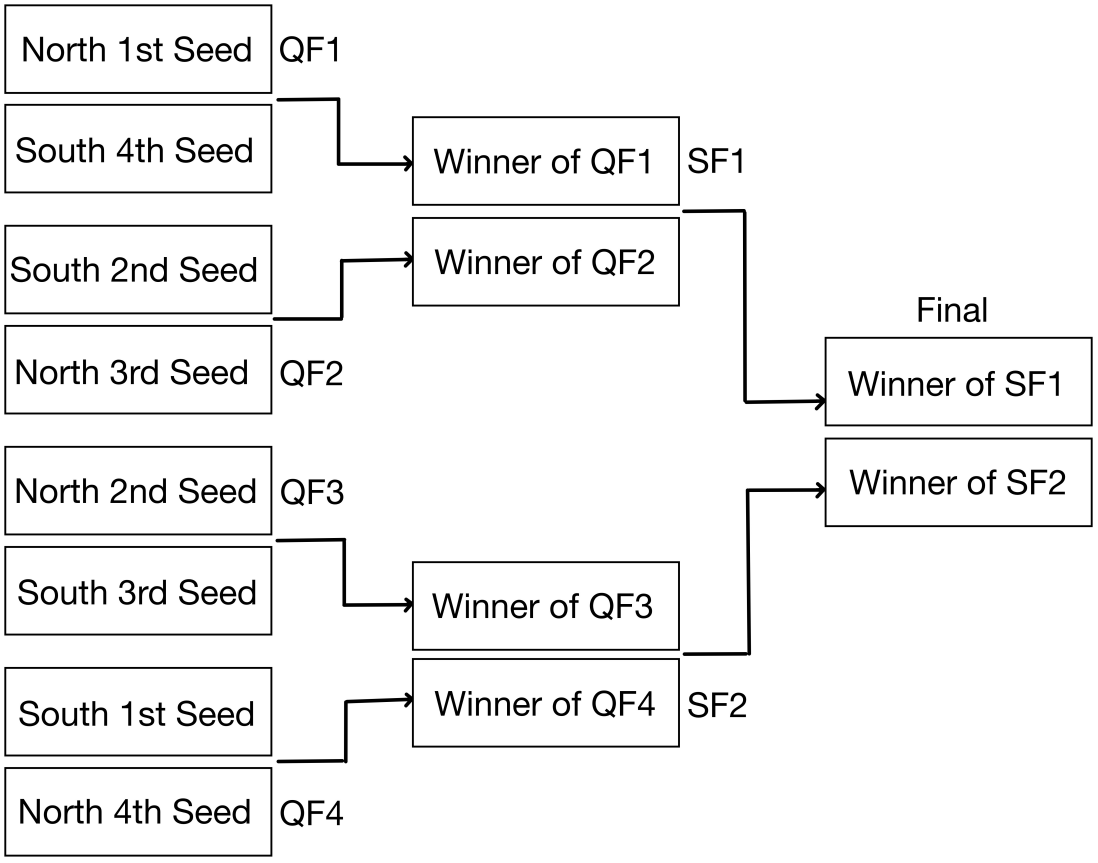


Appendix A: Play-in Tournament Format - NBA



Appendix B: Playoffs Format - NBA

In the entire postseason of NBA, the higher seed gets the home-ground advantage of playing home games in the 1st, 2nd, 5th, and 7th games of a playoff series. In the play-in tournament, the higher seed gets the home-court advantage.



Appendix C: Knockouts format - T20 Blast

Division	Team	Location
Eastern Conference		
Atlantic	Boston Celtics	Boston, Massachusetts
	Brooklyn Nets	New York City, New York
	New York Knicks	New York City, New York
	Philadelphia 76ers	Philadelphia, Pennsylvania
	Toronto Raptors	Toronto, Ontario
Central	Chicago Bulls	Chicago, Illinois
	Cleveland Cavaliers	Cleveland, Ohio
	Detroit Pistons	Detroit, Michigan
	Indiana Pacers	Indianapolis, Indiana
	Milwaukee Bucks	Milwaukee, Wisconsion
Southeast	Atlanta Hawks	Atlanta, Georgia
	Charlotte Hornets	Charlotte, North Carolina
	Miami Heat	Miami, Florida
	Orlando Magic	Orlando, Florida
	Washington Wizards	Washington, D.C.
Western Conference		
Southwest	Dallas Mavericks	Dallas, Texas
	Houston Rockets	Houston, Texas
	Memphis Grizzlies	Memphis, Tennessee
	New Orleans Pelicans	New Orleans, Louisiana
	San Antonio Spurs	San Antonio, Texas
Pacific	Golden State Warriors	San Francisco, California
	Los Angeles Clippers	Los Angeles, California
	Los Angeles Lakers	Los Angeles, California
	Phoenix Suns	Phoenix, Arizona
	Sacramento Kings	Sacramento, California
Northwest	Denver Nuggets	Denver, Colorado
	Minnesota Timberwolves	Minneapolis, Minnesota
	Oklahoma City Thunder	Oklahoma City, Oklahoma
	Portland Trail Blazers	Portland, Oregon
	Utah Jazz	Salt Lake City, Utah

Appendix D: Teams of the NBA

North Group		**South Group**	
Team	Location	Team	Location
Derbyshire Falcons	Derbyshire	Essex Eagles	Essex
Durham	Durham	Glamorgan	Glamorgan
Lancashire Lightning	Lancashire	Gloucestershire	Gloucestershire
Leicestershire Foxes	Leicestershire	Hampshire Hawks	Hampshire
Northamptonshire Steelbacks	Northamptonshire	Kent Spitfires	Kent
Nottinghamshire Outlaws	Nottinghamshire	Middlesex	Middlesex
Warwickshire Bears	Warwickshire	Somerset	Somerset
Worcestershire Rapids	Worcestershire	Surrey	Surrey
Yorkshire Vikings	Yorkshire	Sussex Sharks	Sussex

Appendix E: Teams of the T20 Blast: Current Hierarchy

North Group	**Midlands/Central Group**	**South Group**
Derbyshire Falcons	Glamorgan	Essex Eagles
Durham	Gloucestershire	Hampshire Hawks
Lancashire Lightning	Northamptonshire Steelbacks	Kent Spitfires
Leicestershire Foxes	Somerset	Middlesex
Nottinghamshire Outlaws	Warwickshire Bears	Surrey
Yorkshire Vikings	Worcestershire Rapids	Sussex Sharks

Appendix F: Teams of the T20 Blast: Previous Hierarchy

Teams	Wins	Losses	PCT
Milwaukee Bucks	251	118	0.6802168
Utah Jazz	207	130	0.6142433
Philadelphia 76ers	214	135	0.6131805
Denver Nuggets	216	141	0.6050420
Toronto Raptors	211	138	0.6045845
Boston Celtics	215	149	0.5906593
Los Angeles Clippers	205	143	0.5890805
Miami Heat	201	151	0.5710227
Golden State Warriors	194	153	0.5590778
Phoenix Suns	189	155	0.5494186
Los Angeles Lakers	183	152	0.5462687
Dallas Mavericks	184	158	0.5380117
Brooklyn Nets	178	156	0.5329341
Memphis Grizzlies	170	159	0.5167173
Portland Trail Blazers	169	169	0.5000000
Indiana Pacers	153	166	0.4796238
San Antonio Spurs	150	166	0.4746835
Charlotte Hornets	138	165	0.4554455
Oklahoma City Thunder	143	177	0.4468750
Atlanta Hawks	146	182	0.4451220
Houston Rockets	145	186	0.4380665
Sacramento Kings	131	177	0.4253247
New Orleans Pelicans	134	182	0.4240506
Minnesota Timberwolves	127	180	0.4136808
Washington Wizards	128	187	0.4063492
Chicago Bulls	122	184	0.3986928
New York Knicks	117	190	0.3811075
Orlando Magic	120	199	0.3761755
Cleveland Cavaliers	104	199	0.3432343
Detroit Pistons	104	202	0.3398693

Appendix G: Win-loss Records in the NBA - 2018-19 to 2021-22

Teams	Home Wins	Home Losses	PCT
Lancashire Lightning	19	4	0.8260870
Somerset	20	9	0.6896552
Surrey	16	10	0.6153846
Gloucestershire	12	8	0.6000000
Worcestershire Rapids	13	9	0.5909091
Hampshire Hawks	13	9	0.5909091
Yorkshire Vikings	14	10	0.5833333
Nottinghamshire Outlaws	16	12	0.5714286
Kent Spitfires	14	11	0.5600000
Essex Eagles	11	11	0.5000000
Sussex Sharks	11	11	0.5000000
Warwickshire Bears	13	14	0.4814815
Glamorgan	10	14	0.4166667
Middlesex	9	13	0.4090909
Derbyshire Falcons	9	13	0.4090909
Durham	11	16	0.4074074
Northamptonshire Steelbacks	10	15	0.4000000
Leicestershire Foxes	8	18	0.3076923

Appendix H: Home Win-loss Records in the T20 Blast - 2018 to 2022. Just like throughout the paper, the 2020 season was ignored due to a difference in the hierarchical structure in the league.

Bibliography

[1] Alan Agresti. *Bradley-Terry Model for Paired Preferences*, page 436–438. John Wiley and; Sons (etc.), III edition, 2013.

[2] Murray Aitkin. A History of the GLIM Statistical Package. 2018.

[3] A. C. Atkinson. A test of the linear logistic and bradley-terry models. *Biometrika*, 59(1):37–42, 1972.

[4] Corporate Author: basketball reference.com.

[5] R. J. Beaver and D. V. Gokhale. A model to incorporate within-pair order effects in paired comparisons. *Communications in Statistics*, 4(10):923–939, 1975.

[6] Blake and Blake. Nba championship favorites entering the playoffs, Apr 2023.

[7] Ralph Allan Bradley. Rank analysis of incomplete block designs: II. additional tables for the method of paired comparisons. *Biometrika*, 41(3/4):502–537, 1954.

[8] Ralph Allan Bradley. Rank analysis of incomplete block designs: III some large-sample results on estimation and power for a method of paired comparisons. *Biometrika*, 42(3/4):450–470, 1955.

[9] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324, 1952.

[10] N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

[11] Manuela Cattelan, Cristiano Varin, and David Firth. Dynamic Bradley–Terry Modelling of Sports Tournaments. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 62(1):135–150, 05 2012.

[12] Christopher F. Chabris and Mark E. Glickman. Sex differences in intellectual performance: Analysis of a large cohort of competitive chess players. *Psychological Science*, 17(12):1040–1046, 2006.

[13] D. R. Cox. *The Analysis of Binary Data*. 1st edition, 1970.

[14] Douglas E. Critchlow and Michael A. Fligner. Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation on GLIM. *Psychometrika*, 56(3):517–533, 1991.

[15] Roger R. Davidson. On extending the bradley-terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328, 1970.

[16] Roger R. Davidson and Robert J. Beaver. On extending the bradley-terry model to incorporate within-pair order effects. *Biometrics*, 33(4):693–702, 1977.

[17] Corporate Author: espncricinfo.com.

[18] Stephen E. Fienberg and Kinley Larntz. Log linear representation for paired and multiple comparisons models. *Biometrika*, 63(2):245–254, 1976.

[19] David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993.

[20] David Firth. Bradley-terry models in R. *Journal of Statistical Software*, 12:10, 01 2005.

[21] Francesca Giambona, Mariano Porcu, and Isabella Sulis. Students mobility: Assessing the determinants of attractiveness across competing territorial areas. *Social Indicators Research*, 133(3):pp. 1105–1132, 2017.

[22] W. A. Glenn and H. A. David. Ties in paired-comparison experiments using a modified thurstone-mosteller model. *Biometrics*, 16(1):86–109, 1960.

[23] Coen Jones. Hierarchical bradley-terry models. Honours thesis, University of Queensland, 2021.

[24] Nils Kousgaard. Models for paired comparisons with ties. *Scandinavian Journal of Statistics*, 3(1):1–14, 1976.

[25] J. N. S. Matthews and K. P. Morris. An application of bradley-terry-type models to the measurement of pain. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(2):243–255, 1995.

[26] A. E. Maxwell. The logistic transformation in the analysis of paired-comparison data. *British Journal of Mathematical and Statistical Psychology*, 27(1):62–71, 1974.

[27] Danny McConnell. American and national league.

[28] The Editors of Encyclopaedia Britannica, May 2023.

[29] Neil Paine. Paine: Unique Denver, Utah Home Advantages, Mar 2013.

[30] W. Katzenbeisser R. Dittrich, R. Hatzinger. Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 47(4):511–525, 1998.

[31] P. V. Rao and L. L. Kupper. Ties in paired-comparison experiments: A generalization of the bradley-terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967.

[32] From NBA.com Staff. Everything you need to know about the 2019-20 nba season restart, Jul 2020.

[33] Steven E Stern. The duckworth-lewis-stern method: Extending the duckworth-lewis methodology to deal with modern scoring rates. *Journal of the Operational Research Society*, 67(12):1469–1480, 2016.

[34] Devi M. Stuart-Fox, David Firth, Adnan Moussalli, and Martin J. Whiting. Multiple signals in chameleon contests: designing and analysing animal contests as a tournament. *Animal Behaviour*, 71(6):1263–1271, 2006.

[35] Heather Turner and David Firth. Bradley-terry models in R: The bradleyterry2 package. *Journal of Statistical Software*, 48(9):1–21, 2012.

[36] E. Zermelo. Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29(1):436–460, 1929.