Hierarchical Bradley-Terry Models: Refining and Improving the Model Mechanism, Considering Neutral Venue Scenario

Project Report - MSc Statistics

Jinxi Zhang

Supervisor: Alan Huang

November 2024

Contents

1. Abstract

1.1. Project Background

This project is primarily based on the article “Hierarchical Bradley-Terry Models: Extension to Incorporate Neutral Venue Scenario” by previous researcher Vanru Ranganath. In his original work, a model was established, and one of the Hierarchical Bradley-Terry models was selected to predict the outcomes of a specific season for the NBA and T20. The results were within expectations. This project made several optimizations and improvements, including code optimization, comparison of predictions from multiple models, and the addition of an interactive interface for users to freely select the models and outcomes they wish to predict. To validate the model's improvement, we also predicted the results of the latest NBA 2023-2024 season, and the model's predictions were consistent with real-world outcomes, demonstrating its value and applicability.

1.2. Project Significance

In this study, we propose an improved statistical model for predicting NBA game outcomes. Previous models primarily relied on the win-loss record of teams, but this single variable has limitations. First, teams often perform better within the same division, and these models generally overlook the home-court advantage across different divisions. For example, some teams are better able to utilize home-court conditions due to schedule arrangements or geographical advantages, improving their win rate. Therefore, solely relying on the win-loss record does not fully reflect the actual performance and potential strength of a team.

The importance of this project also lies in its potential for practical applications. Betting companies can use this model to make more accurate betting decisions, thereby enhancing their future revenue expectations. By integrating more complex factors such as home-court advantage and division differences and deeper statistical analysis, our model can provide more precise predictions for betting companies, increasing their profitability. In terms of predicting future game outcomes, this model has promising applications, providing useful insights not only to NBA fans and data analysts but also offering economic benefits to the betting industry.

2. Project Review

2.1. Overview of the Original Project

The original article reviews the principles of the Bradley-Terry model and applies likelihood functions to build various models for predicting wins and losses in sports events. One of these models, the Hierarchical Bradley-Terry model, was selected to predict a specific season of the NBA and T20. The article introduces multiple data details and presents the results of the model's performance. Finally, multiple simulations were conducted, and the experimental results are convincing.

2.2. Models Involved in the Project

Some common notation that has been used in the models is listed below:

- 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 - The 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 $\alpha_k = \ln \gamma_k$.

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 $IP(i > j)$
- 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 IP$(i > j)$: The initial equation of the model looks like this:

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

Likelihood Function: The likelihood function of the model is:

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

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:

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

Gradient to the log-likelihood Function: Considering the partial derivative with respect to θ_i, the result is:

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

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 IP$(i > j)$: The initial equation of the model looks like this:

$$IP(i > j)_i = \frac{\lambda_i \gamma}{\lambda_i \gamma + \lambda_j}, \quad IP(i < j)_i = \frac{\lambda_j}{\lambda_i \gamma + \lambda_j}$$

Likelihood Function: The likelihood function of the model is:

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

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:

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

Gradient to the log-likelihood Function: The gradient with respect to θ_i is given by:

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

and the gradient with respect to α is:

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

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 λ_s for each team and the n home-ground advantage parameters γ_s. γ_n represents the home-ground advantage at each level of the hierarchy. 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.

Representation of IP$(i > j)$**:** The initial equation of the model looks like this:

$$IP(i > j)_i = \frac{\lambda_i \gamma_{R_{ij}}}{\lambda_i \gamma_{R_{ij}} + \lambda_j}, \quad IP(i < j)_i = \frac{\lambda_j}{\lambda_i \gamma_{R_{ij}} + \lambda_j}$$

Likelihood Function: The likelihood function of the model is:

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

Log-likelihood Function: The log-likelihood function looks as follows:

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

Gradient to the log-likelihood Function: The gradient with respect to θ_i is given by:

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

and the partial derivative with respect to α_n is:

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

2.2.4. Team-specific Home-ground Advantage Model

Number of Parameters: The model has $2N$ parameters, where we try to obtain the parameter λ_s for each team and the individual home-ground advantage parameter γ_i. We assume that the γ_i parameter remains constant throughout the period chosen for analysis.

Representation of IP$(i > j)$: The initial equation of the model looks like this:

$$
IP(i > j)_i = \frac{\lambda_i\gamma_i}{\lambda_i\gamma_i + \lambda_j}, \quad IP(i < j)_i = \frac{\lambda_j}{\lambda_i\gamma_i + \lambda_j}
$$

Likelihood Function: The likelihood function of the model is:

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

Log-likelihood Function: The log-likelihood function looks as follows:

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

Gradient to the log-likelihood Function: The gradient with respect to θ_i is:

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

and the partial derivative with respect to α_i is:

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

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 λ_s value for each team and the $n \times N$ home-ground advantage parameter $\gamma_{i,n}$ for each team and level of hierarchy. R_{ij} represents the level of relationship in the hierarchy between team i and team j.

Representation of IP$(i > j)$: The initial equation of the model looks like this:

$$
IP(i > j)_i = \frac{\lambda_i\gamma_{i,R_{ij}}}{\lambda_i\gamma_{i,R_{ij}} + \lambda_j}, \quad IP(i < j)_i = \frac{\lambda_j}{\lambda_i\gamma_{i,R_{ij}} + \lambda_j}
$$

Likelihood Function: The likelihood function of the model is:

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

Log-likelihood Function: The log-likelihood function looks as follows:

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

Gradient to the log-likelihood Function: The gradient with respect to θ_i is:

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

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

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

2.3. Previous Code Processing

In the previous project, the code is divided into four parts. The first part, `Functions.R`, reads the raw data and converts it into a matrix format suitable for modeling. Additionally, the `ntheta` function is defined to quantify each team's advantage, and six model formulas are constructed through likelihood functions, including the home advantage model.

The second part, `Implementation.R`, uses the `BT_Model` function to integrate normalized data and models, defining the model's operation process. `BT_test` is used for significance testing, `BT_plots` for generating ranking charts, and `BT_predict` for win-rate predictions.

The third part, `Simulation.R`, defines a series of functions for NBA game predictions, including `Simulation_Play_In`, `PlayInTournament`, `Seeds` rankings for

East and West divisions, `playoffs` simulations, and finals simulations. Some code sections were hard-coded, and optimizations will be discussed later.

The fourth part, `SimulationNBA.R`, is prediction code for the NBA 2022-23 season, specifically tailored to the `Common Hierarchical model`. Future work will involve adjustments to predictions across five formulas.

2.4. Limitations of the Original Project

2.4.1. Hard Code

The previous author's code structure was not clear enough, and the explanations were not precise, making it difficult for new users to learn and conduct in-depth research, which could cause significant resistance. Many variable names were unclear, especially when using subsequent equations. When new data was imported, it could not run correctly, and errors were difficult to locate and fix in time.

2.4.2. Number of Models

Only one model was used for a large number of predictions, and no other statistical models were employed for separate predictions.

2.4.3. Lack of User Options

There was no interactive interface, and users could not easily select models or parameters. In the previous project, the author used fixed data and models for analysis without considering making the project available to more users. Thus, new users had no way to choose the model they wanted to simulate. Furthermore, during multiple simulations, the author fixed the simulation count at 10,000, but sometimes this large amount of calculation would take up too much time. When users wanted to run multiple simulations, they might only want a fixed number of simulations, but the author did not provide options for users to select.

3. Code Improvement and Optimization

3.1. Debugging Process and Error Correction

In the original author's project, numerous hard-coded sections were present. In this project, we conducted various debugging adjustments and refined specific code details, resulting in improvements that make the entire code smoother, more understandable, and more streamlined.

3.1.1. Plot Section

First, in the `Implementation.R` code file, within the `BT_plots` function, selecting the "Common Hierarchical Home-ground Advantage Model" currently prevents the plot from being generated, with issues arising in the code execution. When choosing the "Team-specific Home-ground Advantage Model," running the plot function also fails to produce a chart. Furthermore, in the TSH model, only each team's original `Theta` and separate home advantage values are displayed, resulting in plots sorted by original `Theta` values from that season, making it challenging to observe the ranking under the current model directly. Similarly, the code for the "Hierarchical Home-ground Advantage Model" faces this issue, as plot sorting appears unsatisfactory.

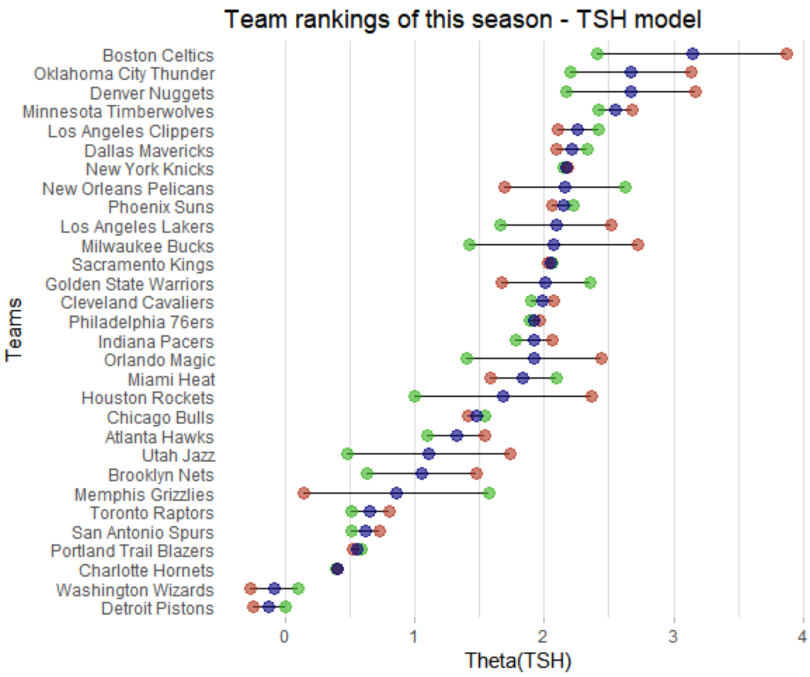

Figure 1: Team rankings of TSH model

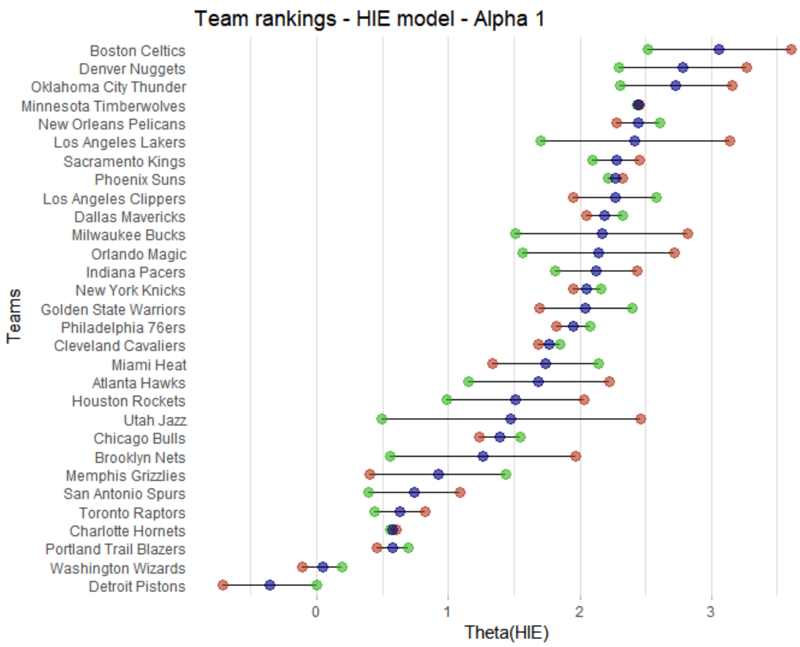

Figure 2: Team rankings of HIE model in Alpha1

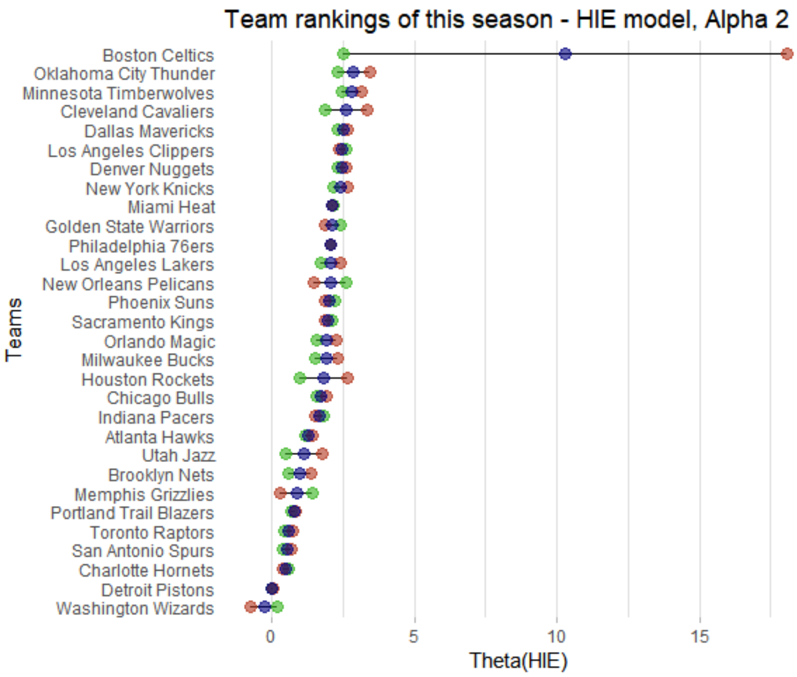

Figure 3: Team rankings of HIE model in Alpha2

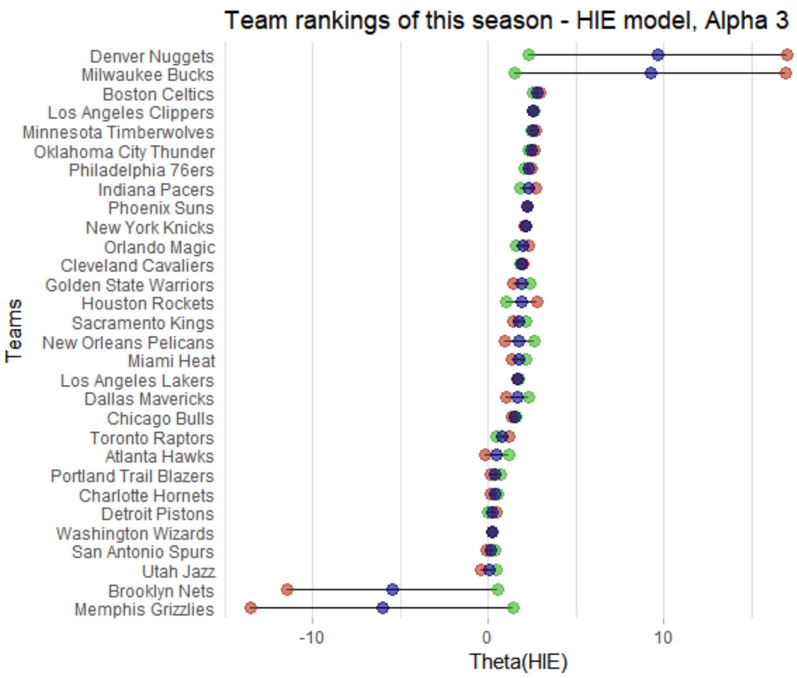

Figure 4: Team rankings of HIE model in Alpha3

We have thus improved the code for the above models. In the TSH and HIE code sections, I introduced `midpoint` to represent the average of `Theta` and home values, which is plotted in navy blue, as shown below. This adjustment enhances visibility of each team's ranking across models, aiding in understanding differences between this and other models. Previously, in the CHI and HIE sections, three different `alpha` values generated one plot per value with a 10-second pause between each. To streamline user experience, I adjusted the plot logic, storing plots in a list, so users see all three `alpha` plots once they're ready.

3.1.2. Seeds order

The ranking orders of the teams that may go in playoffs of the season we focus on. There is a flaw that the final simulation function requires manual input of `Seed_No` to switch the order of the home and away teams in the finals. Previous code manually inputs the top 20 official records predicted for the season, which is understandable as there should be 10 teams from each conference participating in the play-in and playoffs. However, a flaw remains: it does not consider the differences in records between Eastern and Western conference teams.

```r
Seeds_Tournament = function(){
  # Seeds of the entire league with an example of the 2022-23 season
  #https://www.espn.com/nba/standings/_/season/2023/group/league
  Seed_No <- c(1:21)
  Team <- c("Boston Celtics", "Oklahoma City Thunder", "Denver Nuggets", "Minnesota Timberwolves",
            "Los Angeles Clippers",  "Dallas Mavericks", "New York Knicks", "Milwaukee Bucks",
            "New Orleans Pelicans", "Phoenix Suns", "Cleveland Cavaliers", "Orlando Magic",
            "Los Angeles Lakers", "Philadelphia 76ers", "Indiana Pacers", "Miami Heat",
            "Sacramento Kings", "Golden State Warriors", "Houston Rockets","Chicago Bulls",
            "Atlanta Hawks")
  Seeds_Tournament <- data.frame(Seed_No, Team)
  return(Seeds_Tournament)
}

Final = function(East, West, ST){
  # East and West are the 7 games that have been completed in both the playoffs.
  # ST is the seeds of the tournament
  Final <- data.frame(matrix(ncol = 2))
  colnames(Final) <- c("Team.A", "Team.B")
  Final[,1] <- East$Winner[7]
  Final[,2] <- West$Winner[7]
  if ((ST$Seed_No[ST$Team == Final[,1]]) > (ST$Seed_No[ST$Team == Final[,2]])){
    Final[,1:2] <- Final[c("Team.B", "Team.A")]
  }
  return(Final)
  browser()
}
```

Figure 5: Manual input Seeds ranking of 2022-23 season

For example, in 2023, the Hawks were ranked 21st in the entire league and 10th in the East, qualifying for the play-in. The Rockets, on the other hand, were ranked 19th in the league but 11th in the West, thus missing the play-in. Following the previous code's logic, the top 20 would include the Rockets but not the Hawks. Therefore, in one of the 10,000 simulations where the Hawks win the Championship, adjusting the order based on `Seed_No` would not be possible.

The improvement here should involve manually inputting the league-wide rankings in `Seed_No`. If the top 15 teams are all from the Western conference, the names of at least the top 25 teams of the season should be input, especially considering expansion to 32 teams with Alaska and Seattle additions.

Regarding the seeds list, knowing which team has home advantage when different teams compete during the play-in and playoffs simulations is crucial. After the NBA regular season, a total ranking is available. Whether for the Eastern Conference playoffs, Western Conference playoffs, or the finals, the better-ranked team on the total ranking list holds the home advantage. In both previous and current code, seeds are manually obtained. Future researchers may face an issue if they ignore the input order for different seeds, leading to potential mistakes such as missing or misentering names. While the previous author's issue of manually inputting the top 20 list has been noted in this experiment, I also entered the new seeds list manually.

Subsequent improvements could involve directly scraping the official website's ranking list in the corresponding R code and importing it into the code. This approach is both simple and ensures sustainable, error-free seeds for future projects. The previous author did not provide the source website of seeds. Based on the original data, I located the relevant source website.[1] This site also provides the team rankings of the East and West Divisions and the entire league each NBA season.

3.2. Refining function

1. The code for simulating the Eastern and Western Conference playoffs, as well as the Finals, has been modified. In the original code, the simulation results for the Eastern and Western Conferences were stored as a single variable, allowing only one simulation per run, which was inconvenient for subsequent interactive use. After modification, the simulation code for the Eastern and Western Conferences and the Finals has been converted into a function, enabling repeated use for multiple simulations at the same stage.

```
East <- function() {
  result <- Simulation_Playoff_East(playoffs_East, Seedsofeast, ProbabilitiesCHI)
  EastResult <<- result
  return(EastResult)
}
```

Figure 6: New function 1

2. The code for all series matchups has been improved. Not only was the calculation method for the home court advantage corrected, but two additional columns,

[1]https://www.espn.com/nba/standings/_/season/2024/group/league

```
West <- function() {
  result <- Simulation_Playoff_West(playoffs_West, Seedsofwest, ProbabilitiesCHI)
  WestResult <<- result
  return(WestResult)
}
```

Figure 7: New function 2

```
Simulation_Playoff_East = function(df, SE, Probabilities){
  # df is the fixture list of the Eastern Conference playoffs
  # Probabilities is the results of BT_predict for the entire
  # SE is the Seeds of the Eastern Conference obtained from Se
  p <- df
  p[1,2] <- SE[8,2]
  p[2,2] <- SE[7,2]
  p$Home_Win <- NA
  p$Winner <- NA
  p$TeamA_home_win_prob <-NA
  p$TeamA_away_win_prob <-NA
```

Figure 8: New function 3

`TeamA_home_win_prob` and `TeamA_away_win_prob`, were also added in the results section. These columns display the home team's win probabilities in each specific matchup, making each set of results clearer. This not only reduces the chance of unexpected errors but also enables users to better understand the conditions of both competing teams.

3. The final NBA simulation code `NBA()` has been improved. This function is designed to simulate the entire playoff series and return the winner, allowing for repeated use in multiple simulations. The issue with the original code was that when conducting numerous simulations, such as 10,000 times, it required a substantial waiting time, which varied depending on the device used. After testing, the updated code reduces the time required for large-scale simulations by approximately 70%, a significant improvement, without compromising the accuracy of the results.

```
NBA <- function() {
  # Combine all the functions to obtain results of the simulations
```

Figure 9: New function 4

3.3. Implementation of R Interface Interaction

When all the code was corrected, and additional equations were included, a brand new interface was added in RStudio for user interaction, both to facilitate future researchers and to support the ideal prospect of the model—to become an interactive website for users. The main commands of this interface are as follows:

```
> choose_model()
Choose the model you want to use: VAN, CHA, CHI, TSH, HIE: TSH
```

Figure 10: Interaction 1

A very simple interactive interface, where, after executing these lines of code, the user will be prompted with model selection options, allowing them to choose the desired model for further analysis.

After making a selection, the following commands are available:

```
Choose the next step: Modelplot,Seeds, East, West, Finalgame,One playoffs, Mutisimulation: West
```

Figure 11: Interaction 2

- `Modelplot`: This command will display a plot of team rankings for the selected model for the current season, with some examples placed in the directory section.
- `Seeds`: This command will display a ranking list of the top 8 teams from each conference, based on the results of the play-in tournament for the current season, for further simulations.
- `East`: This command will simulate an Eastern Conference playoff once and display the results. The user can repeat this command, and each time a different playoff result will be generated.
- `West`: This command will simulate a Western Conference playoff once and display the results. Similarly, the user can repeat this command for different playoff outcomes.
- `Finalgame`: This command will simulate the final game between the latest simulated Eastern and Western Conference champions. The user can also repeatedly simulate multiple finals.
- `One playoffs`: This command will simulate a complete playoff process for both conferences, leading up to the final champion.
- `Mutisimulation`: After selecting a model, the user can directly generate multiple simulation prediction plots. Example plots are provided in later sections of the paper.

After the user selects any of the above options, the following interactions are available:

```
Simulate another? Choose 'Yes', 'Next', or 'Stop': Yes
Choose the next step: Modelplot,Seeds, East, West, Finalgame,One playoffs, Mutisimulation:
```

Figure 12: Interaction 3

- `Yes`: This command returns the user to the previous level, allowing them to simulate other parts.
- `Next`: This command initiates multiple simulations with the following prompt:

```
Simulate another? Choose 'Yes', 'Next', or 'Stop': Next
The times you want to simulate:
```

Figure 13: Interaction 4

  At this point, the user can choose the desired number of simulations, and the result graph will be generated on the right side.

- `Stop`: This command will end the interaction.

This is a basic interactive logic that has the potential to be elevated to a higher level in the future. Subsequent updates could expand the code into a web-based version, allowing users to make real-time selections for simulations. With this interactive logic, users can independently choose the model they need, simulate specific parts of the playoffs as desired, and even select the number of simulations they wish to run.

4. Simulation After Model Improvement

4.1. Simulation Background

4.1.1. NBA

The National Basketball Association (NBA) was established in 1946 and is the most influential professional basketball league in the world. The NBA brings together the world's finest basketball players and has significantly promoted the development of basketball globally. The league consists of 30 teams, divided into the Eastern Conference and Western Conference, each further subdivided into three divisions.

In recent years, the NBA's global influence has continued to grow, not only in North America but also with a vast fan base in Europe, Asia, and other regions. The league has achieved remarkable success in commercial operations, player development, and global promotion. In recent years, it has also actively advanced reforms in game rules and technological innovations, such as introducing the Replay Center and improving the officiating system, to enhance the fairness and entertainment value of the games.

The design of the season's competition system aims to ensure fairness and intensity of competition. Through the lengthy 82-game regular season, teams need to demonstrate sustained competitiveness and stability. The playoff series format tests teams' on-the-spot performance and strategic adjustments under high-pressure environments. This competition format not only guarantees entertainment value but also provides opportunities for teams of varying strengths to compete.

4.1.2. NBA Schedule

I. Regular Season

The regular season is the first phase of the NBA season, typically starting in October each year and lasting until April of the following year. Each team plays 82 games during the regular season, competing against all other teams in the league. Games are scheduled in a home-and-away format to ensure fair competition. The main purpose of the regular season is to determine each team's ranking and lay the foundation for the contention of playoff spots.

1.Game Schedule. Each team plays 82 games during the regular season, facing the other 29 teams in the league. The opponent distribution is as follows: each team plays 4 games against each of the 4 teams in its own division, totaling 16 games; against the 6 teams from the other two divisions in the same conference, each team plays 4 games (totaling 24 games), and against the remaining 4 teams, they play 3 games each (totaling 12 games). Additionally, each team plays 2 games against each of the 15 teams from the other conference, totaling 30 games. Games are scheduled in a home-and-away format to ensure that each team has equal opportunities to play at home and away.

2.Ranking Rules. Team rankings are primarily based on their winning percentage, calculated as the number of wins divided by the total number of games played. If two or more teams have the same winning percentage, the league ranks them according to the following tie-breakers: first, the head-to-head record, which is the win-loss record among the tied teams; second, the division leader status, with the division leader ranking higher if the teams are in the same division; third, the division record, which is the winning percentage within the same division; fourth, the conference record, which is the winning percentage within the entire conference; fifth, the record against playoff teams, or the winning percentage against teams that have qualified for the playoffs; and finally, the point differential, calculated as total points scored minus total points allowed.

II. Play-in Tournament

1. Background Introduction. The purpose of the Play-in Tournament is to increase the competitiveness for playoff spots, allowing more teams to compete for playoff qualification.

2. Participating Teams. After the regular season, teams ranked 7th to 10th in each conference participate in the Play-in Tournament.

3. Tournament Rules. Game arrangements are as follows: the 7th seed plays the 8th seed, with the winner securing the 7th seed in the playoffs. The 9th seed plays the 10th seed, and the winner of that game faces the loser of the 7th vs 8th game to compete for the 8th seed. The elimination mechanism dictates that teams losing two games are eliminated and cannot enter the playoffs.

4. Impact on Rankings. The results of the Play-in Tournament determine the

7th and 8th seeds in each conference. The tournament increases competitiveness towards the end of the regular season, as teams strive to avoid falling into the Play-in zone or aim to qualify for the Play-in Tournament.

III. Playoffs

The playoffs are the second phase of the NBA season and the most fiercely competitive stage. Both the Eastern and Western Conferences have eight teams entering the playoffs, totaling 16 teams. Teams qualify for the playoffs based on their winning percentage during the regular season. The playoffs adopt a series format, specifically a ”best-of-seven” system, meaning the team that first wins four games advances to the next round. The playoffs are divided into four stages: the first round, conference semifinals, conference finals, and the NBA Finals.

The NBA Finals are the pinnacle of the NBA season, featuring a matchup between the Eastern Conference champion and the Western Conference champion. The Finals also use the ”best-of-seven” format, with the winner earning the NBA championship for that year. Winning the championship not only represents the highest honor for the team in that season but also affirms the team’s overall strength and teamwork.

1. Participating Teams

Directly qualified teams are the top six teams in each conference based on regular season standings. Play-in Tournament winners are the two teams from each conference that win in the Play-in Tournament (the 7th and 8th seeds).

2. Playoff Rules

Matchups in the first round are as follows: 1st seed vs 8th seed, 2nd seed vs 7th seed, 3rd seed vs 6th seed, and 4th seed vs 5th seed. Each round uses a best-of-seven format; the first team to win four games advances to the next round. The game sequence follows a 2-2-1-1-1 format, ensuring that the higher-seeded team has home court advantage.

3. Rankings and Advancement

The advancement path includes progression through the First Round, Conference Semifinals, Conference Finals, and ultimately the NBA Finals. The Eastern Conference Champion and Western Conference Champion compete for the NBA Championship, also using a best-of-seven format.

IV. Interrelation Between Different Stages

1. Connection Between Regular Season and Play-in Tournament

Ranking pressure is present for teams ranked 7th to 10th in the regular season as they must participate in the Play-in Tournament, pushing teams to improve their rankings towards the end of the regular season to avoid the risks associated with the Play-in. Strategic adjustments may also occur, with coaches adjusting player rotations and game strategies to achieve better rankings or prepare for the Play-in Tournament.

2. Connection Between Play-in Tournament and Playoffs

The Play-in Tournament provides teams ranked 9th and 10th with a chance to enter the playoffs, increasing competitive suspense. However, teams participating in the Play-in Tournament may be at a disadvantage in terms of physical condition and preparation compared to teams that directly qualified, which can affect their playoff performance.

3. Continuity of Home Court Advantage

The importance of the regular season is highlighted, as the regular season winning percentage not only determines playoff seeding but also impacts home court advantage in the playoffs. Strategic considerations may lead teams to focus more on winning home games during the regular season to ensure they have more home games in the playoffs.

4.1.3. Home-ground Advantage

For Regular Season, home ground Advantage. Teams usually perform better at home, possibly due to a familiar environment, fan support, and reduced travel fatigue. Strategically, teams aim for more home victories during the regular season to improve their overall winning percentage, secure a higher ranking, and gain home court advantage in the playoffs.

For play-in tournament, higher-ranked teams enjoy home ground advantage in the Play-in Tournament, with games held at their home venues.

For playoffs, home ground advantage is determined based on regular season winning percentage, with the team holding the higher winning percentage receiving home court advantage in the series. Home court advantage can influence the outcome of a series, especially in a decisive Game 7.

4.2. Data Sources and Structure

In this project, we worked on the data of the NBA 2023-2024 season, which mainly included the information records of all regular season team matches. After data cleaning, we obtained 1,231 regular season match records, which contain a lot of relevant information. In order to facilitate model testing, we standardized the data and retained the names of the two teams, the win-loss records, and whether it was home and away.

Historical game data comes from the NBA's official website and reliable public datasets. The data includes results from the regular season and playoffs for the NBA 2023-2024 season.

The sorted data is as follows:

```
> formaldatanew
                    Team.A                Team.B Team.A.win Team.B.win Team.A.home Team.B.home
1       Los Angeles Lakers        Denver Nuggets          0          1           0           1
2             Phoenix Suns Golden State Warriors          1          0           0           1
3          Houston Rockets         Orlando Magic          0          1           0           1
4           Boston Celtics       New York Knicks          1          0           0           1
5       Washington Wizards        Indiana Pacers          0          1           0           1
6            Atlanta Hawks     Charlotte Hornets          0          1           0           1
7          Detroit Pistons            Miami Heat          0          1           0           1
8   Minnesota Timberwolves       Toronto Raptors          0          1           0           1
9      Cleveland Cavaliers         Brooklyn Nets          1          0           0           1
10    New Orleans Pelicans     Memphis Grizzlies          1          0           0           1
```

Figure 14: Head of formal data for 2023-2024 season

Another important dataset is the relationship matrix R, which represents the affiliation between two teams. The `NBA_R()` function takes a formatted data frame of fixtures as input and uses the unique teams in those fixtures to generate the NBA relationship matrix. The relationship levels are assigned as follows:

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

	Boston Celtics	Golden State Warriors	Charlotte Hornets	Detroit Pistons	Indiana Pacers	Orlando Magic	New York Knicks	Toronto Raptors	Houston Rockets
Boston Celtics	0	1	2	2	2	2	3	3	1
Golden State Warriors	1	0	1	1	1	1	1	1	2
Charlotte Hornets	2	1	0	2	2	3	2	2	1
Detroit Pistons	2	1	2	0	3	2	2	2	1
Indiana Pacers	2	1	2	3	0	2	2	2	1
Orlando Magic	2	1	3	2	2	0	2	2	1
New York Knicks	3	1	2	2	2	2	0	3	1
Toronto Raptors	3	1	2	2	2	2	3	0	1
Houston Rockets	1	2	1	1	1	1	1	1	0

Figure 15: Relationship matrix

4.3. Application

After fitting the models to the data of past 5 years 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 2022-23. The hypothesis tests conducted for all the models with some important information about the tests are all provided in the table below. Based on the comparison between high-level and low-level models, we chose the Common Hierarchical Home-ground Advantage mode for simulation in this experiment. In order to make the results more diverse, we again present the running results of different models.

Higher-order Model	Lower-order Model	p-value	^{2} statistic
Common HG	Vanilla	<0.0001	102.6468
Common Hierarchical HG	Common HG	0.0007227554	4.631086
Team-specific HG	Common Hierarchical HG	0.442545	27.39692
Hierarchical HG	Team-specific HG	0.9671788	41.49052

Table 1: LRT for NBA data of 2018-19 to 2022-23

The following figures depict the Common Hierarchical Home-ground Advantage mode and sorted according to the strength of the matchup under different models. We can see that the model fitting results are relatively successful and more realistic. The top five are Boston Celtics, Oklahoma City Thunder, Denver Nuggets, Minnesota Timberwolves and Los Angeles Clippers. Because the NBA 2023-2024 season has ended, we can know from the situation that Boston Celtics won the final championship, so it is reasonable to rank first in the current CHI model. Oklahoma City Thunder, Denver Nuggets, Minnesota Timberwolves, and Los Angeles Clippers are the top four in the Western Conference. Their regular season records are very good, and it is reasonable to rank 2nd to 5th in this model.

Detroit Pistons, Washington Wizards and Charlotte Hornets are the last three teams respectively, which shows that their capabilities throughout the season are relatively weak. Combined with the actual situation, we can know that these three teams are exactly the last three teams in the Eastern Conference, which shows that the model operation is very consistent with the actual situation and the fitting is very successful.

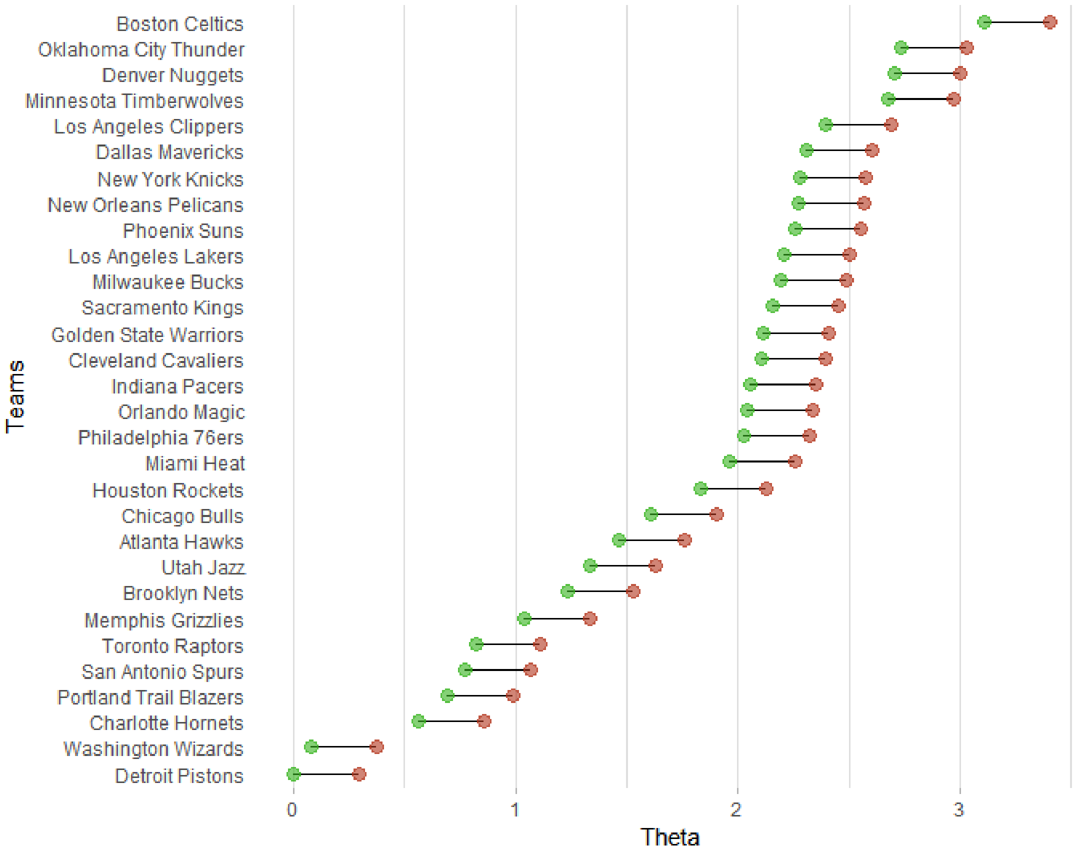

Figure 16: Common Hierarchical Home-ground Advantage model with Alpha 1 NBA 2023-2024

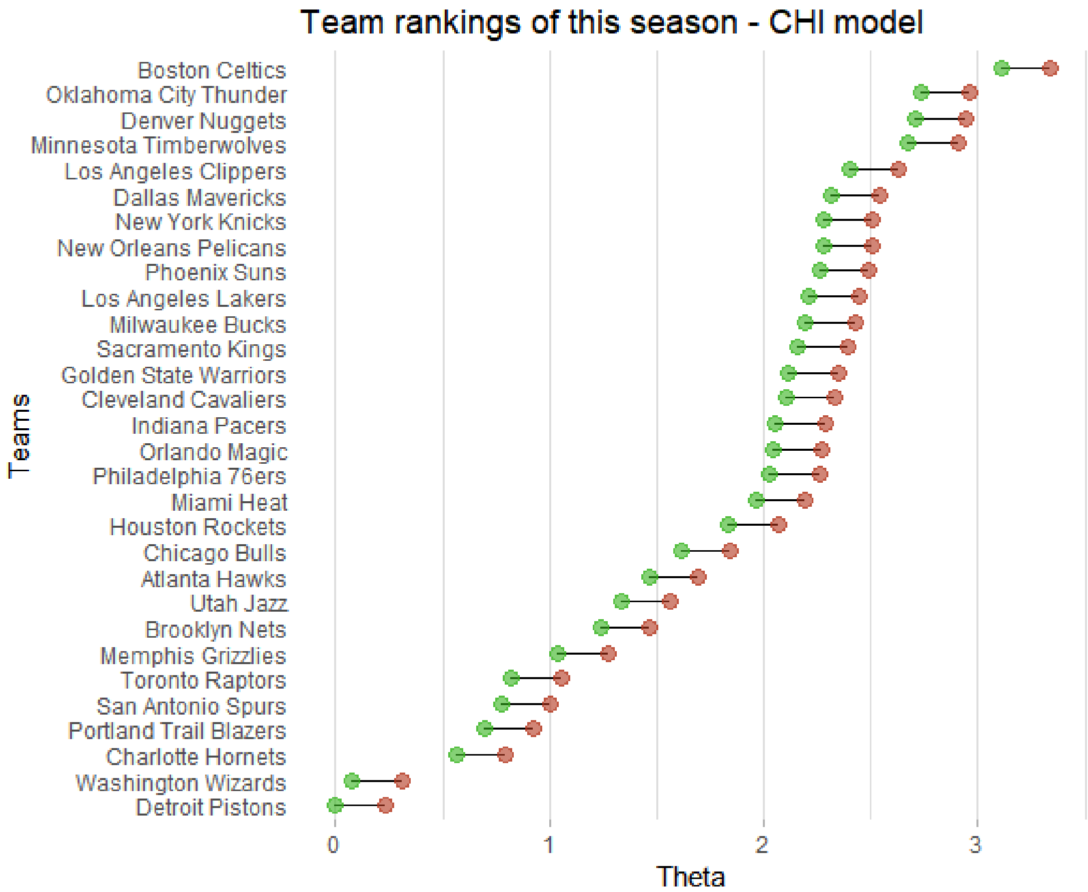

Figure 17: Common Hierarchical Home-ground Advantage model with Alpha 2 NBA 2023-2024

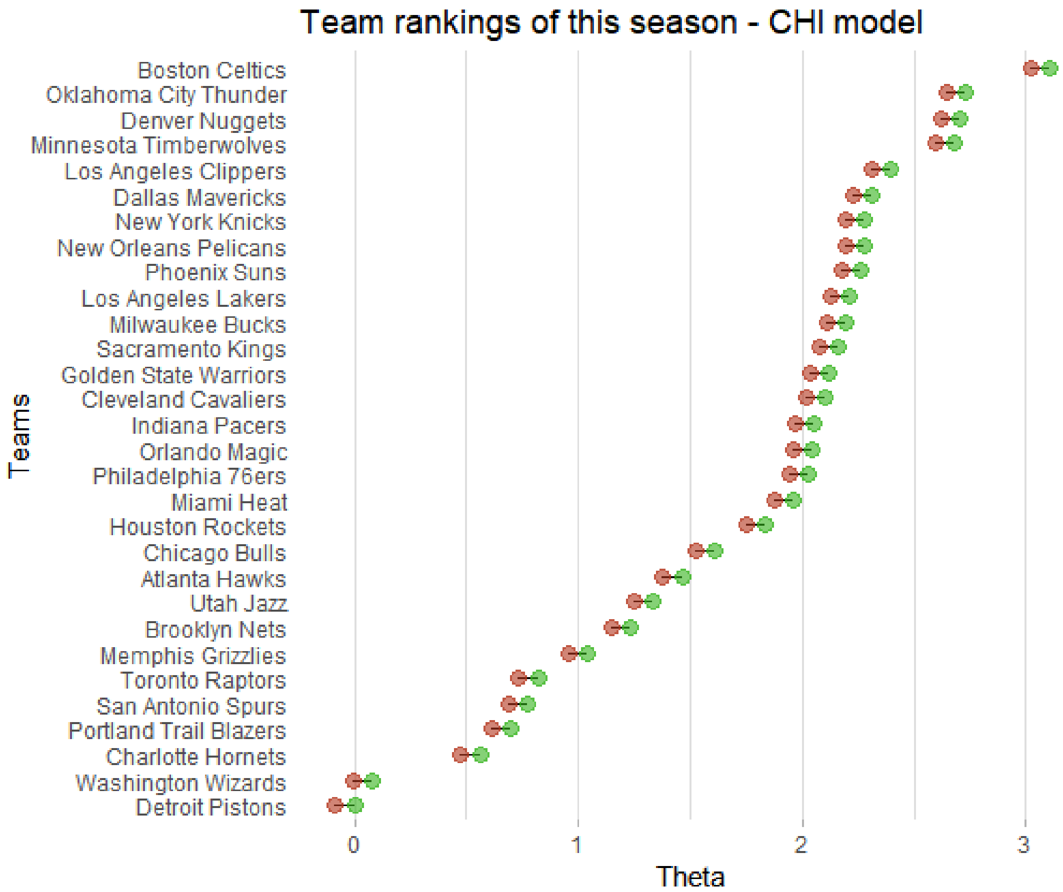

Figure 18: Common Hierarchical Home-ground Advantage model with Alpha 3 NBA 2023-2024

The table below represents the strengths of the teams in season NBA 2023-2024 with the teams arranged from 1 to 30. We can see the strengths of different teams by looking at the numbers.

Teams	Wins	Losses	PCT
Boston Celtics	64	18	0.7805
Denver Nuggets	57	25	0.6951
Oklahoma City Thunder	57	25	0.6951
Minnesota Timberwolves	56	26	0.6829
Los Angeles Clippers	51	31	0.6220
New York Knicks	50	32	0.6098
Dallas Mavericks	50	32	0.6098
Milwaukee Bucks	49	33	0.5976
New Orleans Pelicans	49	33	0.5976
Phoenix Suns	49	33	0.5976
Cleveland Cavaliers	48	34	0.5854
Los Angeles Lakers	48	35	0.5783
Orlando Magic	47	35	0.5732
Philadelphia 76ers	47	35	0.5732
Indiana Pacers	47	36	0.5663
Golden State Warriors	46	36	0.5610
Miami Heat	46	36	0.5610
Sacramento Kings	46	36	0.5610
Houston Rockets	41	41	0.5000
Chicago Bulls	39	43	0.4756
Atlanta Hawks	36	46	0.4390
Brooklyn Nets	32	50	0.3902
Utah Jazz	31	51	0.3780
Memphis Grizzlies	27	55	0.3293
Toronto Raptors	25	57	0.3049
San Antonio Spurs	22	60	0.2683
Charlotte Hornets	21	61	0.2561
Portland Trail Blazers	21	61	0.2561
Washington Wizards	15	67	0.1829
Detroit Pistons	14	68	0.1707

Table 2: NBA Teams' Wins, Losses, and Winning Percentage (PCT) in CHI model of NBA 2023-2024

5. Results and Analysis

5.1. One time simulation and analysis

Next, we will run a simulation, during which we can see how the teams compete with each other according to the existing seeds in the CHI model. The following running results show the simulation of the play-in tournament of the season, a match in the Eastern Conference playoffs, a simulation in the Western Conference playoffs, and a simulation in the finals.

The play-in tournament has a lot of uncertainty. We can see that in this simulation, the Golden State Warriors are ranked as the 8th seed in the West, but in reality, the Golden State Warriors were eliminated by the Sacramento Kings in the NBA 2023-2024 season play-in event and ranked 10th. Likewise, the Chicago Bulls made the playoffs in this simulation.

```
> Seeds
              Seeds                 Teams
1  7th Seed - East            Miami Heat
2  7th Seed - West  New Orleans Pelicans
3  8th Seed - East         Chicago Bulls
4  8th Seed - West Golden State Warriors
5  9th Seed - East    Philadelphia 76ers
6  9th Seed - West    Los Angeles Lakers
7 10th Seed - East         Atlanta Hawks
8 10th Seed - West      Sacramento Kings
```

Figure 19: One time simulation of play-in Tournament rankings with CHI model in NBA 2023-2024 season

When we get the seeds list after the play-in event, we have the top eight playoff teams in the Eastern and Western Conferences. Then we can simulate the Eastern and Western Conferences playoffs multiple times. We will then perform a general simulation to illustrate.

- **Team.A** represents the home team in this game, that is, the team with the highest total record in the entire regular season, and is related to the **Seed_ST** we manually entered according to the official website.
- Conversely, **Team.B** represents the away team.
- **Home_win** represents the number of games in which the home team won the series. When the number is greater than or equal to 4, it means that the

home team won the series. When the number is less than 4, it means that the home team lost the series. For example, in this simulation of the game between Boston and the Bulls, **Home_win = 7**, which means that Boston won the series in the 7th game.

- **Team.A_home_win_prob** represents the probability of the home team winning the game at home, and **Team.A_away_win_prob** represents the probability of the home team winning the game away.

These two probabilities are not fixed; they are calculated based on the two teams each time. I included them here to help users understand better, which is also my innovation this time.

```
> East()
               Team.A          Team.B Home_Win         Winner TeamA_home_win_prob TeamA_away_win_prob
1      Boston Celtics   Chicago Bulls        7  Boston Celtics           0.8491893           0.7797548
2     New York Knicks      Miami Heat        5 New York Knicks           0.6336237           0.5209317
3     Milwaukee Bucks  Indiana Pacers        5 Milwaukee Bucks           0.5139361           0.5559824
4 Cleveland Cavaliers   Orlando Magic        3   Orlando Magic           0.5730590           0.4576826
5      Boston Celtics   Orlando Magic        6  Boston Celtics           0.7854565           0.6971439
6     New York Knicks Milwaukee Bucks        4 New York Knicks           0.5778029           0.4625060
7      Boston Celtics New York Knicks        6  Boston Celtics           0.6780428           0.7137986
```

Figure 20: One time simulation of East playoffs with CHI model in NBA 2023-2024 season

In this Eastern Conference playoff simulation, we can see that the first-round winners are Boston Celtics, New York Knicks, Milwaukee Bucks, and Orlando Magic. The only series that `Team.B` won is Orlando Magic. It can be understood that among these four groups of games, Cleveland Cavaliers is the only one with `Team.A_away_win_prob` less than 0.5, which is 0.4576826. That is to say, Cleveland Cavaliers' win rate is lower than 0.5 when playing away, and there is a greater possibility to lose the series.

In the semifinals, the two teams with home court advantage, Boston Celtics and New York Knicks, won the final victory and entered the Eastern Conference Finals. In the end, Boston Celtics won the Eastern Conference championship in 6 games. We cannot ignore that it has a high winning probability both at home and away, 0.6780428 and 0.7137986.

```
> West()
                  Team.A                 Team.B Home_Win                 Winner TeamA_home_win_prob TeamA_away_win_prob
1  Oklahoma City Thunder  Golden State Warriors        5  Oklahoma City Thunder           0.7002202           0.5949166
2         Denver Nuggets   New Orleans Pelicans        5         Denver Nuggets           0.6602823           0.5499662
3 Minnesota Timberwolves           Phoenix Suns        4 Minnesota Timberwolves           0.6569026           0.5462432
4   Los Angeles Clippers       Dallas Mavericks        2       Dallas Mavericks           0.5788996           0.4636242
5  Oklahoma City Thunder       Dallas Mavericks        6  Oklahoma City Thunder           0.6576643           0.5470812
6         Denver Nuggets Minnesota Timberwolves        3 Minnesota Timberwolves           0.4866907           0.5289336
7  Oklahoma City Thunder Minnesota Timberwolves        4  Oklahoma City Thunder           0.4927304           0.5349511
```

Figure 21: One time simulation of West playoffs with CHI model in NBA 2023-2024 season

In the simulation of the Western Conference playoffs, we can see that the winners of the first round are Oklahoma City Thunder, Denver Nuggets, Minnesota Timberwolves and Dallas Mavericks. According to the team strength rankings generated by our regular season, Oklahoma City Thunder, Denver Nuggets and Minnesota Timberwolves are ranked 2nd, 3rd, and 4th respectively. This is very reasonable, indicating that the model ranking is of reference value. Similarly, the 5th and 6th places are Los Angeles Clippers and Dallas Mavericks, but they are opponents in the first round, and their competitive abilities are very close.

In the end, Oklahoma City Thunder, the strongest team in the West according to the CHI model, won the Western Conference championship.

```
> Finalgame2()
          Team.A                Team.B Home_Win         Winner TeamA_home_win_prob TeamA_away_win_prob
1 Boston Celtics Oklahoma City Thunder        5 Boston Celtics           0.6613124           0.5197017
```

Figure 22: One time simulation of Finals with CHI model in NBA 2023-2024 season

In the finals, Boston Celtics and Oklahoma City Thunder were the two teams, and Boston won the championship in 6 games. It can be seen that in this series, the values of `Team.A_home_win_prob` and `Team.A_away_win_prob` are both over 0.5, which are 0.6613124 and 0.5197017.

5.2. Alignment with Actual Results

1. In the Eastern Conference Series, all simulations are more in line with the real situation. In particular, Boston won the Eastern Conference Championship all the way. In the NBA 2023-2024 season playoffs, Boston showed a super high level of competitive performance and was unstoppable in the East. It is also worth noting that it is very reasonable for the New York Knicks to enter the Eastern Conference Finals in this simulation. They performed well in the real playoffs and played an

eye-catching defensive performance.

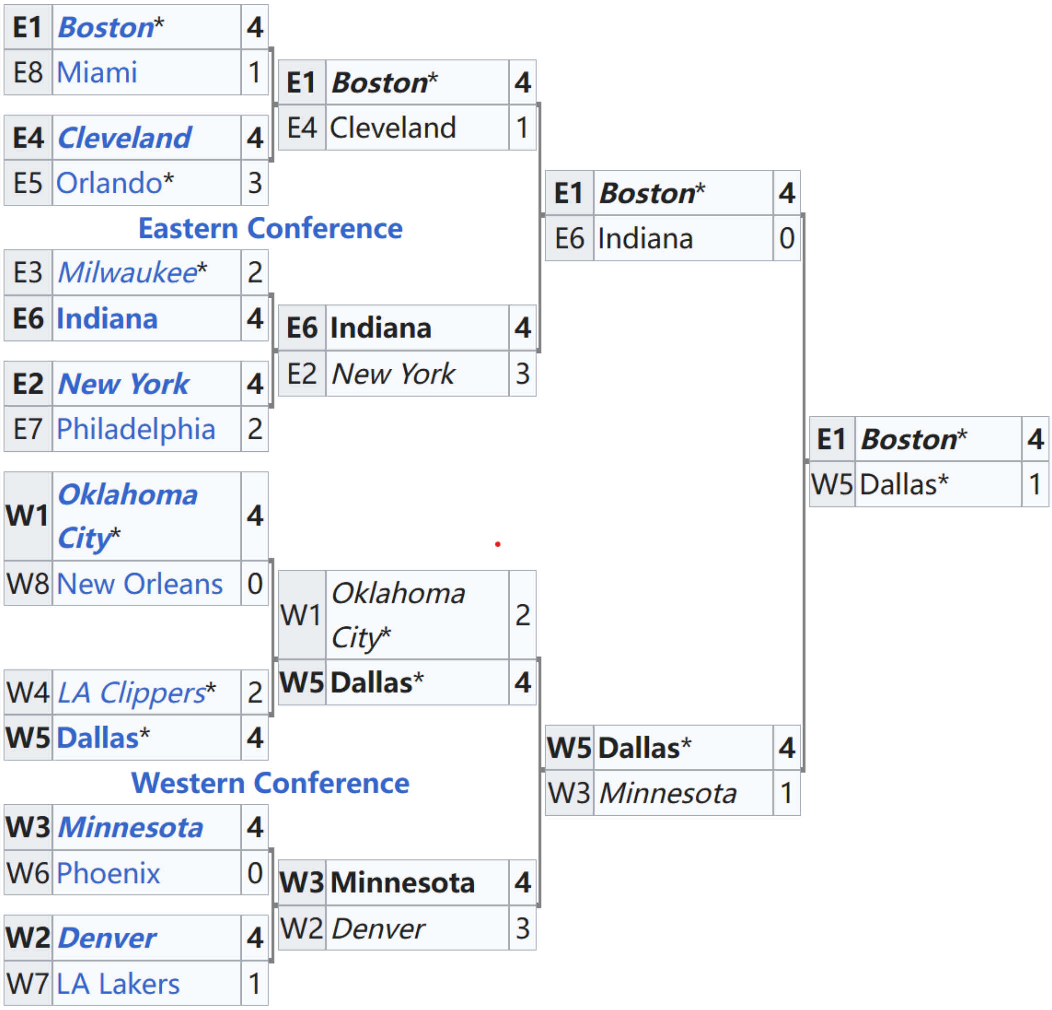

Figure 23: NBA 2023-2024 Season playoffs result

2. In this Western Conference Series, many realistic simulation results were also shown. In the first round of the Western Conference playoffs, the match between Los Angeles Clippers and Dallas Mavericks was a very popular match in reality, especially when the No. 4 seed Los Angeles Clippers had the home advantage, but was eventually overturned by Dallas. In the simulation results, Dallas Mavericks won the Clippers 4:2, which is exactly the same as the real situation, indicating that such simulation is of great reference value.

Similarly, although the Denver Nuggets performed well in most data samples, they also lost to the Minnesota Timberwolves in this simulation with a score of 3:4. Shockingly, in the real playoff situation, the Minnesota Timberwolves did win the game in the much-anticipated "Game 7" with a younger and more energetic competitive state. The score of the simulation result is exactly the same as the real situation, which once again shows the value of the model.

3. In this Final Game simulation, the two teams in the finals are Boston Celtics and Oklahoma City Thunder, which matches our model very well because they are the top two teams in the strength ranking. In the end, Boston Celtics won the championship in 6 games, and the game was very tense.

```
> subset(formaldatanew,
+                   (Team.A == "Boston Celtics" & Team.B == "Oklahoma City Thunder") |
+                   (Team.A == "Oklahoma City Thunder" & Team.B == "Boston Celtics"))
                    Team.A                Team.B Team.A.win Team.B.win Team.A.home Team.B.home
493         Boston Celtics Oklahoma City Thunder          0          1           0           1
1135 Oklahoma City Thunder        Boston Celtics          0          1           0           1
```

Figure 24: The actual records of Boston and Oklahoma in regular season 2023-2024

Although Oklahoma City Thunder did not make it to the finals in reality, according to their record of confrontation in the NBA 2023-2024 season, the two teams played against each other twice in the regular season and won one game each, indicating that the competitive intensity of the two teams is not very different. Therefore, it is easy to understand and logical that they played 6 games in the finals in the simulation.

5.3. Multiple times simulations and Insight

5.3.1. Result

After building the model and applying it to the data, we found that the complete process was successful and reasonable. Therefore, we have a longer-term plan. We want to have sufficient situations to simulate the entire playoffs between the end of the regular season and the day when the champion is born, so as to support the accuracy of the prediction with a large number of experiments, so as to have a greater confidence in predicting who the final champion will be.

Based on the regular season data, we hope to simulate the playoffs of this season, starting from the play-in game to the NBA Finals. Therefore, we save the results of each simulation, repeat a large number of simulations, and finally count the number of times each team that is eligible for the playin event and the playoff team has won the championship. This number can be used as the probability of each team winning the championship. Such a probability result is more effective and convincing than the simple guesses given by some sports media.

Therefore, we conducted 10,000 simulation experiments based on the Common Hierarchical Home-ground Advantage Model and visualized the results as shown in the figure.

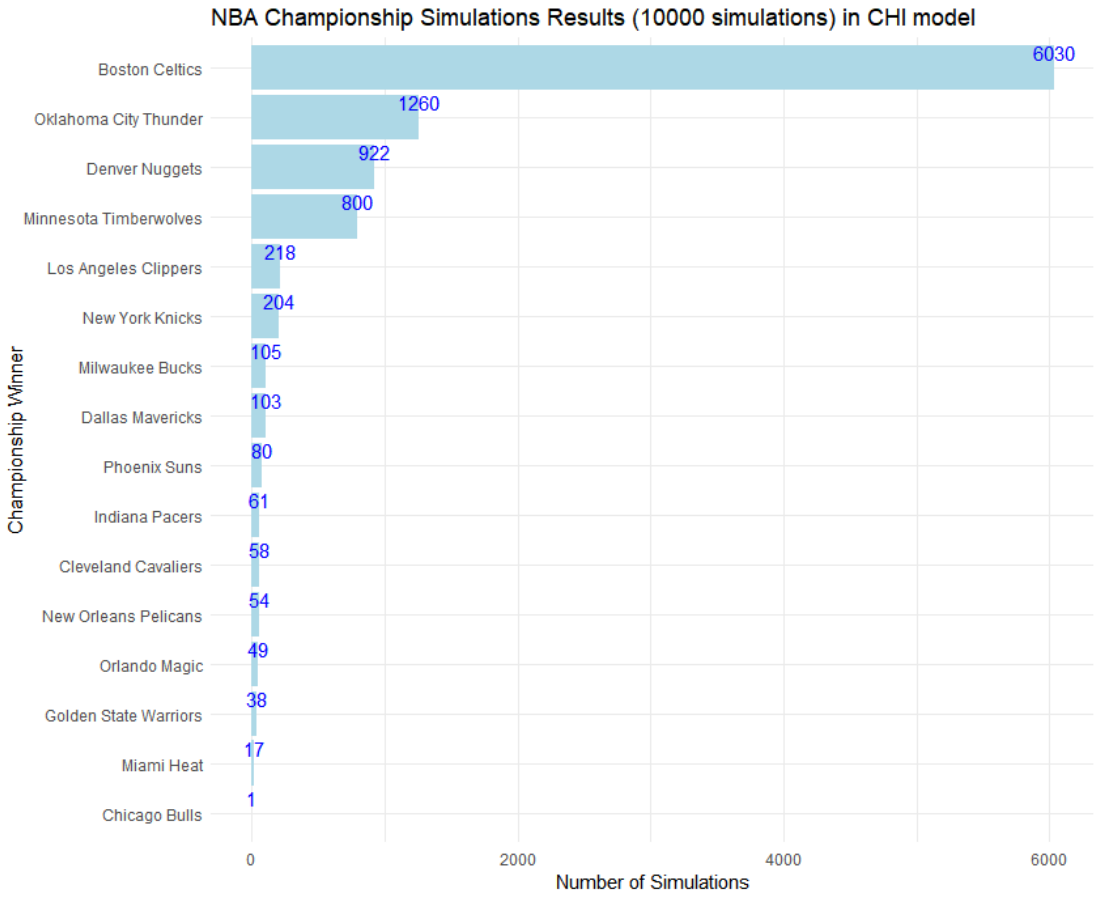

Figure 25: Championship Winners -10000. time Simulation with CHI model

5.3.2. Insight

The final day of the NBA 2023-24 regular season ended on April 14. On April 21, 2024, at 12:00 AM, a sports reporter from the famous American sports company ESPN published an article titled “2024 NBA playoffs betting: Title, conference, Finals MVP picks”, mentioning ESPN’s predictions for various playoff awards after the regular season, including the top teams favored to win the championship. The timing of these predictions aligns with ours, so we conducted a comparative analysis.

NBA Champion

Leaders: Celtics (+145)
In the hunt: Nuggets (+275)
Longer shots of interest: Thunder (+1400), Mavericks (+1500), Bucks (+1600)

Figure 26: ESPN predictions in 2024

Website source:[2]

It can be seen that the top five teams, ranked from highest to lowest probability, are Celtics (+145), Nuggets (+275), Thunder (+1400), Mavericks (+1500), and Bucks (+1600). Our results from 10,000 simulations, as shown in the chart above, have a similar top three, indicating significant reference value.

Although ESPN predicted that the Mavericks would have the fourth-highest probability of winning the championship, in our prediction, this team only ranks around 8th.

On the other hand, it is worth mentioning that in our predictions, the Minnesota Timberwolves achieved excellent results, winning the fourth-highest number of championships in extensive simulations. However, ESPN did not mention this, indicating that this sports media outlet does not favor the Minnesota Timberwolves. However, in reality, the Minnesota Timberwolves had an impressive performance in the 2023-24 playoffs, defeating the Denver Nuggets to advance to the Western Conference Finals, making them one of the top four teams of the season. This aligns perfectly with our prediction, suggesting that we had an advantage in predicting this team.

I believe the reason for this phenomenon is that our model is based on regular-season performance and home-court advantage. Therefore, the Minnesota Timberwolves performed well due to their strong regular-season record and substantial home-court advantage. In contrast, mainstream media companies often consider more realistic factors, such as team stars, team chemistry, or playoff experience—these are parameters that cannot be quantified. For example, the Mavericks have one of the best players of recent years, Luka Dončić, who has repeatedly proven his abilities on the playoff stage. Compared to the Timberwolves' team star, Anthony DeVante Edwards, who lacks playoff experience, it's understandable that the media would be more optimistic about the Mavericks' chances of winning more games than the Timberwolves. Therefore, both mainstream media

[2] https://www.espn.com.au/espn/betting/story/$/_{/}id/$39978961

predictions and this model's predictions hold value; they just reference different aspects.

6. Conclusion and Potential Improvements

6.1. Project Improvement Summary

This project involved substantial work and analysis. The lack of a detailed introduction and purpose analysis from the previous author initially complicated the understanding of the code's logic. After communicating with the original author, clarity was gained regarding some of the hard-coded logic naming.

The main achievements of this modification project include the following aspects in the code:

1. Corrections were made to several pieces of code that could not run directly. The original author successfully ran the project in his own environment, but when new researchers attempted to follow up and improve it in a new environment, they encountered numerous incompatible errors. Modifications were also made to the plotting code, which previously could not generate plots for certain models. After the improvements, successful plot generation was achieved. Additionally, visualization convenience was enhanced, particularly for the Team-specific Home-ground Advantage Model and Hierarchical Home-ground Advantage Model, which originally followed the Theta rankings instead of balanced midpoint values. The modified model plots now present a smoother appearance.

2. Unreasonable aspects in the data were addressed. The original author considered only the top 20 teams from the NBA 2022-2023 season, resulting in errors during the play-in event simulation phase. Corrections were implemented, and this was noted in the paper. Future improvements could expand the list to include the complete 30 or even 32 teams for each season.

3. An interactive page was designed. The entire model operation process was standardized, and obscure code sections were removed. Furthermore, a user interaction page was created, allowing new users to select models and simulation parts. This development offers significant flexibility for subsequent researchers and potential users to make their own choices.

The original code was limited to supporting a single model for one-time simulations. Post-modification, users can now select any model at any time and simulate any part of the playoffs, significantly enhancing flexibility. Additionally, the original model defaulted to supporting only 10,000 simulations; the improved interaction now allows users to simulate any desired number of times.

6.2. Potential Improvements

NBA's official sources have indicated plans for expansion in the 2026-2027 season, possibly adding teams in Las Vegas and Seattle, bringing the total number of teams in the league to 32. To ensure balanced teams between the Eastern and Western conferences, some teams may be reassigned to different divisions. Consequently, data collection and background information matrices cannot remain static; they should be updated in real-time to maintain long-term practicality.

6.2.1. Automatic Data Scraping.

Currently, all data sources require manual acquisition and preprocessing into a specific format. Utilizing API interfaces to obtain real-time data can ensure timely updates. For instance, when simulating play-in events and playoff events, researchers and users currently need to manually look up the rankings for the current season and input them into a table for execution. Future researchers intending to improve this process could implement code to scrape data directly from official websites, thereby simplifying operations for other users.

Moreover, the list of rankings for each season presents a cumbersome challenge when making predictions for a new season. Users must manually input the list and combine the rankings, which is time-consuming. However, this task is one of the simplest data scraping operations that can be implemented through code.

6.2.2. Incorporating More Elements.

This experiment did not advance the Pairwise Home-ground Advantage Model due to the extensive computations and considerations involved. The PAI model has vast potential for further exploration, as it encompasses the crucial fan culture in NBA events. A significant portion of a team's home advantage is derived from its fan base; the attitudes of fans in different cities often influence the home advantage in various matchups. The PAI model would assign a unique home-ground advantage value for each team pair.

For example, the fan culture surrounding the Boston Celtics and the Knicks is very strong, with their mutual desire to win against each other being particularly intense, resulting in a pronounced home-ground advantage. Conversely, the rivalry between Boston and the Suns is relatively mild, leading to a less significant home-ground advantage. This would necessitate more detailed data and greater computational capacity.

6.2.3. Application Expansion.

This model possesses significant potential. First, an online real-time prediction platform based on R code could be developed, allowing fans and businesses to log in at any time during the season and select models to predict the season champion. Further exploration could enhance visualization convenience for a broader user base. Additionally, once the model website matures, collaborating with sports media or betting companies to apply the model in real-world business scenarios could prove beneficial.

A. Appendix

Related Code as below link:
https://github.com/jessezhang2024/NBA-Simulation

B. Appendix

In order to facilitate comparison and readability, the plots for the remaining four models are presented here for reference.

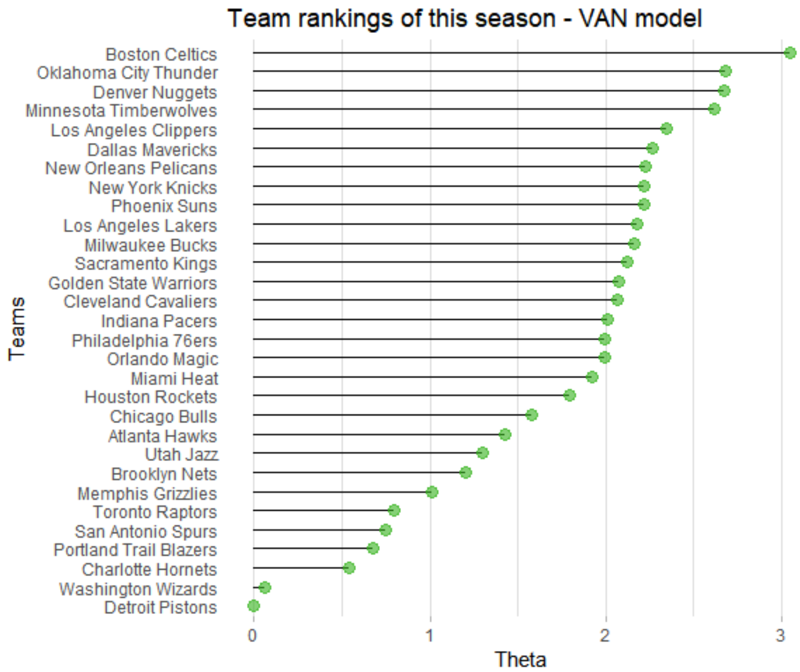

Figure 27: Team Rankings of the season - Vanilla Bradley-Terry Model

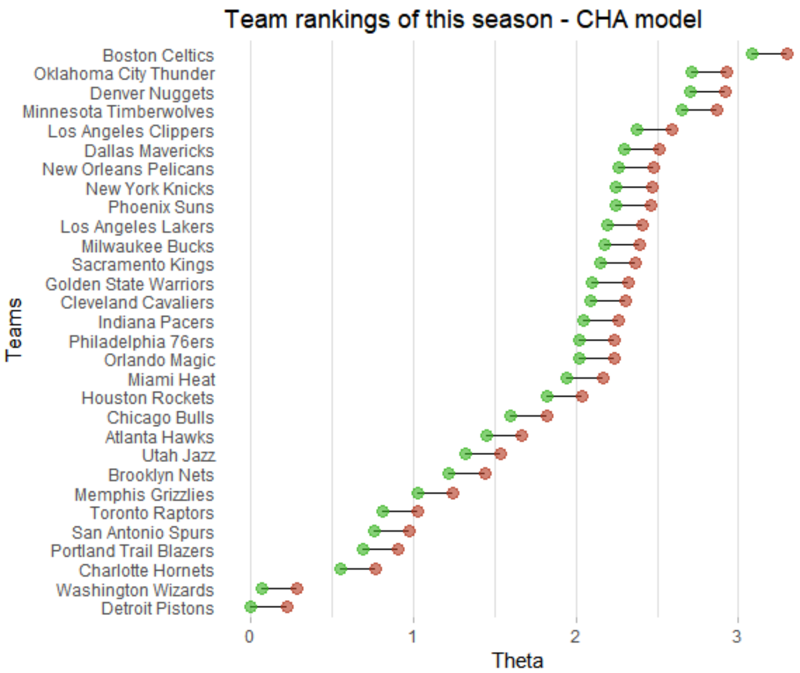

Figure 28: Team Rankings of the season - Common Home-ground Advantage Model

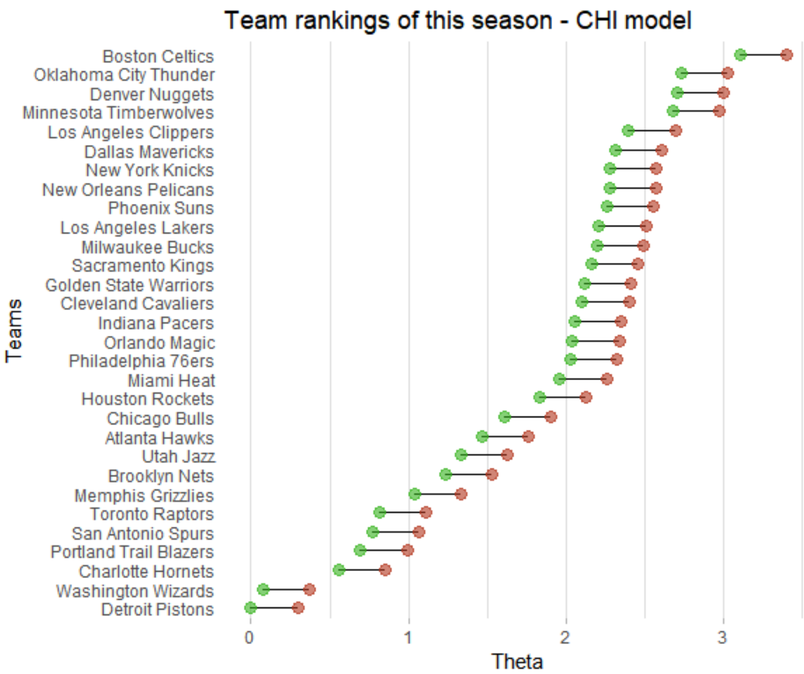

Figure 29: Team Rankings of the season - Common Hierarchical Home-ground Advantage Model with Alpha 1

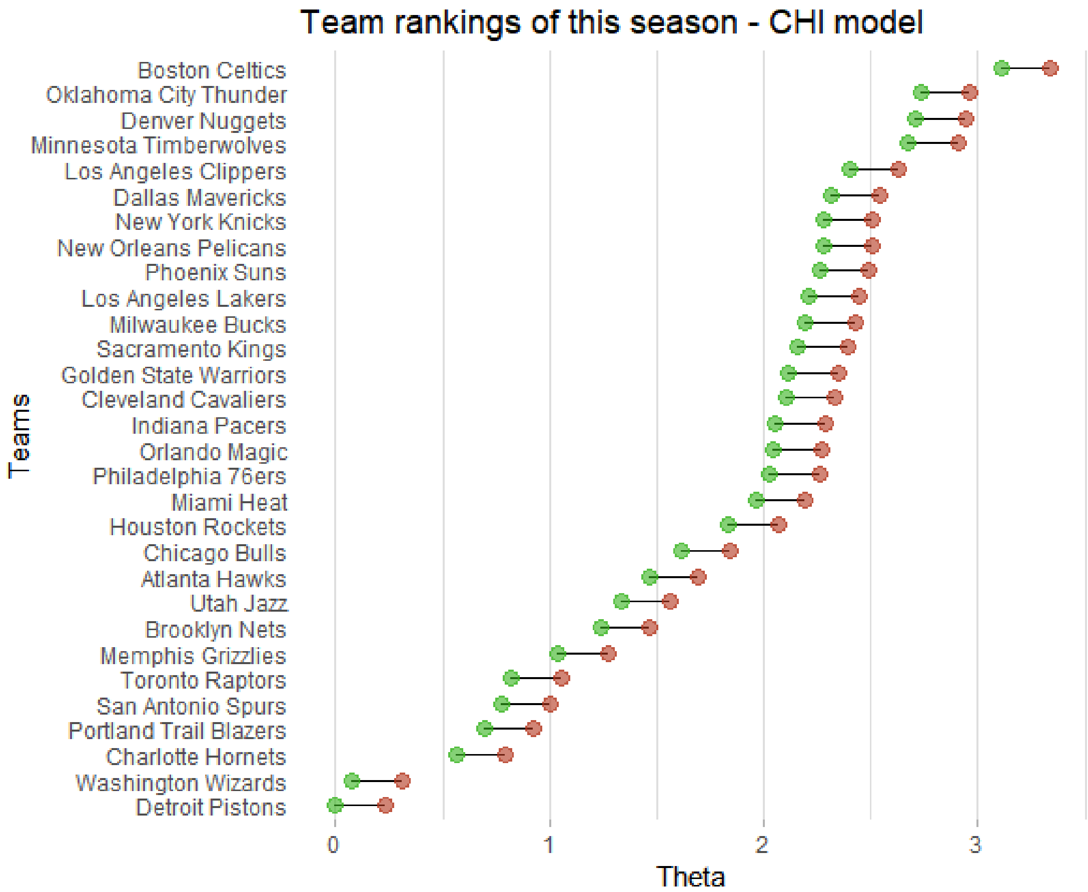

Figure 30: Team Rankings of the season - Common Hierarchical Home-ground Advantage Model with Alpha 2

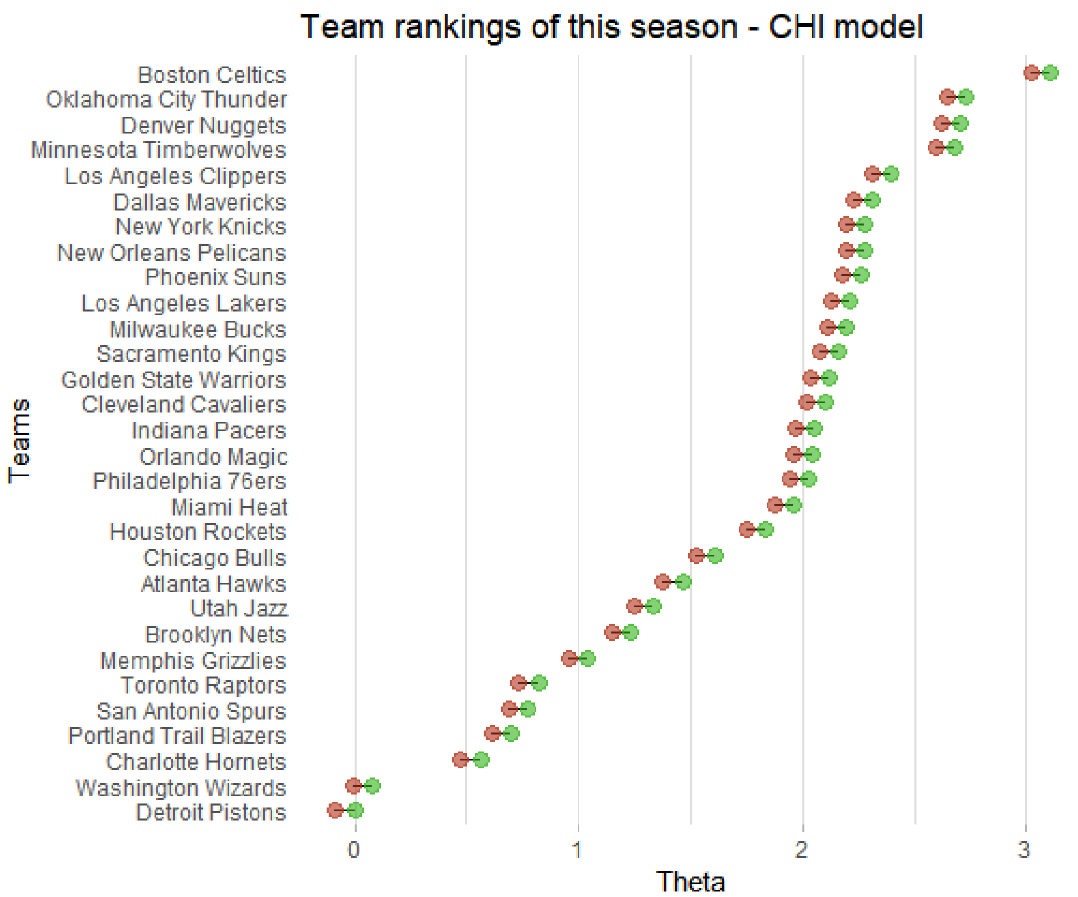

Figure 31: Team Rankings of the season - Common Hierarchical Home-ground Advantage Model with Alpha 3

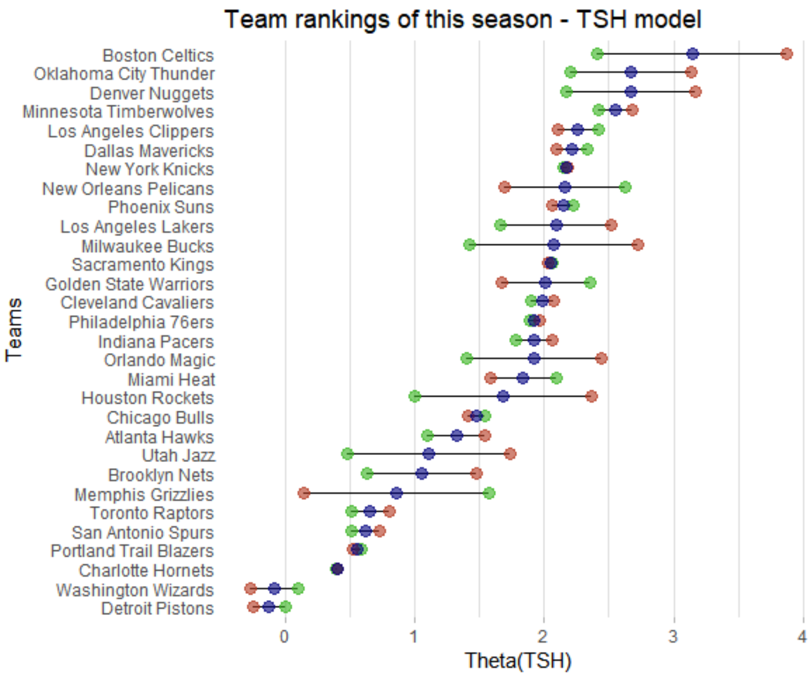

Figure 32: Team Rankings of the season -Team-specific Home-ground Advantage Model

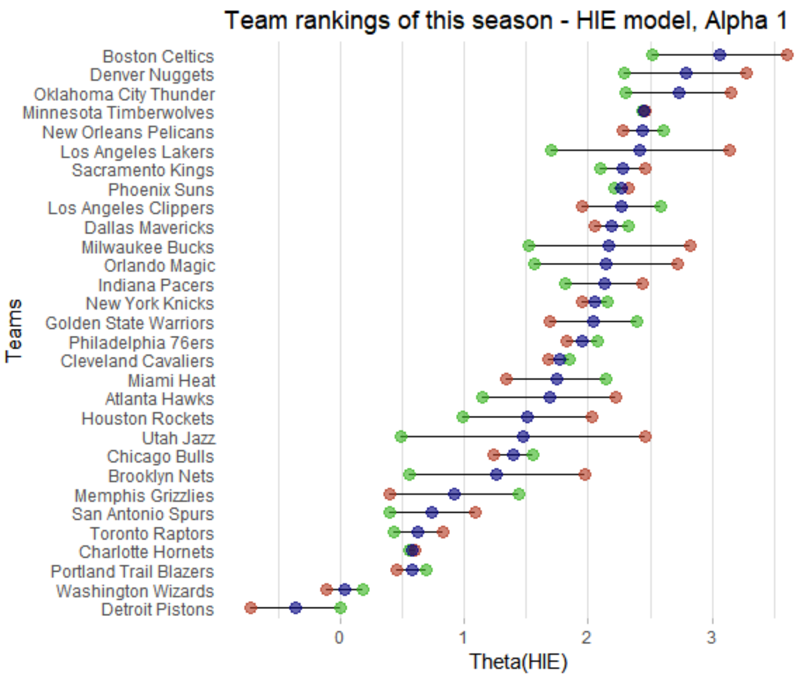

Figure 33: Team Rankings of the season -Hierarchical Home-ground Advantage Model

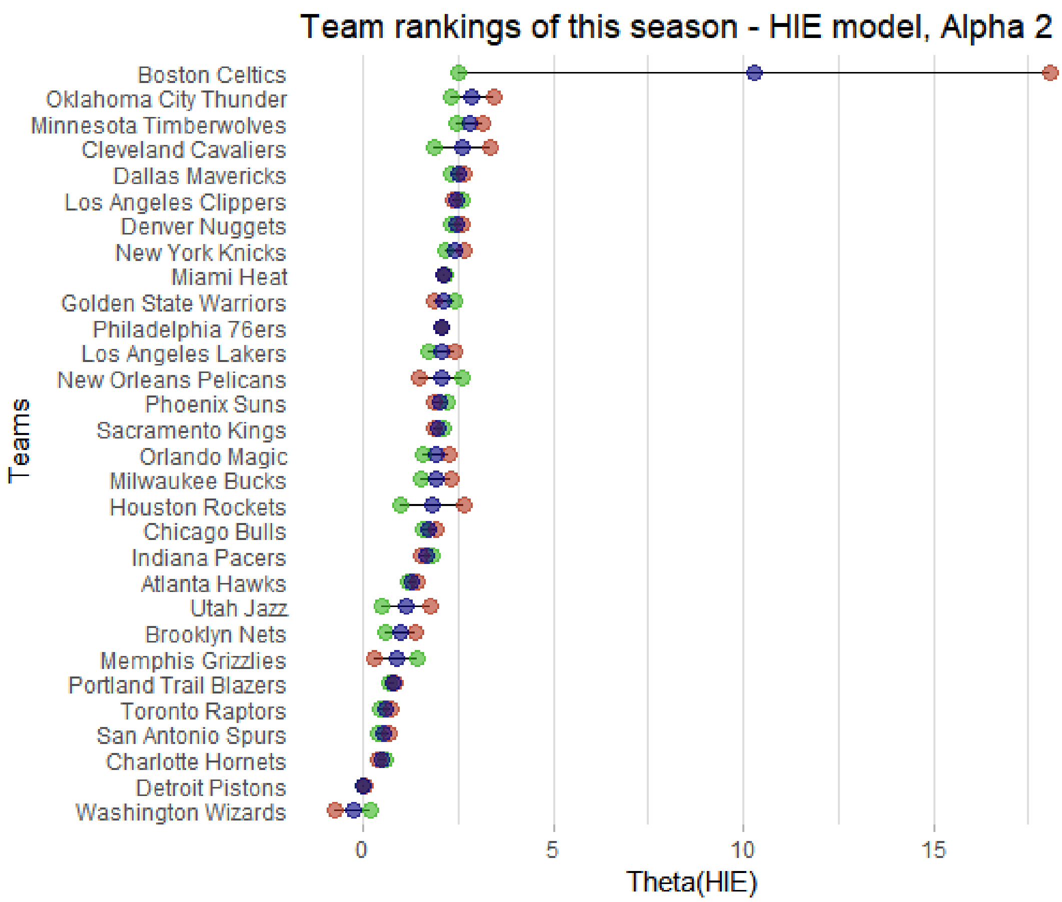

Figure 34: Team Rankings of the season -Hierarchical Home-ground Advantage Model

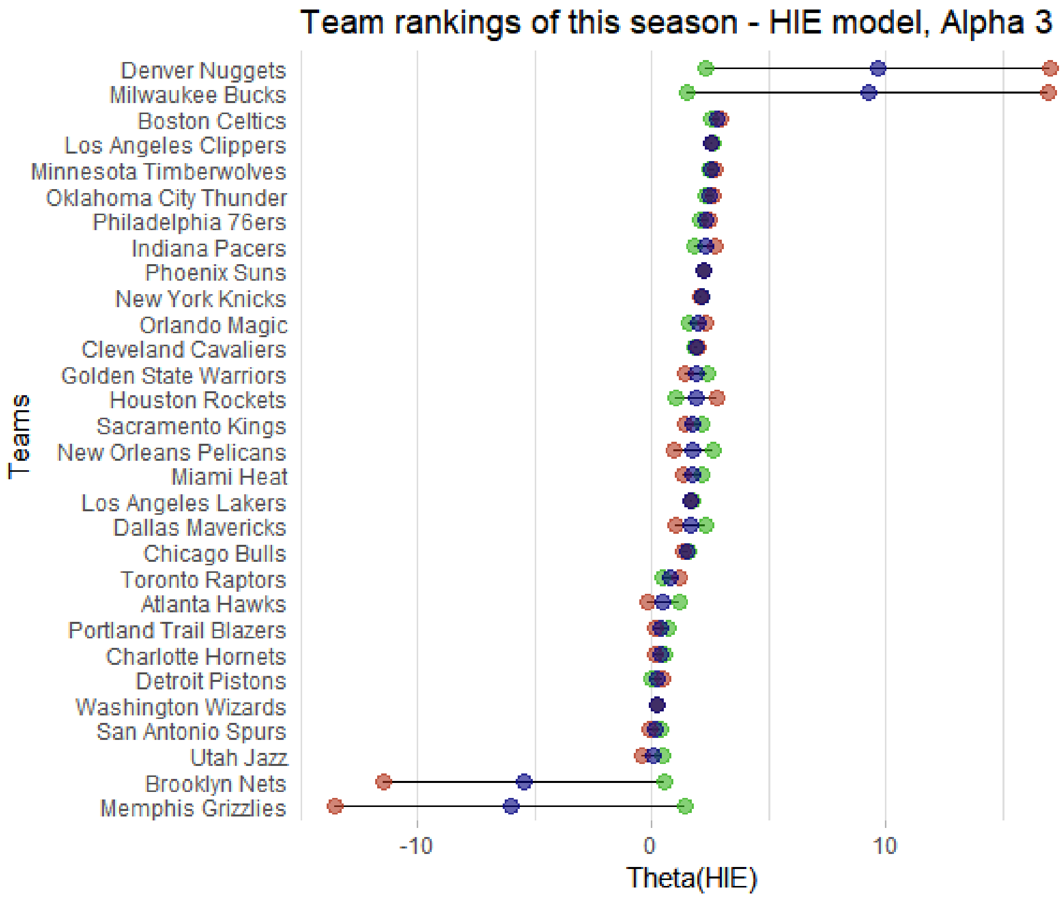

Figure 35: Team Rankings of the season -Hierarchical Home-ground Advantage Model

References

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

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

[3] Atkinson, A. C. "A test of the linear logistic and bradley-terry models." *Biometrika*, vol. 59, no. 1, 1972, pp. 37-42.

[4] ESPN. *NBA Standings - 2023 Season*. Retrieved from https://www.espn.com/nba/standings/_/season/2023/group/league.

[5] Snellings, Andre. "2024 NBA playoffs betting: Title, conference, Finals MVP picks." *ESPN*, Apr 21, 2024.

[6] Jones, Coen. *Hierarchical Bradley-Terry Models*. Honours thesis, University of Queensland, 2021.

[7] Sporting News. *NBA Finals and Playoff Odds 2024: Power Rankings and Picks*. Retrieved from https://www.sportingnews.com/au/nba/news/nba-finals-playoff-odds-2024-power-ranking-picks.

[8] Piva, Sloan. "NBA Finals odds 2024: Updated championship futures, best bets and top values." *Sporting News*, May 17, 2024.

[9] Wikipedia. *2024 NBA Playoffs*. Retrieved from https://en.wikipedia.org/wiki/2024_NBA_playoffs.

[10] Hsieh, Hsin-Ying, Chieh-Yu Chen, Yu-Shuen Wang, and Jung-Hong Chuang. "BasketballGAN: Generating Basketball Play Simulation Through Sketching." *Proceedings of the 27th ACM International Conference*, October 2019.

[11] Tabatabaie, Cameron. "We simulated the upcoming Celtics season on NBA 2K24; Here's what we learned." *Celtics Blog*, September 18, 2023.