Quantitative Essays on Mixed Martial Arts

A Markov Chain Based Forecasting Model and Analyses of the Judges

By

Benjamin Paul Holmes

Department of Mathematics
University of Liverpool

A dissertation submitted to the University of Liverpool in accordance with the requirements of the degree of Doctor of Philosophy.

May 2022

THE UNIVERSITY
of LIVERPOOL

Abstract

Whilst Mixed Martial Arts (MMA) has only recently gained mainstream popularity, the rapid rise of it, particularly the Ultimate Fighting Championship (UFC–the most popular MMA promotion), has been unparalleled in sports. Academic research on MMA is still scarce, and the vast majority has focused on the sport's health implications. This thesis comprises three articles which contribute to the knowledge on MMA, as well as the wider literature regarding sports forecasting and biases.

The first article, now published in the International Journal of Forecasting (Holmes et al., 2022), introduces a Markov chain (MC) based model to predict MMA bouts. The states of the MC are associated with key techniques or positions within MMA. Various models based on the athletes' historical in-fight statistics determine the transition probabilities between states, thus accounting for individual fighting styles. By simulating the chain many times, we obtain probabilities of fight outcomes. These predictions were comparable to the bookmakers, and generated positive returns when used for betting.

Compared to other subjectively judges sports, for instance, diving, the performance data available for UFC fights provides an ideal environment to model the judges' behaviours. Thus, the remaining two papers examined the judges in the UFC. First, we explored the potential of several biases within MMA judging. We find evidence suggesting two biases exist: the judges are influenced by a live audience, thus favouring a home athlete; and the judges favour athletes higher in the official rankings. One issue with previous work was establishing whether the significant effects were due to bias or fighter skill. Under the hypothesis that the betting market is efficient, we address this issue by including the bookmakers' odds to account for unseen skills. Market efficiency suggests the bias variables don't add any information on skills beyond what is contained in the odds, and thus significance is indicative of bias, not skill. We demonstrate that the market is efficient, and thus we can be more confident in our conclusions.

Second, we use a Bayesian hierarchical model to show that the judges have different preferences towards each action. We identify several actions where judges have a wide range of opinions, even to the extent of actions being valued in opposite directions. Using this model, we demonstrate how the judges' preferences may themselves determine the winner of a fight, and also develop a "fair"-scoring model that could be used by promotions or athletic commissions for a number of purposes. We apply the concept of variable significance to determine whether a judge's verdict was mathematically controversial or within reason. Further, we estimate a similar model using scores submitted by fans. This model suggests that fans are more likely to give rarer scores, such as draws. Interestingly, it appears fans are less influenced by bias variables than the judges.

Acknowledgements

First and foremost, I'd like to thank my supervisors, Kamila and Ian, for their guidance, help, knowledge, and friendship throughout the PhD. I will always be grateful for initially backing me and the project, and your enthusiasm with every new idea. Dom, who pushed me to pursue my passion, I can't thank you enough. I've collaborated with some brilliant people in industry during this time, Alex, Andy, and Phill, all deserve special thanks.

None of this would be possible without my friends and family. In particular, Darah, Elliott, Matthew, Tom, Ronan, and Will, thanks for all the UFC related fun. Norma, Ed, and Iz, thank you for being so caring and loving over the years. Beth and Jon, thank you for always being such amazing siblings, friends, and inspirations. Mum and Dad, you're such perfect role-models, and I would never have been in this position without your love and support. Empi, you've been such a wonderful partner the last few years and made what should have been a tough time so enjoyable. Finally, who could forget Coco and Mio.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Mixed Martial Arts (MMA) is a sport that has only recently captivated the mainstream. Yet its rate of expansion, particularly over the last decade thanks to super-stars such as Ronda Rousey and Conor McGregor, has been unprecedented in sports. This thesis is devoted to progressing the academic literature available on MMA, of which there is currently very little available outside the health sciences.

This chapter will serve to introduce MMA. The history and origins will be detailed in Section 1.1. How an MMA contest unfolds and the role of the judges will be described in Section 1.2. Finally, a summary of the remainder of the thesis will be given in Section 1.3.

1.1 History of MMA

As one would expect from the name, MMA is a full-contact combat sport, which "mixes" techniques from all other martial arts. Whilst not an exhaustive list, athletes can use punches from boxing, kicks from taekwondo, knees and elbows from muay Thai, throws from judo and wrestling, or submission moves from Brazilian jiu-jitsu (BJJ).

The exact origins of modern day MMA are hard to pinpoint. Traditional full-contact, unarmed, "mixed" martial arts have long existed, such as vale tudo (Brazil) and combat sambo (Soviet Union). Pankration is a similar sport that was introduced to the Olympics in 648bc[1]. Bouts between practitioners of different martial arts occurred throughout the 20th-century in East Asia. Even Muhammad Ali, widely regarded as one of the greatest boxers to have lived, competed in a hybrid match against Japanese professional wrestler Antonio Inoki in 1976.

For most though, the catalyst to modern-day MMA would be the inaugural event of

[1] https://www.britannica.com/sports/pankration

the Ultimate Fighting Championship (UFC–the world's premier MMA promotion) in 1993, Colorado, USA.

Originally, UFC events followed a knockout tournament structure that would take place over one day. The athletes who took part in these early tournaments were practitioners of a single martial art, with the aim being finding the best martial art. The eight martial arts featured in the first tournament, each represented by one athlete, were: savate, sumo, kickboxing, American kenpo, BJJ, boxing, shootfighting, and taekwondo. The tournament was won by BJJ athlete, Royce Gracie, who would win three of the four knockout tournaments he competed in.

The rules–or the lack of–were also very different at first. There were no doping checks, no mandatory padded gloves to wear, and no weight classes. No judges scored the bouts, meaning fights would continue indefinitely until a knockout, submission[2], or corner stoppage. Further, only three fouls existed, each enforced by a $1,500 fine: biting, eye-gouging, and groin strikes.

In the following years, the UFC became particularly infamous due to Senator John McCain. In 1996, he labelled it as "human cockfighting" (Smith, 2009) and sent letters to all US governors asking them to ban it. This, in part, led to additional rules in an effort to make MMA more of a sport than a spectacle[3]. Some key changes include: introducing a 30-minute time-limit (1995); allowing the judges to determine the outcome of fights which reached the time-limit (1996); the introduction of two weight classes, determined by whether an athlete was above or below 200 lbs (1997); mandatory padded gloves (1997); banning of hair-pulling, kicks to a downed opponent, and headbutts (1997); and finally, introducing five-minute rounds (1999).

It wasn't until 2000 when the Unified Rules of Mixed Martial Arts were established by the New Jersey State Athletic Control Board. These rules became the go-to regulations that athletic commissions would abide by and can be seen as the beginning of modern-day MMA. The rules ensured heterogeneity across promotions on a number of issues: weight classes and the structure of rounds; as well as introducing judging criteria, a comprehensive list of fouls, and a set of prohibited substances.

As the years progressed it became clear what styles worked. A good example is BJJ. Originally, this was a niche martial art developed and used by the Gracie family[4]. As the

[2] A submission is a grappling technique, such as a choke or joint-lock, so that when successfully applied, the opponent must concede defeat to avoid further serious damage. This is signalled to the referee by 'tapping-out', or if unable to tap by verbally conceding.

[3] In fact, McCain forcing legislation in this way is often seen as a key driver in the rise of MMA. McCain would later accept MMA as a sport and is quoted as saying he would have likely took part if it was around during his youth.

[4] https://www.gracieuniversity.com/Pages/Public/About

Gracie family continued to dominate in MMA competitions, more and more athletes began to train BJJ, to the point where it is now a staple of MMA. These days, the majority of successful athletes will be proficient in all areas of MMA, whilst still having their own style and skills. In fact, one can now train MMA as a sport itself rather than training specific martial arts.

Over the last two decades, the UFC went from strength to strength, garnering millions of pay-per-view buys per event, merging with (or even purchasing) competing MMA promotions, and partnering with major sports networks. Two recent milestones highlight the current popularity and potential of the UFC. First, the 2016 sale to WME-IMG (recently rebranded to Endeavor) for $4 billion, which at the time was the largest acquisition in the history of sports. Second, the $1.5 billion deal with ESPN to televise 42 events per-year for five years from 2018. For comparison, the largest television rights contract for football is for the English Premier League, which agreed to a deal from 2019 to 2022 to show 200 games per season for an estimated £5 billion.

The preceding discussion has been focused on highlighting the history and current popularity of MMA. Next, we discuss how an MMA bout plays out, and describe the rules by which MMA contests are run.

1.2 An MMA contest

Since the research will be using data exclusively on UFC fights, we will discuss fights as they occur in the UFC as there may be slight differences across promotions.

As in the majority of combat sports, athletes are split by weight classes. There are currently eight men's weight classes in the UFC (the upper-limit is given in brackets): Heavyweight (265lb), Light Heavyweight (205lbs), Middleweight (185lbs), Welterweight (170lbs), Lightweight (155lbs), Featherweight (145lbs), Bantamweight (135lbs), Flyweight (125lbs). The first women's fight in the UFC took place in February 2013 and now there are four women's weight classes: Featherweight (145lbs), Bantamweight (135lbs), Flyweight (125lbs), and Strawweight (115lbs).

Fights are generally scheduled for three five-minute rounds with a minute break in between each round. Title fights and the majority of main events are scheduled for an extra two rounds. Compared to boxing, which at the highest level typically has 12 three-minute rounds, the round structure is quite different.

Each round starts with both athletes standing and can transition to the ground through various grappling moves that are generally referred to as “takedowns”. Fighters may engage in a “clinch”, that is grappling whilst standing. Strikes are permitted at all times as long as the technique adheres to the rules. Submission attempts, such as chokes and joint locks, can be attempted at any time, but will mostly come whilst on the ground in an advantageous grappling position.

The contest can also transition to the ground through a “knockdown”, which is a strike causing the opponent to fall to the ground. Unlike in boxing, where the opponent will then have 10 seconds to get back to their feet to continue the fight, the fighter can follow their opponent to the ground and attempt more techniques whilst the opponent is still vulnerable. Consequently, knockdowns often lead to the end of the fight.

A fight may end prematurely through a knockout (KO) or technical knockout (TKO). A knockout win occurs when the opponent is physically unable to continue due to loss of consciousness; typically from a powerful strike to the head. A TKO occurs when the referee intervenes to stop the contest, deeming the opponent unable to continue. There is a lot of overlap between what is a KO and TKO, so generally the official outcome is given as KO/TKO.

A successful submission attempt will also end the fight prematurely. Submission moves can cause serious damage: joint and limb locks cause hyper-extension which can lead to fractures and tears, whilst chokes lead to a loss of consciousness. Consequently, when applied successfully the opponent is forced to concede the fight to avoid the unnecessary damage. This is done by physically tapping or verbally signalling to the referee, who then breaks the fighters up.

Intentional fouls that render the opponent unable to continue will lead to a disqualification loss. Examples include a serious eye-poke, strike to the groin, or kneeing a grounded opponent in the head. If the referee believes the foul was unintentional/accidental, then the fight will be declared a no-contest–this essentially voids the bout. A no-contest can also be given retrospectively if an athlete fails a post-fight doping test.

Around 50% of fights will end prematurely. Failing any of these outcomes, the fight will be scored by the judges.

1.2.1 Judging in MMA

Three judges score each round, with the scoring criteria set out in the Unified Rules of MMA (California State Athletic Commission, 2020). There are three criteria that a judge must assess sequentially, if and only if they deem the preceding criteria to be exactly even.

1. *Effective striking and grappling* refers to “legal blows that have immediate or cumulative impact with the potential to contribute towards the end of the match”. The immediate is stated to weigh more, so actions such as knockdowns and submission attempts, which

often lead directly to the end of the fight, are thus weighted highly.

2. *Effective aggression* entails “aggressively making attempts to finish the fight”. Chasing after an opponent and swinging wildly with no effective results would not count towards this criterion.

3. *Fighting area control* is scored by the fighter who is “dictating the pace, place and position of the match”.

The vast majority of rounds can be accurately scored using the first criterion alone, as even just one more effective strike would determine a winner.

Rounds are scored in-line with the “10-point must” system, which is the same system used in professional boxing. In this system, the winner will receive 10 points, and their opponent gets nine or less. Ties, i.e. 10-10, are permitted, but are actively discouraged in the Unified Rules: “a 10-10 round in MMA should be extremely rare and is not a score to be used as an excuse by a judge that cannot assess the differences in the round”.

By far the most common scoreline is a 10-9. This can reflect either an extremely close round, possibly determined by one more effective strike, or a very obvious round in which one fighter clearly outclassed their opponent.

A 10-8 round is the next most likely score, although rare compared to a 10-9. This should be given when an athlete wins the round by a large margin. There are three criteria a judge should assess when determining a 10-8 round. *Impact* is the first criterion, and evidence of this may be visible damage, such as lacerations or swelling, or diminishing the opponent’s energy, confidence, abilities, and spirit through your own striking or grappling. The next criterion, *dominance*, can be evidenced by the opposing fighter being forced to continually defend strikes, or using dominant grappling positions to attempt fight-ending submissions or attacks. Finally, judges will assess the *duration* for which the fighter maintains full control of the effective offense, whilst the opponent offers little effective in return.

Finally, in rare circumstances, a judge may give a 10-7 round. A fighter must completely overwhelm their opponent, demonstrating significant dominance and impact to such an extent the fight could be stopped.

Various fouls exist, including: headbutts, eye gouging, groin attacks, hair pulling, grabbing the cage, and grabbing an opponent’s shorts or gloves. Fouls are called by the referee, and deducting a point is entirely at the discretion of the referee. Any deductions will be applied to all of the judges’ scorecards within the round.

To find a judge’s overall winner of the bout, their scores are totalled and the athlete with the highest tally is the winner. If a majority of the judges score the bout in favour of a fighter, they are declared the winner. A bout can result in an overall draw, but again this is uncommon given the odd numbers of judges, odd numbers of rounds, and discouragement of 10-10 rounds. There are several possible outcomes of the fight, as detailed in Table 1.1.

Table 1.1. Different decisions which can be given based on the verdicts of the individual judges. The fight is between two fighters: Blue and Red.

Judges' overall winner			Blue	Red	Draw	Result	Decision
Blue	Blue	Blue	3	0	0	Blue	Unanimous win
Blue	Blue	Draw	2	0	1	Blue	Majority win
Blue	Red	Red	1	2	0	Red	Split win
Blue	Red	Draw	1	1	1	Draw	Split draw
Draw	Draw	Red	1	0	2	Draw	Majority draw
Draw	Draw	Draw	0	0	3	Draw	Unanimous draw

1.3 Thesis outline

With MMA introduced as a sport, I will now outline the remainder of the thesis.

Chapter 2 discusses the required literature. There are various topics to be covered: the types of models used for sports forecasting; how Markov chains have been applied to sports; betting in sports; combat sports judging; and the existence of biases within sporting officials. We also provide a summary of the findings from this literature review. With the surrounding literature detailed, we discuss the main contributions of this thesis.

Chapters 3, 4, and 5 contain the three research articles written throughout the PhD. Chapter 3 presents a novel Markov chain based forecasting methodology for MMA. We introduce 14 unique models to estimate fighters' skills, which drive the transition probabilities. This model is tested against two benchmark models, as well as the betting market.

Chapter 4 investigates potential biases within the judges. We explore the possibility that the judges are influenced by the crowd (using Covid-19 as a natural experiment) and the official UFC rankings. We suggest future reforms in light of the finding of this article.

Chapter 5 models the individual preferences of the judges and the fans. We use three historical case-studies to highlight different features and applications of the models, in particular we introduce a way of assessing if a fight outcome was fair.

Finally, Chapter 6 provides our conclusions and recommendations for future work.

Chapter 2

Literature review

Much of the academic literature surrounding MMA has been focused on the impacts on athletes' health. Bledsoe et al. (2006) reviewed the frequency of injuries and knockouts in MMA and found the rate of knockouts was lower than in boxing. Crighton et al. (2016) reviewed the methods employed by athletes when weight-cutting, noting the distinct link between dehydration and brain trauma risk. Lockwood et al. (2018) provided a systematic review of brain injuries in MMA athletes, concluding that the available data on the topic is poor. Hubbard et al. (2019) found that after a knockout, technical knockout, choke, or submission, the majority of athletes would perform worse in the King-Devick test–a visual-based test to assess concussion. Tiernan et al. (2020) tracked and recorded the accelerations and velocities of strikes to the head using electronic mouth-guards. The authors found that the impacts sustained by MMA athletes are shorter in duration than those in American football, due to the light gloves and lack of headgear. However, the human tolerance of repeatedly absorbing such blows is not known.

This chapter will introduce the past literature relevant to the thesis. In Section 2.1, we will focus on forecasting literature that is useful for Chapter 3, which presents a novel Markov chain (MC) based forecasting methodology for MMA. Section 2.2 will look at how MCs have previously been used in sports analytics. Chapter 3 compares our forecasting model against the betting market, and with that in mind, Section 2.3 discusses relevant literature on sports betting. Sections 2.4 and 2.5 will look at past research on combat sports judges and biases within sporting officials, respectively. Both sections are relevant to Chapters 4 and 5, which investigate the biases and preferences of MMA judges. Finally, Section 2.6 will detail the contributions of the thesis to the literature.

2.1 Forecasting in Sport and MMA

This section will introduce past literature on forecasting models within sport. The literature reviewed here is useful for Chapter 3 which introduces a Markov chain based forecasting methodology for MMA.

Whilst predictive sports models have been an area of academic interest for many years, the increase in data collection and availability, coupled with the rise of online gambling, have further increased their popularity over the last two decades. Indeed, the public now has access to sophisticated datasets, meaning professionals and hobbyists can develop models as well.

The recency of MMA's popularity–and, perhaps, the perceived randomness of results–means there has been little research into MMA forecasting to date. Finding publications in peer-reviewed academic journals is difficult itself. The only article we found was Hitkul et al. (2019). This study investigated the predictive ability of several machine-learning algorithms (with features based on numerous in-fight statistics). The best performing model achieved an accuracy of 62.84%. Several projects from bachelor's and master's degrees exist online, for instance, Johnson (2012), Ho (2013), Robles and Wu (2019), and Bartoš (2021).

In fact, there is little research into forecasting for any combat sports. Warnick and Warnick (2007) fitted a logistic regression predicting the outcome of boxing matches. The authors found significant positive effects for a fighters' total number of wins and whether they had won their previous bout. Significant negative effects were found for the fighters' age and their total number of losses. A further study found that a previous win against the current opponent, or in the current location, had significant positive effects (Warnick and Warnick, 2009).

We were unable to find relevant forecasting models for any other combat sport besides articles focused more on health and physiology. For instance, Sadowski et al. (2012), Podrigalo et al. (2018), and Kostrzewa et al. (2020) investigate the psychophysiological indicators correlated with success in taekwondo, kick-boxing, and judo, respectively.

Whilst accurate predictions are interesting and useful in their own right, the models can be used in other ways. For instance, Buraimo et al. (2022) used a forecasting model to estimate the significance of a given football match. This measure was then used as an independent variable in a second model to estimate the audience size for a game. Sæbø and Hvattum (2018) simulated full seasons of the English Premier League to evaluate the financial contribution of players. As a final example, Detotto et al. (2018) used a bivariate ordered probit regression to model the number of goals scored by opposing football teams. This enabled the identification of managers' characteristics that correlate with good performance.

Here, we will split the forecasting literature into two areas. Section 2.1.1 will discuss models that estimate the likelihood of different outcomes between competitors (which may be individual players or teams). Such models are often used to rate the relative quality of competitors or establish rankings. One such model (Bradley-Terry) will be used as a benchmark for comparison in Chapter 3. Section 2.1.2 looks at models estimating the scoring rates of competitors. Such models explore the mechanics of the games and allow more advanced predictions (for instance, how many goals will a football team score). Several rate models will be estimated in Chapter 3 to generate transition probabilities in the MC model.

2.1.1 Outcome models

Rating systems are often used within the sports forecasting literature. Perhaps the most famous is the Elo rating system, which was originally intended to rate chess players (Elo, 1978). An appealing aspect of Elo is its simplicity. Suppose competitors A and B play each other and are currently rated as R_A and R_B, respectively. Define S_A as the result from $A's$ perspective, which takes 1 for a win, 0 for a loss, and 0.5 for a draw. The expected result for A when playing B is calculated as

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}},$$

which is based on the logistic curve with base 10. The choice of 400 in the denominator is arbitrary. A's new rating is calculated as

$$R'_A = R_A + K \cdot (S_A - E_A),$$

where the value of K determines the maximum amount a competitor's rating can change after a game. Due to the model's simplicity, the ratings can be modified to better account for the user's sport of choice. For instance, if applying Elo to MMA, K could vary based on the different promotions; for example, fights within the UFC could have the maximum value, other global promotions (such as Bellator, One, and KSW) could take a slightly smaller value, whilst regional promotions are smaller again. A further multiplicative constant could scale the adjutment based on how the fight was won: perhaps weighing a knockout or submission more than a decision. An appealing aspect of Elo is the quick updates where a competitor's rating can be updated using only the new results, rather than re-fitting a model using all previous observations. The ratings also account for the change in competitors' strengths throughout time.

One unaddressed issue in the Elo ratings is the uncertainty of a competitor's rating. The

Glicko ratings (Glickman, 1995) introduce a term that represents this uncertainty, which determines how much a competitor's rating can change after a match.

TrueSkill (Herbrich et al., 2007) is a popular Bayesian system used (and patented) by Microsoft to assist in online match-making for Xbox games. Unlike chess, where the task is to rate individuals who compete in one-on-one matches, TrueSkill was required to rate individuals in games with many competitors, possibly split into teams. Similarly to Glicko, a competitor's skill comprises a mean rating and a measure of uncertainty.

The pi-ratings (Constantinou and Fenton, 2013) are similar to Elo, but account for a few specifics of football. For instance home advantage; that more recent observations are more important for predictions; and that whilst higher goal-difference is indicative of stronger teams, winning is more important overall. Similarly to Elo, the algorithm calculates an expected goal-difference and then adjusts teams' ratings according to the difference between the expected and actual goal-difference.

Comparison models (also referred to as ranking models) are also often used in sports. As with rating models, comparison models can establish relative strengths, generate predictions, or enrich larger predictive models. The Bradley-Terry (BT) model (Bradley and Terry, 1952) is probably the most famous comparison model. Suppose α_i and α_j represent the strength of competitors i and j. Then the BT model stipulates the probability i beats j is

$$P(i > j) = \frac{\alpha_i}{\alpha_i + \alpha_j}.$$

A binomial regression can represent this, where if i and j played each other n_{ij} times, with i winning y_{ij} matches,

$$\begin{aligned} y_{ij} &\sim \text{Binomial}(n_{ij}, p_{ij}), \\ \text{logit}(p_{ij}) &= \alpha_i - \alpha_j. \end{aligned}$$

The Plackett-Luce model (Luce, 1977) extends the problem to a situation with more than two competitors (for instance, horse racing). Suppose there are N competitors, who form a set S, where i_1 won, i_2 came second, and so on. Under Luce's axiom, the probability of selecting j from the set S is

$$P(j|S) = \frac{\alpha_j}{\sum_{i \in S} \alpha_i}.$$

Now the rankings are viewed as a sequence of choices, such that: i_1 is chosen from the full set of items, i_2 is chosen from the set minus i_1, and so on. Denote the set of alternatives to item j as $A_j = \{i_j, \dots, i_N\}$. Consequently, the probability of the full set of rankings is

$$P(i_1 > i_2 > \cdots > i_N) = \prod_{j=1}^{N} \frac{\alpha_j}{\sum_{i \in A_j} \alpha_i}.$$

One problem with the BT model is that it does not allow for ties. Two common ad-hoc solutions exist: defining a draw as half a win and half a loss; or the Bradley-Terry-Davidson model (Davidson and Beaver, 1977), where the parameter $\kappa \geq 0$, determines how likely ties are to occur, and

$$\begin{aligned} P(i > j) &= \frac{\alpha_i}{\alpha_i + \kappa\sqrt{\alpha_i \alpha_j} + \alpha_j}, \\ P(i = j) &= \frac{\kappa\sqrt{\alpha_i \alpha_j}}{\alpha_i + \kappa\sqrt{\alpha_i \alpha_j} + \alpha_j}. \end{aligned}$$

Baker and Scarf (2020) provide a more formal solution to ties. Suppose the discrete score[1] of competitor i is j_i, with probability p_{ij_i}, where lower scores are assumed to be better (for instance, in golf). Then

$$P(j_1 < j_2 < \cdots < j_n) = \sum_{j_1=0}^{\infty} p_{1j_1} \sum_{j_2=j_1+1}^{\infty} p_{2j_2} \cdots \sum_{j_n=j_{n-1}+1}^{\infty} p_{nj_n}$$

is the probability of the ranking $1, 2, \dots, n$ when there are no ties. If competitors 1 and 2 are tied, then

$$P(j_1 = j_2, j_1 < j_3 \cdots < j_n) = \sum_{j_1=0}^{\infty} p_{1j_1} p_{2j_1} \sum_{j_3=j_1+1}^{\infty} p_{3j_3} \cdots \sum_{j_n=j_{n-1}+1}^{\infty} p_{nj_n}.$$

The geometric distribution is then applied. The probability mass function is $p_{ij_i} = (1-p_i)p_i^{j_i}$, whilst the survival function is $S_{ij_i} = \sum_{k=j_i+1}^{\infty} p_{ik} = p_i^{j_i+1}$. Using these identities, for the case of two competitors, the authors obtain

$$\begin{aligned} P(j_1 < j_2) &= \frac{(1-p_1)p_2}{1-p_1p_2}, \\ P(j_1 = j_2) &= \frac{(1-p_1)(1-p_2)}{1-p_1p_2}, \end{aligned}$$

where p_i is associated with the strength of competitor i; which is the value of interest. Using

[1] In the continuous case (for instance, horse-racing times), ties cannot occur, as these measures cannot be equal.

the reciprocal of the geometric distribution's mean[2], the authors define $\lambda_i = (1-p_i)/p_i$, and thus

$$P(j_1 < j_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2 + 1},$$
$$P(j_1 = j_2) = \frac{1}{\lambda_1 + \lambda_2 + 1}.$$

Similar formulae are derived for the three competitor case, and can be extended to any general case.

Whilst the aforementioned models can predict outcomes, authors often use the strengths derived from these models as independent variables within a larger regression or machine-learning model. This methodology allows authors to expand the number of covariates. For instance, Hubáček et al. (2018) used gradient boosted trees to predict football results. The independent variables combined the pi-ratings with various other measures of skill, such as the teams' average goals scored and conceded within their last five games.

2.1.2 Scoring-rate models

The models described in Section 2.1.1 only consider the match outcomes, typically win, draw, or loss. Results in sports are often more granular, and the scores contain further information on competitors' skills. Consequently, researchers often model the scoring rates of competitors using generalised linear models, such as a Poisson regression.

Possibly the most famous scoring-rate model is the Dixon-Coles model (Dixon and Coles, 1997) (which itself is an extension of Maher (1982)), which predicts the scoring rates of football teams. These pioneering models allow teams to have two variables that reflect their attack and defence strengths. Suppose that in game n, teams $h(n)$ and $a(n)$ are playing home and away, and score x_n and y_n, respectively. Then the model can be defined by

$$x_n \sim \text{Poisson}(\lambda_n), \tag{2.1}$$
$$y_n \sim \text{Poisson}(\mu_n), \tag{2.2}$$
$$\log(\lambda_n) = att_{h(n)} + def_{a(n)} + home, \tag{2.3}$$
$$\log(\mu_n) = att_{a(n)} + def_{h(n)}, \tag{2.4}$$

where *home* represents home advantage, att_i is the implied attacking ability of i, and def_i the implied defensive ability of i (such that more negative values are optimal).

Such a representation is appealing as it mirrors the reality of the sport whilst providing

[2] The reciprocal is used since in this case lower scores, and thus lower means, are better.

good predictions. This is an art that has perhaps been lost in the age of big-data and machine learning, where the "black-box" models provide little to no inference beyond predictions.

The model can be extended into a Bayesian framework by setting priors on the parameters. For instance, Baio and Blangiardo (2010) estimated a hierarchical model with priors

$$\begin{aligned} \mathit{home} &\sim \text{Normal}(0, 0.0001), \\ \mathit{att}_i &\sim \text{Normal}(\mu_{att}, \tau_{att}), \\ \mathit{def}_i &\sim \text{Normal}(\mu_{def}, \tau_{def}), \end{aligned}$$

and corresponding minimally-informative hyper-priors,

$$\begin{aligned} \mu_{att}, \mu_{def} &\sim \text{Normal}(0, 0.0001), \\ \tau_{att}, \tau_{def} &\sim \text{Gamma}(0.1, 0.1). \end{aligned}$$

The hierarchical model suggests there is a common distribution from which the team-level parameters are drawn from. This is useful in the case of football, as the teams within a league each season form a natural hierarchy, and the extra information can potentially improve results. The minimally-informative hyper-priors allow the data to drive the inferences on the latent group-level parameters. Estimating this in a Bayesian framework with priors allows the uncertainty in estimates to be propagated into predictions, which is particularly useful at the start of the season when there is less information available on teams.

Often, authors will aim to model the dynamics of team strengths. Most commonly used is an exponential weighting scheme, which down weights older games, meaning future predictions will rely more on recent games. This was the method used by Dixon and Coles (1997). Typically, when estimating at time t, the weight for a game which occurred at time t_n is calculated as $e^{-\phi(t-t_n)}$ where ϕ is a parameter to be tuned on a validation set. In this context, the pseudo-likelihood becomes

$$L(\mathit{home}, \boldsymbol{att}, \boldsymbol{def} | \boldsymbol{x}, \boldsymbol{y}) = \prod_{n=1}^{N} \left(\frac{e^{-\lambda_n} \lambda_n^{x_n}}{x_n!} \right) e^{-\phi(t-t_n)} \cdot \prod_{n=1}^{N} \left(\frac{e^{-\mu_n} \mu_n^{y_n}}{y_n!} \right) e^{-\phi(t-t_n)}.$$

To be explicit, for game n, the likelihood of observing x_n home goals under the model specified by equations 2.1 and 2.3 is $e^{-\lambda_n} \lambda_n^{x_n} / x_n!$. This is then weighted by the time-decay function $e^{-\phi(t-t_n)}$. Equations 2.2 and 2.4 similarly provide the likelihood of observing y_n away goals.

Others have allowed team strengths to change stochastically. For a given team i, with attack strength at time t equal to $\alpha_{i,t}$, we can assume a random walk, whereby $\alpha_{i,t} =$

$\alpha_{i,t-1} + u_t$, for $u_t \sim \mathrm{N}(0, \sigma_\alpha^2)$ (Fahrmeir and Tutz, 1994). This can similarly be modelled by an auto-regressive process (Koopman and Lit, 2015), or Brownian motion (Rue and Salvesen, 2000).

Whether scoring rates between opposing teams are independent has been discussed since Maher (1982), who expanded the model using a bivariate Poisson. Indeed, Dixon and Coles (1997) also used a bivariate Poisson and then included an ad-hoc adjustment to account for low-scoring games, which are estimated poorly. Baio and Blangiardo (2010) argued that the hierarchical Bayesian framework implicitly allows for correlation between teams' scoring rates, as the hyper-parameters drive inferences on the observed goals of both teams. McHale and Scarf (2011) explicitly include the correlation between rates by using copulas. Suppose we assume home and away goals follow a joint distribution with cumulative density function $F(x, y)$. Then by Sklar's theorem (Sklar, 1959), for a copula C, we have $F(x, y) = C(F_x(x), F_y(y))$, where F_x and F_y are their marginal cumulative density functions. The marginal distributions may have the same attack and defence parameterisation as the previously discussed models. Any appropriate marginal distribution can be used: Poisson, negative binomial as in McHale and Scarf (2011), or a Weibull count distribution as in Boshnakov et al. (2017). The likelihood for such a model is

$$\begin{aligned} L = \prod_{n=1}^{N} \big\{ & C(F_x(x_n), F_y(y_n)) - C(F_x(x_n - 1), F_y(y_n)) - \\ & C(F_x(x_n), F_y(y_n - 1)) + C(F_x(x_n - 1), F_y(y_n - 1)) \big\}. \end{aligned}$$

Scoring-rate models such as those discussed apply to any sport with an attack and defence aspect. For example, one could model points won by the server in a tennis match, where the attack and defence strengths could be interpreted as serving and returning ability. As with the Elo ratings, the model is simple and flexible, so one can easily control for other aspects, such as the different surfaces in tennis, or as we will see in Chapter 3, the different weight classes in MMA.

2.2 Markov chains in sports analytics

Now we will review how Markov chains have previously been applied in sports. Again, the following literature is useful for Chapter 3.

MCs and simulation models have also been successfully applied in sports analytics. Turn-based sports with an obvious attacker and defender at any given point–such as in tennis, where one player is serving and the other returning–are more amenable to such a framework, as it is simpler to break down the transitions from state to state.

Carter and Crews (1974), Newton and Keller (2005), and O'Malley (2008) all developed equations to calculate the probability of winning a game in tennis based on a MC representation. The states of the chain represent each possible score in a game, for example, 0-0, 30-15, deuce (40-40), and advantage (Adv-40). Suppose the probability the serving player wins a point is p, and conversely, $q = 1 - p$ represents the probability the returning player wins. These values thus define the transition probabilities between states. A game can be won by winning four points whilst the opponent wins none, one, or two. Otherwise, if both players win three points, the game goes to deuce, and a player must obtain a two-point lead. The probability of winning a game is thus

$$G(p) = p^4 + 4p^4q + 10p^4q^2 + 20p^3q^3 \cdot p^2 \sum_{i=0}^{\infty} (2pq)^i,$$

where $2pq$ is the probability of transitioning from deuce back to deuce after two points. The probabilities of winning a tie-break, set, and finally, a match can be calculated from $G(p)$ and $G(q)$.

An appealing aspect of Markov models is that in-game dynamics can be included. For instance, one could relax the assumption that p is constant throughout the match, Newton and Keller (2005) suggested a model for player's point-win probability that deviates slightly from their mean depending on the current score-line.

In baseball, Bukiet et al. (1997) assumed each pitch could result in six possibilities: an out, walk, single, double, triple, or home-run. The probability of each can be estimated or modelled using past information on the pitcher and batter. The state of the game is described by how many outs the batting team currently has, and whether there is a runner on each base. From this, an optimal batting order can be calculated, as can the expected number of wins.

In American football, each drive by a team ends in one of several ways: the team can score through a touchdown or field goal; a new drive may begin after a down; or possession may switch through an interception, fumble, missed field goal, or punt. Goldner (2012) model the expected points from a given state, which includes information such as the current position on the field. This can be used to determine how particular players or teams are performing.

MCs can allow users to make advanced inferences, such as the optimal batting order in baseball. Predictions can also be more detailed, for instance, how many points, games, or sets each player will win in a tennis match. Uncertainty due to randomness is inherently included in the predictions, so one can easily report prediction intervals. A further advantage and application are "in-play" predictions: predicting an outcome given the current state of the game.

Since these sports are turn-based and composed of many individual plays, MCs are relatively simple to apply. On the other hand, free-flowing and dynamic sports, such as football, basketball, and MMA, are harder to apply MCs to since the states and transitions are not as obvious. First, one has to discretise the free-flowing game into a Markov representation. For instance, Shirley (2007) represented basketball by three factors: which team is in possession (home or away); how they gained possession (inbound pass, steal, defensive rebound, offensive rebound, or a free throw); and how many points were scored on the previous possession (0, 1, 2, or 3); thus giving 40 possible states. Transition probabilities associated with the average home and away team were estimated using corresponding percentages over the data. A retrospective analysis shows that the model provides realistic simulated games. Štrumbelj and Vračar (2012) used the same MC representation, but estimated transition probabilities using multinomial logit models and tested the model on out-of-sample games. The independent variables of the transition models consisted of various team-level summary statistics. By simulating the chain 10,000 times, the authors obtained probabilistic forecasts of match results. However, Elo ratings and bookmaker odds both performed significantly better.

Hirotsu and Wright (2002) model football using just four states which represent when either team is in possession or when either team has scored. The authors detail transition probability models which include an attack and defense component for each team (similar to Dixon and Coles (1997)). Using a hypothetical set of games, they show how the optimal timing of substitutions, or a change in tactics, can be determined using the model, depending on the circumstances of the game. In Hirotsu and Wright (2003), the authors use this MC representation to evaluate the different playing styles and abilities of Premier League teams.

Rudd (2011), introduced the theory of expected possession value in football (and possibly sports as a whole), using MCs to break down a team's possession into a discrete sequence of moves to identify top performers. In total, 39 states were defined: 30 using zonal location and defensive state, seven set-pieces (e.g. penalty, short corner, and long corner), and two absorbing states (goal or end of possession). A team could move the ball through each state by a deliberate action, such as passing or dribbling. Advanced inferences can be made, for instance, the probability of each team scoring from every position on the pitch (which may highlight deficiencies). The key insight was that each action could be valued by finding whether (and by how much) it improves the probability the possession ends with a goal. This idea has been the cornerstone for many more player-ratings systems, for instance in basketball (Cervone et al., 2016), ice-hockey (Schulte et al., 2017), and football (Decroos et al., 2019; Liu et al., 2020).

2.3 Betting literature

Chapter 3 concludes with a comparison of our forecasting model against the betting market. Consequently, here we will discuss relevant literature on sports betting.

Naturally, many papers focused on predicting sports outcomes conclude with an exercise testing whether the model could obtain out-of-sample profits in the betting market. Three main strategies are used: flat unit stakes, Kelly stakes, and portfolio optimisation. To allow fair comparisons of models and results, "units" are the standardised currency used in simulations. Flat unit stakes refers to the most basic strategy of staking one unit on the model's outcome prediction. This is often used as a benchmark strategy as the stakes do not depend on the probability of the selection or the odds.

Perhaps the most well-known strategy is Kelly staking (Kelly, 1956). The Kelly criterion was derived to provide the user with the maximum expected log wealth when continuously reinvesting their wealth. If the probability an event occurs is calculated to be p, and b is the proportion of the bet gained with a win (e.g. a win with decimal odds of 2.00 gives $b = 1.00$, whilst fractional odds of 2-1 gives $b = 2.00$), then

$$k = p - \frac{1-p}{b}$$

is the proportion of one's bankroll that should be staked. In practical settings, the regular Kelly criterion can often lead to bankruptcy. Consequently, authors often implement a fractional Kelly, where $f \cdot k$ of the bankroll is staked for $0 < f < 1$. Several modifications of Kelly have been implemented. For instance, Boshnakov et al. (2017) added a minimum expected value threshold and removed the reinvestment assumption, whilst Matej et al. (2021) included a draw-down constraint to reduce the ruin probability.

Markowitz's portfolio theory (MPT) (Markowitz, 1952) involves optimising the risk-adjusted expected returns of a portfolio of bets. Whereas the Kelly criterion is applied to consecutive single bets, MPT inherently can be applied to multiple games simultaneously. Given several bets and assuming that one's stakes must sum to 1, the Sharpe ratio can be calculated as the expected returns divided by the standard deviation of returns. This can be optimised to establish the best portfolio. Whilst the Kelly criterion is more commonly used in the forecasting literaure, several articles have applied MPT to sports betting. For instance, Fitt (2008) and Hubáček et al. (2019) show how portfolio theory may be applied to create an optimal betting portfolio for football and basketball, respectively; whilst Matej et al. (2021) compared the results of Kelly and MPT across different sports.

Hubáček et al. (2019) argued that the predictions from one's model need to be sufficiently decorrelated from the bookmakers' odds to obtain good profits. The authors suggested two techniques to ensure decorrelation. First, one can include the bookmakers' odds as weights of the observations. When the underdog wins, the weight is the bookmakers' odds. Alternatively, if the favourite won, the weight is 1. The second technique alters the maximisation objective, so the difference between the model's prediction and the bookmakers' odds is considered when fitting. In each method, the model learns to be more accurate when disagreeing with the bookmaker.

2.4 MMA and combat sport judging

In Chapters 4 and 5, we investigate the judges of MMA using two distinct models. Thus, here we will discuss past literature regarding MMA and combat sports judging.

Since the judges can directly determine the outcomes of bouts in combat sports, they play a far more crucial (and often controversial) role than in other sports. We found three peer-reviewed journal articles examining MMA judges.

Gift (2018) fitted both binary and ordered logistic regression models to the scores given by UFC judges. The differences in various in-round statistics were included as independent variables. Additional variables were included to investigate different sources of bias: whether an athlete was the champion, their implied win probability derived from the bookmaker's odds, whether they had an insurmountable lead going into the round, whether they won the previous round, and finally, whether they were deducted more points than their opponent within the round. The coefficients for the in-round statistics were as expected: the only significant negative effect was missing takedowns, the only non-significant variables were missing strikes or stand-ups, and the most influential actions were tight submissions and knockdowns. The author found significant positive effects for the implied win probability and insurmountable lead variables, indicating reputation and recency biases, respectively. However, neither of these conclusions are robust, as both variables are likely correlated with in-round performance and may contain skill information "unseen" by the count variables.

In Collier et al. (2012), the authors modelled the overall winner of the fight (not the round) using a probit model. The differences in the aggregated counts of the various performance measures were used as independent variables, as were several athlete characteristics: height, weight, and age. Given that the judges do not score fights overall (instead, they score each round), the conclusions from this study are limited. Nonetheless, the authors found that visible damage and knockdowns had the largest average marginal effects. The only significant effect from non-performance variables was height.

Feldman (2020) used several logistic regressions to examine how the judges weigh the three sequential criteria specified in the Unified Rules and detailed in Section 1.2.1: effective striking and grappling, effective aggression, and fighting area control. Again, knockdowns were found to be the most influential action. Overall, the author concluded that the judges largely adhere to the criteria.

In Balmer et al. (2005), the authors examined whether judges contribute to home advantage in boxing. The authors modelled whether the home fighter or away fighter won, having removed neutral fights or those that ended in a draw. The relative difference in quality (measured as overall win percentage) and an indicator for whether the fight ended with a points decision were included as independent variables. The significant positive effect of the points indicator suggests that the judges do contribute towards the home advantage.

An experiment in Myers et al. (2012) saw ten experienced muay Thai judges score a fight in two different conditions: with and without crowd noise. The authors found that judges awarded significantly more strikes in the presence of crowd noise. An interaction term describing whether the noise was for the home or away fighter was also significant.

Boxing judges have a long and notorious history of bad decisions. One of the most infamous was the 1999 defeat of Lennox Lewis to Evander Holyfield[3]. This particular match was investigated in Lee et al. (2002). The authors used exact tests and logistic regressions (in both frequentist and Bayesian frameworks) to compare the judges' scores against the scores submitted by the experts in the media. This article is particularly relevant to Chapter 5 as the authors include indicators for each judge within the logistic model. These indicators assess whether (and by how much) each judge favoured Holyfield. The authors pool the scores within each round to obtain round-win probabilities and thus find the probability of each judge's scorecard. For each judge's scorecard, a *p*-value is calculated as the probability of observing at least as many rounds scored for Holyfield as given by the judge. In all their results, the judges scored the bout significantly different to the media.

2.5 Biases literature

The final topic of literature covered here is relevant to Chapter 4, which explores the existence of different biases within MMA judging. In our context, a bias is defined as a tendency to prefer one thing over another that prevents objectivity or that influences outcomes in some way[4].

Detecting biases exhibited by sporting officials is a crucial topic. Four key sources of bias investigated in the literature which we will discuss are:

[3] As per https://bleacherreport.com/articles/1677686-ranking-the-15-worst-judging-decisions-in-boxing-history, this bout was the third worst decision in boxing of all time.

[4] https://sociologydictionary.org/bias/

- Nationalistic bias: when an official favours a competitor representing their own home country.
- Racial bias: when an official favours a competitor based on their race.
- Home bias: when an official favours a competitor playing at home.
- Reputation bias: when an official's prior beliefs on a competitor will impact their observation.

2.5.1 Nationalistic bias

Campbell and Galbraith (1996) investigated whether nationalistic bias existed within the judges assessing Olympic figure skating. First, the authors implemented a non-parametric sign test based on the deviance of each judge's score from the median. Strong biases were shown from this, however, the size of the biases was not found. Consequently, an ordinary least squares regression was set up, whereby the deviation was the target variable, and an indicator for whether the judges shared nationality with the assessed skater was the only independent variable. Again, strong evidence of a nationalistic bias was found, although the effect was small (less than the 0.1 increments between scores). Zitzewitz (2006) found further evidence of nationalistic bias in figure skating at the 2002 Olympics[5].

Heiniger and Mercier (2018) again looked at nationalistic biases within the Olympics, this time focusing on gymnastics. A similar method was utilised, regressing the judges' scores with several variables possibly indicative of bias. Nationalistic bias was assessed by including an indicator variable that represented whether the judges shared nationality with the athlete. A further indicator, representing whether an athlete was of a nationality directly competing against an athlete of the judge's nationality, allowed the authors to examine whether the judge penalises their nation's opponents. The median score of all judges for the performance was included as a control, as were the judges' individual tendencies to score higher or lower. Different levels of variation between scores were found across the other disciples, which was also included in the regression. Significant bias was found in several disciplines and, in some cases, grew more substantial in the later stages of the competition.

Evidence of nationalistic bias has also been found in: Olympic diving (Emerson et al., 2009), in which the authors included interaction terms between specific judges and countries, thus examining more than just same-nationality biases; gymnastics (Heiniger and Mercier, 2018); and muay Thai (Myers et al., 2006), a sport closely related to MMA.

[5] Figure skating has been the subject of much research, due in part to two separate instances of vote trading and collusion at the 1998 and 2002 Olympics. See, for instance, (Zitzewitz, 2014), who show the reforms actually led to more collusion.

2.5.2 Racial bias

Own-race bias was examined in baseball in Parsons et al. (2011). The hypothesis was that umpires would call more strikes when their race matched the pitcher's, having accounted for various controls. Consequently, an indicator representing whether the pitcher's and umpire's race matched was included (UPM: umpire pitcher match). Overall, no significant effect was found; however, the authors suspected the level of scrutiny faced by the umpire may alter their behaviour. Thus, they assess three situations when the umpire would be under more scrutiny: when the stadium features QuesTec technology[6]; when there is a larger attendance; and when the pitch was not "terminal"[7]. In each scenario, UPM was signifciant and positive when the umpire faced less scrutiny.

In Price and Wolfers (2010), the authors were interested in observing own-race bias within basketball (regardless of the player's race). To this end, they modelled the number of fouls earned by individual players. The variable of interest was the interaction between the percentage of white referees and an indicator representing whether the player was black (race was rather loosely defined as a binary variable), and numerous control variables were included. Own-race bias was evidenced by the significant positive effect of the aforementioned variable.

2.5.3 Home team bias

Whilst home advantage is a widely accepted and well-documented phenomenon, the mechanics driving it are less well understood. Nevill and Holder (1999) provide a review of home advantage. They determine the crowd's influence on the officials is likely the most significant factor, whilst they may also provide a psychological lift to the home team. Travel factors, such as fatigue, are also relevant when the away team must travel between time-zones. Nevill et al. (2002) investigate how the crowd can influence the referee. They conclude that social pressure will likely influence the referee, in that they want to avoid the crowd's displeasure. Further, the referee may use the salient crowd noise as a heuristic when unsure on a decision. There is an abundance of research on home advantage in football, which we will discuss in the following.

Garicano et al. (2005) found that referees will award extra-time to favour the home team (e.g. more time if the home team are losing by one goal and less time if they are winning by one goal). Further, they found that when the proportion of away fans increased, the size of bias was significantly reduced. Sutter and Kocher (2004) found similar evidence, whilst

[6] This technology tracks the ball using cameras around the stadium. Consequently, the umpire's calls can be monitored.

[7] That is, if another strike ends the batter's at-plate appearance

they also investigated the distribution of penalties awarded to home and away teams. A significant difference was found, but robust conclusions could not be made as, for instance, the home team will generally attack more, and thus receive more penalties. The authors had details on whether a penalty was "irregular" (that is, shouldn't have been awarded), and whether a regular penalty was "refused". After removing irregular penalties, they hypothesise that the relationship between regular and refused penalties should be the same for home and away teams. A chi-squared test indicates a significant bias towards home teams, who are awarded 81% of their regular penalties, whereas away teams receive 51%.

Behind-closed-doors (BCD) games provide a natural experiment, allowing researchers to observe how the officials' behaviour changes without a live audience. Pettersson-Lidbom and Priks (2010) found that the home team are punished less than the away team when there is a crowd, yet in the absence of a live audience, the home team are punished more than the away team. The authors show that the players' statistics (e.g. number of shots on target) do not significantly change between the two settings, providing evidence this is an effect stemming from the referee, not the players. The Covid-19 pandemic offered further opportunities to study BCD matches. Reade et al. (2020) found that referees awarded significantly fewer yellow cards to away teams during these matches.

Using counts of statistics over the duration of an entire game is a common problem within the research, as such counts do not consider in-game dynamics. For instance, an issue highlighted within Sutter and Kocher (2004) was that although the home team received more penalties, they also attacked more. Buraimo et al. (2010) addressed this issue by using minute-by-minute football data to inform a bivariate probit, which modelled whether the home or away team would receive a card given various controls that wouldn't be available otherwise, for instance, the minute of the match and the goal difference at the time. The authors observed home teams received significantly more cards when their stadium featured a running track around the pitch. They conclude this is because the home team's fans are further from the referee, and thus unable to exert as much pressure. Three teams switched stadium design within the data, providing more robustness to this conclusion.

Whilst we have solely discussed football here, home advantage has been found in many other sports. For instance, boxing (Balmer et al., 2005), muay Thai (Myers et al., 2012), Olympic combat sports (boxing, fencing, judo, taekwondo and wrestling) (Franchini and Takito, 2016), cricket (Morley and Thomas, 2005), whilst Schwartz and Barsky (1977) provide evidence in ice-hockey, basketball, and baseball.

2.5.4 Reputation bias

In gymnastics, a coach will often place their athletes in order of worst to best. Consequently, this has become a “self-fulfilling” prophesy in that judges expect athletes performing later are better. Several experiments have found a significant reputation bias deriving from the order of the athletes (for instance, Scheer and Ansorge, 1975; Scheer and Ansorge, 1979; Plessner, 1999). In each experiment, the authors edited videotape of different performances, so that some athletes would appear in different positions within their respective teams.

An experiment by Findlay and Ste-Marie (2004) demonstrated that reputation bias exists within figure skating judges. In this experiment, 12 judges (six from Ontario and six from Québec) were selected to assess the performance of 14 different skaters. These skaters were chosen so that half were known to the Ontario judges (and unknown to Québec), and the other half known to the Québec judges (and unknown to Ontario). Paired t-tests found that an athlete’s ordinal placement was significantly better when the judge knew their name.

Jones and Erskine (2003) also investigated reputation bias, this time with an experiment involving football referees. Two groups were shown the same clips of various incidents that could be labelled fouls within a football match. The experimental group were told beforehand that the team committing the fouling incident was known for their aggressive style of play. The authors found a significant effect through chi-squared tests, such that the experimental group would penalise the team with more cards than the control group. There was no significant difference in the number of fouls or the response time to the incident.

As discussed in Section 2.4, Gift (2018) found a significant reputation bias within MMA judging, in that judges were influenced by the bookmakers’ pre-fight odds.

Erikstad and Johansen (2020) found further evidence of reputation bias within football. In this case, an expert panel of four referees reviewed 43 potential penalty situations which involved teams with a successful reputation, and 55 situations not involving any successful teams. The authors then compared the experts’ verdicts with the actual match official’s using a chi-square test. The authors observed that the successful teams would receive significantly more penalties than they should, and their opponents would receive significantly fewer.

2.6 Thesis contributions

2.6.1 A Markov Chain Model for Forecasting Results of Mixed Martial Arts Contests

The preceding review highlights the lack of forecasting literature on MMA, and combat sports as a whole. Despite a vast catalogue of such literature in other sports, only one MMA article exists. Only two relevant papers investigating boxing predictions were found, which is interesting given the long-standing worldwide popularity of the sport. We were unable to find relevant peer-reviewed journal articles developing forecasting models for any other combat sports, that weren't focused on the psychophysiological indicators of success. There is a distinct lack of data available within combat sports, which is likely a reason for the scarce amount of literature. Whilst MMA results can be found on websites such as `sherdog.com`, detailed fight statistics are only currently available for the UFC (via `ufcstats.com`). In boxing, `boxrec.com` provide a comprehensive set of historical results, but in-fight statistics are not publicly available.

Despite the advantages of a MC based forecasting model (more detailed predictions, access to in-play predictions based on the current state of the game, and the inclusion of uncertainty due to randomness), compared to more traditional sports forecasting methods, relatively few articles utilise MCs or simulation models. The majority of MCs used in sporting literature focus on player/team-based inferences, and "simpler" turn-based sports, such as tennis. There are two main obstacles when applying MCs. First, to accurately model the transition probabilities between states, more granular data is necessary, which may not be publicly available. Second, there is an increased computational demand; for instance, when simulating the chain to obtain predictions, or estimating various transition models. A final point to make it that whilst one can obtain more detailed predictions from a MC, they do not necessarily out-perform more traditional methods.

In Chapter 3, we will present a forecasting methodology for predicting the outcome of MMA bouts using Markov chain based simulations. We develop a Markov chain representation of the sport, where the states represent the key actions and positions within MMA. The transition probabilities between each state are driven by 13 Bayesian generalised linear models, which account for the different abilities and styles of athletes. From 327 out-of-sample test matches, we find the model predicts comparably to the bookmakers, and can generate positive returns when used as the basis of a betting strategy. As discussed, there are very few MC based models in the forecasting literature, thus our approach is comparatively novel.

Modelling the individual skills of athletes is also an understudied area. Whilst this has been done in other sports (for example, Szczepański and McHale (2016) model the passing ability of individual football players), there has been little application to sports forecasting, with most authors focusing on player-based inferences. We point out that most MC based models use multinomial models trained on event-data (which is typically very large and behind a paywall) to explicitly model the transitions between states. An appealing feature of our skill models is that they use more basic data which is publicly available. This approach is thus flexible enough to be applied to any sport with similar data available.

To generate judges' decisions during the simulations, we use a simple logistic regression that is estimated using historical UFC fights. In Chapters 4 and 5, we will further develop this judging model.

2.6.2 Reputation Bias and Home Crowd Influence in Judging: The Case of Mixed Martial Arts

There has been much research into detecting biases within sporting officials. The most researched forms of bias are nationalistic, racial, home team, and reputation; which have been observed in many different sports. The main motivation for such research is because it matters: integrity and fairness are of paramount importance to sports and its continued appeal. The seriousness can be evidenced by the advancements in technology specifically designed to make sports fairer, including Hawkeye in tennis, video-assisted referees in football, and QuesTec in baseball.

We note that in the three aforementioned papers regarding MMA judges (Gift, 2018; Collier et al., 2012; Feldman, 2020), the authors had access to the proprietary non-public data collected by the company FightMetric. FightMetric allowed academics to apply for research access but unfortunately were no longer registered at the beginning of this PhD. This data data was more detailed than the public data scraped for the work of this PhD. It featured several improvements, including but not limited to: splitting control-time into various positions, such as clinch, side-control, and back-control; indicators for whether a fighter suffered visible damage; and splitting submissions into chokes and locks, or whether they were tight.

Chapter 4 explores the possibility of different biases within the judging of MMA. To accomplish this, we created an expansive database of MMA scores, which included 17,105 unique verdicts given by judges. The data available on MMA makes it an ideal sport to study the judges, as we have counts of various actions we can use to examine the judges' behaviour. This is in contrast to other subjectively judges sports, such as diving, where there are no measures of success other than the judges' scores. Typically, authors will use the average score in a regression as an independent variable; however, the counts in the MMA data allow our modelling to be more robust.

Our results show that there is a significant reputation bias, such that athletes placed higher in the rankings are significantly favoured by the judges. This is an important finding with possible parallels in other sports (do VAR referees favour higher ranked teams in football?), or even other industries such as academia (are well-known academics more likely to have their papers published?). Compared to other forms of bias, reputation bias is relatively understudied. Previous studies on reputation bias have all been experiment-based, thus our study is, to our knowledge, the first to identify reputation bias using data *directly* from professional sports. In what is still a growing sport, issues such as this must be raised and addressed quickly to maintain trust between the promotion, judges, athletes, and fans.

During the Covid-19 pandemic, fights took place with no live audience, which enabled research into the effect of the fans on the judges. We find that the judges significantly favour home athletes in the presence of a crowd, with no significant effect for behind-closed-doors events. The impact of crowds on the officials' behaviour has been investigated across many sports, but we are the first to study it within MMA. We also contribute to the knowledge of how Covid-19 affected sports, which is vital for future planning.

In addition to these findings, our technical contribution is that we are the first to apply the "purposeful selection" model fitting strategy in the literature on biases. In the context of biases, where the interpretation of the final model is crucial, this is a valuable technique which blends key advantages of machine-learning algorithms with the well-understood inferences of logistic regression.

2.6.3 Individual Preferences and Controversial Decisions in Mixed Martial Arts Judges

Although there has been much research into sporting officials and judges, no one has directly modelled their individual preferences. This is despite often acknowledging different opinions likely exist. Some works have included dummy variables for officials: Lee et al. (2002) to assess whether boxing judges were biased in Lewis vs. Holyfield; Parsons et al. (2011) to control for different baseball umpires likelihood of calling a strike; and Buraimo et al. (2010) to control for football referees who may give more or less fouls.

Whilst some sports have explicit and thorough guidelines for judges, MMA is largely left to the interpretation of the judge. This means their opinions, which may derive from their own martial arts background, can influence the outcome of fights, with huge monetary and career implications at stake. With this in mind, the final project of the PhD, contained in Chapter 5, models the scores given by MMA judges in a Bayesian hierarchical framework, allowing each judge to have their own opinions on each action.

We identify several actions where judges have a wide range of opinions, even to the point of opposite signs. Using a historical case study, we show the judges' preferences are large enough to be the deciding factor in a fight. This is an important finding for MMA and can be used to educate athletes, fans, stakeholders, and the judges themselves. We argue that athletes can even use the findings to assist in their game-plan if they know what a judge values most.

We also model the fans' scores in a similar manner and compare with the judge model. Overall, we find the fans score each action similarly to the judges. The biggest discrepancy is how likely fans are to give the rarer scores (10-10, 10-8, and 10-7). The most intriguing finding is that it appears the fans are actually *less* influenced by sources of bias, such as a home fighter or the rankings. Given some promotions allow fans to determine the official results, research into whether the fans are a viable replacement, or useful addition, to the judges is important to the sport.

Whilst a judge may not choose the most likely score, their verdict could be "within reason". To mathematically assess this, we apply the concept of variable significance to the predicted probabilities. We apply this to three different controversial fights, and show that some judges' scores were indeed controversial, whilst others were fair. Despite clear applications–for instance, we may predict a fighter to win but not significantly enough we would want to bet on them–we are yet to see any other discussion regarding this concept.

Finally, we use our model to obtain fair-scores by removing the effects of bias terms and the individual judges' preferences. We re-score a particularly infamous match, and show the wrong athlete won. Given the technological advancements in sport over the last two decades, such scores could be used in a variety of manners by all participants in MMA.

Chapter 3

A Markov Chain Model for Forecasting Results of Mixed Martial Arts Contests

In this chapter, we will present a novel methodology for predicting MMA bouts using Markov chains. The article, presented here, has been published in the International Journal of Forecasting (Holmes et al., 2022). An earlier version was presented at the European Sports Economics Association in 2019, where it received the Best Young Researcher Paper award.

3.1 Introduction

3.1.1 Background of Mixed Martial Arts

Mixed Martial Arts (MMA) is a full-contact combat sport which, as the name suggests, incorporates aspects of all martial arts: throws and submission moves from judo, Brazilian jiu-jitsu, and wrestling; and strikes from boxing, muay Thai, and taekwondo, to name just a few. In modern-day MMA, an athlete needs to be competent in each facet of martial arts to compete at the highest level.

The definitive origins of MMA are up for debate. It is hard to pinpoint the exact moment, as contests between practitioners of different martial arts occurred throughout the 20th century in East Asia. Furthermore, the traditional martial arts Vale Tudo (Brazil) and Sambo (Soviet Union) are both full contact, unarmed combat sports which utilise techniques from many martial arts; in other words, they are “mixed” martial arts.

Within MMA, there are numerous organisations at local, national, and international levels. The Ultimate Fighting Championship (UFC) is considered the top-tier organisation; their first event occurred in 1993 and is when MMA started to gain popularity. The Unified Rules of Mixed Martial Arts were not set until 2001 by the New Jersey Athletic Control

Board, and this can be seen as the beginning of present-day MMA.

MMA has grown rapidly in popularity in recent years. For example, the last television broadcasting rights contract signed between UFC and ESPN was for 30 events to be aired during a five-year deal worth a reported USD$1.5bn in 2018. To put this into perspective and demonstrate the size of the potential audience for the UFC, the largest television rights contract for football is for the English Premier League, which agreed to a deal from 2019 to 2022 to show 200 games per season for an estimated GBP£5bn.

3.1.2 An MMA contest

In scientific work on sports such as football or tennis, a thorough description of how a match is played and what events can occur is, for the most part, unnecessary. Despite its recent surge in popularity, the same assumption of knowledge cannot be made for MMA contests, and with this in mind, we here detail how a bout unfolds.

Typically, contests are fought over three five-minute rounds. Within the UFC, main-event and title fights are extended to consist of five five-minute rounds. Compared with boxing at the highest level, in which there are usually 12 three-minute rounds, the round structure of MMA is quite different. Shorter rounds in MMA would give an advantage to fighters who favour striking, as those who grapple would have a limited amount of time to progress to advantageous positions.

As in boxing and other combat sports, fighters are split according to their weight into different 'weight classes'. There are currently eight men's weight classes in the UFC (the upper limit is given for each): Heavyweight (265lb), Light Heavyweight (205lbs), Middleweight (185lbs), Welterweight (170lbs), Lightweight (155lbs), Featherweight (145lbs), Bantamweight (135lbs), Flyweight (125lbs). The first women's fight in the UFC took place in February 2013 and has now grown into four women's weight classes: Featherweight (145lbs), Bantamweight (135lbs), Flyweight (125lbs), and Strawweight (115lbs).

Athletes can win fights through a strike resulting in a 'knockout' or a successful 'submission' attempt (consisting of various chokes and joint locks). A fight ending by one of these methods is often referred to as a 'finish' victory. In these cases, the fight ends before the time limit has been reached. If neither fighter wins early through a finish, then the contest must be scored by the judges.

Generally, there are three judges who must assign a score to both fighters in each round. Rounds are scored using the '10-point must' system: at least one of the fighters must be awarded 10 points. Usually, rounds are scored as 10-9, with 10 being awarded to the victor. However, if one fighter is deemed to have won by a significant margin, the judge may score the round as 10-8 or lower. A round can be scored as a draw, 10-10, but judges are encouraged not to, making it extremely rare.

Fouls can occur and cover a wide variety of offences, including: illegal strikes such as head-butts and groin strikes, grabbing an opponent's shorts or gloves, or hair-pulling. Referees can issue warnings or deduct points from fighters. Any point deductions will be applied to the scores of all judges.

To find who a judge deemed the fight's winner overall (with contestants Blue and Red), their scores for each round are summed. The possible outcomes of a fight given the verdicts in each round are shown in Table 3.1.

Table 3.1. Different outcomes of a fight based on the verdicts of the individual judges.

Judges' overall winner			Blue	Red	Draw	Result	Decision
Blue	Blue	Blue	3	0	0	Blue	Unanimous decision
Blue	Blue	Draw	2	0	1	Blue	Majority decision
Blue	Red	Red	1	2	0	Red	Split decision
Draw	Draw	Draw	0	0	3	Draw	Unanimous draw
Draw	Draw	Red	0	1	2	Draw	Majority draw
Draw	Blue	Red	1	1	1	Draw	Split draw

Each round begins with both fighters standing. When standing with a reasonable separation between them, the combatants are said to be at 'distance'. Whilst at distance, fighters will try to strike one another with punches, kicks, elbows etc. These strikes can target anywhere on the opponent's body (with exceptions according to the rules).

Some fighters will excel at fighting from range and keep distance between themself and their opponent, while others prefer to get close. Fighters may engage in a 'clinch' (when both contestants are standing and grappling with one another). The clinch can be helpful in limiting your opponent's ability to strike, as well as setting up 'takedowns' (grappling techniques used to bring an opponent to the floor). Once on the ground, some athletes prefer to strike, while others will look to gain an advantageous position and attempt a submission.

The fight can also go to the ground through a 'knockdown' (a strike that causes the opponent to fall to the ground, indicative of a brief loss of consciousness). This is a much different scenario to going to the ground through a takedown, when a fighter will be clear-headed. A knockdown is often followed by a finish victory, as the opponent tries to regain their composure they are still "dazed" and thus vulnerable to subsequent strikes and submission attempts.

MMA is unlike boxing in that a fighter's record is not "protected" by their manager. While fighters may have agents, match-making is done by the organisation they compete in, whose primary concern is to arrange for the most attractive (and lucrative) fights to occur. This is one reason it is scarce to see undefeated records in MMA since the strongest fighters are asked to fight against each other.

3.1.3 Predicting an MMA contest

Predicting the results of Mixed Martial Arts bouts, and indeed other combat sports such as boxing, presents several problems not present in many other sports:

- The low frequency of fights and non-homogenous times in-between for athletes will cause rating systems such as Elo (Elo, 1978) or TrueSkill (Herbrich et al., 2007) to struggle.

- Like boxers, MMA fighters have particular fighting styles and understanding how a pair of fighters' styles will interact during a bout is of critical importance when predicting the fight's outcome. As such, typical ratings models will not capture the nuances of the sport.

- Due to the large pool of fighters spread across numerous organisations, the ever-shifting rankings, and the low frequency of fights, many fighters may never fight each other, making pairwise comparisons difficult.

- The outcome of bouts are often not as simple as a binary win or loss. In the case of disqualifications or no-contests[1] the fight result is "uninformative" in that it doesn't reflect who the better fighter was. Bouts in which the judges disagreed on who won the fight can be classed as "controversial". In the 4,678 fights in 2001-2018, there were 13 disqualifications, 49 no-contests, 32 majority decisions, 454 split decisions, 19 majority draws, and 8 split draws. This gives a total of 575 (12.29%) uninformative or controversial outcomes.

- In combat sports, since any strike can result in a knockout, an athlete is only ever one punch away from winning. The phrase a "puncher's chance" refers to an athlete having at least a small chance of winning, despite being an underdog. This is especially true in MMA since combatants wear 4oz gloves, compared to heavier 8oz (or more) in boxing; lighter gloves mean less padding and lead to more knockouts. This, combined with the considerable variety in stylistic match-ups, means results are particularly vulnerable to noise.

[1] This is the recorded outcome when a fighter can no longer continue in a contest due to an accidental foul by their opponents. Fights may also be retrospectively declared a no-contest if an athlete tests positive for a banned substance.

There is very little literature on forecasting the results of MMA bouts. Johnson (2012), Ho (2013), Hitkul et al. (2019), and Robles and Wu (2019) all use various machine learning algorithms consisting of similar variables to predict the winner of a fight. Varying accuracies across the different models were reported, ranging from 50% to 68%; though, none were tested against the betting market.

Aside from the machine learning models in their paper, Ho (2013) implemented a contest as an adversarial game with random elements. A "fight" appeared to involve three plays (representing three rounds), and allowed three actions: strike, takedown, or submission. The accuracy in this model was reported to be 54%.

Robles and Wu (2019) implemented a k-means algorithm to identify three different styles of fighter: striker, grappler, and well-rounded.

3.1.4 Applying Markov chains to MMA

To circumvent the difficulties of modelling MMA described in Section 3.1.3, our approach is to drill into the mechanics of the sport before simulating a contest using a Markov chain. We estimate fighter skills in various aspects of the sport, for instance, how often a fighter will attempt a strike.

We then build a Markov chain model of a fight with transition probabilities determined by the various skill models. By simulating the chain a significant number of times, we obtain detailed predictions for each fight, beyond what one could achieve from a binary win-lose model.

Markov chains have been successfully applied to sports such as tennis (O'Malley, 2008), baseball (Bukiet et al., 1997), and American football (Blanc et al., 2016). For the most part, these games are turn-based (there are two distinct roles and players/teams swap between them in a structured manner), and they can be described as a sequence of individual plays. Combat sports share the dynamic traits of sports such as football, basketball, and ice hockey: players/teams can swap roles at any time.

Despite increased complexities, Markov chains have also been used to model dynamic sports. In basketball, Shirley (2007) presents a model to estimate the points in a game. In ice hockey (Schulte et al., 2017) and football (Haave and Hoiland (2017) and Szczepanski (2015)), Markov chain models are used to quantify the value of player actions, and thus rate players. Damour and Lang (2015) use Markov chains to model the outcome of set pieces in football.

The paper is organised as follows. The data we will utilise throughout the paper will be introduced in Section 3.2. In Section 3.3, we detail the skill models that will drive the transition probabilities in the Markov chain. The Markov chain representation of MMA is detailed in Section 3.4. A simulation model of MMA would not be complete without a model to simulate the judges' decisions; we present the corresponding model in Section 3.5. Two benchmark models are presented in Section 3.6. We compare our Markov model with actual fight results and statistics, the two benchmark models, and the betting market in Section 3.7. Finally, Section 3.8 contains our conclusions and suggestions for future work.

3.2 Data

We obtained fight statistics of UFC bouts from 2001 to 2018 from two sources: espn.com (ESPN) and ufcstats.com (UFC-Stats), using a combination of the rvest (Wickham, 2020) and rSelenium (Harrison, 2020) packages within the R programming language (R Core Team, 2020). The dataset we collected amassed 4,678 fights and 1,680 unique athletes. There are many fighters who have competed in few contests: 806 athletes competed in three or fewer fights over 2001-2018. Our strategy is to use fights from 2001-2017 as training data, leaving 2018 for testing. There were 4,204 contests during the training period; giving a total of 8,408 observations to be used for training (one per fighter per fight).

Our data includes the following statistics for each fighter, given as totals over the entire bout: significant[2] strikes (split into different positions from which a strike was attempted: distance, clinch, or ground; different targets: head, body, or leg; and finally, whether they landed or not), takedowns (split into whether they were successful or not), knockdowns, submission attempts, control-time (how long a fighter dominated their opponent in grappling situations), and the number of positional advances (moving to a more advantageous grappling position on the ground). We also gathered the fight result (who won), method of result (how they won), date, weight class, duration of the bout, and maximum number of rounds.

We make three simplifications to the striking data to assist our modelling:

- Due to the nature of 'non-significant' strikes, they have very little impact on the fight. Consequently, we only use significant strikes in all our modelling. To avoid needless repetition throughout the remainder of the paper, any strike can be assumed to be significant.

- Without more granular data on the clinch, it cannot be adequately modelled. More granular data could include, for instance, how many times the fighters engaged in a

[2] Significant strikes include all strikes attempted from distance, and any strikes from clinch and ground which are deemed to be powerful enough. The official fight scorers determine what constitutes a "power" strike.

clinch and who was in control of each. Consequently, we chose to amalgamate distance and clinch striking statistics into one "standing" category.

- Whilst head strikes are attempted to knock the opponent out, body and leg strikes aim to slow the opponent down and tire them. With that in mind, we combine body and leg strikes into one "body" category, and any reference to body strikes can be assumed to include both body and leg strikes.

In some cases, generally, when a fighter has left the UFC, these statistics are no longer hosted on ESPN. In such circumstances, we use the data available on UFC-Stats. There is a subtle but important difference in the granularity of the ESPN and UFC-Stats data. As such, in 68 training observations we imputed the striking totals. Appendix A.1 provides a full explanation.

We scraped historical odds from `bestfightodds.com` (BestFightOdds). The data consists of the closing odds from several bookmakers, including but not limited to Bet365, William Hill, Pinnacle, and Intertops.

3.3 Estimating fighter skills

In this section, we will present several models used to estimate the skills of MMA athletes in different aspects of the sport. These models will then drive the transition probabilities of the Markov chain model of the sport, which is introduced in Section 3.4.

A commonly used idea in modelling sports is to estimate the attack and defence strengths for each competitor. For example, in football Maher (1982) and Dixon and Coles (1997) estimate attack and defence strengths of the teams. The attraction of such methods is not only that the results provide accurate forecasts but that the idea mirrors the mechanics of the game. Similarly, in tennis (see Klaassen and Magnus (2001), for example), it is common to model the outcome of points as a function of the serve and return strengths of the two players. Here we adopt this framework in the context of fighter skills in MMA.

Unlike tennis, in which the authors have reduced the number of a player's skills to two, serving and returning, skills in fighting are many and having an advantage in one aspect can be the deciding factor in a bout. We fit 13 different skill models to estimate each of the transition probabilities in the Markov chain depicted in Section 3.4.

Estimating the attack and defence strengths for each of the fighters' skills is complicated since there is often limited data on certain competitors. This can be for two reasons: either the fighter has not competed in many UFC fights, or the athlete fights in a style such that certain actions are rare. For instance, a karate expert may have never attempted a takedown.

In the Bayesian approach used here, estimated skills for these fighters are influenced strongly by the prior distribution, and thus they are pulled from either extreme towards the average. As the amount of data on a fighter increases, the estimated ability will rely less on the prior and more on the outcome of the fighter's attempted actions.

We fit generalised linear models in a Bayesian framework using an expectation maximisation (EM) algorithm through the `bayesglm` function within the `arm` package (Gelman and Su, 2018). In all models, we estimate each parameter using the weakly informative priors recommended in Gelman et al. (2008); that is, a Cauchy distribution with center 0 and scale of 10 for the regression intercept, 2.5 for binary predictors, or $2.5/(2 \cdot sd)$ for numerical predictors (where sd is the standard deviation of the predictor). The recommended prior induces a reasonable amount of shrinkage for coefficients, whilst still allowing some larger coefficients. We believe this is important in the context of our data: with a limited amount of data on each fighter, setting too strong a prior will hinder us from quickly detecting fighters' unique skills.

In the case of the models we fit, the algorithm generates an augmented dataset including pseudo-observations based on the prior distributions. The model is estimated by alternating between one step of iteratively weighted least squares on the augmented data set, and one step of EM. The algorithm estimates approximate posterior modes and the covariance matrix of the coefficients; this allows an approximate posterior density to be generated. More details can be found in Gelman et al. (2008).

It is well known that fighters in different weight classes will possess different attributes; for instance, heavier weight classes will produce more knockouts. Consequently, we include the upper limit of the weight class a fight takes place in as a covariate. The exact weight of each fighter is not available in the data, however, overweight fighters will forfeit a portion of their money to their opponent, and underweight fighters offer their opponent an advantage in the bout; hence fighters usually weigh in at exactly the upper limit of the weight class.

To avoid repetition, we only mathematically define two of our models in the following sections. Table A.3 in Appendix A.2 contains the formal definitions of all models. All in-fight statistics used in the models are displayed in Table A.4.

Our models require us to know the fighter's control-time in the clinch and the ground separately. Since the data we obtained does not split control-time by position, we must estimate both from the available data. We find the proportion of ground control-time as the ratio of ground-based techniques (ground strikes, submission attempts, and positional advances) to the sum of both ground and clinch techniques. The estimated ground control-time follows as the product of the total control-time and the ground control proportion[3]. One can find the clinch control-time similarly.

We now define some notation which will be used in the following sections. For athletes i and j competing in fight k, let T_k be the total bout duration in seconds, C_{ik} be i's total control-time, GC_{ik} be i's estimated ground control-time, and CC_k be i's estimated clinch control-time.

3.3.1 Work-rate models

We model three "work-rates", which predict the number of strikes, takedowns, and submissions a fighter will attempt during a contest. While these may seem like individual traits, the ability to stop one's opponent from attempting techniques is a crucial skill in itself. This can be through range control or rendering the opponent unable to attempt techniques through grappling. Consequently, we allow for an attack and defence parameter in each of these models. We include weight class in the models to allow work-rates to be lower for heavier fighters.

Let SA_{ijk} denote the total number of standing strikes attempted by fighter i against j in contest k. Similarly, denote by GA_{ijk} the number of ground strikes attempted. Let lbs_k be the weight class of the fight in pounds. There are four parameters to be estimated: the intercept (str_int), the attacking ability of i (str_att_i), the defensive ability of j (str_def_j), and finally the effect of weight (str_weight). The abbreviation *str* refers to strike rate and is necessary to identify the different parameters across the various skill models.

We estimate these models using informative offsets with the knowledge that certain actions can only be performed from particular positions: for instance, takedowns can only be performed whilst standing. In the case of strikes, we assume they can be attempted whenever, hence include the total bout duration as the offset. Consequently, we have

$$\mathrm{SA}_{ijk} + \mathrm{GA}_{ijk} \sim \mathrm{Poisson}(str_{ijk}) \tag{3.1}$$

$$\log(str_{ijk}) = str\_int + str\_att_i + str\_def_j + str\_weight \cdot \mathrm{lbs}_k + \log(\mathrm{T}_k). \tag{3.2}$$

We model takedown rates (tdr) similarly, the only change being the choice of offset. As mentioned, fighters can only attempt a takedown whilst standing. Define the "stand-time" to be the amount of time a fighter was standing and their opponent was not in control, $\mathrm{ST}_{ik} = \mathrm{T}_k - \mathrm{C}_{jk} - \mathrm{GC}_{ik}$. We use this as the offset in the takedown rate model.

Submission rates (smr) again follow similarly, using the fighter's ground control-time,

[3] In some cases, a fighter would have zero ground control time despite having landed a takedown. In such circumstances, we assume the fighter had ground control equal to the minimum non-zero value.

GC_{ik}, as the offset. Implicitly, we assume that a fighter can only attempt a submission from a dominant grappling position on the ground. In reality, submissions can be performed from any position, even from standing at distance with moves such as a flying arm-bar or an iminari roll. However, the majority will occur whilst in control on the ground, and we do not have the data on what position a fighter attempted a submission from; hence, we assume all come from top-control.

When estimating the work-rates, we removed fights that were less than one minute in length. We found that there were several fights within this threshold which resulted in unrealistic work-rates. It is fair to assume that fighters' work-rates in the first minute of a fight will not align with most of the fight: either with low rates while "feeling out" their opponent or high rates while fresh.

Note that to simplify the framework, we assume independence between all of the skill models. Consequently, we also assume a fighter's skill parameters are independent.

3.3.2 Strike, takedown, and submission accuracy models

Models for the accuracy of fighters in their strikes, takedowns, and submissions (whether an attempted technique is successful) are required for our Markov chain. We model four different striking accuracies based on two positions (standing or ground) and targets (head or body), and two grappling accuracies: takedowns and submissions. Again, we allow athletes to have an attack and defence rating in each skill and allow weight to have an effect.

Denote by SHL_{ijk} and SHA_{ijk} the standing head strikes landed and attempted, respectively, by fighter i against opponent j in fight k. Then we can model the accuracy using a binomial model with a logit link such that,

$$\mathrm{SHL}_{ijk} \sim \mathrm{Bin}(\mathrm{SHA}_{ijk}, sha_{ijk}) \tag{3.3}$$

$$\mathrm{logit}(sha_{ijk}) = sha\_int + sha\_att_i + sha\_def_j + sha\_weight \cdot \mathrm{lbs}_k, \tag{3.4}$$

where the abbreviation *sha* denotes standing head accuracy. Similarly, we can model the other five accuracies using the corresponding totals: standing body accuracy (*sba*), ground head accuracy (*gha*), ground body accuracy (*gba*), takedown accuracy (*tda*), and submission accuracy (*sma*).

3.3.3 Knockout or knockdown probability model

Being able to throw powerful strikes which can knock an opponent out is a revered skill to possess. Not only can this overcome shortcomings in other skills, but the ensuing highlights will boost the fighter's popularity, leading to more lucrative fights involving higher-ranked opponents. A knockdown can be considered a semi-knockout: as mentioned in Section 3.1.2, they often precede a knockout victory.

To incorporate knockdowns into the simulation, the transition probabilities in the aftermath of one would have to reflect the scenario: that the opponent is vulnerable and on the verge of being finished. Our data does not include information which would enable us to model this scenario accurately.

However, since there is clearly a lot of information on an athlete's power and knockout ability contained within knockdowns, we pool knockouts and knockdowns together[4], estimating the probability of either happening using a binomial model where head strikes landed are the trials.

3.3.4 Strike target models

To model how often a fighter targets their opponent's head with strikes whilst standing, we use the number of standing head strikes as successes in a binomial model whereby the total number of standing strikes are trials. We model the corresponding ground model similarly. The fitted probabilities from these models are denoted by *shp* and *ghp*: standing head strike probability and ground head strike probability, respectively.

These models are the only skill we do not include a defensive component to. We argue that this is an individual tactic and an opponent has no influence on this. Including a defensive term would lead to biases when athletes have fought an opponent who favours one target.

3.3.5 Control-time per takedown and stand-up probability models

The ability of a fighter to keep their opponent grounded after successfully landing a takedown (and vice-versa, the ability for an opponent to get up after being taken down) is of great importance. Keeping an opponent down allows the fighter more time to gain dominant positions to land strikes and attempt submissions whilst limiting the opponent's ability to perform techniques.

We model the ground control-time per takedown landed using a gamma distribution, allowing for attack and defence abilities and an effect for the weight class. Given a fighter's predicted ground control-time per takedown, gc_{ijk}, we can then find their opponent's probability of forcing a stand-up (per second) to be $stnd_{jik} = 1/gc_{ijk}$.

[4] A maximum of one knockout can happen in a fight, but there can be multiple knockdowns to either fighter.

3.4 A Markov Chain Model for MMA Fights

We now define a Markov chain model for an MMA contest between fighters i and j. In Section 3.4.1, we discuss the Markov chain and the simulation procedure in a broad sense. In Section 3.4.2, we detail the states and transitions involved whilst the fighters are standing. In Section 3.4.3, we repeat this in the context of the ground states.

Figures 3.2 and 3.3 display the associated transition probabilities. The underlying models generating these probabilities were described in Section 3.3. Further, Appendix A.2 contains a table summarising the different skill models (Table A.3).

3.4.1 Overview of the chain

Figure 3.1 provides an overview of how a contest can progress and how the states connect. The chain detailed in Figure 3.1 does not represent the full chain used for simulations, as the striking states are more detailed than shown here. Shaded states are displayed in more detail in Figure 3.2. States with a dashed border are displayed in more detail in Figure 3.3.

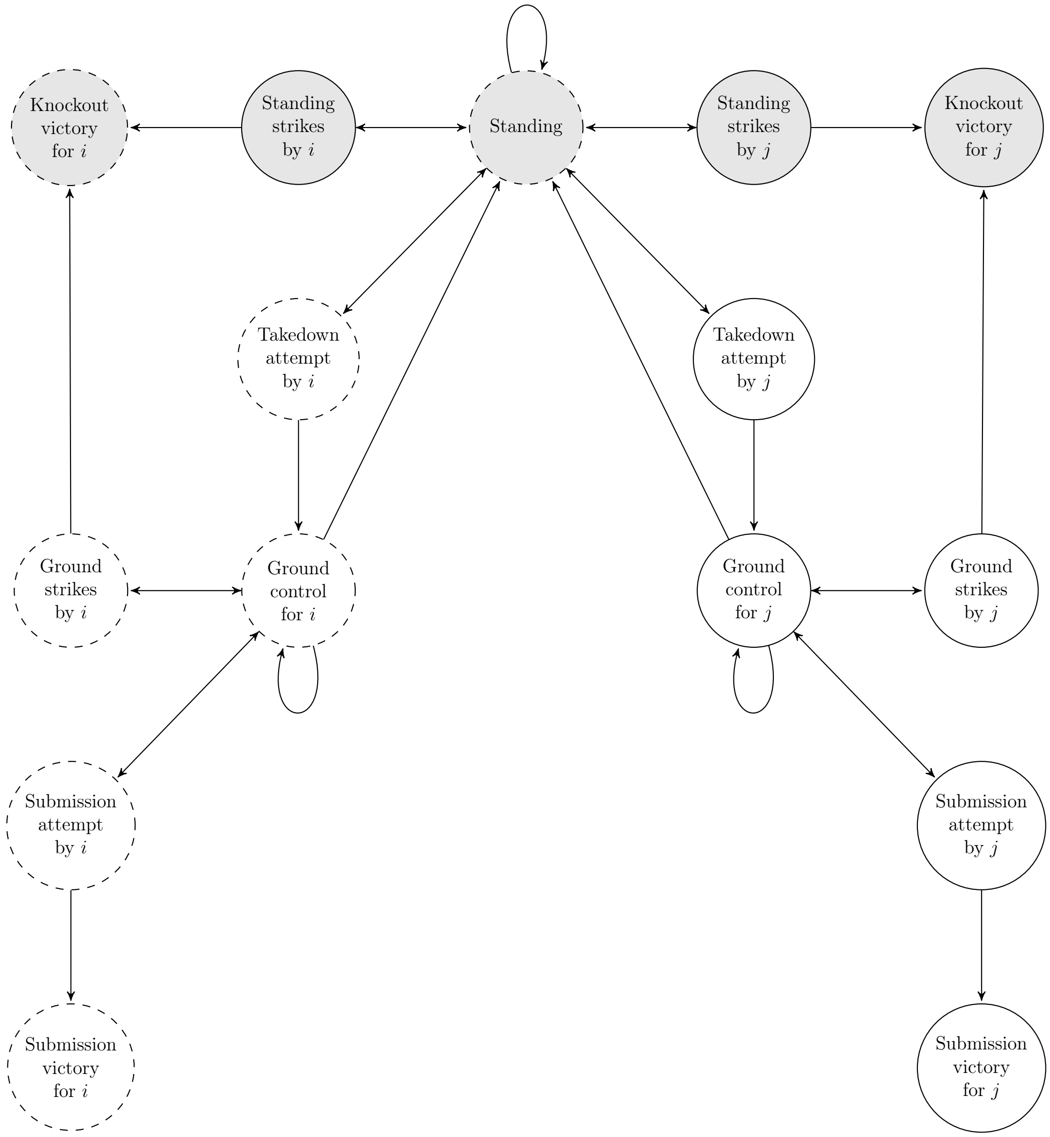
Looking at Figure 3.1, from standing fighters can attempt strikes which can, in turn, lead to knockout victories[5]. Successful takedowns will take the chain to the fighter's ground state, where the state "*Ground control for i*" implies i is in control of j on the ground.

From the ground, the athlete in control can perform strikes and submissions; both can lead to finish victories. The fighter being controlled can force the fight back to standing through a stand-up. While in reality fighters on the bottom can perform strikes, they would likely not be deemed significant by the judges and thus have no impact on the fight. For this reason, we omitted strikes from the bottom in our chain.

We simulate a contest as three or five-minute rounds, depending on the fight's status. Each round begins from the neutral standing position, as in reality. Since our work-rate models in Section 3.3.1 estimate the rate of actions per second, we set the time-lags between iterations of the chain to be one second. Appendix A.3 includes a short example simulation to help clarify how time passes. A chain is run until either it transitions into one of the absorbing finish states–in which case we terminate the chain prematurely–or the time-limit is reached. If the time-limit is reached and neither fighter wins via a finish, the fight must be 'judged'. We present our model for judging in Section 3.5.

[5] Whilst in reality strikes, takedowns, and submissions may be thought more as "actions" than "states", for the Markov chain they serve as states and will be referred to as such.

Figure 3.1. An overview of all states and transitions involved in the Markov chain for simulating an MMA contest. Note that this is not the full chain used to simulate a fight.



3.4.2 Striking States

Figure 3.2 displays how various striking techniques from the neutral standing state are included in the Markov chain. Only transition probabilities associated with fighter i are explained; the transitions for the opponent, j, follow similarly.

From the neutral standing state, strikes are attempted by i at a rate of str_{ij}, thus transitioning from *Standing* to *Standing strike attempt i*. Once in this state, i is committed to attempting a strike. We use the predicted rates from the Poisson GLM described in Section 3.4.2 (equations 3.1 and 3.2) to allow strikes to occur at a constant rate. This is similar to a Markov queuing process–specifically an M/M/1 queue–where customers arrive into the queue at a constant rate λ according to a Poisson process, moving the chain from state S to $S+1$ (Kleinrock, 1975).

From *Standing strike attempt i*, we must determine the target of i's strike. They target the head with probability shp_i, and conversely the body with probability $1 - shp_i$.

Suppose i targets the head, and we transition into state *Standing head attempt i*; this strike lands with probability sha_{ij}, according to the model specified by equations 3.3 and 3.4. If the strike does not land, then the chain transitions back to *Standing*.

Recall from Section 3.3.3 that we pooled knockouts and knockdowns together in one model (since both provide a lot of information on a fighter's striking power). Using this model to generate knockout victories directly would lead to an overestimation (since not all knockdowns lead to knockouts); hence, a successful head strike leads to a knockout victory with probability $\widehat{kdo}_{ij} = adj_{ko} \cdot kdo_{ij}$, where $adj_{ko} < 1$. As with tuning the parameters of a machine learning algorithm, we optimise this value to provide the best predictive accuracy on an out-of-sample validation set; we will discuss this further in Section 3.7.

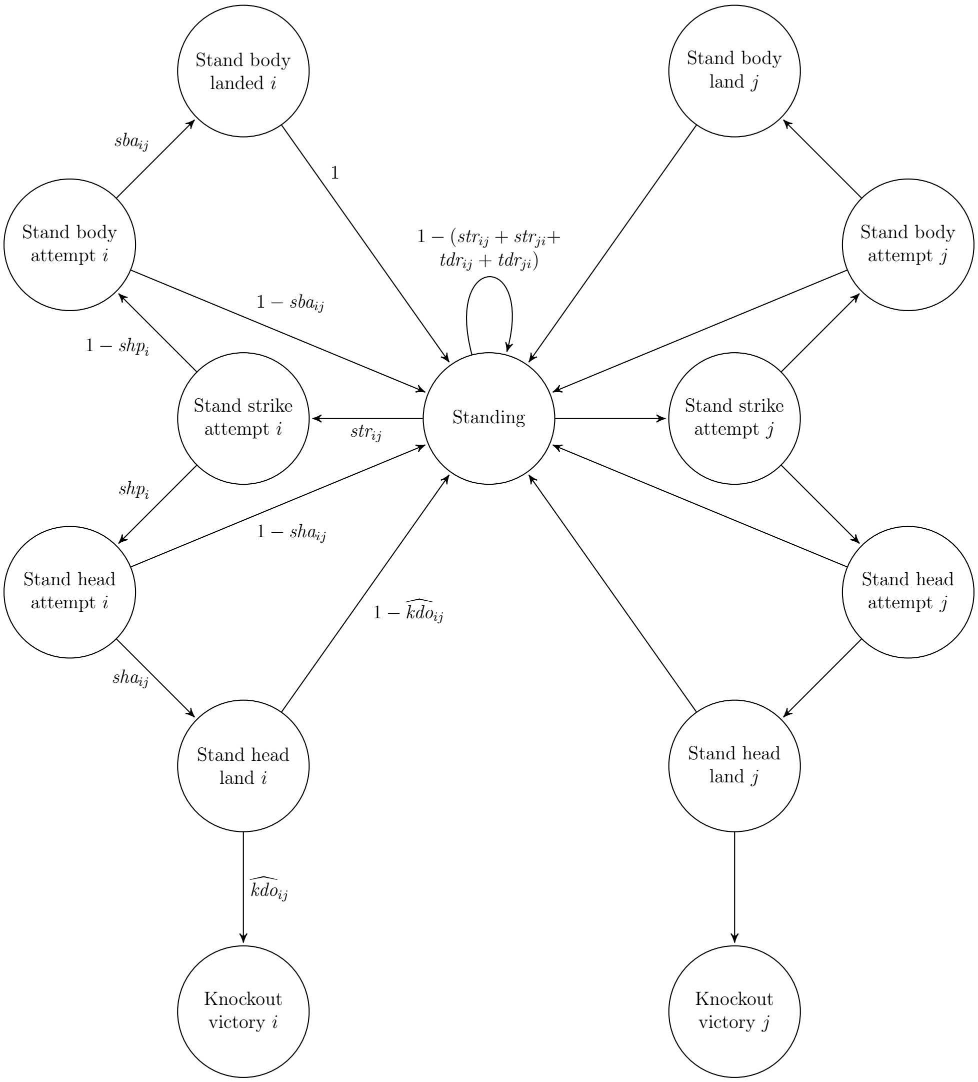
A strike attempt by i targeting the body lands with probability sba_{ij}. We do not allow for body strikes to cause a knockout; thus, the a body strike will transition back to *Standing* whether the strike lands or not.

3.4.3 Ground states

The ground is perhaps the most difficult aspect of MMA to model. There are numerous positions, each has its own advantages and disadvantages. We simplify the situation and for each fighter include one ground state where they are in control, state *Ground control i* implies i is in control and on top of j. Figure 3.3 displays the states involved in getting to the ground and what follows after.

To obtain control on the ground, a fighter must first successfully land a takedown. From the neutral standing state, i attemtps takedowns at a rate of tdr_{ij}. A takedown is successful

FIGURE 3.2. Markov chain diagram displaying the different states and transitions involved in striking techniques for both athletes.



with with probability tda_{ij}, thus transitioning to *Ground control i*. A failed takedown attempt transitions back to the neutral standing state.

From *Ground control i*, i can attempt a strike or submission. Strikes follow a similar flow to the standing equivalents: head and body strikes can miss or land, and landed head strikes can result in a KO victory. However, the associated probabilities differ from the standing equivalents.

From their ground control state, i attempts submissions at a rate of $\widehat{smr}_{ij} = adj_{sm} \cdot smr_{ij}$, where $adj_{sm} > 1$. We found using the 'raw' rates led to simulations with too few submission attempts. We believe this is because not all ground positions are created equal: submissions are only viable from a handful of ground positions, in which an athlete may not spend much time. The time spent in each position is not contained in our data, hence we inflate the submission rate in the simulations.

A submission is successful with probability sma_{ij}. A successful submission transitions to the absorbing *Submission victory i* state and the simulation is complete.

Finally, the chain can transition back to the neutral standing state from i's ground control state with probability $stnd_{ji}$.

3.5 Modelling the judges' decisions

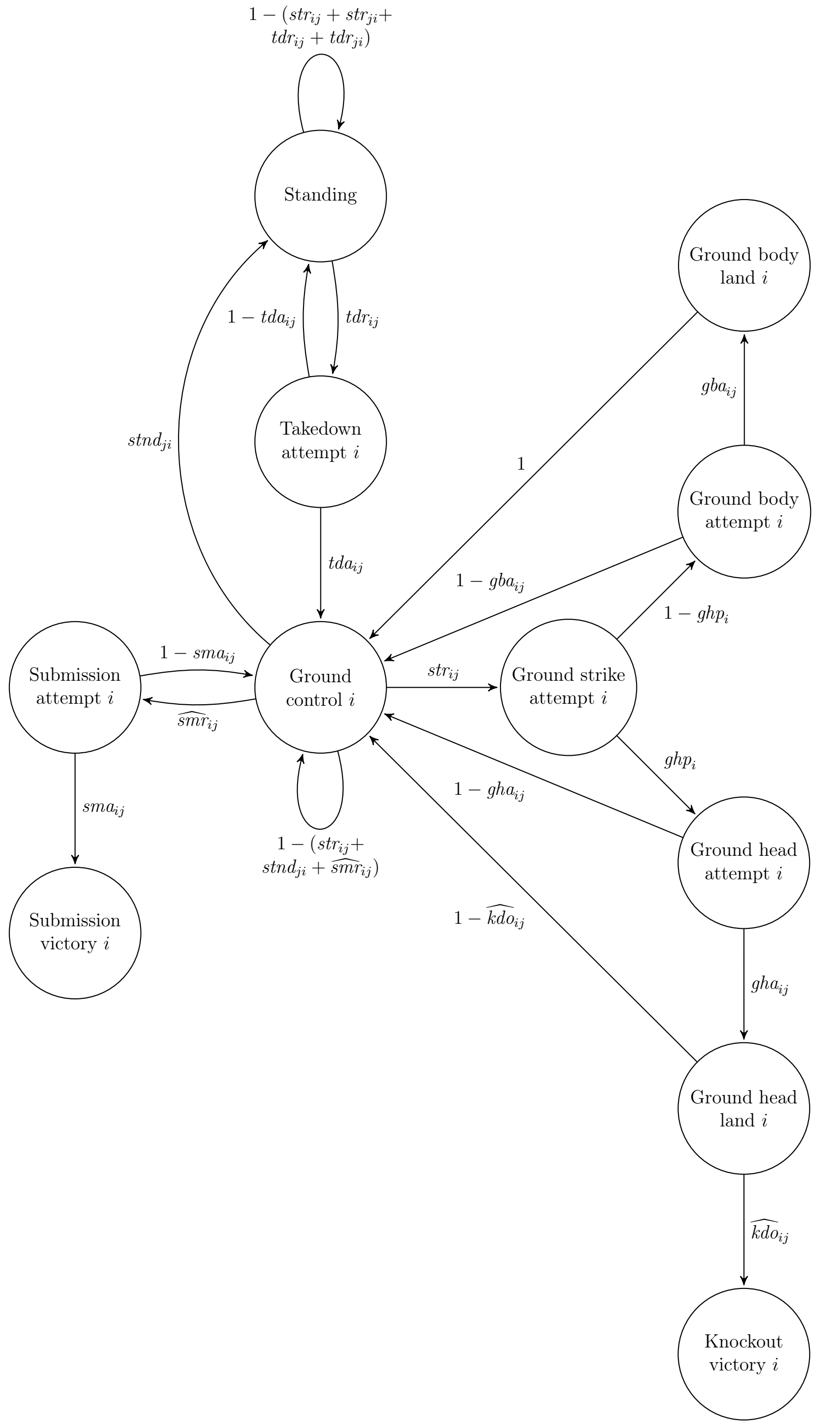
A fundamental aspect of combat sports is the judges' verdict on who wins the fight when neither athlete has won via finish. These decisions are subjective and often the subject of controversy, as discussed in Section 3.1.2. A simulation model of MMA contests would not be complete without a judging model: without it, one would not know how to score simulations meaningfully.

There has been little scientific research investigating the decisions of MMA judges. Gift (2018) estimated two models: one logistic regression modelling only 10-9 scores, and one ordered probit regression modelling the full range of possible scores. In both, the author used the differences of in-round performance statistics for opposing fighters and non-performance variables, such as: whether the fighter is the champion, the bookmaker odds of them winning, and whether they won the previous round.

Collier et al. (2012) modelled the judges overall score using the fighter's statistics for a whole fight using similar variables. Clearly, this is not ideal: rounds are scored individually and should, in theory, be treated independently by each judge.

We scraped all available UFC scorecards from `mmadecisions.com` between 2001 and 2017. For each round of each fight in the UFC, we have the scores of the three judges who

FIGURE 3.3. Markov chain diagram displaying the states and transitions involved on the ground from athlete i's ground control state.



were scoring that fight when that fight did not end due to a finish[6]. We then merged these scores with the round-by-round total statistics available from UFC-Stats.

The vast majority of rounds are scored 10-9 to the winning fighter. With this in mind, we chose to model round victories as a Bernoulli random variable, with success being winning a round by any margin.

We include only three variables: the differences in strikes landed, takedowns landed, and submission attempts[7]. The Unified Rules of Mixed Martial Arts state that the priority in judging a round is to assess the "effective striking and grappling" (California State Athletic Commission, 2020). This includes "legal blows that have immediate or cumulative impact with the potential to contribute towards the end of the match" and the "successful execution of takedowns, submission attempts, reversals and the achievement of advantageous positions". Thus, our model uses the key components of judging available to us, noting that we do not include information on reversals or positional advances in our skill models.

Denote by AL_{ijk_r} the total number of strikes landed by i against j in round r of fight k. Now, define $\Delta\mathrm{AL}_{ijk_r} = \mathrm{AL}_{ijk_r} - \mathrm{AL}_{jik_r}$ as the difference in strikes. Similarly, denote by $\Delta\mathrm{TDL}_{ijk_r}$ and $\Delta\mathrm{SMA}_{ijk_r}$ the difference in takedowns landed and submissions attempted, respectively. Then our logit model for fighter i winning round k_r by any score is:

$$won_{ijk_r} \sim \mathrm{Bernoulli}(p_{ijk_r}), \tag{3.5}$$

$$\mathrm{logit}(p_{ijk_r}) = \beta_1 \Delta\mathrm{AL}_{ijk_r} + \beta_2 \Delta\mathrm{TDL}_{ijk_r} + \beta_3 \Delta\mathrm{SMA}_{ijk_r}. \tag{3.6}$$

We do not include an intercept in equation 3.6; this ensures two combatants with identical statistics will win a round with a probability of 0.5. The model summary and coefficients are shown in Table 3.2. All variables have a positive effect and are significant at the 1% level.

This model is then used to generate fight outcomes when a simulation reaches the time-limit and needs to be judged. According to the in-round statistics, three judge's decisions are simulated per round, with the overall winner of the simulation determined by the fighter who won on most of the judges' scorecards. Since the judging model predicts a binary win/lose variable and there is an odd number of both rounds and judges, draws cannot occur.

[6] Only bouts which required the judges verdicts are available on mmadecisions.com.

[7] We cannot include submissions landed since it implies a finish win. Submission attempts are mostly from a dominant position, and a sign that the opponent is in danger, so we would still expect them to have an impact on judging.

Table 3.2. Summary of the logit model fitted to judge's decisions over 2001-2017 using the differences of key in-round statistics.

	Dependent variable:
	Won round
ΔAL	0.172*** (0.004)
ΔTDL	0.838*** (0.028)
ΔSMA	0.547*** (0.052)
Observations	9,377
Log Likelihood	−3,983.719
Akaike Inf. Crit.	7,973.438
Note:	*p<0.1; **p<0.05; ***p<0.01

3.6 Benchmark models

To compare the predictive performance of our model, we estimate two benchmark models. The first is a Bradley-Terry model fitted in a Bayesian framework. The second is a logistic regression model using the difference between combatants in several cumulative statistics. A comparison of the performance of the three models is presented in Section 3.7.2.

3.6.1 Bradley-Terry model

The Bradley-Terry (BT) model states that the probability of i beating j, p_{ij}, is given as:

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j},$$

where π_i and π_j are the strengths to be estimated by the BT model.

Since the comparison graph of fighters is not fully connected, a finite maximum likelihood estimate of the BT model does not exist. Thus, we find the maximum a posteriori estimate (MAPE), as in Caron and Doucet (2010). To this end, a Gamma(a, b) prior is placed on each π_i (where a and b are shape and rate, respectively, such that the mean of the distribution is a/b), and the MAPE is found using an expectation maximisation algorithm through the `BradleyTerryScalable` package (Kaye and Firth, 2021a).

One assumption to note is that the BT model assumes transitivity, which may not be the case in MMA considering the variety in stylistic match-ups. Nonetheless, the BT model provides a useful comparative benchmark for our Markov model.

We tune the choice of a to maximise the predictive accuracy on an out-of-sample set, which will be explained further in Section 3.7, finding the optimal value to be $a = 1.40$. Given a total of K fighters, b is set to equal $aK - 1$ to improve the speed of convergence

(Kaye and Firth, 2021b). Following, we fit the BT model using data from 2001-2017. Table 3.3 displays the skill estimates of the top 10 fighters.

Table 3.3. Estimates of fighter's abilities in the Bayesian Bradley-Terry model using a gamma prior with shape $a = 1.40$, and UFC fights from 2001-2017.

Fighter	π	Fights	Wins	Losses	Draws	No-contests
Jon Jones	2.63	18	16	1	0	1
Georges St-Pierre	2.52	22	20	2	0	0
Demetrious Johnson	2.41	17	15	1	1	0
Conor McGregor	2.34	10	9	1	0	0
Daniel Cormier	2.32	10	8	1	0	1
Anderson Silva	2.30	22	17	4	0	1
Yoel Romero	2.30	9	8	1	0	0
Tony Ferguson	2.25	14	13	1	0	0
Khabib Nurmagomedov	2.23	9	9	0	0	0
Cain Velasquez	2.19	14	12	2	0	0

The fighters who populate Table 3.3 are as one would expect. Jon Jones, Georges St-Pierre, and Demetrious Johnson are always in conversations for the moniker of greatest of all time[8]. The same can be said of Anderson Silva and Khabib Nurmagomedov, who at this point were in opposite stages of their careers: Silva was ageing and declining, whilst Nurmagomedov was the newly crowned Lightweight champion and would go on to retire with an unprecedented undefeated record of 29 wins and 0 losses. The remaining athletes who populate the rankings are all elite MMA athletes and certainly at the time would be considered some of the best.

3.6.2 Logistic regression

The second benchmark model we fitted was a logistic regression model with a binary dependent variable indicating whether the fighter in question won the fight or not. We calculated several summary statistics for each athlete competing in a fight, using only bouts prior to the observation: strikes landed per second, striking accuracy, strikes absorbed per second, strike defence (percentage of opponent's strikes that don't land), takedowns landed per second, takedown accuracy, takedown defence, and submission attempts per second. Missing data could occur either when the fighter was debuting in the UFC, or when calculating an accuracy with zero attempts. In such instances, we imputed the data using the mean of the

[8] Jon Jones would have an even higher rating were it not for a controversial disqualification loss early in his career in which he dominated his opponent throughout the fight. This highlights one of the key points made in Section 3.1.2: rating systems such as a Bradley Terry model will not account for the context of results. To this day it remains the only fight Jones has lost.

statistic across all non-missing training observations. We then calculated the difference in each of these measures between opposing athletes, which were the final covariates used in the model.

When fitting the model, we centred and scaled the covariates to have mean 0 and variance 1. As in the judging model, we randomised which fighter we would use as an observation to avoid any unwanted biases and did not fit an intercept.

Upon fitting the model, only the differences in *strikes landed per second*, *strike defense*, *takedowns landed per second* and *takedown defence* were statistically significant. We chose to then include only the significant variables, Table 3.4 displays the summary statistics and estimated coefficients of the final logistic model.

Table 3.4. Summary of the logistic model fitted to the binary win variable using the difference in several statistics as covariates.

	Dependent variable:
	Fight won
Δ*Strikes landed per second*	0.1081*** (0.0348)
Δ*Takedowns attempted per second*	0.1766*** (0.0328)
Δ*Strike defense*	0.1446*** (0.0330)
Δ*Takedown defense*	0.1240*** (0.0328)
Observations	4,129
Log Likelihood	−2,819.0840
Akaike Inf. Crit.	5,646.1670
Note:	*p<0.1; **p<0.05; ***p<0.01

3.7 Results

As detailed in Sections 3.4.2, 3.4.3, and 3.6.1 there are two parameters within the Markov model and one in the Bradley-Terry model that need to be set. To tune these parameters, we initially fit our models using the data from 2001-2016, keeping fights in 2017 as the validation data. We find the parameters that maximise the accuracy of fight predictions in the validation data regarding who will win the bout.

We found the optimal choices for the Markov chain to be: $adj_{ko} = 0.4$ and $adj_{sm} = 2$. We tested numerous combinations of these hyper-parameters; however, an exhaustive search of the possible combinations is unfeasible due to the computational demands of obtaining the simulations. As mentioned in Section 3.6.1, we found the optimal value for the shape parameter in the gamma prior to be $a = 1.40$.

Having tuned the required parameters, we then fit our models using the fights from 2001-2017, keeping bouts in 2018 as our test data. This gives us 8,408 observations for training (two observations per fight) and 474 contests for testing. We remove two contests from the test set that the UFC later declared no-contests, as well as contests in which either fighter was debuting. This leaves us with a total of 327 fights for testing.

To obtain predictions of a fight, we simulate 10,000 chains. Transition probabilities are generated by sampling from the posterior distributions of the fitted skill models introduced in Section 3.3, thus propagating the uncertainty in the estimates of the coefficients into the predictions.

3.7.1 Comparisons with in-fight statistics

To compare the transition counts with empirical frequencies, we obtain expected values for the different numbers of transitions in each fight, calculated as the median number of transitions over all simulations[9]. Scatter-plots displaying pairs of observed and expected statistics for several in-fight statistics are displayed in Figure 3.4: strikes attempted and landed, takedowns attempted and landed, submissions attempted, and the total bout duration. We apply a “jitter” function to each observation, that is adding a small amount of random noise, to ensure each point is visible in the plot (since for instance, there are many observations in which there were zero submissions and we predicted zero submissions). The transparency of each point is proportional to the amount of training data we have on the competing athletes; thus, we can see if more data improves the predictions. Finally, a regression line through the points is displayed.

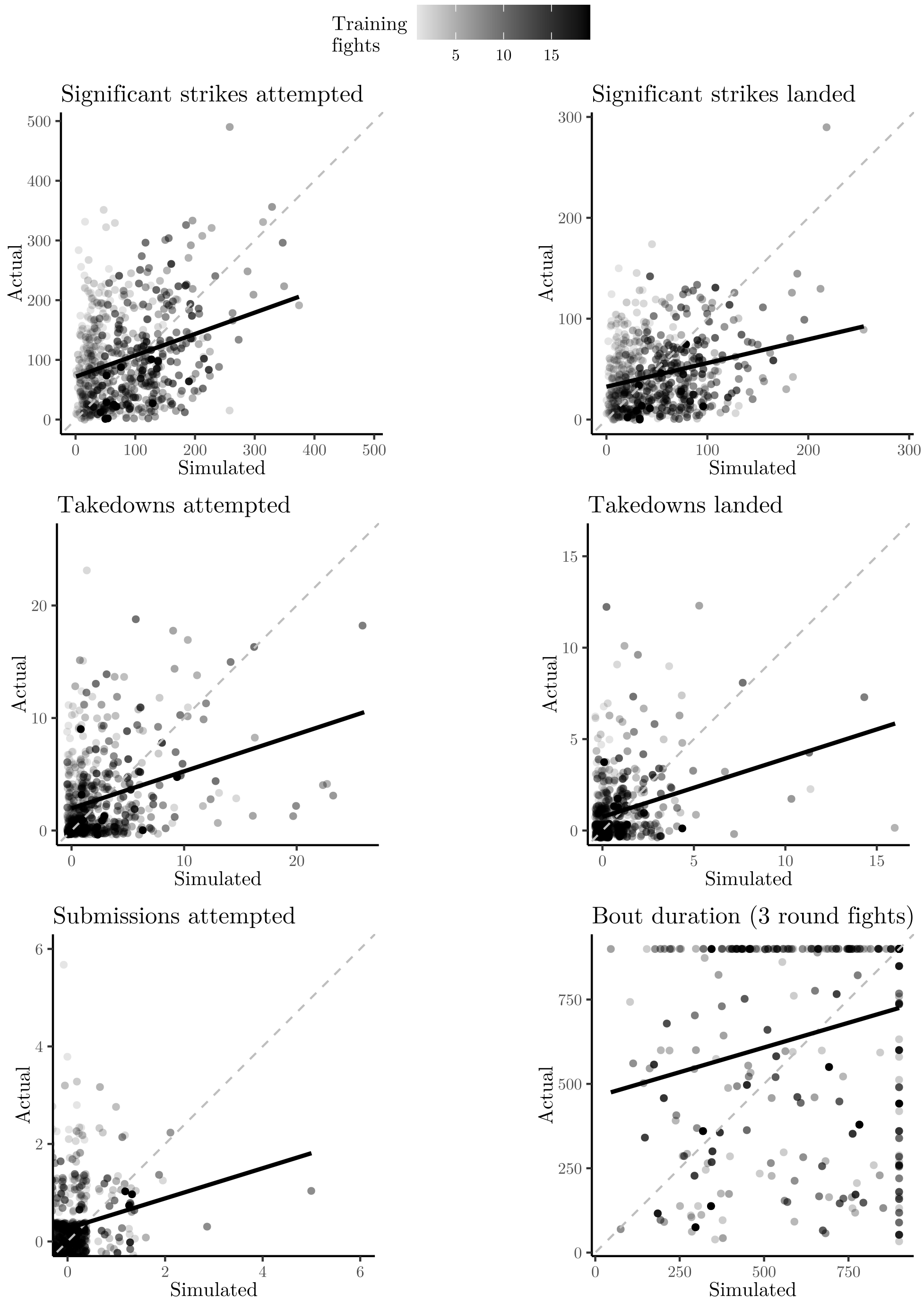
We observe a positive correlation across all statistics indicated by the regression line. Also, the distributions of each statistic match up well. We can conclude that our model is producing realistic simulations which capture individual fighting styles well. The poor predictions which lie well away from the bulk of observations appear to be fought by athletes with limited training data (indicated by the transparency). Given a fight can last anywhere between five[10] and 1500 seconds, and a fighter may be controlled and unable to attempt techniques for large portions, predicting the number of actions athletes will perform in MMA is far more challenging than the equivalent in, for instance, football, where one knows a match will last 90 minutes. Nonetheless, our model generates realistic simulations with a positive correlation across the expected statistics.

Table 3.5 shows the total number of fights our Markov model predicted to end via each

[9] We chose the median rather than the mean since some statistics are highly skewed, for instance, bout duration, where a majority of simulations for a given fight may end in a decision giving exactly 900 seconds.

[10] This is the UFC record held by Jorge Masvidal following his 2019 knockout of Ben Askren.

FIGURE 3.4. Scatterplot comparing the expected and observed values for several in-fight statistics using the averages across 10,000 simulations per fight. A regression through the points is calculated to display the correlation between values. The transparency of points is proportional to how many fights competing athletes have fought in within the training data. Small amounts of random noise have been added to each point to assist with visibility.



method (to either fighter), as well as the observed number. Again, our model is close to empirical data; obviously, we cannot predict a fight ending via disqualification.

Table 3.5. Comparison of the number of fights predicted to end in each possibility using the Markov model with the empirical frequencies.

Method	Actual	Predicted
Decision	162	167
Knockout	107	103
Submission	57	57
Disqualification	1	0

3.7.2 Comparison with benchmark models

We now compare the results predicted by the Markov model with those by the benchmark models: Bradley-Terry, and logistic regression. We compare the accuracy of each model in predicting the correct fighter to win a bout in the test data.

We found the Markov model to predict the correct fighter to win in 61.77% of fights. The Bradley-Terry model achieved 54.13% and logistic regression just 47.71%.

The Markov model is clearly superior to both benchmark models. The Bradley-Terry model performing about as well as a coin toss is interesting; this would imply the UFC are doing a good job of matchmaking: pitting fighters of equal strength against each other.

It is intriguing that although the logistic regression model is utilising similar variables as our Markov model (albeit in a much different way), the performance gap is so apparent. We believe this is because the cumulative statistics calculated in the regression are not accounting for the ability of past opponents. Our skill models, by containing attack and defence components, do account for past opponent strength.

Estimating the uncertainty in the predictive accuracy of the Markov model is hard for two reasons. First, uncertainty exists in our model from two sources: the estimated skill models and the Markov chain simulations. Second, the computational demands of fitting the skill models and obtaining simulations for each fight renders repeated experiments unfeasible.

We opted to obtain uncertainty estimates through a resampling strategy. As discussed in Section 3.4.1, we simulate each fight 10,000 times. We collect a sample (with repetition) of the simulations for each fight, calculating the probability of either fighter winning using these 10,000 resamples. The accuracy across the 327 fights is then calculated. We repeat this 100 times to obtain the uncertainty estimates which follow.

We found the mean accuracy across the resamples to be 61.62%, with a standard deviation of 0.53. The minimum and maximum were 59.94% and 63.00%, respectively. Finally, the lower and upper quartiles were 61.16% and 61.85%, respectively.

3.7.3 Comparison with the betting market

A paper on sports forecasting would not be complete without assessing the model's performance versus the betting market. Interest in betting markets is not solely to do with potential financial rewards but also has ramifications for findings on market efficiency. In the case of MMA, this is particularly interesting since, in comparison to other sports, the betting market on MMA is relatively young.

As described in Section 3.2, we scraped historical odds from BestFightOdds on the fight result (who will win the fight) and the result-method (who will win the fight and how). For brevity, we will refer to the result-method market as simply the "method" market.

We found an average over-round of 4.25% in the result market. The method market was significantly higher, with an over-round of 23.38%.

Our simulations allow us to determine the most probable outcome in each of these markets. It may be the case that we predict one fighter to be the most likely to win, but their opponent has the highest chance of winning by a particular method.

Table 3.6 compares the accuracies of our predictions in both markets over a range of different thresholds implying both fighters have had a minimum of t fights in the training data. Further, we include the "disagreement rate" between us and the bookmakers: that is, the percentage of predictions which differed to the bookmakers.

Table 3.6. Accuracies for both the Markov model and bookmakers odds in the result and method markets when filtering for a range of different thresholds ensuring each fighter has had a minimum number of fights in the training data, t. The disagreement rate (that is, the percentage of predictions which differed from the bookmakers) is also displayed.

		Result			*Method*		
t	n	Markov model	Bookmaker	Disagreement rate	Markov model	Bookmaker	Disagreement rate
1	327	61.77	61.16	40.98	38.84	32.72	63.61
2	260	61.54	61.92	39.62	37.69	33.85	63.08
3	207	60.39	57.97	40.10	34.78	33.33	62.80
4	174	60.34	58.62	40.80	33.33	33.33	62.07
5	143	57.34	58.74	41.96	34.27	34.97	60.14
6	118	60.17	56.78	40.68	35.59	30.51	56.78
7	92	58.70	57.61	40.22	34.78	28.26	58.70
8	78	58.97	55.13	42.31	33.33	24.36	64.10
9	53	56.60	56.60	37.74	32.08	18.87	64.15
10	36	61.11	61.11	38.89	33.33	16.67	61.11
13	13	69.23	76.92	38.46	38.46	30.77	61.54
16	7	71.43	71.43	57.14	42.86	14.29	85.71
19	5	80.00	80.00	40.00	60.00	20.00	80.00

Table 3.6 shows that our Markov model performs well in comparison to the bookmakers. In the result market, we perform comparably against the bookmaker even when there is potentially limited data on the athletes. We out-perform the bookmakers in the method market by large margins across the majority of the minimum fight thresholds.

Since bookmakers apply a margin to their odds, it is often not enough to have a model with a high predictive accuracy; the model must also have a low correlation with the odds (Hubáček et al., 2019). The disagreement rate in the two markets would suggest our model is adequately decalibrated from the bookmaker odds.

Our next investigation is to ascertain whether the model can be used as the basis of a profitable betting strategy. We assess four betting strategies: flat unit betting, expected value betting, fractional Kelly betting, and a modified version of Kelly betting presented in Boshnakov et al. (2017).

Flat unit staking is the most basic strategy. This consists of staking one unit on the selection deemed to be the most likely by the model, irrespective of the odds offered by the bookmaker and the bettor's estimation of the probabilities.

Given n possible outcomes for a bet, expected value betting implies the bettor places stakes equal to their estimate of the expected value for the bet. Given p_k is the bettor's estimate of the probability of selection k occurring, and o_k is the (decimal) odds offered by the bookmaker, the expected value is calculated to be $v_k = p_k o_k - 1$. In our implementation, we bet only on the selection with the largest positive expected value. We place no bets if no selections have a positive value.

Kelly betting is a well-known betting system that maximises the long-run log-utility of the investment, Kelly (1956). Solving the problem mathematically results in placing bets of size

$$f_k = \frac{p_k(o_k - 1) - q_k}{o_k - 1},$$

where $q_k = 1 - p_k$. The stake on a selection is then equal to the product of f_k and the bettor's current bankroll. Fractional Kelly betting is more often used in practice by bettors since previous authors have found that Kelly betting is overly risky. In a fractional Kelly strategy, the proportion of the bettor's bankroll to be staked is the Kelly stake f multiplied by a fixed fraction. Since the Kelly strategy is focused on long-term growth of one's bankroll, in the context of a small number of bets such as we have here, the results would depend greatly on the outcome of the last few bets. Hence, we omit its inclusion from our results.

In lieu of a fractional Kelly strategy, we test a modified Kelly, as presented in Boshnakov et al. (2017). We reset our bankroll to 1 unit before each bet and use the Kelly criterion to decide what fraction of our 1 unit is staked. An additional ‘protection’ is also introduced: we restrict ourselves to ‘quality bets’. For a potential wager, we place a bet if the expected value of that wager exceeds some threshold, v. In choosing the optimal value of v, we wanted to optimise for return on investment and still bet on a reasonable number of contests. We performed this tuning on the validation set, which we then applied to our results on the test set.

The results of the various betting strategies are shown in Table 3.7. Having found the optimal value in the modified Kelly strategy, we then test this same threshold with the flat and expected value stakes; to see if including only quality bets improves results. A final variation was to test flat stakes when only betting on selections with a positive expected value. This further test is not required with the other schemes, which limits the bettor to positive expected values by their design. A value in column v indicates we only stake on selections with value exceeding the given threshold.

Table 3.7. Summary of the results from several betting strategies using the Markov model to generate predictions for the results market. Values in column v indicate that we only bet when the selection has an expected value exceeding the given threshold. All results are based on filtering out fights in which either fighter was debuting in the UFC.

Strategy	v	Bets	Wins	Acc.	Stakes	Gross	Net	ROI
Flat		327	202	61.77	327.00	360.36	33.36	10.20
Flat	0.00	292	179	61.30	292.00	323.98	31.98	10.95
Flat	0.26	144	80	55.56	144.00	165.99	21.99	15.27
Expected value	0.00	292	140	47.95	121.19	134.05	12.86	10.61
Expected value	0.26	144	59	40.97	102.81	114.18	11.37	11.06
Modified Kelly	0.00	292	140	47.95	86.88	96.44	9.56	11.01
Modified Kelly	0.26	144	59	40.97	61.48	68.10	6.62	10.77

Table 3.7 shows that we can achieve positive returns with all betting schemes. Flat, expected value, and modified Kelly all perform comparably. We found the optimal threshold in the modified Kelly strategy to be 0.26. The greatest returns were achieved using flat stakes when only betting on selections with a minimum value of 0.26.

We now turn our attention to the method market. We investigate the same strategies as with the result market and present the results in Table 3.8. There were two fights we were unable to obtain the odds for the method market; hence, they are not included in the following results.

We achieve positive returns with the majority of the strategies. This time, expected value betting results in losses. Again, flat stakes with a minimum value threshold performs best.

Table 3.8 shows our optimal choice for v was 0.76 in the method market, higher than for

Table 3.8. Summary of the results from several betting strategies using the Markov model to generate predictions for the method market. Values in column v indicate that we only bet when the selection has an expected value exceeding the given threshold. All results are based on filtering out fights in which either fighter was debuting in the UFC.

Strategy	v	Bets	Wins	Acc.	Stakes	Gross	Net	ROI
Flat		325	127	39.08	325.00	411.06	86.06	26.48
Flat	0.00	323	125	38.70	323.00	406.63	83.63	25.89
Flat	0.76	180	62	34.44	180.00	234.61	54.61	30.34
Expected value	0.00	323	55	17.03	379.91	356.99	−22.91	−6.03
Expected value	0.76	180	22	12.22	313.02	292.17	−20.85	−6.66
Modified Kelly	0.00	323	70	21.67	66.00	73.25	7.25	10.98
Modified Kelly	0.76	180	27	15.00	42.47	47.17	4.70	11.07

the result market where $v = 0.26$. We believe this is due to the higher over-rounds of the method market: one needs higher-quality predictions in such markets.

Achieving such results in a difficult market to predict, with high over-rounds making it even more difficult to generate a profit, is a strong indication that our model has excellent predictive power.

3.8 Conclusions

The paper has presented a Markov Chain model for predicting the results of Mixed Martial Arts contests. Our approach first entails estimating the skills of athletes in various key fundamentals of the sport. These models generated transition probabilities that are used to simulate realistic MMA contests.

We developed a model for predicting the decisions of judges given the in-round statistics for opposing fighters. We implemented this judging model within our fight simulations to mimic how MMA contests are decided when they need to be assessed by the judges.

Forecasting MMA results is difficult for several reasons, not least the small numbers of fights each competitor takes part in. Our modelling approach is to drill down to MMA mechanics and model the quantity and quality of each action by fighters. In addition to performing well compared to benchmark models, our model can produce positive returns when used as part of a betting strategy.

Despite the clear success of our model, we see opportunities for several improvements, though the majority can only be implemented with a more detailed dataset. First, we have simplified the ground state of a fight and allowed for one ground position when there are numerous positions in reality, each having different advantages. Limitations in our dataset meant it was impossible to model all, or even some, of these ground positions. With the right data, one could model the likelihood of fighters advancing to more advantageous positions and allow different positions to have different transition probabilities. This would surely improve the predictive capabilities of the model.

Second, we have accounted for only one type of strike (although we allow for the strike to land in different areas of the opponent's anatomy). In reality, fighters can perform a strike with a knee, hand, or arm. Different striking techniques will have different probabilities of landing and different probabilities of a knockout. For instance, a successful knee strike is likely to inflict much more damage than a strike with the hand, but it is much harder to land a knee strike. Again, we feel this would improve the model, but more granular data are required.

Issues relating to cardio, damage, and time passing in a fight are areas we have not yet addressed. There are numerous questions to investigate: which fighters tire and struggle in later rounds; do body and leg strikes slow down an opponent; does knockout power fade over the course of a fight; will a fighter attempt more strikes if they lost the previous round. Again, this requires more detailed data.

Due to these models' large computation time and storage size, we chose only to update the skill models until the end of 2017 to predict all of 2018. When using this model in practice, one could update the skill estimates as much as possible to include all available fights. Updating the models throughout the test-set period would surely improve the results further.

There is potential for future work to investigate the use of different priors in the models. Currently, we only used the recommended weakly-informative scaled Cauchy prior for all skill models. However, one may find more informative priors to be useful. Whilst more informative priors would induce more regularisation, one could directly investigate the use of regularisation through ridge, lasso, or elastic-net regression.

The skill-estimation framework could be made more realistic in two manners: by allowing correlation between an athlete's skill parameters, and inducing some hierarchy between the models. Unfortunately, due to the size of the estimation problem, we were limited to simpler GLMs available through the `arm` package. However, future work could implement these structures using a smaller subset of the data, perhaps focusing on a handful of weight classes.

Despite our simplifications and potential areas for improvement, the model is a realistic representation of an MMA contest. It performs well in terms of accuracy of the predictions, the total counts of in-fight statistics, and when used as the basis for a betting strategy. MMA is a rapidly growing sport, and we hope that our model could be of use to several stakeholders, including bookmakers, bettors, media, and fans. There is even potential for

MMA athletes to utilise the model in their preparation for facing a particular opponent.

Chapter 4

Reputation Bias and Home Crowd Influence in Judging: The Case of Mixed Martial Arts

In Chapter 3, we used a simple judging model to assess the winner of a simulated bout based on just three counts: the differences in strikes landed, takedowns landed, and submissions attempted. In this paper, we expand our judging model and investigate the existence of different biases.

4.1 Introduction

Humans are subject to many biases in many areas of society. Psychologists have discovered and studied a plethora of cognitive biases such as confirmation bias (the tendency for individuals to recall information more readily when that information confirms a prior belief), or anchoring bias (the tendency to weight the first piece of information more highly than subsequent details) (Bunn, 1975). Discovering and detecting cognitive biases has become a popular research theme in several disciplines, including Economics, Operational Research, Statistics and Finance. In this paper, we use sport as our laboratory for detecting biases. Specifically, we look for evidence of bias in the judging of the top level of Mixed Martial Arts (MMA) competition, known as the Ultimate Fighting Championship (UFC).

Detecting biases in sports–particularly when judges directly influence the outcome–is important, as this strikes at the very heart of sports' integrity and fairness. Indeed, following scandals during the 1998 and 2002 Winter Olympics involving vote trading and collusion between judges of figure skating contests[1], a great emphasis has been placed on detecting biases in judging, and, subsequently, on improving judging and scoring mechanisms to minimise the likelihood of bias, or even cheating. Following scandals at the 1998 and 2002 Winter Olympics, new scoring procedures were implemented in 2004. However, in 2014, following further issues with anonymous scoring by judges, the scoring process changed again. Having fair and true scoring mechanisms in sports is a core tenet of operation if they are to maintain their integrity.

In the literature on bias in sports, several biases have been detected. Nationalistic bias (where judges favour athletes from their own country) has been found in figure skating (Campbell and Galbraith, 1996), gymnastics (Heiniger and Mercier, 2018), diving (Emerson et al., 2009), and muay Thai (Myers et al., 2006).

Racial bias in sport has been widely investigated in various contexts. Own-race bias was found to be exhibited by basketball referees in Price and Wolfers (2010), such that officials penalise players of the opposite race to them (assuming race to be either black or white); the study did not distinguish the race of the referee themselves. Similarly, umpires in lower levels of baseball were found to be less likely to call a strike when their race did not match that of the pitcher (Parsons et al., 2011).

Whilst home advantage is an accepted fact in almost all sports, the mechanism by which it occurs is less well understood. One explanation is that referees suffer from a bias in that they succumb to social pressure exerted by the crowd. Indeed, evidence for this has been found by, for example, Buraimo et al. (2010), Sutter and Kocher (2004), and Garicano et al. (2005), by examining the distribution of cards given to the home team versus the away team or how much time will be added on depending on the scoreline. The absence of crowds throughout the Covid-19 pandemic and during the 2007 Italian football season, enabled further research confirming the theory that crowd pressure causes, in some parts, home advantage. Pettersson-Lidbom and Priks (2010) found that in the presence of a crowd, referees penalised the away team more than the home team; however, the opposite was true for behind-closed-doors matches. Reade et al. (2020) found that in the absence of a crowd, referees penalised the away team significantly less; however, they found no significant change towards the home team.

An experiment by Myers et al. (2012) saw ten experienced muay Thai judges score a fight in two different conditions: with and without crowd noise. The authors found the judges scored significantly more strikes when in the presence of noise. Home advantage was also observed, as the judges would score significantly more strikes to the home fighter in the

[1] see, for example, https://en.wikipedia.org/wiki/2002_Winter_Olympics_figure_skating_scandal

presence of noise.

Balmer et al. (2005) found that judges in boxing favoured home fighters. Having controlled for the athletes' relative skills, it was found that when a fight requires the judges' verdicts, the home fighter won significantly more often than in fights which ended prematurely via knockout.

Reputation bias (when judges favour known athletes or those expected to perform well) has been detected in figure skating in Findlay and Ste-Marie (2004). When a series of athletes are to perform in a competition, coaches typically order them such that the best perform last: Plessner (1999) found a reputation bias such that judges overly favoured those performing later in the event. In football, a team's prior reputation for aggression was found to significantly affect the likelihood of a referee penalising a team for a foul in Jones and Erskine (2003).

In the aftermath of uncovering biases (or indeed collusion) within sports, it is common for there to be widespread changes designed to combat their existence. As discussed above, judging of figure skating went through a complete reform–which, in fact, was found to further enable some aspects of the colluding (Zitzewitz, 2014). Technological solutions have been implemented to minimise bias/corruption in boxing and taekwondo following separate controversies. Judges in amateur boxing now must report points or warnings using an electronic button system. To be officially recorded, three or more judges must then "agree" with the verdict within a short time window. Perhaps taekwondo has had the most advanced solutions: adding electronic sensors into protective gear, implementing electronic scoring as in boxing, and not allowing judges of the same nationality as a combatant.

In this paper, we investigate the presence of bias in how judges score contests in the highest level of MMA competition, the UFC. The UFC has grown enormously in popularity in recent years. The latest broadcasting rights contract was signed in 2018 between the UFC and ESPN to air 30 events over five years and was worth a reported USD$1.5bn. To put this into perspective and to demonstrate the size of the UFC, the largest television rights contract for soccer is for the English Premier League, which agreed a deal from 2019 to 2022 to show 200 games per season for an estimated GBP£5bn. That the broadcasting rights deal is of the same order of magnitude as soccer's top league indicates just how popular the UFC has become amongst sports fans.

The paper is organised as follows. Section 4.2 details the system and scoring criteria implemented by MMA judges before providing an overview of the scientific literature on MMA. The data used throughout the paper is introduced in Section 4.3, and descriptions of the independent variables are given in Section 4.4. Section 4.5 introduces the model and the purposeful selection methodology we use. Finally, results from our model and discussion on the implied biases are contained in Sections 4.6 and 4.7, respectively.

4.2 Introduction to MMA and the Literature on Judging

The Unified Rules of Mixed Martial Arts (California State Athletic Commission, 2020) were originally set in 2001 in an effort to protect fighters, whilst also legitimising MMA as a sporting spectacle. Unlike boxing, fighters can combine punching with ways to strike the opponent with their arms, legs and feet, and even wrestle and grapple.

In the UFC, most fights are contested over three five-minute rounds. Title fights and main-events are scheduled for five rounds.

Around 50% of contests end before the scheduled number of rounds have been completed because one fighter has won via a knockout, submission, or disqualification[2]. For contests that are not ended, the scores of three judges determine the bout's winner. A referee monitors the action within the ring: it is their job to call fouls, deduct points for fouls, or end the fight if they deem one athlete unable to continue.

The judges score each round of the fight, and award a score of 10 to the winner of the round, their opponent then receives 10 points or less. A score of 10-10 implies a tie, and occurs very rarely in practice. By far, the most common scoreline is 10-9. This can be quite a broad score, and may be given in extremely close rounds when one fighter lands one strike more than their opponent, or very obvious rounds when one fighter clearly outclassed their opponent.

A score of 10-8 is less common and suggests either a dominant performance of one fighter, or that a fighter has been penalised for lack of action. The final plausible scoreline is 10-7. This is only given when an opponent has been completely overwhelmed, such that the judges consider the fight could be stopped. In our dataset (see below), 0.08% of rounds were scored 10-10, 95.06% were scored 10-9, 4.52% were scored 10-8, and 0.03% were scored 10-7. A further 0.30% were scored as 9-9; such rounds are only possible when the referee deducted a point for a foul.

To decide the winner of the fight, each judge's round scores are summed, with the winner according to each individual judge being the fighter with the most points. A majority verdict across the three judges is then used to identify the winner. There are several possible outcomes of the fight, as detailed in Table 4.1.

Although the UFC is now extremely popular, it is still a relatively new phenomenon, and the research community is only just starting to study MMA. Despite being such a young field

[2] A fight may also end early through a "no-contest". In this rare situation, a fight is essentially voided, the most common reason being an accidental foul rendering a fighter unable to continue.

Table 4.1. Different decisions which can be given based on the verdicts of the individual judges. The fight is between two fighters: Blue and Red.

Judges' overall winner			Blue	Red	Draw	Result	Decision
Blue	Blue	Blue	3	0	0	Blue	Unanimous win
Blue	Blue	Draw	2	0	1	Blue	Majority win
Blue	Red	Red	1	2	0	Red	Split win
Blue	Red	Draw	1	1	1	Draw	Split draw
Draw	Draw	Red	1	0	2	Draw	Majority draw
Draw	Draw	Draw	0	0	3	Draw	Unanimous draw

of research, there already exist three papers investigating the judges. This perhaps indicates just how critical a role the judges play in the sport.

Collier et al. (2012) and Feldman (2020) looked at how in-fight statistics (such as the number of head strikes landed) influenced the judges' verdicts. The studies found that knockdowns (strikes by a fighter that led to the opponent falling to the floor) are the most influential[3]. However, there are limitations to both studies. Although judges score each round of a fight, Collier et al. (2012) used statistics over the duration of a fight to model the overall fight outcome; whilst Feldman (2020) only used a small subset of the variables available to them.

Gift (2018) used in-round statistics amongst other variables potentially indicative of biases to estimate the probability of a fighter winning a particular round in MMA. The odds of a fighter winning were found to have a significant effect which the authors suggested indicated reputation bias. Significant effects when a fighter won the previous round, or had an insurmountable lead, were also found, which the author concluded showed further biases. However, robust conclusions from the Gift (2018) paper cannot be made. The in-round statistics cannot fully capture what happened during a round. Consequently, any significant effect of the odds in a model estimating the judges' scores is likely to be informative of the athletes' relative skills, rather than a reputation bias. The same is true for the effects of winning the previous round and having an insurmountable lead: they are likely to be indicators of how well an athlete was performing on the night.

In this paper, we build on the existing MMA judging literature, and investigate the existence of two biases: reputation bias and home athlete bias.

[3] This is in-line with what one would expect since knockdowns are relatively rare actions, and often lead to the end of the fight.

4.3 Data

Assembling a dataset for this study was non-trivial and required merging data from several sources for each of six ‘families’ of data. First, scores awarded by judges in each round of UFC fights were obtained from mmadecisions.com. Only fights which ended via a decision are available, limiting the dataset to fights where the judges’ scores were used to decide on the outcome.

Second, round-by-round fight statistics of all UFC fights were scraped from ufcstats.com. The in-round statistics covered a variety of actions.

- ‘Strikes’: various types of strikes were included, such as significant[4] head, body, or leg strikes landed or missed; non-significant strikes landed or missed. There is a lower level of detail available for non-significant strikes. This is not surprising since they are often disregarded entirely in post-fight analysis. Further, there are generally fewer non-significant strikes in a fight, and they will likely weigh less to the judges. We note that non-significant and significant strikes are exclusive categories.

- ‘Takedowns’ are grappling techniques used to take an opponent to the ground, split by those landed or missed in the data.

- ‘Control time’ is how long a fighter was in a dominating grappling position; fighters on “top” are said to be in control of those on the “bottom”.

- ‘Submission attempts’ are techniques which involve various joint locks or chokes. These techniques finish the fight if successful, as the opponent is forced to concede to avoid passing out or damaging their joints.

- ‘Reversals’ are techniques used to take a fighter from being controlled to being in control of a grappling exchange.

- ‘Knockdowns’ are strikes which cause an opponent to fall to the ground. Like submission attempts, knockdowns often lead to the end of the fight as the opponent is dazed and thus vulnerable to subsequent attacks.

Our third ‘family’ of data was information on the fighters. This was scraped from ufcstats.com with information such as height, reach, and date of birth.

The fourth family of data on historical bookmaker odds was collected from bestfightodds.com. These data consist of the closing odds from several bookmakers, including William Hill, DraftKings, and Unibet. We obtained the odds in the result market (that is, the odds

[4] A ‘significant’ strike is one that is deemed to have been of an adequate amount of power.

of each fighter winning) and the result-method market (the odds of each fighter winning by each possible method, i.e. fighter A/B winning by decision/knockout/submission).

The fifth family of data we collected was the rankings of fighters. Since 04/02/2013, the UFC has maintained official rankings, which we obtained from historicalufcrankings.com. Following each event, the rankings within each weight class are updated using the results of a poll of a select panel of media members.

Finally, the sixth family of data was obtained from wikipedia.org/wiki/List_of_UFC_events and gave the attendances at each UFC event. The record attendance was 57,127 for UFC 243: Whittaker vs. Adesanya in 2019. During the Covid-19 pandemic, there were 51 events (281 fights) with no fans. This provides us with a unique opportunity to assess whether the crowd influences the judges, and specifically, whether judges favour fighters competing in their home country. We note there were 24 events (109 fights) where the UFC did not release the attendance. We omitted these events from the analysis since the attendance figures are a crucial aspect of the investigation.

Once all six families of data were merged, the final dataset amassed a total of 17,105 unique judge's scores over 5,800 rounds in 1,840 fights spanning from 16/02/2013 to 18/06/2022. This included 309 unique judges who scored a median of 12 (mean of 55.36) rounds, with a minimum and a maximum number of rounds of 3 and 1,573, respectively.

There were 38 rounds in which a fighter was deducted one point, and three rounds where a fighter was deducted two points. Since these deductions were applied at the referee's discretion, we chose to model the "adjusted" score of the judges and unapply these deductions. Of these adjusted scores, 16,359 (95.64%) rounds were scored as 10-9. There were 728 (4.26%) scored as 10-8, just two (0.01%) scored as 10-7, and finally, 16 (0.09%) draws

Our objective is to use these data to investigate the presence of two biases: home bias, and reputation bias. In the next section, we describe the variables used in our study.

4.4 Variables

The focus of our paper is the score awarded by a judge to a round. Given the distribution of scores is over 95% of rounds scored as 10-9, and just 0.09% scored as draws, we choose to simplify the modelling framework and model which fighter was deemed to have won the round (by a score of either 10-9, 10-8 or 10-7). Consequently, we omit the tied rounds from our analysis.

The variables used to predict which fighter won a round are as follows:

- *In-round statistics*: knockdowns, significant head/body/leg strikes landed/missed, takedowns landed/missed, reversals, control-time, submission attempts, non-significant strikes landed/missed[5].

- *Fighter ability*: bookmaker implied probability of winning (Win IP), and winning by decision (Decision-win IP), knockout (Knockout-win IP), or submission (Submission-win IP).

- *Fighter information*: age, height, reach, official ranking, whether the fighter is competing in their home country, and whether the fighter is the weight category champion. 'Stance' describes how the fighter stands in a fight. The different categories of fighter stance are: left foot forward (orthodox, 77.09% of fighters), right foot forward (southpaw, 18.21% of fighters), a mix of the two (switch, 4.55% of fighters), or neither (open, 0.15% of fighters). We used a single binary variable, 'orthodox stance', to assess how judges interpreted different styles.

- *Crowd information*: we include an indicator variable representing whether a live audience was present, and the actual attendance. Further, we include the interaction of these terms with the home fighter variable.

In many sports using judges, athletes compete independently of one another such that a judge's score for one athlete should be independent of the score awarded to another. This is the case in, for example, gymnastics and diving, where the athletes do not interact. This is not the case in MMA and the UFC: two athletes interact, and the identity of the winner of a round is intrinsically connected. Consequently, we model the differences between contestants' performances in a given round, as measured by the variables described above, rather than the 'raw' counts.

This issue also applies to the binary variables: home country, champion, and non-orthodox stance. For example, if either both or neither athlete are in their home country, it is a neutral venue, and the variable takes 0. In the case of a home and away fighter, the variable takes 1 and -1, respectively.

The bookmaker odds are included in the model to control for fighter skill which the in-fight statistics cannot capture. There are two sets of odds available: result odds (which fighter wins the fight?), and result-method odds (which fighter wins the bout and by which method (a decision, a knockout, or a submission)?). We remove the bookmaker's margin and include the odds as implied probabilities.

We investigate the potential of reputation bias amongst the judges by including the difference in the fighters' rank according to the official UFC rankings. Each weightclass has

[5] The non-significant strikes in the data are not as granular as the significant counter-parts and only split by whether they landed or not.

separate rankings with a cut-off (usually 15 fighters) such that anyone outside the cut-off is classed as ‘unranked’. We assign unranked fighters a rank one more than the weightclass maximum. For instance, if there are 15 numbered ranks in a weightclass, all unranked fighters are assigned rank 16. It is convention in the UFC that the current champion is ranked at rank 0 (the other ranked fighters are effectively challengers to the champion). To make interpretations simpler, we reverse the order of the rankings, so that higher rankings have larger numerical values.

Further, we explore the effect of being the champion on the judges’ decision-making using an indicator variable describing whether an athlete is fighting the champion (-1), whether neither fighter is the champion (0), or whether the fighter is the champion (1).

To assess whether the crowd influence the judges, we calculate the interaction term between the home country variable and the crowd indicator. We create a further term using the interaction with the crowd size to see if the impact increases with the size of the crowd.

4.5 Model

Modern-day machine learning algorithms allow users to automatically fit models which implement feature reduction (variable selection), create interaction terms, and identify optimal shapes of the relationships between features and the dependent variable. However, these algorithms come with a cost, chiefly difficult implementations and interpretations. In a study such as this, where we aim to inform stakeholders of potential biases with a view to reforms, the interpretation of the final model is crucial. Nevertheless, the machine learning approach has many positives, such as objective modelling choices determined by predictive accuracy, and model parsimony.

We aim to address the gap in the literature between traditional methods of identifying model specifications and machine learning algorithms, by implementing a technique first introduced in Hosmer and Lemeshow (2000) known as “purposeful variable selection”. This methodology combines the flexibility of machine learning algorithms with the simplicity of interpretation of logistic regression. The methodology is presented in detail in Appendix B.1.

To summarise here, first, each variable is screened individually using a univariate logistic regression and kept if their p-value is below 0.25; this wide threshold for inclusion ensures that all variables of note remain in the model to begin with. A preliminary model is fitted, composed of all variables selected through the initial screening. Variables are then sequentially removed if the model is not significantly affected (using likelihood ratio tests), and they are not a ‘confounding’ variable (a variable which is required to adjust for the effect of another variable). The optimal shape of each variable is then assessed. The recommended method is to use fractional polynomials; one must test whether the inclusion of power terms significantly improves the model fit. Plausible interaction terms are then created and assessed for inclusion through sequential likelihood ratio tests. Finally, the adequacy and fit of the model should be ensured before making any inferences.

For the models presented herein, the set of variables selected, the shapes of the relationship between these variables and the dependent variable, and the inclusion of interaction terms, were all established through *purposeful selection*.

We use logistic regression to model the probability of a fighter being judged as the winner of a round. To ensure fighters with differences of zero across all variables are estimated to win a round with probability 0.5, we do not fit an intercept in the model. Since each variable is calculated as a difference between the two fighters, we keep only one observation per round, randomising whether the observation will be from the winner's or loser's perspective.

4.6 Results

Table 4.2 displays the fitted model. The table includes the coefficient estimates, p-values and average marginal effects (AME).

Looking at the average marginal effects given in Table 4.2, the most influential fighter action is a knockdown. This is not surprising since knockdowns are rare events that often lead to the end of the fight and are thus known to be crucial to the outcome of a fight. It is therefore reasonable that judges value these highly in determining the winner of a round. Submissions are a clear second, which also often lead to the end of a fight.

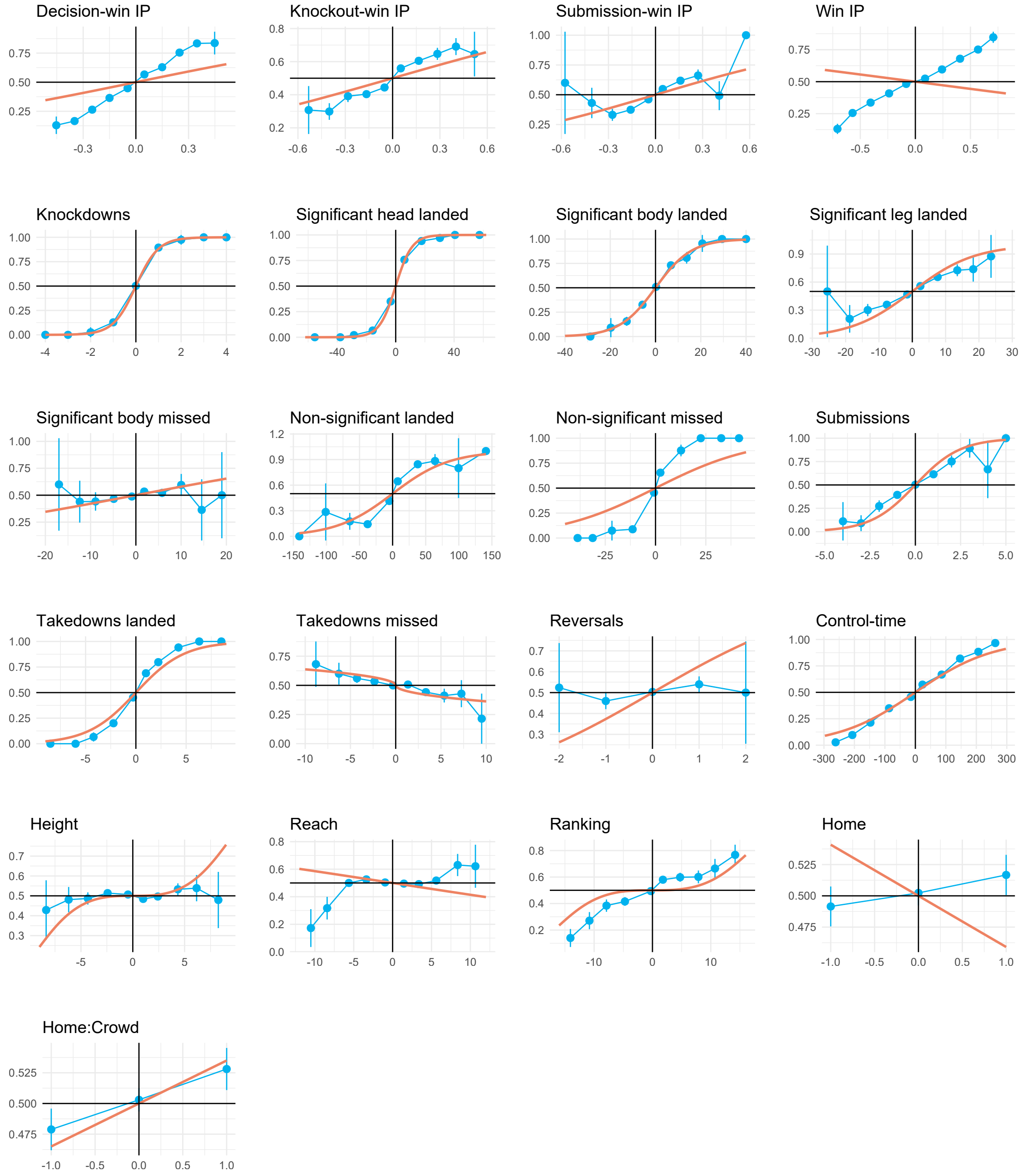
Through the use of purposeful selection, we have identified several interesting interactions. We find that if a fighter has a higher probability of winning via submission (from the odds), the judges value their control-time more. This is likely since that athlete will be more skilled in grappling, and consequently, there is a greater threat of submission from dominant grappling positions. We will discuss the interaction between knockdowns and rankings in Section 4.7.3. Recall that interaction terms were identified and included in the model solely through the use of purposeful selection, and would not have been identified otherwise.

Having used purposeful selection to identify the shapes of the relationships between the dependent variable (win probability) and the covariates, it is useful to examine what exactly those shapes are. Figure 4.1 displays the effect of each covariate on the probability of winning a round at all observed values (holding all other covariates equal to zero). Also displayed are the mean win percentage and value of the covariate once split into several bins (shown as blue dots/line).

Table 4.2. Final logistic regression model predicting the winner of a round given various in-round statistics combined with other informative covariates. AME shows the average marginal effect for each variable. Note that all variables are differences between the counts of the two opponents. Variables were selected through purposeful selection.

Term	AME	Estimate	p-value
Knockdowns	0.2382	2.0059	0.0000***
Submission-win IP	0.1863	1.5683	0.0034**
Decision-win IP	0.1476	1.2427	0.0279*
Knockout-win IP	0.1299	1.0941	0.0442*
Submissions	0.0958	0.8063	0.0000***
Reversals	0.0615	0.5176	0.0000***
Takedowns landed	0.0492	0.4144	0.0000***
Home $*$ Crowd	0.0361	0.3041	0.0052**
Reversals $*$ Home	0.0295	0.2485	0.0391*
Significant head landed	0.0214	0.1801	0.0000***
Significant body landed	0.0149	0.1254	0.0000***
Significant leg landed	0.0127	0.1067	0.0000***
Non-significant missed	0.0048	0.0402	0.0010**
Significant body missed	0.0038	0.0319	0.0002***
Non-significant landed	0.0028	0.0232	0.0000***
Control-time	0.0009	0.0078	0.0000***
Control-time $*$ Submission-win IP	0.0008	0.0068	0.0047**
Height^3	0.0002	0.0016	0.0000***
Knockdowns $*$ Ranking^3	0.0001	0.0012	0.0129*
Ranking^3	0.0000	0.0003	0.0000***
Reach	−0.0041	−0.0347	0.0000***
Home	−0.0195	−0.1639	0.1102
Takedowns missed$^{1/2}$	−0.0212	−0.1785	0.0000***
Win IP	−0.0536	−0.4514	0.3295
Observations		17089	
Log likelihood		−6308.57	
AIC		12665.14	
Note:	*p<0.05; **p<0.01; ***p<0.001		

Figure 4.1. Plots displaying the relationship between each covariate and the win probability. Blue points and lines indicate the observed averages for the binned data. Red curves display the effect of each variable holding all other covariates equal to zero.



We see that, for the most part, our coefficients are as one would expect. The observed values and win probability increase as, for example, *Significant body landed* (significant strikes landed on the body) increases. However, there are two exceptions: *Win IP* and *Reach*. We believe that in both instances, it is due to high positive correlations with other variables. *Win IP* is correlated with the other bookmaker implied probabilities (*Decision-win IP*, *Knockout-win IP*, and *Submission-win IP*); whilst *Reach* is correlated with *Height*

Before moving to the paper's main focus and examining the existence of biases in the judges' decision-making, we comment on the use of purposeful selection for model specification. We compare the model's fit to that of a 'naïve' model in which all original variables, with no interaction terms or shape adjustments, were included. The AIC and AUC scores were 12665.14 and 12718.44, and 0.9151 and 0.9144, for the purposeful and naïve models, respectively. Thus, in both performance measures, the purposeful model outperforms the naïve.

4.7 Evidence of Bias

We now discuss the findings from the model presented in Section 4.6 concerning the paper's primary objective: assessing the existence of home fighter bias and reputation bias in judging. However, before we can make any conclusions on biases, we first establish whether variables in the model are merely indicators of skill. This is done by examining the betting market's efficiency under our final model.

4.7.1 Market Efficiency and Accounting for Unobserved Fighter Skills

Gift (2018) was unable to definitively conclude whether the significant variables were indicative of bias or whether they represented some unaccounted fighter skills. We first examine whether the betting market is efficient to ensure we can make more robust conclusions from our model. If the market is efficient, then only the bookmaker odds will be significant in a model to predict the winner of each fight prior to the contest, and our other variables, such as home advantage, height, reach, and the difference in ranking, will fail to attract statistical significance, as that information has already been accounted for in the betting odds. This is an important issue to examine as the betting market on the UFC is still relatively young and has never been tested for efficiency. Indeed, it should arguably be the focus of a research paper on its own.

To determine whether the market is efficient, we fit a logistic regression using all variables that would be available prior to a fight. For fairness, we have included variables within the function they appear in in the judging model (Table 4.2). The odds are included as 'implied probabilities' with the bookmaker overround removed. The fitted model is displayed in Table 4.3.

Table 4.3. Logistic regression model estimating the probability of a fighter winning a contest given the variables that would be available before the fight selected through the purposeful selection methodology. AME shows the average marginal effect for each variable.

Term	AME	Estimate	p-value
Decision-win IP	0.8705	4.1257	0.0009***
Submission-win IP	0.1812	0.8589	0.4683
Knockout-win IP	0.1259	0.5967	0.6179
Win IP	0.1043	0.4942	0.6263
Home $*$ Crowd	0.0680	0.3221	0.1855
Height	0.0015	0.0072	0.8387
Reach	0.0004	0.0020	0.9222
Ranking^3	0.0000	0.0002	0.2588
Height^3	0.0000	-0.0001	0.9601
Home	-0.0455	-0.2154	0.3470
Observations		1805	
Log likelihood		-1100.86	
AIC		2221.72	
Observations		3792	
Log likelihood		-2302.75	
AIC		4625.50	
Note:	*p<0.05; **p<0.01; ***p<0.001		

We see that the betting market is efficient as the only significant effect is the implied probability of a decision-win. In fact, the other three implied probabilities (win, knockout-win, and submission-win) are non-significant themselves[6].

The official rankings, home fighter, and the interaction between the home fighter and a live audience are all non-significant. Having now established that these variables add no further information on fighter skills beyond what is contained in the odds, we can make more robust conclusions concerning biases in the judging model.

We now examine each potential source of bias.

[6] We believe this is due to the large correlation between the result and result-method odds. Indeed, in a model using only the result odds, *Win IP* is statistically significant; whilst in a model using only the result-method odds, *Decision-win IP*, *Knockout-win IP*, *Submission-win IP* are all significant

4.7.2 Home advantage

The two variables of interest in Table 4.2 when assessing whether the presence of a crowd influences the judges' decision-making are the *Home* main effect and $Home * Crowd$ interaction. We see that in the absence of a crowd, there is no statistically significant effect. However, there is a statistically significant positive effect when the event has a crowd.

The interaction terms including the size of the audience have been dropped during the purposeful selection process. This suggests that it is not the size of the crowd which matters, merely the presence of one.

Crowds in MMA are unlike those in football. Unlike football, there are frequently no 'away' fans at all. A handful of 'super-stars' could expect a small number of fans to travel to another country to see them perform. As such, the crowd will likely side with a home fighter, and each action will be met with cheers, whilst each action of the away fighter will be met with derision and boos.

The average marginal effect of being a home fighter in the presence of a live audience is estimated to be 0.0361. Thus, holding all other variables equal, such a fighter receives a 3.61% increase in their win probability.

4.7.3 Reputation bias

Given the lack of statistical significance of the ranking variable in the pre-fight prediction model (see Table 4.3), that the ranking variable is statistically significant and positive in the judges' model is evidence of a reputation bias. The bias is such that fighters with a better reputation (and are thus ranked higher than their opponent) are favoured by the judges, even having controlled for what actions each fighter performed in the round. Given the AME of Ranking[3], the size of the reputation bias is such that a fighter ranked ten places higher than an opponent has a probability of 3.43% higher of winning a round than if there were no reputation bias.

We also identified a significant interaction between knockdowns and higher-ranked fighters. The AME of this interaction implies a knockdown by a fighter ranked ten places higher than their opponent is awarded an additional 14.1%.

This is in line with past literature investigating reputation biases in sport discussed in Section 4.1: those with better reputations are overly favoured. A further example of reputation bias is football referees awarding more penalties to successful teams (Erikstad and Johansen, 2020).

A long-standing cliché in combat sports discourse is that "you have to beat the champ to be the champ". This most commonly alludes to the belief that the judge will favour the champion in close rounds. Since the champion variable was dropped during the fitting process, we can conclude there is no significant *additional* reputation effect of being the champion.

4.8 Conclusions

By collecting a large dataset on MMA scores and including novel variables, we have established that judges exhibit bias towards home fighters, and bias towards fighters with a higher reputation, when determining the winner of fights in the UFC. A major unaddressed problem with past research was determining whether a significant variable was indicative of bias, or merely fighters' skills. We controlled for fighters' differing skill sets using the bookmaker implied probabilities, thus accounting for skills unseen by the in-round statistics. By showing that the bias variables are captured within these implied probabilities, we can definitively conclude the existence of two biases.

First, we found evidence that judges favour home fighters when there is a live audience. In the absence of fans, we found no significant home advantage. This dovetails with the previous work on home advantage in football, where authors have found that the crowd significantly influences the presence and size of a home advantage, as discussed in Section 4.1.

Second, we found evidence of reputation bias. Controlling for fighter skill and the actions occurring in a round, a higher ranked fighter is more likely to be judged to have won the round. In addition, we have identified a bias such that knockdowns by higher-ranked fighters are overly favoured. Despite the consensus that there is a bias towards champions, we found no additional effect. However, finding a significant effect may be hindered by the small number of title fights within the dataset.

Our findings can help implement judging reforms to limit these biases, as has been done in other combat sports. Having found an influence from the home crowd, perhaps judges could wear headphones to limit their exposure to noise. In football, the video assistant referees are situated in a booth completely removed from the stadium–perhaps a similar idea could be used to reduce the effect of a crowd on judges.

Eliminating reputation bias may prove more problematic since the officials will likely follow the UFC closely and know most or all fighters. Even if the fighters are unknown to the judges, the rankings are displayed throughout an event build-up and during the television broadcast. As such, judges are highly likely to learn of the rankings of the two fighters (if they did not already know them). Nevertheless, knowing the bias exists and informing the judges may help defend against it. Specific training could be set up to teach the judges about this implicit bias. For instance, 21% of individuals on a medical admissions committee stated that awareness of their implicit biases influenced their admission decisions (Capers, 2020). However, one must be careful that decision-makers do not overcompensate for their biases.

Chapter 5

Individual Preferences and Controversial Decisions in Mixed Martial Arts Judges

In Chapter 4, we presented a model estimating the likelihood an athlete would win a round (by any score) based on the different in-round statistics and other variables, some possibly indicative of bias. In this chapter, we build on the previous work, using Bayesian hierarchical models to investigate the decision-making of judges of MMA contests at an individual level.

5.1 Introduction

Every day humans make decisions and judgements, be they conscious or subconscious, which for the most part, are inconsequential. However, every weekend sporting officials across the globe determine the fortunes of individual athletes, teams, stakeholders, and loyal supporters, sometimes resulting in controversy. Accurate and fair officiating is of paramount importance to the integrity of sport, and it is for this reason that public scrutiny of fairness within sport is a pursuit worthy of study.

Whilst referees within sports such as football and ice-hockey, or umpires within cricket and tennis, can indirectly influence outcomes, sports in which judges decide the final result are particularly vulnerable to spurious outcomes. Perhaps out of all these sports, mixed martial arts (MMA) judges face the most challenging task.

Many other sports have features making the judges' work more straightforward. For example, athletes compete independently in numerous Olympic sports, such as figure skating. Or in some judged sports, athletes have a set number of attempts to perform their best techniques, for instance, ski jumping. Finally, it is often the case that judges are assessing techniques which are the same, or at least very similar, and simple to rank. For example, in diving, a dive consisting of two somersaults is deemed better than a dive consisting of one;

or in boxing, a judge only has to assess one type of offensive action: a punch.

However, MMA judges do not have things so simple. Fighters compete against one another, meaning judges have to assess the performance of two athletes simultaneously. Bouts can be 25 minutes long, resulting in hundreds of actions to assess, which can occur at any given second. Typically these actions are of many different types (e.g. punches, kicks, or throws). MMA comprises the full spectrum of martial arts, allowing athletes to implement strikes from sports such as boxing or taekwondo, throws from judo and wrestling, or chokes and joint-locks from Brazilian jiu-jitsu. These techniques are not always similar in their function, and can have varying degrees of impact, so it is often unclear how to score each. For instance, how does one score a punch aiming for the opponent's head that is partially blocked, versus a kick that lands flush on the opponent's leg?

Not only is their job extremely difficult, but they are under some of the most intense scrutiny of any judges. We believe two reasons contribute to this scrutiny: the sport's popularity, and the consequences of winning and losing. First, compared to other subjectively judged sports, other than boxing, there is a much larger audience. The largest live crowd attendance for the top tier of MMA, the UFC (Ultimate Fighting Championship), was 57,127 fans in 2019 for UFC 243: Whittaker vs. Adesanya[1]. The largest pay-per-view event saw 2,400,000 buys in 2018 for UFC 229: Khabib vs. McGregor. The popularity of MMA means many fans will be scrutinising the judges' verdicts.

The second reason for the scrutiny experienced by UFC judges is that the consequences of winning and losing can be significant. Given the large sums of money on offer for an athlete to win a contest, the immediate impact on an individual can be staggering. For example, consider the title fight between Jon Jones and Dominick Reyes in August 2020. This fight resulted in a highly contentious decision made by the judges. All three judges scored the bout in favour of Jones, despite 76.4% of the public believing Reyes was victorious[2]. Whilst the payout to both athletes was in the hundreds of thousands, Reyes reportedly lost an estimated $150,000 win-bonus[3].

The consequences of winning and losing are not limited to short-term financial gain. Losing a fight can drastically affect an athlete's future career prospects. Since Jones is widely regarded as the best MMA athlete of all time, had Reyes won, he could have begun to build a legacy as one of the sport's greatest competitors. His future fight(s) as the champion would have certainly been accompanied by larger payouts. MMA is an unforgiving sport: given the ever-changing rankings and increasing pool of talent within each organisation, and the

[1] https://www.tapology.com/search/mma-event-figures/ppv-pay-per-view-buys-buyrate

[2] http://www.mmadecisions.com/decision/10877/Jon-Jones-vs-Dominick-Reyes

[3] https://www.sportekz.com/mma/jon-jones-vs-dominick-reyes-purse-payouts

limited amount of fights an athlete can compete in each year, losing a single bout can set an athlete back years as they have to begin climbing the rankings to achieve a title-shot again.

Given this background, and the importance placed on judges making good and fair decisions, in this paper, we develop a Bayesian hierarchical model to investigate judging within the UFC. We use our model to show that individual judges have different preferences regarding the techniques that athletes may attempt. These preferences can be the difference between winning and losing a fight. We believe this is an important finding to the sport. In the aftermath of controversial verdicts, athletes, fans, and stakeholders need to understand how an official may have come to their conclusion. But ultimately, our model can be used to homogenise judges' preferences such that unfair or controversial decisions are less commonplace. Our judging model can help train new judges, identify current judges who may be performing poorly, or even provide a benchmark score to assist when giving verdicts on fights.

In gymnastics Heiniger and Mercier (2021) developed tools to assess individual judges' scores objectively. Gymnastics is scored by penalising athletes for various errors, and it is a judge's task to accurately detect such errors during a given routine. Control scores can be derived post-competition by an outside judging panel using video reviews. Judges' scores can then be compared with the median of all other panel and control scores to assess their skill.

Judging in figure skating has come under much scrutiny. Accusations of corruption during the 1998 and 2002 Olympics led to a new scoring system being introduced[4], whilst the scoring system itself came under fire in Frederiksen and Machol (1988) who showed the system had paradoxical properties such as intransitivity.

Boxing judges have a long and notorious history of poor decisions. One of the most famous was a draw between Lennox Lewis and Evander Holyfield–in which the media and public believed Lewis clearly won the fight. This particular fight was the subject of research in Lee et al. (2002). The authors used exact tests, logistic regression, and a direct Bayesian model to demonstrate that two of the three judges scored the bout significantly different from other professionals. Interestingly, the authors acknowledge the key point we address in this paper: that judges may weigh the various criteria of boxing differently. However, they do not explicitly investigate or include this aspect.

Despite its popularity, the relative youth of MMA means that its judges have been the subject of just a handful of papers. Collier et al. (2012) and Feldman (2020) explore the effect of the various actions on the judges' decisions (finding that knockdowns are the most

[4] See, for instance, https://en.wikipedia.org/wiki/2002_Winter_Olympics_figure_skating_scandal

influential); Gift (2018) extends these models to include variables possibly indicative of bias. These three papers have a commonality: they model the population-level effects of actions on the judges' decisions. We progress the literature by implementing hierarchical models, allowing each judge to have their own effects.

Although identifying judges' preferences is clearly an important issue, not only has this not been discussed in the MMA judging literature, we are yet to find examples from any sport.

5.2 Data

We collected the scores within each round of UFC fights submitted by judges (and fans) from mmadecisions.com. Only fights which ended via a decision are available, limiting the dataset to fights where the judges' scores were used to decide on the outcome.

To model the judges' decisions as a function of the events occurring in a fight, and to identify how each judge valued each type of action, in-round fight statistics were scraped from ufcstats.com. The in-round statistics covered a variety of actions, including: 'strikes', the location of the strike (e.g. head, body, or leg), and the strength of the strike (split into two categories: significant or non-significant); 'takedowns' (actions used by a fighter to bring an opponent to the ground); 'control time' (how long a fighter was in a dominating grappling position in the round); 'submission attempts' (e.g. chokes and joint locks); 'reversals' (actions used to take a fighter from being controlled to being in control of a grappling exchange); and 'knockdowns' (strikes which cause an opponent to fall to the ground).

The in-round statistics reveal much of what has gone on in each round of a fight. Nevertheless, there may be some unmeasured events that might influence the judges. Consequently, we collected the official rankings of the fighters (at the time of the fight), and bookmakers' odds for the contest. The rankings of fighters were obtained from kaggle.com/datasets/martj42/ufc-rankings. The rankings are given for the top 15 fighters in each weight category, with all other fighters classed as 'unranked'. We assign unranked fighters a rank one more than the weight category maximum (i.e. we give them a ranking of 16 if the maximum rank was 15). It is convention in the UFC that the current champion is ranked at rank 0 (the other ranked fighters are effectively challengers to the champion). To make interpretations simpler, we reverse the order of the rankings, so that better rankings have larger numerical values.

In addition to the rankings of the fighters, we propose using bookmaker odds as a further proxy of unmeasured characteristics of the fight. Historical bookmaker odds were collected from bestfightodds.com. These data consist of the closing odds from several bookmakers, including William Hill, DraftKings, and Unibet. We obtained the odds in the result market (that is, the odds of each fighter winning) and the result-method market (the odds of each fighter winning by each possible method, i.e. fighter A/B to win by decision/knockout/submission).

We scraped each fighter's height, reach, and date of birth from ufcstats.com as these may affect how the judges view the fight.

Finally, we obtained information on the attendances at each event from wikipedia.org/wiki/List_of_UFC_events. During the Covid-19 pandemic, there were 51 events (281 fights) with no fans. These fights provide a unique opportunity to assess whether, and to what extent, the crowd influences individual judges, and whether judges favour fighters competing in their home country. We note there were 24 events (109 fights) where the UFC did not release the attendance. We omitted these events from the analysis.

Data from these separate sources needed to be merged, resulting in a final dataset of 17,105 unique judge's scores from 5,800 rounds in 1,840 fights spanning from 16/02/2013 to 18/06/2022. This included 309 unique judges who scored a median of 12 (mean of 55.36) rounds, with minimum and maximum rounds of 3 and 1,573, respectively.

There were 38 rounds in which a fighter was deducted one point, and three rounds where a fighter was deducted two points. Since these deductions were applied at the referee's discretion, we chose to model the "adjusted" score of the judges and unapply these deductions. Of these adjusted scores, 16,359 (95.64%) rounds were scored 10-9. There were 728 (4.26%) scored as 10-8, just two (0.01%) scored as 10-7, and finally, 16 (0.09%) draws.

In addition to the main dataset, we were able to obtain the judgemental scores of fans for 1,832 of the fights. We found a median of 32 (mean of 94.35) fans submitted scores for a fight, the maximum was 4,030 (interestingly, for the aforementioned Jon Jones vs. Dominick Reyes fight), whilst the minimum was four. Again, we modelled the adjusted scores, and found 93.49% of scores were 10-9, 4.77% were 10-8, 0.20% were 10-7, and 1.54% were draws. This immediately shows that the fans are much more likely to submit rarer scores, particularly in the case of 10-10. We will use the fans' scores in a separate model from the judges' scores to compare how fans value actions to how judges do.

The variables used to predict which fighter won a round are as follows:

- *In-round statistics*: knockdowns, significant head/body/leg strikes landed/missed, takedowns landed/missed, reversals, control-time, submission attempts, non-significant strikes landed/missed[5].

[5] The non-significant strikes in the data are not as granular as the significant counter-parts and only split by whether they landed or not.

- *Fighter ability*: bookmaker implied probability of winning (Win IP), and winning by decision (Decision-win IP), knockout (Knockout-win IP), or submission (Submission-win IP). The bookmaker odds are included in the model to control for fighter skill which the in-fight statistics may not have captured. We remove the bookmaker's margin and include the odds as implied probabilities (IP).

- *Fighter information*: age, height, reach, official ranking, and whether the fighter is the weight category champion. 'Stance' describes how the fighter stands in a fight. The different categories of fighter stance are left foot forward (orthodox, 77.09% of fighters), right foot forward (southpaw, 18.21% of fighters), a mix of the two (switch, 4.55% of fighters), or neither (open, 0.15% of fighters). We used a single variable, 'orthodox stance', to assess how judges interpreted different styles.

- *Crowd information*: we include a binary indicator representing whether the athlete is fighting in their home country, as well as the interaction of this term with an indicator representing whether a live audience was present.

- *Judge identity*: we know the name of the judge awarding the scores, and use these to identify differences between judges' valuations of the different variables with regards to the variable's contribution to the judge's round score.

In the remainder of the paper, we will broadly refer to any variables that are not in-round statistics as 'bias' variables, since they should not directly influence the fight's outcome. This includes the bookmaker implied probabilities, as although they likely do contain unseen skill information, strictly speaking, they were not indicators of events which happened.

5.3 Methodology

We fit a hierarchical ordered-logit model in a Bayesian framework using the STAN software Stan Development Team (2021a) within the R statistical programming language R Core Team (2020). Utilising a hierarchical Bayesian framework means that the common prior distributions will more heavily influence the coefficients of judges with limited data. Intuitively this makes sense since judges with a small number of observations will have their coefficients shrunk towards the common prior mean. A frequentist approach could result in large and unrealistic coefficient estimates.

Let $y_{rj}^{ab} \in \{$7-10, 8-10, 9-10, 10-10, 10-9, 10-8, 10-7$\}$ denote the score given by a judge j in round r from the perspective of fighter a facing opponent b. Suppose we have $n = 1, \ldots, N$

observations of judges' scores of unique rounds within fights, and for brevity, we will refer to these as y_n. Suppose there exists $j = 1, \ldots, J$ judges and $k = 1, \ldots, K$ predictors.

For each variable, we use the difference between the opposing athletes' values in that variable. This applies to any binary variables as well. For instance, if a home fighter is fighting an away fighter, they will have $+1$ and -1, respectively. Since the observations for opposing fighters are now mirror-images of each other, we randomly sample one observation to be used for model fitting. Further, we rescaled each variable by dividing by its maximum absolute value (this ensures differences of zero still have zero effect).

We model y_n using an ordered-logit regression with mean λ_n and thresholds indicating the cutoffs of each category denoted by $t = (t_1, \ldots, t_6)$.

To ensure our model is realistic, the probability of a fighter getting a 10-9 must be identical to the probability their opponent receives a 9-10. This is one issue not discussed by Gift (2018). To implement this in our ordered logit, we do not directly estimate the cutoffs of each threshold, but instead estimate the spacing. Imagine both fighters having not attempted any techniques start at 10-10. Any subsequent actions shift the predicted score probabilities away from 10-10 in either direction. Consequently, s_1 denotes the spacing between zero and the threshold of a fighter winning 10-9. Then, s_2 denotes the space between the 10-9 and 10-8 thresholds. Finally, s_3 denotes the space between winning 10-8 and 10-7. The vector $t = (-s_1 - s_2 - s_3, -s_1 - s_2, -s_1, s_1, s_1 + s_2, s_1 + s_2 + s_3)$ denotes the six cutoffs. We place a weakly informative half-normal prior on these spacings, $s_i \sim \text{Half-Normal}(0, 5)$, for $i = 1, 2, 3$. Consequently, we ensure the spacings are positive, the cutoffs are ordered correctly, and there is the required symmetry.[6]

Each judge has an individual set of parameters, representing the value they attribute to each action, denoted by $\beta_j = (\beta_{j1}, \ldots, \beta_{jK})$. We place a multivariate-normal prior on the $J \times K$ parameters–as suggested by Gelman and Hill (2006, ch. 13)–enabling correlation between judge's preferences. One can imagine such correlations exist as judges may favour grappling or strikes, perhaps due to their background in martial arts.

A weakly-informative hyper-prior is placed on the mean of the MVN prior, such that $\mu_k \sim N(0, 5)$. The covariance matrix, Σ, is decomposed into a correlation matrix, Ω, and vector of coefficient scales $\sigma_{1,\ldots,K} \sim \text{Half-Normal}(0, 2.5)$ (Barnard et al., 2000). The correlation matrix is given a prior of LKG(2), as recommended in Stan Development Team (2021b, ch. 1.13).

[6] Whilst we are surely not the first to implement a symmetrical ordered-logit model such as this, we note that we found no statistical packages implementing it within `R`. For instance, the most widely used function for implementing ordered logistic regression, `polr` within the `MASS` package, has no such feature.

The model in full is thus as follows:

$$y_n \sim \text{Ordered-Logit}(\lambda_n, t) \tag{5.1}$$

$$\lambda_n = \beta_{j_n} x_n \tag{5.2}$$

$$t = (-s_1 - s_2 - s_3, -s_1 - s_2, -s_1, s_1, s_1 + s_2, s_1 + s_2 + s_3) \tag{5.3}$$

$$\beta_j \sim \mathcal{N}_K(\mu, \Sigma) \tag{5.4}$$

$$\mu_k \sim \mathcal{N}(0, 5) \tag{5.5}$$

$$\Sigma = \text{Diag}(\sigma)\Omega\text{Diag}(\sigma) \tag{5.6}$$

$$\sigma_k \sim \text{Half-Normal}(0, 2.5) \tag{5.7}$$

$$\Omega \sim \text{LKJ}(2) \tag{5.8}$$

$$s_{1,2,3} \sim \text{Half-Normal}(0, 5). \tag{5.9}$$

We will briefly summarise the different components of the model. Equation 5.1 gives the main observation-level ordered-logit model, consisting of an unobserved latent variable λ_n and a vector of thresholds t. Equation 5.2 shows that the unobserved latent variable for a given observation is the linear combination of the independent variables x_n and the judge's individual set of coefficients β_{j_n}. From equation 5.4, each judge's vector of coefficients come from a multivariate-normal prior distribution, with group-level mean and covariance matrix, μ and Σ, respectively. The group-level mean for each of the k coefficients come from a $\mathcal{N}(0, 5)$ hyper-prior distribution (equation 5.5). In equation 5.6 the covariance matrix is decomposed into a correlation matrix, Ω, and vector of coefficient scales, σ. We place weakly-informative hyper-priors on these in equations 5.7 and 5.8. The thresholds in the ordered-logit model are actually defined using three spacings (equation 5.3), to ensure the required symmetry and ordering is adhered to. Finally, weakly-informative priors are placed on these spacings in 5.9.

To increase efficiency and improve the likelihood of convergence, we re-parameterise the model using the 'non-centred parameterisation' as described in Stan Development Team (2021b, ch. 23.7). A summary of the non-centred parameterisation is given in Appendix C.1.

We run four chains, each with 2,000 samples and 2,000 warm-up iterations. The smallest effective sample size was 1482.13, and the largest $\hat{R}$ (the potential scale reduction factor) was 1.002; both indicate convergence.

We also have information on how fans have scored each round. The data is in the form of the number of fans awarding each score. We thus fit a model with y as the round score but weight the observation according to the proportion of fans who scored that particular round in that way. The full fans model has been included in Appendix C.2, and we discuss comparisons with the judges model in the next section.

5.4 Results

5.4.1 Population effects

Table 5.1 presents a summary of the latent population-level effects of each variable on the judges' scores, indicated by μ in the model. These parameters indicate the average effect of each action. We also report the 2.5% and 97.5% Highest Density Intervals (HDI) for each coefficient.

Table 5.1. Summary of population-level effects in the model. Recall that each variable was rescaled by dividing by its maximum absolute value. Consequently, we display the "unit effect", that is, the effect of a one unit increase in each variable, and order the table by these values.

Variable	Mean, μ	Unit effect	SD	HDI (2.5%)	HDI (97.5%)
Knockdowns	6.705	1.676	0.461	5.809	7.623
Submission-win IP	0.590	1.023	0.303	−0.007	1.184
Knockout-win IP	0.581	0.983	0.316	−0.009	1.216
Submissions	3.929	0.786	0.324	3.281	4.551
Decision-win IP	0.376	0.730	0.277	−0.162	0.910
Reversals	0.767	0.384	0.162	0.439	1.081
Takedowns landed	3.070	0.341	0.267	2.577	3.623
Home∗Crowd	0.336	0.336	0.107	0.132	0.545
Significant head landed	9.902	0.160	0.293	9.325	10.465
Significant body landed	4.360	0.109	0.337	3.678	4.993
Significant leg landed	2.786	0.099	0.167	2.440	3.098
Win IP	0.061	0.074	0.361	−0.656	0.749
Non-significant missed	2.370	0.053	0.614	1.156	3.567
Significant body missed	0.644	0.032	0.192	0.244	1.007
Champion	0.029	0.029	0.135	−0.240	0.290
Non-significant landed	3.078	0.022	0.361	2.358	3.755
Height	0.192	0.021	0.109	−0.023	0.400
Ranking	0.257	0.016	0.124	0.020	0.504
Significant leg missed	0.100	0.009	0.163	−0.214	0.421
Control-time	2.406	0.008	0.125	2.155	2.644
Significant head missed	0.312	0.005	0.172	−0.030	0.650
Age	0.059	0.003	0.092	−0.120	0.238
Reach	−0.381	−0.032	0.114	−0.602	−0.158
Orthodox	−0.040	−0.040	0.039	−0.115	0.036
Takedowns missed	−1.027	−0.103	0.160	−1.332	−0.705
Home	−0.184	−0.184	0.107	−0.386	0.029

All in-round actions, other than *Takedowns missed*, have a positive effect, in-line with the Unified Rules and past literature. As one would expect, "big" moves such as *Knockdowns* and *Submissions*–which both have the potential to finish a fight immediately–have the largest unit effects.

The bookmaker implied probabilities (*Win IP*, *Decision-win IP*, *Knockout-win IP*, and *Submission-win IP*) also have large positive effects, thus likely accounting for various unmeasured actions of the fighters. We note that these odds variables have a non-significant effect (at the 5% level), suggesting that the in-round actions capture the vast majority of what happened.

We see there is a non-significant effect for the home fighter main effect (*Home*), but a significant positive effect for the interaction with a live audience (*Home**Crowd*). This would suggest that the crowd influences the judges.

We also find a significant positive effect from the official rankings (*Ranking*), suggesting higher ranked fighters are overly favoured (recall, we reversed the order of rankings).

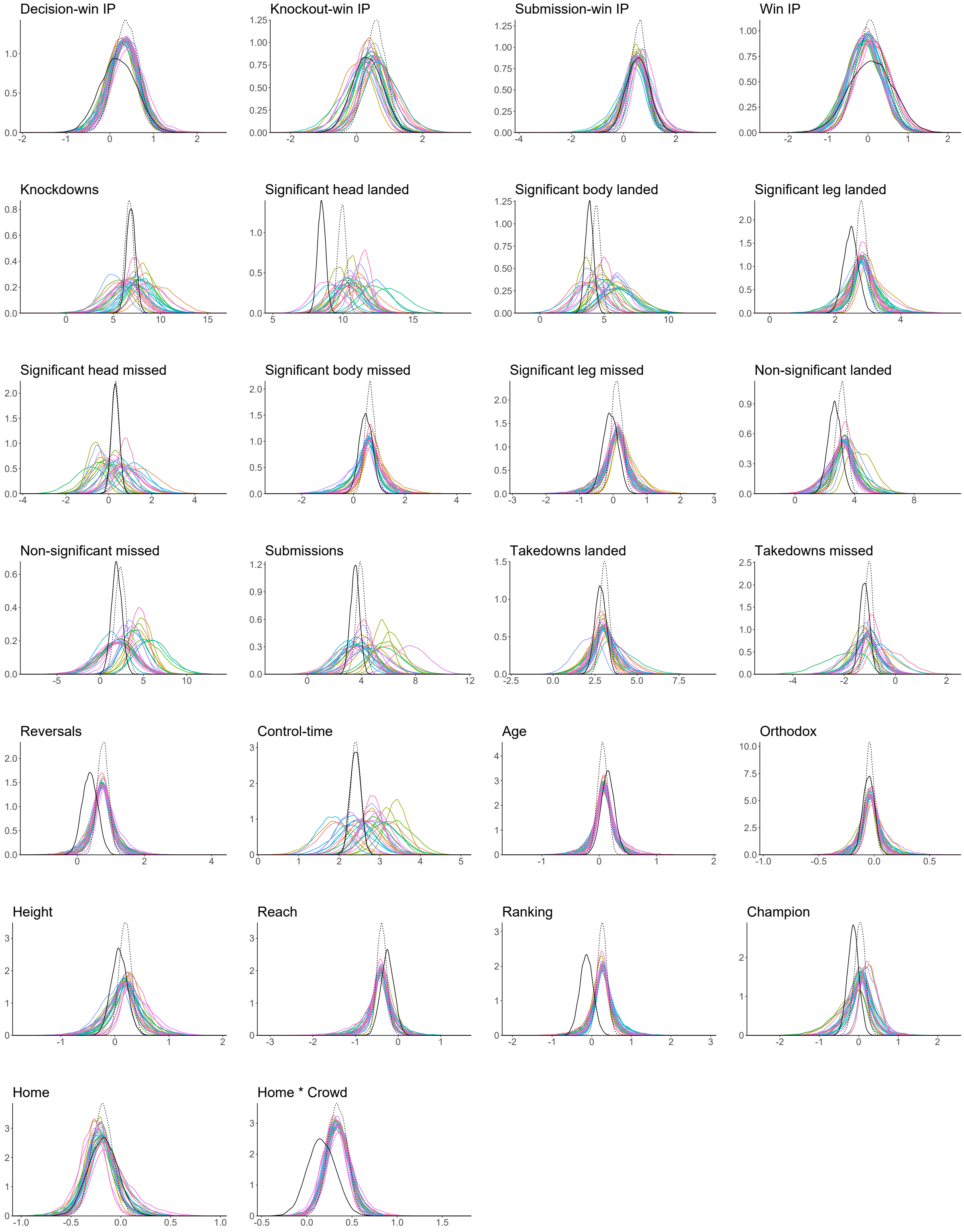
5.4.2 Individual preferences

The focus of this research is to investigate judges' preferences at an individual level and discern whether significant differences in judges' decision-making exist. Figure 5.1 displays the posterior densities of each coefficient for the 25 judges who scored the most rounds within the data. The latent population-level effect is shown as the dotted black line for reference. Further, the corresponding density based on the model fitted to the fans' scores for rounds is displayed as a solid black line.

We can immediately see the disagreement in how judges value several actions. The most striking difference is for *Significant head missed*, where some judges deem this as a positive effect, yet others see it as a negative. This itself is not surprising: whilst landing strikes should clearly have a positive impact on a fighter's score, it is harder to definitively say who, if anyone, should benefit from missing strikes. Should the defending fighter benefit from good defence in dodging the incoming strike? Or should the attacking fighter be awarded for being aggressive despite missing? The Unified Rules state: "No scoring is given for defensive manoeuvres. Using smart, tactically sound defensive manoeuvres allows the fighter to stay in the fight and to be competitive". This would suggest that whilst neither benefits in the 'effective striking' criteria; perhaps the attacker would benefit through 'aggressiveness'. Consequently, by the official rules, we would argue that those who value missed significant head strikes as a negative are incorrect.

Control-time also has a wide spread of densities; at least in this case, all judges agree

Figure 5.1. Plots of the posterior densities for the 25 judges who scored the most rounds, the latent population density (dotted black), and the fans (solid black).



on the sign of the effect. Control-time is one of the more complex and subjective actions to assess. The Unified Rules state that “top and bottom position fighters are assessed more on the impactful/effective result of their actions, more so than their position”. So merely being in control of an opponent should not weigh more than establishing an offence from a dominant position.

Looking at *Submissions*, there appear to be a few judges who are far from the others. In particular, one judge is almost entirely separate from the population-level density.

There are several actions which are largely agreed upon both in size and sign of effect: *Significant leg landed*, *Takedowns landed*, *Takedowns missed*, and the ‘bias’ variables *Height*, *Reach*, *Ranking*, and *Home**Crowd* interaction.

5.4.3 Comparison with the fans

In this section, we look at how fans value each in-round action when judging the winner of the round, and compare the fans’ valuations with the judges’. The fans effectively act as a crowd, and there have been several studies on the wisdom of crowds in sports. Brown and Reade (2019), for example, look at the wisdom of amateur crowds, like ours is, in predicting the outcome of sporting events, including martial arts. As is mostly the case in such studies, the crowd proves to be ‘wise’. Here, we do not know the ground truth (what the round should have been scored), but we can compare the crowd (the fans) with the judges.

For all variables, the fans’ opinions are within the observed densities of the individual judges, and for most variables, there is an overlap with the overall population effects. Several of the variables are almost identically weighted: *Knockdowns*, *Significant head missed*, and *Control-time* for instance. Some variables have more obvious differences: *Significant head landed* and *Reversals*, but, for the most part, it appears the fans and judges value actions very similarly to the judges.

One interesting finding is that the fans appear to be *less* influenced than the judges by the bias variables. The effects for the official rankings and being a home fighter are of particular interest. We see notable differences between the fans and the judges, whereby the coefficients for the fans are closer to zero than the judges. The most likely reason for bias towards home fighters in front of a live audience is that noise sways the judges. The fans who submit scorecards to `mmadecisions` will (most likely) not be present in the audience and thus will be less exposed to the noise. Perhaps the powers that govern UFC might consider having one or more judges away from the arena, similar to how Video Assisted Referees (VAR) operate in football.

We will now compare the differences in the scoreline thresholds, s_i (which have not been plotted). We will create several pseudo-fights to assess the probabilities of the different scorelines. For a given value q, we will find the q'th quantile of the absolute value of each variable[7]. These quantiles become the variables in the associated pseudo-observation. Consequently, $q = 0$ represents the closest round possible, in which the athletes were even across all variables, whilst $q = 1$ is the most one-sided round possible, in which the dominant fighter scored the maximum of each variable. We note that whilst we explicitly allowed correlations between the variables in the model, we have not included such correlations in this set-up.

The predicted probabilities for each scoreline for several different values of q are shown in Table 5.2.

Table 5.2. Probabilities of each scoreline predicted by the judge and fan models for several different pseudo-rounds. For each q we calculate the q'th quantile of the absolute value of each variable and include these values as the variables for the observation. Consequently, $q = 0$ represents a round in which both fighters were exactly even across all variables, whilst $q = 1$ represents the most one-sided round possible in which the fighter scored the maximum in each variable. We ensured that variables with an estimated negative coefficient were accounted for.

	Judges				*Fans*			
q	10-10	10-9	10-8	10-7	10-10	10-9	10-8	10-7
0.00	0.002	0.498	0.000	0.000	0.029	0.483	0.002	0.000
0.25	0.002	0.693	0.001	0.000	0.026	0.651	0.005	0.000
0.50	0.001	0.927	0.007	0.000	0.009	0.884	0.023	0.000
0.75	0.000	0.855	0.142	0.000	0.001	0.739	0.250	0.003
0.85	0.000	0.401	0.599	0.000	0.000	0.303	0.676	0.018
0.90	0.000	0.080	0.920	0.000	0.000	0.070	0.824	0.096
0.95	0.000	0.001	0.967	0.032	0.000	0.002	0.166	0.832
0.98	0.000	0.000	0.444	0.556	0.000	0.000	0.009	0.991
1.00	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000

One “oddity” to note from Table 5.2 is that in close rounds ($q = 0$), it is more likely that the judge picks one of the fighters to win the round than give a draw score (10-10). It would seem more natural to score a close round as a tie. However, this is an effect of the Unified Rules which actively discourage the use of 10-10 scorelines: “A 10-10 round in MMA should be extremely rare and is not a score to be used as an excuse by a judge that cannot assess the differences in the round... If there is any discernible difference between the two fighters during the round the judge shall not give the score of 10-10”. Even when examining a fake fight in which each fighter was exactly even across all variables (that is, $q = 0$), the mean posterior predictive probability of a 10-10 round is just 0.002.

However, the fans are much more likely to give a 10-10 round: for $q = 0$, fans have a 0.029

[7] For variables with a negative coefficient, we then multiply this value by -1.

probability, whilst for the judges, it is just 0.002. Proportionately this is a massive difference, which perhaps demonstrates the fans are not aware of the rules actively discouraging tied rounds, or they are more willing to ignore the rules.

For a given round, generally speaking, a judge will be choosing between 10-9 and 9-10, or 10-8 and 10-9. Due to the rules, it is hard to imagine a scenario whereby a judge could choose between 10-8 and 8-10, or 10-8 and 9-10. Indeed, in all of our experiments, we found that each round was essentially a pick between two scores.

Fans are also much more likely to give big scores, i.e. 10-8 or 10-7. Recall that just 0.01% of rounds were scored 10-7, but even at $q = 0.90$, a quite one-sided round, the judges have a zero probability of awarding 10-7, whilst for the fans, the probability is 0.096.

5.5 Case-studies

In this section, we will use the model to scrutinise judges' actual scores given within a round, overall scores, and overall decisions (that is, win, lose, or draw). Each of these case-studies will serve to highlight a particular use or feature of the model:

- In Section 5.5.1, we introduce the concept of a "significant prediction" to determine whether a judge's decision was valid based on the predicted posterior distributions of the probability for each outcome.

- In Section 5.5.2, we use our model to demonstrate that the judges' individual preferences can lead to different verdicts.

- Finally, in Section 5.5.3, we implement a "fair" model that removes the effect of the bias variables. We use this model to determine who should have won Jon Jones' and Dominick Reyes' infamous fight.

5.5.1 Zhang Weili defeats Joanna Jedrzejczyk (2020)

Regarded as one of the greatest and most competitive fights of all time, in 2020, the UFC's strawweight champion, Zhang Weili, defended her belt against number one contender and former champion, Joanna Jedrzejczyk. The bout was a back-and-forth affair, with each round being extremely close, but in the end, Zhang won via a split decision (48-47, 48-47, 47-48).

The fan scores highlight how close this fight was, as 48.6% gave the battle to Jedrzejczyk and 48.2% to Zhang (based on 1,291 scorecards[8]). The most common score was 47-48, which

[8] http://www.mmadecisions.com/decision/10984/Weili-Zhang-vs-Joanna-Jedrzejczyk

was given by 36.8% of fans; however, 48-47 was the verdict of 33.2%.

We will use this close fight to introduce the concept of “significant predictions”. Testing the significance of a variable within a model is a staple of quantitative research, yet, to our knowledge, the concept has not been applied to predictions, despite obvious uses. In the context of MMA judging, we want to see whether a judge’s decision was valid, even if it may not have been the most likely choice.

Figures 5.2a, 5.2b, and 5.2c, display the posterior predictive probabilities for each score within the round, each score overall, and the overall result (win/draw/lose), for each of the three judges (Bell, Cleary and Colon). These plots are based on our model that accounts for their individual preferences. For brevity, we will refer to these plots as the *round*, *score*, and *result* plots, respectively. When calculating the overall score and result probabilities, we ensured the same posterior sample of the coefficients were used across the rounds. The chosen score is shown in blue.

Looking at Figure 5.2a, the distributions for each scoreline reveal much overlap in each round. Round three is a good example as the densities for Colon scoring 10-9 and 9-10 are practically identical. Consequently, although Colon has a higher probability of choosing 10-9, it is not controversial that he decided to score the round 9-10. To formalise this concept, we apply the concept of statistical significance.

Having estimated our Bayesian model through MCMC, we obtain N posterior samples for each coefficient and parameter of the model. Thus, we can obtain the seven probabilities corresponding to each score within a round for a particular sample.

Suppose we want to compare whether the probability of scoring the round as i is significantly different to j. For posterior sample s, denote the probability of scoring the round as i, as p_{si} (similarly, p_{sj} denotes the probability of scoring the round as j). Now, for $s = 1, \dots, N$, we calculate the difference between these two probabilities, $p_{si} - p_{sj}$, and denote the distribution of these differences over all s, as $\mathbf{d}$.

In an extremely close round, $\mathbf{d}$ will be centred at 0, whilst in the most extreme case, it will be close to either -1 or 1. If the majority of $\mathbf{d}$’s mass is on one side of 0, then that suggests a significant difference exists between the two sets of probabilities. We then calculate the proportion of $\mathbf{d}$ on each side of 0, and find the minimum of these to be p. If $p < \alpha$, then we can say that the sets of probabilities are significantly different at the α level.

We display a p-value in each plot. In cases where the judge’s decision was different to the most likely predicted by the model, we calculate the p-value associated with the difference of these two sets of probabilities. If the judge submitted the most likely score, then we give the p-value comparing that score with the second most likely score.

We see from Figure 5.2c that although the model predicts Zhang actually should have lost[9], this result was not significant for any of the judges. Consequently, we can conclude that all of their final verdicts were within reason.

5.5.2 Edson Barboza defeats Danny Castillo (2013)

We use this fight to demonstrate how the individual preferences of the judges themselves may influence the final outcome.

In this bout, most fans believed the outcome was a 28-28 draw (61.8% of 152). The first round was dominated by Castillo, with the majority thinking it was an 8-10 (87.5%). The remaining two rounds were then clearly Barboza, with 73.7% and 96.7% giving him rounds two and three as 10-9, respectively.

Judge Derek Cleary scored the bout this way, arriving at the 28-28 consensus scorecard. However, Michael Bell and Wade Viera arrived at 29-28, having given 9-10 in the first round. Figures 5.3a, 5.3b, and 5.3c display the round, score, and result plots for this fight, respectively.

Looking at Figure 5.3a, we see the model would have predicted Bell and Cleary to score the first round as 8-10, but interestingly, Vierra would have most likely given it a 9-10. There are no further disagreements in rounds two or three. Consequently, the most likely scores for Bell and Cleary were 28-28, but 29-28 for Vierra. Correspondingly, the most likely result for Bell and Cleary was a draw, whilst Vierra was a win for Barboza. We note that all these results are significantly different by the p-values introduced in Section 5.5.1.

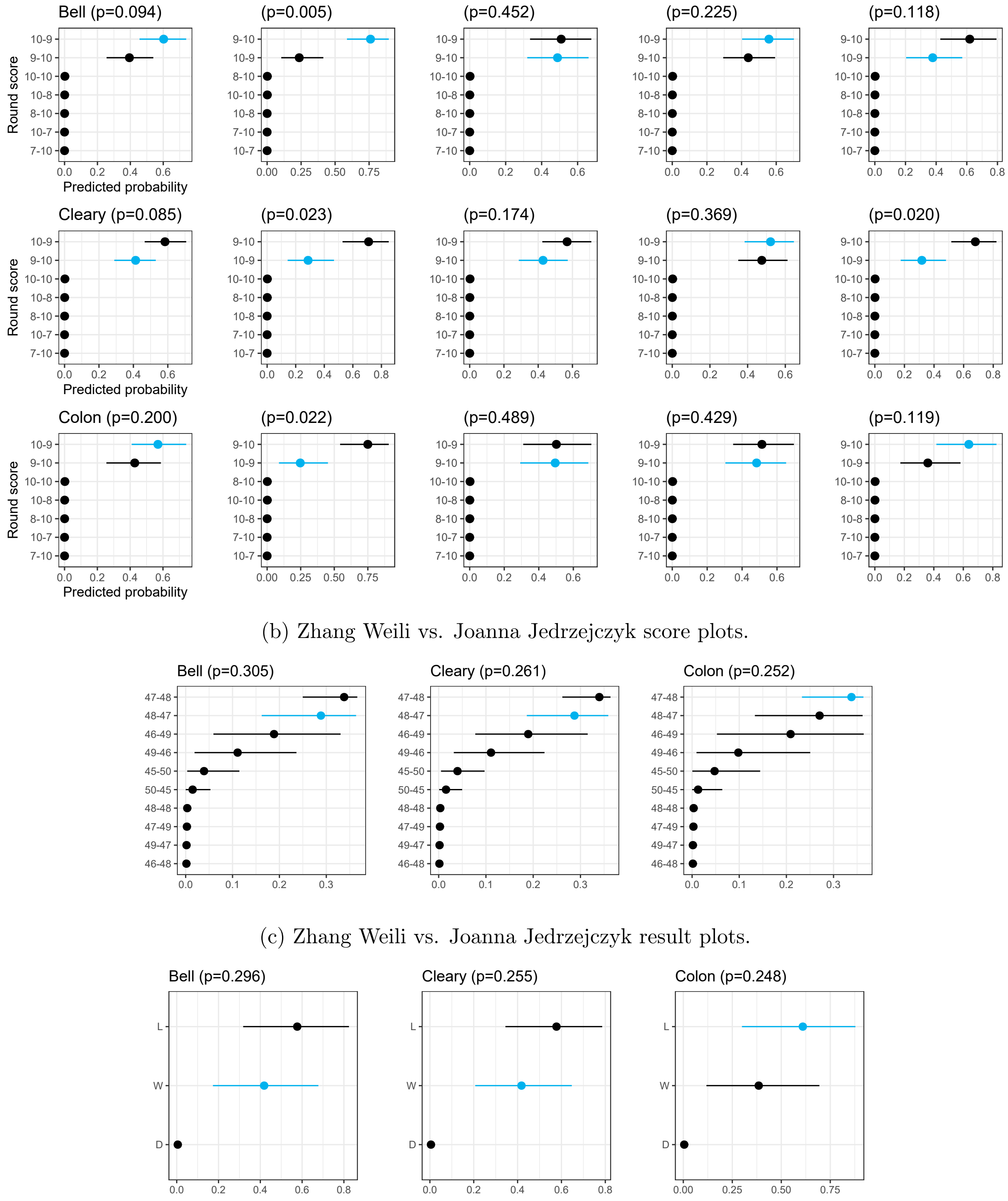
Given that we predicted the judges to predict different outcomes, this demonstrates how their individual preferences may influence the result of a fight. We believe this is important for all participants of MMA–stakeholders, athletes, fans, and even judges themselves–to understand. Given the often high-stakes nature of bouts, it is common to see judges receive backlash for their decisions. Understanding these judges have their own opinions helps everyone appreciate the complexities of judging in what is still a relatively young sport, and avoid unnecessary bad publicity.

5.5.3 Jon Jones defeats Dominick Reyes (2020)

We now return to the infamous bout between Jon Jones and Dominick Reyes, to introduce the concept of the "fair-score" model, which stakeholders could use in various ways to compare or calibrate judges.

[9] An interesting point to make is that although in the round plots we would have predicted Zhang to win–having been favoured in three rounds–we wouldn't predict her to win overall. This is because the rounds that Jedrzejczyk won were won with a much higher probability.

Figure 5.2. Plots detailing the predictive posterior probabilities for the scores within each round, the overall scores, and the overall result, for the three judges in the bout between Zhang Weili and Joanna Jedrzejczyk. All scores are given from the perspective of Weili. Associated p-values of the predicted probabilities, introduced in Section 5.5.1 are also given. The chosen score is shown in blue.



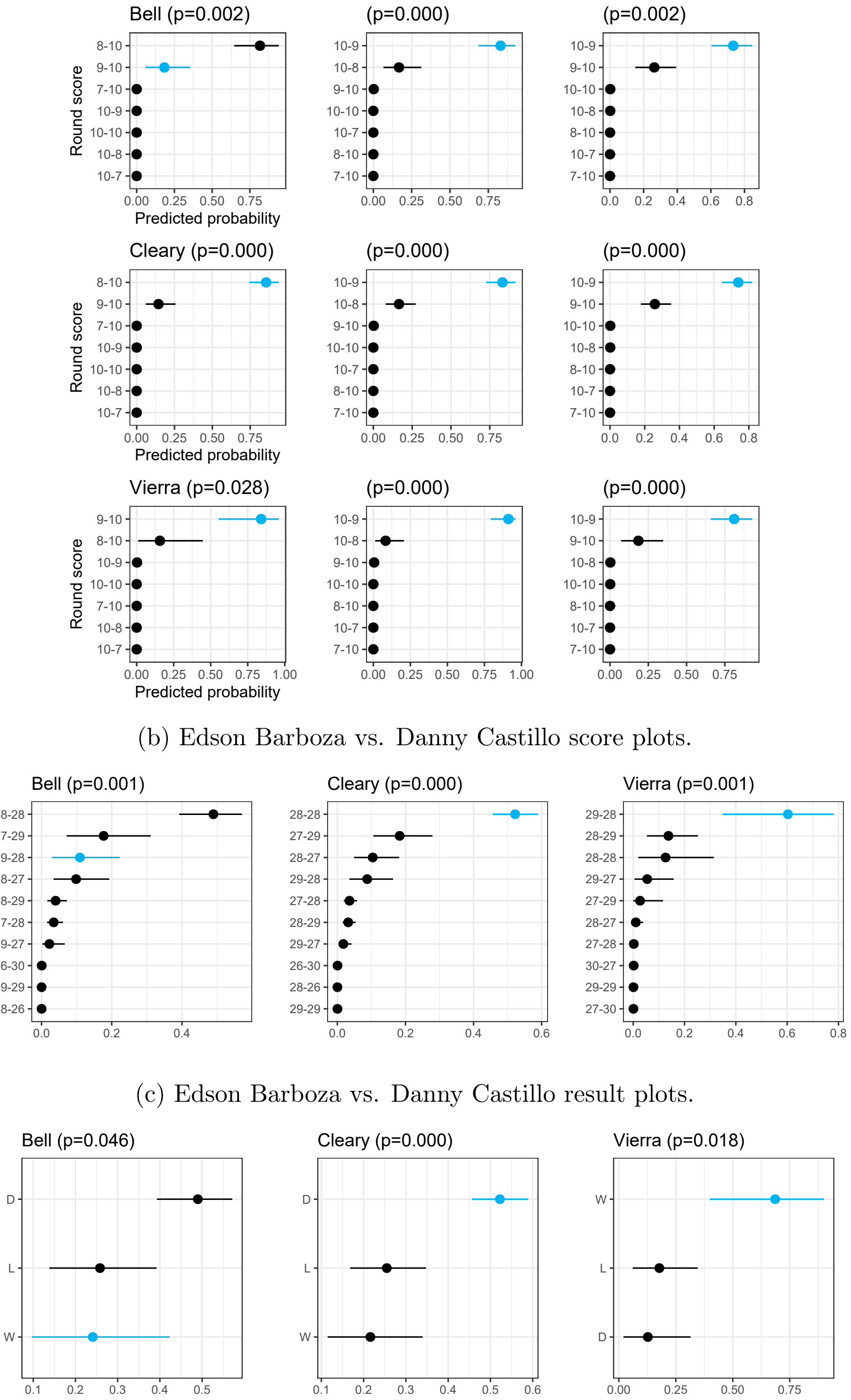
(a) Zhang Weili vs. Joanna Jedrzejczyk round plots. The left plots are Round 1, moving to Round 5 on the right of the figure.

(b) Zhang Weili vs. Joanna Jedrzejczyk score plots.

(c) Zhang Weili vs. Joanna Jedrzejczyk result plots.

Figure 5.3. Plots detailing the predictive posterior probabilities for the scores within each round, the overall scores, and the overall result, for the three judges in the bout between Edson Barboza and Danny Castillo. All scores are given from the perspective of Barboza. Associated p-values of the predicted probabilities, introduced in Section 5.5.1 are also given. The chosen score is shown in blue.

(a) Edson Barboza vs. Danny Castillo round plots. The left plots are Round 1, moving to Round 3 on the right of the figure.



(b) Edson Barboza vs. Danny Castillo score plots.

(c) Edson Barboza vs. Danny Castillo result plots.

The first step in establishing the fair score is removing the bias variables' effect. Recall from Section 5.2, these are any variables which aren't in-round statistics, including the pre-fight bookmaker implied probabilities.

To remove the effect of the bias terms, we could fit a new model which uses only the in-round statistics as independent variables. However, we have established several significant effects from the bias variables. Consequently, removing them entirely would introduce inherent "omitted-variable" bias. Instead, we use the original model and set any bias variable to 0 when making the fair predictions. This has the desired effect of removing their effects, whilst not introducing biases.

In Section 5.4.2, we demonstrated that judges have individual preferences towards each action, and in Section 5.5.2, we showed how these preferences might determine who wins a fight. Consequently, in the fair model, we aim to remove these individual preferences, to establish an average score. With that in mind, we use the model's latent population effects, represented by μ, rather than the judge effects, β_j.

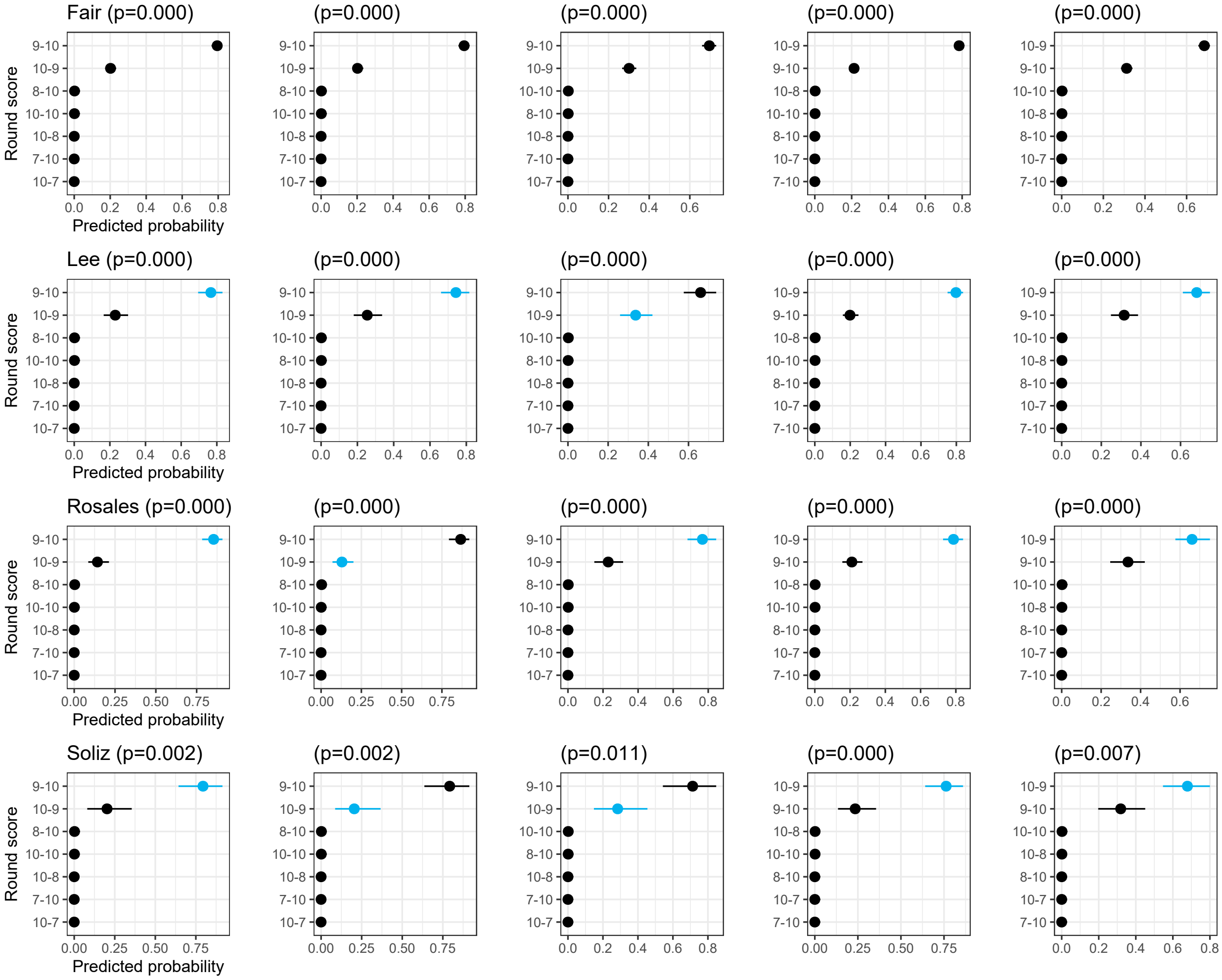
Figures 5.4a, 5.4b, and 5.4c display the round, score, and result plots. We include the posterior predicted probabilities associated with the fair model. For comparison to the fair model, we remove the effect of the bias terms from the judges' probabilities, but keep their individual preferences.

As was the consensus fan opinion, according to the fair model, Reyes should have won the first three rounds, and Jones the last two. The most likely score of the fair model in all of these rounds was significantly different from the next most likely score, with $p = 0.000$ in each.

From Figure 5.4b, we see that the fair model predicts the consensus score, 47-48, the score of 69.8% of fans. The 48-47 by Rosales and 49-46 by Soliz were significantly different from what we would have predicted they would score the fight. Lee's 48-47 was also significantly different, but to a lesser extent.

Finally, looking at the result plots in Figure 5.4c, the fair model would have predicted that Jones lost the fight. The model also predicts that each judge should have given the fight to Reyes overall. However, it is interesting that the decisions of Lee and Soliz are not significantly different from the predicted result. From the model's predictions of their behaviours (after removing bias terms), we see that either fighter winning would have been a just decision. The same cannot be said for Rosales, whom we would predict to side with Reyes.

(a) Jon Jones vs. Dominick Reyes round plots.



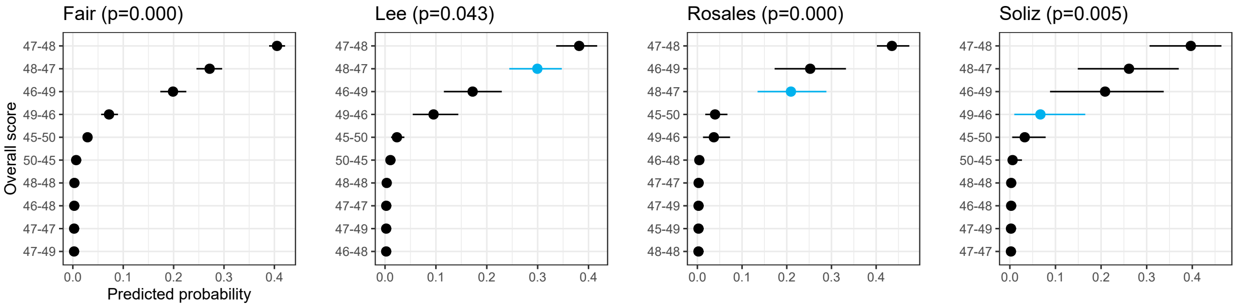
Fair (p=0.000) (p=0.000) (p=0.000) (p=0.000) (p=0.000)

Lee (p=0.000) (p=0.000) (p=0.000) (p=0.000) (p=0.000)

Rosales (p=0.000) (p=0.000) (p=0.000) (p=0.000) (p=0.000)

Soliz (p=0.002) (p=0.002) (p=0.011) (p=0.000) (p=0.007)

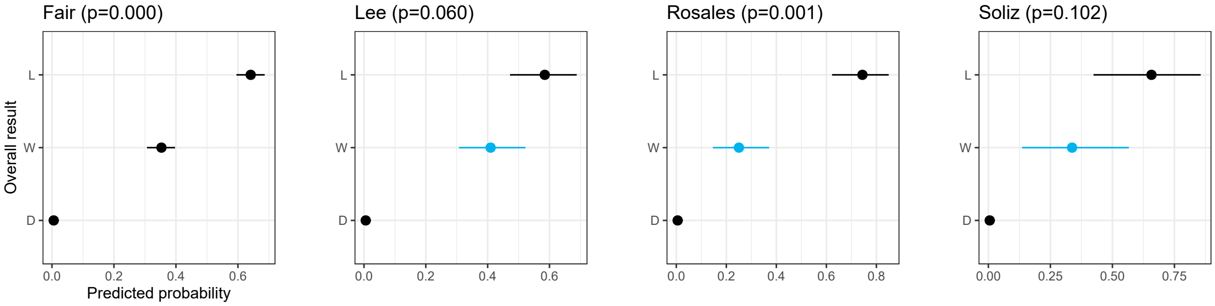
(b) Jon Jones vs. Dominick Reyes score plots.



Fair (p=0.000) Lee (p=0.043) Rosales (p=0.000) Soliz (p=0.005)

Figure 5.4. Plots detailing the predictive posterior probabilities for the scores within each round, the overall scores, and the overall result, for the three judges in the bout between Jon Jones and Dominick Reyes. All scores are given from the perspective of Jones. Associated p-values of the predicted probabilities, introduced in Section 5.5.1 are also given. We include the predictions made by the *fair* model, which removes the effect of bias terms and the judges' individual preferences. For the judges, we keep their preferences, but remove the effect of the bias variables (that is, any variable that is not an in-round statistic).

(c) Jon Jones vs. Dominick Reyes result plots.



5.6 Conclusions

In this paper, we investigated whether individual preferences exist between judges of MMA contests. Whilst there has been some research into judging in sports, we believe this is the first research to directly explore the different opinions of individual judges and how these differences can influence the outcomes of competitions and contests.

Using a Bayesian hierarchical model, trained on a large set of MMA scores including several novel variables, we found various levels of disagreement between judges across the different in-round actions. The most notable was in scoring missed significant head strikes. We found that some judges deemed these as positive actions, whilst others believed they were negative. It is stated in the rules that fighters should not be rewarded for successful defensive manoeuvres. Consequently, we believe the judges who assess them as positive are correct. Using a real-life example, we demonstrated that these preferences can be the deciding factors in a fight, and may lead different judges to declare different winners.

Whilst these findings point towards different preferences amongst the many MMA judges, they perhaps also evidence rather vague and subjective judging criteria. Whilst some subjectivity is inherent in all live judging, it should not be the case that an action can be positive with one judge and negative with another.

We demonstrated the use of our models in potentially detecting erroneous decisions and establishing who should have won a fight, given the data available to us. Technology has recently been successfully implemented in football and tennis to assist the officials, namely

VAR and Hawkeye. Whilst a mathematical model cannot entirely replace live judges (particularly until the level of available data is improved), we maintain there are several potential uses for our model: as a tool for training or calibrating the judges; detecting consistently problematic judges; gauging whether a fight was indeed controversial, and if so, how controversial; or demonstrating to judges that they may hold biases, as this might help to reduce them.

We introduced the concept of significant predictions in this paper. The judges' scores are particularly suited to this idea, as we want to see whether a given score is mathematically controversial or within reason.

A similar model was estimated to explain the scores submitted by fans on mmadecisions. Using this model, we could examine whether the fans and judges agree on the values of each action. Considering the recency of the sport's mainstream popularity, we were pleased to find that the fans weigh each in-round action comparably to the judges. The biggest difference is in the thresholds for giving each score: the public is much more likely to submit ties and big scores. An interesting finding was that the fans appear to be less influenced by bias variables, such as home-crowd influence and the official rankings.

Whilst investigating the fans' scores is interesting in its own right, there are real-world applications. The Professional Fighting Championship recently partnered with Verdict (who, like mmadecisions, allow fans to submit scores) so that in certain fights, the fans' scores are used as the official result. Our findings suggest that this potentially controversial approach may be a valid solution. However, further research should investigate the presence of other biases within the fan scores, for instance, biases towards more popular athletes, or disadvantages when fighting in a different timezone (as your fans may not be awake to submit scores).

Chapter 6

Conclusions

This thesis has served to advance the academic literature and knowledge of MMA in several ways.

The Markov chain-based forecasting model presented in Chapter 3 is the most advanced model available in peer-reviewed journals (Holmes et al., 2022), or the broader space driven professionals and hobbyists on websites such as kaggle.com. The models that drive the transition probabilities are the first of their kind within MMA. These transition models are the first attempt to scrutinise the atheltes' differing styles and abilities. The methodology used has several advantages. The most notable is the access to more detailed predictions, for instance, how many takedowns will fighter A attempt. We demonstrate that the simulations achieve a positive correlation with such counts. Consequently, whilst more exotic betting markets have been available in more mainstream sports (for instance, how many corners in a football match), our model opens up such markets for MMA. The analysis of our model against the betting market is also a first in MMA forecasting literature. The methodology is flexible enough to be applied to other sports. Indeed, similar approaches have already been implemented for tennis, and we hope to apply the methods to more complex sports such as football in due course.

We noted several possible improvements to the model; the majority rely on more granular data. Companies such as Opta have established "event-data" within football, which provides details on every on-ball action (e.g. passes, tackles, and shots). With similar data, we could include states such as the clinch, various ground positions, and different types of strikes. If the data is time-stamped, we can model various dynamics within a fight: the cardio of athletes, how different strikes impact them, or whether they will "coast" when winning a bout. "Tracking-data", which provides the coordinates of each player and the ball, is quickly revolutionising sports, and this would further improve the modelling. Tracking data would allow us to model the distance control of athletes, their movement capabilities, and how the cage is used, for instance. Future researchers could improve the transition modelling framework. Whilst we estimated each model independently of one another, it is likely that including correlation between an athlete's skills would lead to more realistic models. Optimising the prior distributions on an out-of-sample validation set is another simple, but time-consuming, improvement. A full MCMC framework, perhaps with different hierarchical levels, could also be explored despite the severe computational demands.

As discussed in Section 2.1.1, there are many applications of forecasting models beyond predictions. Our Markov model could be used by athletes and their teams to help prepare for a particular opponent. Promoters of MMA organisations can use the model to determine the competitiveness and significance of fights. This in-turn could improve match-making. The competitiveness and significance estimates could be used in a second model estimating the number of pay-per-view buys for an event. This model could assess the financial worth of current, or prospective, athletes.

We built an expansive database to scrutinise the judges of MMA, which included all the available in-round fight statistics and other variables potentially indicative of bias. In Chapter 4, we found a significant reputation bias, such that athletes placed higher in the rankings are significantly favoured. Compared to other forms of bias, reputation bias is relatively understudied; the few articles which do explore it use experiments consisting of expert officials. Hence, our study is the first to identify reputation bias using data from professional sports. We find evidence suggesting the judges significantly favour home athletes in the presence of a crowd, with no significant effect for behind-closed-doors events. We show these effects are not due to differences in fighter skills by including pre-fight bookmaker odds into the model (to account for skills unseen by the fight statistics) and demonstrating that the market is efficient. Market efficiency suggests that the skill information within the bias variables does not go beyond what is contained in the odds. Thus significant effects are likely due to bias, not skill. This check for robust conclusions is a novel contribution which can be used for similar research in any sport. We are the first to apply the "purposeful selection" model fitting strategy in the literature on biases. In the context of biases, where the interpretation of the final model is crucial, this is a valuable technique which blends key advantages of machine-learning algorithms with the well-understood inferences of logistic regression.

The conclusions from our research could help to reform judging within MMA. To limit the judges' exposure to noise from the crowd, they could be made to wear headphones or sit in booths isolated from the crowd (similar to VAR referees in football). Addressing reputation bias is a more difficult task, as the judges will likely know the athletes' rankings or learn them during the event. However, identifying and informing the judges of the bias may help reduce its effect. Future work could look at other potential biases. Nationalistic bias has been detected within many sports, and with a database of the judges' nationalities, we could examine that. Similarly, data on the athletes' races could be used to identify any racial prejudices. The official rankings are based on the opinions of different experts within the media so that one could investigate them for biases.

Finally, Chapter 5 investigated the preferences and behaviour of MMA judges at an individual level. Whilst authors have often acknowledged the likelihood of the judges having differing opinions, to our knowledge, we are the first to explicitly model this. We found varying levels of disagreement between the judges; in one case, judges disagreed on whether the action was positive or negative. We used three historical case studies to highlight different aspects of the model. First, we showed how the judges opinions may themselves be the deciding factor within a bout. Next, using the logic of a significant variable, we introduced an approach to determine whether a judge's decision was mathematically controversial or within reason. This is a novel contribution with applications beyond the article; for instance, a forecasting model may predict one competitor to win, but not significantly enough that you would bet on them. Finally, we used our model to generate "fair" scores by removing the effects of bias variables and the judges' preferences. We similarly modelled the fans' scores and found that overall they score a fight similar to the judges; however, they are much more likely to give ties or big scores. We found the fans are *less* influenced by the bias variables than the judges.

Given recent technological advances in sport, for instance, VAR and Hawkeye, we believe our model has real-world use in training and calibrating the judges. Further research could investigate whether the fans' scores could be a viable alternative, or at least a useful addition, to live judges. Other potential forms of bias would have to be explored. For instance, are there biases towards more popular athletes or when competing in a different time-zone.

Appendix A

A Markov Chain Model for Forecasting Results of Mixed Martial Arts Contests

A.1 Interpolating the UFC-Stats data

Recall from Section 3.2 that we made three simplifications to the striking data: all strikes are assumed to be significant, 'standing' refers to distance and clinch, and 'body' refers to body and leg. Thus, our striking data consists of three indicators:

- *Position*: whether the strike was executed from the standing (S) or ground (G) positions;
- *Target*: whether the strike was aimed at the head (H) or body (B);
- *Landed*: counting landed strikes (L) or all strike attempts (A).

A combination of these indicators then gives the specific statistic. For example, standing head strikes attempted is denoted by SHA and ground strikes landed is GL (which would constitute all targets). Similarly, takedowns are split into takedowns landed (TDL) and attempted (TDA); submissions are split by landed (SML) and attempted (SMA).

Tables A.1 and A.2 show the striking data for the athletes involved in the UFC's most lucrative fight to date: Khabib Nurmagomedov against Conor McGregor for the Lightweight title in 2018.

We can see the data from ESPN is more granular: strikes are split by the position *and* the target. In contrast, the data from UFC-Stats is split by the position *or* the target. Thus, we have to convert the striking data from UFC-Stats to the same format through interpolation.

We chose to perform this in a simple manner. First, we find the proportion of strikes a fighter performed from each position. To find the estimated strike targets from each position,

Table A.1. Example of strike data from ESPN for one fight

Fighter	SHA	SHL	SBA	SBL	GHA	GHL	GBA	GBL
Conor McGregor	54	28	21	17	6	6	0	0
Khabib Nurmagomedov	50	21	6	4	55	37	8	8

Table A.2. Example of significant strike data from UFC-Stats for one fight

Fighter	SA	SL	GA	GL	HA	HL	BA	BL
Conor McGregor	75	45	6	6	60	34	21	17
Khabib Nurmagomedov	56	25	63	45	105	58	14	12

we multiply the total strikes for a particular target by the calculated positional proportion.

As an example, suppose we are estimating the standing head strikes landed, SHL. The overall proportion of strikes which were aimed for the opponent's head are calculated as $hp = \mathrm{HL}/(\mathrm{HL} + \mathrm{BL})$. Then we can calculate $\mathrm{SHL} \approx hp \cdot \mathrm{SL}$.

A.2 Summary of the skill models

Table A.3 presents a summary of the skill models we estimated. The variables used are found in Table A.4.

A.3 Example simulation

This section will give a short example simulation to help understand how the chain works. Suppose a chain goes: *Standing* → *Standing* → *Stand strike attempt i* → *Stand head attempt i* → *Stand head land i* → *Standing* → *Standing* → *Takedown attempt j* → *Ground control j* → *Ground control j* → *Ground strike attempt j* → *Ground body attempt j* → *Ground control j* → *Ground control j* → *Submission attempt j* → *Submission victory j*.

There are a few interesting features to note. First, there are neutral transitions in which nothing happens (*Standing* → *Standing* and *Ground control j* → *Ground control j*). Second, the process of landing a strike is made up of several different states: choosing to strike (*Stand strike attempt i*), choosing a target (*Stand head attempt i*), whether it lands (*Stand head land i*), and whether it results in a knockout (in this example not, hence the chain transitions back to *Standing*). Finally, the chain terminates once in the absorbing *Submission victory j* state.

The transitions which use up time, each taking one second, are four neutral transitions, two strike attempts, one takedown attempt, and one submission attempt. Although a single

Table A.3. Summary of the skill models estimated. The variables used are summarised in Table A.4.

Skill	Model
Strike rate	$\mathrm{SA}_{ijk} + \mathrm{GA}_{ijk} \sim \mathrm{Poisson}(str_{ijk})$ $\log(str_{ijk}) = str\_int + str\_att_i + str\_def_j + str\_weight \cdot \mathrm{lbs}_k + \log(\mathrm{T}_k)$
Takedown rate	$\mathrm{TDA}_{ijk} \sim \mathrm{Poisson}(tdr_{ijk})$ $\log(tdr_{ijk}) = tdr\_int + tdr\_att_i + tdr\_def_j + tdr\_weight \cdot \mathrm{lbs}_k + \log(\mathrm{ST}_{ik})$
Submission rate	$\mathrm{SMA}_{ijk} \sim \mathrm{Poisson}(smr_{ijk})$ $\log(smr_{ijk}) = smr\_int + smr\_att_i + smr\_def_j + smr\_weight \cdot \mathrm{lbs}_k + \log(\mathrm{GC}_{ik})$
Standing head strikes accuracy	$\mathrm{SHL}_{ijk} \sim \mathrm{Binomial}(\mathrm{SHA}_{ijk}, sha_{ijk})$ $\mathrm{logit}(sha_{ijk}) = sha\_int + sha\_att_i + sha\_def_j + sha\_weight \cdot \mathrm{lbs}_k$
Ground head strikes accuracy	$\mathrm{GHL}_{ijk} \sim \mathrm{Binomial}(\mathrm{GHA}_{ijk}, gha_{ijk})$ $\mathrm{logit}(gha_{ijk}) = gha\_int + gha\_att_i + gha\_def_j + gha\_weight \cdot \mathrm{lbs}_k$
Standing body strikes accuracy	$\mathrm{SBL}_{ijk} \sim \mathrm{Binomial}(\mathrm{SBA}_{ijk} + \mathrm{SLA}_{ijk}, sba_{ijk})$ $\mathrm{logit}(sba_{ijk}) = sba\_int + sba\_att_i + sba\_def_j + textitsba\_weight \cdot \mathrm{lbs}_k$
Ground body strikes accuracy	$\mathrm{GBL}_{ijk} \sim \mathrm{Binomial}(\mathrm{GBA}_{ijk} + \mathrm{GLA}_{ijk}, gba_{ijk})$ $\mathrm{logit}(gba_{ijk}) = gba\_int + gba\_att_i + gba\_def_j + gba\_weight \cdot \mathrm{lbs}_k$
Takedown accuracy	$\mathrm{TDL}_{ijk} \sim \mathrm{Binomial}(\mathrm{TDA}_{ijk}, tda_{ijk})$ $\mathrm{logit}(tda_{ijk}) = tda\_int + tda\_att_i + tda\_def_j + tda\_weight \cdot \mathrm{lbs}_k$
Submission accuracy	$\mathrm{SML}_{ijk} \sim \mathrm{Binomial}(\mathrm{SMA}_{ijk}, sma_{ijk})$ $\mathrm{logit}(sma_{ijk}) = sma\_int + sma\_att_i + sma\_def_j + sma\_weight \cdot \mathrm{lbs}_k$
Knockout or knockdown probability	$\mathrm{KD}_{ijk} + \mathrm{KO}_{ijk} \sim \mathrm{Binomial}(\mathrm{SHL}_{ijk} + \mathrm{GHL}_{ijk}, kdo_{ijk})$ $\mathrm{logit}(kdo_{ijk}) = kdo\_int + kdo\_att_i + kdo\_def_j + kdo\_weight \cdot \mathrm{lbs}_k$
Standing head strike probability	$\mathrm{SHA}_{ijk} \sim \mathrm{Binomial}(\mathrm{SA}_{ijk}, shp_{ijk})$ $\mathrm{logit}(shp_{ijk}) = shp\_int + shp\_att_i + shp\_weight \cdot \mathrm{lbs}_k$
Ground head strike probability	$\mathrm{GHA}_{ijk} \sim \mathrm{Binomial}(\mathrm{GA}_{ijk}, ghp_{ijk})$ $\mathrm{logit}(ghp_{ijk}) = ghp\_int + ghp\_att_i + ghp\_weight \cdot \mathrm{lbs}_k$
Ground control per takedown landed	$\mathrm{GC}_{ijk}/\mathrm{TDL}_{ijk} \sim \mathrm{Gamma}(gc_{ijk}, \phi)$ $\log(gc_{ijk}) = gc\_int + gc\_att_i + gc\_def_j + gc\_weight \cdot \mathrm{lbs}_k$
Stand-up probabilitiy	$stnd_{ijk} = 1/gc_{jik}$

Table A.4. Summary of the variables used in the skill models. Recall from Section 3.2 that all strikes are assumed to be significant, 'standing' refers to distance and clinch, and 'body' refers to body and leg.

Variable	Name
AL	Total strikes landed
C	Total control-time
CC	Clinch control-time
GA	Ground strikes attempted
GBA	Ground body strikes attempted
GBL	Ground body strikes landed
GC	Ground control-time
GHA	Ground head strikes attempted
GHL	Ground head strikes landed
GL	Ground strikes landed
KD	Number of knockdowns inflicted on the opponent
KO	Number of knockouts inflicted on the opponent
lbs	Upper-limit of the weight class for a given contest
SA	Standing strikes attempted
SBA	Standing body strikes attempted
SBL	Standing body strikes landed
SHA	Standing head strikes attempted
SHL	Standing head strikes landed
SMA	Submissions attempted
SML	Submissions landed
SL	Standing strikes landed
ST	Standing-time: time not on ground and opponent not in control
TDA	Takedowns attempted
TDL	Takedowns landed
T	Total bout duration

strike transitions through several states to complete, they take one second. Also, whether a technique is successful or not, it takes one second. This means this whole chain would have lasted eight seconds.

Appendix B

Reputation Bias and Home Crowd Influence in Judging: The Case of Mixed Martial Arts

B.1 Purposeful variable selection

Whilst recent advancements in machine learning algorithms have allowed statisticians to forego manual feature selection, it is still an important stage of model development; particularly for "simple" generalised linear models.

Many authors argue that all variables should remain in the model. A "confounding" variable has an impact on the effect of other variables, despite potentially being non-significant. Including all variables, regardless of significance, controls for confounding as much as possible. However, this produces its own problems, chiefly "overfitting", whereby the model's predictive capacity is hindered. The impact of an unreduced model was apparent in Collier et al. (2012), whose models included several negative effects for positive actions.

A different approach designed to overcome some of the limitations of stepwise selection methods is "purposeful selection", introduced in Hosmer and Lemeshow (2000, Chapter 4.2). The method proceeds as follows:

1. Fit univariable models predicting the target variable with each of the predictors individually. A variable is kept if the p-value of the associated Wald test is below 0.25. Classically used lower thresholds often fail to identify important variables, hence the wider threshold. With a wider threshold, however, the user must review the validity of each variable.

2. Fit a 'larger' model, which consists of all variables kept in Step 1. For each variable

with a p-value greater than 0.10, assess if the variable is necessary by removing it and comparing this smaller model with the larger model by way of a likelihood ratio test.

3. Check for confounding between variables by re-adding removed variables one-by-one, observing the change in the coefficients of the variables which remain in the model. Any removed variable which produces a change of 20% or higher should be added back into the model.

4. Check if adding any variables removed in Step 1 now leads to a significantly improved fit through the likelihood-ratio test.

5. The optimal shape representing a variable's effect on the outcome should now be assessed. The recommended method is through fractional polynomials, and one should evaluate if a polynomial representation of each variable significantly improves the fit.

6. Check for interactions between variables. These interactions should be plausible within the context of the model. If an interaction is found to be significant at 0.05, then it should be included. Repeat Step 2 to simplify the model again, focusing only on removing interaction terms.

7. Assess the adequacy and fit of the final model. If the model passes, then one can begin to make inferences.

The outlined steps allow one to account for any confounding. Variables are also initially screened at a much broader threshold than the standard $p \leq 0.05$. There is also a great emphasis on user input and critical thinking regarding the variables at each stage, something which perhaps has been overlooked in the era of big data and machine learning.

When determining the optimal shape for each variable, we follow the suggested methodology in Hosmer and Lemeshow (2000, Chapter 4.2.1) of using fractional polynomials. To implement this method, one will fit numerous models using various specifications of powers. For a given covariate, x, one can generalise the logit model such that

$$g(x, \beta) = \beta_0 + \sum_{j=1}^{J} \beta_j \times F_j(x),$$

where, for a power, p_j, under the convention that $x^0 \equiv \ln(x)$,

$$F_j(x) = \begin{cases} x^{p_j}, & p_j \neq p_{j-1}, \\ F_{j-1}(x)\ln(x), & p_j = p_{j-1}. \end{cases}$$

Typically, $p \in \{-3, -2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, and one then fits eight models for $J = 1$, and 36 models for $J = 2$. In our case, since there are many zero observations for each covariate, we restrict our powers to $p \in \{0, 0.5, 1, 2, 3\}$. Given this restriction, we estimate models for $J \in \{1, 2, 3\}$. Table B.1 gives some example formulations.

Table B.1. Several fractional polynomial forms for a general covariate x.

J	p_1	p_2	p_3	Polynomial
1	0			$\ln(x)$
1	2			x^2
2	1	2		$x + x^2$
2	3	3		$x^3 + x^3 \ln(x)$
3	1	2	3	$x + x^2 + x^3$
3	2	2	0.5	$x^2 + x^2 \ln(x) + x^{0.5}$

Furthermore, we make a slight adjustment due to the negative and zero values in the dataset. We model what we refer to as the "pseudo-powers", that is,

$$f(x,p) = \begin{cases} x^p & x \geq 0 \ \& \ p \neq 0, \\ -(|x|^p) & x < 0 \ \& \ p \neq 0, \\ \ln(x+1) & x \geq 0 \ \& \ p = 0, \\ -\ln(|x|+1) & x < 0 \ \& \ p = 0. \end{cases}$$

Clearly, squaring negative values directly would not have the desired effect. Adding 1 when taking the log overcomes problems of $\ln(0)$.

Appendix C

Individual Preferences and Controversial Decisions in Mixed Martial Arts Judges

C.1 Non-centered parameterisation

Often, particularly when estimating hierarchical models, Stan can struggle to efficiently sample from the full state-space. A well-documented example is Neal's funnel (Neal, 2003), where the scale of the density changes over the state-space. Consequently, the optimal step-size changes as you move around the density.

In the so-called "centered" parameterisation, one may wish to model

$$\begin{aligned}
\beta &\sim \mathcal{N}(\mu, \sigma), \\
\mu &\sim \mathcal{N}(0, 2.5), \\
\sigma &\sim \text{Half-Normal}(0, 2.5).
\end{aligned}$$

Now, depending on the amount of data available, there will be high correlation in the posterior between β, μ, and σ, thus leading to similar problems as Neal's funnel.

We can remove the dependencies between the parameters and hyper-parameters by parameterising β as a deterministic transformation of μ and σ. To do this, we introduce an offset term $\alpha \sim \mathcal{N}(0, 1)$, such that

$$\beta = \mu + \sigma * \alpha.$$

The remainder of the model is as defined originally.

This line of thinking easily extends to a multivariate prior on β, for instance

$$\beta \sim \mathcal{N}(\mu, \Sigma),$$

where μ is a vector of mean values and Σ a covariance matrix. In this case $\alpha \sim \mathcal{N}(0, 1)$ is a vector of independent identically distributed standard normal variables, such that

$$\beta = \mu + L * \alpha,$$

where $LL^T = \Sigma$ is the Cholesky decomposition of Σ.

The Stan user guide recommends modelling the covariance as a correlation matrix multiplied from both sides by a diagonal matrix of standard deviations. Suppose our covariance matrix is Σ, correlation matrix is Ω, and the standard deviations are denoted by the vector σ. Then, if $LL^T = \Omega$, the Cholesky factor of Σ is equal to $\text{Diag}(\sigma) L \text{Diag}(\sigma)$. Thus, if α is as defined before

$$\beta = \mu + \text{Diag}(\sigma) \cdot (L \cdot \alpha).$$

C.2 Fan model

Here, we explicitly state the fan-score model used in Section 5.3.

$$\begin{aligned}
y_n &\sim \text{Ordered-Logit}(\lambda_n, t) \\
\lambda_n &= \beta x_n \\
t &= (-s_1 - s_2 - s_3, -s_1 - s_2, -s_1, s_1, s_1 + s_2, s_1 + s_2 + s_3) \\
\beta &\sim \mathcal{N}_K(\mu, \Sigma) \\
\Sigma &= \text{Diag}(\tau) \Omega \text{Diag}(\tau) \\
\mu_k &\sim \mathcal{N}(0, 5) \\
\tau_k &\sim \text{Half-Normal}(0, 2) \\
\Omega &\sim \text{LKJ}(2) \\
s_{1,2,3} &\sim \text{Half-Normal}(0, 5).
\end{aligned}$$

For observation i and score s, we then include the proportion of fans who scored the round as s as the observation weight.

Bibliography

Baio, Gianluca and Marta Blangiardo (2010). “Bayesian hierarchical model for the prediction of football results”. In: *Journal of Applied Statistics* 37, pp. 253–264. DOI: 10.1080/02664760802684177.

Baker, Rose and Philip Scarf (2020). “Modifying Bradley–Terry and other ranking models to allow ties”. In: *IMA Journal of Management Mathematics* 32.4, pp. 451–463. DOI: 10.1093/imaman/dpaa027.

Balmer, NJ, AM Nevill, and AM Lane (2005). “Do judges enhance home advantage in European championship boxing?” In: *Journal of sports sciences* 23.4, 409—416. DOI: 10.1080/02640410400021583.

Barnard, John, Robert McCulloch, and Xiao-Li Meng (2000). “Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage”. In: *Statistica Sinica* 10.4, pp. 1281–1311. URL: http://www.jstor.org/stable/24306780.

Bartoš, Mikoláš (2021). “Machine learning in combat sports”. Bachelor’s thesis. Czech Technical University. URL: https://dspace.cvut.cz/bitstream/handle/10467/96672/F3-BP-2021-Bartos-Mikolas-machine_learning_in_combat_sports.pdf.

Blanc, Guy, Eric S Luxenberg, and Stanley C Xie (2016). “NFL score difference prediction with Markov modeling”. Bachelor’s project. Stanford University. URL: https://cs229.stanford.edu/proj2016/report/BlancLuxenbergXie-NFLScoreDifferencePredictionWithMarkovModeling-report.pdf.

Bledsoe, Gregory, Edbert Hsu, Jurek Grabowski, Justin Brill, and Guohua Li (2006). “Incidence of injury in professional mixed martial arts competitions”. In: *Journal of sports science & medicine* 5, pp. 136–42. URL: https://pubmed.ncbi.nlm.nih.gov/24357986.

Boshnakov, Georgi, Tarak Kharrat, and Ian G. McHale (2017). “A bivariate Weibull count model for forecasting association football scores”. In: *International Journal of Forecasting* 33.2, pp. 458–466. DOI: 10.1016/j.ijforecast.2016.11.006.

Bradley, Ralph Allan and Milton E. Terry (1952). “Rank analysis of incomplete block designs: I. the method of paired comparisons”. In: 39.3-4, pp. 324–345. DOI: 10.1093/biomet/39.3-4.324.

Brown, Alasdair and J. James Reade (2019). "The wisdom of amateur crowds: Evidence from an online community of sports tipsters". In: *European Journal of Operational Research* 272.3, pp. 1073–1081. DOI: 10.1016/j.ejor.2018.07.015.

Bukiet, Bruce, Elliotte Rusty Harold, and José Luis Palacios (1997). "A Markov chain approach to baseball". In: *Operations Research* 45.1, pp. 14–23. DOI: 10.1287/opre.45.1.14.

Bunn, D. W. (1975). "Anchoring bias in the assessment of subjective probability". In: *Journal of the Operational Research Society* 26.2, pp. 449–454. DOI: 10.1057/jors.1975.94.

Buraimo, Babatunde, David Forrest, Ian G. McHale, and J.D. Tena (2022). "Armchair fans: Modelling audience size for televised football matches". In: *European Journal of Operational Research* 298.2, pp. 644–655. DOI: 10.1016/j.ejor.2021.06.046.

Buraimo, Babatunde, David Forrest, and Robert Simmons (2010). "The 12th man?: Refereeing bias in English and German soccer". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173.2, pp. 431–449. DOI: 10.1111/j.1467-985X.2009.00604.x.

California State Athletic Commission (2020). *Unified rules of mixed martial arts*. URL: https://www.abcboxing.com/wp-content/uploads/2020/02/unified-rules-mma-2019.pdf.

Campbell, Bryan and John W. Galbraith (1996). "Nonparametric tests of the unbiasedness of Olympic figure-skating judgments". In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 45.4, pp. 521–526. DOI: 10.2307/2988550.

Capers, Quinn (2020). "How clinicians and educators can mitigate implicit bias in patient care and candidate selection in medical education". In: *ATS Scholar* 1. DOI: 10.34197/ats-scholar.2020-0024PS.

Caron, Francois and Arnaud Doucet (2010). "Efficient Bayesian inference for generalized Bradley-Terry models". In: *Journal of Computational and Graphical Statistics* 21. DOI: 10.1080/10618600.2012.638220.

Carter, Walter H. and Sharon L. Crews (1974). "An analysis of the game of tennis". In: *The American Statistician* 28.4, pp. 130–134. DOI: 10.2307/2683337.

Cervone, Daniel, Alex D'Amour, Luke Bornn, and Kirk Goldsberry (2016). "A multiresolution stochastic process model for predicting basketball possession outcomes". In: *Journal of the American Statistical Association* 111.514, pp. 585–599. DOI: 10.1080/01621459.2016.1141685.

Collier, Trevor, Andrew L. Johnson, and John Ruggiero (2012). "Aggression in mixed martial arts: An analysis of the likelihood of winning a decision". In: *Violence and Aggression in Sporting Contests: Economics, History and Policy*. New York, NY: Springer New York, pp. 97–109. DOI: 10.1007/978-1-4419-6630-8_7.

Constantinou, Anthony Costa and Norman Elliott Fenton (2013). “Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries”. In: *Journal of Quantitative Analysis in Sports* 9.1, pp. 37–50. DOI: 10.1515/jqas-2012-0036.

Crighton, Ben, Graeme L Close, and James P Morton (2016). “Alarming weight cutting behaviours in mixed martial arts: a cause for concern and a call for action”. In: *British Journal of Sports Medicine* 50.8, pp. 446–447. DOI: 10.1136/bjsports-2015-094732.

Damour, G. and P. Lang (2015). “Modelling Football as a Markov Process”. Master’s thesis. Stockholm, Sweden: Royal Institute of Technology. URL: https://www.diva-portal.org/smash/get/diva2:828101/FULLTEXT01.pdf.

Davidson, Roger R. and Robert J. Beaver (1977). “On extending the Bradley-Terry model to incorporate within-pair order effects”. In: *Biometrics* 33.4, pp. 693–702. DOI: 10.2307/2529467.

Decroos, Tom, Lotte Bransen, Jan Van Haaren, and Jesse Davis (2019). “Actions Speak Louder than Goals”. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery &amp Data Mining*. DOI: 10.1145/3292500.3330758.

Detotto, Claudio, Dimitri Paolini, and J. D. Tena (2018). “Do managerial skills matter? An analysis of the impact of managerial features on performance for Italian football”. In: *Journal of the Operational Research Society* 69.2, pp. 270–282. DOI: 10.1057/s41274-017-0215-6.

Dixon, Mark J. and Stuart G. Coles (1997). “Modelling Association Football Scores and Inefficiencies in the Football Betting Market”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 46.2, pp. 265–280. DOI: 10.1111/1467-9876.00065.

Elo, Arpad E. (1978). *The Rating of Chessplayers, Past and Present*. New York: Arco Pub.

Emerson, John W., Miki Seltzer, and David Lin (2009). “Assessing Judging Bias: An Example From the 2000 Olympic Games”. In: *The American Statistician* 63.2, pp. 124–131. DOI: 10.1198/tast.2009.0026.

Erikstad, Martin and Bjørn Johansen (2020). “Referee bias in professional football: Favoritism toward successful teams in potential penalty situations”. In: *Frontiers in Sports and Active Living* 2.19. DOI: 10.3389/fspor.2020.00019.

Fahrmeir, Ludwig and Gerhard Tutz (1994). “Dynamic stochastic models for time-dependent ordered paired comparison systems”. In: *Journal of the American Statistical Association* 89.428, pp. 1438–1449. DOI: 10.1080/01621459.1994.10476882.

Feldman, Todd (2020). “The way of the fight: An analysis of MMA judging”. In: *Journal of Applied Sport Management* 12.2. DOI: 10.7290/jasm120205.

Findlay, Leanne and Diane Ste-Marie (2004). “A Reputation Bias in Figure Skating Judging”. In: *Journal of Sport and Exercise Psychology* 26.1, pp. 154–166. DOI: 10.1123/jsep.26.1.154.

Fitt, Alistair D. (2008). “Markowitz portfolio theory for soccer spread betting”. In: *IMA Journal of Management Mathematics* 20.2, pp. 167–184. DOI: 10.1093/imaman/dpn028.

Franchini, Emerson and Monica Takito (2016). “Home advantage in combat sports during the Olympic Games”. In: *Sport Sciences for Health* 12, pp. 287–290. DOI: 10.1007/s11332-016-0286-9.

Frederiksen, Jesper S. and Robert E. Machol (1988). “Reduction of paradoxes in subjectively judged competitions”. In: *European Journal of Operational Research* 35.1, pp. 16–29. DOI: https://doi.org/10.1016/0377-2217(88)90375-X.

Garicano, Luis, Ignacio Palacios-Huerta, and Canice Prendergast (2005). “Favoritism under social pressure”. In: *The Review of Economics and Statistics* 87.2, pp. 208–216. DOI: 10.1162/0034653053970267.

Gelman, Andrew and Jennifer Hill (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press. DOI: 10.1017/CBO9780511790942.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su (2008). “A weakly informative default prior distribution for logistic and other regression models”. In: *The Annals of Applied Statistics* 2.4, pp. 1360–1383. DOI: 10.1214/08-aoas191.

Gelman, Andrew and Yu-Sung Su (2018). *ARM: Data Analysis Using Regression and Multilevel/Hierarchical Models*. R package version 1.10-1. URL: https://CRAN.R-project.org/package=arm.

Gift, Paul (2018). “Performance evaluation and favoritism: Evidence from mixed martial arts”. In: *Journal of Sports Economics* 19.8, pp. 1147–1173. DOI: 10.1177/1527002517702422.

Glickman, Mark E (1995). *The Glicko System*. URL: http://www.glicko.net/glicko/glicko.pdf.

Goldner, Keith (2012). “A Markov model of football: Using stochastic processes to model a football drive”. In: *Journal of Quantitative Analysis in Sports* 8.1. DOI: 10.1515/1559-0410.1400.

Haave, H. S. and H. Hoiland (2017). “Evaluating association football player performances using Markov models”. Master’s thesis. Trondheim, Norway: Norwegian University of Science and Technology. URL: https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2469351/16981_FULLTEXT.pdf?sequence=1.

Harrison, John (2020). *RSelenium: R Bindings for 'Selenium WebDriver'*. R package version 1.7.7. URL: https://CRAN.R-project.org/package=RSelenium.

Heiniger, Sandro and Hugues Mercier (2018). "National Bias of International Gymnastics Judges during the 2013-2016 Olympic Cycle". Unpublished. DOI: 10.48550/ARXIV.1807.10033.

Heiniger, Sandro and Hugues Mercier (2021). "Judging the judges: Evaluating the accuracy and national bias of international gymnastics judges". In: *Journal of Quantitative Analysis in Sports* 17.4, pp. 289–305. DOI: 10.1515/jqas-2019-0113.

Herbrich, Ralf, Tom Minka, and Thore Graepel (2007). "TrueSkill(TM): A Bayesian Skill Rating System". In: *Advances in Neural Information Processing Systems 20*. MIT Press, pp. 569–576. URL: https://www.microsoft.com/en-us/research/publication/trueskilltm-a-bayesian-skill-rating-system/.

Hirotsu, N and M Wright (2002). "Using a Markov process model of an association football match to determine the optimal timing of substitution and tactical decisions". In: *Journal of the Operational Research Society* 53.1, pp. 88–96. DOI: 10.1057/palgrave.jors.2601254.

Hirotsu, Nobuyoshi and Mike Wright (2003). "An evaluation of characteristics of teams in association football by using a Markov process model". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 52.4, pp. 591–602. DOI: 10.1046/j.0039-0526.2003.00437.x.

Hitkul, Karmanya Aggarwal, Neha Yadav, and Maheshwar Dwivedy (2019). "A comparative study of machine learning algorithms for prior prediction of UFC fights". In: *Harmony Search and Nature Inspired Optimization Algorithms*. Vol. 741. Singapore: Springer Singapore, pp. 67–76. DOI: 10.1007/978-981-13-0761-4_7.

Ho, Christopher (2013). "Does MMA math work? A study on sports prediction applied to mixed martial arts". Undergraduate project. Stanford University. URL: https://cs229.stanford.edu/proj2013/Ho-DoesMMAMathWorkAStudyonSportsPredictionAppliedtoMixedMartialArts.pdf.

Holmes, Benjamin, Ian G. McHale, and Kamila Żychaluk (2022). "A Markov chain model for forecasting results of mixed martial arts contests". In: *International Journal of Forecasting*. DOI: 10.1016/j.ijforecast.2022.01.007.

Hosmer, David W. and Stanley Lemeshow (2000). *Applied logistic regression*. Wiley Series in Probability and Statistics. John Wiley and Sons. DOI: 10.1002/9781118548387.

Hubbard, Ryan, Gene Stringer, Ken Peterson, Mario Roberto Filho Vaz Carneiro, Jonathan T. Finnoff, and Rodolfo Savica (2019). "The King-Devick test in mixed martial arts:

The immediate consequences of knock-outs, technical knock-outs, and chokes on brain functions". In: *Brain Injury* 33.3, pp. 349–354. DOI: 10.1080/02699052.2018.1553068.

Hubáček, Ondřej, Gustav Šourek, and Filip Železný (2018). "Learning to predict soccer results from relational data with gradient boosted trees". In: *Machine Learning* 108.1, pp. 29–47. DOI: 10.1007/s10994-018-5704-6.

Hubáček, Ondřej, Gustav Šourek, and Filip Železný (2019). "Exploiting sports-betting market using machine learning". In: *International Journal of Forecasting* 35.2, pp. 783–796. DOI: 10.1016/j.ijforecast.2019.01.001.

Johnson, Jeremiah Douglas (2012). "Predicting outcomes of mixed martial arts fights with novel fight variables". Master's thesis. Athens, Georgia, USA: University of Georgia. URL: https://getd.libs.uga.edu/pdfs/johnson_jeremiah_d_201208_ms.pdf.

Jones, Marc and John Erskine (2003). "The impact of a team's aggressive reputation on the decisions of Association football referees". In: *Journal of sports sciences* 20.12, pp. 991–1000. DOI: 10.1080/026404102321011751.

Kaye, Ella and David Firth (2021a). *BradleyTerryScalable: Fits the Bradley-Terry Model to Potentially Large and Sparse Networks of Comparison Data.* R package version 0.1.0.9200. URL: https://github.com/EllaKaye/BradleyTerryScalable.

Kaye, Ella and David Firth (2021b). *btfit: Fits the Bradley-Terry model.* URL: https://ellakaye.github.io/BradleyTerryScalable/reference/btfit.html.

Kelly, J. L. (1956). "A new interpretation of information rate". In: *The Bell System Technical Journal* 35.4, pp. 917–926. DOI: 10.1002/j.1538-7305.1956.tb03809.x.

Klaassen, Franc and Jan Magnus (2001). "Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model". In: *Journal of the American Statistical Association* 96, pp. 500–509. DOI: 10.1198/016214501753168217.

Kleinrock, Leonard. (1975). *Queueing systems.* New York: Wiley. ISBN: 0471491101.

Koopman, Siem Jan and Rutger Lit (2015). "A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League". In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 178.1, pp. 167–186. DOI: 10.2139/ssrn.2154792.

Kostrzewa, Maciej, Radosław Laskowski, Michal Wilk, Wiesław Błach, Angelina Ignatjeva, and Magdalena Nitychoruk (2020). "Significant predictors of sports performance in elite men judo athletes based on multidimensional regression models". In: *International Journal of Environmental Research and Public Health* 17.21. DOI: 10.3390/ijerph17218192.

Lee, Herbert K. H., Daniel L. Cork, and David J. Algranati (2002). "Did Lennox Lewis beat Evander Holyfield?: Methods for analysing small sample interrater agreement problems".

In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 51.2, pp. 129–146. DOI: 10.1111/1467-9884.00306.

Liu, Guiliang, Yudong Luo, Oliver Schulte, and Tarak Kharrat (2020). "Deep soccer analytics: Learning an action-value function for evaluating soccer players". In: *Data Mining and Knowledge Discovery* 34.5, pp. 1531–1559. DOI: 10.1007/s10618-020-00705-9.

Lockwood, Joel, Liam Frape, Steve Lin, and Alun Ackery (2018). "Traumatic brain injuries in mixed martial arts: A systematic review". In: *Trauma* 20.4, pp. 245–254. DOI: 10.1177/1460408617740902.

Luce, R.Duncan (1977). "The choice axiom after twenty years". In: *Journal of Mathematical Psychology* 15.3, pp. 215–233. DOI: 10.1016/0022-2496(77)90032-3.

Maher, M. J. (1982). "Modelling association football scores". In: *Statistica Neerlandica* 36.3, pp. 109–118. DOI: 10.1111/j.1467-9574.1982..

Markowitz, Harry (1952). "Portfolio selection". In: *The Journal of Finance* 7.1, pp. 77–91. DOI: 10.2307/2975974.

Matej, Uhrín, Šourek Gustav, Hubáček Ondřej, and Železný Filip (2021). "Optimal sports betting strategies in practice: An experimental review". In: *IMA Journal of Management Mathematics* 32.4, pp. 465–489. DOI: 10.1093/imaman/dpaa029.

McHale, Ian and Phil Scarf (2011). "Modelling the dependence of goals scored by opposing teams in international soccer matches". In: *Statistical Modelling* 11.3, pp. 219–236. DOI: 10.1177/1471082X1001100303.

Morley, Bruce and Dennis Thomas (2005). "An investigation of home advantage and other factors affecting outcomes in English one-day cricket matches". In: *Journal of Sports Sciences* 23.3, pp. 261–268. DOI: 10.1080/02640410410001730133.

Myers, Tony, Nigel Balmer, Alan Nevill, and Yahya Al-Nakeeb (2006). "Evidence of nationalistic bias in muay Thai". In: *Journal of Sports Science & Medicine* 5, pp. 21–7. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3863918.

Myers, Tony, Alan Nevill, and Yahya Al-Nakeeb (2012). "The influence of crowd noise upon judging decisions in muay Thai". In: *Advances in Physical Education* 2, pp. 148–152. DOI: 10.4236/ape.2012.24026.

Neal, Radford M. (2003). "Slice sampling". In: *The Annals of Statistics* 31.3, pp. 705–767. DOI: 10.1214/aos/1056562461.

Nevill, Alan and R Holder (1999). "Home advantage in sport: An overview of studies on the advantage of playing at home". In: *Sports Medicine* 28, pp. 221–236. DOI: 10.2165/00007256-199928040-00001.

Nevill, A.M, N.J Balmer, and A Mark Williams (2002). “The influence of crowd noise and experience upon refereeing decisions in football”. In: *Psychology of Sport and Exercise* 3.4, pp. 261–272. DOI: 10.1016/S1469-0292(01)00033-4.

Newton, Paul K. and Joseph B. Keller (2005). “Probability of winning at tennis I. Theory and data”. In: *Studies in Applied Mathematics* 114.3, pp. 241–269. DOI: 10.1111/j.0022-2526.2005.01547.x.

O’Malley, A. James (2008). “Probability formulas and statistical analysis in tennis”. In: *Journal of Quantitative Analysis in Sports* 4.2. DOI: 10.2202/1559-0410.1100.

Parsons, Christopher, Johan Sulaeman, Michael Yates, and Daniel Hamermesh (2011). “Strike three: Umpires’ demand for discrimination”. In: *American Economic Review* 101.4, pp. 1410–1435. DOI: 10.1257/aer.101.4.1410.

Pettersson-Lidbom, Per and Mikael Priks (2010). “Behavior under social pressure: Empty Italian stadiums and referee bias”. In: *Economics Letters* 108.2, pp. 212–214. DOI: 10.1016/j.econlet.2010.04.023.

Plessner, H. (1999). “Expectation biases in gymnastics judging”. In: *Journal of Sport & Exercise Psychology* 21.2, pp. 131–144. DOI: 10.1123/jsep.21.2.131.

Podrigalo, L.V., A.A. Volodchenko, O.A. Rovnaya, O.V. Podavalenko, and T.I. Grynova (2018). “The prediction of success in kickboxing based on the analysis of morphofunctional, physiological, biomechanical and psychophysiological indicators”. In: *Physical education of students* 22, p. 51. DOI: 10.15561/20755279.2018.0108.

Price, Joseph and Justin Wolfers (2010). “Racial discrimination among NBA referees”. In: *The Quarterly Journal of Economics* 125.4, pp. 1859–1887. DOI: 10.1162/qjec.2010.125.4.1859.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Reade, J James, Dominik Schreyer, and Carl Singleton (2020). “Echoes: What happens when football is played behind closed doors?” Unpublished. DOI: 10.2139/ssrn.3630130.

Robles, I. and J. Wu (2019). Undergraduate project. Stanford, California, USA: Stanford University. URL: http://cs229.stanford.edu/proj2015/121_report.pdf.

Rudd, Sarah (2011). “A framework for tactical analysis and individual offensive production assessment in soccer using Markov chains”. In: *New England Symposium for Statistics in Sport*. Conference presentation. URL: http://nessis.org/nessis11/rudd.pdf.

Rue, Håvard and Øyvind Salvesen (2000). “Prediction and retrospective analysis of soccer matches in a league”. In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 49.3, pp. 399–418. DOI: 10.1111/1467-9884.00243.

Sadowski, Jerzy, Dariusz Gierczuk, Jerzy Miller, Igor Cieśliński, and Mariusz Buszta (2012). "Success factors in male WTF taekwondo juniors". In: *Journal of Combat Sports and Martial Arts* 3, pp. 47–51. DOI: 10.5604/20815735.1047647.

Scheer, John K. and Charles J. Ansorge (1975). "Effects of naturally induced judges' expectations on the ratings of physical performances". In: *Research Quarterly. American Alliance for Health, Physical Education and Recreation* 46.4, pp. 463–470. DOI: 10.1080/10671315.1975.10616704.

Scheer, John K. and Charles J. Ansorge (1979). "Influence due to expectations of judges: A function of internal-external locus of control". In: *Journal of Sport Psychology* 1.1, pp. 53–58. DOI: 10.1123/jsp.1.1.53.

Schulte, Oliver, Mahmoud Khademi, Sajjad Gholami, Zeyu Zhao, Mehrsan Javan Roshtkhari, and Philippe Desaulniers (2017). "A Markov game model for valuing actions, locations, and team performance in ice hockey". In: *Data Mining and Knowledge Discovery* 31, 1735–1757. DOI: 10.1007/s10618-017-0496-z.

Schwartz, Barry and Stephen Barsky (1977). "The home advantage". In: *Social Forces* 55.3, pp. 641–661. DOI: 10.1093/sf/55.3.641.

Shirley, Kenny (2007). "A Markov model for basketball". In: *New England Symposium for Statistics in Sport*. Poster presentation. URL: https://www.nessis.org/nessis07/Kenny_Shirley.pdf.

Sklar, M. (1959). *Fonctions de Répartition À N Dimensions Et Leurs Marges*. Université Paris 8. URL: https://books.google.co.uk/books?id=nreSmAEACAAJ.

Smith, Jordan T. (2009). "Fighting for regulation: Mixed martial arts legislation in the United States". In: *Drake Law Review* 58, pp. 617–655. URL: https://lawreviewdrake.files.wordpress.com/2015/06/irvol58-2_smith2.pdf.

Stan Development Team (2021a). *Stan Modeling Language Users Guide and Reference Manual*. URL: https://mc-stan.org.

Stan Development Team (2021b). *Stan User's Guide*. URL: https://mc-stan.org/docs/2_27/stan-users-guide-2_27.pdf.

Štrumbelj, Erik and Petar Vračar (2012). "Simulating a basketball match with a homogeneous Markov model and forecasting the outcome". In: *International Journal of Forecasting* 28.2, pp. 532–542. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2011.01.004.

Sutter, Matthias and Martin G Kocher (2004). "Favoritism of agents – The case of referees' home bias". In: *Journal of Economic Psychology* 25.4, pp. 461–469. DOI: 10.1016/S0167-4870(03)00013-8.

Szczepanski, L. (2015). "Assessing the skill of football players using statistical methods". PhD thesis. Greater Manchester, England: University of Salford. URL: http://usir.salford.ac.uk/id/eprint/34027/2/thesis.pdf.

Szczepański, Łukasz and Ian McHale (2016). "Beyond completion rate: Evaluating the passing ability of footballers". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179.2, pp. 513–533. DOI: 10.1111/rssa.12115.

Sæbø, Olav and Lars Magnus Hvattum (2018). "Modelling the financial contribution of soccer players to their clubs". In: *Journal of Sports Analytics* 5.1, pp. 1–12. DOI: 10.3233/JSA-170235.

Tiernan, Stephen, Aidan Meagher, David O'Sullivan, Eoin O'Keeffe, Eoin Kelly, Eugene Wallace, Colin Doherty, Matthew Campbell, Yuzhe Liu, and August Domel (2020). "Concussion and the severity of head impacts in mixed martial arts". In: *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine* 234.12, pp. 1472–1483. DOI: 10.1177/0954411920947850.

Warnick, Jason and Kyla Warnick (2007). "Specification of variables predictive of victories in the sport of boxing". In: *Perceptual and Motor Skills* 105.1, pp. 153–8. DOI: 10.2466/pms.105.1.153-158.

Warnick, Jason E. and Kyla Warnick (2009). "Specification of variables predictive of victories in the sport of boxing: II. Further characterization of previous success". In: *Perceptual and Motor Skills* 108.1, pp. 137–138. DOI: 10.2466/pms.108.1.137-138.

Wickham, Hadley (2020). *rvest: Easily Harvest (Scrape) Web Pages*. R package version 0.3.6. URL: https://CRAN.R-project.org/package=rvest.

Zitzewitz, Eric (2006). "Nationalism in winter sports judging and its lessons for organizational decision making". In: *Journal of Economics & Management Strategy* 15.1, pp. 67–99. DOI: https://doi.org/10.1111/j.1530-9134.2006.00092.x.

Zitzewitz, Eric (2014). "Does transparency reduce favoritism and corruption? Evidence from the reform of figure skating judging". In: *Journal of Sports Economics* 15.1, pp. 3–30. DOI: 10.1177/1527002512441479.