Slay the Dragon, a Study of Upsets in International Rugby Union[1, p. 19]

Arthur, Guillaume (N11371200)  
IFN703 Assessment 1

# Abstract [1, Ch. 3]

This is a template for IFN703/4 Assessment 1, a *“Written project plan providing a description of the background, aim, significance, expected deliverables and development schedule of the project.”*

# Introduction [1, p. 28]

The continuous increase in computer capabilities has brought data analytics as a key companion for decision-makers in all areas of business and society. In the sports landscape, the study of on-the-field action reflects an increase in the adoption of data analysis. Sports have an important quantitative aspect. Be it in team or individual events, games are won or lost based on scores. Actions such as goals, tackles, and distance covered are all quantitative measures that help describe a game and the performance of a team or a player.

Data analytics takes such measurements and derives qualitative information that can then help inform game decisions. Instead of relying on general wisdom, it is now more common to apply the scientific method to gain a deeper understanding of a game. Data collection and analysis are useful tools that can be used to inform decisions by multiple stakeholders, including coaches, trainers, managers, or even players.

The enthusiasm for sports analytics has been building since 1950 and emerged from distinct communities: operations researchers, freelance sports journalists, and internet hobbyists [2]. "Operations Research" was a new branch of warfare in which statisticians analysed military operations to improve their efficiency. And what better thing could those OR officers do to relax than apply their knowledge to various sports? A 1959 article from a Royal Canadian Air Force officer, for example, described how right-handed batters had a higher batting average against left-handed pitchers, and vice versa [3]. He naturally suggested that managers should substitute a player depending on the handedness of the opposing pitcher.

Baseball focused a lot of interest from statisticians in the 1970s in the US through, not exclusively, the Society for Baseball Research founded in 1971, and the movement never lost momentum. In the nineties, it was the turn of American football with the publication of the pivotal work "The Hidden Game of Football" in 1988. This work illustrates the creation and use of performance metrics, such as expected points depending on the position on the field.

Booming in North America, the rest of the world was not ignorant. Charles Reep, a former wing-commander in Britain’s Royal Air Force, dismayed by the poor performance of his beloved Swindon team, started recording game actions on a notepad. This led to the study of more than 578 soccer games between 1953 and 1967 and can be seen as the spark of game analysis in football.

Rugby is a close-contact team "invasion" sport, in which teams try to infiltrate the opposition’s territory to score a goal. A team is composed of 15 players, using an oval-shaped ball on a rectangular field. Born in 1845 in England, it was split into rugby league and rugby union in 1895. Contrary to its counterpart, there is no limit to the amount of time a team can keep possession of the ball. Rugby Union remained an amateur sport until 1995, when the International Rugby Board declared it a professional sport.

Compared to the other sports mentioned previously, rugby is a relatively new federation. Since then, it has garnered interest worldwide, which has increased the need for performance analysis and game modeling.

Notational analyses refer to the identification of patterns and events in a match. The aim is to identify strengths and weaknesses by notating numerous events that take place in a field. At the team level, this form of analysis can help describe collective behavior and outcomes.

Modeling game outcomes is an increasingly popular endeavor, whether for managers or betting companies. Consumers are gaining wider access to sports betting, which requires increasingly complex models to offer more intricate bets. However, little work has been done in rugby union to model game outcomes. This can be attributed to the lower amount of data available compared to other sports, such as soccer.

## Literature Review

In Rugby, various approaches have been utilized to study the relationships between performance indicators and outcomes. Performance indicators are quantifiable metrics used to assess performance of a player or team. In this review, our aim is twofold: first, to outline the research done exploring the relationship between one or multiple performance indicator and a positive outcome.

In a second time, we will compare the statistical models employed in rugby union with those used in other sports to predict match outcomes. Since we are given in game commentary, we will review how neural networks have been used to study sequences of plays.

Our objective is to provide an overview of data analysis in rugby by showcasing results and highlighting potential unexplored paths.

Part one: Performance Indicators

The research in rugby union has primarily focused on single or combinations of performance indicators deemed relevant to team success, such as tackle success, while other studies have concentrated on physical and technical requirements, such as running. These studies often assume a linear, causal, and direct relationship between the performance indicator and the outcome. However, there appears to be little implementation of the conclusions drawn from these studies in practice by coaches, which may indicate a potential reluctance to adopt performance analyses, or more realistically, a lack of strength and consensus in the observations made [4].

Research is often conducted without sufficient consideration of the applicability and utility of the findings. Rugby performance is complex and requires coordination between players as well as individual excellence. In a recent literature review published by Colomer et al. [4], it was noted that there is a lack of data available from the period between 1997 and 2019; only 41 studies were included after screening. Most of the data were collected from games played between 2000 and 2008, and various levels of play were studied. These factors make it challenging to reach a definitive consensus and raise doubts about tactical analyses, especially considering that sports strategies are rapidly evolving.

Out of the 41 studies, only 22 contextualized the data, and a mere 5 utilized multiple variables, highlighting once again the lack of complex analyses. Most performance indicators could be classified as attacking. Overall, it is observed that little has changed over the covered time span regarding performance indicator analyses, with most studies relying on univariate measures of performance.

However, it was suggested that match outcomes were better predicted by relative datasets. Bennett et al.[5] described a model consisting of 10 performance indicators that have a significant relationship with match outcomes: kicks from hand, clean breaks, average carry distance, penalties conceded when the opposition has the ball, turnovers conceded, total meters carried, defenders beaten, and the ratio of tackles. Their models only predicted the results of the premier with 80% accuracy at best, suggesting that a good proportion of the complexity is not captured using these performance indicators.[4]

Part two: overview of technique used for performance Analyses

Analysts can measure specific technical skills in each sport. These measurements help decision-makers understand different aspects of play. Further application of notational analysis can elucidate collective behavior, manifested in common styles of play, which can then model relative outcomes like match success. It was noted by Lago-Peñas et al.[6] and argued that this information can be used by coaches to counter opposition. However, the amount of data being collected is enormous, and no consensual framework of analysis exists. How do we interpret the information we receive? Multiple techniques can be employed [7].

The first step is to reduce the amount of data treated by selecting meaningful predictors and filters. In this process, large multidimensional datasets are condensed into smaller sets without losing information. Techniques commonly used in sports include Principal Component Analysis (PCA)and multidimensional scaling, both of which produce factors representing groups of similar variables and have been used to study performance indicators. Clustering, specifically two-step clustering, can also be utilized to reveal 'natural' groupings in a dataset, as seen in rugby league to establish groups of players for performance assessment.

Although these techniques help us understand successful performance, coaches and analysts are more interested in which features can be modified to achieve better outcomes. Decision support analysis helps interveners identify interactions and factors, aiming to determine the probabilities of certain outcomes.

Modeling the probability of events in a team match is an area of vivid interest. Being able to account for multiple variables is a powerful tool. Logistic regression is used to predict dichotomous events (two possible outcomes). A major advantage of regression is that it provides the direction and strength of interaction between the variables and the outcome event. Parmar et al. [8] demonstrated how to combine PCA and logistic regression to model the probability of success in rugby league, noting that a team had a 91% chance of winning if outperforming its opponent on certain metrics [7].

Part Three: Modelling game outcome in Sport

The method described above for modeling the complex game of rugby is just a small sample of what is available to statisticians. Attempts to model rugby union are rare and challenging. Multiple modeling techniques can be employed, including statistical models, simulation systems, and artificial neural networks. Analyzing the model parameters as well as the model's use provides insights into the strengths and limitations of this approach.

O’Donoghue et al.[9] compared different linear regression models to predict the results of the Rugby World Cup 2015 matches and ranked the models based on the percentage of simulations yielding accurate results over the 48 matches of the tournament. Results showed that incorporating the results of previous World Cups to represent relative team strength improved accuracy, while separating pool and knockout games had no impact. However, it is noteworthy that the best accuracy achieved was 74.4%, which is lower than the results from a similar model published in 2004 on the games of WCR 2003, with an accuracy of 92.7%, suggesting that these tournaments have become harder to predict. Overall, statistical models may not be the best choice.

Rugby Union is a sequential game, and examining the sequence of play may yield important information for decision-makers. Decision trees are commonly used because they can be easily understood by coaches when kept simple and offer heuristics easy to follow. CHAID (Chi-squared Automatic Interaction Detection), for example, has been used to model non-binary outcomes, such as phases of play. It was used to predict the outcome of closely contested games during COVID-19 in different elite leagues and study the impact of playing at home [10].

Artificial Neural Networks (ANN) are intriguing due to their capacity to model non-linear behavior. Watson et al. used action sequence descriptions to predict the outcome of sequences of play, based on ordered action sequences and where these actions occur. However, as reported by the authors, a common argument against the use of neural networks in sport performance analysis is the challenge in making actionable recommendations since we don’t know which parameters are used. To address this issue, it could be interesting to integrate Operations Research and Machine Learning. In the previous study, the trained model was used to predict the probabilities of hypothetical sequences of actions. Predictions were visualized on a heat map. This method would enable stakeholders to assess any sequence of play [8].

It is exciting for performance analysts to work in rugby, as the field is wide open and the challenge of such a complex sport is enticing for all aspiring data analysts. In this project, we hope to shed light on a small corner of that sport.

International competition is a different beast compared to national leagues. The tournament format makes it harder for the favorite teams to prevail in the end; the knockout format makes upsets more likely. The winning team needs to finish among the best teams of its group in the knockout stage. The five teams of each group play only four games at that stage. Three direct elimination games follow before the best team can be crowned champion. Gaining a tactical advantage against a specific opponent therefore holds a lot of value.

Exploration of tactical performance has been conducted qualitatively, and three Rugby World Cup coaches reported unsuccessful performance connected with an irrational kicking game, weak defensive line, and losing possession of set-piece play [11]. Vaz et al. hypothesized that in a close game in the Super Twelve, a kicking-based game supported by an effective defensive structure is more likely to win than a possession-based one [12]. These studies hint at the necessity of a better understanding of tactical analysis in Rugby Union. We will therefore try to develop an approach to use whole-game and in-game performance indicators to study if tactics are detectable in international rugby, and how this knowledge can be used by coaches of weaker teams to help progress in a knockout tournament.

How to slay the giants? The overarching try of this study is to look at whole game data and in-game play patterns to investigate if certain tactics are more efficient at creating upset.

# The Project

## Definition

Our goal will be to look for identifiable playing styles in the Rugby League World Cup 2023. Firstly, we will examine the whole game data for common behaviors. Secondly, we will study the game description to characterize what makes for a scoring sequence, considering different styles of play.

The primary objective of this project is to illustrate how multivariate analysis and neural networks can be used in conjunction to assist coaches in their in-game decision-making. Reviewing existing literature, we can conclude that a common flaw in many studies is focusing solely on the relationships between single performance indicators and match outcomes. Such an approach often lacks depth and fails to capture the nuanced complexities of the game.

In international rugby, a sport characterized by multiple tiers and significant disparities in team strength, defining distinct team styles of play may provide insights into whether certain tactics confer advantages over others. **Our first question aims to investigate if specific strategies offer a better chance of victory in mismatched games at the international level**. Understanding this could greatly aid coaches of weaker teams in devising approaches to face stronger opponents.

While assisting in pregame preparations is undoubtedly valuable, providing real-time insights during critical moments of play could be equally vital. While much research has focused on modeling expected points based primarily on field position, we advocate for extending this approach to consider outcomes beyond scoring. Helping at crucial junctures on the field regarding the likelihood of various outcomes could empower coaches to make more precise decisions.

The nonlinear nature of rugby poses challenges to this analysis, and neural networks seem an appropriate approach. **As a follow-up to the first question, we aim to investigate whether certain sequences of play are more likely to result in a try for weaker teams**. This inquiry complements our initial question and could provide valuable insights into whether different styles of play exhibit distinct patterns that are more conducive to scoring.

Our analysis will focus on possessions gained in the opposition half, and through simulation, we will evaluate the probabilities of success for different sequences of play, considering factors such as rucks and passes. A significant challenge will be determining the most effective way to present these results comprehensively. Diagrams illustrating probabilities of success may not suffice; therefore, identifying the optimal representation is a key consideration at this stage.

## Plan

### Performance Indicators Study

In rugby, the athletic makeup of a team dictates how it should play. We could even suggest that the style is embedded in the team's identity for certain countries. Do some teams play a more pass-heavy game? Do others put a greater emphasis on good defense and avoiding penalties? Do others try to kick the opponent to death? Here, the amount of available data will be key.

To isolate potential tactical choices, we could sample games where the score is close (2 possessions, or less than 14 points). Keeping only the close games might enable us to dissociate tactic and sheer strength. The downside of this approach is the possible low number of games available matching that condition. Moreover, we potentially remove games where the tactical advantage was so overwhelming that it resulted in a blowout. We will instead analyze the higher number of games available, in a given period of time (the last 3-4 years?) to be sure that samples are related. A long period of time means more occasion for changes in styles and personnel.

Are some styles of play more likely to cause upsets. For this, using an Elo scoring to estimate the strength difference between teams, we could look at the games where the winner had a far greater chance of losing, and look for patterns. An ELO style scoring system has as constraint that a 400 points difference means the better rated team wins 10 games out of 11 in a confrontation. We will select a difference in score that will enable us to have a large enough sample, maybe between 200 and 400.

Next, we’ll use a principal component analysis to reduce the amount of data and determine the number of variables that should be kept. Can the teams be defined along a defense and attack axis? Rojas-Valverde et al. [13] found 6 studies published between 2000 and 2020 in rugby (no discrimination between league and union) that on average selected 11.2 variables out of 17.3. Since PCA results can lead to several solutions depending on the parameters used, we’ll clearly define our method to ensure the best outcome.

The first step is to select only the uncorrelated variables. We will set a threshold favorability of 5, since some authors suggested that r > 5 is practically significant. We will then test the suitability for factor analysis with KMO for sample adequacy and Bartlett's test for sphericity. We want KMO > 5 and a significant Bartlett's test.

Finally, for factor extraction, we will use Kaiser's criteria, considering an eigenvalue > 0.75 as a rule. We aim to obtain vectors that could correlate with different team statistics. If possible, some defensive features could be extracted from the game description data; otherwise, all variables will be pooled. The features of interest are: lineouts stolen, number of scrums, number of free kicks.

Variables:

* Offensive: possession, territory, tries, meters, defenders beaten, clean breaks, gain line carries, passes, offloads, kicks from hand, rucks won.
* Defensive: turnovers won, penalties, opponent points.

We will perform the analysis using R with the princomp() library and apply the verification steps described earlier. Once armed with this knowledge, we will look at upsets and look for similarities in the tactics of winning teams.

We will also, as a side interrogation, compare close games, with a gap of 2 possessions, and investigate if some styles have an advantage over others. We might not get positive results with the main question, however the framework if the study will enable other question to be investigate easily.

### Modelling

A previous interesting study utilized an artificial neural network (ANN) "to model game outcomes as a function of sequences of actions as well as the field locations where these occur" [8] in rugby union. The author argues it is a great tool to model non-linear behaviour. It has been used to model the outcome of sport events.

however, the goal of this project is to investigate if some outcomes are more likely knowing the events occurring in a possession phase. We will define outcomes that are relevant to decision-makers, such as "Was a try scored?", "Was a penalty awarded?", or "Goal kick?". We will then demonstrate how this method could yield useful information for coaches. We will aim to find tactics that are successful in achieving their desired outcomes. We will display the visualizations on heat maps to be able to easily convey a message. As a follow up to the question asked in the first section, we will look at the composition of sequence with the greater probability to end in a try for teams that are in a disadvantage in a match. We ponder looking at other questions related to in-game decision making, such as: if given the choice of a scrum or lineout in our half, which decision incurs less risk of a penalty? Can we discern offensive patterns that are more likely to ensure the possession is prolonged? Results from the first part could be integrated as parameters in those analyses. However, the author decided that looking at the same question but from different angles in the two parts of the project made for better reading. The sampling will be limited to international matches from the last four years. This time, we should select with a strength imbalance, similar to part one. However, we don't care if the underdog won, which should increase the sample size. This is an important consideration for neural networks, which require a large dataset to prevent overfitting. Then, suggest improvements.

Sample: The most possible, from international rugby.

Variable: Zone, Origin, Pass, Ruck, Clear, Contest.

Outcomes: try, kick-out, turnover, penalty.

We will try to compare different ANN models able to model sequences: a convolutional neural network and a recurrent neural network. For both types of models, we will use the TensorFlow library in Python. The type of RNN to use is not decided, but a good starting point could be to test different methods to remember the data: gated recurrent units, long short-term memory, Bi-directional Recurrent units, as performed by Watson and Dubach [8]. We will first test the usefulness of this method by comparing predictions with a decision tree, probably a random forest. For each model of NN, we will fit: 1) only the number of events in a sequence, 2) the number of events and the field location, 3) the sequence description and the field location. For the decision tree, only the description of the sequences. We will, of course, fine-tune the following hyperparameters: dropout rate, embedding dimension, kernel size, batch size, and number of layers. We will then compare the models based on their error and improvement over the decision tree. We will compute for each passage the validation and training error, area under the receiver operator curve (AUC), as well as the F1-score, which is the harmonic mean of sensitivity and precision.

Once armed with a model, we will evaluate different scenarios at different starting points and predict the probability for all outcomes. To answer the question asked, we will investigate the sequences leading to a try for weakest team, and look for method of starting the play, patterns, starting that have the highest probability to yield a try. We will then display a heatmap with the different outcomes' probability depending on the initial starting point and how the play started. The color of the map will represent probability. A main point to resolve is how to convey the information about the sequence composition. Should the length be displayed? The percentage of rucks or pass? The absolute number of pass or contest? The exploration of the data will inform our choice.

# Time Planning

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Week | Date | Goal | Figures | Script | Tables | Report |
| 5 | 25-03-2024 | Hopefully Have access to the Data. Completion of the Elo scoring system for team strength. Scrap Wikipedia for International Games since 2020. | Display ranking with score as x and countries as y => hbar | ELOR + Scrap Wiki | NA | Method for Scoring system with formulas |
| 6 | 01-04-2024 | With Whole Game Data, do the preparation and cleaning. Write script for PCA. Preparation will be extracting the variables described earlier, selecting the time period. Here We need to think about how the whole game data while be inputted: percentage or raw values. Maybe it is even more meaningful to use the difference in performance between teams for a given indicator. | Display the components selected. Not sure how to display info: the goal would be to have 2 axes on which teams will be located. The closer they are, the more similar they should be. | PCA – Data\_prep | Result tables with eigenvalue and the vector for each. | Write Method |
| 7 | 08-04-2024 | Prepare game commentary data. Decide what outcomes will be. Do evaluation and training for DT. This involves finding a way to keep the order of sequence, had a column for total events in a play. Keep the most possible. What could Data augmentation be? | Description of the design for method | Script For Neural Networks and Decision Tree. Do pretrained model exist? Would save time if only fine tuning is needed. Test multiple Neural network, find a good architecture. |  | Write Method |
| 8 | 15-04-2024 | Train Models. Try CNN and one RNN. Do 3 models, one with only the number of events, one with starting location and number of events, and one with description+number of events+ starting yard. Compute F1-score and error loss (validation and training). Tune hyperparameters. | Na |  | Results, score for each | Add tuning results and model accuracy results(improvement over DT) |
| 9 | 22-04-2024 | Answer Main question. If results are inconclusive, think of other questions to answer, or other method | heatmap | Heatmap |  |  |
| 10 | 29-04-2024 | If up-to-date, create statistical model to predict match outcome and try to integrate results form the 1 analysis in whole game data. |  |  |  | Write Results |
| 11 | 06-05-2024 |  |  |  |  |  |
| 12 | 13-05-2024 |  |  |  |  |  |
| 13 | 20-05-2024 | Prepare Presentation |  |  |  |  |
| 14 | 27-05-2024 | Finish Report – Everything Should be done |  |  |  |  |
| 15 | 03-06-2024 | Review and correct work |  |  |  |  |
| Due Date | 09-06-2024 |  |  |  |  |  |

# References

[1] H. Silyn-Roberts, *Writing for science and engineering: papers, presentations, and reports*, Second edition. in Elsevier insights. Amsterdam: Elsevier, 2013.

[2] E. Hintz, ‘Sports Analytics Before Moneyball’, Smithsonian, Museum of American History. [Online]. Available: https://invention.si.edu/sports-analytics-moneyball

[3] G. R. Lindsey, ‘Statistical Data Useful for the Operation of a Baseball Team’, *Oper. Res.*, vol. 7, no. 2, pp. 197–207, 1959.

[4] C. M. E. Colomer, D. B. Pyne, M. Mooney, A. McKune, and B. G. Serpell, ‘Performance Analysis in Rugby Union: a Critical Systematic Review’, *Sports Med. - Open*, vol. 6, no. 1, p. 4, Jan. 2020, doi: 10.1186/s40798-019-0232-x.

[5] M. Bennett, N. Bezodis, D. A. Shearer, D. Locke, and L. P. Kilduff, ‘Descriptive conversion of performance indicators in rugby union.’, *J. Sci. Med. Sport*, vol. 22, no. 3, pp. 330–334, Mar. 2019, doi: 10.1016/j.jsams.2018.08.008.

[6] M. G.-R. Carlos Lago-Peñas and G. Yang, ‘Styles of play in professional soccer: an approach of the Chinese Soccer Super League’, *Int. J. Perform. Anal. Sport*, vol. 17, no. 6, pp. 1073–1084, 2017, doi: 10.1080/24748668.2018.1431857.

[7] C. J. Wedding, C. T. Woods, W. H. Sinclair, and A. S. Leicht, ‘Operational Insights into Analysing Team and Player Performance in Elite Rugby League: A Narrative Review with Case Examples’, *Sports Med. - Open*, vol. 8, no. 1, p. 140, Dec. 2022, doi: 10.1186/s40798-022-00535-7.

[8] T. S. Neil Watson Sharief Hendricks and I. Durbach, ‘Integrating machine learning and decision support in tactical decision-making in rugby union’, *J. Oper. Res. Soc.*, vol. 72, no. 10, pp. 2274–2285, 2021, doi: 10.1080/01605682.2020.1779624.

[9] P. O’Donoghue, D. Ball, J. Eustace, B. McFarlan, and M. Nisotaki, ‘Predictive models of the 2015 Rugby World Cup: accuracy and application’, *Int. J. Comput. Sci. Sport*, vol. 15, no. 1, pp. 37–58, 2016, doi: doi:10.1515/ijcss-2016-0003.

[10] A. N. Ungureanu, C. Lupo, and P. R. Brustio, ‘A Machine Learning Approach to Analyze Home Advantage during COVID-19 Pandemic Period with Regards to Margin of Victory and to Different Tournaments in Professional Rugby Union Competitions’, *Int. J. Environ. Res. Public. Health*, vol. 18, no. 23, 2021, doi: 10.3390/ijerph182312711.

[11] C. M. E. Colomer, D. B. Pyne, M. Mooney, A. McKune, and B. G. Serpell, ‘A qualitative study exploring tactical performance determinants from the perspective of three Rugby World Cup coaches’, *Int. J. Sports Sci. Coach.*, vol. 17, no. 4, pp. 734–741, Aug. 2022, doi: 10.1177/17479541221087384.

[12] L. Vaz, M. V. Rooyen, and J. Sampaio, ‘Rugby Game-Related Statistics that Discriminate Between Winning and Losing Teams in Irb and Super Twelve Close Games.’, *J. Sports Sci. Med.*, vol. 9, no. 1, pp. 51–55, 2010.

[13] D. Rojas-Valverde, J. Pino-Ortega, C. D. Gómez-Carmona, and M. Rico-González, ‘A Systematic Review of Methods and Criteria Standard Proposal for the Use of Principal Component Analysis in Team’s Sports Science.’, *Int. J. Environ. Res. Public. Health*, vol. 17, no. 23, Nov. 2020, doi: 10.3390/ijerph17238712.