Title [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 disentangle what can be perceived as a good approach from the reality of what actions could lead to the desired outcome. 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 analyzed military operations to improve their efficiency. And what better thing could those OR officers do 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. 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.

The data collected can be separated into two categories: the sport-specific functional components or technical skills, such as working out ways to offload the ball or capture it. This helps coaches understand how different aspects of the game unfold. Notational analyses at the team level can help describe collective behavior and outcomes, such as match success.

Modeling game outcomes is an increasingly popular endeavor, whether for managers or betting companies. Prediction and modeling in rugby league have increased in prevalence since the professionalization of the sport and the constant growth of revenue.

We also observe 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.

Litterature Review [1, Ch. 4]

In Rugby, various approaches have been utilized to study the relationships between performance indicators and outcomes. 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.

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., 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. 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. However, 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. 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 [5].

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 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. 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. As an example, 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 [6].

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. 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 [5].

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, presenting a hurdle for any sport. 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.[7] 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. Rugby is an invasion sport, with the goal of infiltrating the opposition’s territory to score a goal. 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, we should aim to integrate Operations Research (OR) and Machine Learning (ML). 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. Our main 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 and analyze the game as Markov chains to further characterize what makes for a scoring sequence, considering different styles of play. Armed with this knowledge, we will more precisely examine how the length of a sequence affects the outcome: does the length increase the probability of scoring for the attacking team? Does it increase the likelihood of scoring for the other team? And how does that vary during the game?

The Project

Definition

In this project the author will try to reach three goal. The first one will be to use the whole-game data from the Rugby world Cup 2023, and more, if more is available, to define different styles of play amongst the team involved. In a second time, the author will use in game sequences and neural-networks to find relationships between between the events, the starting location of each action to evaluate the most probablie outcome. Then to showcase how this model can be use to support decision making, we will investigate 2 questions: X and Y.

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, I would like to examine games where the score is close (1 possession, or 6 points or less). Then, I am interested in investigating if a particular style of play is used 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.

I will use principal component analysis to reduce the amount of data and determine the number of variables that should be used. I would like to define the team along a defense and attack axis.Rojas-Valverde et al. [9] 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 possible features I would like to extract 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 compare close games, with a gap of 2 possessions, and investigate if some styles have an advantage over others.

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 knowing the events occurring in a possession phase gives some more likely outcomes. We will define outcomes that are relevant to decision-makers, such as "Was a try scored?", "Was a penalty awarded?", or "Goal kick?" Armed with this knowledge, we aim to 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. We will pick a question to demonstrate how the method would work: 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? In this study, it would be ideal to integrate the results from the first part.

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 as well as 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. 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.

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. | Display the components selected. | 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 | Description of the design for method | Script For Neural Networks and Decision Tree. |  | Write Method |
| 8 | 15-04-2024 | Train Models. Try I 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. | heatmap | Heatmap |  |  |
| 10 | 29-04-2024 |  |  |  |  | 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 |  |  |  |  |  |

In general, your report should conform to the Research Proposal style set out by Silyn-Roberts [1, Ch. 5] (note that this reference is available from the QUT library) including:

* a Title [1, p. 19]
* an Abstract [1, Ch. 3]
* an Introduction [1, p. 28]
* a Literature Review [1, Ch. 4]. This is relevant to defining the project scope and could include review of methods as well as the topic domain
* the body of the report, in which you should present your
  + Project definition: what you intend to do,
    - e.g., a *“...description of the expected stages of the research and an outline of the techniques you expect to use during each one. It may be effective to describe each expected stage and its procedures under an appropriate series of headings”* [1, Ch. 5]
  + Project plan: how you intend to do it (your project)
    - *“For those topics that are less well defined (such as Ph.D. projects and projects where you will follow research leads and possibly construct equipment or devise methods of which you may not have any clear idea at present):  
      State clearly how you propose to tackle the first stages of the project.*

*Then follow with a reasoned description of the framework that the research is likely to follow and the possible procedures that may be needed.”* [1, Ch. 5]

* + Time planning [1, p. 33]
* References [1, Ch. 15], [10]
  + Please use IEEE style referencing
  + We strongly encourage using bibliography management software (e.g., [Zotero](https://www.zotero.org/) [11]) to automate citation and bibliography generation.

# More guidance for your project plan

This purpose of your plan is

* to show that you have a clear idea of previous work in the area, the research problem and the procedures you will use to tackle it
* to convince someone else (e.g., your client/collaborator/partner) that your work will be of value.

*“Design your [proposal] with both specialists and non-specialists in mind”* [1, p. 78]

* *"Embedding your detail within a framework of cleverly designed headings, subheadings and listed points will make it much more easily accessible to all your assessors, both specialist and non-specialist.*
* *It is a much greater achievement to be able to design a readily navigable document with a clear logical pathway – the red thread – through it, than to bombard your assessors with solid detail"*

## Use of Microsoft Word styles

**Please use, but do not change the styles of this template in your report**. They are deliberately plain.

If you do not know what Microsoft Word® styles are, or how to use them, please visit <https://support.office.com/en-us/article/Style-basics-in-Word-d382f84d-5c38-4444-98a5-9cbb6ede1ba4>.

## Length of report

Your plan should be as long as you think you need to successfully communicate your intentions. It is important that your plan has a logical flow of ideas, i.e., tells a meaningful "story". You should strike an effective balance between detail and meaning... and the reader's attention span. Our guess is that this will involve around 5-10 pages of text references, appendices and figures.

## Use of figures and tables

Microsoft Word® is good at many things. Unfortunately, placing figures and tables is not one of them. Our advice is to place figures and tables after you are satisfied with the text.

Follow Silyn-Roberts’ [1] guidelines on Illustrations for figures (p.44) and tables (p.47). All figures and tables should be captioned using Word’s captioning (see <https://support.office.com/en-us/article/Add-captions-in-Word-82fa82a4-f0f3-438f-a422-34bb5cef9c81>) like this:

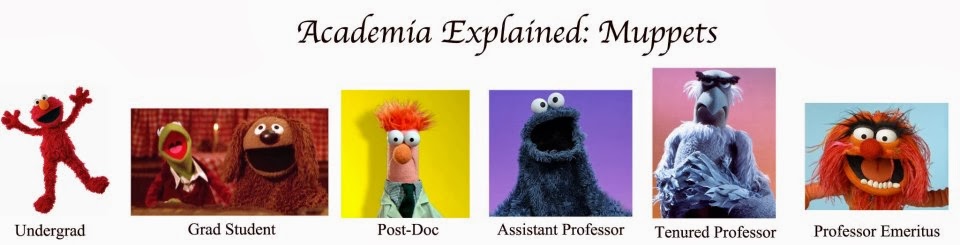


Figure 1. Academics come in all shapes and sizes; however, they are generally not as funny as Muppets.

|  |  |  |
| --- | --- | --- |
| Academic Level | Title | Analogous Muppet |
| A | Associate Lecturer | Robin (Kermit’s nephew) |
| B | Lecturer | Scooter |
| C | Senior Lecturer | Fozzie Bear |
| D | Assistant Professor | Cookie Monster |
| E | Professor | Sam the Eagle |

Table 1. Existing academic-Muppet analogues (Levels D and E) plus conjectured analogues (Levels A-C) for the Australian academic system.

## Use of bulleted lists

Bulleted lists should use the basic Word style as follows:

* First level item
  + Second level item
* First level item.

## Use of numbered lists

Numbered lists should use the basic Word style as follows:

1. First level item
   1. Second level item
2. First level item.

## Headers and footers

Please replace the existing page header with the title of your report, or a shortened version thereof.

Please do not change the page footer.

## How to help readers navigate their way through your document

Silyn-Roberts [1, p. 11] has useful advice on this topic suggesting that the document and its sections start and end with information that is brief, focused and concise.

# 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] 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.

[6] 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.

[7] 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.

[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] 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.

[10] ‘QUT cite|write - QUT cite’. Accessed: Jun. 20, 2016. [Online]. Available: http://www.citewrite.qut.edu.au/cite/

[11] Various authors, *Zotero*. Roy Rosenzweig Center for History and New Media, 2020. [Online]. Available: https://www.zotero.org/