



Institiúid Teicneolaíochta Chorcaí
Cork Institute of Technology

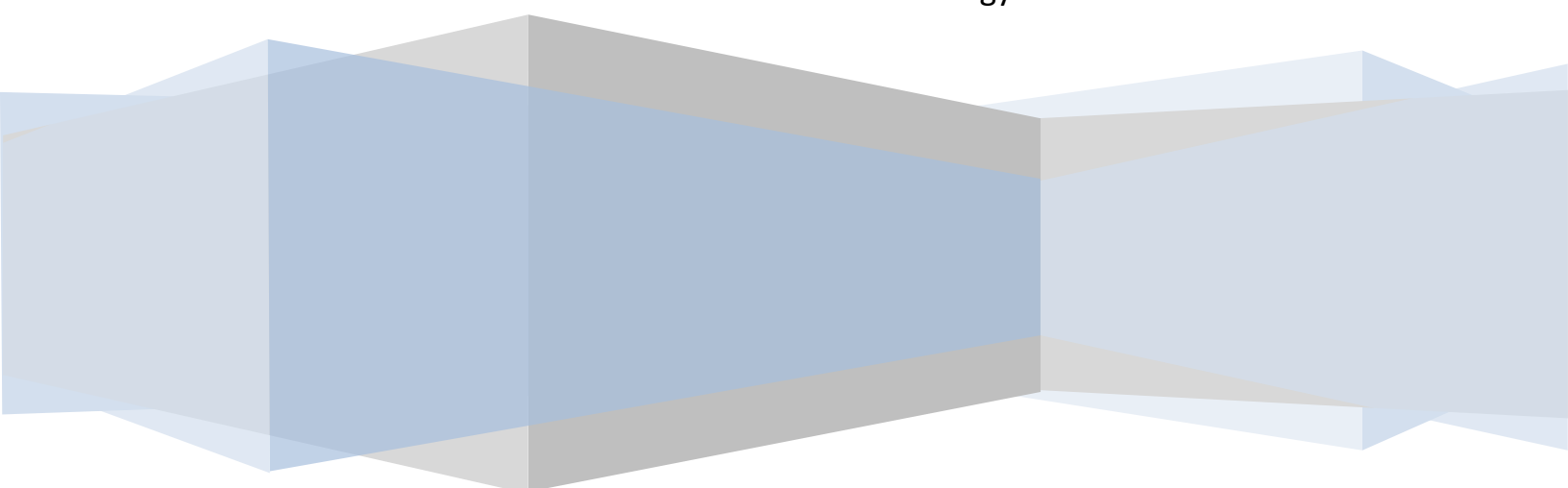
Sporting Performance Analysis Utilising Big Data

Fergal O Farrell (BBS, H.Dip. CSC)

Supervisor: Aengus Daly

MSc in Cloud Computing

Cork Institute of Technology – CIT



Disclaimer: This report represents entirely the work of the author except where specifically referenced from an alternative source. The report may not be copied or distributed without the consent of the author due to the personal identifiable information contained within. An alternative anonymous paper will be made freely available.

**Sporting Performance Analysis
Utilising Big Data**

Fergal O Farrell (BBS, HDip CSc)

Abstract

This research was completed in collaboration with the Strength and Conditioning Department of Munster Rugby. Each of the Munster players' activities is being measured by GPS units and creates various different datasets. These datasets include distance, speed and impact data for these players for each of their games and training. This research looks to mine these data sets for analysis purposes. It looks to analyse various focus areas including player difficulty grading, game segmenting, intensity impacts, and differences between the two halves of games.

This research highlights very interesting results in terms of behavioural and correlation activity. The following paper will detail, for each of the focus areas; the background, methodology, analysis, limitations and findings associated with the research undertaken and the value extracted for Munster Rugby. This is a primary research paper and attempts to explore areas that have limited, if any, published analysis to date. It is seeking to go beyond the traditional standardised GPS outputs from the GPS providers to gain deeper understanding and more applied usage of the GPS generated data.

These results provide a platform from which Munster can apply different preparation strategies in order to best manage their players. The results allow Munster to span different levels which scale from Competition through to Game to Individual level for insights and application. Munster is using these results for the current season to prepare the team for upcoming games.

Acknowledgements

I would like to thank Aengus Daly, William Douglas and Bryce Cavanagh, all of whom were of great assistance through the lifetime of this research project and dedicated a lot of their time to the workings and findings of this project.

- Aengus Daly – from the Mathematics Department in CIT was the project supervisor and provided great support, feedback and recommendations for alternative views and aspects to generate the data modelling and analysis. He was also very helpful at recommending guiding principles in terms of deliverables within each specific research objective.



- William Douglas – was the representative from the Strength and Conditioning Team from Munster Rugby. This project was driven by this team's desire to explore new insights from their own data. This team's enthusiasm and clarity of vision provided a unique and refreshing platform from which to work.
- Bryce Cavanagh – was the Head of the Strength and Conditioning Team from Munster Rugby. His enthusiasm and drive to unearth value from these information sources was the catalyst behind much of the research requirements. This created an environment that allowed for many different focal points to be potential areas of value.



Table of Contents

Abstract	i
Acknowledgements	ii
Glossary of terms	1
Introduction	2
Rugby Union – A Short Introduction to the Game & the Players’ Positions	4
Background Research	5
1. Measurement Variables Overview	7
1.1 GPS Units	7
1.2 Units of Measurements	8
2. Data Preparation	11
2.1 Data Extraction Requirements	11
2.2 Data Cleansing	11
2.2.1 Match Data	12
2.2.2 RPE Game Grading	13
2.2.3 GPS Raw Data	15
2.3 Data Modelling	15
3. Performance Analysis	18
3.1 10 Minute Segments and Load Analysis	18
3.2 Load Change Analysis	33
3.3 RPE Ratings Vs GPS Output	40
3.4 High Intensity Minutes & Clusters of Overall Load by Player	55
3.5 High Intensity Minutes & Clusters of Four Load Variables by General Position	64
4. Results	69
4.1 Result Limitations	69
4.2 Result Sets	69
4.2 Future Proofing	70
5. Conclusion	72
References	73
Appendices	75

Glossary of Terms

GPS	Global Positioning System
RPE	Rate of perceived exertion
RDP	Rabo Direct Pro League
ERC	European Rugby Championship
OD	Odometer
Accel \ Decel	High intensity acceleration \ deceleration
AD	High intensity acceleration \ deceleration
HSR	High speed running
TMO	Third match official
Field	An attribute of a record
Record	A set of fields that represent one instance
Data set	A set of records
Filter	A limitation applied to a data set
IM	Intense Minutes

Introduction

“The results of the analysis conducted by Fergal have a number of implications to the way we prepare our squad. Analysis of the opposition gives insight into what to expect from each team and competition, allowing optimisation of training in light of the predicted match intensity. Understanding of which physical demands, both individuals and positions find most difficult during match play can be used to specify conditioning / fitness drills in order to harden and overcome these difficulties during match play. Furthermore, first and second half load analysis has tactical implications for coaches in terms of the game plan employed for the duration of play. This detailed and unique analysis produced by Fergal ultimately allows us to ask better questions in the way we train our squad, with the ultimate goal of performance optimisation.”

William Douglas – Strength and Conditioning Team, Munster Rugby.

[1] A Global Positioning System (GPS) is a satellite based navigation system that can be used to locate positions anywhere on earth. It consists of satellites, control and monitor stations, and receivers. GPS receivers take information transmitted from the satellites and uses triangulation to calculate a user’s exact location. Individual GPS units are used to gather data from each of the rugby players from the Munster Rugby team.

GPS analysis is a relatively new research topic in a variety of different fields of study. Sporting organisations are amongst the leaders in adopting a GPS research strategy. The scale of GPS generated values within field sporting events makes it a popular tool for data gathering. Many sporting organisations are collecting this data and producing primary information such as distance travelled by player etc. However, the GPS data can contain more insight other than basic distance information and can be matched to other variables both objective and subjective to allow the analysts to extract patterns. The depth and volume of the data means that it is difficult to manage and manipulate it relative to the desired information requests. The ability to create and apply algorithms as well as eco-systems for the processing of this data has proven elusive to many organisations. The ability to extract information from these large amounts of data in a timely manner is an avenue that many are attempting to pursue in order to enable organisations to get a deeper insight into their players’ performance levels.

This research study has been requested by Munster Rugby. Munster Rugby is a globally recognised Rugby Union franchise entity and has a long history of sporting success in its illustrious existence. The organisation has turned to a scientific approach of utilising GPS units to assist in gauging player performance levels. It has recognised the value of Big Data and the ability to extract from this data pool will assist in applied learning from past experience.

This research looks at different subsets within the data and specifically explores both the GPS variables internal relationships and also their linkages to other subjective collected material. Munster had an original set of requirements and based on some of the initial feedback from the exploratory research that was provided to them, more areas were then explored. Some of these areas became obvious and flowed naturally from these results. Others areas were looked at in conjunction with the Munster representatives and were based more on intuition and challenging perceived norms. This project looks to assist Munster Rugby in transforming their collected data into their information requirements so that it can be utilised for process planning. The following sections outline these different research areas.

Rugby Union – A Short Introduction to the Game & the Players' Positions

Rugby Union is a contact field sport played over 80 minutes at multiple levels by underage, school, college, club, provincial and national teams. It is a professional sport having moved from away from amateur status officially in 1995. At any given point, each team is only allowed to have a maximum of 15 players on the pitch. Each player plays in a designated position which provides structure to the game. Each of the players' positions has a different role and the demands on the players can vary depending on these roles. This results in differing physical attributes and skill set requirements. Fig 1 below outlines the position names on the left and the general position groups on the right. The teams are also segmented into forwards and backs. The forwards tend to be closest to the ball at most points in the game and would generally be larger physically. [2] The IRB has been the governing body for Rugby Union since its formation in 1886. As of November 2012 the IRB recorded 118 unions in its membership, 100 full members and 18 associate member countries.

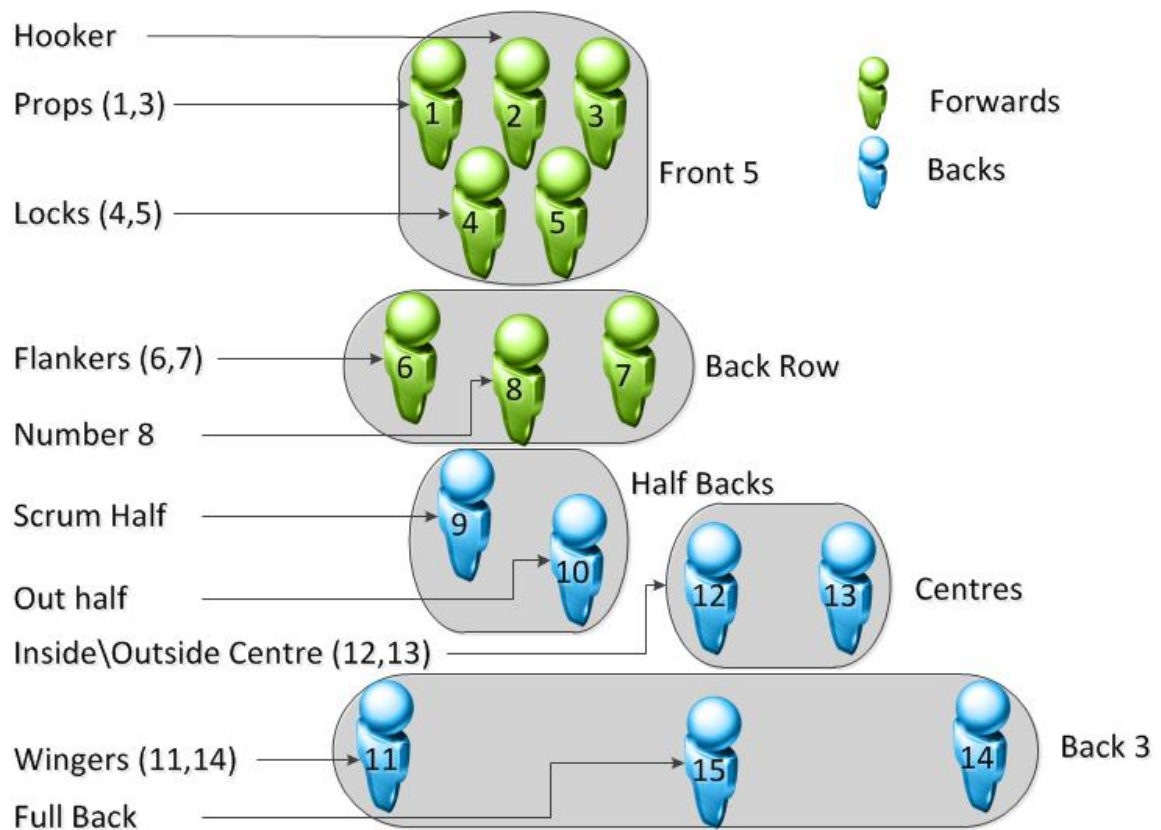


Fig 1 – Positional Outline for Rugby Union

Background Research

This document focuses on new research areas and similar previous research specific to these areas is limited. There have, however, been various studies of GPS monitoring of professional sporting athletes in an attempt to gain behavioural and performance related insights into match and training environments. GPS studies continue to become part of standard processes within the modern sporting organisation. Some research to date looks at the validity of assuming a relationship between any performance indicators and GPS data. Varley et al looked specifically at the measure of velocity in acceleration, deceleration and constant motion. They focused heavily on the GPS units themselves as an appropriate unit to capture the necessary data. In their study they outlined the importance of the data gathering mechanisms. [3] GPS sampling at 5Hz was incapable of detecting the smallest worthwhile change during all phases of these tests. In contrast, 10Hz GPS was able to detect the smallest worthwhile change during the constant velocity and acceleration phase and during the deceleration phase.

The GPS [1] satellites, which are equipped with atomic clocks, transmit radio signals that contain their exact location, time and other information. The radio signals from the satellites, which are monitored and corrected by control stations, are picked up by the GPS receiver. This enables the service provider to output relevant GPS data for all connected devices. The higher the rates of this activity, the more reliable the data generation becomes. The use of GPS units with more frequent sampling rates has been shown to vastly improve measurement precision during activities of consisting of high movement speed and short durations. [4] The evolution of technology has provided greater confidence in captured data and given rise to a much more detailed understanding as to the nature of the intermittent, short-duration, high-intensity sprint activities. The Catapult GPS units that are used by Munster Rugby are 10Hz.

Venter et al have conducted research into players' positions relative to their movements and impacts at an underage level in rugby union. Their research concluded that the different players' groups displayed differing movement characteristics in terms of

speed of movement. [5] Players covered on average a total distance of 4469.95 ± 292.25 m during a game. The front row forwards covered the greatest total distance (4672.00 ± 215 m), followed by outside backs (4597.93 ± 210.18 m), inside backs (4307.78 ± 214 m), and then back row forwards (4302.1 ± 529.82 m). Maximum speeds reached did not differ significantly between the groups of players. Although, this was an interesting paper, it tended to focus largely on limited basic GPS output variables.

Because of the nature and pace of change in the areas of technology, the focus of the previous research was for validation rather than data extraction, modelling and applied learning. Munster Rugby's use of the 10Hz units is supported, in terms of accuracy and reliability, by the research conducted to date as being of a valid rate to capture the variables at a high level of consistency. Any concerns over data validity generated by the GPS units can be underwritten by these aforementioned studies. The authentication of the 10Hz capture rates allows for this research to focus on the data extracts of the GPS units rather than the data generation.

1 Measurement Variables Overview

Munster Rugby is using the Minimax S4 Catapult brand of GPS data gathering equipment. The S4 is part of the Minimax product family, which all share the same basic components. Catapult has a wide variety of sporting franchises on their customer listing from Rugby Union and League, to American Football and Soccer. It is known as being the market leader in GPS measurement in sporting activity and has some of the top sporting franchises amongst its customer base including the 2013 British and Irish Lions, AC Milan and the New York Knicks. They have been providing these units and support software since 2006. Munster first started using the Minimax S4 in 2012. These units are located within a pouch on the back of specifically designed vest worn by the players.

These units are able to produce GPS data that is read by satellites and hosted within an environment managed by Catapult. This environment allows for users to access real-time data sets while games are on-going. Catapult's packaged software enables 'on the spot' processing and provides basic information sets to the end users. This supports the decision-making process during game activity. There is however a limited amount of analytical tools to measure behavioural patterns over the long-term \ extended periods. The raw data is available to Munster, who are then responsible for mining the data for information insights and value extraction. This research project is exploratory and uses these data files to try to formulate intuitive beliefs and understandings of the game.

1.1 GPS Units

These Minimax S4 GPS units that Munster is using are currently gathering 360 variables at 10Hz for all of the training and match sessions. There are 22 GPS units currently in use. The Minimax technology combines 11 sensors in a small wearable electronics device for tracking and monitoring the players. [6] The GPS engine provides 10Hz data for position, velocity, acceleration and distance. 10Hz sampling allows more accurate measurement of short sprints, reliably track complex movements involving multiple direction changes, accurately measure peak and instantaneous velocities. Three axis 100Hz accelerometers are used to measure linear motion, impact forces, jump height,

airtime, acceleration, deceleration and more. Three axis 100Hz gyroscopes are used to measure angular motion, rotation and three axis 30Hz magnetometers to measure direction and orientation. They are now used in more than twenty sports at the elite level for fitness, tactical, rehabilitation and technical analysis. [7] Minimax V4 has a completely new generation GPS engine targeted for team sports. It has better accuracy, a faster data rate (10Hz) and a shorter lock on time in difficult GPS environments such as stadiums.

These units are not worn by prop forwards due to the positioning of the devices as they tend to interfere with scrummaging. For the 2012/13 season, the wearing of the units was optional for the players. The players that did wear the units did so with the assistance of the Strength and Conditioning Team in Munster for each game as well as training sessions. All of this data was captured and collected by Catapult. All players in the current season, except for the props now choose to wear these devices.

1.2 Units of Measurements

The GPS output files were able to gather information on 360 different variables. Some of these variables have sub variables. For example, there are different graded zones to measure different levels of activity. For velocity there are different scales which are broken into various segments to represent different running speeds. This is replicated across other variables of similar scaled nature also.

Munster has configured their system's variables and zones for different collation levels. It was agreed with Munster that there were four major variables that were to be focused on for the purposes of this research. Munster had completed previous internal research and found that these variables provided the most insightful data for their requirements. There was also a requirement to review the other variables to see if there was additional value that could be extracted from their data sets. This requirement was subsequently excluded from this research due to the expansion of the scope from the original requirements. The analysis of these variables should form some part of any subsequent analysis.

The variable outputs that are used in this research are listed below. There was also a variable created from a combination of all of these outputs to give a high level Overall Load variable.

- Odometer: the odometer reading gives [8] the distance travelled for the session in metres. The odometer is set to 0 at the start of each session and summed for each measurement period within the session i.e. per minute per game or per training session. This is the most basic of the GPS outputs and one of the most commonly used variables.
- High Acceleration \ Deceleration: the high acceleration \ deceleration is calculated from summing the two outputs from these readings to give one combined variable. These readings are taken from [8] the percentage of distance in the top \ bottom zone of accelerations as defined by the user. Munster have configured their zones with Catapult as per Fig 1.2.1

	Zone Start	1	2	3	4	5	6	7	8
Acceleration (m/s ²)	-20	-5	-4	-2	0	2	4	5	20

Fig 1.2.1 – Acceleration \ Deceleration Zone Configuration

- High Speed Running: this is the summing of the Velocity Zones of 5 – 8. The zones are configured by Munster as per Fig 1.2.2 and all readings that are within the top 4 zones are summed to give one variable value per defined period.

	Zone Start	1	2	3	4	5	6	7	8
Velocity (m/s)	0	0.5	2	2.8	4.4	5.6	7.5	9	12

Fig 1.2.2 – High Speed Running Zone Configuration

- Tackle Load: this indicates the [8] intensity of effort or exertion during an event that the software has marked as a tackle. The software uses a combination of inertial sensor signals to detect when a tackle has occurred. For a potential tackle to qualify it must first involve an impact event greater than a pre-defined limit. Also there must also be a significant change in the athlete's orientation. These two conditions together indicate that there has been a significant collision. Munster has configured

its Tackle settings as per Fig 1.2.3. The Tackle Load is then calculated by summing zones 3 – 8 to give one output.

	Zone Start	1	2	3	4	5	6	7	8
Tackles (arb)	0	1.5	3	4.5	6	7.5	9	10.5	12

Fig 1.2.3 – Tackle Zone Configuration

- Overall Load: this is a calculated field and is a representation of the four variables above in one single value. The logic behind the creation of the values is contained in section 3.1 within the Methodology section.

These five variables are referred to throughout the document and may appear in graphs and charts under abbreviated synonyms Od, Accel \ Decel or AD, HSR and Tack. Overall Load may be referred to as OL. The configuration of the settings of the zones is owned by Munster and is outside the scope of this research in terms of zone altering. This research takes the outputs from these variable configurations.

2 Data Preparation

2.1 Data Extraction Requirements

The research for this project looked at activities over an extended period of time. There was not any reliance of system performance in terms of speed of data retrieval or output processing delays. The analysis focused on complete data extracts from the entire 2012/13 season. The Catapult system allows Munster to extract large volumes of data for various activities. For this research, there were two data extracts although the second extract 'GPS Raw Data' was removed from the scope of the research due to scope creep directly as a result of new insight requests from the initial findings of the research. A third file, RPE Game Grading was generated and manually managed by Munster for subjective player measurements.

2.2 Data Cleansing

Three files were provided by Munster Rugby. All of the raw data that was required for this analysis was contained within these files. As the project progressed, the requirements were amended and the research only focused on two files. The original requirement of the training load analysis, which would have required utilising the GPS Raw Data, was withdrawn in order to focus on spin-off requests from the other analysis sets. The three files provided were:

- Match Data
- RPE Game Grading
- GPS Raw Data

These files needed substantial cleansing in order to get consistency and integrity within the datasets. The depth of some of the issues would need to be corrected for future analysis and would undermine any automated reporting from the source systems. All of the data quality issues are highlighted in the following sections.

2.2.1 Match Data

The match data consisted of a variety of data variables from each game played. Every player's GPS readings for each game played were captured for every minute of play in all of their games for the 2012/13 season. For this season the wearing of the GPS units was not mandatory and was instead optional for the players. The prop forwards are excluded from wearing the units due to the position of the GPS unit on the upper back \ neck region as it may impact on any scrum activity. Subsequently there was no raw data generated for these players.

The match data consisted of a data cube of 33,777 records with each containing 81 data fields. This data was just a raw extract from the Catapult system and required extensive corrections to make the data suitable for analysis. The biggest problem within the data sets was the time period assigned to each record. The GPS units were not completely synchronised with each other and this resulted in time stamps with 1 second differentials. This resulted in records receiving different time periods to records which started 1 second earlier. In some instances there were 150+ time periods for a match because of this issue. The data in this state was unsuitable for any time-series analysis. Fig 2.2.1.1 outlines this issue, which shows the number of records in each time period.

Count of Game		Labels												
Row Labels	Cardiff	Connacht	Dragons	Edinburgh	Glasgow	Leinster	Ospreys	Racing Metro	Saracens	Scarlets	Treviso	Ulster	Zebre	
1	13	15	10	20	14	14	12	16	15	9	10	14	11	
2	13	14	1	17	14	1	12	15	15	9	1	14	1	
3	13	1	10	3	14	14	12	1	15	9	10	14	11	
4	13	14	1	17	14	1	12	15	15	9	1	13	1	
5	13	1	10	3	14	14	12	1	15	9	10	1	11	
6	13	14	1	18	14	1	12	15	15	9	1	14	1	
7	13	1	10	2	14	14	12	1	15	9	10	14	11	
8	13	14	1	18	14	1	12	15	15	9	1	14	1	

Fig 2.2.1.1 – Data Anomalies

Any value with 1, 2 or 3 has a synchronisation issue. This issue is further compounded by the fact that all of the subsequent readings are offset and not reflecting the correct time period. Correcting this consumed a substantial amount of effort. It did require all 33+k records to be individually checked for integrity. If this went uncorrected, the analysis for consecutive 10 minutes periods could result in a range from 10 to 5 actual

minutes per period. This issue would also undermine any current ‘off the shelf’ analysis provided by the vendor. Munster Rugby has been advised of this issue and they have in turn highlighted it to Catapult. In Fig 2.2.1.2 the corrected data is summarised showing all corrected minute sets.

Count of Game Column Labels																
Row Labels	Cardiff	Clermont	Connacht	Dragons	Edinburgh	Glasgow	Harlequins	Leinster	Ospreys	Racing Metro	Saracens	Scarlets	Treviso	Ulster	Zebre	
1	22	11	22	19	33	20	12	21	8	19	16	15	18	18	7	
2	21	12	20	18	35	20	12	20	9	17	18	14	17	23	7	
3	19	11	22	17	32	21	14	18	9	17	16	13	17	23	8	
4	20	11	20	20	31	22	12	22	9	17	16	13	17	21	7	
5	19	12	20	21	34	20	12	18	9	17	17	13	17	22	7	
6	19	12	20	18	32	24	13	20	10	18	16	13	19	21	7	
7	19	11	22	20	33	20	12	21	9	19	17	13	18	22	7	
8	19	11	21	17	34	21	12	22	9	18	16	13	17	22	7	

Fig 2.2.1.2 – Data Corrections

Other issues found with the data required the complete removal of RDP 5 game data against the Ospreys as all of the GPS readings were returning values inconsistent with the balance of the file. All of the time periods, which were supposed to end one minute after the start had a 0 end time. There seems to have been an issue particularly with the Odometer readings which were returning very low values.

For the Ulster match RDP 12, there were various minutes that had duplicate records, namely period 76 and 126. For the Edinburgh match RDP 1 the same issue occurred for the time periods 51 to 59. Both of these sets were removed from any subsequent analysis. For the ERC 1 match against Racing Metro there was an issue with Ronan O Gara’s GPS readings and these values were removed.

Other anomalies included no second half GPS output for the Dragons RDP 4 home match. There is no data for the Zebre away RDP match. The Treviso game RDP 16 was showing as being against the Scarlets, which was the opposition the previous week for RDP 15. This was corrected prior to the analysis.

2.2.2 RPE Game Grading

RPE (Rate of Perceived Exertion) is a generic measurement used across many different sports. Each player grades each game from 1-10 after the game has been completed. This is not the typical RPE scale but it is the one that is used by Munster.

Scores closer to one represents easier games while the scores at the upper end of the scale represent the more difficult games in terms of physical demand. The players grade is then multiplied by the minutes played (stops at 80) to give the RPE Load value. The RPE data is available for 22 of the 28 games consisting of 466 rows of data. The cut off between a Hard and an Easy categorised game was agreed with Munster as being 7.5. As this measurement is subjective and idiosyncratic to each of the players' perceptions, we adjusted down the cut-off line to 7 for two of the players, Mike Sherry and Denis Hurley, so that they have a more balanced representation of easy \ hard games.

There were two sets of criteria that were agreed to see whether or not a record should be considered for analysis at an individual level. The threshold set for the number of games graded to be consider was set to > than 4. Each of the time measurements were categorised as follows: 80 minutes = Full, 55+ minutes = Most and below 55 = Sub. Only grades that were recorded for Full and Most were considered to be valid. Because of the nature of games nowadays, this limited the readings from two positions in particular; Scrum Half and Hooker.

Because the RPE data was to be used along with the GPS units, there was a further cut in some of the records. Even though, one data set may have RPE grades, there were many that didn't have the corresponding GPS records. The wearing of the GPS units was optional and therefore a further 226 records were removed which included the props who do not wear the GPS units for any game. There were also instances of the opposite, whereby the GPS readings of a player could not be matched with any equivalent RPE grades. This can be for a variety of reasons including player injury or human error.

This left the data set with 240 records of which 164 met all the criteria to be included in the analysis. This included data for seventeen different players in all seven different general positions. This allowed for in position analysis (four different players in the "Back 3" category) in some instances as well as positional comparison analysis.

2.2.3 GPS Raw Data

The GPS raw data file was an output per player per training session, game or activity that was measured throughout the season. This file was not used in this research due to changing requirements. Each of the players was monitored by the GPS units for all activity and the result outputs were totalled in this file. It provided a data set that could compare different weeks of training activity against the match data to analyse the impacts of the training loads on the outcomes of the games. It could also be used for analysing the training loads and RPE grades per player to seek out behavioural trending. The requirement from Munster specific to this data set was to issue recommendations as to the optimal dose of training to elicit optimal performance on the game weekends. This may be now included in any future analysis as it was removed from the output objectives of this research.

2.3 Data Modelling

The data modelling is explained in each of the different sets of analysis in section 3 that under the 'Methodology' heading. For the most part Munster were trying to establish correlative information and much of the analysis sought to extract information pertaining to higher level groups and seek correlation with a dependable variable. The correlation value is [9] where a correlation coefficient is reported representing the degree of linear association between two variables.

Munster Rugby had all of this information to hand but with the quality, size and breadth of the information, struggled due to time constraints and systems' expertise, with outputting useful summary data that would assist in any sort of high-level behavioural analysis. The Catapult technology does have a set of standardised outputs but these are more for real-time readings for in-game and training analysis to assist with the decision-making process. However, some of the exploratory requirements requested by Munster, were not standard outputs and therefore not readily available from the system. The Munster requirements were built from strength and conditioning requirements rather than the data set itself. These requirements were well defined and very specific to core

focus areas. This greatly assisted in the data modelling process and while some of the analysis was an iterative process, most results did not need major re-building from initial feedback.

As well as correlation, Decision Tree analysis was used to build up a data model that would allow for Munster to input raw data to a model to assist in predicting the outcome of Hard or Easy games for players that were less strong in this subjective measure. This model was built from the data that was measured in this analysis which was based on the data from 2012/13 season. The Decision Tree analysis can be seen further in section 3.2.

All of the modelling outputs were put into graphical format where possible to ease the interpretation requirement of any processing. This is important to show non-analytical based communication to management or to the individual players. Where possible, trending graphs were used for the output for example run charts were used for time-series events. Colour coding was used for results of measurements of note. Also major time milestones such as halves of a game were added to dissect output again to limit the possibility for misinterpretation. For any sustained viewing of the output, supplementary information was provided on the graphs. For example in the 10 minute segments the score trends were added so that an additional level of understanding and interpretation could be applied. A lot of the supplementary information additions were part of the iteration process of producing the output and were as a result of predicting typical questions that would have come from those interested in the output.

From a system's support perspective, all of the modelling and analysis was performing on a limited number of systems with MS Excel being the main one. Excel did struggle with the size of the files and would not be recommended for future repeat activity. However, it has the capability to extract most of the requirements sought. Also, the user profile of Excel would need to be at an Expert level. As well as advanced formulae such as Fig 2.3.1, Visual Basic and Excel Analytical add-ins were used to spin the data into information. One of the add-ins used was the Lean Six Sigma analysis tool kit

which has a suite of analytical measurement and output processing VBA modules. The data can be fed through this to return desired results for the data sets. This allows for standardised output for different analysis sets.

`= (Z2/MAX(Z:Z)*LookUps!I$3)+(AI2/MAX(AI:AI)*LookUps!J$3+(AT2/MAX(AT:AT))*LookUps!K$3+BO2/MAX(BO:BO)*LookUps!L$3)`

Fig 2.3.1 – Advanced Formula Example

The Decision Tree modelling and analysis was performed on Rapidminer. This required the base data to be cleaned, labelled and uploaded to the platform to be used to build up the data model. The model would output the credibility of the analysis and the base data would be limited or refined depending on the % credibility. The % credibility was calculated by pushing the data through a validation operator function. [10] This operator performs a cross-validation in order to estimate the statistical performance of a learning operator (usually on unseen data sets). It is mainly used to estimate how accurately a model (learnt by a particular learning operator) will perform in practice. Once the model was satisfactory, this would allow for a predictive element to be utilised by uploading new raw data and processing it through the model to predict the outcome variable.

Other systems used include Tableau for modelling of the outputs to more human readable formats for ease of understanding. The Tableau software utilises raw / summary data files and applies various data visualisation techniques to represent the same data. Their goal is to [11] provide data visualization software that helps you understand and communicate information in the most effective way possible.

Overall the difficulty with the data sets provided by Munster was not a technical limitation. The biggest challenges were in the cleansing, processing and outputting of it from a logical perspective rather than a systems perspective. This research focused on the results sets rather than on the systemic mechanisms by which the results were obtained.

3 Performance Analysis

3.1 10 Minute Segments and Load Analysis

Background

Munster Rugby's requirements included analysis of the various loads placed on the players and team in 10 minute segments so that they can analyse game patterns. By segmenting the games into 10 minutes windows, it would show consecutive periods in each match and allow them to map out the games and behavioural activity of the load variables. It was hoped that this would highlight the opportunities which may be taken advantage of or periods where they consistently came under pressure.

The analysis was to be provided at a game level as well as at a competition and result level. This allows for game comparisons and analysis against different patterns for example in the ERC (European Rugby Competition) versus RDP (Rabo Direct Pro) competitions. It would also provide a template for opposition profiling whereby Munster could study previous game patterns for either the Home or Away game from this analysis to best design a game plan for upcoming games. Most of the games would have a Home and Away match with the exception of the ERC knock-out rounds and the games for which there were GPS data issues.

Along with the games and opposition, each individual player or general player position was available so that positional behavioural profiles could be built. Although that information was available, for this analysis Munster were more interested in extracting the analysis for the team level. This analysis goes on to highlight some very interesting results which Munster can directly apply to their squad preparation processes, which is one of the core purposes of this research.

Methodology

There are challenges associated with drawing comparisons with the timings of both the match clock and the GPS units. Both time recordings are not 'like for like' so

there are further considerations to be made. The match clock is stopped by the referee \ time keeper at each prolonged stop in play. The GPS units are recording data on a continuous basis and are only stopped at the end of the half or game. Therefore, comparison of the GPS timings and the match minutes is not possible. Each of the 28 games is segmented into 10 minute time segments. These time segments are consistent for the first 40 minutes from the GPS units in each half by taking consecutive 10 minute segments. The extra time played at the end of each half is put into its own category. These categories 1HEX & 2HEX have values that range from 3 minutes to 13 minutes depending on additional time played.

A straight minute by minute analysis was also conducted by segregating each match timeline into 10 minutes and disregarding any half time break. This resulted in eleven different 10 minute time segments where the data segmented into the 5th or 6th group could be either 1st or 2nd half. There were similar issues for the 9th, 10th and 11th segments where there was no clear like for like comparison in terms of counts of records per 10 minute record set as each game has differing additional minutes in both halves as can be seen in Fig 3.1.2.1.

Count	Back 3	Centre	No. 10	Half	B'Row	Lock	Hook
10 Min (1)	630	508	139	247	521	359	310
10 Min (2)	673	561	143	306	564	396	372
10 Min (3)	682	571	147	298	560	392	400
10 Min (4)	667	551	148	271	546	384	342
10 Min (5)	673	582	150	312	568	413	398
10 Min (6)	653	520	142	290	533	380	365
10 Min (7)	659	514	136	283	529	380	367
10 Min (8)	636	485	122	249	530	388	324
10 Min (9)	578	469	108	214	468	370	300
10 Min (10)	199	177	37	87	160	121	101
10 Min (11)	18	20	3	12	14	26	16

Fig 3.1.2.1 – Consecutive Time Series

Because of the inconsistency with the straight measurement which did not distinguish between the halves, this method was disregarded. It was agreed that the games halves were critical in terms of segregation. Also, it was believed that player

behaviour could be significantly altered before and after a half time break. Therefore the extra time method was used so the measurements could definitively be comparable for all time periods as a like for like with the exception of the two periods of additional time at either end of the half.

This results in 10 different time segments for each game. Each of the 10 game segments is then subjected to comparative analysis regarding the load variables. The agreed load variables to be analysed were the Odometer Load, Acceleration\Deceleration Load, High Speed Running Load and the Tackle Load. Each individual record was divided by the maximum value for that variable in the entire data set. This gave each record a reading as a percentage of the top value for each of the variables. A combined Overall Load was then calculated per minute per player based on multiplying each of the above four loads by a weight factor and summing the results. For the purposes of this analysis, the loads were weighted equally until further insight could be gained into weighting groups differently.

Every GPS data row had a criteria applied as to whether or not it met with inclusion logic to be considered for analysis. If the odometer reading was equal to 0, it meant that there was no movement from the player which is interpreted as the player not being involved in the play (sitting on the bench) and so this information was discounted.

Each of these four agreed load variables as well as the Overall Load was used in the analysis. Munster Rugby required the analysis to be done per opposition, per game, per ERC \ RDP, per result – win \ loss \ draw. All of the result outputs were plotted in various run charts to show the time series or pareto charts to show decreasing quantities for the analysis. For each of the opposition analysis, the scoring pattern by half was represented in a table above the run chart information. All opposition graphs were plotted showing each game played as well as an average load for visual comparison. Each game half was separated by a heavy line for visual interpretation. Munster played against Edinburgh four times in the 2012/13 season and so there was further opportunity to analyse opposition per competition for this opposition. Munster won all four of these games so the mix may

not be representative of their season given that they won 14, lost 12 and drew 1 of their 27 games for that entire competitive season.

The season's results were poor overall and can be viewed in terms of result and the margin of the result (with the exception of the 1 drawn game) in Fig 3.1.2.2. This outlines how balanced the season was in terms of wins in green and losses in grey. There is far too much grey representation on the graph which signifies an average season. The size of the circles indicates the margin of the result in absolute terms.

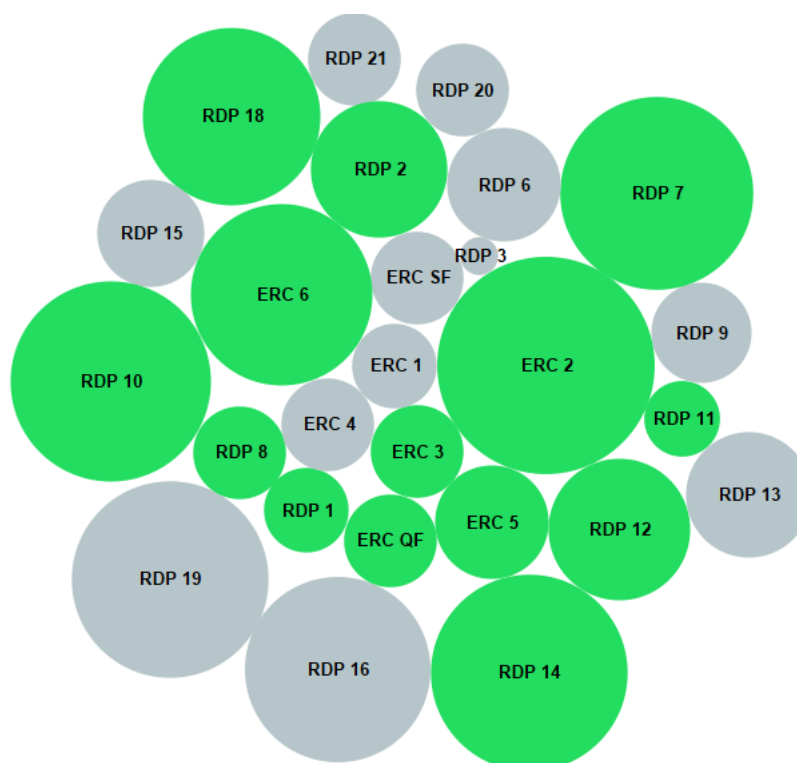


Fig 3.1.2.2 – Results and Margins Representation

Analysis

This analysis looks to profile teams so that Munster can anticipate the flow of the games against upcoming teams which are included here. These insights can have a substantial influence on game plan management. The analysis was carried out for all games for the season but for illustrative purposes only one of the opposing teams has been focused on. All of the charts in this section represent the analysis for the opposition team Racing Metro. See Appendix A Ref 3.1.A for the output files. Racing Metro is a rugby

franchise from the French Top 14 league and they were drawn in the same group as Munster for the ERC Heineken Cup pool stages. There were two games between these two sides in this season. Both of these games are represented in these analysis results.

In Fig 3.1.3.1 the graph shows the overall calculated load (calculated from the 4 agreed variables) for the two ERC round games and plots them against the average Overall Load for all of the 28 games of the entire competitive season. The table along the top shows the scoring sequence per half both for points scored for and against. It also categorises the scores in terms of Tries, Penalties, Conversions and Drop Goals. The total is summed for each half as well as for the game overall and the result is provided. For ease of viewing, the final score always represents the Munster score first regardless of the Home or Away status of the game. This table is reproduced for all of the individual opposition analysis outputs. In general, the Overall Load across all games peaks at the start of each half and subsides as the halves progress. This is consistent with the exception of the 30-40 minute time period in the first half which sees an increase in Overall Load relative to its adjoining time segments.

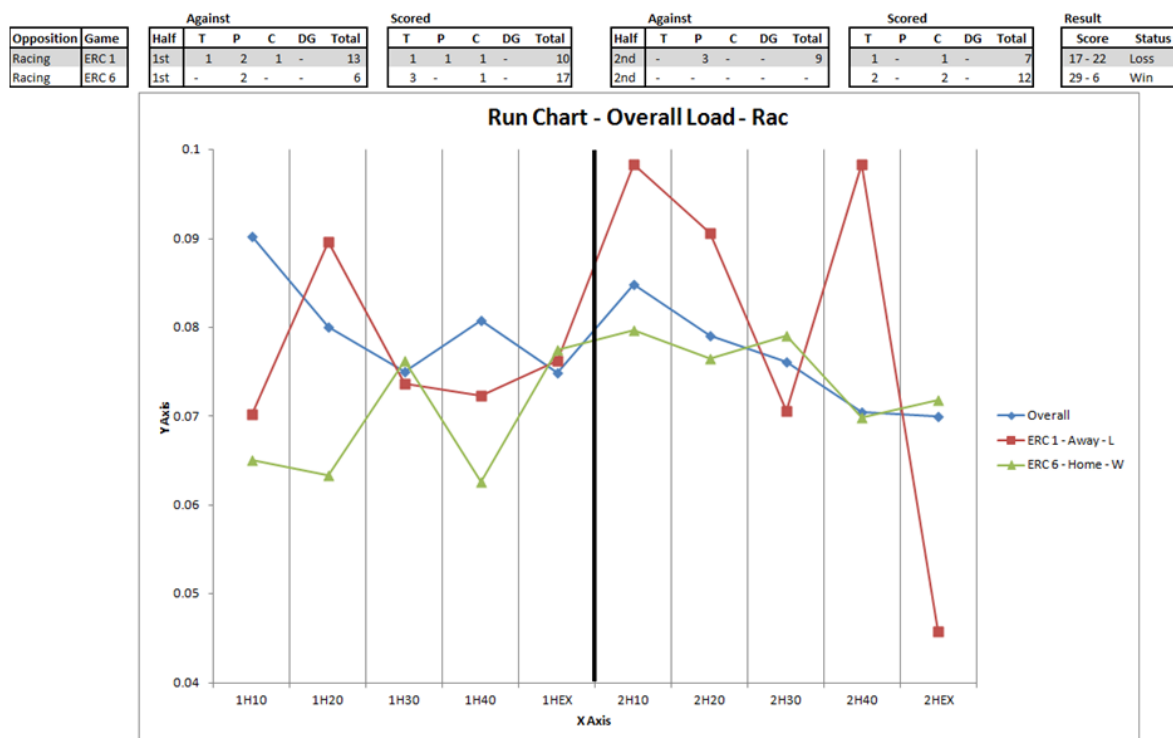


Fig. 3.1.3.1 – Overall Load – Racing Metro

In this graph we can see that ERC 1 was above the average Overall Load of five and below the overall average load of five of the 10 time segments. The first half is less load bearing with fewer spikes than the second half. For ERC 6 we can see that overall the game placed less Overall Load both than the average and the ERC 1 round. The first half looks to be well below average, with the second half more aligned to average Overall Load. Both of these games register either side of the average consistently and neither of these games appears to have applied the greatest load on the players.

In Fig 3.1.3.2 the graph looks at the Odometer Load between both of the Racing Metro games and the average Odometer Load placed across all 28 games. Again the home game looks to align with the average readings for the other games but the Away game shows spikes for a lot of the second half. For both the 1HEX and the 2HEX time periods, these may contain less \ more data sets than the other timing segments so drawing conclusions based on these may not be accurate.

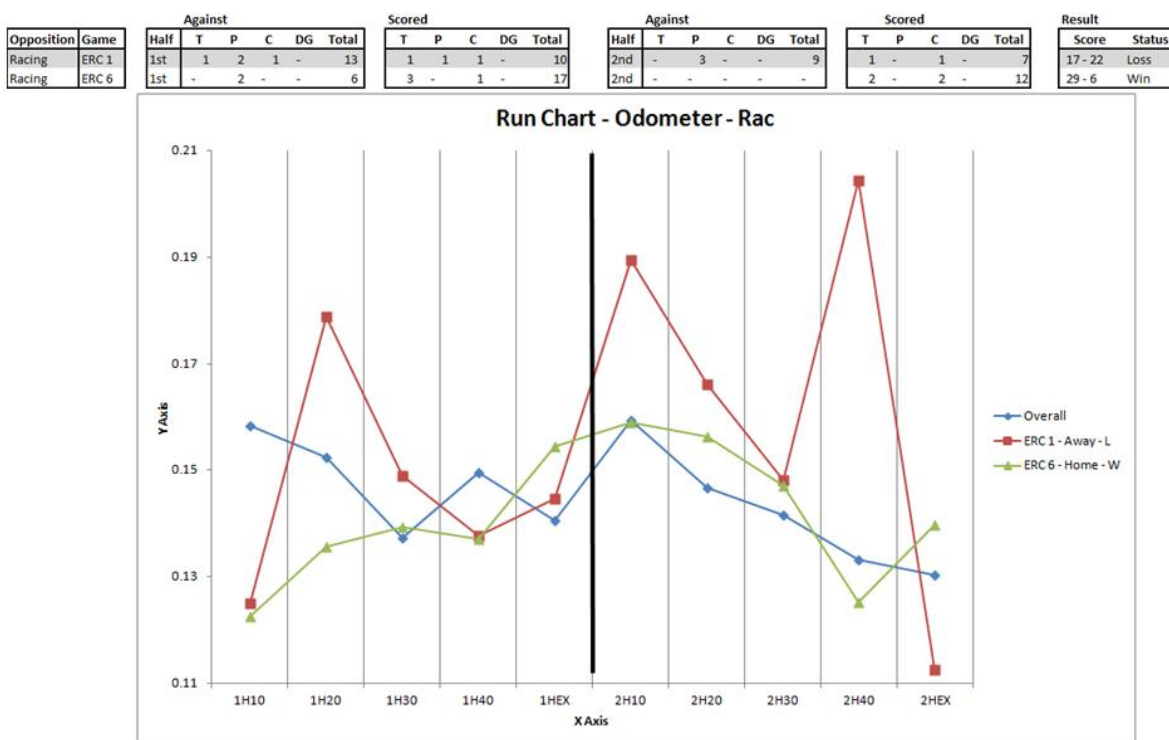


Fig 3.1.3.2 – Odometer Load – Racing Metro

In fig 3.1.3.3 the graph represents the Acceleration \ Deceleration Loads placed on the players relative to the average load for the same variable across all of the games.

Looking at the graph, we can see that both games plot below the average in general with the exception of two periods in a row in the ERC 6 game. These readings measured below average right up to the 2H20 time segment with three of the last four readings showing above the average. For the ERC 1 game, only two of the readings measured above the average and one reading measured about average with the balance falling below the average measurement.

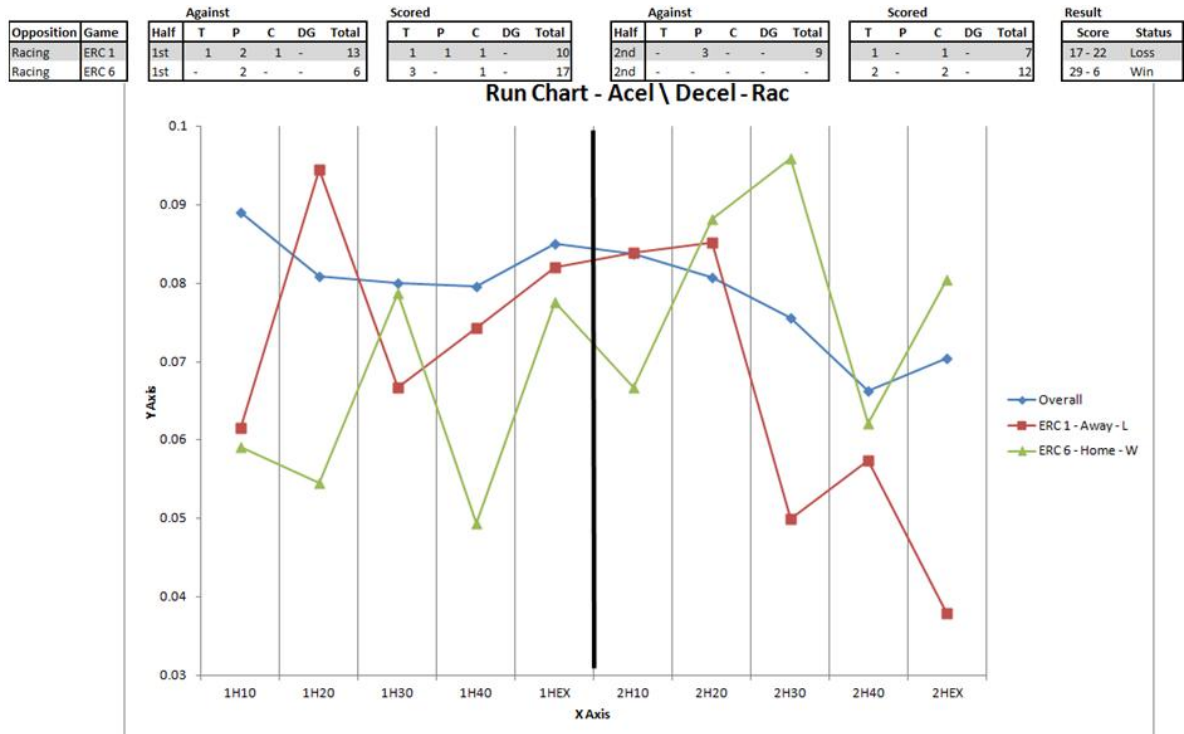


Fig 3.1.3.3 – Acceleration \ Deceleration Load – Racing Metro

The graph in Fig 3.1.3.4 represents the high speed running for these two games against the average measurement for this same variable. The first half looks to be on or below average across both games. It is apparent that the load applied across all measured variables is well below the average for the first two time segments in the ERC 6 game. The second half for the ERC 1 shows high levels of high speed running measurements which are registering well above the average reading.

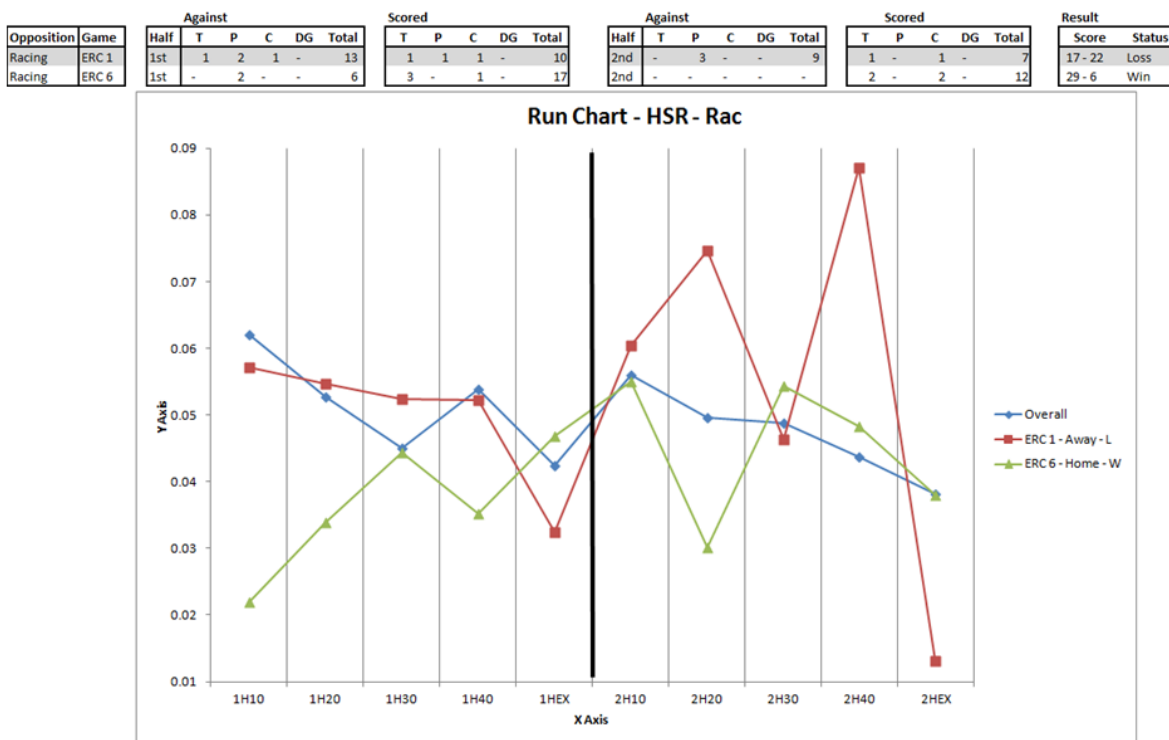


Fig 3.1.3.4 – High Speed Running Load – Racing Metro

The Fig 3.1.3.5 graph represents the Tackle Load for these two games against the average overall Tackle Load across all of the games in the season. The Tackle Load can be impacted by other variables such as possession and weather conditions etc. which were not available for this analysis. The ERC 6 game is aligning to the average measurements in the first half with a significant drop off in the load in the second half. The ERC 1 game looks like the typical game of two opposing halves with regard to Tackle Load. The load is limited for the first half which may suggest that Munster had more possession or did more attacking or the least amount of defending. Even though that may be the case, the score at half time showed Racing as having three point lead which would suggest that these scores didn't come from sustained periods of pressure but opportunist or break-away scoring. Either side of half time shows a dramatic spike in the Tackle Load placed on the Munster team which would suggest a period of dominance by the Racing Metro team.

This opposition analysis was completed for all the opposing teams within the season comprising 15 sets of analysis. Each set had five different graphs for the load variables, with the Edinburgh game having fifteen sets of graphs as there was more

opportunity for analysis due to the higher number of games through the season against this opposition. This allows Munster to build a profile per opposition.

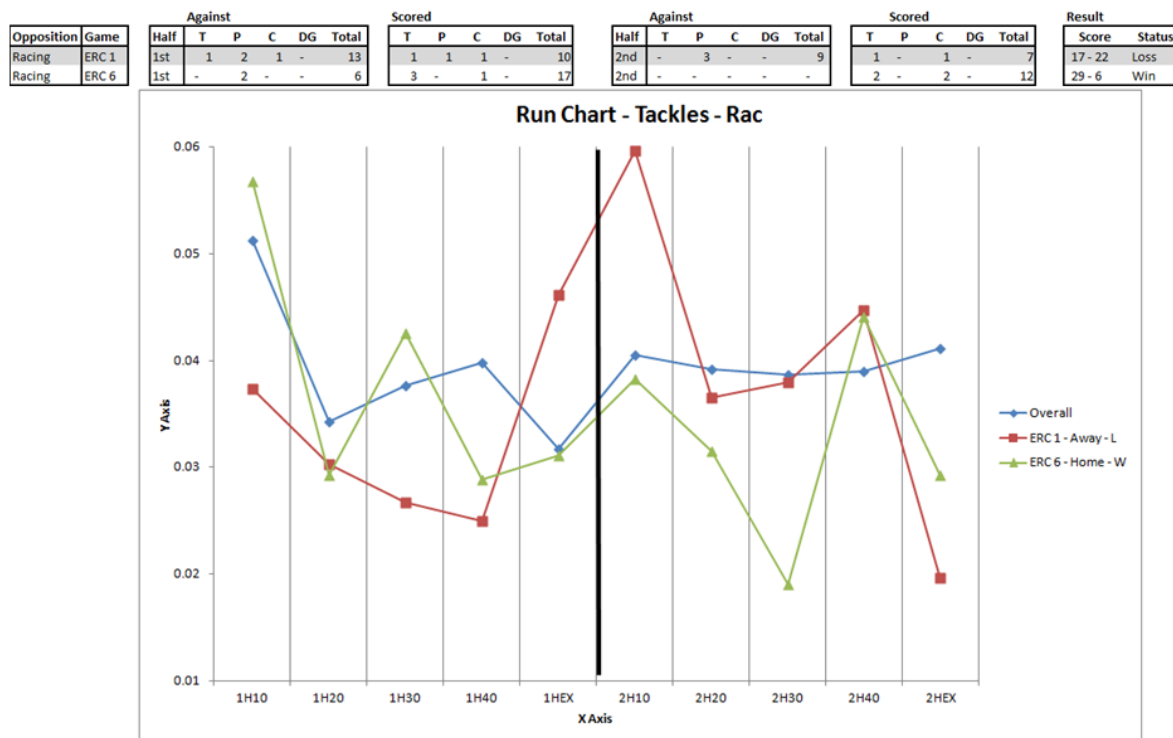


Fig 3.1.3.5 – Tackle Load – Racing Metro

Other areas of the time segmentation analysis included comparative load analysis for the RDP Vs ERC games as shown in Fig 3.1.3.6. Again this was completed for the four agreed variables as well as for the Overall Load calculation. The Overall Load shows that the RDP games on average assert more load on the players than the ERC games. The only exceptions to this are the periods at either side of half time. Also, the 70-80 minute shows a higher load for the ERC games. The analysis included 8 ERC games and 19 RDP games. Some of the ERC games would typically be 'tighter' games with less open periods of play. The ERC games also look more balanced over both halves. These games seem to be played under a more tactical influence than the RDP games as there is more at stake pending the game outcomes. This would align with opinion and the tense circumstances under which these games are typically played.

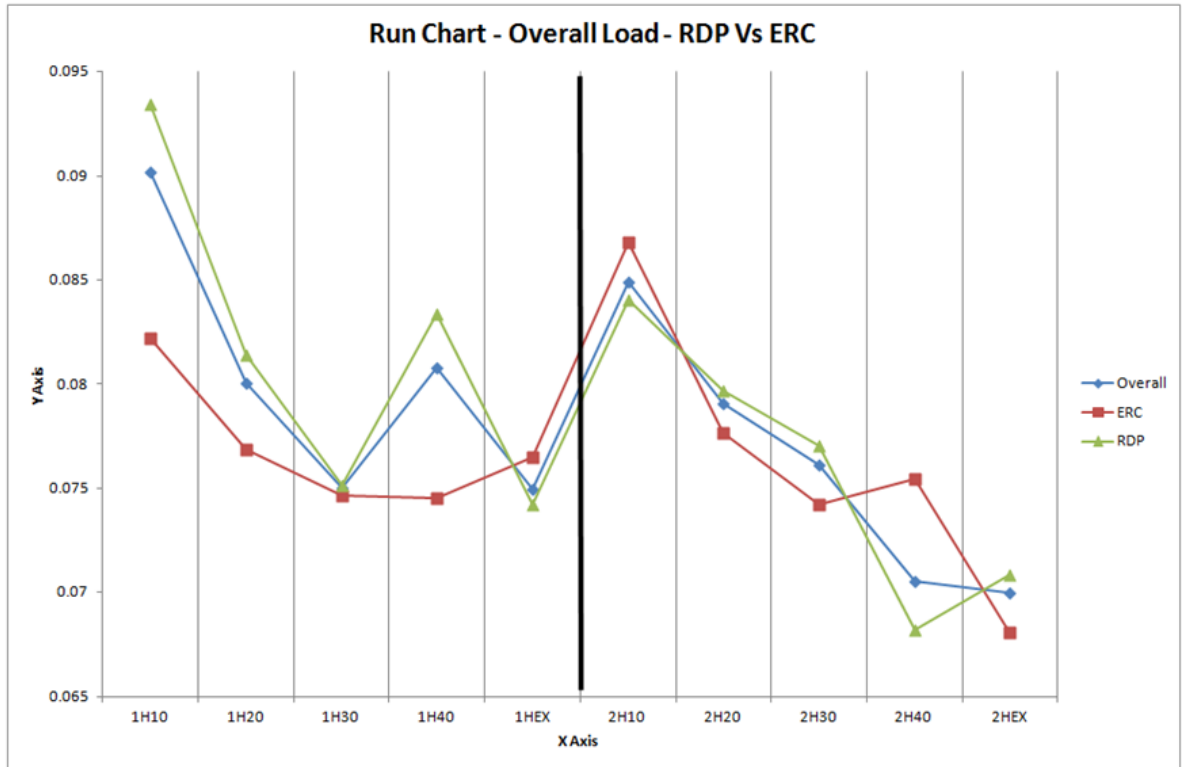


Fig 3.1.3.6 – Overall Load – RDP Vs ERC

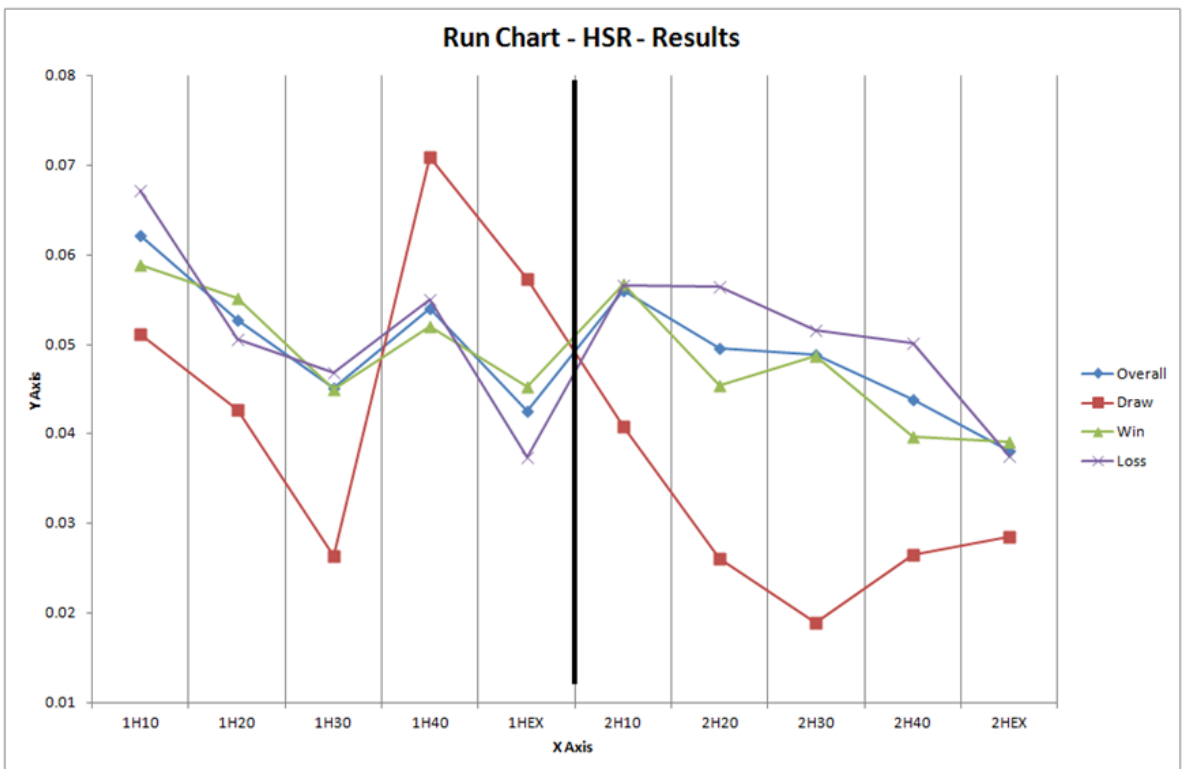


Fig 3.1.3.7 – High Speed Running Load per Result

Analysis was completed for the same load variables relative to the results of all games. This showed some interesting results. There was only one instance of a draw so some of the measurements for this game are extreme and because of the nature of this game are not aligned to the average load grades. The most notable load differential was that high speed running increased in the second half for games that Munster lost as can be seen in Fig 3.1.3.7. This would suggest that Munster tried to chase the game and maybe abandoned any degree of conservatism to try to achieve a result. It is also worth noting that the difference in the result can be due to the width of the post, a TMO (Third Match Official) decision or the bounce of a ball etc. Also, games can be won and lost in the last play of the game, so the result of the game may not reproduce like for like patterns. Munster can take these results sets and profile up and coming games for similar circumstances. It allows them to bring a certain level of predictability to their game preparations.

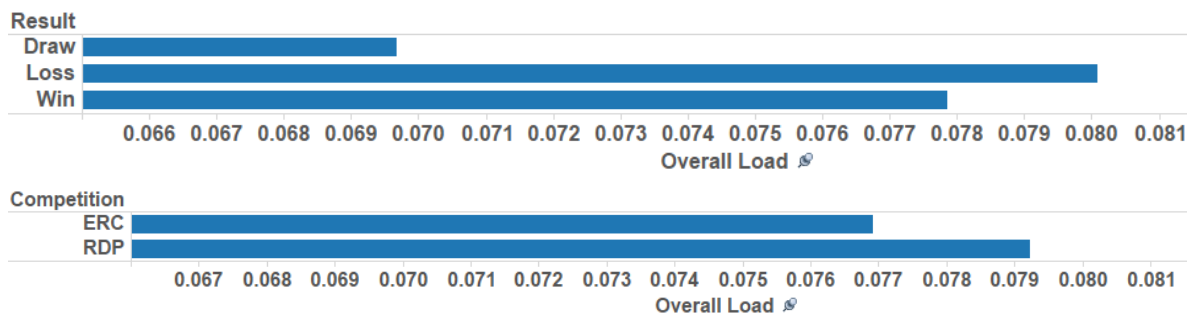


Fig. 3.1.3.8 – Overall Load by Result and Competition

The next requirement was to summarise at a high level the load behaviours outside of the 10 minute segments to give perspective to the 10 minute analysis. It would have been expected that the ERC games would apply more load than the RDP games. Some of the results can be seen in Fig 3.1.3.8. Overall Load results showed higher load being placed on the players for the RDP games over the ERC games as is also the case for games that were lost. The higher RDP load was a bit surprising and this flagged the request to analyse load changes over both halves of the games to seek out the sustenance of load across the entire match rather than spikes and troughs. As it turned out, the ERC

games showed much more consistency in the Overall Load application which can be seen in the following section 3.2.

This section of the analysis concluded with looking at each of the individual games and the opponents. Each game and each component was graded from heaviest to lightest load for all the variables as can be seen in Figs 3.1.3.9 and 3.1.3.10. Cardiff proved to be the opposition that provided the heaviest load on average for the Overall Load measurement.

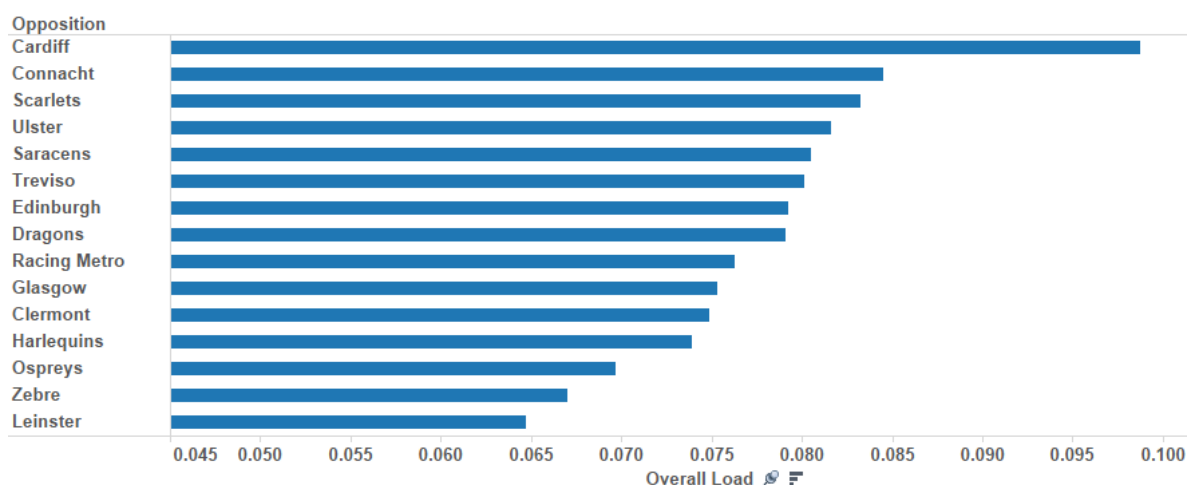


Fig 3.1.3.9 – Overall Load by Opposition

Most notably, Leinster was graded so far down the listing of the opposition that measured as placing the least Overall Load on the Munster team. For the individual variable analysis, again Leinster graded towards the end with the exception of the Tackle Load where they were graded fourth. This can be interpreted as an opposition that have lots of possession and the games seem to be tight affairs with little movement around the pitch and potentially consists of a lot of scrums, rucks and mauls. These games would contain less overall broken play and less end-to-end activity. Munster lost both of these games against Leinster in this season. Incidentally both games were played in testing weather conditions that didn't allow for a very expansive game plan.

The individual game analysis showed the two Cardiff games as being the most demanding games in terms of Overall Load. Cardiff also scored highly in the Odometer

Load, Acceleration \ Deceleration and High Speed Running loads. For the Tackle Load, they are less prominent but still in the top half for both games. Munster won one and lost one of the two games played against this opposition.

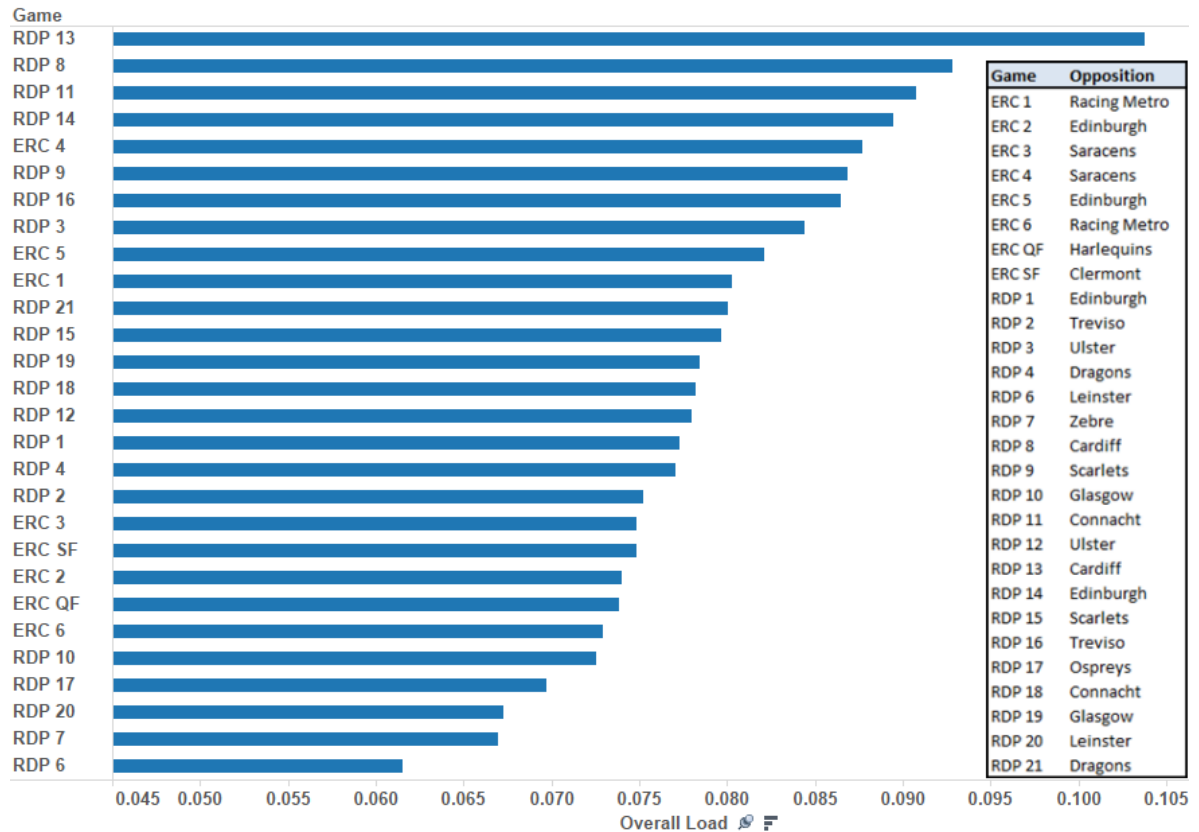


Fig 3.1.3.10 - Overall Load by Game

When the analysis was completed for this section, Munster Rugby was presented back with 106 graphs outlining the analysis summaries for all games played in the season. Pivot table analysis was also provided to ensure that Munster were able to extract all logical comparisons from the data set.

Specific result sets were provided to match activity against the next opponents so that Munster could prepare as well as possible. Munster has and will continue to use this information for the 2013/14 season for up and coming games.

Limitations

Aside from the timing calculations issues as outlined in section 2.2 there are also challenges with the time recordings. The match clock not being synchronised with the GPS units presents issues to the integrity of the data. For example if there is a period of prolonged stoppage due to a player injury or referee communications issue, this period of the measurement is skewed by this inactivity. Ideally the GPS readings should be aligned to the match clock. This would also limit the time periods of the 1HEX and 2HEX to only the time that the referee is waiting for the ball to go dead.

The load variables as reported by the GPS units are unable to define other elements of the match that can have large impact to the load on the players such as position on the pitch, possession of the ball, repeated scrummaging, weather conditions etc. These assumptions are throughout all of the research and it is under these limited grounds that all of the research is undertaken. This is a common issue with GPS sporting analysis.

Of the variables that were used within this research, it is unclear as to whether Munster has the possession of the ball for three of the four of them. For the Tackle Load it is assumed that they do not have the possession at that point. This gives us a bit more background when drawing assumptions. The Tackle Load gives us a bit more insight when reviewing the figures.

Games that are very tight or may be decided on a single incident are difficult to grade in terms of GPS output. Generally, these will be closed games in terms of movements and more like sparring affairs. These feature lowly on the GPS outputs but can be some of the tougher games in the season. We see this with the Leinster figures.

Findings

This analysis provides Munster with a profile for all opposing teams for the entire season that the information was available for. This allows them to better prepare for up and coming repeat fixtures against the same opposition. There is definite trending across

games, variables and opposition. It also provides a ranking system in terms of opposition and competition that previously had not been defined.

The segmentation of the games allows Munster to study each match against their average grades within each time segment to see if the load was above or below. It also allows them to prepare better for games for example the first period in the second half of the ERC games places the most load on the players on average for the ERC games. For the RDP games, the load applied usually spikes at the start of the games and tapers off as the games progress. This allows Munster to know what to expect in each competition. This is also available at the opposition and variable level which means that they can plan for different periods within the game and set expectations.

Having each of the games and opposition split out by variable allows Munster, for example with regard to tackle count, assume that they may not have as much possession for teams at the higher end of this load variable than for teams at the lower end as can be seen in Fig 3.1.5.1. More general analysis information is available in Appendix A Ref 3.1.B.

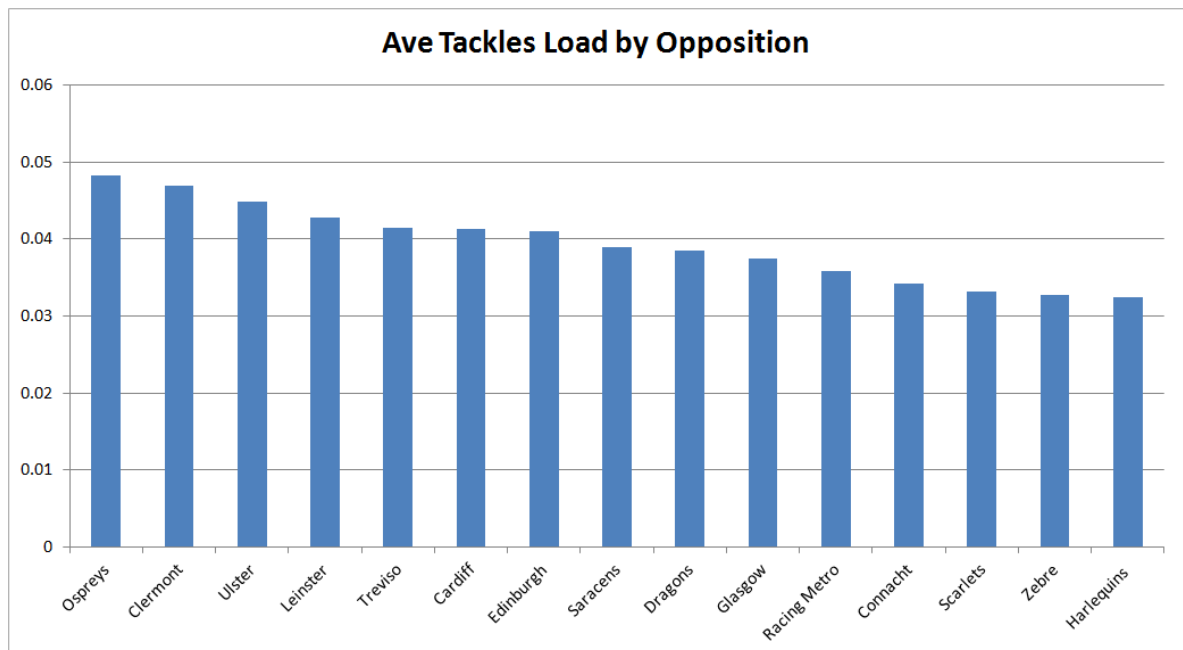


Fig 3.1.5.1 – Tackle Load by Opposition

3.2 Load Change Analysis

Background

Following on from the game analysis, there was an additional requirement for Load Change Analysis. The Load Change Analysis is an attempt to try to view the change in the load variables between the first and the second half of games. Munster Rugby was keen to explore the relationships with these changes to the loads and the other varying data elements in an attempt to spot any trending or consistency in behaviour. This would allow them to prepare the team for more prolonged periods of load placement or may emphasise the increasing of the load variables for certain games. This may have impacts on the preparation of the squad in the weeks prior to games.

Methodology

The Half column which was added to the Match Data file was used as the aggregation point. All of the four variables were averaged for these two periods for all players for each game. The differences between both halves were then analysed. The second half load was then compared against the first half to see if there was trending in terms of a positive increase or a negative decrease of load for the four agreed variables. Where it made sense, percentage changes were applied. Summing percentages to reflect higher aggregation points was avoided as this may be misleading. The results were calculated for below:

- Average drop off per game per variable
- Average drop off per result per variable
- Average drop off per competition per variable
- Average drop off per competition per result per variable
- Average drop off per opposition per variable

Analysis

These results did show some clear distinctions in behaviours. Fig 3.2.3.1 and Fig 3.2.3.2 show the delta values as well as the percentage change in load respectively across

the agreed variables for the ERC games Vs the RDP games. From the result set we can see that there were more than twice as many games for the RDP data set. The Overall Load on average increases for the ERC games but it drops off more significantly for the RDP games. For the ERC games there is an increase in the load across two of the variables with the Overall Load remaining static or slightly increasing. The most significant increase comes from the Tackle Load for the ERC games where on average there is an 8.7 increase in second half Tackle Load activity. This would suggest less possession and more defending being done by the team. The ERC games would in general be regarded as a higher quality of opposition and therefore more expectation of higher defensive load. The general increase in Overall Load would suggest that there is a concerted effort by teams to win the games in the second half. The results in the ERC games are more critical in terms of advancing within the competition and therefore teams up the effort to try to obtain the desired results. For the variables that reduced in the ERC, the drop in load is less dramatic than in RDP. This contributes to the increase in Overall Load demand.

	Count	Ave Overall	Ave Od	Ave AD	Ave HSR	Ave Tackles
ERC	8	0.0001	0.0015	- 0.0042	- 0.0004	0.0032
RDP	19 -	0.0050 -	0.0063 -	0.0082 -	0.0047 -	0.0006

Fig 3.2.3.1 – Values Changes between 1st and 2nd Halves per Competition

	Count	Ave Overall	Ave Od	Ave AD	Ave HSR	Ave Tackles
ERC	8	0.1%	1.1%	-5.0%	-0.9%	8.7%
RDP	19	-6.0%	-4.2%	-9.7%	-8.7%	-1.6%

Fig 3.2.3.2 – Percentages Changes between 1st and 2nd Halves per Competition

By contrast the RDP loads drop off across the board. Munster's RDP campaign in this season was disappointing having won as many games as they lost. Overall the drop-off signifies a lesser pace to the second half in terms of odometer, acceleration \ deceleration and high speed running as well as less defending in the form of Tackle Load. This may be because teams had either won the game and Munster has more possession or the opposition had given up the fight in the second half.

The Load Change Analysis by opposition shows some significant changes in the load variables. This analysis is shown as a percentage change to the first half. All of the percentage changes needed to be recalculated from the base file rather than apply the percentages to the changed values. This then means that there is no averaging of percentages for opposition that had more than one game against Munster in this season. Clermont who Munster only played once in the ERC semi-final, were ranked as the one of the toughest in the RPE grading. The overall load differential between the two halves is very significant at 10.9%. This game shows all of the loads increasing with the exception of the Tackle Load.

	Count	Ave Overall	Ave Od	Ave AD	Ave HSR	Ave Tackles
Cardiff	2	-3.4%	-4.7%	1.3%	-4.7%	-5.8%
Clermont	1	10.9%	5.9%	11.9%	50.8%	-2.7%
Connaght	2	-5.1%	-3.2%	-11.1%	-6.9%	7.8%
Dragons	1	-15.8%	-7.9%	-22.9%	-29.9%	-8.7%
Edinburgh	4	1.0%	1.8%	-8.3%	-2.5%	25.1%
Glasgow	2	-8.1%	-5.8%	-15.3%	-10.6%	1.8%
Harlequins	1	0.3%	1.5%	-4.3%	-0.3%	7.6%
Leinster	2	-7.8%	-7.8%	-5.2%	18.7%	-26.4%

Fig 3.2.3.3 – Variable Load Changes per Half per Opposition

This game was a game of two contrasting halves with Clermont having a lot of the possession and exerting huge pressure in the first half but tiring significantly in the second half. Although Munster lost this game they threw the ball about trying to ‘chase down the result’. This is reflected in the doubling of High Speed Running Load in the second half and also in the reduction of the tackle count.

Analysis was also done for both results and results per competition. There was only one drawn game all season against the Ospreys in RDP 17. Because the drawn result reflects only one game, there is no levelling of the data and there are big variations in the load variables. In this summary in Fig 3.2.3.4, the Overall Load drops across all games in the second half. The mixing of the RDP with the ERC games has decreased significance of the Overall Load and turned into a negative value. There is a significant increase in the Tackle Load for games that are won which could conclude that Munster were sitting on

their lead as all of the other load variables decreased. The decrease in the Tackle Load for the lost games would suggest that Munster had more possession and were trying to retrieve the result.

	Count	Ave Overall	Ave Od	Ave AD	Ave HSR	Ave Tackles
Win	14	-4.0%	-3.7%	-7.6%	-9.6%	11.2%
Loss	12	-3.1%	0.1%	-8.9%	-0.6%	-5.6%
Draw	1	-23.4%	-20.2%	-13.0%	-42.0%	-27.3%

Fig 3.2.3.4 – Variable Load Changes per Half per Result

In Fig 3.2.3.5, the break-out of win \ loss per competition gives us further insight into the change in the games' load variables between the two halves of play. There are less overall data sets per sub group and changes are more dramatic.

	Count	Ave Overall	Ave Od	Ave AD	Ave HSR	Ave Tackles
ERC						
Win	5	-3.5%	-3.0%	-6.0%	-11.6%	10.9%
Loss	3	5.9%	7.8%	-3.4%	17.4%	5.4%
RDP						
Win	9	-4.3%	-4.1%	-8.5%	-8.6%	11.3%
Loss	9	-5.9%	-2.4%	-10.7%	-5.6%	-9.1%
Draw	1	-23.4%	-20.2%	-13.0%	-42.0%	-27.3%

Fig 3.2.3.5 – Variable Load Changes per Half per Competition per Result

Interestingly, the only grouping that is showing an average increase in Overall Load is the lost category in the ERC competition. The most significant of these factors is the high speed running again suggesting more broken play possibly with Munster trying to chase the game and attacking from further out the pitch. The reversal of this example for the ERC win category shows a large drop in the HSR value which could signify Munster having more control of the game and the result and are subsequently attempting to 'kill off the game'.

Fig 3.2.3.6 is a graphical representation of both the Overall Load variables per competition game per result as well as the change in load per competition game per result. This allows for a visual comparison of both competitions as well as highlighting upper and lower limits. It also plots the same data sets in logical groupings which allows

for tiered analysis purely on visual inspection i.e. it can be seen that on average there are more percentage decreases in load change in RDP games than ERC accounting for relative data sizes. It can also be seen that for two of the three ERC games that Munster lost, there was a considerable change in the Overall Load in the second half. This analysis was repeated for the four other variables and can be viewed in Appendix B Ref 3.2.A

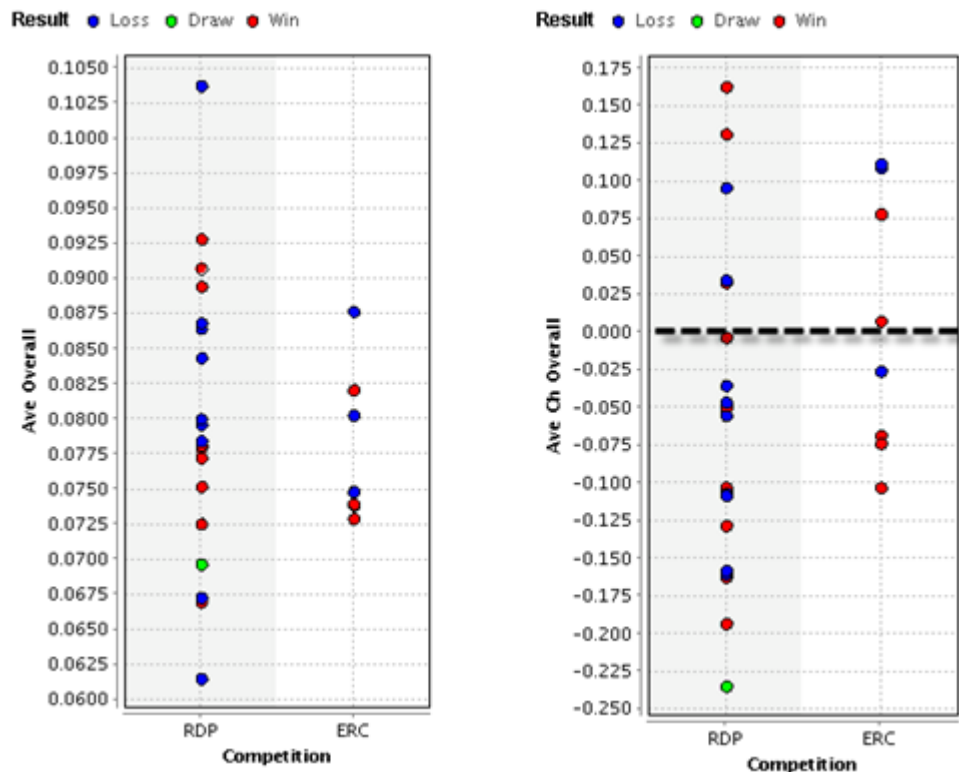


Fig 3.2.3.6 – Overall Load and Changes per Overall Load per Competition per Result

When the results were discussed with Munster, these figures aligned to the intuitive beliefs that the ERC games applied a greater standard or consistency of load. While the Overall Load may be lower for the ERC games, the variance is much less suggesting tighter matches in the ERC competition which is a common perception.

Limitations

The data set used for this analysis consists of 27 different values. This is a relatively low set of values to research in order to draw conclusive findings. Also, Munster had a particularly poor season in terms of results so the win:loss data sets may be more

balanced now but limited in terms of future analysis. The analysis variables that are used are averaged for all the halves of the games so a half is the lowest level of granularity. The values used to calculate the half values vary in size and are not consistent. For example one half of information can consist of $45 * 1 * 12$ readings as is the case with Harlequins ERC QF for 45 minutes played, 1 minute GPS readings and 12 players using the GPS which gives averaged values of 540 records. This can be compared to $53 * 1 * 11$ for Glasgow or $50 * 1 * 7$ for the Scarlets game which gives the averages over 583 and 350 records respectively.

Some games can behave very differently in terms of activity. For example if there are extreme weather conditions, first half possession playing into a strong wind would expect to be high and result in a lower tackle count as opposed to playing a half with less possession. Games that are badly broken up by player errors due to wet conditions resulting in scrums tend to get lost within the GPS readings. Certain factors like fatigue, 'player drain' or 'period of the season' can also impact the GPS readings as well as game intensity and occasion. Games may be won or lost by half time and this can lead directly to a change in tactics resulting in different GPS behaviours over the two halves. Also categorising games GPS activity as a win or a loss when the game could have been won or lost with the last play of the game and expecting different trending may be difficult. Nevertheless the figures did give very useful insight into the load change activities.

Findings

The Load Change analysis provides some very valuable information that can be utilised for planning and preparing players for games that may require a more concerted effort over both halves of a game. Seeing the sustenance of load for the ERC cup games would correspond with the player's rankings of games in terms of difficulty. It also suggests that the games are fought for in a much more consistent and persistent manner than the RDP league games. This may lead the Strength and Conditioning Team to adopt a less intensive build up to ERC game weeks. It can also help to pinpoint optimal conditioning and stamina levels to peak for these periods of the season. Spotting

behaviours like increased second half Tackle Loads for games in which Munster would have won would suggest that they had less possession in the second half of those games or the games were a bit more broken up.

The ability to see the oppositions' attitude towards first and second half and sustenance of applied load is a valuable piece of information. For example, if there was to be similar type of match to the ERC semi-final, Munster could plan for a first half of high tackle count which would reduce in the second half. There would be a serious break-up of the game in the second half which is reflected in a 50% increase in HSR. This would be typical of most French opposition in cup games on their own home grounds. They would usually come out strong and attempt to 'blow teams away' and lay down an early marker for the opposition. Most teams cannot deal with the intensity and suffer both on the scoreboard and also in fitness and ability to recover physically and apply their game plan due to fatigue.

The information gathered here should be re-applied to subsequent season's data to build a bigger sample size to see if there is consistency across competitions, games and opposing teams. Munster was pleased with these results sets as they confirmed beliefs related to this area. The value of the information sets returned provides a different focus on the fitness requirements in terms of endurance and game pace management.

3.3 RPE Ratings Vs GPS Output

Background

Rate of Perceived Exertion (RPE) is a [12] psycho-physiological scale, meaning it calls on the mind and body to rate one's perception of effort. The RPE scale measures feelings of effort, strain, discomfort, and \ or fatigue experienced during both aerobic and resistance training. One's perception of physical exertion is a subjective assessment that incorporates information from the internal and external environment of the body. The greater the frequency of these signals, the more intense the perceptions are of physical exertion. Munster's Strength and Conditioning Team are tasked with gathering measurements from each player after each game.

The players are asked to assign a difficulty grade from 1 to 10 based on the game just played. This is called the RPE value. This is usually performed immediately in the aftermath of a game. This has been agreed as being the best timeframe to gather the information. There are some difficulties in the logistics of this task e.g. players who are injured or unavailable. Some of these measurements were also provided in half units. The purpose of this is to be able to gauge what games players find difficult for grading of opposition. The RPE is then multiplied by the number of minutes that the player was on the pitch. This gives an overall RPE load score. RPE is a common classification methodology used across various sports. It is relatively simple to gather the data sets for the measurement and there is also little overhead associated. The more discipline and understanding of the requirements the players have, the more beneficial the use of the analysis can be applied. It is subjective in nature and therefore does have limitations regarding the idiosyncratic nature of individuals providing the source data. The requirement is to seek correlation between the GPS readings and the RPE file.

Methodology

Both the GPS and RPE files are at different baselines so there is a need to align both files. The GPS file is at a 'per minute per player' level whereas the RPE file is at a

‘game per player’ level. The GPS readings were averaged per player per game. All of the players whose RPE grades had corresponding GPS readings were used in this analysis. All of the players whose GPS readings didn’t have an equivalent RPE grading were eliminated from the analysis. It was agreed with Munster that any player who had four or less readings or who played for less than 55 minutes was not included in this analysis. Players are each categorised into groups for each game. The players who play all of the minutes in the game are classified as ‘full’, the players who have participated in 55+ minutes are classified as ‘most’ and the balance are classed as ‘subs’.

Games were also categorised in terms of difficulty by player. As discussed with Munster, the grade of 7.5 for the RPE was agreed as being the middle value in terms of easy and hard. Any value above this grade was classified as hard for that player and any game below that value were classified as easy for that player. There were two exceptions to this where players RPE grades were limited in terms of range of values. We agreed that for these two players, Mike Sherry and Denis Hurley that the value should be set to 7. The logic was then applied to all measureable data.

Player	Position	Time	RPE	LOAD Category
BJ Botha	Prop	80	7	560 Easy
David Kilcoyne	Prop	79	8	632 Hard
Donnacha Ryan	Lock	80	6	480 Easy
Ian Keatley	Fly Half	80	9	720 Hard
JJ Hanrahan	Fly Half	8	4	32 Easy
Paddy Butler	Back Row	7	6	42 Easy
Paul O'Connell	Lock	61	8	488 Hard
David Kilcoyne	Prop	57	5	285 Easy
Marcus Horan	Prop	23	5	115 Easy

Fig 3.3.2.2 – RPE and Categorisation per Player per Game

The range as well as the average of the RPE values can be seen in the Treemap in Fig 3.3.2.3 with the area block representing the averages and the shading representing the variance with the darkest being the highest variance. In this graphic, Paul O Connell has the highest level of RPE grade on average for all of the available records within the data set. In contrast, Denis Hurley has the lowest average RPE value. Conor Murray has the

highest range in terms of RPE values within the data set while both James Downey and Felix Jones show lower levels of variance. These are very important to the authenticity of the data that is provided by these players and is important to understand and control as this is a subjective variable. For example providing low variance in RPE values implies that the player does not distinguish between Hard and Easy games. This undermines any comparative analysis with GPS outputs.

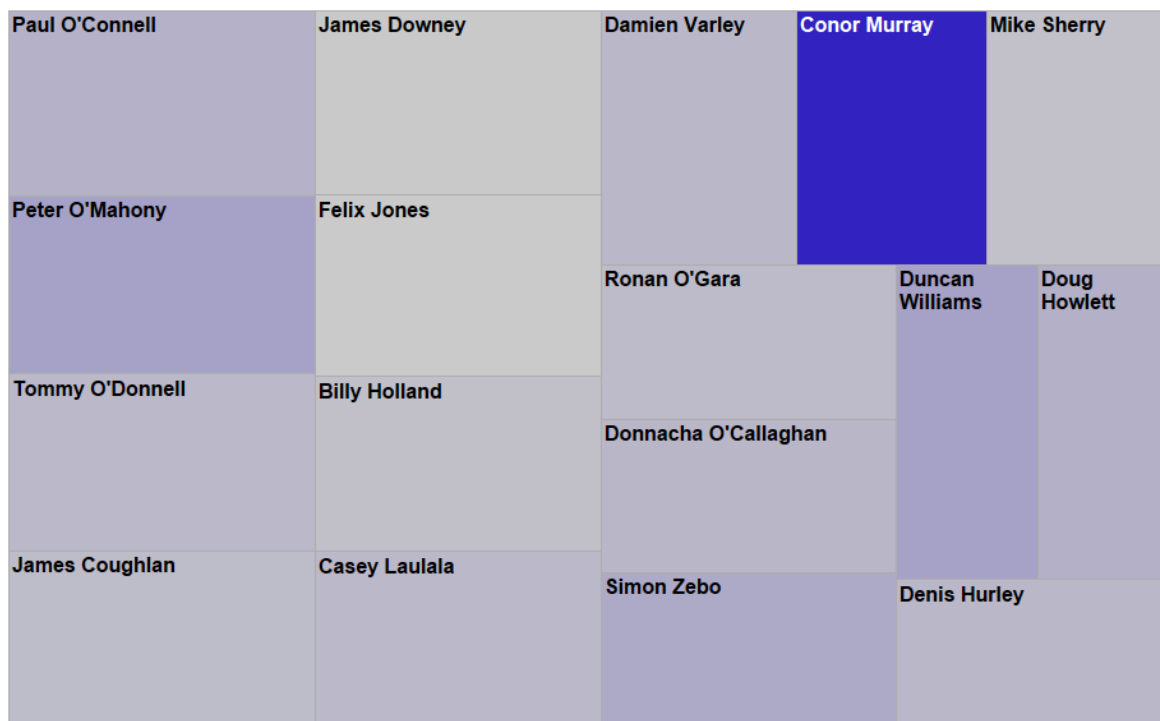


Fig 3.3.2.3 – Treemap for RPE Averages (Volume) and Range (Shade)

Analysis

The analysis is done at various different levels – game, opposition, general position and at an individual level. There are 176 workable RPE records that had corresponding GPS values that were used for the analysis. For the player analysis, there were 164 available data sets as some of the players did not have the required number of values to do individual analysis.

The first set of analysis is limited to the available values from the RPE file. It looks at the games that were played throughout the season and ranked opposition and games accordingly as seen in Fig 3.3.3.1. The results were not surprising with the games at the

latter end of the ERC featuring amongst the highest ranks in terms of difficulty ranking. Also, the Italian teams who are traditionally poor performers in the RDP are towards the end of the table. These figures are for the accumulated values against these teams. For example, the figure for Clermont represents all available values for the one game that Munster played against Clermont in the season. However, the Edinburgh figures represent all the values for the four different games that Munster played against that team in the season. These two teams were in the same ERC pools as each other.

Row Labels	<input type="button" value="▼"/> Ave LOAD	Count LOAD	Row Labels	<input type="button" value="▼"/> Ave LOAD	Count LOAD
Clermont	663.20	10	ERC SF	663.20	10
Harlequins	632.27	11	ERC 5	647.38	8
Ospreys	612.00	9	ERC QF	632.27	11
Leinster	612.00	9	RDP 20	612.00	9
Edinburgh	592.29	21	RDP 17	612.00	9
Saracens	569.00	16	ERC 4	605.71	7
Connacht	565.50	20	RDP 11	583.20	10
Racing Metro	548.00	7	ERC 2	565.60	10
Ulster	535.44	9	ERC 6	548.00	7

Fig 3.3.3.1 – RPE Loads and Counts per Opposition and Game

The game analysis ranks the individual games as they were graded by the players, which is also shown in Fig 3.3.3.1. There were a total of 22 games graded that have been included in the rankings. The RPE analysis was not available for some of the games as the measurement was only implemented after the season had begun. The consistency of the timing of the measurement was something that Munster put focus on so that the grades recorded would be comparable. The ERC games are prominent towards the higher end of the grades with the 7 ERC games played in the season in the top 11 values of the overall 22. The players are grading these games as being more difficult than the RDP with the last 11 values in the ranking all coming from the RDP games.

For the game analysis, the Easy \ Hard grades were grouped to review the corresponding GPS data for these readings. There were three different levels of playing time and the analysis was done for each. All levels were also analysed as one. In Fig

3.3.3.2 and Fig 3.3.3.3 respectively the output can be seen for the Full and Most minutes played.

Play Time	Full				
	Ave Overall	Ave Od	Ave AD	Ave HSR	Ave Tackle
Easy	0.0830	0.1528	0.0885	0.0526	0.0382
Hard	0.0845	0.1520	0.0888	0.0500	0.0472

Fig 3.3.3.2 – Game Categorisation for 4 Variables of Players for Full Game Minutes

Play Time	Most				
	Ave Overall	Ave Od	Ave AD	Ave HSR	Ave Tackle
Easy	0.0806	0.1489	0.0822	0.0567	0.0347
Hard	0.0745	0.1390	0.0645	0.0447	0.0496

Fig 3.3.3.3 – Game Categorisation for 4 Variables of Players for Most Game Minutes

There was only one variable that gave consistent results for all of the analysis. For the Hard category the Tackle Load was higher in all instances than the Easy category. All other variables showed inconsistency as follows. For Odometer, the Hard:Easy ratio had higher value ratios of 1:2. For Acceleration\Deceleration, the Hard:Easy ratio had higher value ratios of 2:1. For HSR, the Hard:Easy ratio had higher value ratios of 1:2. For Tackle, there was consistency for the Hard value in all instances being higher than the Easy value.

This can also be highlighted in the box plots as outlined in 3.3.3.4. Here we can see minimal overlap for the variables by position and by game difficulty category. The only exception here is for the Tackle Load whereby there is more distinction in the values across the Easy and Hard categories. The distinction between Hard and Easy games at a positional level shows inconsistency in terms of the GPS datasets across three of the four variables. There is no GPS output variable that correlates with the game grading. The closest potential variable as can be seen in this graphic is the Tackle Load.

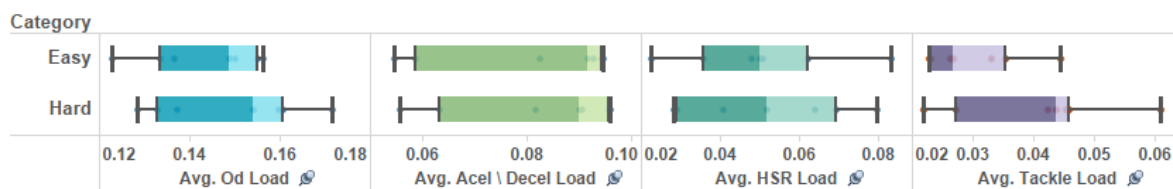


Fig 3.3.3.4 – Box Plots of Positional Groups Game Categorisation against Variable Loads

After the grades for Mike Sherry and Denis Hurley were normalised, the table output in Fig 3.3.3.5 is the resulting Easy and Hard grades for each of the general positions. This summary is only inclusive of the players that played the Full or Most parts of each of the games and also only includes players that have had a dataset of 4 or more in quantity.

	Easy	Hard	Grand Total
Back Three	25	16	41
Centre	10	17	27
Fly Half	7	2	9
Scrum Half	7	4	11
Back Row	12	21	33
Lock	15	9	24
Hooker	8	11	19
Grand Total	84	80	164

Fig 3.3.3.5 – Game Categorisation Counts by General Position

The games level analysis looked at two different areas, the first one being a results-based analysis and the second one being a competition based analysis. For the results analysis, there was only one draw throughout the season which skews the results slightly as shown in Fig 3.3.3.6. The loads placed on the players are higher for losses than it is for the wins. Again in every instance, the loads are lower for the hard category across all load variables with the exception of Tackle Load. Incidentally the only drawn game counters this behaviour and the loads placed look to align much more directly with the RPE ratings with the hard games having higher values for all variables.

	Ave Overall		Ave Od		Ave AD		Ave HSR		Ave Tackle	
	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
DRAW	0.0629	0.0771	0.1405	0.1439	0.0543	0.0607	0.0236	0.0431	0.0334	0.0607
LOSS	0.0835	0.0819	0.1516	0.1498	0.0903	0.0810	0.0581	0.0462	0.0341	0.0505
WIN	0.0812	0.0802	0.1500	0.1469	0.0845	0.0839	0.0525	0.0460	0.0378	0.0440

Fig 3.3.3.6 – Game Categorisation by 4 Variables by Result

For the competition comparisons in Fig 3.3.3.7 again there is consistency with the Tackle Load from previous analysis where the hard games are reflected as having a higher value. In the RDP games, there is more parity with the result sets than the ERC games. The Odometer Load is higher for the hard category than the easy category and the Overall Load value is higher as a result of both the odometer and tackle activities. For the ERC

games, the values again show higher on the easy games with the exception of the Tackle variable.

	Ave Overall		Ave Od		Ave AD		Ave HSR		Ave Tackle	
	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
ERC	0.0807	0.0775	0.1492	0.1439	0.0824	0.0786	0.0526	0.0419	0.0385	0.0455
RDP	0.0819	0.0831	0.1508	0.1509	0.0870	0.0827	0.0541	0.0492	0.0356	0.0497

Fig 3.3.3.7 – Game Categorisation by 4 Variables by Competition

The individual analysis looks at the individuals RPE grades and loads against each individual's equivalent GPS readings. All of the players that met the criteria of minimal requirements in terms of counts of records and times spent playing for each of those records were then analysed on both a player and an individual record basis.

Row Labels	Ave Overall		Ave Od		Ave Accel \ Decel		Ave HSR		Ave Tackle	
	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
Back Three	0.0891	0.0922	0.1614	0.1713	0.0942	0.0875	0.0636	0.0640	0.0373	0.0460
Denis Hurley	0.0971	0.0998	0.1626	0.1630	0.1186	0.1236	0.0687	0.0684	0.0383	0.0440
Doug Howlett	0.0812	0.1059	0.1593	0.1941	0.0715	0.0828	0.0575	0.0940	0.0367	0.0527
Felix Jones	0.0892	0.0893	0.1838	0.1811	0.0657	0.0637	0.0721	0.0625	0.0352	0.0499
Simon Zebo	0.0890	0.0818	0.1494	0.1548	0.1087	0.0843	0.0598	0.0502	0.0381	0.0377
Centre	0.0896	0.0851	0.1544	0.1533	0.1137	0.0926	0.0522	0.0503	0.0384	0.0442
Casey Laulala	0.0938	0.0949	0.1485	0.1545	0.1366	0.1297	0.0494	0.0515	0.0409	0.0438
James Downey	0.0833	0.0764	0.1632	0.1521	0.0793	0.0597	0.0564	0.0492	0.0345	0.0446
Fly Half	0.0796	0.0839	0.1511	0.1597	0.0911	0.0901	0.0528	0.0640	0.0234	0.0218
Ronan O'Gara	0.0796	0.0839	0.1511	0.1597	0.0911	0.0901	0.0528	0.0640	0.0234	0.0218
Scrum Half	0.0969	0.0852	0.1641	0.1550	0.0990	0.0847	0.1025	0.0715	0.0219	0.0297
Conor Murray	0.0922	0.0834	0.1673	0.1563	0.0848	0.0893	0.0975	0.0656	0.0192	0.0225
Duncan Williams	0.0987	0.0906	0.1628	0.1509	0.1046	0.0709	0.1045	0.0891	0.0230	0.0513
Back Row	0.0808	0.0787	0.1397	0.1357	0.0864	0.0818	0.0468	0.0369	0.0504	0.0604
James Coughlan	0.0777	0.0797	0.1370	0.1431	0.0773	0.0743	0.0462	0.0435	0.0504	0.0579
Peter O'Mahony	0.0659	0.0711	0.1184	0.1214	0.0707	0.0857	0.0271	0.0213	0.0474	0.0558
Tommy O'Donnell	0.0922	0.0829	0.1545	0.1364	0.1030	0.0887	0.0591	0.0396	0.0523	0.0669
Lock	0.0687	0.0697	0.1413	0.1355	0.0606	0.0660	0.0385	0.0308	0.0346	0.0465
Billy Holland	0.0718	0.0711	0.1500	0.1418	0.0673	0.0727	0.0451	0.0340	0.0249	0.0359
Donnacha O'Callaghan	0.0672	0.0758	0.1367	0.1469	0.0570	0.0542	0.0352	0.0409	0.0401	0.0610
Paul O'Connell	0.0669	0.0668	0.1386	0.1264	0.0602	0.0622	0.0352	0.0251	0.0337	0.0534
Hooker	0.0613	0.0665	0.1314	0.1327	0.0545	0.0601	0.0217	0.0289	0.0376	0.0443
Damien Varley	0.0611	0.0671	0.1265	0.1282	0.0577	0.0773	0.0153	0.0155	0.0448	0.0477
Mike Sherry	0.0615	0.0663	0.1363	0.1337	0.0513	0.0563	0.0280	0.0319	0.0304	0.0436
Grand Total	0.0816	0.0805	0.1504	0.1477	0.0858	0.0808	0.0537	0.0458	0.0364	0.0478

Fig 3.3.3.8 – Game Categorisation by 4 Variables by General Position and by Player

Table 3.3.3.8 outlines this data. This also shows the general positional values as well as the total values for each of the four variables and the corresponding Easy \ Hard grading as calculated from the individuals RPE ratings. These figures are averaged based on the records for each of the players. This allows for analysis at an individual level of the

players grades relative to the RPE categorisation. For example, all of Damian Varley's figures are reading as higher values for the Hard category than for the Easy category. On the other hand, if we look at Simon Zebo's figures the readings are for the hard categorised games are shown to be lower than the Easy games for three of the four games with the only exception being the odometer figures as the Hard games are showing as a much higher value. This table can help to compare and contrast different player's figures as well as different positions to see what players grade the games differently depending on the different loads from the games.

For the individual player analysis, each record of the players is analysed to seek correlation between the player's GPS and RPE on a per match basis. RPE Load (RPE * Minutes Played) is also measured to seek correlations with the GPS. All of the individual player's data can be seen in the Appendix C Ref 3.3.A.

Peter O'Mahony	Correlation				
	Overall	Od	Accel / Decel	HSR	Tackles
Correlations of RPE	0.71649	0.51722	0.74196	-0.71997	0.71977
Correlations of Load	0.67314	0.44614	0.64585	-0.47124	0.67329

James Coughlan	Correlation				
	Overall	Od	Accel / Decel	HSR	Tackles
Correlations of RPE	0.23207	0.40086	-	0.10380	-0.09426
Correlations of Load	0.18941	0.40814	-	0.17932	-0.10516

Fig 3.3.3.9 – Correlation Summary by 4 Variables and RPE

Fig 3.3.3.9 shows result samples for Peter O Mahony and James Coughlan showing the correlations. If the cells are not highlighted, there is no correlation between the Load or the RPE and the GPS readings. For the yellow highlighted, there is a mild correlation between the sets of data. For these particular examples there is not any instance of a strong correlation. For James Coughlan there is no correlation between his grades applied to each game and those corresponding games' GPS figures. For Peter O Mahony there is correlation across three of the four variables. Note that there are also inverse correlations i.e. the higher a load variable value was, the easier the game was graded by the player. In

this instance, there is a correlation between the player grading games as lower difficulty even though the HSR for those games has increased.

All of the players' correlation analysis was centralised to one summary file for ease of reading and comparison. This allowed for players in similar positions to be directly compared in terms of their grades and correlations to their GPS outputs. Overall there were seven of the seventeen players that showed no correlation at all against each of the eight readings – RPE and Load against the four variables ($2 * 4$). There were seven players that showed a limited level of correlation with less than 50% of GPS variables. There were three players that showed correlation for 50% or higher against the variables. The Overall Load was also used in the analysis but this is a representation of the four individual variables.

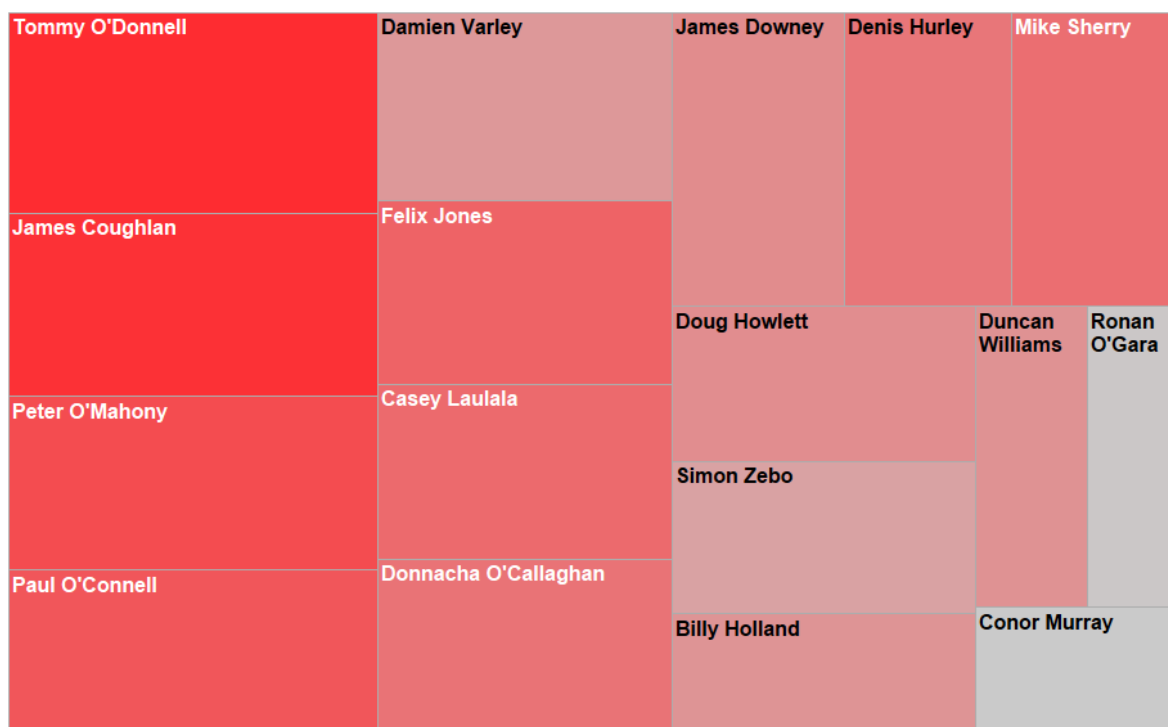


Fig 3.3.3.10 – Treemap Tackle Load by Average (Volume) and Maximum (Shade) by Player

All of the loads were graphed at an individual level to give both the highest average readings as well as the maximum reading per variable. Fig 3.3.3.10 shows an example of the tackle count. It is not surprising that the Back Row players feature strongly in the output. The area represents the highest averages and the shading represents the

maximum recorded values. See Appendix C Ref 3.3.B for the other Treemap examples showing the other variables. Tackle Load is recurring in its significance in terms of impact and Fig 3.3.3.10 shows the spread of this load. The Half Backs feature towards the lower end of both scales. This is in direct contrast to the Back Row and would suggest that they are well protected by the workloads of the Back Row players. The Half Backs' positions require them to manage the game. Getting involved in physical contact particularly without possession should be avoided where possible to help protect them and the team's tactical game plans.

For the positional analysis, the generic positions of Back Three, Centre, Fly Half, Scrum Half, Back Row, Lock and Hookers were used. Sample result sets can be seen in Fig 3.3.3.11. Each player from within their respective group had their loads (RPE * Minutes) combined and measured against the combined GPS. This then allowed for individual comparison against the positional values. All positions showed some examples of correlation with 4 of the 7 showing a strong correlation.

Lock	Variables	Ave Overall	Ave Od	Ave Accel \ Decel	Ave HSR	Ave Tackle
Correlation with Average RPE	24 -	0.1058	0.1663	0.0930	0.4688	0.2877
Correlation RPE with Losses	9 -	0.0298	0.4246	0.0331	0.5813	0.6857
Correlation RPE with Wins	14 -	0.0332	0.0334	0.2643	0.2357	0.4006
Correlation RPE with RDP	19 -	0.0694	0.0876	0.1420	0.4277	0.2204
Correlation RPE with ERC	5 -	0.1456	0.2368	0.2623	0.6495	0.3068

Hooker	Variables	Ave Overall	Ave Od	Ave Accel \ Decel	Ave HSR	Ave Tackle
Correlation with Average RPE	18	0.3718	0.1953	0.6160	0.0662	0.2256
Correlation RPE with Losses	9	0.4408	0.3035	0.6295	0.2866	0.5079
Correlation RPE with Wins	8	0.4827	0.4455	0.7778	0.0287	0.0401
Correlation RPE with RDP	12	0.5227	0.3046	0.5055	0.2305	0.6231
Correlation RPE with ERC	7	0.6217	0.4864	0.8363	0.4472	0.1340

Fig 3.3.3.11 – General Position RPE Correlations with 5 Variables

Interestingly for the Hooker readings, there is a consistent level of correlation between the players' RPE and the Acceleration \ Deceleration Load. This would suggest that the more high intensity acceleration \ deceleration activity, the tougher the Hookers find the game. This kind of information can be used to re-focus training for particular player groupings on areas of difficulty.

Based on the values within this positional data set, the Decision Tree methodology was used to try to apply a model to the data. [13] Decision Trees are a predictive model used to determine which attributes of a given data set are the strongest indicators of a given outcome. They [14] classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute. In this instance the possible values are Easy or Hard which represent the level of game difficulty. A file containing each of the players RPE and variables totals per game is used in this analysis.

The data was manipulated to allow the model to conform to compliance percentages. All attempts that didn't conform can be seen in Appendix C Ref 3.3.C. The data sets that did conform were for two different groups within the forwards; the Front 5 players less Donnacha O Callaghan (DOC) as seen in Fig 3.3.3.12 and the Back Row only as seen in Fig 3.3.3.14. Both of these sets of data show a total count of 55 and 43 and accuracy levels of 70.33% and 68.55 respectively. Both sets of data show sufficient variable compliance percentages to ensure the models' accuracies. Another decision tree with multiple tiers for the Front 5 can be seen in Appendix C Ref 3.3.D which has accuracy levels of 76%.

accuracy: 70.33% +/- 15.52% (mikro: 70.91%)			
	true Easy	true Hard	class precision
pred. Easy	29	11	72.50%
pred. Hard	5	10	66.67%
class recall	85.29%	47.62%	

Fig 3.3.3.12 – Decision Tree Accuracy - Front 5 less Donnacha O Callaghan

Fig 3.3.3.13 shows the decision tree output for the Front 5 players less DOC data set. It outlines the decision-making variables \ nodes and the tolerance to which the decisions are dependent. The leaf nodes then signify a category for which the inputted data should then reside.

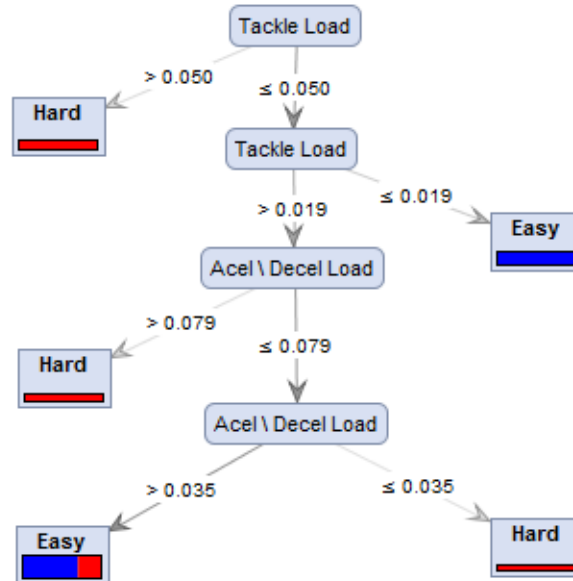


Fig 3.3.3.13 – Decision Tree for Front 5 Less DOC for Easy Hard Game Classification

accuracy: 68.50% +/- 20.25% (mikro: 67.44%)			
	true Hard	true Easy	class precision
pred. Hard	18	9	66.67%
pred. Easy	5	11	68.75%
class recall	78.26%	55.00%	

Fig 3.3.3.14 – Decision Tree Accuracy for Back Row Only

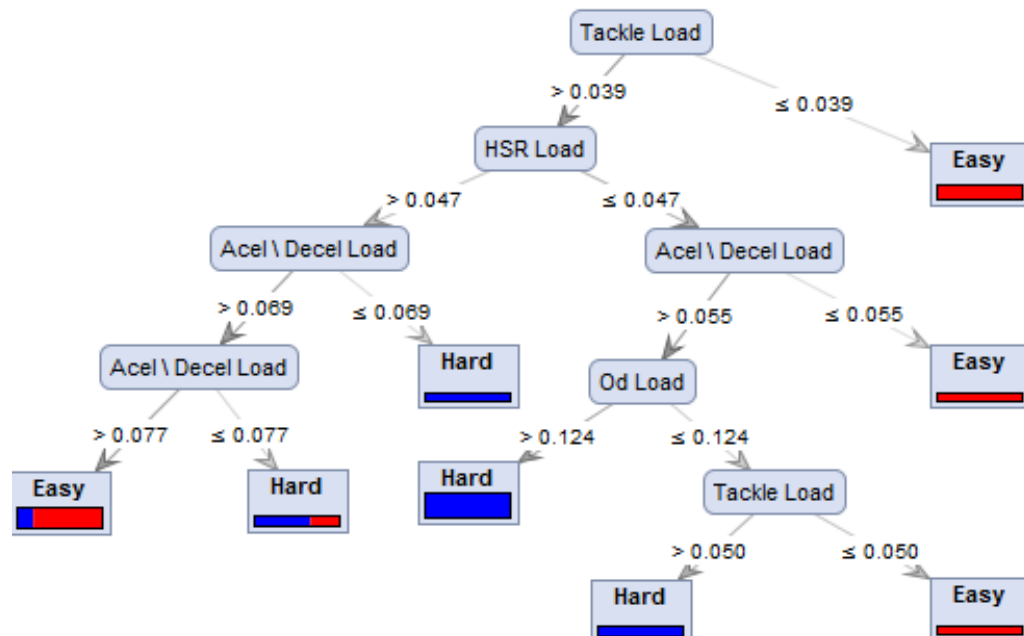


Fig 3.3.3.15 – Decision Tree for Back Row for Easy Hard Game Classification

Fig 3.3.3.15 represents the decision tree for the Back Row players. Each time the data is put through an If:Else loop it dictates the flow of the data record through the

decision tree. If the record meets the branch criteria it will flow through to the next node point. If this is a leaf node, then it will be assigned an attribute value. If it is not a leaf node, it moves to the next decision node which will dictate its next course through another If:Else set of criteria. The most notable item from both trees is that both begin with a tackle value check and this is sufficient to assign a value to the data record. This aligns with the previous results that Tackle Load is consistent in its level of impact to the players in terms of load. [15] When classification cost is important, decision trees may be attractive in that they ask only for the values of the features along a single path from the root to a leaf. In terms of accuracy, decision trees have been shown to be competitive with other classifiers for several learning tasks.

Limitations

RPE (Rate of Perceived Exertion) is a commonly used approach in sporting analytics. [16] RPE is a recognized marker of intensity and of homeostatic disturbance during exercise. It is typically monitored during exercise tests to complement other measures of intensity. It is known to have its flaws due to its subjective nature. Any subjective variable can be impacted by multiple factors. Lamb et al looked at the reliability of RPE and found that [17] ratings recorded during graded exercise testing do not match the levels of relative physiological intensity that they are assumed to. Fundamental to this, concern over the validity of the RPE scale is the issue of its reliability, as a measurement tool cannot be deemed valid without also being reliable.

There are circumstances under which the RPE values assigned can be influenced. Often psychological factors like emotions and feelings such as post match elation \ deflation or personal disappointment in performance levels can impact the RPE grade provided. Also the consequential analysis of the RPE value may push certain players to 'play safe' and declare similar RPE values across varied levels of game difficulties.

For some players the range in values for games played in the full does not allow for enough variation to allow for the data quality to be useful. The GPS will provide much larger variances, which when matched against non-varied RPE grades limits the amount of

analytical value that can then be applied. Also the scale of 1 to 10 may not be wide enough for players to accurately apply their scores. Typically RPE is measured within a 6 to 20 range. With a wider data range, variance should grow relatively which would increase the data quality. Some of the players did use half value units which help in the analysis; however of the useable RPE data only three values used a .5 grade.

The fact that the RPE is generated by the players, the data collection can be difficult. For example if players are in difficulty or have post-match obligations their accessibility may be limited. Any player that has suffered injury or needs immediate treatment in the aftermath of the match may also have limited accessibility. Any retrospective applying of an RPE grade brings a level of inconsistency to the RPE collection method and could potentially impact on the values provided.

The wearing of GPS units for this particular season was optional for the players to wear. Some of the players declined, which impacted any form of match-up analysis. There were 226 RPE grades that could not be tied back to any GPS readings. Of these 76 were from prop forwards who are exempt from wearing the GPS units. However, there is also the opposite issue where the GPS data is available however the RPE value has not been gathered. All of these one-sided match-ups leave a big gap in terms of completeness of information and impact on the overall analysis. It also shows the missed potential value of a more robust and consistent data collection policy.

For the Decision Tree analysis, the count of the data records in each of the models is low. This may have an impact to the robustness of the models and would ideally need a higher volume of records to be deemed as a stable model. If the RPE quality was to be improved as per the previous comments within this section, then the Decision Tree model should reflect higher levels of accuracy and percentage compliance.

Findings

RPE is a very useful data collection technique. [16] Given the robust relationship between the RPE and measures of exercise intensity, particularly if this is known for an

individual, the RPE is commonly used as a guide to the subjective assessment of exercise intensity. There is a reason why it is widely used, however its variance, unreliability and inconsistency makes it difficult to apply for objective data analysis. Nevertheless, it does provide valuable information for players who are competent at judging game difficulty and load. Consistency in both grading by the players and the collection by the strength and conditioning team would benefit this in terms of applicability. Also, greater scaling of the range within which the RPE values reside would achieve a higher level of accuracy. It is a good measurement for comparing like for like general positions in terms of players GPS outputs and their subsequent grades.

For areas that have strong correlations, it can be a focal point for the Strength and Conditioning Team to work on improving individual's capabilities in areas that they find most difficult in games. For example, there are correlations with the Hookers and the Out Halves for the difficulty grade in games that had high quantities of Acceleration \ Deceleration and Tackles respectively.

Re-applying similar research to the next season's activities with revised control over the data collection would prove beneficial to see if the results are consistent with the findings in this analysis. The improvement in the data quality would allow for cleaner consistency and may further help to highlight existing correlations.

The newer sets of data can be pushed through the Decision Tree model to assign game categories to players that have perceived low levels of accuracy of RPE grading. This will then allow for two things:

- to be able to align known problematic game graders so that a more scientific approach could be taken to measure the loads placed per activity.
- to actively apply player management strategies by reviewing the quantity of the easy or hard games they have played within any period of time or when recovering from injury etc. and initiate a course of action i.e. lightening of the training loads

3.4 High Intensity Minutes & Clusters by Overall Load by Player

Background

The High Intensity Minutes Analysis attempted to look at entire games in one minute segments to gauge the levels of intensity per minutes. These intensity periods were then analysed to see if there was any impact of that intensity on the Overall Load for all of those games. Some games would have a visibly higher level of intensity and there was a requirement to try to get further understanding of these. There is no intensity reading from the GPS units, so these had to be defined within the research.

There was also a requirement to seek out clustering of intensity. [18] Cluster Analysis groups data objects based only on information found in the data that describes the objects and their relationships. The goal is that the objects within a group be similar (or related) to one another and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group and the greater the difference between groups the better or more distinct the clustering. The Clustering Analysis attempts to look at extended periods within all of the games to gauge the levels of clustered intensity. Once these are calculated they are plotted against the individual's RPE and load readings to see if these had an impact to the perceived difficulty as reflected in the players' grades.

Methodology

The analysis required the intense minutes as well as the clustering to be defined and calculated. Because this was open to interpretation there were two different methodologies used but first the game profile per player needed to be used as the base data. This data was retrieved from the Match Minutes file. For each player, the Overall Load from the four variables was calculated on a minute by minute basis for all games played within the season. This only included information that met the inclusion logic for the Match Minutes file i.e. that there was some sort of movement within that particular

minute's data. This data set was then moved to the individual's tab for analysis. For the first method of Intense Minutes (IM) and Clustering calculation, the median value was calculated from the data set. A mirror data set was then used but the values were replaced with 1s or 0s depending on whether the value was greater or less than the median. See Fig 3.4.2.1 showing the original data set and the IM calculation. The count of the IMs per game was then averaged to give an overall score of IM for that game.

Doug Howlett		Median		0.0644											
								IM Score							
Game Minutes	ERC 1	ERC 2	ERC 3	ERC 4	ERC 6	RDP 1	ERC 1	ERC 2	ERC 3	ERC 4	ERC 6	RDP 1			
1	0.0248	0.0846	0.1867	0.0722	0.1657	0.0627	0	1	1	1	1	0			
2	0.0820	0.0448	0.0507	0.0189	0.1199	0.0242	1	0	0	0	1	0			
3	0.1534	0.0432	0.0644	0.1343	0.0332	0.0177	1	0	1	1	0	0			
4	0.0532	0.1859	0.0537	0.0876	0.0206	0.1238	0	1	0	1	0	1			
5	0.1785	0.1157	0.3535	0.0615	0.0953	0.1944	1	1	1	0	1	1			
6	0.0195	0.1930	0.1123	0.0759	0.0616	0.1635	0	1	1	1	0	1			
7	0.0397	0.0395	0.0563	0.0301	0.0313	0.0799	0	0	0	0	0	1			

Fig 3.4.2.1 – IM Calculation Example

The clustering was calculated in windows of 5 minutes. The current minute looked at the previous 4 minutes activity and summed the values that were in the IM values for those corresponding minutes as shown in Fig 3.4.2.2. This gave a maximum output count of 5, where there were 5 consecutive minutes where the Overall Load reading was above the median value for all Overall Load values for that player. Alternatively, the lowest reading of 0 was calculated if the player had a 5 minute consecutive period of play where the Overall Load was below this median value. All clustering values that had a minimum value of 4 were considered as a true high intensity cluster and these are summed for each game.

IM Score						Clusters					
ERC 1	ERC 2	ERC 3	ERC 4	ERC 6	RDP 1	ERC 1	ERC 2	ERC 3	ERC 4	ERC 6	RDP 1
0	1	1	1	1	0	0	1	1	1	1	0
1	0	0	0	1	0	1	1	1	1	2	0
1	0	1	1	0	0	2	1	2	2	2	0
0	1	0	1	0	1	2	2	2	3	2	1
1	1	1	0	1	1	3	3	3	3	3	2
0	1	1	1	0	1	3	3	3	3	2	3
0	0	0	0	0	1	2	3	3	3	1	4

Fig 3.4.2.2 – Intense Cluster Calculation Example

Both of these calculations that met the minimum requirements for inclusion were repeated for all games that the player had partaken in. These minimum requirements were needed for consistency in the output and included

- 55 minutes or more played in a game
- RPE recorded value for that game from that player

The second IM and Clustering calculation looked to extract the same information using a different technique. For each base data set, a window of 5 minutes Overall Load values was summed to give a rolled IM value for that particular minute as shown in fig 3.4.2.3. This represented the previous 5 minutes activity. These values were then summed and divided by the minutes played for that particular match to calculate the secondary IM value.

Game Minutes							IM Rolled II					
	ERC 1	ERC 2	ERC 3	ERC 4	ERC 6	RDP 1	ERC 1	ERC 2	ERC 3	ERC 4	ERC 6	RDP 1
1	0.0248	0.0846	0.1867	0.0722	0.1657	0.0627	0.0248	0.0846	0.1867	0.0722	0.1657	0.0627
2	0.0820	0.0448	0.0507	0.0189	0.1199	0.0242	0.1068	0.1294	0.2374	0.0911	0.2856	0.0868
3	0.1534	0.0432	0.0644	0.1343	0.0332	0.0177	0.2602	0.1726	0.3019	0.2254	0.3188	0.1045
4	0.0532	0.1859	0.0537	0.0876	0.0206	0.1238	0.3135	0.3586	0.3555	0.3129	0.3394	0.2283
5	0.1785	0.1157	0.3535	0.0615	0.0953	0.1944	0.4919	0.4743	0.7090	0.3744	0.4347	0.4227
6	0.0195	0.1930	0.1123	0.0759	0.0616	0.1635	0.4866	0.5827	0.6345	0.3781	0.3307	0.5236
7	0.0397	0.0395	0.0563	0.0301	0.0313	0.0799	0.4443	0.5774	0.6401	0.3893	0.2420	0.5793

Fig 3.4.2.3 – Alternative IM Calculation Example

Using the IM rolled data set, representing the sum of the previous 5 minutes, a percentile value of 60% was extracted from the entire set. Each rolled IM minute per match was checked against this percentile value and if it was greater, then it was counted as a high intensity cluster. These clusters were totalled for all games to give an overall secondary cluster value.

The two data sets that were calculated for the IM and Clustering were not mirror representations and when values are compared one to one they did not in any way look related as shown in fig 3.4.2.5 which shows the output for Doug Howlett. This table shows that there is no obvious comparison between the two sets of values. However, when there is a correlation analysis done on the models against both sets of values we can see that there is a strong correlation with 83.5% and 78.5% for IM and Clusters respectively.

Game	ERC 1	ERC 2	ERC 3	ERC 4	ERC 6	RDP 1	RDP 10	RDP 12	RDP 13	RDP 2	RDP 3	RDP 4	RDP 6	RDP 8	RDP 9
IM	2.77	1.96	2.30	2.96	2.01	2.66	2.08	2.47	3.13	2.58	2.79	2.61	1.91	2.64	2.66
Clusters	23	5	11	30	9	29	12	14	32	20	29	8	3	10	18
IM II	0.45	0.35	0.46	0.45	0.37	0.46	0.37	0.43	0.53	0.43	0.43	0.40	0.35	0.48	0.41
Clusters II	46	26	39	49	25	51	32	34	65	42	43	12	20	51	30
Minutes	92	84	89	98	97	101	102	75	91	93	95	46	97	91	92
Correlation of IM	83.5%														
Correlation of Clusters	78.5%														

Fig 3.4.2.5 – IM and Cluster Results Example for 2 Methods

Once the IM and Clusters were extracted from the player's data set they were then compared against the RPE and Load for that player for various different scenarios. The margin of victory was also added to the file. Any record that did not have an RPE \ Load grade was removed as were all records where game times are below 55 minutes played. The analysis was completed for two sets of figures; IM correlation and Cluster correlation.

Analysis

The analysis was conducted in two ways. The same analysis was done for both IM and for Clusters individually against the same measurements. The measurements included correlation of IM or Clusters against the following: Overall, ERC, RDP, Wins and Losses. The comparisons were done for the two different sets of IMs and Clusters Model's values as outlined in the Methodology. Any correlation result that was greater than +/- 0.6 was considered to have a relationship either directly or inversely. For any value that had a correlation of +/- .75, the relationship was considered to be a strong one. These are highlighted in the results set as yellow and green respectively. This analysis was carried out for all of the players for which this information was available which is 17 in total. Each player then had a correlative profile of RPE \ Load measurement for ERC \ RDP and Wins \ Losses against games that had high or low numbers of IMs or Clusters relative to Overall Load. The expectation from Munster was that games that had high values in IM and Clusters would be graded as a higher difficulty by the player.

Some players didn't have sufficient values within the particular categories of the correlations and these were therefore not considered and represented as N/A in the tables. Also some correlations had minimum comparative data sets of three values which

tended to over-state the results due to the low value count. In Fig 3.4.3.1 Peter O Mahony's results are shown with the correlation's strengths highlighted in green and yellow. Any values that had two or less in the group were disregarded. There is consistency in the analysis for both sets of data in terms of the primary and secondary IM and Clustering methods of calculation. The results with the higher count of variables can be considered as true non-skewed results. From this example we can see a correlation between the player's grades and load and the Intensity Minutes across all games that meet the inclusion criteria. It can also be seen that the player's grade would have direct correlation with the Clusters for the ERC games. This is the kind of typical result set that Munster would have been expecting. See all players' results in the Appendix C Ref 3.4.A.

Peter O Mahony	Value #s	Primary		Secondary	
		RPE	Ave Load	RPE	Ave Load
Correlation IM	8	0.8334	0.8064	0.7169	0.6719
Correlation IM ERC	6	0.7024	0.7770	0.5567	0.6486
Correlation IM RDP	2	N/A	N/A	N/A	N/A
Correlation IM Wins	7	0.8973	0.8961	0.7602	0.7336
Correlation IM Loses	1	N/A	N/A	N/A	N/A
Correlation Clusters	8	0.5902	0.5709	0.7174	0.7995
Correlation Clusters ERC	6	0.9993	0.8810	0.6585	0.6580
Correlation Clusters RDP	2	N/A	N/A	N/A	N/A
Correlation Clusters Wins	7	0.6678	0.6743	0.7479	0.8527
Correlation Clusters Losses	1	N/A	N/A	N/A	N/A

Fig 3.4.3.1 – IM & Cluster by Overall Load Correlation with RPE by Player

Within this analysis there was also a request to repeat the same analysis for the 4 individual loads that the Overall Load is comprised of. This was done for one player to review the results which then resulted in a request for a spin off requirement which is captured by the next set of analysis in section 3.5. The player selected was Felix Jones whose Overall Load analysis did not provide much in the form of correlation. When this was further explored it can be seen that some of the variables showed strong inverse relationships as well as strong direct relationships. These then would have conflicted when analysed under the Overall Load. As Fig 3.4.3.2 shows, there is a strong correlation

with the RPE and games where there where the Tackle Load had high levels of IMs and Clusters.

Tackles Felix Jones	Value #s	Primary		Secondary	
		RPE	Ave Load	RPE	Ave Load
Correlation IM	11	0.6331	0.5055	0.6730	0.5130
Correlation IM ERC	5	0.8272	0.7759	0.8238	0.7158
Correlation IM RDP	6	0.6120	0.3506	0.5704	0.3120
Correlation IM Wins	6	0.4290	0.4322	0.5200	0.4351
Correlation IM Loses	4	0.6972	0.6972	0.7163	0.7163
Correlation Clusters	11	0.4364	0.1832	0.5313	0.3302
Correlation Clusters ERC	5	0.8018	0.6617	0.7903	0.6580
Correlation Clusters RDP	6	0.5522	0.1225	0.3197	0.0110
Correlation Clusters Wins	6	0.4472	0.4101	0.3166	0.1775
Correlation Clusters Losses	4	0.7071	0.7071	0.5970	0.5970

Fig 3.4.3.2 – IM & Cluster by Tackle Load Correlation with RPE by Player

This lower level detail provided Munster Rugby further insight into how the players' perception of difficult and easy games was impacted on by the activity on the pitch. This instance was a pilot only and can be seen in the appendix Ref – 3.4.B. Interestingly there were more inverse relationships for the athleticism based variables which would suggest that the more the intensity of the running type of activity, the easier the player found the game. Given this player's profile and position this is something that can be easily rationalised.

Limitations

There are some limitations to the IM and Cluster analysis and some of these are similar to the issues outlined previously in this document. Taking correlation values for data sets that have limited variables is not robust in terms of how applicable the findings are. If another inconsistent value was presented into a limited data set, correlations could be skewed due to the low count of values. However, there is only a small quantity of games within a season and these values will always be relatively low when you are using the game unit as a baseline metric.

Again this analysis uses the RPE and Load measurement which has its limitations in terms of subjective validity and consistency. These, however, were the balancing data to the GPS data and were necessary for this analysis. Included also in the files were the record sets that were not considered for the analysis due to unavailability of the RPE and Load values. If these values could have been obtained at the time of data collection this would mitigate somewhat against the low count of comparable values.

By using the Overall Load, any conflicting correlations would neutralise and hide any sub category correlations that may exist. This was something that was highlighted when the analysis was done for Felix Jones for the lowest level of variable. Each of the variables showed both direct and inverse correlations. Completing the analysis at the player level by the five different variables – Overall, Odometer, Acceleration \ Deceleration, High Speed Running and Tackle loads – would have required $17 * 5$ sets of analysis and this was deemed as too much of an overhead for this particular research so a pilot was agreed upon. When the result sets were seen, it was agreed to do the Positional $* 5$ variable analysis which was 35 different sets of analysis. This is discussed in the following section 3.5.

For the most part the two different methods of the IM and Cluster calculation were consistent in the results they returned, however some of the of the results especially for the data sets that did have the lower quantities could return many contrasting values. This would suggest that the strength of the correlation result set was at least questionable.

Findings

The information that was provided by this analysis showed for all of the players, the impact of the Cluster and IM activity of the Overall Load had on their perception of each game played. This can provide valuable high-level information with regard to the impact that these activities can have on the players. The results show that of the seventeen players that were analysed, four showed no correlation across all the different measurements. These players were Billy Holland, Denis Hurley, Casey Laulala and Conor

Murray. For each of these players there are other individuals who play in a similar general position and who returned some level of correlation. There may be two factors that would help to explain this; the subjective nature of the RPE has high levels of inconsistency and suffers from this or the variables that were used to calculate the Overall Load and their corresponding intensity or clustering had no impact on the difficulty grade as marked by the player.

Of the 40 different measurement calculations as highlighted in Fig 3.4.3.2, certain players did not meet some of the criteria to calculate the correlation. The table in Fig 3.4.5.1 shows the summary of the results outlining the count of values, the Good Correlations counts and the Strong Correlation counts for all of the players that were included in this analysis. The count of the possible values depended on whether or not there was a sufficient data set size to complete the particular measurement.

	Possible Values	Total Correlation	Good Correlation	Strong Correlation
Denis Hurley	32	0	0	0
Doug Howlett	40	14	6	8
Felix Jones	40	7	1	6
Simon Zebo	40	14	0	14
James Coughlan	40	8	0	8
James Downey	40	7	2	5
Casey Laulala	40	0	0	0
Ronan O Gara	32	9	3	6
Conor Murray	24	0	0	0
Duncan Williams	32	8	1	7
Peter O Mahony	24	21	11	10
Tommy O Donnell	40	13	7	6
Billy Holland	24	0	0	0
Donncha O Callaghan	32	8	3	5
Paul O Connell	24	9	3	6
Damien Varley	24	16	4	12
Mike Sherry	40	9	6	3

Fig 3.4.5.1 – Summary Correlation Table

From the table we can see that Peter O Mahony and Damien Varley had a high number of total correlations of which a high percentage were in the strong correlation

category. When their results are further analysed, it can be seen that both players grade the game's level of difficulty with the amount of clustering and intensity. Conclusions could be drawn that both players are competent subjective benchmarks for their respective positions. The balance of the result sets had a lower level of correlation and some of these correlations were based on small quantities of values. There are examples where the smaller data sets would contradict the larger sets such as for Simon Zebo as shown in Fig 3.4.5.2.

Simon Zebo	Value #s	RPE	Ave Load	RPE	Ave Load
Correlation IM	10	0.1795	0.2357	0.0804	0.1504
Correlation IM ERC	6	0.0560	0.1059	- 0.0601	- 0.0191
Correlation IM RDP	4	0.1801	0.1826	0.0422	0.0432
Correlation IM Wins	7	0.9052	0.9142	0.7927	0.8343
Correlation IM Loses	3	- 0.9163	- 0.9163	- 0.8652	- 0.8652

Fig 3.4.5.2 – IM by Overall Load Correlation with RPE by Player Example

For the individual player's measurement against the four individual variables, the findings definitely presented a more complete picture for the variances within the individual loads. These variables had both comparable and contrasting levels of correlation. This level of detail allowed Munster insight into specific areas of potential focus based on the correlation results. It also identified the value of going to such levels for any future similar analysis with new data sets from the forthcoming seasons.

3.5 High Intensity Minutes & Clusters of Four Load Variables by General Position

Background

This requirement is a spin-off request from the previous 3.4 analysis. In the 3.4 analysis, a pilot case was researched for the four individual variables for Felix Jones to see if there was a better understanding of the relationships at a lower level. These results returned some very interesting sets of figures and based on these, this new requirement was derived.

Because of the workload associated with exploring the individual player level within the timelines associated with this project, it was agreed with Munster Rugby that this would be completed for the General Position level. There are seven general position categories. This would still provide enough information to Munster to be applicable for any proactive group focus areas.

Methodology

The methodology used here was very similar to the analysis carried out in 3.4 with some small differences. The base files were done for each of the four variables Odometer, Acceleration \ Deceleration, High Speed Running and Tackle Loads as well as the Overall Load for every general position – Back 3, Centres, Out Half, Scrum Half, Back Row, Locks and Hookers. This required a unique analysis for Back 3 for Odometer and another unique analysis for Back 3 for Acceleration \ Deceleration etc. This totalled 35 different sets of analysis each with 40 different measurements where the data sets were complete.

Each of these was processed in an identical manner as the player Overall Load analysis in terms of the Intense Minutes and Clusters. Again there were two different calculation methods used to maintain the credibility of the model and to ensure integrity. A summary file was generated with all of the result sets for ease of comparison. See Appendix D Ref 3.5.A. In this file, all of the 35 results sets are positioned side by side in the rows for the variables and ordered in the columns by position. Correlations of any significance are colour-coded for high-level visual interpretation.

Analysis

The analysis replicated that of the Overall Load per player analysis, but was completed for all of the variables and for the general positions. The same measurements were used against IM and Clusters as a whole, ERC Vs RDP and Wins Vs Losses. If the positional group data sets were complete, the maximum output results would be 40 per set of analysis. This gives a total of 1350 correlation measurements against the agreed data sets if all of the information was available. This is 35 individual sets of analysis * 40 results per analysis less 50 data sets that did not have the sufficient number of records.

Limitations

This set of analysis had identical limitations as outlined in the previous section 3.4. The data sets were drawn from the same pools of information, however were filtered on different criteria so the same limitations exist.

Findings

The results sets for this analysis provided a level of granularity beyond the scope of the Overall Load analysis. There are positions that returned no levels of correlations across some of the variables. In total six Positions \ Variables didn't show any correlations with the intensity of that particular variable. These were the results sets:

- Back 3 \ Acceleration - Deceleration
- Centres \ Acceleration - Deceleration
- Back Row \ Odometer
- Back Row \ HSR
- Hooker \ Tackle Load
- Lock \ Overall Load

It is also worth noting that Centres has only one Good Correlation value against the Tackle Load variable. Even though these results are the reverse of what the research is seeking to find, they can be somewhat justified from a rugby perspective. A lot of these variables would be the standard role for the general position and games that have high

levels of these variable loads in terms of clusters and intensity do not impinge on the RPE grades and loads from the players within those particular general positions.

Additionally, there are negative correlations for variables that again form part of the general positions core expertise. For example, a Centre in Rugby typically would spend a lot of time in the high speed running zones both with and without the ball for attacking and defence or leading the line for kick chases. There is also a lot of positional re-alignment as they are impacted in terms of proximity to both sidelines more so than any other position both with and without possession. From Fig 3.5.5.1 it can be seen that the more the intensity and clustering of the HSR, the easier these players grade the difficulty of the games.

HSR Centres	Value #s	Primary		Secondary	
		RPE	Ave Load	RPE	Ave Load
Correlation IM	18	-0.5067	-0.4652	-0.6134	-0.5044
Correlation IM ERC	6	-0.0295	0.0668	0.1356	0.2579
Correlation IM RDP	12	-0.6718	-0.6338	-0.8126	-0.7351
Correlation IM Wins	9	-0.0309	-0.1797	-0.2264	0.0101
Correlation IM Loses	8	-0.6564	-0.5775	-0.6865	-0.6115
Correlation Clusters	18	-0.5681	-0.5366	-0.5178	-0.4348
Correlation Clusters ERC	6	-0.5397	-0.4464	0.0935	0.2015
Correlation Clusters RDP	12	-0.6723	-0.6261	-0.7757	-0.7295
Correlation Clusters Wins	9	-0.1013	-0.4021	-0.1585	-0.0276
Correlation Clusters Losses	8	-0.7394	-0.6188	-0.5911	-0.5247

Fig 3.5.5.1 – IM & Cluster by HSR Correlation with RPE by General Position

There are positions that show low levels of correlation that have less than 10 overall values that are above the 0.6 threshold. Some of these have low levels of value sets within the analysis sample and therefore may not be taken as very emphatic. There are also some positions that have high levels of correlation with the intensity and the clustering of the variable as can be seen in table Fig 3.5.5.2. Of the 35 data sets, 6 did not show any correlation for the set of 40 measurements, 21 had less than 10 sets of correlation and the balance, 8 had relatively high numbers of correlation. There was only 1 data set that had 10 or more strong correlations. This was the Out Half position and the

Tackle Load variable. This result set was also limited to only 30 measurements which in turn enhanced this relationship.

The analysis with the highest set of correlations is the Scrum Half and the HSR variable with 20. Interestingly, the majority of these are inverse relationships meaning the more the games contain high levels of intensity or clusters of HSR, the easier the scrum halves grade the games. This could be relayed to a more broken up game with more focus on positional movements at pace rather than games that are 'dogfights' played around the fringes which typically the scrum halves dislike.

	Possible Values	Total Correlation	Good Correlation	Strong Correlation
Scrum Half HSR	40	20	12	8
Back Row TACK	40	18	17	1
Out Half TACK	30	16	4	12
Centre HSR	40	14	12	2
Hooker AD	40	13	6	7
Out Half OL	32	11	2	9
Back Row AD	40	11	5	6
Scrum Half OL	40	10	3	7
Lock TACK	40	9	3	6
Out Half OD	32	9	2	7
Hooker OD	40	9	5	4

Fig 3.5.5.2 – Correlation Summary Table

The Out Half tackle relationship is very interesting in that the number 10 shirt is a big target for most opposition when building attacking phases especially from first phase possession. Ronan O’Gara has often been highlighted as a player in terms of his tackling capability. The results here show that for the out half position, the games with high levels and concentration on Tackle Load were graded as more difficult by both the Out Halves that were measured within this category. This is not surprising and in a way reinforces the accuracy and applicability of the result sets.

Other interesting results include the Back Row and the Tackle Load impacting on the game grading difficulty. Tackling is a core part of the back row’s responsibility and the Tackle Load would be expected to be above most of the other players. The results tell us

that when this load increases in intensity and clustering it takes its toll on these players in the form of difficulty grading.

This analysis allows Munster to delve deeper into the positional loads and their impacts on the players where these loads had intense activity. This allows them to better prepare the players for such activities. Getting the break-down for the ERC \ RDP allows them to apply short term preparation in the weeks prior to the games to mitigate against fatigue and overload of the players.

4 Results

This section outlines the various limitations, results and future proofing activity that should be considered for this research application.

4.1 Results Limitations

Data integrity – The importance of the data integrity, specifically from an objective data source needs to be highlighted as paramount to the analytical activities. Some of the issues with the raw data files could be avoided by liaising with the solution provider and stressing any data anomalies.

Analysis Limitations – The results as outlined within the research topics all have a limitations section. These should be taken into consideration prior to any application of the research. The lack of high volumes of data at a game level, due to the limited number of games a season, does impact some of the research. Any future analysis should build on these existing data sets to give a broader mass of data. Most of the analysis can be applied across multiple seasons.

Subjective Data Gathering – there is opportunity to improve the quality and consistency of the subjective RPE data that was used in this research. Ensuring completeness of the information through better standardised means would be of great benefit. Also encouragement of the players to use the half grades would also help to get deeper insight. Any players that are showing consistent grades should be spoken to about the importance of the activity so that a more balanced view is taken to the games.

4.2 Result Sets

This research project provided Munster Rugby with a wide variety of analysis results sets for various activities for the season 2012/13. It provided them with both the result sets and the internal base files so that any alternative filtration of the data elements could be reset for any further specific requirements. The results were, where possible, outputted into an end user friendly format for ease of understanding to those for whom it needed to be discussed with or explained to. All findings were shared with Munster

Rugby and the refined master data files and analysis applied were returned to allow Munster to maintain autonomy for any future similar requirements.

The results sets within the different areas in the research allow Munster to:

- segment games and expectations into blocks
- understand what loads are the most consistent across opposition teams
- plan for the changes in the games over the halves
- prepare for the loads that weigh heaviest in terms of player game grade difficulty
- understand the impacts on the players of games with periods of high intensities

The research allowed the Strength and Conditioning Team to gain further insight into the elements that make up their role within the Munster organisation and allows them to plan for future similar types of activities by utilising the modelling that was used here. The game, opposition, positional and individual analysis allows them to plot and overlay new preparation behaviours to best suit circumstances.

4.3 Future Proofing

As this was a one off primary research project, there was no software development for re-processing of new data sets. The research looked to understand the data elements and extract logical reasoning rather than building a processing system. All of the methodologies were documented, shared and discussed with Munster so that a clear understanding of how the data was manipulated is understood. These discussions were not limited to the data sets but also included considerations specific to rugby scenarios, players, positions and games.

A high level work instruction was created to re-trace the calculation steps for Munster so that if there was a requirement to amend any of the existing analysis or spin or limit it in a different way, then they would have the autonomy to do so. This work instruction can be seen in the Appendix E Ref 4.2.A

If this was to be a longer term solution for repetitive activity, I would at a minimum, liaise with the vendor for a cleaner data set and build out templates and macros to ease the processing burden. The overhead of this would be minimal and if Munster sees the benefit of the analysis they should seek to put some form of processing infrastructure in place. The requirements of the analysis were bespoke to Munster and I would imagine something that a vendor would not seek to solution because of this uniqueness. Also, for some of the analysis, the information was outside of the GPS supplied data so there would be an element of collaboration with Munster if it was to be vendor packaged.

5. Conclusion

The critical part of this research was the understanding of the data and what limitations and restrictions it was to be manipulated under. All of the areas within the research have been documented in terms of Background, Methodology, Analysis, Limitations and Results and have been discussed in depth with Munster Rugby.

The deliverable was a deeper understanding for the Munster Strength and Conditioning team as was outlined in their project brief: “This amount of data is analysed to the best of our ability but we recognise our limitations in this area and wish to work with those who have more expertise in data Analytics. We hope to explore this data deeper and confirm thoughts, dispel myths and develop new cutting edge philosophy in sport science that will give Munster Rugby and advantage few hold.” The original project brief, for which this research looked to solution some of these requirements, is available in Appendix F Ref 5.0.A

There have been so many interesting findings in this research that the scope of the original requirements needed to be amended to allow for spin off analysis based on the results of initial research. Having the vision and creativity of seeking out relationships within past activity data sets is something that can unlock huge understanding and potential to a better prepared approach for Munster Rugby. The results of this research have already provided new insights as well as categorisation and structure to existing beliefs. These results have provided a greater understanding of the GPS output and its relationship with player performance and have assisted in the preparation process for this season’s games. The Strength and Conditioning Team of Munster Rugby has a keen interest in the scientific and analytical side of the game and one would suspect that this will not be the last such endeavour to unlock more hidden insights.

References

- [1] A Publication of the National Wildfire Coordinating Group. *“National Interagency Incident Management System Basic Land Navigation”*. PMS 475, NFES 2865, June 2007 Page 5.1
- [2] Rugby Union Wikipedia Website – http://en.wikipedia.org/wiki/Rugby_union – Accessed Dec 2013
- [3] Matthew C. Varley, Ian H. Fairweather & Robert J. Aughey. *“Validity and reliability of GPS for measuring instantaneous velocity during acceleration, deceleration, and constant motion”* Journal of Sports Sciences June 2011.
DOI:10.1080/02640414.2011.627941
- [4] Pyne DB, Petersen C. Higham DG, Cramer M; *“Comparison of 5- and 10 Hz GPS technology for team sport analysis”*. Med Sci Sports Exerc 2010: 42(5):78
- [5] Rachel E. Venter, Eben Opperman, Simon Opperman. *“The use of Global Positioning System (GPS) tracking devices to assess movement demands and impacts in Under-19 Rugby Union match play”*. African Journal for Physical, Health Education, Recreation and Dance (AJPHERD) Vol. 17, No. 1 (March) 2011, pp. 1-8.
- [6] Catapult Product Technical Specifications. Catapult S4 10Hz Athlete Monitoring System. www.catapult.com. Accessed November 2013
- [7] Catapult Innovations. *“Minimax V4 GPS performance - Sprints.”* White Paper. October 2009 Page 1
- [8] Catapult Sports. *“Sprint Help - For Sprint 5.0 and subsequent releases”*. January 2013 www.catapult.com. Accessed November 2013
- [9] Richard Taylor, EDD, RDCS. *“Interpretation of the Correlation Coefficient: A Basic Review”* – 1990. JDMS 1:P35
- [10] Rapidminer – Help function notes
- [11] Tableau website - <http://www.tableausoftware.com/products/desktop> - Accessed Dec 2013
- [12] Alan C. Utter, Ph.D., M.P.H., FACSM, Jie Kang, Ph.D., FACSM, Robert J. Robertson, Ph.D., FACSM. *“Perceived Exertion”*. American College of Sports Medicine
- [13] Dr. Matthew North. *“Data Mining for the Masses”*. P 9. 2012 ISBN-13: 978-0615684376

- [14] George A. Vouros, Themistoklis Panayiotopoulos. *"Methods and Applications of Artificial Intelligence"* Third Hellenic Conference on AI, SETN 2004. P 175
- [15] Saher Esmeir, Shaul Markovitch. *"Anytime Learning of Decision Trees"*. Journal of Machine Learning Research 8 (2007) 891-933. Submitted 1/06; Revised 12/06; Published 5/07
- [16] Roger Eston. *"Use of Ratings of Perceived Exertion in Sports"*. International Journal of Sports Physiology and Performance, 2012, 7, 175-182
- [17] Kevin L Lamb, Roger G Eston, David Corns. *"Reliability of ratings of perceived exertion during progressive treadmill exercise"*. Br J Sports Med 1999;33:336–339
- [18] Pang-Ning Tan, Michigan State University, Michael Steinbach, University of Minnesota, Vipin Kumar, University of Minnesota. *"Introduction to Data Mining"*. March 25th 2006 CH.8 P490

Appendix A

3.1.A

Racing Metro 10 Minutes Analysis



App - RAC Load
Analysis.pptx

3.1.B

General Analysis of the Load Variables



App - General
Analysis.pptx

Appendix B

3.2.A

Loads & Changes to Loads



App - Load & Change
Analysis.pptx

Appendix C

3.3.A

All Players – RPE Vs GPS

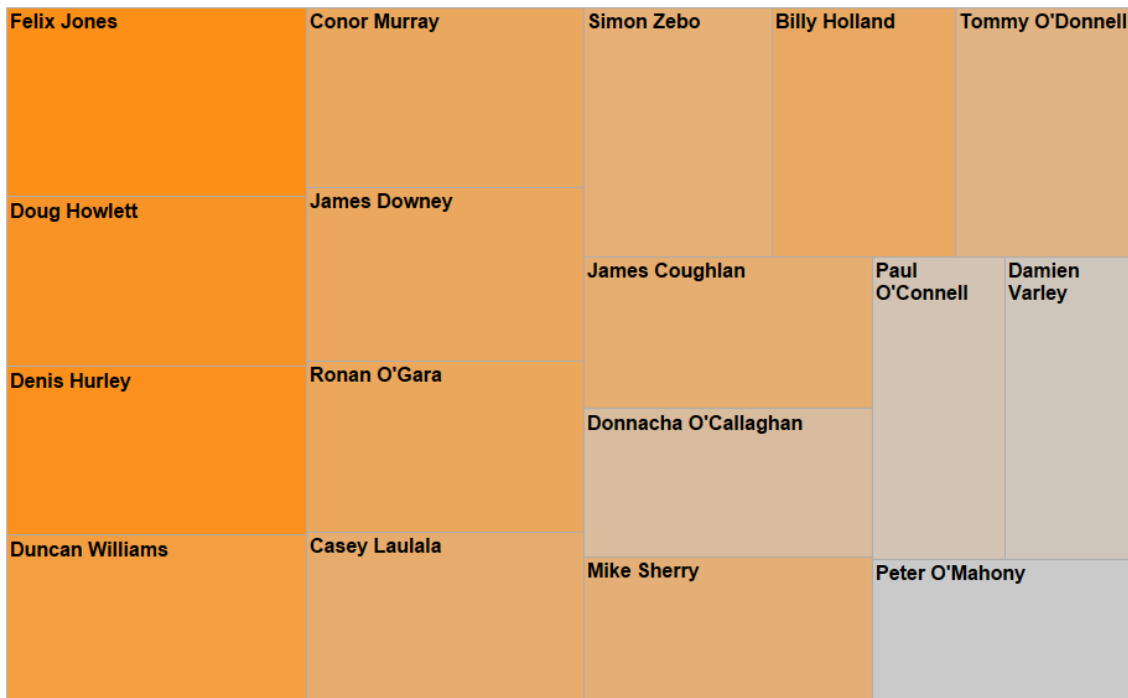


App - RPE Per
Player.xlsx

3.3.B

Treemaps for each of the variables showing highest averages (volume) and maximum readings (shading)

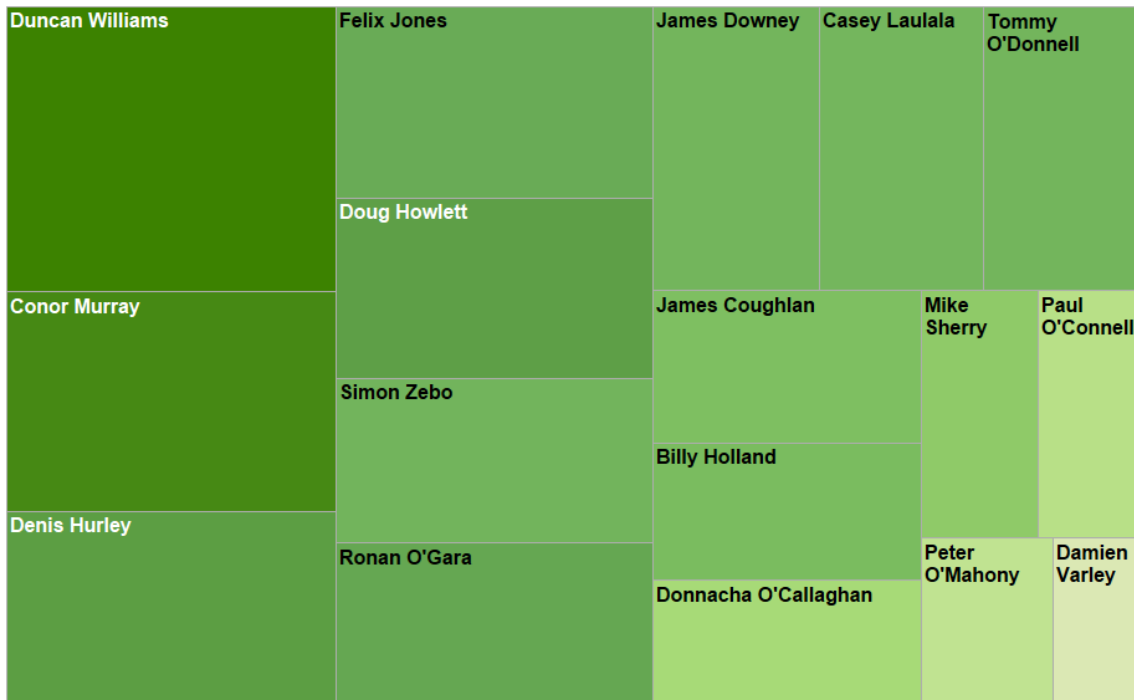
Odometer



Acceleration \ Deceleration



HSR



3.3.C

Decision Tree Incompatible Data Sets

3 Grade Variables

accuracy: 54.78% +/- 2.97% (mikro: 54.81%)				
	true Easy	true Hard	true Very Hard	class precision
pred. Easy	1	3	0	25.00%
pred. Hard	78	130	27	55.32%
pred. Very Hard	0	0	0	0.00%
class recall	1.27%	97.74%	0.00%	

All Values + 2 players weighted

accuracy: 68.21% +/- 7.47% (mikro: 68.20%)			
	true Easy	true Hard	class precision
pred. Easy	139	71	66.19%
pred. Hard	5	24	82.76%
class recall	96.53%	25.26%	

All values – equally weighted

accuracy: 70.31% +/- 4.64% (mikro: 70.29%)			
	true Easy	true Hard	class precision
pred. Easy	148	62	70.48%
pred. Hard	9	20	68.97%
class recall	94.27%	24.39%	

Back 3 Only

accuracy: 70.33% +/- 18.76% (mikro: 69.81%)			
	true Hard	true Easy	class precision
pred. Hard	6	1	85.71%
pred. Easy	15	31	67.39%
class recall	28.57%	96.88%	

Backs Only

accuracy: 65.90% +/- 3.64% (mikro: 65.89%)			
	true Easy	true Hard	class precision
pred. Easy	85	44	65.89%
pred. Hard	0	0	0.00%
class recall	100.00%	0.00%	

Centres

accuracy: 44.50% +/- 15.72% (mikro: 43.90%)			
	true Hard	true Easy	class precision
pred. Hard	6	10	37.50%
pred. Easy	13	12	48.00%
class recall	31.58%	54.55%	

Forwards Only

accuracy: 67.27% +/- 7.27% (mikro: 67.27%)			
	true Hard	true Easy	class precision
pred. Hard	9	7	56.25%
pred. Easy	29	65	69.15%
class recall	23.68%	90.28%	

Change of Grade Ranks

accuracy: 68.19% +/- 3.91% (mikro: 68.20%)			
	true Easy	true Hard	class precision
pred. Easy	4	1	80.00%
pred. Hard	75	159	67.95%
class recall	5.06%	99.38%	

Halves

accuracy: 63.33% +/- 14.53% (mikro: 62.86%)			
	true Hard	true Easy	class precision
pred. Hard	0	4	0.00%
pred. Easy	9	22	70.97%
class recall	0.00%	84.62%	

2 Players Removed – Denis Hurley & Mike Sherry

accuracy: 68.67% +/- 4.53% (mikro: 68.60%)			
	true Easy	true Hard	class precision
pred. Easy	121	58	67.60%
pred. Hard	7	21	75.00%
class recall	94.53%	26.58%	

Front 5

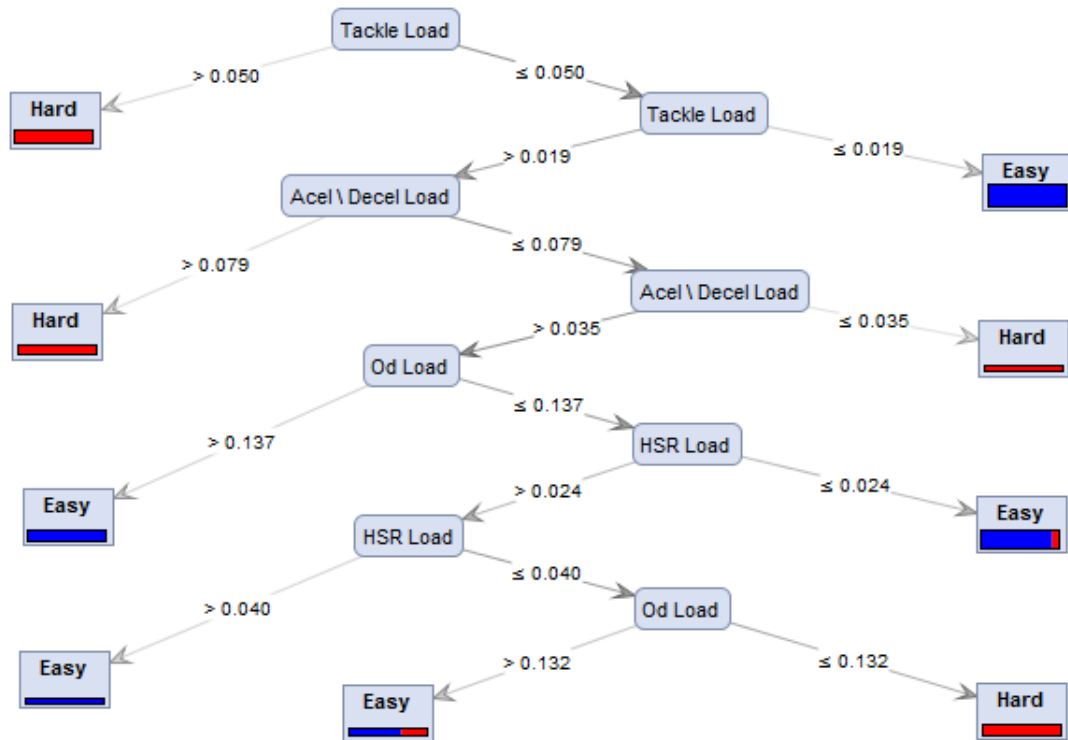
accuracy: 77.86% +/- 13.39% (mikro: 77.61%)			
	true Hard	true Easy	class precision
pred. Hard	11	3	78.57%
pred. Easy	12	41	77.36%
class recall	47.83%	93.18%	

Forwards less DOC

accuracy: 56.00% +/- 12.25% (mikro: 56.12%)			
	true Easy	true Hard	class precision
pred. Easy	24	13	64.86%
pred. Hard	30	31	50.82%
class recall	44.44%	70.45%	

3.3.D

Front 5 less DOC – Decision Tree



Front 5 less DOC – Multiple Levels

accuracy: 76.00% +/- 18.25% (mikro: 76.36%)			
	true Easy	true Hard	class precision
pred. Easy	29	8	78.38%
pred. Hard	5	13	72.22%
class recall	85.29%	61.90%	

Appendix C

3.4.A

All Players - Intensity Minutes and Clustering Analysis by Overall Load



App - IM OL by
Player.xlsx

3.4.B

Felix Jones - Intensity Minutes and Clustering Analysis by 5 Variables



App - FJ x 5
Variables.xlsx

Appendix D

3.5.A

General Positions - Intensity Minutes and Clustering Analysis by 5 Variables



App - IM 4 Vars + OL
by Positions.xlsx

Appendix E

4.2.A

Work Instruction



Calculations
Outlines.docx

Appendix F

5.0.A

Requirements Document



Sports Data
Analytics Research.pd