Cluster analysis of players from rugby world cup 2015

# Introduction

The Rugby World Cup was one of the biggest sporting events of the 2015 calendar year. The tournament saw records broken on and off the pitch. Bryan Habana matched Jonah Lomu’s Rugby World Cup try scoring record with 15 tries while on the stands over 2.47 million tickets were sold. Television numbers were also outstanding; the final was estimated to have been enjoyed by audiences of 120 million worldwide.

This study used player statistics from the 2015 Rugby World Cup to try to group (or cluster) player into meaningful groups. Cluster Analysis is used in many other areas including social network analysis to recognize communities within large groups of people. I first chose cluster analysis was to investigate if the groups created by the cluster analysis would place players into groups with other players that play in the same positions, this will be explained further in the next sections.

# methods used

Data collection

For this analysis there were two data sets used:

1. RWC 2015 player statistics dataset provided by Kitman Labs (<https://gist.github.com/itsakettle/7f62833a149a0b09424d>)
2. RWC 2015 player sizes dataset (<http://www.rugbyhow.com/2015RWC-all-player-sizes.html>)

The player sizes dataset is only used in this analysis to provide the position of each of the players that took part in the competition.

data processing

After some exploratory analysis of the data I decided that the number of matches played by each player introduced an unwanted bias to the data e.g. A player that has played in only one game in the competition is likely to have much lower statistics than one that has played in all of his team’s games. I therefore decided divide all players’ stats by the number of games they played to create a more level playing field.

Clustering players

“ I’m going to try to use K-means clustering to see if we can group players by positions by using just the 10 fields provided in the data. With K-means you have to specify the number of groups to use for clustering I’m going to start with 9, one for each position: prop, hooker, lock, back row, scrum half, fly half, center, wing and full back. “

The above quote shows my initial intention of using clustering algorithm, having some knowledge of Rugby I knew that k-means was unlikely to be able to create 9 separate clusters not due to a fault in the algorithm but because of the nature of the sport. Many players can interchange between multiple positions for example a player like Australia’s Kurtley Beale can basically play any position in the back line. As well as players being interchangeable in modern day rugby the role of players in different positions can be very similar for example back-row forwards and centers can end up doing a lot of the same jobs and when looking at their stats it can be difficult to differentiate between position.

After a lot of experimentation with the number of clusters I decided that 3 clusters made the most sense, while this didn’t split the players into each individual position it did create a pretty interesting division of the player.

# exploring results

Exploring clusters

This section of the analysis was split into two parts:

1. I first looked at the players in each cluster to develop some intuition about the clusters.
2. I created visualizations to compare the three clusters of players using some of the statistics provided in the data. I will explain further some of the most interesting ones.

###### developing an intuition about the clusters

In this part of the analysis I looked at the players that were in each cluster and a table of cluster vs. position to develop some intuition on each cluster.

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | **1** | **2** | **3** |
| **back-row** | 27 | 2 | 73 |
| **centre** | 32 | 6 | 28 |
| **fly-half** | 19 | 1 | 18 |
| **fullback** | 11 | 13 | 2 |
| **hooker** | 2 | 0 | 43 |
| **Lock** | 2 | 0 | 59 |
| **prop** | 0 | 0 | 85 |
| **scrum-half** | 6 | 1 | 40 |
| **wing** | 30 | 15 | 20 |

Table1 – Cluster vs. Position table

cluster 1

* Mostly backs with some back-rows
* Interestingly 2 locks were placed in this cluster they are Ian Henderson and Leone Nakarawa, probably the two most skillful ball carriers in this position group

Cluster 2

* All backs with just two back-rows
* Mostly wingers and fullbacks
* Contains some of the most electrifying players in the competition including Israel Folau, Mike Brown, and Julian Savea amongst others.

cluster 3

* All tight 5 players except four in cluster 1 are in this group
* Interestingly all but 7 scrum-halves are in this group
* 20 wingers were assigned to this group, they have just one try between them.

###### analysing statistics per cluster

meters made per game (mmpg)

Out of all the statistics I used in this analysis the one that produced the most interesting divide between the groups was meters made per game (*MMpG*). The below density plot (fig1) shows the distributions of *MMpG* for each cluster. We can see that there is an obvious divide between the 3 clusters with some small overlap. Cluster 3(blue) has the lowest values followed by Cluster 1(Red) with Cluster 2(Green) having the highest number of meters made per game. This follows the intuition I gained by looking at the players in each cluster.



fig1 – *MMpG* density plot

carries over gain line per game (coglpg)

Similar to *MMpG* carries over gain line per game also shows a good divide between the three clusters. The density plot (fig2) shows the distribution of COGLpG for each cluster. The divide in this case is nowhere near as clear-cut as the one provided by *MMpG.* The order of the three clusters remains the same with Cluster 2(Green) leading the way again followed by Cluster 1(Red) and then Cluster 3(Blue). One of the reason COGLpG shows a greater overlap between the clusters is the scale of the variable compared to MMpG. For the entire population of players COGLpG only varies from 0 to 8 while MMpG goes from 0 to 129.

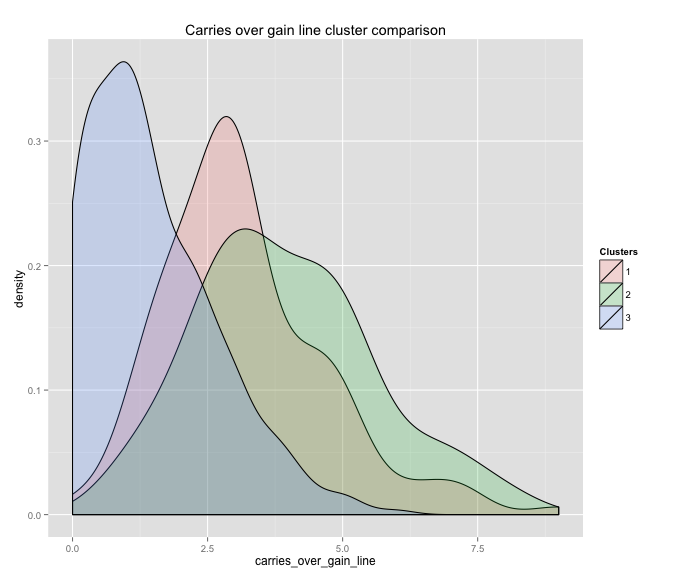


fig2 – COGLpG density plot