European Journal of Operational Research 268 (2018) 339–347

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

Identifying key players in soccer teams using network analysis and pass difficulty

Ian G. McHale[a,*], Samuel D. Relton[b]

[a] Centre for Sports Business, University of Liverpool Management School, University of Liverpool, L69 7ZH, UK
[b] Leeds Institute for Health Sciences, The University of Leeds, Leeds LS2 9LU, UK

ARTICLE INFO

Article history:
Received 3 February 2017
Accepted 8 January 2018
Available online 31 January 2018

Keywords:
Sport
Big data
Football
Moneyball
Random effects

ABSTRACT

We use a unique dataset to identify the key members of a football team. The methodology uses a statistical model to determine the difficulty of a pass from one player to another, and combines this information with results from network analysis, to identify which players are pivotal to each team in the English Premier League during the 2012–13 season. We demonstrate the methodology by looking closely at one game, whilst also summarising player performance for each team over the entire season. The analysis is hoped to be of use to managers and coaches in identifying the best team lineup, and in the analysis of opposition teams to identify their key players.

1. Introduction

The use of quantitative analysis in sports is, like in many industries, on the rise. A combination of increased computing power, and better recording and availability of data, has led to an increase in the awareness of the contribution that analytics can make to success in the sporting arena. Across sports globally, there are success stories to be cited. Perhaps the first, and certainly the most well-known case of analytics being used successfully in sports, is the story of how the Oakland Athletics were able to compete at the very highest echelons of Major League Baseball on a budget of around a tenth that of the bigger teams. Elsewhere, in cycling for example, much of the success of British riders in the Tour de France and at the Olympics Games in recent years has been attributed to Sir Dave Brailsford's adoption of analytical methods.

In soccer however, there is as at the time of writing, no such "success story". Further, there is very little written evidence documenting the adoption of advanced quantitative methods in the pursuit of gaining an advantage over the opposition by any professional team. The seemingly slow take up of analytics in soccer is very likely a consequence, at least to some extent, of the complexity of the game: 22 players moving and interacting continuously for over 90 minutes is certainly not a simple setting for an analyst, and makes it particularly difficult to gain insight above what an expert eye can achieve.

But recent advances in data collection has meant that rich, detailed data on the locations and timings of all actions on the pitch are now available. Such attractive data sets have caught the eye of academics, and in the academic literature there are now some examples of utilising such data. McHale and Szczepanski (2013) and Szczepanski and McHale (2015) present models for identifying goal scoring ability and pass making ability, respectively. Meanwhile, Peña and Touchette (2013) take an entirely novel approach to the analysis of football strategy and make use of network analysis to identify the important players on each team.

In addition to data detailing the location of events, the richest datasets available in soccer also give the locations of the 22 players themselves. Recorded at a frequency of up to ten times per second, the 'player tracking data' can be used to measure distance covered, top speed, and the acceleration of players. Indeed, to date, the majority of academic work using such data has been descriptive in nature. For example, Castellano, Alvarez-Pastor, and Bradley (2014) assess the accuracy of two systems for tracking players on the pitch, whilst Rampinini, Coutts, Castagna, Sassi, and Impellizzeri (2007) tabulate speeds and distances run by players and compare these between the first and second halves of matches.

Outside of soccer, player tracking data has been used by Dan Cervone et al. (2014) to calculate the expected possession value in basketball. This methodology estimates the impact each player has on the probability that a series of passes (the possession) results in points being scored. The probability is updated continually as players move around the court and the players are 're-

* Corresponding author.
E-mail addresses: ian.mchale@liverpool.ac.uk (I.G. McHale), s.d.relton@leeds.ac.uk (S.D. Relton).

https://doi.org/10.1016/j.ejor.2018.01.018
0377-2217/

warded' for their actions which contribute to increases in the expected value of the possession.

In this paper, we also make use of player tracking data. Collected and made available to us by Prozone, we have information on the location ($x-y$ coordinates) of each player at a frequency of ten times per second. The data also include the events occurring in the match (such as passes, tackles, dribbles and shots etc.). We have these data for all 380 matches in the 2012–13 season of the English Premier League season.

Our objective is to use this unique dataset to learn about which players are key to each team. Such analysis and information could be used by team managers and coaches to aid decision making in team selection, and where to concentrate effort on the pitch in order to thwart the opposition's strengths. Our model fills a gap in the literature since player tracking data have, until now, not been used in soccer in any meaningful way to inform team strategy or recruitment. Further, by utilising such rich data, the resulting tools we develop should be able to identify key players more accurately than previously available models.

To achieve our objective, we combine two tools: network analysis and statistical modelling. The use of network analysis is intended to identify the key passers in the team – those players which are heavily involved in passing moves, and who are central to how the team plays. However, to take account of the impact a player's passes have on the team, we weight the passes in terms of importance. We do not know the importance of the pass, but we proxy it using a measure of pass difficulty, which we take as the probability of the intended pass being successful. And this is our second tool, a statistical model to estimate the probability of a pass being successful.

We use this weighting scheme because it should, in principle, reflect players frequently involved in passing moves from which the ball enters key areas that are heavily defended by the opposition team. Thus the measure of pass difficulty, should in theory be related to pass importance. A player making 5 yard passes to the side, or even backwards, in his own half is likely to be much less effective in generating goal scoring opportunities than a player passing into the opposition penalty area. Such a weighting scheme should help identify players at the heart of a team's "shot generation engine", and knowing which players these are has clear advantages when selecting which players to field, and how to nullify the attacking threat posed by opposition teams.

The paper is structured as follows. First we present the data and give some descriptive statistics in Section 2, before discussing our model for estimating the probability of a pass being successful in Section 3. Section 4 presents the network analysis tools we employed to generate our results in Section 5. We conclude with some closing remarks in Section 6.

2. Player tracking data

The depth of the analysis that can be performed to analyse player and team performance is enormously dependent upon the data that one has available. In soccer, the most widely available data are simple summary statistics for each game. For example, we might know that a player had a 95% pass completion rate in one game. However, without knowing the context of each pass, it is hard to judge whether or not this is an impressive feat, and as such, the insight that can be gleaned is massively diminished. For example, a midfielder making lots of passes back towards the defence will have a high completion rate but few of these passes would contribute towards winning the game. The rich nature of player tracking data makes much deeper analysis possible.

The data used in this research, provided by Prozone, gives the $x-y$ coordinates of each player ten times per second to 10 centimetres accuracy, and was made available to us for all 380 games in the English Premier League during the 2012–13 season, leading to a dataset containing over 451 million player positions and over 960,000 events. In the era of 'big data', this data must qualify. In the remainder of this section, we present some descriptive statistics on this unique and rich dataset.

Table 1
Average distance run by each playing position during the English Premier League 2012–13. Running is defined as moving at more than 3 metres per second, whilst a sprint is more than 6 metres per second.

Position	Total distance (kilometres)	Running distance (kilometres)	Sprints per 90 minutes
Goalkeeper	5.4	0.5	0
Centreback	9.6	3.6	62
Fullback	9.9	4.1	84
Wide midfield	8.5	3.8	72
Centre midfield	9.2	4.2	80
Attacker	7.6	3.1	64

Table 1 shows the average distance covered by players, for each playing position. As one would expect, goalkeepers cover the least distance, though it is perhaps surprising to see that even they cover over 5 kilometres per game as they protect their 6 yard wide goal. The most ground covered is by fullbacks (left backs and right backs). In the modern game, this is again unsurprising as fullbacks are charged with both attacking and defending duties. Of the outfield positions, attackers cover the least distance.

Also shown in Table 1 are the running distances and number of sprints per 90 minutes. The story is similar to the total distance covered – goalkeepers run much less and do no sprints per 90 minutes, whilst fullbacks and centre midfielders do the most.

Table 2 shows the distances covered per 90 minutes by each team (per player) over the season, ranked by final league position. Also shown are the number of passes into the final third of the opposition pitch, the number of shots, and the number of goals. It is interesting to see how well the descriptive statistics 'explain' league position. A simple way to do this is to calculate Spearman's rank correlation, ρ, between each of the descriptive statistics and the league position. Unsurprisingly, goals have the strongest relationship with league position ($\rho = 0.85$), followed by shots ($\rho = 0.69$), then passes ($\rho = 0.43$), all of which are statistically significant. The intriguing result here is that distance covered and number of sprints have negative rank correlations with league position of -0.08 and -0.11, respectively, though not statistically significantly different from 0. It appears then, that successful football teams are doing something other than simply running more, or performing a higher frequency of sprints.

These results are somewhat contradictory to the popularly held belief that running more than the opposition improves the team's chances of success, and to examine this relationship further, we performed the following experiment. For the moments in each match when the scores were level (e.g. $0-0$, $1-1$, $2-2$ and so on), we calculated the total distance covered and the number of sprints by each team. We then used the difference between the home and away teams' distances covered per minute and numbers of sprints per minute as two covariates in a logistic regression model to predict whether the home team scored the next goal. We chose to use only moments in the match when the scores were level to guard against any bias that might be introduced since once a team goes ahead (behind) in a match, it is likely to change its behaviour. The reason for the bias is because it is not random which team takes the lead – it is more likely that the better team scores first.

The results of the model concur with the above result – the coefficients on *difference in distance covered per minute* and *difference in sprints per minute* are both statistically significant and negative. In other words, if the home team runs more than the away team,