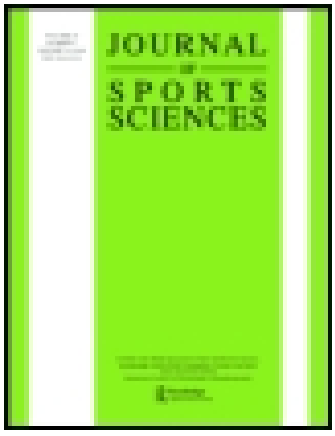
This article was downloaded by: [Memorial University of Newfoundland]
On: 17 July 2014, At: 18:41
Publisher: Routledge
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Journal of Sports Sciences
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rjsp20

Analysis of football game-related statistics using multivariate techniques

Felipe Arruda Moura[a], Luiz Eduardo Barreto Martins[b] & Sergio Augusto Cunha[b]
[a] Sport Sciences Department, Laboratory of Applied Biomechanics, State University of Londrina, Londrina, Brazil
[b] University of Campinas, College of Physical Education, Campinas, Brazil
Published online: 17 Apr 2014.

To cite this article: Felipe Arruda Moura, Luiz Eduardo Barreto Martins & Sergio Augusto Cunha (2014): Analysis of football game-related statistics using multivariate techniques, Journal of Sports Sciences, DOI: 10.1080/02640414.2013.853130

To link to this article: http://dx.doi.org/10.1080/02640414.2013.853130

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

Journal of Sports Sciences, 2013
http://dx.doi.org/10.1080/02640414.2013.853130

Routledge
Taylor & Francis Group

Analysis of football game-related statistics using multivariate techniques

FELIPE ARRUDA MOURA[1], LUIZ EDUARDO BARRETO MARTINS[2], & SERGIO AUGUSTO CUNHA[2]

[1]*Sport Sciences Department, Laboratory of Applied Biomechanics, State University of Londrina, Londrina, Brazil, and* [2]*University of Campinas, College of Physical Education, Campinas, Brazil*

(*Accepted 1 October 2013*)

Downloaded by [Memorial University of Newfoundland] at 18:41 17 July 2014

Abstract
The purpose of this study was to explore football game-related statistics during a competition, using principal component and cluster analyses to determine if it is possible to distinguish the winning teams from the drawing and losing ones. We collected the game-related statistics of the group phase matches of the 2006 World Cup and organised them into a matrix. The principal components of the covariance matrix were calculated. The scores of the first and second components were used to represent the new data, and cluster analysis was applied to separate the elements in two groups (G1 and G2). To analyse the degree of separation between the groups, we calculated the Silhouette Coefficient for each group. Finally, we checked if the winning teams were classified into the same group. The Silhouette Coefficients found for G1 and G2 were 0.54 and 0.55, respectively. Results showed that 70.3% of the winning teams were classified into the same group (G1). Similarly, 67.8% of the drawing and losing teams were classified in G2. This study presented a different way to analyse game-related statistics that allowed the multivariate differences to be shown between successful and unsuccessful teams.

Keywords: *principal components, cluster analysis, notational analysis, match performance*

Introduction

A considerable amount of research has been devoted to establishing the need for objective forms of sports analysis and their importance to the coaching process (Hughes, 1996). Notation systems have been developed to register and store information about players' actions during competition periods and then to define and identify critical elements of the performance (Hughes & Franks, 1997). For football analysis, some of these systems provide game-related statistics about players' actions (such as control, passing, shots on goal, fouls, etc.) that may provide valuable information about the performance of successful and unsuccessful teams during a match.

Related literature has presented some contradictory findings about the differences in the performance and the features of build-up plays between successful and unsuccessful teams during official matches. A previous study (Garganta, Maia, & Basto, 1997) showed that the scoring movements of top-level European football teams often start on their own attacking third of the pitch, revealing a short attacking time (10 s or less) and performing few passes. These findings are in agreement with the "direct-play" idea (Bate, 1988), which supported a game strategy with fewer passes per team possession. Several teams have achieved a measure of success using this strategy, particularly in progressing from lower divisions of English football (Hughes & Franks, 2005). Recent studies also suggest that counter-attacks are more effective than elaborate attacks when playing against an unbalanced defence (Tenga, Holme, Ronglan, & Bahr, 2010a, 2010b). On the other hand, a previous study reported that successful teams touched the ball more times than unsuccessful teams (Hughes, Robertson, & Nicholson, 1988). Additionally, some recent studies also suggest that successful teams attempt more shots on goal per game (Grant, Willians, & Reilly, 1999; Lago-Penas, Lago-Ballesteros, Dellal, & Gomez, 2010).

Ball possession has also been reported as a variable that discriminates between winning, drawing and losing teams (Lago-Penas et al., 2010). Previous research has suggested that successful teams have longer possessions than unsuccessful teams.

Correspondence: Felipe Arruda Moura, State University of Londrina, Sport Sciences Department – Rodovia Celso Garcia Cid, Pr 445 Km 380, Campus Universitário, 86057-970, Londrina, Brazil. E-mail: felipeunesp@yahoo.com.br

© 2013 Taylor & Francis

Downloaded by [Memorial University of Newfoundland] at 18:41 17 July 2014

However, both successful and unsuccessful teams had longer durations of possession when they were losing matches compared to when they were winning (Jones, James, & Mellalieu, 2004; Lago, 2009; Lago & Martin, 2007).

Great part of these literature studies used hand notation and computational notational systems to obtain the data to be processed. Equally, football coaches and assistant coaches use notational analysis to evaluate their teams and opponents. Nevertheless, these analyses generally provide a huge quantity of raw data from the consideration of an entire championship that is quite difficult to analyse or to determine which teams had the best performance or what the relationship is between the quantified variables. Additionally, due to the contradictory findings on the literature, one can argue if the variables analysed are really good estimators of successful performance. If these variables are good estimators, detailed analysis is necessary in order to identify which of them are determinants of winning performance. Multivariate analysis can be particularly helpful in better understanding the kinds of data collected during football matches.

Multivariate analysis methods represent a mixture of matrix algebra, geometry and statistics. Multivariate techniques are useful for discovering regularities in the behaviour of two or more variables and for testing alternative models of association between variables, including the determination of whether and how two or more groups differ in their "multivariate profiles" (Carroll, Green, & Chaturvedi, 1997).

One of the most used multivariate analysis techniques is the principal component analysis (PCA). The central idea of PCA is to reduce the dimensions of data that have a large number of interrelated variables while preserving the maximal variance. This reduction is achieved by the transformation of new uncorrelated data, the principal components, which are ordered in such way that the first components retain the greater part of the variance present on the original variables (Jolliffe, 2002).

The principal component analysis has a great number of applications, and some previous research has employed this technique to analyse data collected in official sporting competitions. Barros, Cunha, Magalhães, and Guimarães (2006) applied PCA to represent and quantify the pitch region that different football players played and, using these analyses, to provide tactical information about the team. Additionally, two previous studies (Dawkins, 1989; Naik & Khattree, 1996) applied different PCA techniques to analyse the men's and women's track records in the 1984 Olympic Games. Using the results of PCA, the authors discussed and classified countries with the best performance based on all races.

Since the purpose of using notational analysis is to provide a performance evaluation of teams during matches, a very important issue to investigate is if the interrelated data provided by notational analysis can discriminate winning teams from the drawing and losing ones. For such classification, cluster analysis can be applied with PCA in order to reach different objectives. The PCA can provide a reduction of data dimensions and, at the same time, can highlight the most important variables that should be analysed from the game-related statistics. On the other hand, cluster analysis is a technique used for data classification in which similar data elements are partitioned into groups called clusters (Kaufman & Rousseeuw, 1990). Thus, cluster analysis can yield a classification of different teams into groups, according to their performance similarities.

Therefore, the aim of this study was to analyse game-related statistics during an official competition using multivariate techniques. We hypothesised that the use of principal component and cluster analyses on the notational analysis data would allow the discrimination of winning teams from drawing and losing teams. If the hypothesis is confirmed, as a practical application, the more relevant variables in this classification will be presented.

Methods

The game-related statistics data from all the group phase matches of the 2006 FIFA World Cup, which were publically available from the FIFA website (FIFA, 2012), were analysed in the present study. To perform the proposed multivariate analysis, the following actions of the players were organised by team during each match: shots, shots on goal, goals performed, fouls committed, fouls suffered, corner kicks, direct free kicks to goal, offside, own goals, yellow cards, expulsions due to second yellow cards, direct expulsions (red cards), actual playing time with possession of the ball and percentage of ball possession in relation to the total time played.

Since the 2006 World Cup had 32 participant countries and each team played 3 matches during the group phase, we stored all the game-related statistics in a matrix M of 96 lines (32 teams playing 3 matches) by 14 columns (14 variables analysed).

Principal component analysis

The principal component analysis consisted of the calculation of eigenvectors and eigenvalues from the covariance matrix of M. Eigenvectors are the vectors of coefficients corresponding to eigenvalues and were used to calculate the results. Thus, the coefficients represent the loading factors of each original variable to obtain the new transformed data, and the

Downloaded by [Memorial University of Newfoundland] at 18:41 17 July 2014

positive or negative value represents a direct or inverse proportionality, respectively. The eigenvalues represented the variances of each component, so that the first eigenvalues retained the greater part of the variance. Then, the scores were calculated by multiplying the original data (centred in the mean value) by the eigenvectors. Finally, we selected the first and second components to represent the new data.

Cluster analysis

Cluster analysis was applied in order to classify the teams into two groups, labelled G1 and G2, using the scores calculated for the first principal components. We applied the *k-means* method, indicating previously that the data would be separated into two groups (Gan, Ma, & Wu, 2007). Then, we checked how many winning teams were classified into the first group and how many drawing and losing teams were classified into the other group.

To analyse the degree of separation between groups, we calculated the Silhouette Coefficient (SC) for each group (Kaufman & Rousseeuw, 1990) as follows:

Let $k = 2$ be the number of clusters defined previously and $j = 1,\ldots,$ the number of points of cluster A;
$a(j)$ = the average distance between the point j and all the other points of cluster A to which j belongs;
$d(j,B)$ = the distances between point j and each point of cluster B.
After computing $d(j,B)$, we selected the smallest value of those:
$b(j) = \min d(j,B)$, where $B \neq A$.

Then, we calculated $s(j)$, the silhouette value of each point j:

$$s(j) = \frac{b(j) - a(j)}{\max\{a(j), b(j)\}}, -1 \leq s(j) \leq 1$$

Thus, if the element was appropriately classified into its group, the silhouette value would be positive. Conversely, negative values represented a poor classification of the element into its group.

Finally, the Silhouette Coefficient for each group was calculated:

SC = the average $s(j)$ value for all points j.

Results

From the covariance matrix of the game-related statistics from all matches, we found the eigenvalues and the eigenvectors. Since 14 variables were analysed, Table I presents the 14 eigenvalues found for the principal components, representing the variance retained by each of them. The results showed that first two components accounted for 64.5% of the variation. Since the first two components represent the greatest variance, Table II presents the coefficients of the first and the second principal components, referent to each variable.

Table I. Eigenvalues referent to each principal component.

Principal components	Eigenvalues
1st	86.5
2nd	25.9
3rd	23.0
4th	16.4
5th	6.4
6th	4.9
7th	4.2
8th	3.3
9th	1.6
10th	1.1
11th	0.8
12th	0.1
13th	0.0
14th	0.0

With the first and second eigenvectors, we obtained a scores matrix that represented the new transformed data. Score values are represented in Figure 1. Beside each point, the first letters of the team name were presented. Using these score values, we applied the cluster analysis.

Figure 2 presents the results of how each team in each match was classified as either G1 or G2, according to the *k-means* cluster analysis. Furthermore, Figure 3 shows the silhouette value for each point, classified into the cluster groups. The silhouette value allowed for verification that the Silhouette Coefficient for G1 (SC_1) was 0.54 and for G2 (SC_2) was 0.55.

Finally, we analysed how many winning teams were classified into a same group and how many

Table II. Coefficients for the first and the second principal components referent to each variable.

Variables	First component	Second component
Shots	−0.41	−0.06
Shots on goal	−0.25	−0.04
Goals performed	−0.05	−0.03
Fouls committed	0.17	0.87
Fouls suffered	−0.11	0.43
Corner kicks	−0.19	−0.06
Free kicks to goal	−0.01	−0.02
Offside	0.01	−0.01
Own goals	0.00	0.00
Yellow cards	0.04	0.05
Second yellow cards	0.01	0.01
Red cards	0.00	0.00
Playing time with ball possession	−0.41	0.05
Percentage ball possession	−0.72	0.19

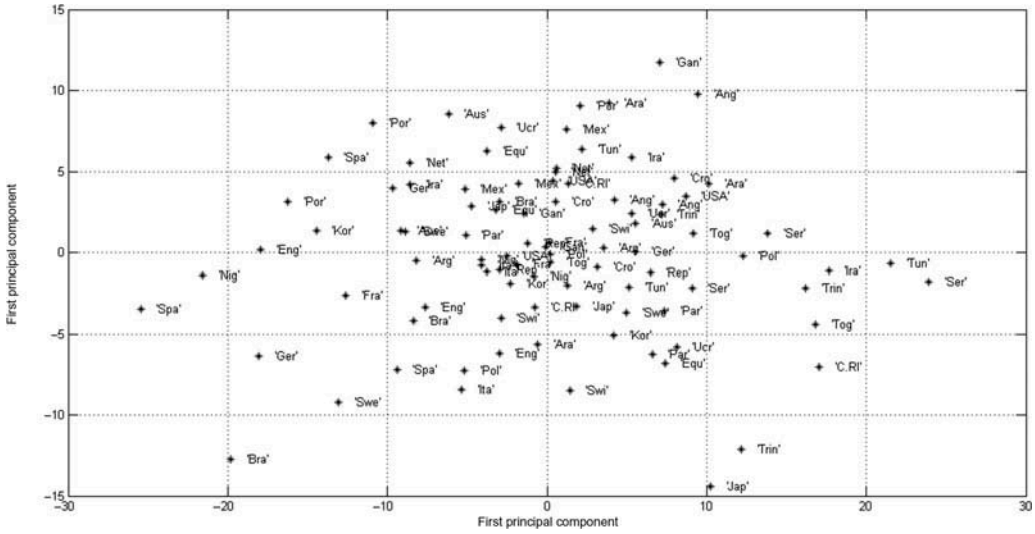


Figure 1. Scores values for the first and second components after the new data transformation.

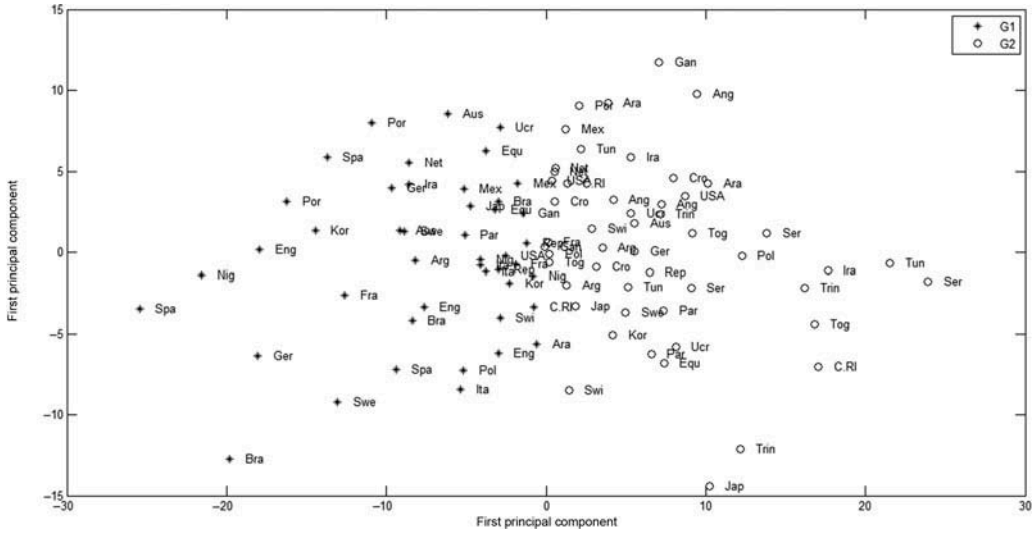


Figure 2. Results of the data classification after the cluster analysis using the *k-means* method.

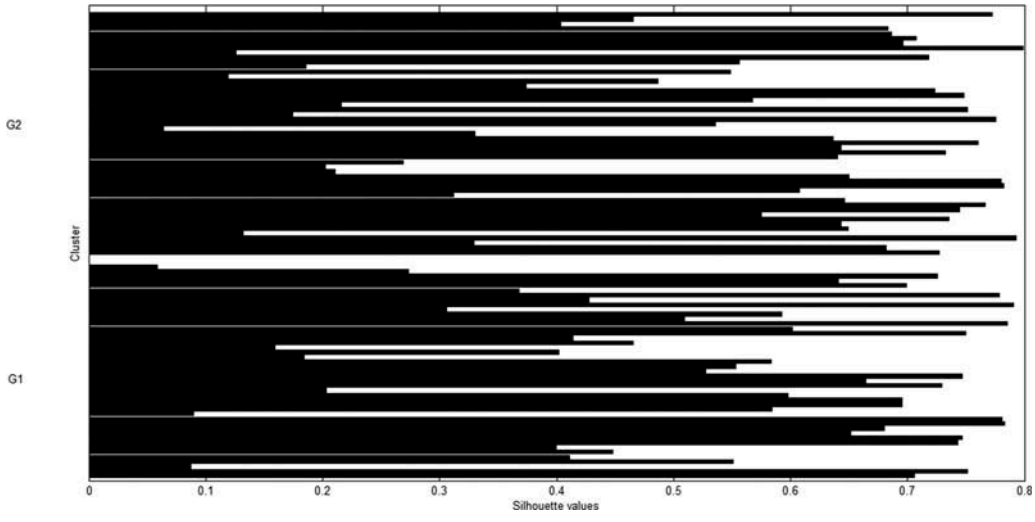


Figure 3. Results of the silhouette value found for each element of G1 and G2.

Downloaded by [Memorial University of Newfoundland] at 18:41 17 July 2014

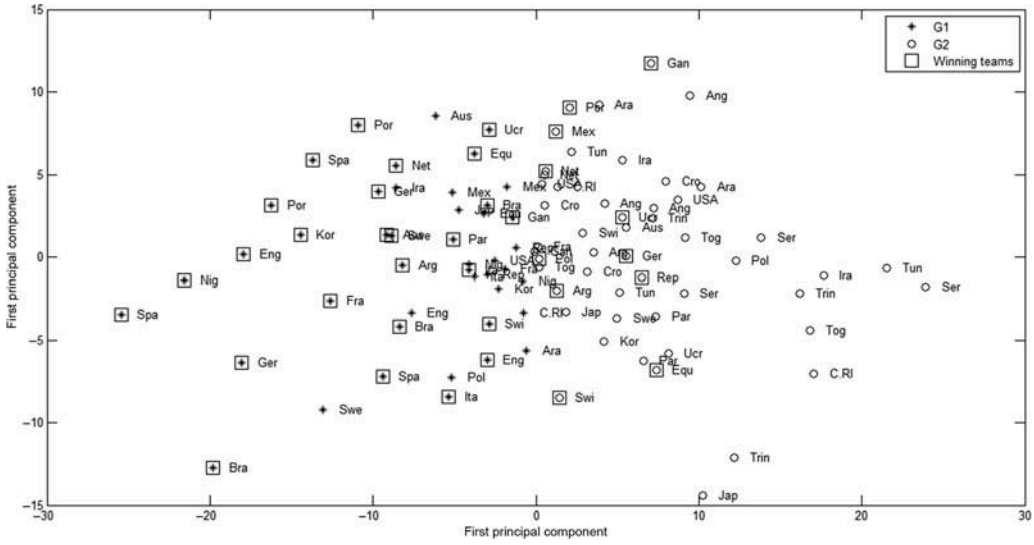


Figure 4. Representation of the winning teams inside each group created after the cluster analysis.

drawing and losing teams were classified together in other group. Figure 4 illustrates the cluster classification and the elements which represent the winning teams. Results showed that 70.3% of the winning teams were classified into the same group (i.e., into the G1). Similarly, 67.8% of the drawing and losing teams were classified into the same group (G2), showing that the cluster analysis was efficient.

Discussion

The purpose of a football match analysis is to correlate technical elements with the match score (Dufour, 1993). However, the major problems are the identification of the critical elements for a successful performance (Hughes & Franks, 1997) and the interpretation and reduction (if necessary) of the large amount of data provided by the notational systems. In order to solve these problems, we presented an exploratory analysis of football game-related statistics using principal component and cluster analyses. Results allowed verifying that it is possible to classify the winning teams according to their game-related statistics similarities. Furthermore, the PCA showed that there are specific variables that are determinants for this classification.

After applying PCA to the game-related statistics for all teams during the group phase of the 2006 FIFA World Cup, we found that the first two principal components retained more than 64% of the total variance (Table I). Although a sensible cut-off is very often in the range of 70–90%, it can sometimes be higher or lower depending on the practical details of a particular data set. When the number of variables is very large, choosing the number of principal components reaching 70% may give an impractically large quantity of components for further analyses. In such cases, the threshold should be set somewhat lower (Jolliffe, 2002). Since our original data had a total of 14 variables, we accepted the cumulative percentage of the total variance provided by the first and second components. Further complex investigations can be performed by adding more components to the analysis. However, it may complicate visual analysis of the results, such as the one provided by Figure 4, since adding more components requires an increase in the number of graph axes.

Cluster analysis was applied to determine if the classification of the elements (represented by each team, in each match) in two groups is able to discriminate winning teams from those that draw or lose. It was possible to verify that the greatest number of winning teams presented a negative score value for the first component (Figure 4). This result may be attributed mainly to the variables "shots", "shots on goal", "playing time with ball possession" and "percentage of ball possession", which presented the smallest coefficient values (Table II). These results are partially in accordance with a recent study (Lago-Penas et al., 2010) which showed that "shots", "shots on goal" and "percentage ball possession" are some variables that discriminate among winning, drawing and losing teams. However, the authors used discriminant analysis within a different group of variables from the present study that can explain why they also concluded that "crosses", "crosses against" and "venues" are important variables for the teams' classification according to match scores.

In the present study, the percentage of ball possession was the variable that presented the coefficient with the greatest absolute value, showing that

Downloaded by [Memorial University of Newfoundland] at 18:41 17 July 2014

Downloaded by [Memorial University of Newfoundland] at 18:41 17 July 2014

it was the variable with the greatest effect on negative values in the first component and, therefore, was an important variable for the team's classification in the winning group. Some researchers affirmed that ball possession is the most popular performance indicator in football and that successful teams have longer possessions than unsuccessful teams (Jones et al., 2004; Lago & Martin, 2007). Furthermore, "total shots" presented a great absolute coefficient value, a result which is in accordance with studies that support a greater number of shots on goal for successful teams (Grant et al., 1999; Lago-Penas et al., 2010).

In order to evaluate the degree of separation between the groups, we calculated their Silhouette Coefficients (SC), attaining values of 0.54 and 0.55 for G1 and G2, respectively. According to Kaufman and Rousseeuw (1990), when SC has a value between 0.51 and 0.7, a reasonable structure has been found. Therefore, the SC for both groups indicated that the elements were well classified in their respective groups.

In order to evaluate the effectiveness of cluster analysis, we checked if winning teams were classified into the first group, while drawing and losing teams were classified into the second group. We found that the greater part (70.3%) of winning teams was correctly classified into the same group (Figure 4). This result was greater than the value of 58.6% into which winning teams were correctly classified in a previous study (Lago-Penas et al., 2010). However, the authors used discriminant analysis, with some different variables, to classify the teams in three groups (winning, drawing and losing), while in our study, two groups were formed. Similarly, we found that 67.8% of the drawing and losing teams were appropriately classified into the same group. Therefore, our initial hypothesis that principal component and cluster analyses are effective in the discrimination of winning teams from drawing and losing teams was confirmed. However, further studies should be addressed to improve the investigations about what are the determinant features of winning teams. Even though our findings had shown a correct classification of roughly 70% of the winning teams into the same group, which is a promising result compared to the literature findings, higher values of this percentage may provide to coaches even greater understanding about the indices related to successful performance that can help to increase winning chances. Furthermore, the present study classified teams into two groups, considering the drawing and losing teams as unsuccessful performance. Indeed, in some specific situations, drawing can be a good score according to the team purpose during the match. In other words, a draw could be enough for a given team to be qualified to the next phase of a competition. Thus, future cluster analysis can consider classifying teams into three groups (wining, drawing and losing teams).

Finally, it is important to emphasise that this study analysed data from a secondary source from 2006, and more updated data should be evaluated to check if the results are reproducible for more recent championships.

Conclusions

The main purpose of this study was to analyse the game-related statistics of football teams during an official competition using principal component and cluster analyses to evaluate whether it is possible to distinguish winning teams from drawing and losing teams. The results showed that the analyses were effective because they facilitated the correct classification of the greater part of winning teams in the same group. As a practical application, our findings suggest that the match analysis performed by coaches and assistant coaches is really important to evaluate team performance, but some variables have greater relevance. Specifically, the variables "shots", "shots on goal", "playing time with ball possession" and "percentage of ball possession" are important to discriminate the winning teams from the drawing and losing ones. Thus, these variables should receive special attention during training sessions and competition.

Future investigations should attempt to apply principal component and cluster analyses to the game-related statistics of other sports when it is desirable to reduce the dimensions of the data and to classify the elements into groups according to their performance similarities.

References

Barros, R. M. L., Cunha, S. A., Magalhães, Jr., W. J., & Guimarães, M. F. (2006). Representation and analysis of soccer players' actions using principal components. *Journal of Human Movement Studies, 51*, 103–116.

Bate, R. (1988). Football chance: Tactics and strategy. In T. Reilly, A. Lees, K. Davis, & W. J. Murphy (Eds.), *Science and football* (pp. 293–301). London: E & FN SPON.

Carroll, J. D., Green, P. E., & Chaturvedi, A. (1997). *Mathematical tools for applied multivariate analysis* (rev. ed.). San Diego, CA: Academic Press.

Dawkins, B. (1989). Multivariate analysis of national track records. *The American Statistician, 43*(2), 110–115.

Dufour, W. (1993). Computer assisted scouting in soccer. In T. Reilly, J. Clarys, & A. Stibbe (Eds.), *Science and football II* (pp. 160–166). London: E & FN SPON.

FIFA. (2012). *Matches statistics*. Retrieved from http://www.fifa.com/worldcup/archive/germany2006/results/index.html

Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: Theory, algorithms, and applications*. Philadelphia, PA: SIAM & American Statistical Association.

Garganta, J., Maia, J., & Basto, F. (1997). Analysis of goal-scoring patterns in European top level soccer teams. In T. Reilly, J. Bangsbo, & M. Hughes (Eds.), *Science and football III* (pp. 246–250). London: E & FN SPON.

Grant, A. G., Willians, A. M., & Reilly, T. (1999). Analysis of the goals scored in the 1998 World Cup. *Journal of Sports Sciences, 17*(10), 826–827.

Hughes, M. (1996). Notational analysis. In T. Reilly (Ed.), *Science and soccer* (pp. 343–361). Great Britain: E & FN Spon.

Hughes, M., & Franks, I. (2005). Analysis of passing sequences, shots and goals in soccer. *Journal of Sports Sciences, 23*(5), 509–514.

Hughes, M., & Franks, I. M. (1997). *Notational analysis of sport* (1st ed.). London: E & FN SPON.

Hughes, M. D., Robertson, K., & Nicholson, A. (1988). An analysis of the 1984 World Cup of Association Football. In T. Reilly, A. Lees, K. Davids, & W. J. Murphy (Eds.), *Science and football* (pp. 363–367). London: E & FN SPON.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York, NY: Springer.

Jones, P. D., James, N., & Mellalieu, S. D. (2004). Possession as a performance indicator in soccer. *International Journal of Performance Analysis in Sport, 4*(1), 98–102.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York, NY: Wiley.

Lago, C. (2009). The influence of match location, quality of opposition, and match status on possession strategies in professional association football. *Journal of Sports Science, 27*(13), 1463–1469.

Lago, C., & Martin, R. (2007). Determinants of possession of the ball in soccer. *Journal of Sports Science, 25*(9), 969–974.

Lago-Penas, C., Lago-Ballesteros, J., Dellal, A., & Gomez, M. (2010). Game-related statistics that discriminated winning, drawing and losing teams from the Spanish soccer league. *Journal of Sports Science and Medicine, 9*(2), 288–293.

Naik, D. N., & Khattree, R. (1996). Revisiting olympic track records: Some practical considerations in the principal component analysis. *The American Statistician, 50*(2), 140–144.

Tenga, A., Holme, I., Ronglan, L. T., & Bahr, R. (2010a). Effect of playing tactics on achieving score-box possessions in a random series of team possessions from Norwegian professional soccer matches. *Journal of Sports Science, 28*(3), 245–255.

Tenga, A., Holme, I., Ronglan, L. T., & Bahr, R. (2010b). Effect of playing tactics on goal scoring in Norwegian professional soccer. *Journal of Sports Science, 28*(3), 237–244.

Downloaded by [Memorial University of Newfoundland] at 18:41 17 July 2014