Journal of the Korean Data & Information Science Society
2016, **27**(5), 1413–1421

http://dx.doi.org/10.7465/jkdi.2016.27.5.1413
한국데이터정보과학회지

Graphical exploratory data analysis for ball games in sports[†]

Seongbaek Yi[1] · Dae-Heung Jang[2]

[12]Department of Statistics, Pukyong National University

Received 5 July 2016, revised 12 August 2016, accepted 7 September 2016

Abstract

In this paper graphical exploratory data analyses are proposed for ball games in sports. The plot of sequence of scoring points of each team can be used to see how the playing game has been processed until the end of each set or quarter. With the plot of sequential score differences through all the games we can see a dominance of each team and the times of score changes, i.e., turnovers. The ternary plots show the contours of scoring compositions for each player and enable us to compare the scoring patterns of each team if any. Using the score sequence plot we also can see the score pattern distribution of players. For demonstration we use the results of the gold medal match between Russia and Brazil for men's volleyball and between USA and Spain for men's basketball at the London 2012 Summer Olympics.

Keywords: Basketball, graphical exploratory data analysis, ternary plot, volleyball.

1. Introduction

There have been multivariate statistical methods to describe any characterizations of sports matches. Dawkins (1989) used principal components analysis to investigate the American national track records in the 1984 Los Angeles Olympic games, and Dawkins *et al.* (1994) applied an incremental clustering algorithm of Andreae *et al.* (1993) to the women's heptathlon data of the 1992 Barcelona Olympic games. Kim *et al.* (2008) analyzed records of Korean professional basketball matches to identify factors influencing each teams winning or losing using logistic regression and regression trees. Pak (2016) did clustering Korean baseball teams based on the records to investigate any pattern of each team against a specific team using time series clustering. Hong *et al.* (2016) proposed a hitting ability index for representing batting ability of Korean pro-baseball players with regression technique.

Especially graphical representation of the results of sports matches is a good tool to understand them intuitively. Westfall (1990) summarized the scoring activity of an American basketball game using a real-time bar plot of the score difference, not the usual score, versus the elapsed time, which showed interesting features such as the time points of largest leads, advance from a losing position, and time length for which a specific team holds a leading

[†] This work was supported by a Research Grant of Pukyong National University (CD-2015-0813).
[1] Corresponding author: Professor, Department of Statistics, Pukyong National University, Busan 608-737, Korea. E-mail:sbyi0108@gmail.com.
[2] Professor, Department of Statistics, Pukyong National University, Busan 608-737, Korea.

position. Chun (2009) analysed mens Korean pro-volleyball games, using the plot showing the process of obtaining scores, but which does not show time points of advancing from a losing position.

In this paper we attempt another graphical representations, among which there exist modifications of existing methods in a sense, for the London 2012 Olympic games data of men's volleyball and men's basketball matches as well.

2. Graphical exploratory data analysis for volleyball

In the volleyball the first team to score 25 points by a two-point margin is awarded the set. Matches are best-of-five sets and the fifth set is played to 15 points. Russia rallied from two sets down to a five-set victory over Brazil, the world No.1 and captured its first Olympic gold medal. Russia's stunning 19-25, 20-25, 29-27, 25-22, 15-9 win prevented Brazil from a volleyball sweep at the London Games. The relevant data can be found at the site (`http://www.london2012.com/volleyball/event/men/match=vom400101/index.html`). The results of the first two sets are represented graphically in Figure 2.1.

Brazil was consistently superior in the first set, which is shown in Figure 2.1. by the fact that the steps are never below the 45-degree line. At the second set Brazil was good overall, but Russia contended fiercely with Brazil at the early of the game and at 15 points as well. This can be explained by the steps in Figure 2.1 which are below the 45-degree line until the 5th point and get close to the 45-degree line at the point 15.

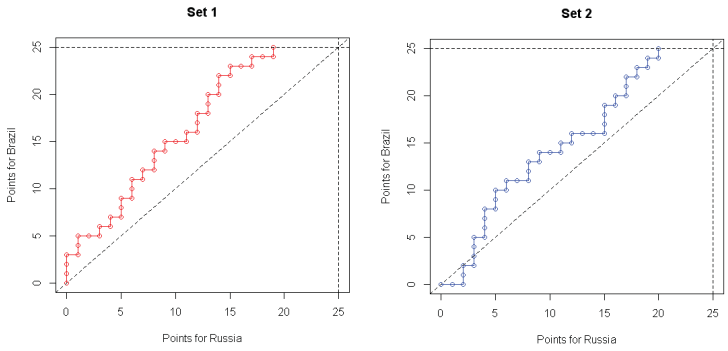

Figure 2.1 The first two sets of the final match at men's volleyball

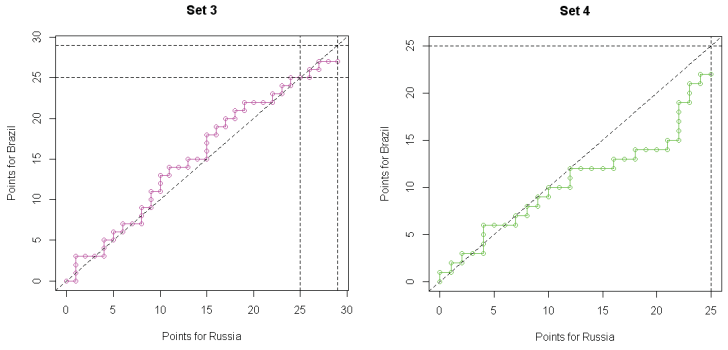

Figure 2.2 The sets 3 and 4 of the final match at men's volleyball

The third set was a close competitive game for the two teams. They had 15 times of tie count and did not have 25 points win, finally Russia had 29-27 win. In the fourth set both teams had comparable plays until the point 12, but onwards Russia got the dance of the game. Figure 2.2 stipulates the reality. The steps in the plot of the third set move upwards along the 45-degree line, not going away. This pattern continues in the plot of the set 4 until the point 12, but afterwards the steps never move across but below the 45-degree line.

Russia closed out the final set at 15-9 without ever being led. This is represented in Figure 2.3 by the constant position of the steps below the 45-degree line. In this way we can see the score patterns of each set at Figures 2.1-2.3. This plot can be considered similar to those in the blog `http://analyticsdemystified.com` by Tim Wilson. However, the plot in this paper describes the process using the 45-degree line, which means the closer to the line, the more competitive the game is.

The game can be examined in another way using Figure 2.4, the plot of the sequential score differences through all the five sets. This plot has been used a lot in many articles, thus we show its explanation here. In the first set were there 44 times of score changes, but Brazil took it 25-19 without ever reversing and had advanced up to 8 points in the middle of the set. At the beginning of the second set Brazil had been led but had overall superiority to take 25-20 win. In the third set the score differences were not big with 3 points at most, which implies the set was very close. Russia continued to catch up Brazil from the latter part and eventually took the set 29-27. At the start of the fourth set the score gap was one or two points at most, but Russia took one-way traffic after 8-7 and took the set 25-22. In the final fifth set Russia continued to dominate, and eventually closed it out at 15-9.

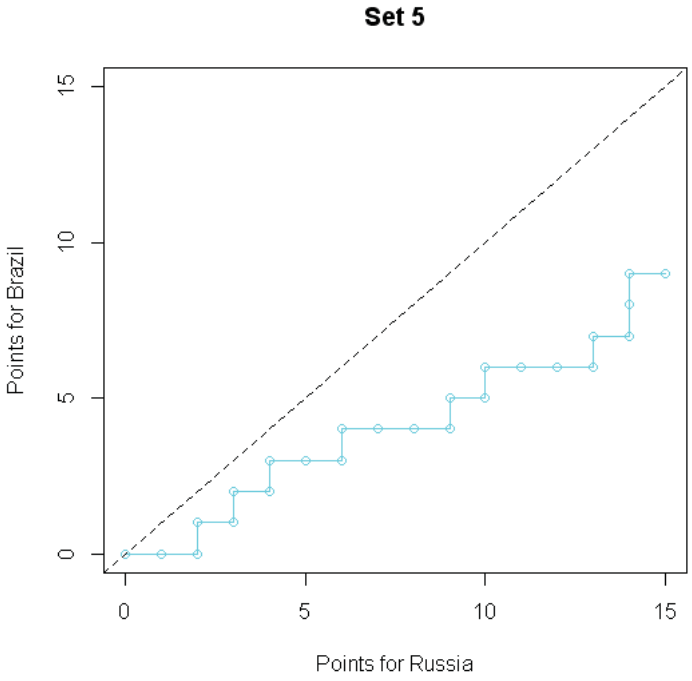

Figure 2.3 The final set of the gold medal match at men's volleyball

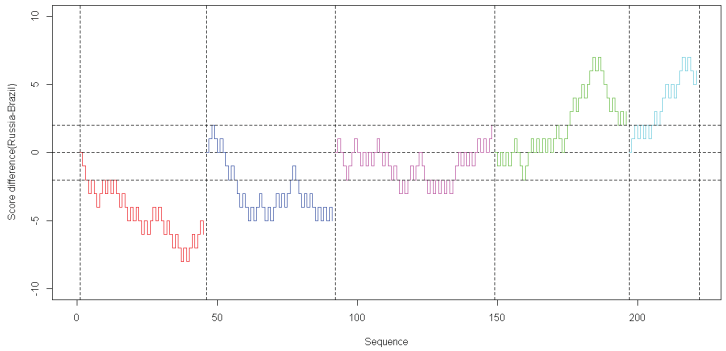

Figure 2.4 The score differences for the final match at men's volleyball

Volleyball has three sources of scoring such as serves, spikes and blocks. The scoring sources of the two teams at the gold medal match of the 2012 London Summer Olympics are in Table 2.1. It is by spikes that the two teams got more than 70% points among the scores by attack and counter-attack techniques.

In order to see the composition of scores by the techniques for each player we can use a ternary plot with the following compositional data:

$$x_1 + x_2 + x_3 = 1,\ 0 \le x_i \le 1,\ \text{for } i = 1, 2, 3,$$

here x_1 is the proportion of scoring by serves, x_2 by spikes, and x_3 by blocks. A ternary plot is a barycenteric plot on the three variables which sum to one (see Aitchison 1986). It shows the composition of scores by individual players, depicting graphically the ratios of three compositions as positions in an equilateral triangle. Figure 2.5 shows the contours of three types of scorings for each player of the two teams. The ternary plot shows that leading players got scores mainly by spikes. But the patterns of the two teams are different.

Table 2.1 Contingency table for the three scoring techniques: scores (%)

Country	Serves	Spikes	Blocks	Scores
Russia	4 (4.94)	62 (76.54)	15 (18.52)	81 (100.00)
Brazil	8 (10.39)	59 (76.62)	10 (12.99)	77 (100.00)

Figure 2.6 is a scatter matrix representing scoring patterns for players of each team. The two scatter plots of spikes vs blocks and blocks vs points indicate that Russian team has two groups of the nine exclusively defense-oriented players and the three defense-offense players. One player with high scoring by services lies at the border line of the two groups. Brazil consists of the two groups with similar pattern who are the ten defensive players and the two defense-offense players. Two players with high scoring in services are located at the borderline. But the pattern is not as clear as in Russia.

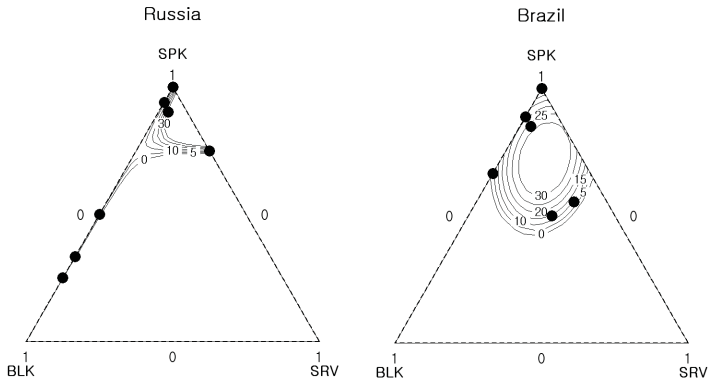

Figure 2.5 Ternary plot for the score patterns of players at the final match of men's volleyball (SRV:service, SPK:spike, BLK:block)

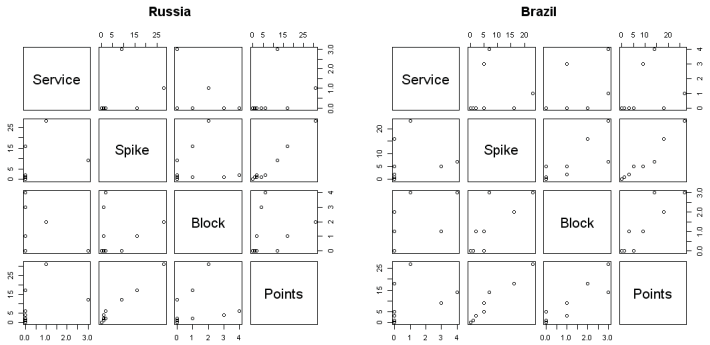

Figure 2.6 Scatter matrix plot for score patterns of players for the final match at men's volleyball

3. Graphical exploratory data analysis for basketball

The basketball games are played in four quarters of 10 minutes at the Olympics. The winner is the team with more points. In the men's basketball final of the 2012 London Olympics did the USA defeat Spain with the score 107-100. The relevant data can be obtained from the website (`http://www.london2012.com/basketball/event/men/match=bkm400101/index.html`).

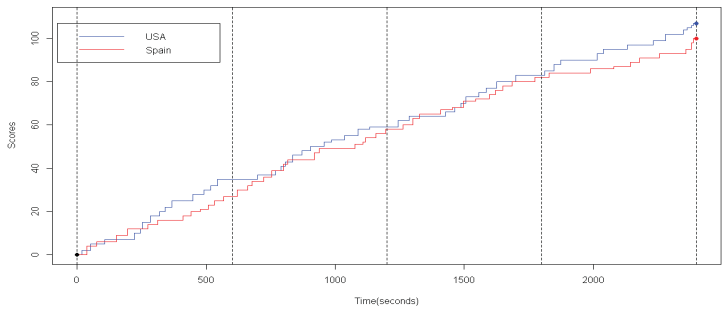

Figure 3.1 The scores for the final match at men's basketball

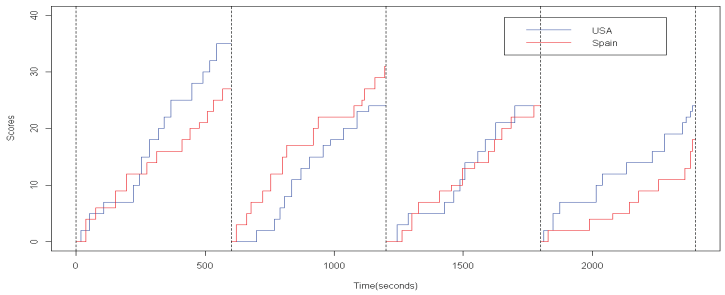

Figure 3.2 The quarter scores for the final match at men's basketball

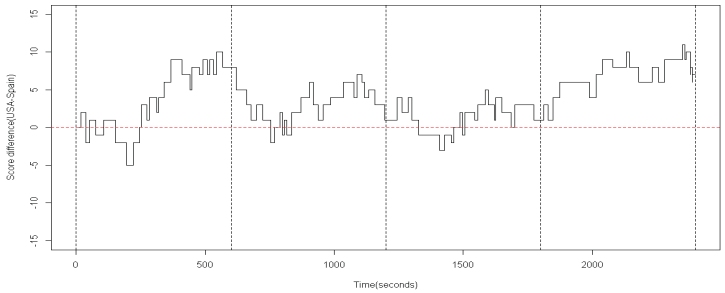

Figure 3.3 The score differences for the final match at men's basketball

The results of the game is represented graphically in Figure 3.1. The first quarter had many turnovers in the early part but the USA took 8 point lead. The gap of scores got narrow up to one point in the second quarter. In the third quarter the two teams had turnovers again and the USA took it one point lead. At the final quarter the USA continued to dominate and closed out the game by 7 points lead.

Figure 3.2 shows the scores of each quarter for the closer look at this pattern. The score differences over sequential time is plotted at Figure 3.3. It shows there were 32, 36, 27 and 23 times of score changes, and 6, 6, 3 and zero times of reversion, respectively in the four quarters. This kind of plots are shown in Tim Wilson's blog `http://analyticsdemystified.com`. If we examine the plots by decomposing into four quarters, we can get deeper understanding of the game.

The basketball has scoring by three shot methods such as two points (P2), three points (P3) and free throws (FT). The scoring points of each method for the two teams are in Table 3.1. The USA had 42% of points by shooting three points, while Spain did 52% by two-points. Since the basketball has three sources of shooting, we can express the scoring composition of each player by the following compositional data:

$$x_1 + x_2 + x_3 = 1,\ 0 \le x_i \le 1,\ \text{for } i = 1, 2, 3,$$

here x_1 is the proportion of scoring by two-points, x_2 by three-points, and x_3 by free-throws.

Table 3.1 Contingency table for scores (%) by three shot methods

Country	2-points (P2)	3-points (P3)	Free Throws (FT)	Score
USA	38 (35.51)	45 (42.06)	24 (22.43)	107 (100.00)
Spain	52 (52.00)	21 (21.00)	27 (27.00)	100 (100.00)

Figure 3.4 is a ternary plot, showing the contours of scoring composition for each player of the two teams. The high-score players in the USA team got scores mainly with three-points, while the high-score players in the Spain team obtained by two points or three points. However the scoring patterns of the two teams are different. The ternary plot shows pattern of obtaining scores for each team, and thus it is helpful for prospective counter teams to set up their strategies toward corresponding team.

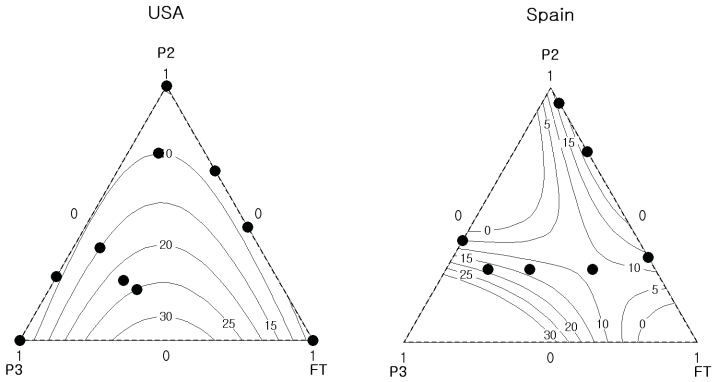

Figure 3.4 Ternary plot of the score patterns of players for the final match at men's basketball

Figure 3.5 is a scatter matrix showing scoring pattern for individual player of the two teams. The USA has two outliers (one player with most two points and the other with most three points) and the Spain three outliers (two players with two points and the other with most three points).

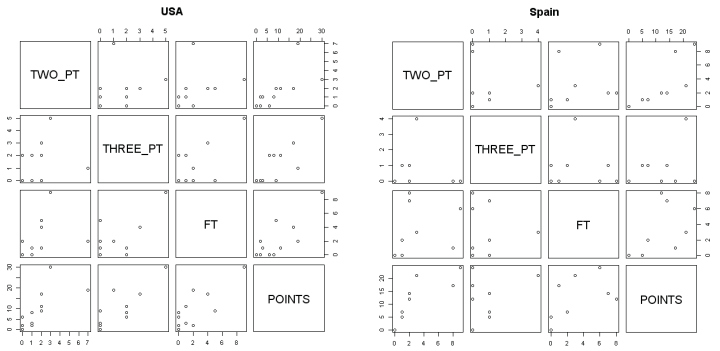

Figure 3.5 Scatter matrix plot for score patterns of players for the final match at men's basketball (P2: two-points, P3: three-points, FT: free-throws)

We can show the scoring pattern of each team using score sequence plot. In Figure 3.6 yellow, brown and red colors represent free throws, two points, and three points respectively. It shows that the USA has more three points, while the Spain more two points. This plot also reveals the scoring patterns of individual players at each quarter. The names of players with the uniform numbers of Figure 3.6 are listed in Table 3.2.

Figure 3.6 Score sequence plot representing the score pattern distribution of players for gold medal match of men's basketball

Table 3.2 Players table

Country	Number	Name	Country	Number	Name
USA	1	Chandler T	Spain	1	Gasol P
	2	Durant K		2	Fernandez R
	3	James L		3	Rodriguez S
	4	Westbrook R		4	Navarro J
	5	Williams D		5	Calderon J
	6	Iguodala A		6	Reyes F
	7	Bryant K		7	claver V
	8	Love K		8	San Emeterio F
	9	Harden J		9	llull S
	10	Paul C		10	Gasol M
	11	Davis A		11	Ibaka S
	12	Anthony C		12	Sada V

4. Conclusion

In this paper we proposed graphical exploratory data analysis for the sports of volleyball and basketball. For the data of the 2012 London Summer Olympics we explored the results by graphical methods. The graphical expressions enabled us to figure out easily the process of each game and the scoring patterns of each team and each player as well.

References

Aitchison, J. (1986). *The statistical analysis of compositional data.* Chapman & Hall, London.

Andreae, P. M., Dawkins, B. P. and O'Connor, P. M. (1993). *DySect: An incremental clustering algorithm.* Technical Report 33, Victoria University of Wellington, ISOR, New Zealand .

Chun, Y-Y. (2009). The game analysis with score obtained process from man pro-volleyball game. *Journal of Sport and Leisure Studies*, **38**, 1305-1312.

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

Dawkins, B. P., Andreae, P. M. and O'Connor, P. M. (1994). Analysis of olympic heptathlon data. *Journal of the American Statistical Association*, **89**, 1100-1106.

Hong, C. S., Kim, J. Y. and Shin, D. S. (2016). Alternative hitting ability index for KBO. *Journal of the Korean Data & Information Science Society*, **27**, 677-687.

Kim, S. H., Kang, S. J., Park, J. H. and Kim, H. J. (2008). The factor of victory and defeat through analyzing the data of the pro-basketball. *The Korean Journal of Measurement and Evaluation in Physical Education and Sports Science*, **10**, 1-12.

Pak, R. J. (2016). Categorical time series clustering: Case study of Korean pro-baseball data. *Journal of the Korean Data & Information Science Society*, **27**, 3, 621-627.

Westfall, P. H. (1990). Graphical representation of a basketball game. *The American Statistician*, **44**, 305-307.