

Normalising Data: Definitions and Differences

Benjamin Stanway¹

¹Faculty of Life Sciences and Education, University of South Wales

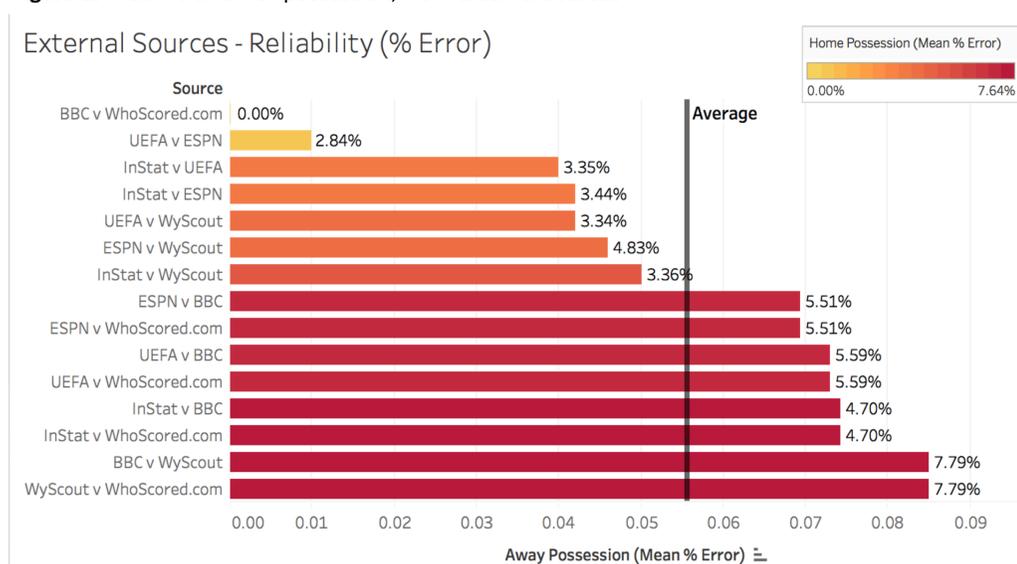
Introduction

When data is analysed in performance analysis (PA) research, technical and tactical actions should be normalised (Hughes and Bartlett, 2002; 2004). In certain invasion games, possession has been used to normalise data, with Soccer (Russell *et al.*, 2013; Tenga and Sigmundstad, 2011; Lui *et al.*, 2015) and Basketball (Csataljay *et al.*, 2009; Sampaio and Janeira, 2003) being the most prominent. Normalising data enables accurate differences to be established, for tactical/technical variables, limiting the amount the amount of error (Lames and McGarry, 2007).

Aim

The study will investigate the mean percentage (%) error, for possession, from six external sources of statistical data, paying particular attention to the operational definitions. When analysing sporting performance, definitions are of importance with a need for consensus, for comparison between data sets (Sarmiento *et al.*, 2017).

Figure 1. Mean % error for possession, from external sources



Results

When considering all external sources for possession statistics Figure 1. indicates the highest mean % error to be WyScout v WhoScored.com (7.79%), with the lowest mean % error being BBC v WhoScored.com (0.00%). Figure 2. suggests more mean % error for the Frequency v Time (6.53%) definitions and the least being Time v No Definition (2.58%). Data used by elite soccer organisations (InStat, UEFA and WyScout) indicated differences in definitions. Although WyScout and UEFA both indicated a time definition, a mean % error was established (3.34%). When comparing InStat (frequency definition) to UEFA (time definition, 3.35%) and WyScout (time definition, 3.36%) both sources highlighted differences below the overall sample average (Figure 1.).

References

- Csataljay, G., James, N., Hughes, M and Dancs, H. (2009) 'Performance indicators that distinguish winning and losing teams in basketball', *International Journal of Performance Analysis in Sport*, 9, pp.60-66
- Hughes, M. and Bartlett, R. (2004) *The use of performance indicators in performance analysis*. In *Notational Analysis of Sport*, 2nd edition (Edited by M. Hughes and I. Franks). London: Routledge, pp.166-188.
- Hughes, M. and Franks, I. (2007) *The essentials of performance analysis: an introduction*. Routledge. London and New York.
- Lames, M. and McGarry, T. (2007) 'On the search for reliable performance indicators in game sports', *International Journal of Performance Analysis in Sport*, 7(1), pp.62-79.
- Liu, H., Gomez, M.A., Lago-Peñas, C. and Sampaio, J. (2015) 'Match statistics related to winning in the group stage of 2014 Brazil FIFA World Cup', *Journal of Sports Sciences*, 33(12), pp.1205-1213.
- Mackenzie, R. and Cushion, C. (2013) 'Performance analysis in football: A critical review and implications for future research', *Journal of Sports Sciences*, 31(6), pp.639-676.
- Pojškić, H., Šeparović, V., and Užičanin, E. (2009). 'Differences between successful and unsuccessful basketball teams on the final Olympic tournament', *Acta Kinesiológica*, 3(2), pp.110-114.
- Russell, M., Rees, G. and Kingsley, M.I. (2013) 'Technical demands of soccer match play in the English championship', *The Journal of Strength & Conditioning Research*, 27(10), pp.2869-2873.
- Sampaio, J. and Janeira, M. (2003) 'Statistical analysis of basketball team performance: understanding teams' wins and losses according to a different index of ball possessions', *International Journal of Performance Analysis in Sport*, 3(1), pp.40-49.
- Sarmiento, H., Clemente, F.M., Araújo, D., Davids, K., McRobert, A. and Figueiredo, A. (2017) 'What Performance Analysts Need to Know About Research Trends in Association Football (2012–2016): A Systematic Review', *Sports medicine*, pp.1-38.
- Tenga, A. and Sigmundstad, E. (2011) 'Characteristics of goal-scoring possessions in open play: Comparing the top, in-between and bottom teams from professional soccer league', *International Journal of Performance Analysis in Sport*, 11(3), pp.545-552.
- Worsfold, P. and Macbeth, K. (2009) 'The reliability of television broadcasting statistics in soccer', *International Journal of Performance Analysis in Sport*, 9(3), pp.344-353.

Method

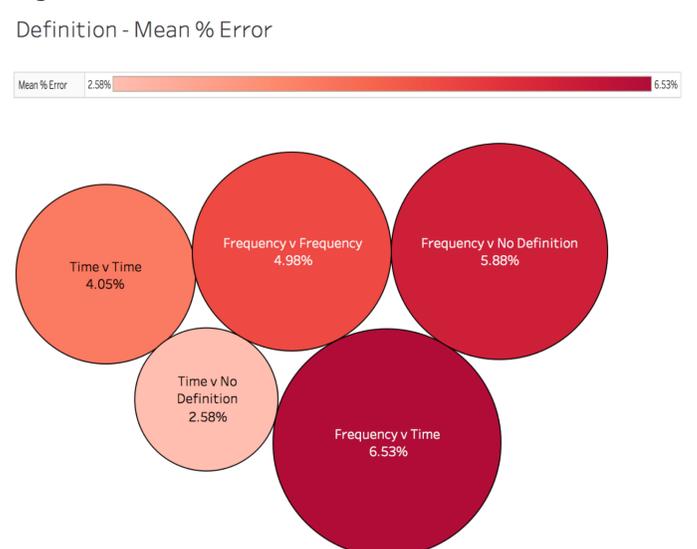
Performance data has been gathered from the UEFA (2016) technical report, BBC (2016), ESPN (2016), WhoScored.com (2016), InStat (2016), WyScout (2016) to gather data for possession from the UEFA European Championships (n=36) in 2016, for each commercial (n=3) and elite (n=3) data source. The performance indicator chosen includes: Possession, definitions for possession are as follows:

Table 1. Definitions of possession for external sources

Source	Definition
ESPN	No definition available.
BBC	Pass frequency, number of passes attempted.
WhoScored.com	Pass frequency, number of passes attempted.
UEFA	Reflective of time with the ball, irrespective of being under control.
WyScout	Reflective of time with the ball, irrespective of being under control.
InStat	Deliberate movement of a player possessing the ball (not less than 3 touches).

Data was analysed using mean % error (Worsfold and Macbeth, 2009; Hughes and Franks, 2007) to determine differences between sources and definitions.

Figure 2. Mean % error between different definitions for



Discussion

Previously, commercial data has been tested for its reliability. Worsfold and Macbeth (2009) found between 10% and 60% mean % error between companies (BBC, ESPN, Sky Sports, Eurosport and Post Match Analyst), compared to the current study, where mean % error was found between 0% and 34%. With a lack of operational definitions and conflicting classifications of activity or playing positions that make it difficult to compare similar groups of studies. Similarities can be drawn for the pair ESPN v BBC, 6.95% mean % error from Worsfold and Macbeth (2009) and 5.51% for the current study. The results suggest a difference in operational definitions, for possession, with two types having been distinguished (attempted passes and time). Other research (Pojškić *et al.*, 2009; Sampaio *et al.*, 2010) suggests that possession could be established using a frequency count. As a consequence, using different possession definitions could lead to errors when analysing technical/tactical variables and negatively influence professional practice (Lames and McGarry, 2007; Worsfold and Macbeth, 2009). Results show disparity, of up to 7.79%, between commercial and elite data sources, for possession statistics. Therefore, practitioners and academics need to use the external commercial and elite data with caution (Worsfold and Macbeth, 2009), and understand how differently data could be interpreted, when normalised with different possession statistics.