Analyzing Collapsibility in Performance Analysis in Sports

Maximilian Klemp(✉), Robert Rein, and Daniel Memmert

German Sport University Cologne, Cologne, Germany
m.klemp@dshs-koeln.de

Abstract. Performance analysis in elite football aims at linking the performance of teams to outcomes. To this end, success indicators (SI) are modelled as a function of performance indicators (PI) using statistical techniques. The aggregation of data on a certain level, however, poses the risk of overlooking effects of contextual factors, which might affect the association between PI and SI. If a contextual factor affects an association in this way, the association is said to be non-collapsible across the respective contextual variable. We analyzed 862 segments from the 1st German Bundesliga and examined the association between running performance and goal scoring with respect to the scoreline. We showed that the effect of running performance on success varies for different levels of scoreline, therefore this association is non-collapsible across scoreline. This finding has important implications for match analysis research, stating the need to account for scoreline and consequently adopting a segmented approach to match analysis.

Keywords: Match Analysis · Collapsibility · Confounding · Football

1 Introduction

1.1 A Subsection Sample

Performance analysis in elite football aims at linking the performance of teams to outcomes. To this end, success indicators (SI) are modelled as a function of performance indicators (PI) using statistical techniques [1]. According to Lepschy and colleagues [1], The design of a match analysis study can be described by specifying the outcome variable (success indicator), the independent variables (performance indicators) examined as well as the statistical method used to model the relationship between PI and SI. It is also common practice to account for certain contextual variables to avoid biases, such as playing location or the teams' quality. However, in order to account for certain contextual information, another aspect of study design has to be considered, namely the level of observation. Mostly, studies examine relationships between PI and SI on the match level, i.e. modelling match outcome as a function of PIs aggregated over the course of a match. However, there are properties of a match that do not remain constant over its course, for example the teams' steady change between attack and defense or the scoreline, which both drastically affect the aims and therefore behavior of the teams. For

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Baca and J. Exel (Eds.): IACSS&ISPAS 2022, AISC 1448, pp. 108–111, 2023.
https://doi.org/10.1007/978-3-031-31772-9_23

example, trailing teams show increased ball possession rates [2] and physical efforts [3]. Since these contextual factors may also affect the probability of further success, ignoring context might pose the risk of introducing confounding effects to performance analysis. In fact, normalizing defensive performance indicators by possession rates reversed the direction of their relationship with a defensive success indicator in a study by Phatak and colleagues [4]. This finding might be grounded in the fact that better teams spend more time in-possession, therefore accumulate a smaller number of absolute defensive actions, while at the same timing winning matches with a higher probability. When accounting for the time in possession, however, the true effect of defensive performance on match success is made visible.

This potential source of bias has already been acknowledged and exhaustively researched in other areas of observational research [5] and is more appropriately denoted by the term non-collapsibility. For example, the statistical phenomenon known as "Simpson's Paradox" can be explained in terms of non-collapsibility [6]. Non-Collapsibility refers to the fact that ignoring a certain contextual variable while examining the effect between two variables will introduce a bias into this examination. Formally, collapsibility of an association between variables X and Y across a variable Z requires that this association is constant across all levels of Z. In the context of performance analysis in sports, an association between a PI and an SI is collapsible across a contextual variable Z, if across all possible expressions of Z the association between PI and SI remains unchanged. If certain contextual variables in a soccer match are found to be non-collapsible with respect to the association between PIs and SIs, this has important implications for performance analysis in soccer, since these contextual variables should then be considered in every study examining such an association. If the value of the contextual variable potentially varies over the course of a match (like scoreline), match-wise aggregations of data and subsequent analysis would not be robust against non-collapsbility bias. In this case, a segmented analysis of soccer matches would be warranted.

2 Methods

We analyzed 302 matches of the 1st German Bundesliga (season 2016/2017), splitting every match after a goal was scored. In this way, we created 1145 segments in which the scoreline remained unchanged during the segment. Goalless matches were excluded from the analysis, leaving 1145 segments out of 283 matches. For every segment, the running distance of both teams was aggregated as the PI under examination. The ratio of the home team's and the away team's running distance was calculated. For every segment, the next goal being scored was defined as the success indicator, with the goals being coded as home goal (1) or away goal (0). Since in every match, the last segment could not be associated with a next goal, these segments were excluded, resulting in 862 segments for further analysis. For every segment, the current scoreline was determined by discretizing the goal difference (GD) from the perspective of the home team into *trailing far* (GD≤ -3)*, trailing close* ($-2 \leq$ GD < 0), *draw* (GD $= 0$)*, leading close* ($0 <$ GD ≤ 2), and *leading far* (goal difference ≥ 3). By framing all variables from the perspective of the home team, the home advantage is taken into account by the base rate of scoring and the results of the subsequent modelling can be interpreted easily. Additionally, a match-level dataset was created, calculating the teams' running distance ratio across the whole match as PI and determining the match outcome (home or away win) for all non-draw matches as the SI. By leaving out draws, 230 observations remained. The association between PI and SI was examined using logistic regression. To analyze collapsibility of this association across scoreline, three different models were fit. The first model (**Model 1**) was fit using the match-level data ($n = 230$), predicting match outcome as a function of match running distance ratio. The second model (**Model 2**) was fit using the segmented data ($n = 862$), predicting the next goal as a function of running distance ratio (per segment). The third model (**Model 3)** was also fit using the segmented data, but predicting the next goal as a function of both running distance ratio and scoreline, including only the interaction effect between running distance ratio and scoreline (which returned a separate coefficient of running distance ratio for every level of scoreline). To assess collapsibility across scoreline, the Odds Ratio of the effect of running distance ratio on winning/scoring probability were compared across models. As stated above, collapsibility across scoreline would require Odds Ratios to be constant across all levels of scoreline and equal to the Odds Ratio in the aggregated data.

3 Results

Table 1 shows the results of the three different logistic regression models. The Odds Ratio of running distance for Model 1 is 0, which means that for match-level data, the probability of winning the match is strongly reduced if the team runs more than the opponent. In Model 2, the Odds Ratio for running distance is 0.69, which still represents a negative effect of running performance on success, while this effect is less pronounced than in Model 1. In Model 3, finally, Odds Ratios for the different levels of scoreline differ considerably, ranging from 0.57 to 4.01. This shows, that the effect of running performance on success changes, depending on the current scoreline. With respect to the initial question, this shows that the association between running performance and success is non-collapsible across different levels of scoreline.

4 Discussion

On the match level, running performance was negatively correlated with winning. In the segmented, but not contextualized data, running performance was still negatively correlated with success, but with a lower effect size. In the segmented, contextualized data, running distance was both positively and negatively related to success, depending on the level of scoreline. This observation is in line with the previously cited findings by Phatak and colleagues [4], who showed a reversal of effects after contextualizing their analysis. In practical terms, given the effect of scoreline on tactical and physical performance [2, 3], it is reasonable to assume that better teams are, on average, ahead for most of the time during a match, therefore covering less distance due to the effect of scoreline. Those teams are still more likely to win the match in the end, which would explain the negative effect on the match level. After segmenting and contextualizing the data, the effect changes considerably.

Table 1. Coefficient table for models examining association between running distance and success

Parameter	Estimate	Std. Error	z	*p*	Odds Ratio
Model 1					
Intercept	19.62	3.69	5.31	< 0.001	-
Running Distance	-18.90	3.66	-5.16	< 0.001	0
Model 2					
Intercept	0.67	0.92	0.73	0.46	-
Running Distance	-0.37	0.91	-0.40	0.69	0.69
Model 3					
Intercept	0.41	0.93	0.43	0.66	-
Running Distance:Leading Far	1.39	1.20	1.15	0.25	4.01
Running Distance:Leading Close	-0.12	0.93	-0.13	0.89	0.88
Running Distance:Drawing	0.05	0.94	0.05	0.96	1.05
Running Distance:Trailing Close	-0.43	0.94	-0.45	0.65	0.65
Running Distance:Trailing Far	-0.57	0.95	-0.60	0.55	0.57

More importantly, from these findings it can be concluded that the association between running distance and scoring the next goal is not collapsible across different scorelines. This implies that scoreline is an important contextual variable potentially affecting associations between other PI and SI. In order to find collapsible segments, matches should be segmented by scoreline. Further work should assess whether other contextual factors, such as ball possession, have to be taken into consideration as well.

References

1. Lepschy, H., Wäsche, H., Woll, A.: How to be successful in football: a systematic review. Open Sports Sci. J. **11**(1), 3–23 (2018)
2. Lago-Peñas, C., Dellal, A.: Ball possession strategies in elite soccer according to the evolution of the match-score: the influence of situational variables. J. Hum. Kinet. **25**, 93–100 (2010)
3. O'Donoghue, P., Robinson, G.: Score-line effect on work-rate in English FA premier league soccer. Int. J. Perform. Anal. Sport **16**(3), 910–923 (2016)
4. Phatak, A.A., et al.: Context is key: normalization as a novel approach to sport specific preprocessing of KPI's for match analysis in soccer. Sci. Rep. **12**(1), 1117 (2022)
5. Greenland, S., Pearl, J., Robins, J.M.: Confounding and collapsibility in causal inference. Stat. Sci. **14**(1), 29–46 (1999)
6. Hernán, M.A., Clayton, D., Keiding, N.: The Simpson's paradox unraveled. Int. J. Epidemiol. **40**(3), 780–785 (2011)