Statistical & Methodological research in sport sciences and sport informatics

Traineeship Report

Christoph Völtzke (9769870)

Supervisors: Fabian Wunderlich (Deutsche Sporthochschule Köln) & Mirjam Moerbeek (Utrecht University)

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Utrecht University

18.01.2023

1 Introduction

In many team sports, the home-field advantage has a positive effect on the success of the home team [1–3]. In particular, the effect of spectators with their supportive chants appears to influence the performance of the home team [4]. For example, spectators have been shown to induce a referee bias that can lead referees to favor the home team [5, 6]. Due to the COVID-19 pandemic, spectator attendance in professional European football has been banned or limited for most games from March 2020 to the end of the 2021/2022 season. Since ghost games (games without spectators) were a rare phenomenon prior to the COVID-19 outbreak, the pandemic provided a unique opportunity to examine the impact of spectator attendance on home-field advantage in a natural experiment [7]. By conducting this natural experiment and distinguishing games with and without spectators on outcome variables relevant to a team's success, causal inferences can be obtained [8].

A recent literature review by Leitner et al. (2022) [9] including 16 empirical studies that examined the aforementioned effect, showed that most studies found a decrease in home advantage in the absence of an audience. However, the decrease in home advantage was measured with different dependent variables (e.g. yellow cards, goals scored, points gained), varying sample sizes of matches included, deviating divisions, and diverse modeling choices. For example, McCarrick et al. (2021) [10] analyzed matches played during the 2019/2020 season, in 15 different leagues spanning 11 countries. As outcome variables, they used the ratio of points, goals, fouls, and cards at home. For statistical modeling, they suggested a random intercept for the home team and fixed effects for the home team's league standing and performance. Whereas, Wunderlich et al. (2021) [11] analyzed matches played in 10 different leagues spanning 6 countries for all seasons from 2010/2011 to 2019/2020. As outcome variables, they choose differences in points, goals, shots, shots on target, betting odds, fouls, and cards between the home and away teams. The statistical model included a random intercept for the respective division and the season as a fixed effect. Therefore, the conclusion of spectator-induced home advantage in professional European football highly depends on the used data, outcome variables, and modeling choices.

Another issue concerns that for all examined studies, the classes were highly imbalanced with a low proportion of ghost games. In fact, when considering the 2019/2020 season alone, as in McMarrick et al. (2021) [10], 3515 games (72.56%) were played with a crowd and 1329 (27.44%) were played without a crowd. Since the publication of the review by Leitner et al. (2022) [9] additional games with banned or limited attendance were played in the 2020/2021 and 2021/2022 seasons that were not included in the reviewed studies.

Therefore, this report aims to validate the effect of spectator-induced home advantage in professional European football by replicating and extending one of the reviewed studies by Leitner et al. (2022) [9]. An exact replication is performed to use it as a starting point for further extensions. Extensions are the inclusion of the 2020/2021 and 2021/2022 seasons as well as extra model-

building choices [12] and exploratory analyses based on other reviewed studies. For this report, the study of Wunderlich et al. (2021) [11] is exactly replicated as it is a peer-reviewed study that analyzed the major professional European football leagues for multiple seasons. In addition, they examined the effect of several dependent variables relevant to a team's success, and they provide a reproducible workflow.

2 Methods

2.1 Data

2.1.1 Replication study

The original data used in the study by Wunderlich et al. (2021) [11] was retrieved from the open source website http://www.football-data.co.uk. The replicated data set includes matches played during the seasons 2010/2011 to 2019/2020. Where the seasons from 2010/2011 to 2018/2019 were analyzed as matches with spectator attendance (N = 36,882) and matches played after March 2020 during the 2019/2020 season as ghost games (N = 1,006). A country with their respective professional leagues was included if it is ranked under the top ten in the UEFA country coefficients at the end of season 2018/2019 https://www.uefa.com/nationalassociations/uefarankings/country/seasons/#/yr/2019. Only the divisions in France, Belgium, Russia, and the Netherlands were excluded due to early league termination in the 2019/2020 season or diverging spectator attendance decisions. Games for professional European divisions in Spain (1st and 2nd Division), England (1st and 2nd Division), Italy (1st and 2nd Division), Germany (1st and 2nd Division), Portugal, and Turkey are included. Eight dependent variables that influence the match outcome are extracted from different categories. These are disciplinary sanctions (fouls, yellow cards, red cards), match dominance (shots, shots on target), market expectation (betting odds), and match results (goals, points).

2.1.2 Additional data

Data for the 2020/2021 and 2021/2022 seasons are considered for the same divisions and variables as described in 2.1.1. Games in the 2020/2021 season are analyzed as ghost games (N = 3,872) and games in the 2021/2022 season as games with attendance (N = 3,832). Additionally, the number of spectators attending games from the 2010/2011 to 2021/2022 seasons for the English Premier League is retrieved from https://www.footballwebpages.co.uk/premier-league/attendances. The Premier League attendance data is used as an exploratory analysis to check whether the number of spectators yields different results compared to a binary classification of matches in attendance and non-attendance.

2.2 Data Processing

As described in Wunderlich et al. (2021) [11] dependent variables are processed to represent the difference between the home team and the away team e.g. Home Goals - Away Goals. This is a common way to assess differences between teams in a single value as positive values represent a higher value for the home team and negative values a higher value for the away team [11]. Betting odds needed to be transformed to a forecasted probability to exclude the bookmaker margin that is generally incorporated in the betting odds [13]. Eight dependent variables are used from four different categories disciplinary sanctions (fouls, yellow cards, red cards), match dominance (shots, shots on target), market expectation (betting odds), and match results (goals, points). Spectator attendance for the replication study is coded as a binary variable with spectator attendance coded as 1 and non-attendance as 0. For the exploratory analysis, the number of spectators per game was converted to a ratio of spectator attendance to stadium capacity to mitigate the problem of the best teams having the highest absolute spectator attendance. Season is coded as -12 to 0 for seasons 2010/2011 to 2021/2022.

2.3 Multilevel model building

2.3.1 Replication study

To account for differences between the leagues Wunderlich et al. (2021) [11] apply a multilevel structure. The used model includes a random intercept for the different leagues and fixed effects for the season and the binary variable of spectator attendance. The authors argue that controlling for seasonal effects as a random effect is not intended as the home advantage has a long-term steady decrease over the seasons and also shifts between seasons. These shifts might be due to new composition of teams or rule changes, like the introduction of the video assistant referee [6]. To mitigate changes due to seasonal changes Wunderlich et al. (2021) [11] propose an additional model that only includes data from the season 2019/20. Therefore, resulting in the following two multilevel equations for one of the eight dependent variables:

$$\begin{aligned} Points_{ij} = \gamma_{00} + \gamma_{10} Season_{ij} + \gamma_{20} Spectator_{ij} + \upsilon_{1j} Season_{ij} \\ + \upsilon_{2j} Spectator_{ij} + \upsilon_{0j} + \epsilon_{ij}, \end{aligned} \tag{1}$$

$$Points_{ij} = \gamma_{00} + \gamma_{10} Spectator_{ij} + \upsilon_{1j} Spectator_{ij} + \upsilon_{0j} + \epsilon_{ij}, \tag{2}$$

where $Points_{ij}$ is the difference in points obtained between the home and away team for match i in division j. *Season* is a variable ranging from -9 to 0, and *Spectator* indicates spectator attendance. ϵ is the first level residual variance and υ is the second level residual variance.

2.3.2 Extended replication study

For the extended models, an additional hierarchical step for the home team is applied. This step accounts for differences between the home teams within their specific division [10]. This extension is chosen due to a substantial increase in the intra-class coefficient (ICC) through all outcome variables. The ICC indicates the variation in the outcome variables that are accounted for by the clustering structure of the data. This procedure provides evidence of whether a multilevel model is necessary. In the social sciences, an ICC of 0.10 generally denotes the necessity to perform multilevel analyses [12]. The average ICC in the initial model is $ICC = 0.003$, whereas the model including the home team as a hierarchical step has an average ICC of $ICC = 0.116$. Only using the home team as a random effect lead to an average $ICC = 0.88$. In order to keep the original level, the three-level model is preferred instead of switching the second level to the home team. Including random slopes or interaction effects between the first-level predictors did not yield improved models.

Therefore, resulting in the multi-level equation for one of the eight dependent variables:

$$\begin{aligned} Points_{ijk} = \gamma_{000} + (\gamma_{100} + \upsilon_{10k} + \nu_{1jk}) * Spectator_{ijk} \\ +(\gamma_{200} + \upsilon_{20k} + \nu_{2jk}) * Season_{ijk} \\ +\epsilon_{ijk} + \upsilon_{0jk} + \nu_{00k}, \end{aligned} \tag{3}$$

where $Points_{ijk}$ is the difference in points obtained between the home and away team for match i with a specific home team j, in a specific division k. *Season* is a variable ranging from -11 to 0 and *Spectator* indicates spectator attendance. ϵ is the first level residual variance and υ is the second level residual variance.

2.3.3 Exploratory analyses

For the exploratory model, the home team is included as the second level as only one division is examined. Therefore, resulting in the single multi-level equation for one of the eight dependent variables:

$$\begin{aligned} Points_{ij} = \gamma_{00} + \gamma_{10} Season_{ij} + \gamma_{20} Occupancy_{ij} \\ +\upsilon_{1j} Season_{ij} + \upsilon_{2j} Occupancy_{ij} + \upsilon_{0j} + \epsilon_{ij}, \end{aligned} \tag{4}$$

where $Points_{ijk}$ is the difference in points obtained between the home and away team for match i with a specific home team j. *Season* is a variable ranging from -11 to 0 and *Occupancy* is the rate of spectators to the stadium capacity. ϵ is the first level residual variance and υ is the second level residual variance.

2.4 Statistical analyses

All analyses were conducted with R-studio version 4.1.3 [14]. For data manipulation, *tidyverse* package is used and *lme4* package is used for multi-level

model. Likelihood ratio tests are used to check whether a model with a new input e.g. random intercept, becomes significantly worse compared to the most recent model. The significance level was set at $p<0.05$.

3 Results

3.1 Replication study

The model from equation 1 for the ten seasons yields the same results as the results obtained by Wunderlich et al. (2021) [11] and are depicted in Table 1. To compare the results with the initial study, see Wunderlich (2021) [11] Table 2 and 3 (pp.6). The replication results of the one-season example are shown in Appendix A.

The results indicate that *Spectator* has a significantly positive effect on the dependent variables expected points, shots, and shots on target. Meaning that the home team has a higher value for e.g. shots on target compared to the away team. *Spectator* has also a significantly negative coefficient for fouls, yellow cards, and red cards. Meaning that the home team has a lower value for e.g. the number of yellow cards given by the referee compared to the away team. For the two dependent variables, goals, and points no significant results for *Spectator* are found. *Season* has a significantly negative effect on expected points, shots, and shots on target indicating that the home advantage decreased over the past seasons. The visual trend over the seasons is shown in Appendix B. The variance estimates of the division-level residual errors, symbolized by σ_u^2, range from 0.00 to 0.11, indicating only marginal differences due to the divisions.

3.2 Extended replication study

The results for the model from equation 3 are depicted in Table 2. In comparison to the results of Table 1, the extended results yield different results for some variables. For the dependent variables, expected points, shots, and shots on target the significant positive influence of *Spectator* remains. In addition, *Spectator* has now a significantly positive effect on the two dependent variables' goals and points. Moreover, *Season* shows a significantly negative trend for all tested variables. *Spectator* has still a significant negative effect on fouls, yellow cards, and red cards and *Season* is now significant and shows a positive trend over time. *Season* remains not significant for yellow cards and red cards. The variance estimates of the home-team-level residual errors, symbolized by σ_v^2, range from 0.00 to 6.37, indicating that there are more differences due to different home teams. Overall, these results indicate that spectator attendance now influences all tested dependent variables in a direction that the advantage for the home team decreases.

This trend is also depicted in Figure 1, where the mean difference between the home and away teams is shown for all dependent variables. The start of the

ghost game series is in the second part of the 10th season $(10 - N)$ and the end after the 11th season. The 12th season shows the average for the first season with readmitted attendance.

3.3 Exploratory analyses

An exploratory analysis including a new independent variable was conducted that measures *Occupancy* instead of the binary *Spectator* variable. Data is used for the English Premier League for the 2010/2011 to 2021/2022 seasons. The results are depicted in Table 3. Compared to the replicated and extension results slight changes are observed. The effect of *Occupancy* is significant for the dependent variables goals, points, expected points, shots, and fouls. The effect on shots on targets, yellow cards, and red cards are not significant anymore. *Season* has a significant negative effect on goals, points, expected points, shots, and shots on target and a significant positive effect on yellow cards. *Season* is not significant for fouls and red cards. The variance estimates of the home-team-level residual errors, symbolized by σ_u^2, range from 0.00 to 11.32, indicating differences due to the home teams. The visual trend over the seasons is shown in Appendix C.

Figure 1: Mean of all dependent variables for each season

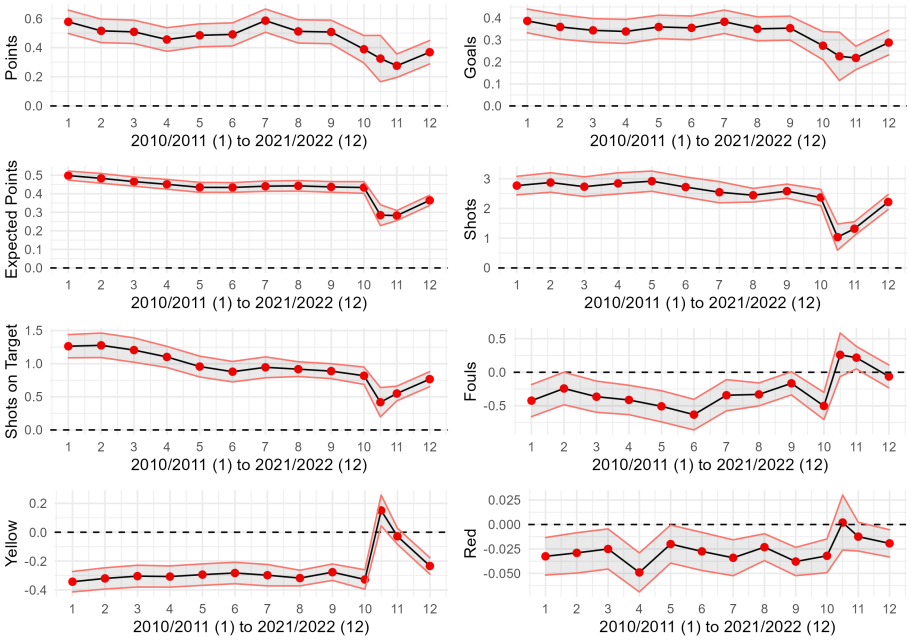


Table 1: Replication results over ten seasons for all dependent variables

Difference H/A	Goals	Points	Expected points	Shots
	(1)	(2)	(3)	(4)
SpectatorsYes	0.108	0.152	0.140***	1.406***
	(0.056)	(0.083)	(0.028)	(0.251)
Season	−0.004	−0.007	−0.006***	−0.045*
	(0.003)	(0.005)	(0.002)	(0.017)
Constant	0.224***	0.322***	0.280***	1.034***
	(0.055)	(0.081)	(0.029)	(0.243)
σ_u^2	0.00	0.00	0.00	0.02
σ_e^2	2.90	6.33	0.70	56.65
Observations	37,888	37,888	37,888	25,270
Akaike Inf. Crit.	147,903.700	177,432.700	94,139.760	173,741.500
Bayesian Inf. Crit.	147,946.400	177,475.500	94,182.470	173,782.200

Difference H/A	Shots Target	Fouls	Yellow Cards	Red Cards
	(5)	(6)	(7)	(8)
SpectatorsYes	0.384**	−0.589***	−0.447***	−0.033*
	(0.123)	(0.177)	(0.057)	(0.015)
Season	−0.051***	−0.009	0.003	−0.0001
	(0.008)	(0.013)	(0.004)	(0.001)
Constant	0.418***	0.251	0.151**	0.002
	(0.118)	(0.198)	(0.055)	(0.014)
σ_u^2	0.00	0.11	0.00	0.00
σ_e^2	13.62	28.31	2.94	0.20
Observations	25,270	25,270	25,270	25,270
Akaike Inf. Crit.	137,729.600	156,230.800	98,962.080	31,171.130
Bayesian Inf. Crit.	137,770.300	156,271.500	99,002.770	31,211.810

Note: *p<0.05; **p<0.01; ***p<0.001;
σ_e^2 = Variance of first level residual error;
σ_u^2 = Variance of second level residual error

Table 2: Extended results over twelve seasons for all dependent variables

Difference H/A	Goals	Points	Expected points	Shots
	(1)	(2)	(3)	(4)
SpectatorsYes	0.092**	0.145***	0.105***	0.996***
	(0.028)	(0.042)	(0.012)	(0.125)
Season	−0.012***	−0.019***	−0.012***	−0.084***
	(0.003)	(0.004)	(0.001)	(0.015)
Constant	0.061	0.071	0.125**	0.625**
	(0.044)	(0.064)	(0.048)	(0.199)
σ_v^2	0.25	0.39	0.25	6.37
σ_u^2	0.01	0.02	0.02	0.09
σ_e^2	2.67	5.98	0.44	50.05
Observations	45,592	45,592	45,589	32,973
Akaike Inf. Crit.	175,114.400	211,776.900	94,092.890	223,438.800
Bayesian Inf. Crit.	175,166.800	211,829.300	94,145.260	223,489.300

Difference H/A	Shots Target	Fouls	Yellow Cards	Red Cards
	(5)	(6)	(7)	(8)
SpectatorsYes	0.245***	−0.402***	−0.279***	−0.018*
	(0.061)	(0.091)	(0.030)	(0.008)
Season	−0.063***	0.028**	0.007	0.001
	(0.007)	(0.011)	(0.003)	(0.001)
Constant	0.145	0.255	0.041	−0.008
	(0.101)	(0.136)	(0.032)	(0.007)
σ_v^2	1.39	1.29	0.07	0.00
σ_u^2	0.03	0.08	0.00	0.00
σ_e^2	11.85	27.29	2.93	0.20
Observations	32,973	32,973	32,973	32,973
Akaike Inf. Crit.	175,913.300	203,158.900	129,429.800	40,881.940
Bayesian Inf. Crit.	175,963.700	203,209.300	129,480.200	40,932.360

Note: *p<0.05; **p<0.01; ***p<0.001;
σ_e^2 = Variance of first level residual error;
σ_u^2 = Variance of second level residual error;
σ_v^2 = Variance of third level residual error

Table 3: Exploratory results over twelve Premier League seasons for all dependent variables

Difference H/A	Goals	Points	Expected points	Shots
	(1)	(2)	(3)	(4)
Occupancy	0.267**	0.447**	0.146**	1.500***
	(0.098)	(0.138)	(0.047)	(0.446)
Season_c	−0.021*	−0.028*	−0.022***	−0.118**
	(0.009)	(0.013)	(0.004)	(0.042)
Constant	−0.276*	−0.419*	−0.117	−0.649
	(0.127)	(0.165)	(0.108)	(0.670)
σ_u^2	0.34	0.50	0.38	11.32
σ_e^2	3.01	6.01	0.68	62.53
Observations	4,558	4,558	4,558	4,558
Log Likelihood	-9,023.024	-10,594.960	-5,670.078	-15,947.010
Akaike Inf. Crit.	18,056.050	21,199.910	11,350.160	31,904.010
Bayesian Inf. Crit.	18,088.170	21,232.040	11,382.280	31,936.140

Difference H/A	Shots Target	Fouls	Yellow Cards	Red Cards
	(5)	(6)	(7)	(8)
Occupancy	0.291	−0.893***	−0.101	−0.014
	(0.219)	(0.256)	(0.087)	(0.020)
Season_c	−0.150***	0.005	0.029***	0.001
	(0.021)	(0.024)	(0.008)	(0.002)
Constant	−0.798*	0.545*	0.072	−0.004
	(0.336)	(0.272)	(0.085)	(0.017)
σ_u^2	2.90	0.96	0.06	0.00
σ_e^2	15.14	20.93	2.44	0.13
Observations	4,558	4,558	4,558	4,558
Log Likelihood	-12,715.650	-13,430.670	-8,527.484	-1,894.340
Akaike Inf. Crit.	25,441.290	26,871.350	17,064.970	3,798.679
Bayesian Inf. Crit.	25,473.410	26,903.470	17,097.090	3,830.803

Note: *p<0.05; **p<0.01; ***p<0.001;
σ_e^2 = Variance of first level residual error;
σ_u^2 = Variance of second level residual error

4 Discussion

This report investigates the impact of spectator attendance on the home advantage in professional European football leagues by replicating and extending a peer-reviewed study that analyzed a series of ghost games played during the COVID-19 pandemic from March 2020 in a natural experiment. This report successfully replicated the study by Wunderlich et al. (2021) [11] and extended the original study with additional modeling and the inclusion of more recent data. The main finding of this report is that the home advantage decreases in the absence of an audience. Therefore, successfully validating the initial effect investigated by the replicated study. This can be observed in the categories of disciplinary sanctions (fouls, yellow cards, red cards), match dominance (shots, shots on target), market expectation (betting odds), and also match results (goals, points). In particular, the findings on match results are of key interest, as they contradict the results of Wunderlich et al. (2021) [11]. Their findings state that home teams receive fewer disciplinary sanctions and are able to create more offensive actions compared to the away teams. However, these differences between home and away teams do not translate into a home advantage in terms of more points won or more goals scored.

A possible reason for these discrepant results is that the replicated study had highly unbalanced data and used a small set of actual ghost games played (N = 1,006) compared to games with spectators (N = 36,882). The extension included more games in the absence of spectators (N = 4,878), which supports balancing the classes (N = 40,714). In addition, the extension includes the first season after the COVID-19 pandemic with readmitted spectators. This provides an opportunity to see if home advantage increases again when spectators return to the stadium. From the visual inspection of Figure 1 it can be concluded that the home advantage indeed increases again for all dependent variables. However, a decreasing home advantage over the seasons is observed for most dependent variables as can be seen in Table 2. An alternative explanation could be the model changes made to account for the hierarchical structure of the data. The ICC must be considered when using random effects as it indicates the variation in the outcome variables accounted for by the cluster structure of the data. The ICC of the three-level hierarchical structure of matches played by a given home team, in a given division, warranted the use of random effects. However, the hierarchical structure of matches in a given division applied in the replicated study did not justify the need for a hierarchical structure based on the ICC.

This report also focuses on additional exploratory analyses. This approach is based on the idea that the binary classification of games with spectators and games without spectators leads to a significant loss of information. The inclusion of the occupancy rate is intended to overcome the loss of information, as it is a continuous variable that has varying values even within seasons. This approach is of key interest for the 2021/2022 season, as capacity was limited to a fixed spectator size for several games. However, the exploratory analysis was based exclusively on Premier League data, as data for other divisions is scarce.

Potential future research with this data set could examine whether book-

makers set their betting odds correctly during the COVID-19 pandemic, as this report showed a decline in home team advantage that should have been accounted for. Considering the differences in points per game as the dependent variable, expected points as the independent variable would be expected to be around the value of 1, as an increase in expected points should exactly match the actual increase in points. Coefficients greater or less than one would indicate deviations from efficiency.

A limitation of this study is the absence of potential confounding variables. In particular, with observational data, it is difficult to control for variables of interest since the data is not collected in an experimental setting. Therefore, many other variables could confound the results and should be considered in the analysis. Several studies that investigated related effects considered variables such as a team's market value, points scored in the four games prior to the game in question, and a team's standings [1, 4, 8, 10]. All of these variables appear to have a significant impact on the outcome of the game, and to establish a causal relationship between spectator attendance and home team advantage, all confounding variables must be accounted for. Obtaining this type of information often requires additional data collection, which is costly in terms of time and money. Nevertheless, to achieve the goal of causality, these additional steps should be considered.

5 Conclusion

In this report, an existing peer-reviewed study examining a series of ghost games during the COVID-19 pandemic in a natural experiment was successfully replicated and extended. New insights were gained regarding the impact of spectator absence on actual game outcomes, which contradicted the results of the replicated study. Additional findings resulted from the inclusion of new variables in the model, as well as the inclusion of the 2021/2022 season to show the impact after the COVID-19 pandemic when spectators were readmitted. This report also discusses exploratory analyses and the advantages and disadvantages of using an open-source observational data set.

6 Reflection on the research report

6.1 Additional Project

The second project was to determine the current state of methodological and statistical procedures for studies using observational data with large sample sizes in the field of sports science.

Since a full reflection and report on this second project is beyond the scope of this traineeship report, I will provide a brief summary of the literature review process and draw conclusions based on the first project.

6.2 Project summary

The goal of this research project is to identify studies in the Journal of Sports Sciences that analyze observational data with large sample sizes and to evaluate their statistical and methodological procedures.

The focus on observational data with large sample sizes was chosen because of the rapidly growing trend toward open-source data. Open-source data provides a broad audience with the opportunity to easily apply various statistical and methodological procedures to answer upcoming research questions. This can lead to problems as the characteristics of observational data can quickly lead to violations of research practice. Therefore, this report aims to assess the current state of sports science, raise awareness, and suggest a workflow on how to handle this type of data.

Eligible studies were identified for the years 2018 to 2022 in the Journal of Sports Sciences. Manuscripts were examined in more detail and study information (year, authorship, title, sports discipline) and specific methodological characteristics were extracted (number of observations, statistical software, statistical tests, number of statistical tests, number of outcome variables, alpha levels, and effect sizes).

The results of this brief literature review showed that methodological problems occurred in various sports such as football, swimming, basketball, biathlon, judo, and golf. In particular, similar methodological flaws occurred repeatedly in the identified studies.
Following is a brief summary of the main identified issues:

- Large number of outcome variables tested [15–18]
- Missing explanation and theory-based inclusion of outcome variables [15, 16, 19, 20]
- Large number of applied tests [15–23]
- Application of tests not suited for very large sample sizes [17, 19, 20, 24–27]
- Report of significant results only [16, 22, 23, 25]

- Missing data without imputation and causal inference [16, 20]
- Missing adjustment of significance levels in case of many applied tests [15–19, 21, 23–25, 27]
- Missing differentiation between confirmatory and exploratory research [15–17, 19, 20]
- Missing report and reflection on effect sizes [15–18, 21–23, 26]
- Missing report and reflection on confounding variables [15–17, 21, 23, 24, 26]

Since many of these methodological errors occurred repeatedly, lack of statistical and methodological training could be one of the most important explanations. The next step after identifying these methodological errors is to propose a workflow on how to handle this type of open-source observational data and to avoid questionable research practices.

6.3 Implications between projects

The theoretical approach of the second project is to identify weaknesses in the statistical process of sports science studies and precisely targets some of the weaknesses of the first project. In the first project, the study uses open-source observational data on a large scale and attempts to draw causal inferences based on these results. To this end, eight different outcome variables are tested for two different models, resulting in a total of 16 different statistical tests. In this case, the use of the outcome variables is based on theory and should be warranted. However, consideration should have been given to including fewer outcome variables, focusing on a single model, or making adjustments to the significance level, such as using the Bonferroni correction. The use of multilevel regression models appears to be appropriate for this type of data since it allows for the inclusion of potentially confounding variables, and all tests used were reported accordingly. Effect sizes were also neglected in the replicated study. However, the inclusion of the regression results and also the visualization of the effects seem appropriate to report and interpret the results. Nevertheless, the inclusion of R^2 or other relevant effect sizes would have been useful. Finally, one of the major drawbacks of the replicated and extended study is the lack of inclusion of confounding variables. As pointed out in the first part of the report, several other studies have included several related confounding variables, whereas this study is limited to the inclusion of a small number of covariates. In this case, the inclusion of new confounding variables would have been costly in terms of time and money, leading to the decision to exclude additional variables. This still sheds light on this general issue of observational data and the inclusion of covariates that are difficult to obtain. In order to draw causal conclusions, it would be desirable if these additional efforts had been made to obtain more valid and reliable results.

In summary, the replicated study and its extension essentially follow the proposed workflow based on their statistical procedure. However, they have some methodological errors in terms of the number of statistical tests applied and the treatment of covariates. With this knowledge of common statistical errors, future research based on the first part of this report can take this into account and attempt to mitigate these problems.

References

[1] Jeffrey Cross and Richard Uhrig. "Do fans impact sports outcomes? A COVID-19 natural experiment". In: *Journal of Sports Economics* (2020), p. 15270025221100204.

[2] Giovanni Angelini and Luca De Angelis. "Efficiency of online football betting markets". In: *International Journal of Forecasting* 35.2 (2019), pp. 712–721.

[3] Richard Pollard and Gregory Pollard. "Long-term trends in home advantage in professional team sports in North America and England (1876–2003)". In: *Journal of sports sciences* 23.4 (2005), pp. 337–350.

[4] Michela Ponzo and Vincenzo Scoppa. "Does the home advantage depend on crowd support? Evidence from same-stadium derbies". In: *Journal of Sports Economics* 19.4 (2018), pp. 562–582.

[5] Thomas Dohmen and Jan Sauermann. "Referee bias". In: *Journal of Economic Surveys* 30.4 (2016), pp. 679–695.

[6] Michael Christian Leitner and Fabio Richlan. "No fans–no pressure: Referees in professional football during the COVID-19 pandemic". In: *Frontiers in Sports and Active Living* (2021), p. 221.

[7] Carlos Cueva. "Animal Spirits in the Beautiful Game. Testing social pressure in professional football during the COVID-19 lockdown". In: (2020).

[8] Álvaro Jiménez Sánchez and José M Lavın. "Home advantage in European soccer without crowd". In: *Soccer & Society* 22.1-2 (2021), pp. 152–165.

[9] Michael Leitner et al. "The cauldron has cooled down: a systematic literature review on home advantage in football during the COVID-19 pandemic from a socio-economic and psychological perspective". In: *Management Review Quarterly* (2022), pp. 1–29.

[10] Dane McCarrick et al. "Home advantage during the COVID-19 pandemic: Analyses of European football leagues". In: *Psychology of sport and exercise* 56 (2021), p. 102013.

[11] Fabian Wunderlich et al. "How does spectator presence affect football? Home advantage remains in European top-class football matches played without spectators during the COVID-19 pandemic". In: *Plos one* 16.3 (2021).

[12] Joop J Hox, Mirjam Moerbeek, and Rens Van de Schoot. *Multilevel analysis: Techniques and applications*. Routledge, 2017.

[13] Fabian Wunderlich and Daniel Memmert. "The betting odds rating system: Using soccer forecasts to forecast soccer". In: *PloS one* 13.6 (2018), e0198668.

[14] RStudio Team. "RStudio: Integrated development environment for R. RStudio, PBC". In: *Boston, MA. Accessed* (2020), pp. 06–07.

[15] Glenn Björklund et al. “The balancing act between skiing and shooting–the determinants of success in biathlon pursuit and mass start events”. In: *Journal of Sports Sciences* 40.1 (2022), pp. 96–103.

[16] Bryan Le Toquin et al. “Is the visual impairment origin a performance factor? Analysis of international-level para swimmers and para athletes”. In: *Journal of Sports Sciences* 40.5 (2022), pp. 489–497.

[17] Enrique Ortega-Toro et al. “Effect of scaling basket height for young basketball players during the competition: seeking out positive sport experiences”. In: *Journal of Sports Sciences* 39.24 (2021), pp. 2763–2771.

[18] Emerson Franchini, David H Fukuda, and Joao Paulo Lopes-Silva. “Tracking 25 years of judo results from the World Championships and Olympic Games: Age and competitive achievement”. In: *Journal of Sports Sciences* 38.13 (2020), pp. 1531–1538.

[19] Aaron Koenigsberg, Jarred Pilgrim, and Joseph Baker. “Generational differences in the ranking pathways of top 100 ranked golfers”. In: *Journal of Sports Sciences* 38.18 (2020), pp. 2047–2053.

[20] Quentin De Larochelambert et al. “Relative age effect in French alpine skiing: Problem and solution”. In: *Journal of Sports Sciences* 40.10 (2022), pp. 1137–1148.

[21] Sebastian Zart and Arne Güllich. “In-season head-coach changes have positive short-and long-term effects on team performance in men’s soccer—evidence from the Premier League, Bundesliga, and La Liga”. In: *Journal of Sports Sciences* 40.6 (2022), pp. 696–703.

[22] Colin W Fuller and Aileen Taylor. “Ten-season epidemiological study of match injuries in men’s international rugby sevens”. In: *Journal of sports sciences* 38.14 (2020), pp. 1595–1604.

[23] Pantelis T Nikolaidis and Beat Knechtle. “Russians are the fastest and the youngest in the “Comrades Marathon””. In: *Journal of Sports Sciences* 37.12 (2019), pp. 1387–1392.

[24] Mat Herold et al. “Off-ball behavior in association football: A data-driven model to measure changes in individual defensive pressure”. In: *Journal of Sports Sciences* (2022), pp. 1–14.

[25] Daniel A Hackett et al. “Effects of age and sex on field-based measures of muscle strength and power of the upper and lower body in adolescents”. In: *Journal of Sports Sciences* 39.9 (2021), pp. 955–960.

[26] Sam McIntosh, Karl B Jackson, and Sam Robertson. “Apples and oranges? Comparing player performances between the Australian Football League and second-tier leagues”. In: *Journal of Sports Sciences* 39.18 (2021), pp. 2123–2132.

[27] Robin C Jackson and Gavin Comber. “Hill on a mountaintop: A longitudinal and cross-sectional analysis of the relative age effect in competitive youth football”. In: *Journal of sports sciences* 38.11-12 (2020), pp. 1352–1358.

Appendices

A

Table 4: Replication results for one season for all dependent variables

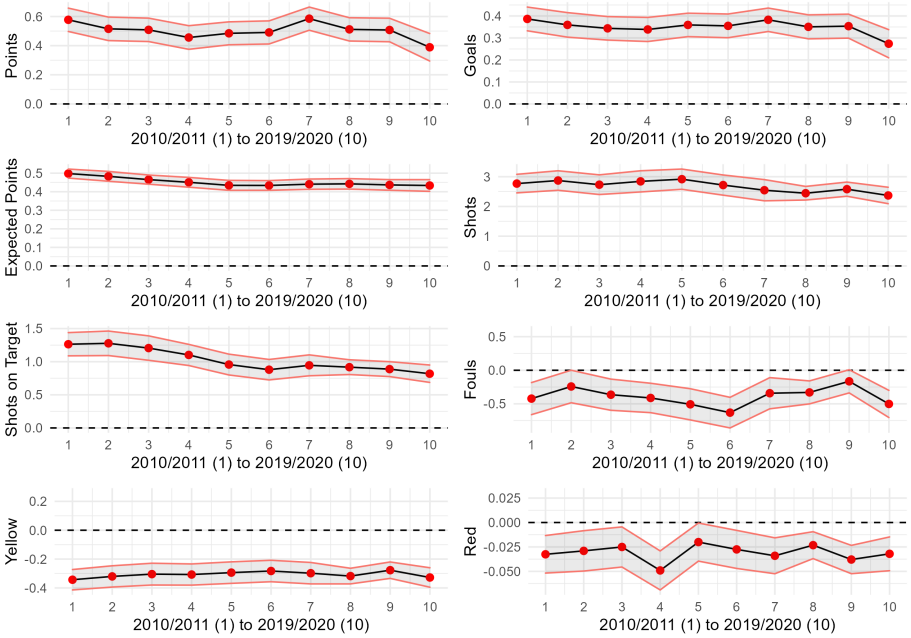
Difference H/A	Goals	Points	Expected points	Shots
	(1)	(2)	(3)	(4)
SpectatorsYes	0.048	0.064	0.149***	1.333***
	(0.064)	(0.093)	(0.032)	(0.270)
Constant	0.226***	0.325***	0.284***	1.037***
	(0.054)	(0.082)	(0.028)	(0.231)
σ_u^2	0.00	0.00	0.00	0.00
σ_e^2	2.99	6.42	0.74	53.49
Observations	3,752	3,752	3,752	3,752
Log Likelihood	-7,375.806	-8,812.464	-4,767.040	-12,789.320
Akaike Inf. Crit.	14,759.610	17,632.930	9,542.081	25,586.650
Bayesian Inf. Crit.	14,784.530	17,657.850	9,567.001	25,611.570

Difference H/A	Shots Target	Fouls	Yellow Cards	Red Cards
	(5)	(6)	(7)	(8)
SpectatorsYes	0.400**	−0.745***	−0.479***	−0.034*
	(0.130)	(0.198)	(0.066)	(0.017)
Constant	0.418***	0.261	0.151**	0.002
	(0.111)	(0.179)	(0.056)	(0.014)
σ_u^2	0.00	0.03	0.00	0.00
σ_e^2	12.43	28.73	3.17	0.21
Observations	3,752	3,752	3,752	3,752
Log Likelihood	-10,052.040	-11,625.370	-7,489.554	-2,408.555
Akaike Inf. Crit.	20,112.080	23,258.740	14,987.110	4,825.109
Bayesian Inf. Crit.	20,136.990	23,283.670	15,012.030	4,850.030

Note: *p<0.05; **p<0.01; ***p<0.001;
σ_e^2 = Variance of first level residual error;
σ_u^2 = Variance of second level residual error

B

Figure 2: Replicated - Mean of all dependent variables for each season



C

Figure 3: Exploratory - Mean of all dependent variables for each season