Home advantage of European major football leagues under COVID-19 pandemic

Eiji Konaka

Meijo University
konaka@meijo-u.ac.jp

January 5, 2021

Abstract

After March 2020, the environment surrounding football has changed dramatically due to COVID-19 pandemic. After a few months' break, the re-scheduled matches have been held behind the closed door without spectators. The main objective of this study is a quantitative evaluation of "crowd effects" on home advantage, using the match results of the closed matches. The proposed analysis uses pairwise comparison method in order to reduce the effects caused by unbalanced schedule. The following conclusion can be drawn from some statistical hypothesis tests: In four major European leagues, the home advantage is reduced in closed matches compared to normal situation with spectators. The reduction amount among leagues is different. In Germany, it shows the negative home advantage during the closed period. On the other hand, in England, statistically significant difference cannot be found between closed and normal situations.

I. Introduction

After March 2020, the environment surrounding football has changed dramatically — due to COVID-19 pandemic. Most of football leagues had been forced to be suspended, similar to many other crowd-pleasing events. Some of the leagues had been calcelled, while others resumed after a few months' break. The resumed leagues, however, no decision was made to bring the spactators as before the suspension. The re-scheduled matches has been held behind the closed door without spactators.

The objective of this paper is to analyse how this unfortunate and unacceptable closed situation affected the match outcome of football. In particular, our main objective is a quantitative evaluation of "crowd effects" on home advantage.

Today, it seems that the existence of "home advantage" in sports, especially football, cannot be questioned[1][2][3]. However, there is still no definitive evidence on the factors that produce home advantage. Of course, the quantitative impact of each factor on home advantage has also been discussed.

"Home advantage" can be defined as that home teams consistently win more than 50 percent of the games under a blanced home-and-away schedule[4]. Quantitative research on home advantage in football dates back nearly 40 years ago to Morris[5], then, Dowie[6] and Pollard[2] followed. In a very brief and thorough review paper on home advantage[7], consisting of only three pages, Pollard presented eight main factors that could generate home advantage and explained conventional studies for each (of course, he did not forget to mention those interactions

and other factors). The first one is "'crowd effect", i.e, effect caused by spactators. In [7], he wrote that "This is the most obvious factor involved with home advantage and one that fans certainly believe to be dominant[8][9]." Studies by Dowie[6] and Pollard[2] could find very little evidence that home advantage depends on crowd density (spectators per stadium capacity).

This paper utilizes the results of matches behind closed door held under the influence of the COVID-19 pandemic to determine the relationship between the presence of spectators and home advantage. Reade et al. used the results of the matches behind closed door since 2003 and reported a study on crowd effect[10]. The number of the matches in empty studium, however, is small (160 out of approx. 34 thousand matches). In addition, in these closed matches, teams have typically been banned from admitting supporters into their stadiums as punishments for bad behaviour off the football pitch (e.g., due to corruption, racist abuse, or violence). Therefore, in this analysis, spactators' characteristics could be biased, e.g., they were excessively violent or doing overly aggressive supporting behavior.

When making use of the results of these closed matches, it should also be noted that the schedule is unbalanced. Many European league matches are about two-thirds completeed by mid-March suspension. It should be considered the possible bias for the strength of home teams in the reseumed matches after the break. Most of famous and extensive previous studies on home advantage[11] and the latest report [12] used only several basic statistics, e.g., numbe of goals, fouls, and wins. It should be pointed out that they could be biased under unbalanced schedule.

For the problem on biased scheldule, in this paper, assume a statistical model the explains match results from the team strength parameters for each team and the home advantage parameter that is common for every team in the league. By fitting the paramters in this model to minimize explanation error, separate the home advantage even from the unbalanced schedule.

The specific techniques are as follows: each team i has one strength evaluation value r_i called "rating". And one league has one parameter $r_{homeAdv}$ that expresses home advantage. The strength difference Δr is expressed by the difference of rating values between the teams indexed by i and j added by the home advantage, i.e., $\Delta r = r_i + r_{homeAdv} - r_j$. The strength difference explains the score ratio in a match via logistic regression model, i.e., $1/(1 + \exp(-\Delta r))$.

The rating values r_i and the home advantage value $r_{homeAdv}$ at a particular date is estimated by using themost recent match results, e.g., five matchweeks. This calculation process is repeated for every matchweek. The home advantage value estimated only from the closed matches are compared to those of from the past "normal" match results.

This paper is organized as follows: Section II states on the data and the detaied algorithm used in this study. This paper analysis on the five major top divition of European football leagues – England, France, Germany, Italy, and Spain (in France, the top division has not been resumed after the suspension). The match results are collected from 2010-2011 season. Section III states statistical analysis results. The following conclusion can be drawn from some statistical hypothesis tests.

- In four major European leagues, the home advantage is reduced without spactators comaperd to normal situation with spactators.
- The reduction amount among leagues is different.
- For all four leagues, the home advantage still remains even in closed matches.

Section IV summarizes and concludes this paper.

II. Methods

In this section, at first, describe the leagues studied and the content of the used data. Then, describe a mathematical method for estimating home advantage.

i. Data set

Table 1 shows the leagues and the number of matches used in this study.

Table 1: *Matches used in this study*

Country	League	Teams	Matches (2010/11-2018/19)	Matches (2019/20) Normal	Closed
England	Premier League	20	3420	290	90
France	Ligue 1	20	3420	279	0
Germany	Bundesliga	18	2754	216	90
Italy	Serie A	20	3420	240	140
Spain	LaLiga	20	3420	270	110
Total			16434	1295	430

This study analyses the home advantage in the top divisions in five European countries – England, France, Germany, Italy, and Spain. They are considered as the most major and highest-quality football leagues among the world. The match results from 2010-2011 season are collected from worldfootball.net (`https://www.worldfootball.net/`). The number of maches are 17729, including 430 closed matches.

All five leagues were suspended from mid-March due to COVID-19 pandemic. Four leagues except France resumed until late-June, and finished until early-August. Ligue 1 in France quickly decided and announced its calcellation at the end of April[13].

Table 2 summarizes their closed period.

Table 2: *Closed period*

Country	League	Matchweeks	Closed from	Closed matchweeks
England	Premier League	38	30	9
Germany	Bundesliga	34	25	10
Italy	Serie A	38	25	14
Spain	LaLiga	38	28	11

ii. Mathematical model

We propose a unified and simple statistical estimation method of scoring ratios based on the score in each match, which is always officially recorded and is common to all ball games. This method is an extention of [14] as to incorporate home advantage. The study[14] reported that his propsed method achieved higher prediction accuracy for ten events of five sports e.g., basketball, handball, hockey, volleyball, and water polo, in the Rio Olympic Games compared to the official world rankings.

Assume that the scoring ratio of a home team i in a match against an away team j (i and j are team indices), denoted as $p_{i,j}$, is estimated as

$$p_{i,j} = \frac{1}{1 + e^{-(r_i + r_{homeAdv} - r_j)}}, \tag{1}$$

where r_i is defined as the *rating* of team i and $r_{homeAdv}$ is quantitative value of home advantage.

Given (s_i, s_j), the actual scores in a match between i and j,

$$s_{i,j} = \frac{s_i + 1}{s_i + 1 + s_j + 1} = p_{i,j} + \epsilon_{i,j}, \tag{2}$$

where $s_{i,j}$ and $\epsilon_{i,j}$ are the modified actual scoring ratio and the estimation error, respectively. In football, shut-out results such as $1-0$ or $3-0$ occur frequently. Thus, a simple scoring ratio, i.e., $s_{i,j} = \frac{s_i}{s_i + s_j}$, can result in invalid strength evaluation. Therefore, the score of each team is added by one. This modification is known as Colley's method[15], and was originally used to rank college (American) football teams.

This mathematical structure is the well-known *logistic regression model*. It is widely used in areas such as the winning probability assumption of Elo ratings in chess games[16], and the correct answer probability for questions in item response theory[17].

The update method is designed to minimize the sum of the squared error E^2 between the result and the prediction , defined by the following equation:

$$E^2 = \sum_{(i,j) \in \text{all matches}} (s_{i,j} - p_{i,j})^2. \tag{3}$$

It is straightforward to obtain the following update based on the steepest-descent method:

$$r_i \leftarrow r_i - \alpha \cdot \frac{\partial E^2}{\partial r_i}, \quad r_{homeAdv} \leftarrow r_{homeAdv} - \alpha \cdot \frac{\partial E^2}{\partial r_{homeAdv}}, \tag{4}$$

where α is a constant.

By definition, the rating is an interval scale. Therefore, its origin, $r = 0$, can be selected arbitrarily and a constant value can be added to all r_i. For example,

$$\boldsymbol{r} \leftarrow \boldsymbol{r} - (\max \boldsymbol{r}) \cdot \mathbf{1} \tag{5}$$

implies that $r = 0$ always shows the highest rating, and $r < 0$ shows the distance from the top team.

ii.1 Convert rating on scoring ratio to winning probability

The rating r_i in (1) explains the scoring ratio. Once we have the scoring ratio $p_{i,j}$ given in (1), assume that the following independent Bernoulli process is executed N times, starting from $(s_i, s_j) = (0, 0)$ and with the parameter $0 < \beta \leq 1$.

$$\begin{cases} s_i \leftarrow s_i + 1 & \text{with probability} \quad \beta p_{i,j}, \\ s_j \leftarrow s_j + 1 & \text{with probability} \quad \beta \left(1 - p_{i,j}\right), \\ s_i \leftarrow s_i, s_j \leftarrow s_j & \text{with probability} \quad (1 - \beta). \end{cases} \tag{6}$$

This is a unified (and approximated) model of a scoring process for all ball games, where s_i and s_j model the scores of teams i and j, respectively.

The parameters N and β vary among the sports and between definitions of a unit of play. For example, in basketball, if a unit of play is defined as 10[s], we have $N = 40[\text{min}] \times 60[\text{s}]/10 = 240$. In football, if a unit of play is defined as 1[min], we have $N = 90$. For both sports, β is determined as $\beta = E(s_i + s_j)/N$.

At the end of the match, $s_i > s_j$ shows that team i wins against team j. Figure 1 shows the simulated winning probability for different $N\beta$ and rating gap $(r_i - r_j)$, with $N = 240$. This probability is expressed by the cumulative distribution function for a normal distribution. In many applications, it is common to use a logistic regression model rather than a cumulative distribution[18].

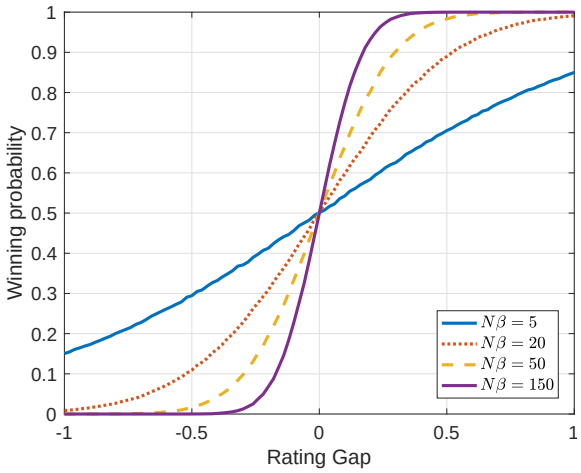

Figure 1: *Rating gap and winning probability*

Based on the discussions above, we convert the rating on the scoring ratio to that of a winning probability, as follows:

$$w_{i,j} = 1 \ (i \text{ wins}), \ \ 0.5 \ (\text{draw}), \ \ \text{or} \ \ 0 \ (j \text{ wins}), \tag{7}$$

which denotes a win, draw, or loss for team i against team j. Find D_k^*, where k is an index of sports, that satisfies

$$\hat{w}_{i,j} = \frac{1}{1 + \exp\left(-D_k\left(r_i + r_{homrAdv} - r_j\right)\right)}, \tag{8}$$

$$D_k^* = \arg\min_{D_k} \sum \left(w_{i,j} - \hat{w}_{i,j}\right)^2. \tag{9}$$

Then, r_i is converted as follows:

$$\bar{r}_i = D_k^* r_i, \ \ i = 1, 2, \cdots, N_T, \text{and } homeAdv, \tag{10}$$

where N_T denotes the number of teams. Therefore, $\bar{r}_{homeAdv}$ is a quantitative home advantage estimation that explains the effect on the winning probability.

In Equations (3) and (9), the sum of squared errors are used as a loss function instead of the cross-entropy. This is because these problems are regression problems, not classification ones.

iii. Short-term estimation of home advantage

The proposed method in the previous section is used to estimate the rating of each team and the home advantage in the league using the results of the last five matchweeks, including itself, for every matchweek.

By using five matchweeks, we can estimate the average of each team's strength and league-wide home advantage over approximately from three weeks to one month.

The calculated home advantage is classified into the following four classes based on the spactator attendence.

- Past: The home advantage calculated using the matches from 2010-2011 to 2018-2019 seasons. It includes closed matches as punishments, if exist.
- Normal: The home advantage calculated using the matches before suspension in 2019-2020. The matches are basically with spactators. Note that closed matches as punishments are inclused here, if exist.
- Mixed: The home advantage calculated using the matches in which with and without spactators are mixed.
- Closed: The home advantage calculated using the matches without spactators.

III. Results and discussions

This chapter describes the analysis results and discussions.

i. Basic stats

Figure 2 depicts the basic stats, e.g., goals difference per match and win ratio difference, in normal and closed periods.

From this counting, home teams goal approximately from 0.3 to 0.5 more per match than away teams in average under the "past and normal" situation. As a result, the home teams win more approximately 0.15 to 0.25 against the away teams. In addition, it seems obvious that the home advantage is reduced under the closed situation. Especially, in Bundesliga, the goals difference and win ratio difference are both negative in the closed matches. In this counting, however, possible schedule unbalance for the closed matches is not considered. In other words, it is possible that the most of the home teams are consistently weak (or strong) in the closed matches.

ii. Estimation of home advantage

Figure 3 shows the estimation results of the home advantage $\bar{r}_{homeAdv}$ in five leagues. The medians are all positive for every four classes. This means that the home advantage remains even if the match is closed and without spactators. The medians in the past and normal periods looks similar. On the other hand, the median in the closed period is smaller than those of the past and the normal periods.

We tested null hypotheses that the home advantage $\bar{r}_{homeAdv}$ from two different categories are samples from continuous distributions with equal medians. Wilcoxon's rank sum test[19, 20] is used as a test method because any assumption on the shape of the distribution of $\bar{r}_{homeAdv}$ can be posed. $p-$values between classes are depicted in Figure 3, and Table 3 summarizes the test result.

From these test results, the followings can be concluded.

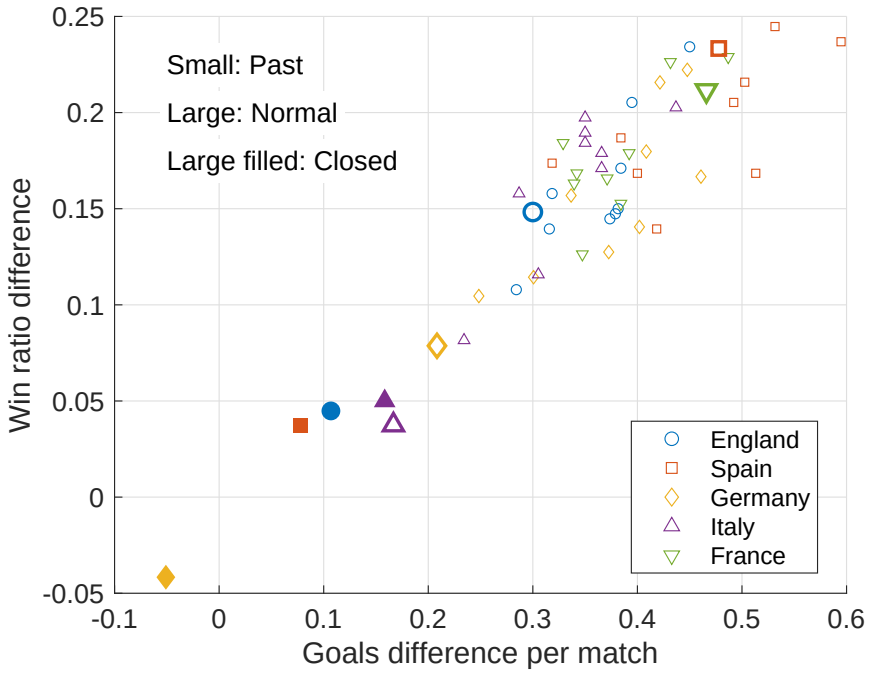

Figure 2: *Goals difference and win probability difference*

Table 3: *Test results*

Sample X	Sample Y	N_X	N_Y	$p-$value	$z-$value	ranksum
Past	Normal	1494	112	1.18×10^{-1}	1.562	1207821
Past	Mixed	1494	16	2.45×10^{-3}	3.029	1133973
Past	Closed	1494	28	8.37×10^{-7}	4.927	1149033
Normal	Mixed	112	16	2.77×10^{-2}	2.201	7530
Normal	Closed	112	28	6.26×10^{-4}	3.420	8553
Mixed	Closed	16	28	4.28×10^{-1}	0.792	393

- There is no significant difference in home advantage between the past and 2019-2020 season's normal matches ($p > 0.1$).
- There is significant difference in home advantage between the normal matches and the closed matches ($p < 10^{-3}$). The significant difference is more obvious between the past and the closed matches ($p < 10^{-6}$). The median of the home advantage in the closed matches is clearly smaller.

Therefore, this study provides strong quantitative evidence of the impact of the crowd effect on home advantage in the top leagues in Europe. It should be noted that, however, all of the closed matches in this season have been with several months' suspension, and held in summer when the matches have not normally been played. It can not be rejected the possible effect by these factors.

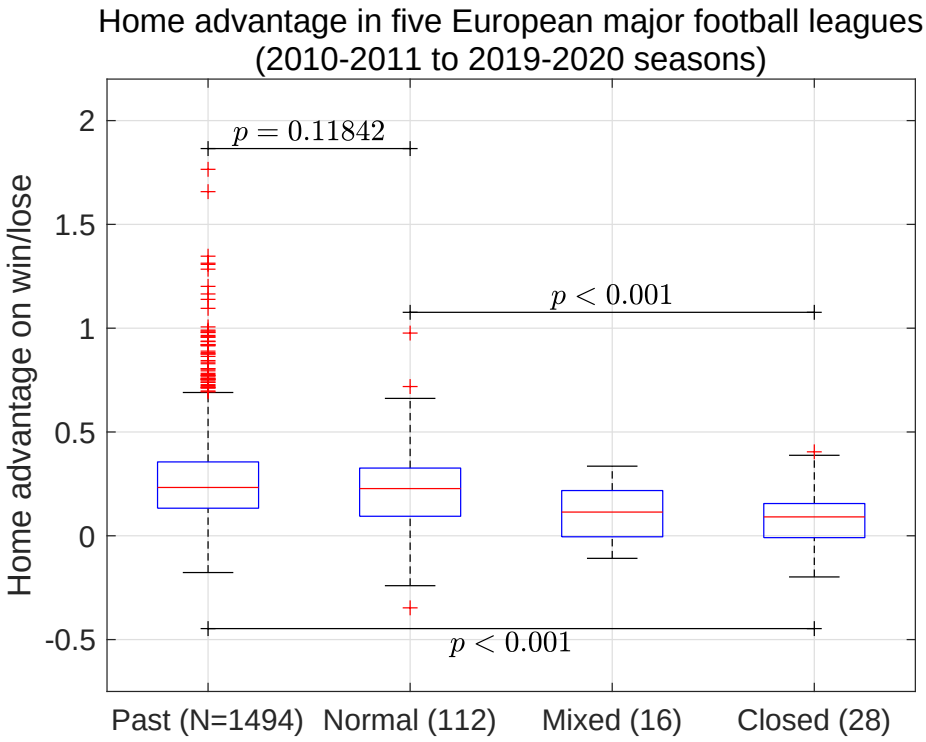

Figure 3: *Distribution of $\overline{r}_{homeAdv}$ in five major European football leagues*

ii.1 Detailed analysis for each league

The last section described the results for all five leagues. In this section, the estimation results will be explained for each league. Table 4 lists the detailed test results of Wilcoxon's rank sum test for each league. Figure 4 show the number of home matches in the closed period. The teams are ordered by the final standings. Figures from 5 to 9 show the estimated value of $\overline{r}_{homeAdv}$ for each league from 2010-11 season.

England In England, the home advantage value in the closed period obtained by the propose method has no significant difference to those of in the past and normal periods. This result shows that the basic stats, such as the goals difference and the win ratio difference shown in Figure 2, are smaller in the closed period because of the unbalanced schedule. In other words, the most of the home teams are consistently weak (or strong) in the closed matches. Figure 4 supports this assertion. In England, there are weak correlation between the number of home matches in the closed period and the final standings. The correlation value is 0.4052. In the other leagues, Germany, Italy, and Spain, the correlation values are much smaller (0.2863, 0.0867, and -0.1561).

France In France, there is not significant difference of $\overline{r}_{homeAdv}$ between past and normal periods.

Germany In Germany, the home advantage value in the closed period has significant difference to those of in the past and normal periods. Both $p-$values are less than 1.00×10^{-2}. In the closed period, Bundesliga showed home "dis"advantage, i.e., $\overline{r}_{homeAdv} < 0$.

Table 4: *Test results: League breakdown*

League	Sample X	Sample Y	N_X	N_Y	$p-$value	$z-$value	ranksum
England	Past	Normal	306	25	1.74×10^{-1}	-1.3596	3524
England	Past	Closed	306	5	4.42×10^{-1}	0.7696	47890
England	Normal	Closed	25	5	8.67×10^{-1}	-0.1669	384
France	Past	Normal	306	24	8.79×10^{-1}	-0.1522	3903
Germany	Past	Normal	270	20	1.24×10^{-1}	-1.5378	2353
Germany	Past	Closed	270	6	4.07×10^{-5}	4.1034	38189
Germany	Normal	Closed	20	6	3.84×10^{-3}	2.8908	318
Italy	Past	Normal	306	20	1.56×10^{-2}	-2.4181	2282
Italy	Past	Closed	306	10	7.48×10^{-2}	1.7815	49008
Italy	Normal	Closed	20	10	5.53×10^{-1}	-0.5939	296
Spain	Past	Normal	306	23	3.87×10^{-2}	2.0673	4705
Spain	Past	Closed	306	7	1.48×10^{-4}	3.7952	48941
Spain	Normal	Closed	23	7	8.76×10^{-5}	3.9227	437

Italy In Italy, the home advantage value in the closed period has significant difference to that of in the past periods $(p < 0.01)$.

Spain In Spain, the home advantage value in the closed period has significant difference to those of in the past and normal periods. Both $p-$values are less than 1.00×10^{-2}. In the closed period, similar to Bundesliga, laLiga showed home disadvantage in the closed period.

IV. Conclusion

In this paper, the results of matches behind closed door held under the influence of the COVID-19 pandemic are used to determine the relationship between the presence of spectators and home advantage. In order to reduce the effect of schedule unbalance in the closed period, this paper proposed a short-term (e.g., five matchweeks) rating method considering home advantage. The proposed method has been applied to the match results in five major European football leagues (England, France, Germany, Italy, and Spain) from 2010-2011 to 2019-2020 seasons.

The distributions of the home advantage for both the past normal and the closed periods are calculated. Their median values are compared by using statistical hypothesis test. A null hypothesis, "the home advantage $\bar{r}_{homeAdv}$ from two different periods are samples from continuous distributions with equal medians" are rejected with sufficiently small $p-$value $(p < 10^{-3})$. The home advantage became smaller behind the closed door.

Our future work is to extend the proposed method to the match results all over the world. This work could clarify the crowd effect to home advantage.

References

[1] Alan M Nevill and Roger L Holder. Home advantage in sport. *Sports Medicine*, 28(4):221–236, 1999.

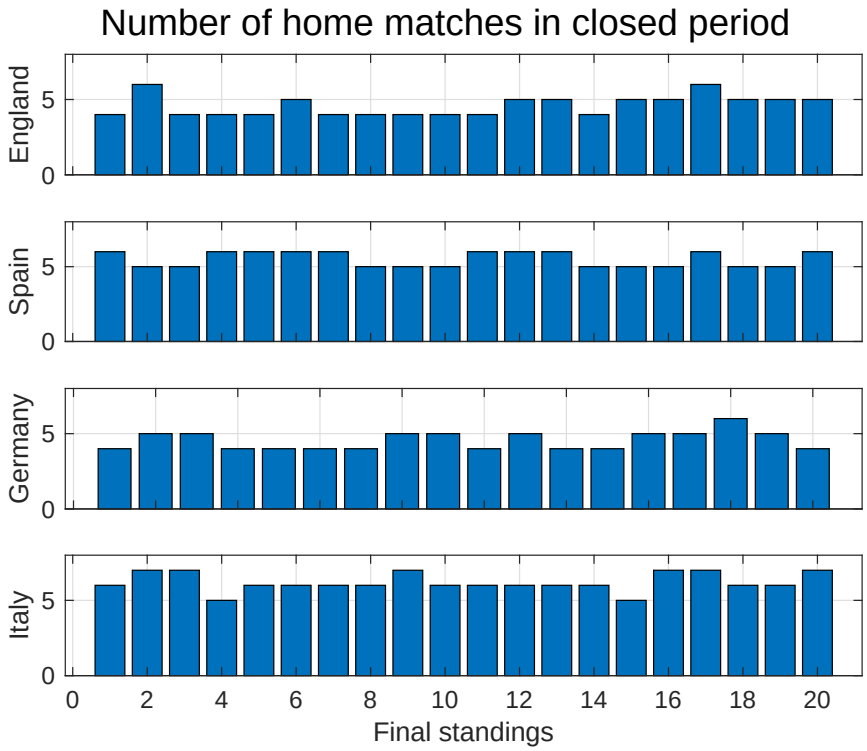

Figure 4: *Number of home matches in closed period*

[2] Richard Pollard. Home advantage in soccer: A retrospective analysis. *Journal of sports sciences*, 4:237–48, 02 1986.

[3] Alan Nevill, Sue Newell, and Sally Gale. Factors associated with home advantage in english and scottish soccer matches. *Journal of sports sciences*, 14:181–6, 04 1996.

[4] K. Courneya and A. Carron. The home advantage in sport competitions: A literature review. *Journal of Sport & Exercise Psychology*, 14:13–27, 1992.

[5] D. Morris and J. Mourinho. *The Soccer Tribe*. Cape, 1981.

[6] Jack Dowie. Why spain should win the world cup. *New Scientist*, 94:693–695, 01 1982.

[7] Richard Pollard. Home advantage in football: A current review of an unsolved puzzle. *The Open Sports Sciences Journal*, 1:12–14, June 2008.

[8] Sandy Wolfson, Delia Wakelin, and Matthew Lewis. Football supporters' perception of their role in the home advantage. *Journal of sports sciences*, 23:365–74, 05 2005.

[9] M. Lewis and Vicki Goltsi. Perceptions of contributions to the home advantage by english and greek football fans. 2007.

[10] J. James Reade, Dominik Schreyer, and Carl Singleton. Echoes: what happens when football is played behind closed doors? https://www.carlsingletoneconomics.com/uploads/4/2/3/0/42306545/closeddoors_reade_singleton.pdf. accessed 2020/7/26.

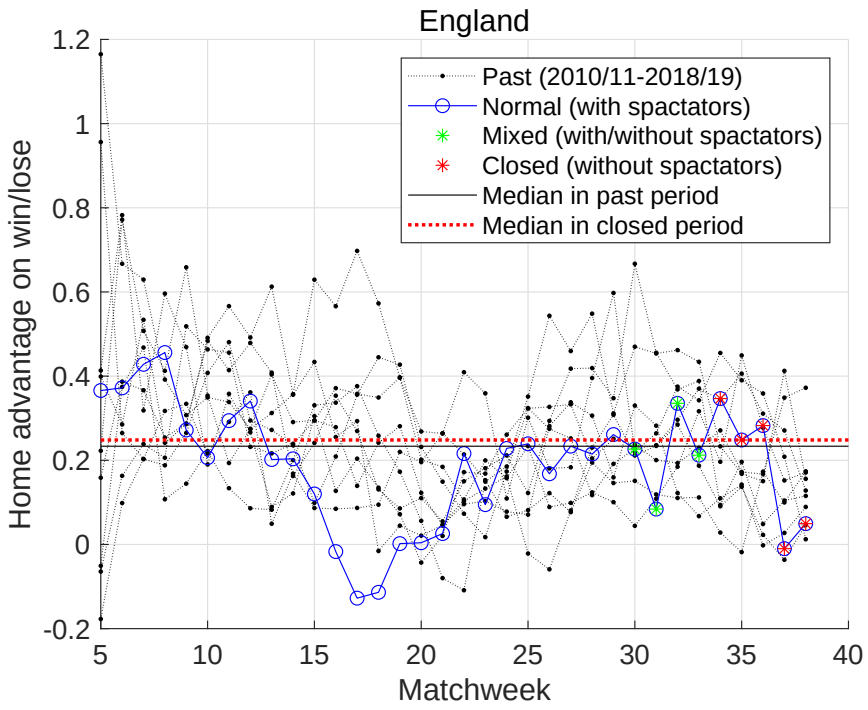

Figure 5: *History of $\bar{r}_{homeAdv}$: England*

[11] Richard Pollard and G Pollard. Long-term trends in home advantage in professional team sports in north america and england (1876-2003). *Journal of sports sciences*, 23:337–50, 05 2005.

[12] 21st Club. Empty stadiums have shrunk football teams' home advantage. https://www.economist.com/graphic-detail/2020/07/25/empty-stadiums-have-shrunk-football-teams-home-advantage, July 2020. accessed 2020/7/29.

[13] Ligue 1. PSG champions as season ended. https://www.ligue1.com/Articles/NEWS/2020/04/30/psg-champions-season-ended-ligue-1, April. accessed 2020/7/29.

[14] E. Konaka. A unified statistical rating method for team ball games and its application to predictions in the Olympic Games. *IEICE TRANSACTIONS on Information and Systems*, E102-D(6):1145–1153, June 2019.

[15] Wesley N Colley. Colley's bias free college football ranking method: The colley matrix explained. *Princeton University, Princeton*, 2002.

[16] Arpad E. Elo. *Ratings of Chess Players Past and Present*. Harper Collins Distribution Services, 1979.

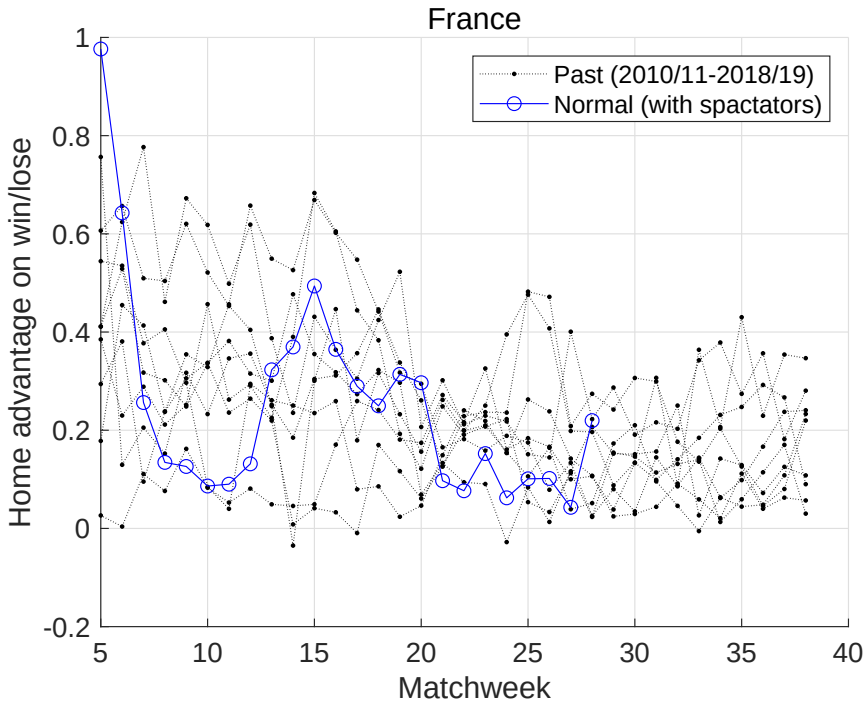

Figure 6: *History of $\bar{r}_{homeAdv}$: France*

[17] R. Hambleton, H. Swaminathan, and H.J. Rogers. *Fundamentals of Item Response Theory (Measurement Methods for the Social Science)*. Sage Publications, Incorporated, new. edition, 9 1991.

[18] J. Lasek, Z. Szlávik, and S. Bhulai. The predictive power of ranking systems in association football. *Int. J. of Applied Pattern Recognition*, 1(1):27–46, 2013.

[19] MathWorks. ranksum (Wilcoxon rank sum test). https://mathworks.com/help/stats/ranksum.html, 2020. accessed 2020/7/31.

[20] J.D. Gibbons and S. Chakraborti. *Nonparametric Statistical Inference, Fifth Edition*. Taylor & Francis, 2010.

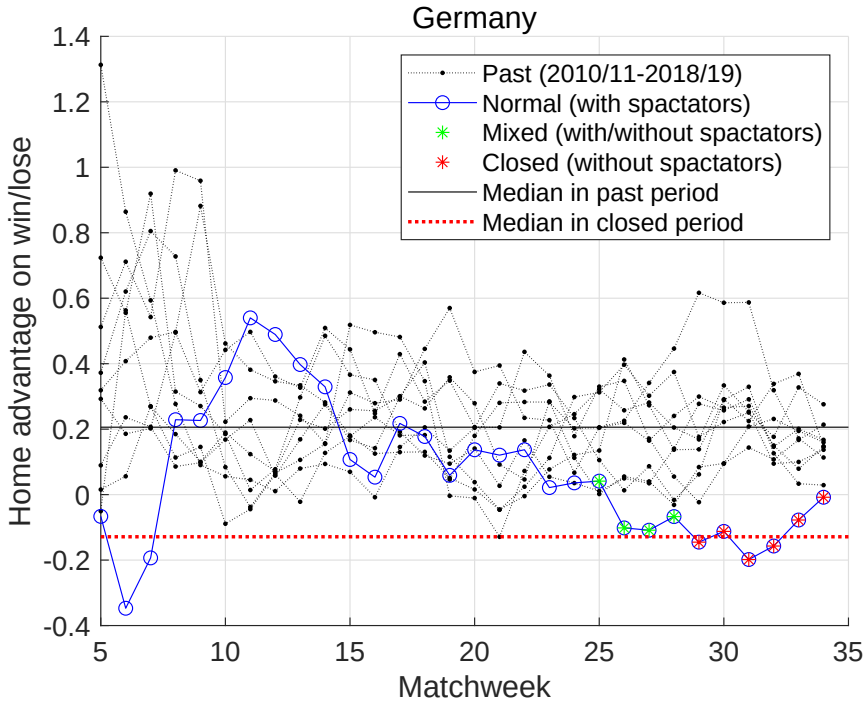

Figure 7: *History of $\bar{r}_{homeAdv}$: Germany*

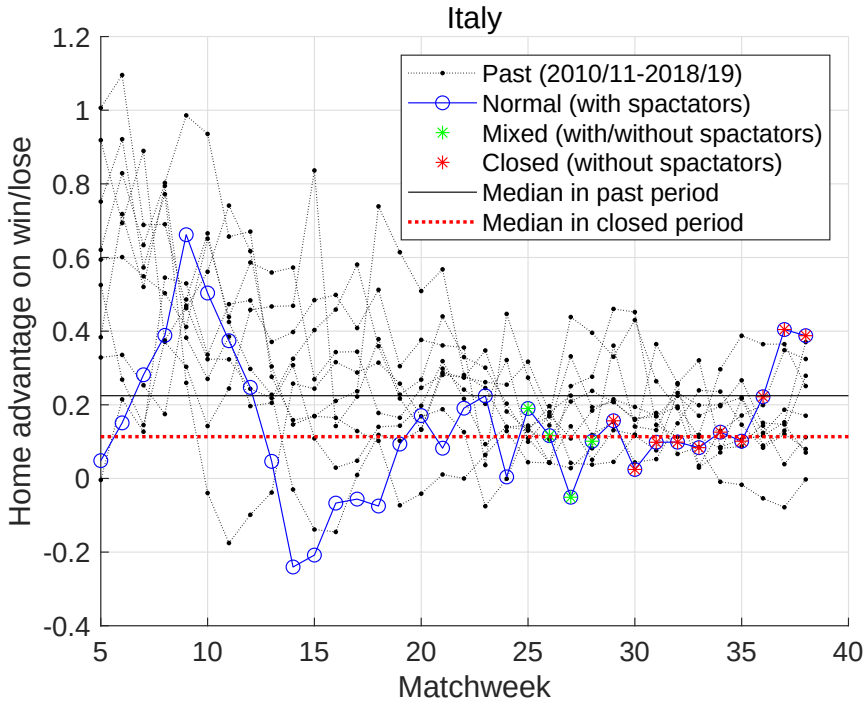

Figure 8: *History of $\bar{r}_{homeAdv}$: Italy*

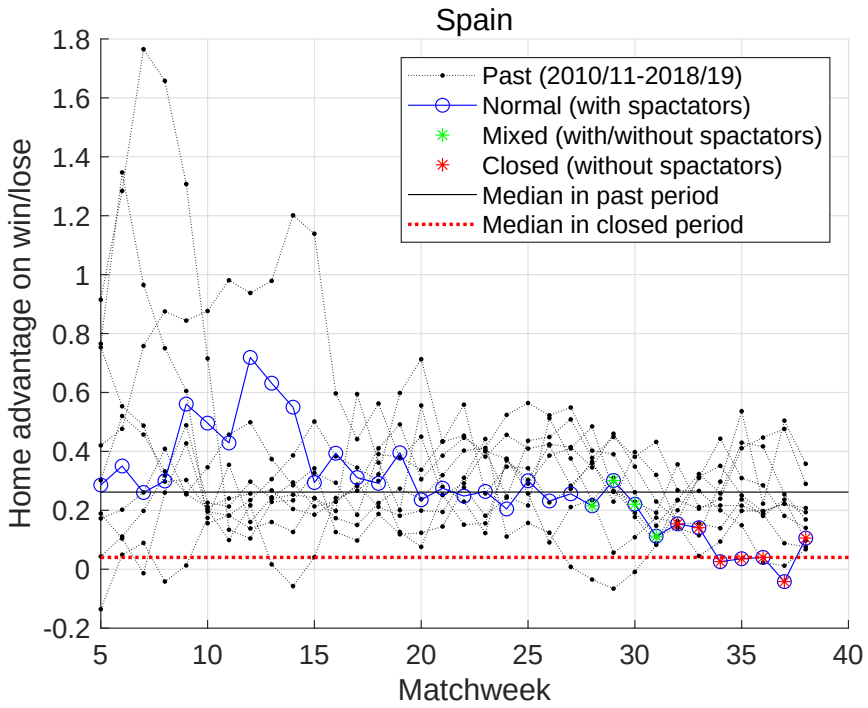

Figure 9: *History of $\bar{r}_{homeAdv}$: Spain*