Coscia *EPJ Data Science* (2024) 13:8
https://doi.org/10.1140/epjds/s13688-024-00448-3

EPJ Data Science
a SpringerOpen Journal

RESEARCH Open Access

Which sport is becoming more predictable? A cross-discipline analysis of predictability in team sports

Michele Coscia[1*]

*Correspondence: mcos@itu.dk
[1]CS Department, IT University of Copenhagen, Rued Langgaards Vej 7, Copenhagen, 2300, Denmark

Abstract

Professional sports are a cultural activity beloved by many, and a global hundred-billion-dollar industry. In this paper, we investigate the trends of match outcome predictability, assuming that the public is more interested in an event if there is some uncertainty about who will win. We reproduce previous methodology focused on soccer and we expand it by analyzing more than 300,000 matches in the 1996-2023 period from nine disciplines, to identify which disciplines are getting more/less predictable over time. We investigate the home advantage effect, since it can affect outcome predictability and it has been impacted by the COVID-19 pandemic. Going beyond previous work, we estimate which sport management model – between the egalitarian one popular in North America and the rich-get-richer used in Europe – leads to more uncertain outcomes. Our results show that there is no generalized trend in predictability across sport disciplines, that home advantage has been decreasing independently from the pandemic, and that sports managed with the egalitarian North American approach tend to be less predictable. We base our result on a predictive model that ranks team by analyzing the directed network of who-beats-whom, where the most central teams in the network are expected to be the best performing ones. Our results are robust to the measure we use for the prediction.

Keywords: Sport; Network; Centrality

1 Introduction

Following sport events is one of the main past times of society. As much as 70% of people could be considered anything from a casual to an avid sport fan, according to a number of surveys [1, 2]. As a result, professional sports are a massive global industry, with global revenues estimated at more than $500 billions for 2023 [3]. Given its societal and economic relevance, it is therefore interesting to investigate how and why sport is so enthralling – and remunerative –, as well as understanding its historic trends and – possibly – forecast its evolution.

In this paper, we want to study the historic trends of outcome predictability of sporting events. Research shows that audiences are more entertained if they think the outcome of

a match has some degree of uncertainty [4–6], although that is not necessarily the only factor – talent and stardom seem to be prominent factors as well [7, 8] – and might not hold for physical attendance but only for televised audiences [9–11] – which, it should be noted, is where most of the advertising money comes from. Uncertainty and excitement are in the mind of sport regulators. A clear case of this can be seen in Formula 1, where regulations constantly evolve to optimize close racing and overtakes – with the introduction of the Drag Reduction System and the ground effect rules.

We are inspired by a previous work which revealed that soccer is becoming more predictable over time [12]. With our main research question we aim at widening the scope: we want to expand our research to the most popular team sports. Are all disciplines getting more predictable over time? If not, which ones are?

This leads us to two secondary research questions. The first focuses on home advantage, which also affects the engagement of the public [4, 9]. Home advantage is the tendency of teams to outperform when playing in their home venue. Is home advantage changing? Has the recent COVID-19 pandemic affected the home advantage effect?

The other secondary research question involves the management of the sport by the regulators. There are two main models followed in the sport industry: the baseball model – characterized by strongly egalitarian aspects –; and the soccer model – dominated by a rich-get-richer logic [13]. Is it true that the latter leads to more predictable outcomes?

The scientific literature involved in questions related to sport outcome prediction is vast. We cannot give a complete overview here and we refer to specialized surveys for an in-depth review of the field [14, 15]. We note that most of the works focus on single disciplines and aim at creating a good predictor for the match outcomes, mostly to evaluate the efficiency of betting markets [16–19].

The literature investigating the effect of the COVID-19 pandemic on sport outcomes and home advantage effect is understandably huge [20–28], notwithstanding the fact that the event is still relatively recent. Most of these works find a large impact of COVID-19 – and, as a consequence, games without physical attendance – on the home advantage effect. The pandemic has also highlighted the sociological relevance – and issues – of modern professional sports [29]. Interestingly, studies of virtualization of sport in cycling has shown that sports and their electronic version could be comparable [30], hinting at potential future and more equal developments of professional sports.

Here, we are focusing on multiple disciplines. We are not interested in betting markets because we do not put emphasis on maximizing the prediction quality: we need predictors good enough to give us confidence that we are seeing patterns in the data and not random noise from bad predictions. In general, predicting single outcomes is hard due to the great importance of luck, and predictors are only good for long term tasks – across seasons, rather than match by match [31, 32]. A closely related work [33] only looks at upsets and focuses only on few leagues (five, while we consider 49).

Note that here we focus on team sports where teams face each other directly, and thus our framework is not applicable to races [34–36]. We also do not focus on predicting the outcome of a given match by analyzing its evolution as the event unfolds [37, 38].

We analyze data from more than 300,000 matches, spread across more than 1000 seasons, 49 leagues, and nine disciplines (baseball, basket, cricket, football, handball, hockey, rugby, soccer, and volleyball), in the 1996-2023 period. Our results show that there is no generalized trend in predictability across sport disciplines: some sports (soccer, volley-

ball) are getting more predictable over time, others (cricket, handball) are getting less predictable, and others have no clear trend. On the other hand, there has been a clear decline in home advantage effect that predates COVID-19 by a decade. Contrary to what most of the literature shows, we see that the effect of the pandemic is unclear – it might have, if anything, merely accelerated a process that was already in motion. Finally, the egalitarian aspect of the baseball model of sport organization leads to more uncertain outcomes and to weaker home advantage effects.

Our predictor relies on techniques from network science, which have been used successfully in the past to analyze sport match outcomes [39–42]. Methodologically, this method is a replication of [12]: we model a season as a network, with weighted directed edges going from the defeated team to the winner. In this network, the PageRank of a team is directly proportional to its performance and we use it to predict future match outcomes, in a simple logit model. To guarantee the robustness of our results, we use alternative scores from PageRank, using Bayesian team quality updates with an Elo-like ranking system, and also a simpler naive predictor using a frequentist approach to team quality estimation. All predictors return highly correlated predictions and they agree on all research questions, showing the robustness of our results.

We do not use official ranking systems as our predictors because information might not be available for all disciplines and leagues and, in any case, Elo-like systems usually perform better than official rankings [43].

Data and code to reproduce our results are publicly available as Supplementary Material (see Additional files 1 and 2).[1]

2 Material & methods

2.1 Data

2.1.1 Sources

We obtain data from a variety of sources. The sources are either open or proprietary. The open sources cover the following disciplines: rugby, soccer, basket, and cricket.

For rugby, our source is Wikipedia. Wikipedia contains the final score results for a number of competitions and long timelines. We scraped the content available by parsing the publicly available HTML pages. For soccer, the source is https://www.football-data.co.uk/, which provides downloadable CSV files for all major leagues starting in the 90s. For basketball, our source is https://www.basketball-database.com/, which provides systematic result tables for all major leagues. Finally, for cricket our source is https://stats.espncricinfo.com/, which reports all results for international games among national teams.

The proprietary data comes from Enetpulse https://enetpulse.com/. Enepulse is a company that provides live scores and result archives for a number of disciplines. We purchased data from Enetpulse covering the sports of hockey, baseball, handball, volleyball, and football. The company played no role in the research besides providing access to the data.

The attendance data used for the case study in the Discussion comes from https://www.european-football-statistics.co.uk/attn.htm.

We make available the data and code necessary to reproduce our results as Supplementary Material, and also at http://www.michelecoscia.com/?page_id=2258. For convenience, we provide a version of the data already cleaned and in the format necessary to be

[1] Also available at http://www.michelecoscia.com/?page_id=2258.

used with our code. This means the data does not contain the final scores – which we cannot share given that it belongs to Enetpulse –, but only contains who won which match, which is sufficient for reproducing the paper.

2.1.2 Preprocessing

The main guiding criterion to decide which data to select for our analysis is that we should focus on the cases with the highest possible level of professionalism. The reason is because amateurism introduces a certain amount of randomness that would pollute whatever signal is there. In amateur scenarios, players and teams can appear/disappear regardless of their skill level, records are kept less accurately, and so on.

We follow this criterion to guide the selection of all aspects of the data, namely:

- We select the most popular disciplines, since larger audiences means more investment and thus professionalism.
- We select the most important national leagues for each discipline, since professional teams will attract the best players.
- We only consider male leagues. While female leagues have recently greatly progressed and many are equally as professional as their male counterparts, professionalism is a recent development which does not give us a long enough timeline to detect a pattern.
- We select only seasons with at least 50% of games on record at the time of data collection, since they give us enough data to reach meaningful conclusions. In some cases, this means we need to drop the 2023 season, while in other cases this is not necessary.

To clean our data, we need to take into account that, in some sports, teams frequently change names for sponsorship reasons. This is not an issue for the disciplines included in the proprietary data, since the data providers associates each team with a unique id which is constant regardless of name changes. However, for the open data sources, team names change. We perform a data cleaning step aggregating teams with different names that refer to the same organization.

For consistency reasons, following previous work [12], we discard draws. The purpose is to have a simpler binary classifier predicting whether the home team will win or lose. This is a lesser issue than in the cited work, because the cited work focused exclusively on soccer, which is the discipline producing the highest number of draws by far. All other disciplines considered here either do not allow for draws at all, or draws are much rarer than in soccer – like in the case of rugby.

2.2 Prediction model

For our predictors, we use the same general simple architecture – which we depict in Fig. 1. The framework has four phases – numbered accordingly in Fig. 1:

1. Construction of the training dataset;
2. Feature extraction from the training dataset;
3. Model training on the extracted features;
4. Prediction and evaluation.

Steps #1 and #3 are identical no matter the features we use to train the model. For step #2 the default choice in the paper is to use PageRank, and for robustness we use Elo and naive features. The differences between these alternatives is explained below. We now describe the framework step by step.

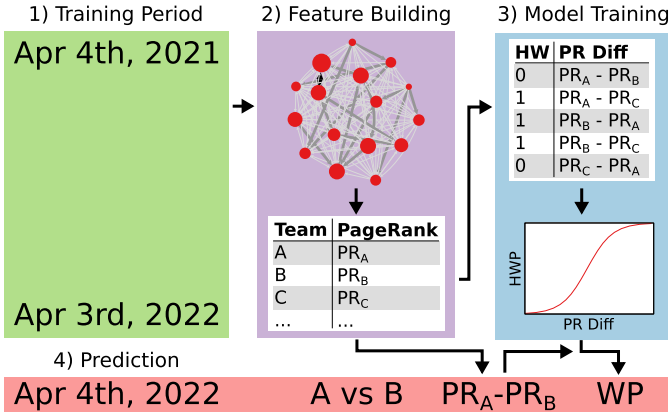


Figure 1 The workflow of our prediction model. Acronyms: HW = Home Win (binary feature); PR Diff = PageRank Difference (PageRank of home team minus PageRank of away team); PR_X = PageRank of team X; HWP = Home Win Probability (as outputted by the trained model)

2.2.1 Training dataset

In step #1, we create the training dataset on which we want to train the model. For the training dataset, we select all the matches that took place in a given league in the preceding year of the match that needs to be predicted. So to predict a match happening on April 4th, 2022, we use all matches from April 4th, 2021 until April 3rd, 2022. This means we use a dynamic sliding window, which allows us to make a prediction for the first game of a season.

The main reference work [12] had a different setup, using the first n matches of a season to predict the remaining ones. The disadvantage of that setup is that the first n matches of a season cannot be part of the test set. In our case they can, provided they are not part of the very first season in the data.

On the other hand, during the off season many teams change a significant portion of their staff – players, coaches, etc – which might cause a lower accuracy for our predictor. We think this is an acceptable trade off for three reasons. First, in many leagues, staff changes can happen regularly during the season, so this is an issue inherent to the phenomenon we study. Second, there is empirical evidence supporting dynamic performance across tournaments [44], which is captured by our sliding window setup but it is ignored by using a season-based train-test split – in this latter case, once the model is trained for a season, it will not change, assuming static performances. Finally, just like in [12], our objective is not to reach the highest possible prediction score, but rather to verify whether there are historic prediction patterns. In this sense, a simpler predictor is preferable to a more accurate one.

As a result of having a different architecture than the previous work [12], we can verify whether their results are robust to these slight perturbations – which they are, as we show in the results section.

Note that, in our framework, we do not distinguish between regular season and play-off/playout matches. Both types of matches are indifferently part of the training and/or test set, when required by the sliding window.

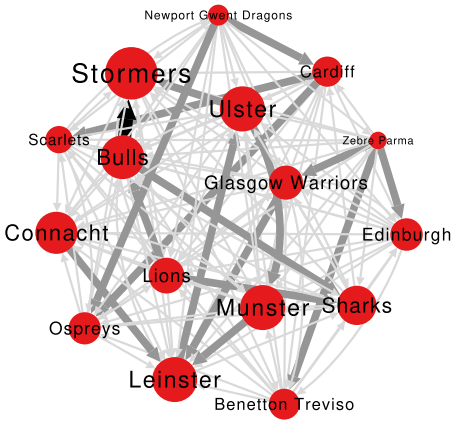


Figure 2 The network for the 2022 season of the United Rugby Championship. Each node is a team and directed links flow from the losing team to the winning team for each match. The size and color of the edges is proportional the edge's weight – the number of matches with the same winner and loser. The size of the nodes is proportional to their PageRank score. Note how the highest PageRank is assigned to the eventual champions (Stormers)

2.2.2 Feature extraction

In step #2, we extract features from the training dataset. We can do this in different ways. We focus on three: PageRank, Elo, and Naive, with PageRank being our default choice.

PageRank. We select PageRank as our default feature based on previous work using network analysis to rank teams [45, 46]. The idea is to create a directed network. In such a network, each match generates an edge. Edges flow from the defeated team to the winning team – we ignore draws. If two teams have faced each other multiple times during the training period, each match will increase their edge weight.

For instance, assume team A and team B played against each other four times in the preceding year – our training period. In three of these matches, A won, and in the remaining match B won. In this case, there will be two reciprocal edges. The $A \rightarrow B$ edge will have weight three, while the $A \leftarrow B$ edge will have weight one. Figure 2 shows one of the networks used for prediction.

Once the network is built, we use the PageRank algorithm [47] to create the training scores. In such a setup, PageRank will give the highest centrality to the nodes that are pointed to the most. The algorithm will take into account the weights for this calculation. Finally, to predict the outcome of the match, the model will consider the difference between the home team's PageRank score minus the away team's score.

Elo. The Elo system is a popular way of ranking players and teams in competitions [48]. It was originally developed to rank chess players. In this paper, we use a development of Elo called Trueskill [49, 50]. Trueskill is a generalization of Elo that can handle teams with more than two players facing each other. However, Trueskill follows the same general principles of Elo which we outline here.

Each time a new player/team appears in the competition, they are awarded the same initial score. Then, following the results of each match, the scores of both players/teams are updated. The winner gains a certain amount of points which depends on the difference in scores they originally had with their opponent. For instance, if the winner had a lower score than the opponent they defeated, they gain a large amount of points – and the defeated opponent loses a large amount of points as a result.

In practice each match outcome in the Elo system represents new data with which the system can update its priors about the skill levels of the two competitors. Every time we need to predict the outcome of a match, we take the difference between the Elo score of the home team minus the score of the away team as our prediction feature. Note that,

since step #1 is the same for all features, we use exclusively the previous year's results to calculate the Elo scores. We keep step #1 fixed to make this predictor fair – Elo cannot use more information than the PageRank predictor, even though with more historic data it would provide more accurate predictions. This increase in predictability, however, would not be a feature of the discipline – the thing we are interested in studying here – but a mere artifact of using more data.

Naive. Both PageRank and Elo are relatively sophisticated ways of ranking teams. While neither is particular complex, we can introduce a naive ranking system that is the simplest way we can use to predict the outcome of a match. We do this to provide a simple alternative that can be easily understood, to avoid a situation in which the results may be ascribed to the complexity of the predictor.

In our naive predictor, for every match we need to predict the outcome of, we look at the training period – the year preceding the match – and calculate both teams' probability of a win. If team A in the training period won 60 matches out of 100 played, their score is 0.6. If team B won only 40 of their 100 matches played, their score is 0.4. Then, just like in the previous cases, the predictor will use the difference between the home team's and the away team's scores as the predictive feature.

2.2.3 Using the scores

In step #3 of our framework, we use the features calculated in the previous step to train the model. Again, for simplicity's sake – and consistency with previous work [12] –, we decide to use a simple setup. Our model is a logit regression:

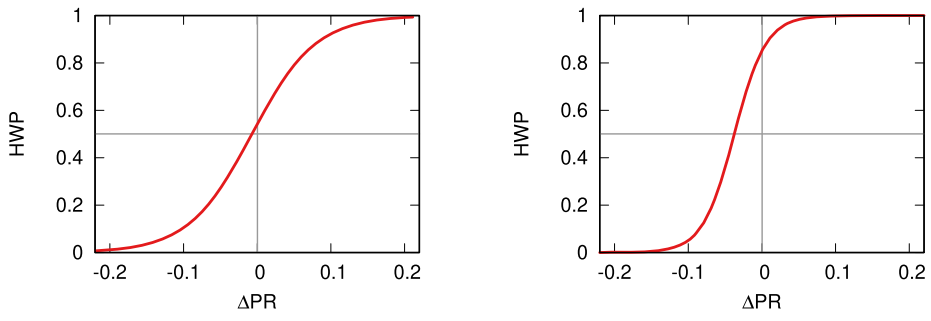
$$HW \sim \alpha + \beta(\Delta PR) + \epsilon,$$

where:

- HW is a binary variable with value 1 if the home team won, and 0 if the away team won;
- $\Delta PR = HPR - APR$ is the difference between the PageRank of the home team (HPR) and the PageRank of the away team (APR);
- α is the intercept;
- ϵ is the error term.

Once trained, the model can return a HWP (Home Win Probability) when it receives as input an arbitrary ΔPR number from a new match. The model's prediction is based on the learned sigmoid function. Figure 3 shows two examples from the model trained on two different leagues from two different sports. Looking at the figure, we can also see how the model estimates the home advantage effect in practice.

In Fig. 3(a) we look at the model trained for the 2015 season in the Japanese Baseball League, which is characterized by a small home advantage effect. If we draw a vertical line at $x = 0$, we can identify the home win probability that the model would output by recording the y value at which the line intersect with the red sigmoid. If there were no home advantage, the vertical line should intersect the sigmoid exactly where the line encounters the horizontal line at $y = 0.5$, which symbolizes even victory odds. Instead, for no PageRank difference, the model assigns a probability higher than 0.5 to the home team to win.



(a) JBL (b) Top 14

Figure 3 The home win probability (y axis) for a given Home-Away PageRank difference (x axis) for (a) the Japanese Baseball League in 2015 and (b) the French Rugby tournament in 2006

The distance between the $(0, 0.5)$ point and the sigmoid value at $x = 0$ is small for Fig. 3(a). For the model, the home team will win a match against an equally ranked opposition with only slightly higher odds than 50/50 – specifically, in this case the probability is around 54%, the value of the sigmoid at $x = 0$. On the other hand, in Fig. 3(b) the distance between the sigmoid and the $(0, 0.5)$ point is high. In this case, the model assigns a >80% probability of home win to equally skilled oppositions.

2.2.4 Prediction and evaluation

After the model is trained, in step #4 we can deploy it to obtain a home win probability for each match by looking at the home-away PageRank difference. Once we do so for each match in a season, we can evaluate how well the model performed its prediction.

The standard approach in machine learning is to build a Receiver Operating Characteristic (ROC) curve and calculate the Area Under the Curve (AUC). To build a ROC curve, one has to sort all the predictions in decreasing order of confidence: we start with the match the model is most confident will result in a home win and we end with the most likely away win match. Then we draw the ROC curve: for each match, we have a new point on the curve located at the cumulative False Positive Rate (FPR) on the x axis and at the cumulative True Positive Rate (TPR) on the y axis.

A completely random predictor will move along the identity line – it predicts a home win randomly with a 50/50 chance. Its corresponding AUC is by definition 0.5. AUCs higher than 0.5 indicate that the model performed better than random guessing. If that is the case, we can infer that the underlying data was predictable. Note that a bad model might underestimate the predictability of a dataset, but a high AUC is always a sign of high predictability.

Figure 4 shows the ROC curves across all leagues and seasons for baseball and handball which, according to Table 2, are the least and most predictable disciplines, respectively.

3 Results

3.1 Overall predictability

3.1.1 Summary statistics

We start our investigation by looking at some summary statistics across disciplines. Table 1 shows the number of matches, teams, seasons, and leagues included in our dataset. Match

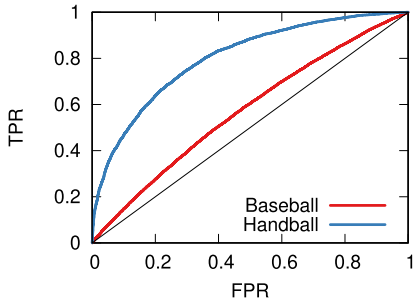


Figure 4 The ROC curve recording the True Positive Rate (y axis) against the False Positive Rate (x axis) for all matches in baseball (red) and handball (blue)

Table 1 Summary statistics for the studied disciplines

Sport	# Matches	# Teams	# Seasons	# Leagues
Baseball	44,065	52	37	3
Basket	80,631	390	246	12
Cricket	3986	96	61	3
Football	5889	32	22	1
Handball	17,852	148	83	4
Hockey	65,595	134	130	6
Rugby	12,405	102	113	5
Soccer	70,802	483	308	11
Volleyball	9918	110	54	4

Table 2 The general levels of predictability and home advantage for the studied disciplines, across all considered seasons and leagues

Sport	AUC	Home Adv.
Baseball	0.5723	0.5148
Basket	0.7078	0.6405
Cricket	0.6628	0.5837
Football	0.6432	0.5526
Handball	0.8013	0.6146
Hockey	0.6086	0.5633
Rugby	0.7170	0.6938
Soccer	0.7065	0.6418
Volleyball	0.7643	0.5421

count is after we remove the draws, as we detail in the Material & Methods section. The discipline for which we have the highest count of matches is basketball, although the count is inflated by the fact that some leagues (like the NBA) schedule teams to play against each other more than twice per season, and have lengthy playoffs. On the other hand, soccer is the most deflated discipline, as it generates the highest number of discarded draws. Soccer tops in terms of number of teams and seasons in the data.

Cricket has the least number of matches, since we only consider international encounters. Since we only have one football league, the NFL, this is also the discipline with the least number of teams and seasons.

3.1.2 Overall predictability & home advantage

Table 2 shows the predictability levels of all the disciplines considered in this paper. We only show the AUC values from our network predictor using PageRank centrality [47]. However, both the Elo and the naive ranking predictors score very similarly. The linear correlations between all three predictors are around 0.99. Interestingly, naive predictor

works best across all sports, except cricket (PageRank is best) and volleyball (Elo is best, although the difference with naive is negligible). The naive predictor is also the predictor showing the strongest home advantage bias.

We decide to show the PageRank accuracies because we do not necessarily care about showing the highest AUC scores. PageRank makes intuitive sense for predictions fans could make – e.g. one would think "My team will play team A and I think it will win, because team A lost to team B and my team already won against team B". The naive predictor – "the team who has won most so far will win" – is too simple, while the Bayesian updates in the Elo system are too complex.

All predictors agree that handball is the most predictable discipline and volleyball is the second most predictable. They also all agree that the most unpredictable discipline is baseball, followed by ice hockey. When it comes to home advantage, the discipline experiencing this effect the strongest is by far rugby, with around 70% chance of a home team winning a game against an equally strong team. The second sport for all predictors, soccer, has only a 65% home win chance in the same scenario. Baseball has by far the weakest home advantage – almost no advantage at all with 51% chance of winning for a home side against an equally skilled opponent.

Note that home field advantage is more sophisticated than simply calculating the ratio of home wins in the data. That is because many games are played in playoffs where better performing teams will play more home than away games as a reward for performing well during the regular season. For instance, only 66% of games in rugby were won by the home team but our estimation of rugby's home field advantage is 70%.

3.1.3 Predictability & home advantage by league

Table 3 shows the predictability and home advantage for each league independently – across time. In general, the table contextualizes some of the previous results. All baseball leagues in our data are the least predictable ones and they also have the weakest home advantage effect. On the other hand, handball's high predictability is driven by the Spanish and German leagues, which are extremely predictable.

The French rugby league is an outlier when it comes to home advantage, giving almost 80% chance of a home win for a team facing an equally ranked team. This confirms previous results in the literature about home advantage, which found the same outlier [51]. The reason is likely a combination of two factors. First: the physically demanding nature of the

Table 3 (a) The five most and least predictable leagues, across all considered disciplines and seasons. (b) The five leagues with the strongest and weakest home advantage effect

(a)				(b)			
#	Sport	Country	AUC	#	Sport	Country	Home Adv.
1	Handball	Spain	0.8233	1	Rugby	France	0.7976
2	Handball	Germany	0.8211	2	Basket	Greece	0.7057
3	Basket	Lithuania	0.7998	3	Rugby	URC	0.7010
4	Volleyball	Russia	0.7869	4	Basket	Adriatic	0.6856
5	Basket	Greece	0.7865	5	Soccer	Greece	0.6854
...	...	...	...	...	...	...	...
45	Cricket	International	0.5967	45	Hockey	USA	0.5244
46	Hockey	USA	0.5859	46	Volleyball	Poland	0.5213
47	Baseball	USA	0.5772	47	Baseball	USA	0.5210
48	Baseball	South Korea	0.5713	48	Baseball	Japan	0.5097
49	Baseball	Japan	0.5485	49	Baseball	South Korea	0.5089

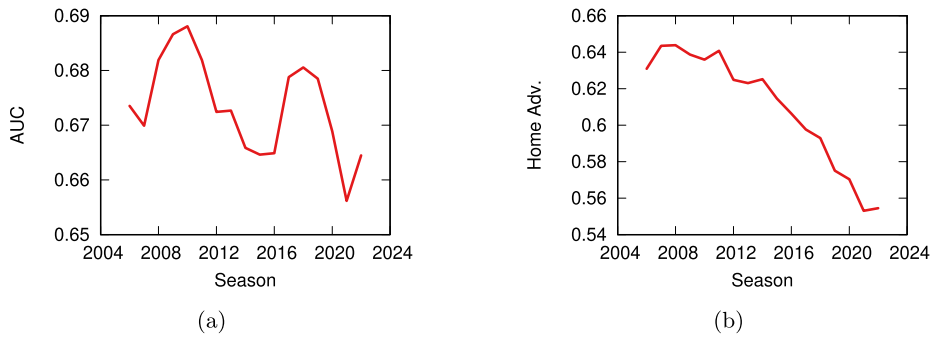


Figure 5 The predictability (a) and home advantage effect (b) across all disciplines, per season (x axis)

sport requires large rosters including more than two full teams – it is rare for a player to play more than a couple matches in a row before resting. Second: French teams find it important not to disfigure in front of their home crowd. The result is that, effectively, French rugby teams will field their best squad home and their second best squad away, effectively boosting the home advantage effect.

3.1.4 Predictability & home advantage by season

We now take a look at predictability and home advantage effect over time across all disciplines and leagues. The idea behind this analysis is to see if there is some common evolution in sport that transcends disciplines. Figure 5 shows the result. We limit ourselves to the 2006-2022 period, because that is the period when we have the most leagues available (more than 30 out of 49 total for each single year).

Figure 5(a) shows the predictability levels. It might be tempting to conclude that there is some sort of 7-8 year cycle with peaks and a general trend of reduced predictability. After all, a linear regression shows a negative slope for the season predicting AUC (with $p = 0.047$). However, we reject this hypothesis on two basis. First, the p-value is too close to 0.05 to really say anything with confidence. Second, the variation in AUC scores is very low – values range from 0.66 to 0.69. As a consequence, it is much more likely that AUC values are flat over time, and the fluctuation we observe is random. So a first answer to our main research question is that, if there are predictability trends over time, they must be discipline-specific and they are not universal.

The same cannot be said for the home advantage evolution over time – Fig. 5(b). In this case we see a decisive constant trend across sport of vanishing home advantage. The home advantage was around 64% until 2011, but it collapsed to 55% post-pandemic. Even if we were to cut the data at the 2019 season, well before COVID-19 started, the home advantage already dropped to 57.5%. The relationship is beyond any statistical doubt ($R^2 = 88\%$, $p < 0.001$). The decline of home advantage is studied and well documented in the literature, at least for soccer [12, 52].

As for the previous analysis, we rely on the PageRank predictor noting that the AUCs correlates at 0.9 with both other predictors, and home advantage correlates at 0.99.

3.1.5 COVID-19 & home advantage

Many studies have used the COVID-19 lockdown as the source of a natural experiment to establish the effect of crowds on home advantage, since during COVID-19 most (if not all)

Table 4 The time series analysis identifying the effect of COVID-19 on home advantage

	Dependent variable: Home Adv.
Season	–0.008***
	(0.0005)
% Closed Games	–0.015*
	(0.007)
Constant	15.751***
	(0.958)
Observations	12
R^2	0.976
Adjusted R^2	0.971
Residual Std. Error	0.005 (df = 9)
F Statistic	186.617*** (df = 2; 9)

Note: *p < 0.1; **p < 0.05; ***p < 0.01.

matches were played behind closed doors or with much reduced crowds [20–22, 53–55]. This is not part of the main research question of our paper – for a complete literature review we refer to [22], but we can also shed some light on this question.

For this analysis, we perform a time series regression starting for the 2011-2022 period – we narrow down the start year because that is when, historically, the home advantage effect started to decline. We then add another variable per season: the share of games played under COVID-19 restrictions. This is not data that is consistently available for all leagues and disciplines, so we make the simplifying assumption that all games were impacted in some significant fashion from mid March 2020 until June 2021.

Table 4 shows the result of the regression for the twelve seasons we consider. The table show that there is evidence for a decline in home advantage during COVID-19 that goes beyond the historic trend. However, crucially, this decline is only significant at $p < 0.1$, which is not enough to be confident about it being a solid result. The prudent conclusion should be that COVID-19 might have slightly accelerated an already-existing trend, but it did not have the large effect people might expect. This result is consistent regardless our choice of predictor variable and ranking system, as the Supplementary Sect. 1.2 shows.

Our finding goes counter most published research on the topic, for good reasons. First, many of the papers look exclusively at soccer, and only at a handful of leagues [22, 53], because of data availability and of the popularity of the discipline. Instead, we study an effect across disciplines over more leagues – we look at discipline-specific effects in the next section. Second, in many cases the research setup is to compare the COVID-19 seasons with immediately preceding seasons [20, 21], which misses the larger picture of a consistent historic decline of home advantage. Yes: 2020/2021 season had a much lower home advantage than the 2018/2019 season, but because it is part of a general decline that started in 2012. Third, many other studied did not look necessarily at wins, but at other factors such as point difference [54].

Notwithstanding our disagreement with the COVID-specific effects detected in the literature, most of our general home advantage results are consistent with the literature [51]. However, it is important to note that [51] collects data only until 2015, therefore missing the effect we found about the rapidly shrinking home advantage effect – and the COVID-19 pandemic. The home advantage decline over time is consistent with Table 3 in [51], as well as the observations in [12] about soccer.

3.2 Predictability over time

We now analyze predictability over time by discipline. In practice, we repeat the analysis behind Fig. 5, but splitting the prediction by discipline. We are interested in spotting consistent trends across disciplines in predictability and/or home advantage.

Figure 6 shows the result. The figure gives an overview that allows to compare disciplines against one another, as the y axis is consistent across sub figures. There is a heterogeneity of slopes and volatility, which seems to imply that there is no common pattern across disciplines. Leagues follow in most cases the overall trend, with some additional volatility due to the smaller dataset and the fact that a single upset impacts the predictability of a league more than the discipline in general.

As a note of color, while the various predictors are all correlated across leagues (at >0.91) they have slightly different leagues as the most and least predictable in absolute. Overall, the most predictable season was the Spanish's handball ASOBAL league in 2010, with an AUC of ∼0.9. The least predictable was instead cricket's 2018 official test season: its 16 test matches resulted in an AUC of ∼0.35, well below random guessing.

We require a more systematic way to sum up the information in the timelines. We then perform another time series regression: if we find a significant coefficient for the slope we can determine that a discipline became more predictable – if the slope is positive –, or unpredictable – if the slope is negative.

We report the result in Table 5. The table confirms that there is no overall consistent behavior across disciplines. Baseball has a weak increase in predictability that could be ascribed to a reversion to the mean, since it is the most unpredictable sport – see Table 2.

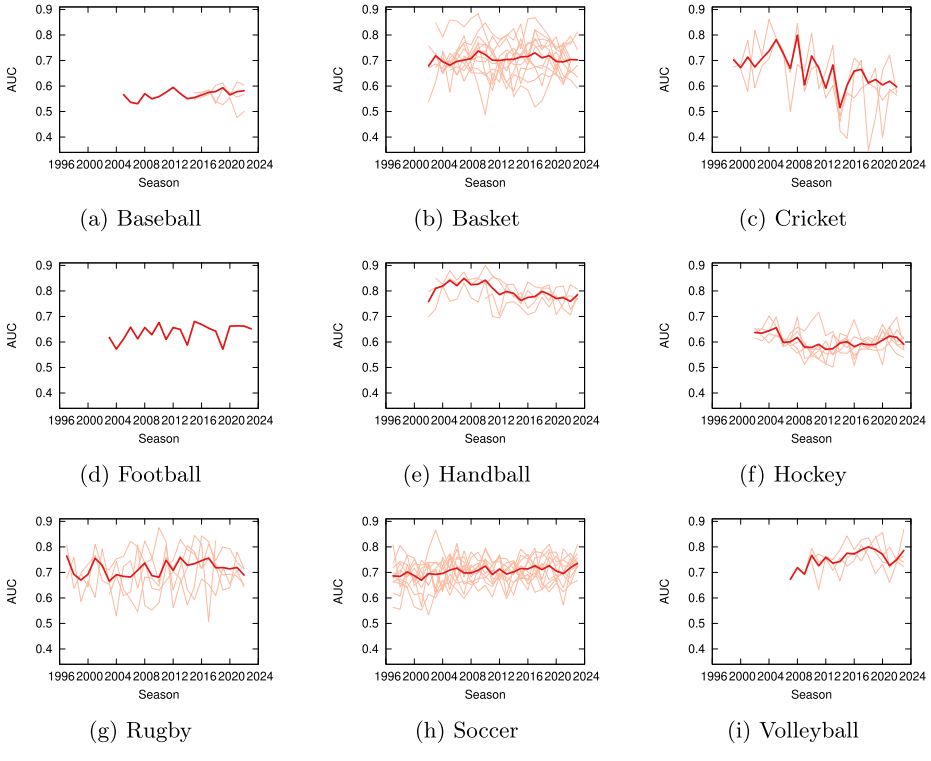


Figure 6 The evolution of AUC for the PageRank predictor (y axis) across seasons (x axis). Thick line in foreground for the sport overall, split per league in the faded background

Table 5 The predictability trends for all disciplines. ↑, ↓, –: increased, decreased and unchanging predictability level, respectively. ***p < 0.001, **p < 0.01, *p < 0.05

Sport	Pred.	*p*	R^2
Baseball	↑	*	0.27
Basket	–		0.00
Cricket	↓	**	0.34
Football	–		0.08
Handball	↓	**	0.31
Hockey	–		0.13
Rugby	–		0.01
Soccer	↑	***	0.43
Volleyball	↑	**	0.38

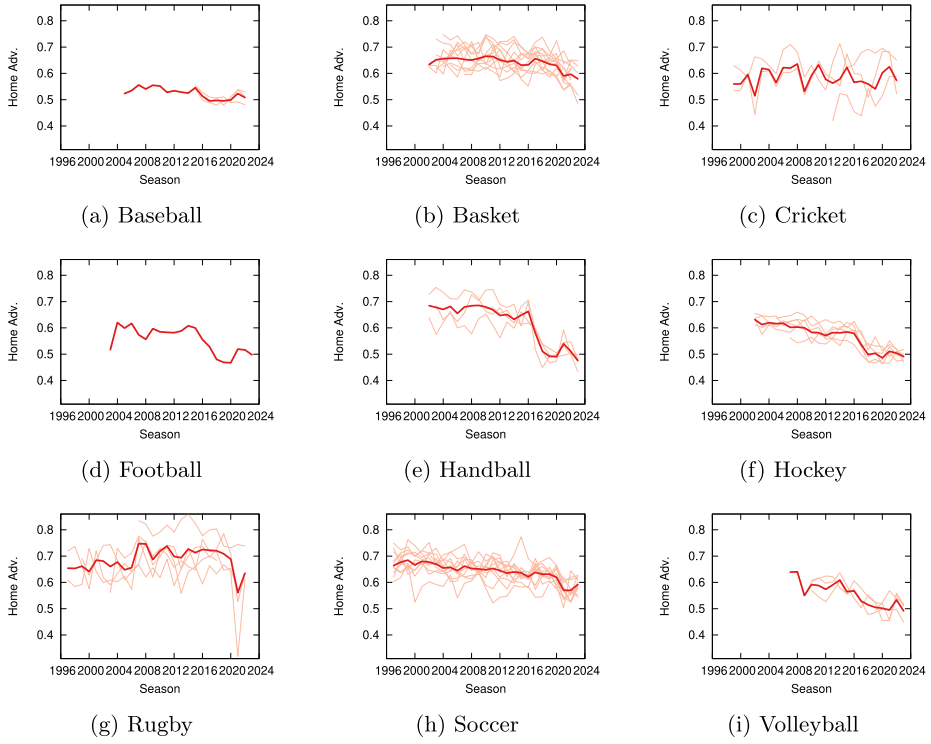


Figure 7 The evolution of the home advantage effect for the PageRank predictor (y axis) across seasons (x axis). Thick line in foreground for the sport overall, split per league in the faded background

Soccer and volleyball, instead show a strong trend of increased predictability. Handball is becoming less predictable but it might also be reversion to the mean, given that it is the most predictable discipline. Cricket is becoming less predictable. All other disciplines – basket, football, hockey, and rugby – do not seem to be changing. This complements our discussion of Fig. 5a: there are indeed predictability trends and they are discipline-specific.

Our soccer result confirms previous work [12], which was not a given since our methodology slightly differs – we have a continuous sliding window rather than the train-test split on a seasonal basis used in the cited work.

Elo and naive predictors return the same picture, with slightly different levels of confidence. In fact, the correlation of their AUCs with the ones coming from PageRank is 0.95.

Table 6 The home advantage effect trends for all disciplines. ↑, ↓, –: increased, decreased and unchanging home advantage effect level, respectively. ***p < 0.001, **p < 0.01, *p < 0.05

Sport	Home	p	R^2
Baseball	↓	**	0.47
Basket	↓	***	0.44
Cricket	–		–0.04
Football	↓	***	0.42
Handball	↓	***	0.73
Hockey	↓	***	0.86
Rugby	–		–0.02
Soccer	↓	***	0.77
Volleyball	↓	***	0.74

We repeat the analysis for the home advantage effect. Figure 7 shows the timelines of home advantage. Notwithstanding some exceptions – most notably cricket –, we see that the shrinking of the home advantage is widespread across disciplines. We again look at the slope of the time series to confirm this impression.

Table 6 shows the result. In every discipline, we either see a drop in home advantage or no change – for cricket and rugby. To be fair, the timeline for rugby seems to imply a strengthening of the home advantage effect but, at the same time, the strong drop due to COVID-19 in 2021 nullified that earlier trend.

We can see from Fig. 7 that rugby's home advantage drop is mostly driven by one league in 2021: the Super Rugby bringing together Australia's and New Zealand's clubs. Due to COVID, the Super Rugby 2021 was extremely weird and resulted in a strong home disadvantage: the home team had only a 32% of winning against an equally skilled opposition, according to the average of our predictors. Unsurprisingly from what we saw in the previous section, rugby French Top14 has most of the strongest home advantage effects over the years, topping in 2014, giving an 85% chance of a home win between equally skilled teams. To give an idea, that season's eventual champion, Toulon, only won 4 out of its 13 away games, and two of them were against the bottom two teams of the league that got relegated that season.

Once again, these results are confirmed when using Elo and naive predictors, due to the high correlations in their estimation of the home advantage effect across leagues and seasons (>0.98).

3.3 Leveling the playing field

There are two main models to organize the sport industry, which have been studied in the economics literature: the baseball model, popular in North America; and the soccer model, popular everywhere else [13]. The baseball model is characterized by strongly egalitarian aspects: players have strong unions, there is a draft system where the best new players are assigned to the worst performing teams to level the playing field, etc. The soccer model is ultra-capitalistic: players do not have unions, the richest teams can outbid everybody else for the best players, etc. The soccer model is more likely to create rich-get-richer effects: success brings in more money, which can in turn buy more success via higher budgets. This has generated calls for regulations to reduce inequalities, e.g. in the form of salary caps, although their effect is unclear [56].

The literature has explored the question of whether the baseball model leads to more unpredictable results, concluding that it does [13]. Here we can investigate the same ques-

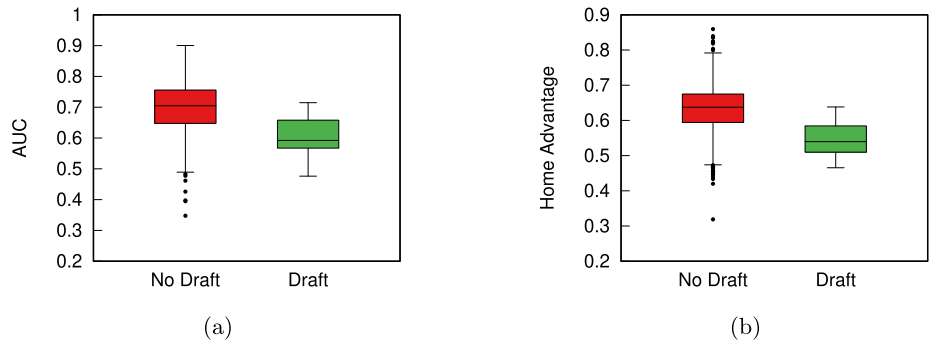


Figure 8 The distributions of (a) AUC and (b) home advantage effect, in the two classes of leagues (x axis): soccer (red) and baseball (green)

tion, to see whether the last 20 years have brought a change or not – and with a wider range of disciplines and leagues.

We classify each league into "draft" or "no draft" classes depending whether they follow the baseball or the soccer model, respectively. Then we look at their distribution of AUC values and home advantage effects over the seasons. Figure 8 shows these distributions.

From Fig. 8(a) we can see that the leagues with a draft system have lower level of predictability: their AUCs tend to be lower. This is statistically significant, after testing the difference between the two distributions with the Mann-Whitney U test ($p < 0.001$). At the same time, Fig. 8(b) shows that the leagues with a draft system also have a statistically significant lower home advantage effect – again tested with the Mann-Whitney U test ($p < 0.001$).

We obviously cannot establish causality this way, but the correlation is present. This correlation is not driven by the accident of baseball being the least predictable and with the weakest home advantage discipline in our dataset. The difference in predictability is statistically significant even if we were to remove baseball from the dataset.

4 Discussion

In this paper, we investigate the trends in predictability across disciplines. We find that there is no overall agreement: some disciplines become more predictable over time, some less, and some have no clear trend. On the other hand, we find that home advantage is eroding in almost all disciplines and this is not due to the effect of the COVID-19 pandemic. The decline in home advantage started somewhere around 2012. Finally, we establish that sports managed under the baseball model tend to be less predictable and have weaker home advantage effects.

Our results should be contextualized by considering a number of limitations. First, we have only looked at a handful of disciplines. More disciplines like water polo, field hockey, and others should be considered, to give an even broader picture. The great increase in popularity of e-sports [57] should also not be overlooked, although we would need to consider what does it mean for an e-team to be playing with "home advantage".

We are also only considering a fairly recent slice of sport history, from 1996 to 2023 – with some years providing data only to one or two disciplines. We should also expand to more leagues – e.g. the Mexican baseball league –, and include the women leagues. While all of these expansions are worthwhile and might provide a cleaner picture, their exclusion

is grounded in our need to use the leagues characterized by the highest possible degree of professionalism. Professionalism leads to more consistent and less random results, which will make historic trends more clear.

Having a more accurate predictor could also lead to more confident results. Any imperfection in the predictor necessarily lowers our confidence, as any trend – or lack thereof – could be ascribed to imperfect predictions. However, all the predictors seem to agree with each other to a very high degree, which in turn is a sign that our results should be robust to the specific predictor used.

Our hypothesis that the soccer model leads to more predictable results than the baseball model because its effect on team budgets should also be tested. Specifically, we should see if there indeed is a correlation between a team's budget and the number of points it can score during a season. While a simple correlation is interesting, we should also take into account the network structure – more successful teams play more against each other because of playoffs –, something that only recently has been possible thanks to the development of network correlation techniques [58–60].

The most important drawback of our study is that it only points at a potential effect without investigating its causes. Future works could investigate a potential causal link behind changes in predictability and/or home advantage. Basing ourselves on the literature, we can sketch one of the possible explanations: home advantage is related to stadium attendance [52, 61]. We can provide some suggestive evidence by comparing the case of the Spanish Liga and the Italian Serie A in soccer.

Figure 9(a) shows that the Spanish Liga has maintained a constant average match attendance in the 1996-2020 period. At the same time, it is a case where we cannot find a statistically significant indication of an actually decreasing home advantage effect over the same period. Vice versa, in Italy attendance has dwindled – with the exception of a pre-COVID-19 recovery which, due to the pandemic, we cannot know if it was a random fluctuation or a real trend. Italy also represents one of the strongest and most significant case of a declining home advantage.

This fact lends some credence to our hypothesis. However, we could not include the full analysis in the paper as we could not access to reliable data about the attendance records across disciplines for the entire period of our data. Moreover, quantitative audience might not tell the full story: some effect could also be due to stadium layouts – e.g. a smaller

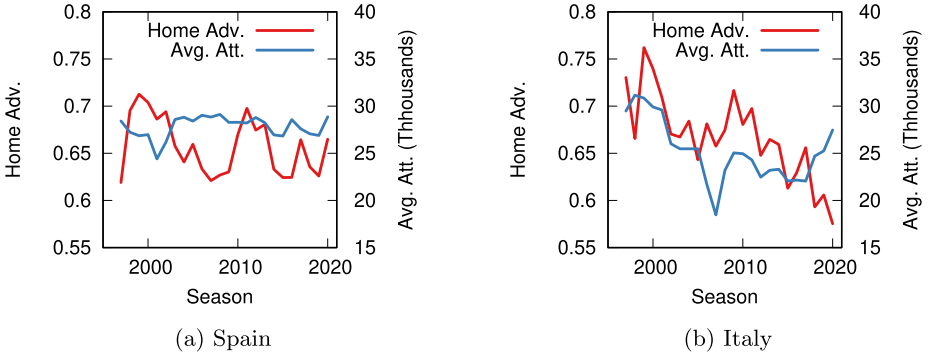


Figure 9 Comparing home advantage effect (red) and average attendance (blue) for the (a) Spanish and (b) Italian soccer leagues, across seasons (x axis)

crowd might have more impact if it is closer to the action, as in the case of stadiums without an athletics track around the field, an explanation that so far has led to mixed results in the literature [21, 62]. We leave this investigation for future work.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-024-00448-3.

Additional file 1. (PDF 152 kB)
Additional file 2. (ZIP 2.2 MB)

Acknowledgements
We thank Luca Pappalardo for insightful conversations and suggestions. We thank Enetpulse and specifically Peter Holst for making available their sport data to us.

Funding
No funding has been used for this project.

Abbreviations
APR, Away PageRank; ASOBAL, Asociación de Clubes de Balonmano de EspaÑa; AUC, Area under the ROC Curve; COVID-19, Coronavirus disease 2019; CSV, Comma-separated values; FPR, False Positive Rate; HPR, Home PageRank; HTML, HyperText Markup Language; HW, Home win; HWP, Home Win Probability; JBL, Japanese Baseball League; NBA, National Basketball Association; NFL, National Football League; PR, PageRank; ROC, Receiver Operating Characteristic; TPR, True Positive Rate.

Data availability
The dataset supporting the conclusions of this article is included within the article (and its additional files). The dataset supporting the conclusions of this article is also available at http://www.michelecoscia.com/?page_id=2258.

Declarations

Competing interests
The author declares no competing interests.

Author contributions
As solo author, MC took care of all the tasks necessary for this project. The author read and approved the final manuscript.

Received: 15 August 2023 Accepted: 18 January 2024 Published online: 29 January 2024

References

1. Siena College Research Institute (2023) American sports fanship survey. https://scri.siena.edu/american-sports-fanship-survey/. Accessed 2023-04-03
2. Morning Consult (2023) Sports fans share in the U.S. https://www.statista.com/statistics/300148/interest-nfl-football-age-canada/. Accessed 2023-04-03
3. The Business Research Company (2023) Sports global market report. https://www.thebusinessresearchcompany.com/report/sports-global-market-report. Accessed 2023-04-03
4. Forrest D, Simmons R (2002) Outcome uncertainty and attendance demand in sport: the case of English soccer. J R Stat Soc, Ser D, Stat 51(2):229–241
5. Forrest D, Simmons R, Buraimo B (2005) Outcome uncertainty and the couch potato audience. Scott J Polit Econ 52(4):641–661
6. Schreyer D, Schmidt SL, Torgler B (2018) Game outcome uncertainty in the English premier league: do German fans care? J Sports Econ 19(5):625–644
7. Buraimo B, Simmons R (2015) Uncertainty of outcome or star quality? Television audience demand for English Premier League football. Int J Econ Bus 22(3):449–469
8. Wills G, Tacon R, Addesa F (2022) Uncertainty of outcome, team quality or star players? What drives TV audience demand for UEFA Champions League football? Eur Sport Manag Q 22(6):876–894
9. Coates D, Humphreys BR (2012) Game attendance and outcome uncertainty in the national hockey league. J Sports Econ 13(4):364–377
10. Mills B, Fort R (2014) League-level attendance and outcome uncertainty in US pro sports leagues. Econ Inq 52(1):205–218
11. Cox A (2018) Spectator demand, uncertainty of results, and public interest: evidence from the English Premier League. J Sports Econ 19(1):3–30
12. Maimone VM, Yasseri T (2021) Football is becoming more predictable; network analysis of 88 thousand matches in 11 major leagues. R Soc Open Sci 8(12):210617
13. Szymanski S (2003) The economic design of sporting contests. J Econ Lit 41(4):1137–1187
14. Stekler HO, Sendor D, Verlander R (2010) Issues in sports forecasting. Int J Forecast 26(3):606–621

15. Wunderlich F, Memmert D (2021) Forecasting the outcomes of sports events: a review. Eur J Sport Sci 21(7):944–957
16. Goddard J (2005) Regression models for forecasting goals and match results in association football. Int J Forecast 21(2):331–340
17. McHale I, Morton A (2011) A bradley-terry type model for forecasting tennis match results. Int J Forecast 27(2):619–630
18. Goddard J, Asimakopoulos I (2004) Forecasting football results and the efficiency of fixed-odds betting. J Forecast 23(1):51–66
19. Peeters T (2018) Testing the wisdom of crowds in the field: Transfermarkt valuations and international soccer results. Int J Forecast 34(1):17–29
20. Tilp M, Thaller S (2020) COVID-19 has turned home advantage into home disadvantage in the German soccer Bundesliga. Front Sports Act Living 2:593499
21. Fischer K, Haucap J (2021) Does crowd support drive the home advantage in professional football? Evidence from German ghost games during the COVID-19 pandemic. J Sports Econ 22(8):982–1008
22. Leitner MC, Daumann F, Follert F, Richlan F (2023) 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. Manag Rev Q 73:605–633
23. McCarrick D, Bilalic M, Neave N, Wolfson S (2021) Home advantage during the COVID-19 pandemic: analyses of European football leagues. Psychol Sport Exerc 56:102013
24. Higgs N, Stavness I (2021) Bayesian analysis of home advantage in North American professional sports before and during COVID-19. Sci Rep 11(1):14521
25. Bilalić M, Gula B, Vaci N (2021) Home advantage mediated (HAM) by referee bias and team performance during COVID. Sci Rep 11(1):21558
26. Correia-Oliveira CR, Andrade-Souza VA (2022) Home advantage in soccer after the break due to COVID-19 pandemic: does crowd support matter? Int J Sport Exerc Psychol 20(4):1245–1256
27. Wunderlich F, Weigelt M, Rein R, Memmert D (2021) How does spectator presence affect football? Home advantage remains in European top-class football matches played without spectators during the COVID-19 pandemic. PLoS ONE 16(3):0248590
28. Sors F, Grassi M, Agostini T, Murgia M (2021) The sound of silence in association football: home advantage and referee bias decrease in matches played without spectators. Eur J Sport Sci 21(12):1597–1605
29. Rowe D (2020) Subjecting pandemic sport to a sociological procedure. J Sociol 56(4):704–713
30. Westmattelmann D, Grotenhermen J-G, Sprenger M, Schewe G (2021) The show must go on-virtualisation of sport events during the COVID-19 pandemic. Eur J Inf Syst 30(2):119–136
31. Aoki RY, Assuncao RM, Melo PO (2017) Luck is hard to beat: the difficulty of sports prediction. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1367–1376
32. Pappalardo L, Cintia P (2018) Quantifying the relation between performance and success in soccer. Adv Complex Syst 21(03n04):1750014
33. Ben-Naim E, Vazquez F, Redner S (2006) Parity and predictability of competitions. J Quant Anal Sports 2(4):1–14
34. Smith TB, Hopkins WG (2011) Variability and predictability of finals times of elite rowers. Med Sci Sports Exerc 43(11):2155–2160
35. Spencer M, Losnegard T, Hallén J, Hopkins WG (2014) Variability and predictability of performance times of elite cross-country skiers. Int J Sports Physiol Perform 9(1):5–11
36. Malcata RM, Hopkins WG (2014) Variability of competitive performance of elite athletes: a systematic review. Sports Med 44:1763–1774
37. Merritt S, Clauset A (2014) Scoring dynamics across professional team sports: tempo, balance and predictability. EPJ Data Sci 3:4
38. Clauset A, Kogan M, Redner S (2015) Safe leads and lead changes in competitive team sports. Phys Rev E 91(6):062815
39. Cintia P, Rinzivillo S, Pappalardo L (2015) A network-based approach to evaluate the performance of football teams. In: Machine learning and data mining for sports analytics workshop, Porto, Portugal
40. Yucesoy B, Barabási A-L (2016) Untangling performance from success. EPJ Data Sci 5(1):17
41. Cintia P, Coscia M, Pappalardo L (2016) The haka network: evaluating rugby team performance with dynamic graph analysis. In: 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE, pp 1095–1102
42. Fraiberger SP, Sinatra R, Resch M, Riedl C, Barabási A-L (2018) Quantifying reputation and success in art. Science 362(6416):825–829
43. Lasek J, Szlávik Z, Bhulai S (2013) The predictive power of ranking systems in association football. Int J Appl Pattern Recogn 1(1):27–46
44. Cattelan M, Varin C, Firth D (2013) Dynamic Bradley–Terry modelling of sports tournaments. J R Stat Soc, Ser C, Appl Stat 62(1):135–150
45. Motegi S, Masuda N (2012) A network-based dynamical ranking system for competitive sports. Sci Rep 2(1):1–7
46. Lazova V, Basnarkov L (2015) Pagerank approach to ranking national football teams. arXiv preprint. arXiv:1503.01331
47. Page L, Brin S, Motwani R, Winograd T (1999) The PageRank citation ranking: bringing order to the web. Technical report, Stanford infolab
48. Elo AE (1978) The rating of chessplayers, past and present. Arco Pub.
49. Herbrich R, Minka T, Graepel T (2006) Trueskill™: a Bayesian skill rating system. Adv Neural Inf Process Syst 19
50. Minka T, Cleven R, Zaykov Y (2018) Trueskill 2: an improved bayesian skill rating system. Technical report
51. Pollard R, Prieto J, Gómez M-Á (2017) Global differences in home advantage by country, sport and sex. Int J Perform Anal Sport 17(4):586–599
52. Peeters T, Ours JC (2021) Seasonal home advantage in English professional football; 1974–2018. Economist 169(1):107–126
53. Hill Y, Van Yperen NW (2021) Losing the home field advantage when playing behind closed doors during COVID-19: change or chance? Front Psychol 12:658452
54. Steinfeldt H, Dallmeyer S, Breuer C (2022) The silence of the fans: the impact of restricted crowds on the margin of victory in the NBA. Int J Sport Finance 17(3):165–177

55. Meurs E, Rehr J-P, Raue-Behlau C, Strauss B (2023) No relevant spectator impact on home advantage in male and female professional volleyball – a longitudinal multilevel logistic model analysis over 25 years. Psychol Sport Exerc 66:102401
56. Peeters T, Szymanski S (2014) Financial fair play in European football. Econ Policy 29(78):343–390
57. Neidhardt J, Huang Y, Contractor N (2015) Team vs. team: success factors in a multiplayer online battle arena game. In: Academy of management proceedings, vol 1. Academy of Management, Briarcliff Manor, p 18725
58. Coscia M (2020) Generalized Euclidean measure to estimate network distances. In: Proceedings of the international AAAI conference on web and social media, vol 14, pp 119–129
59. Coscia M (2021) Pearson correlations on complex networks. J Complex Netw 9(6):036
60. Devriendt K, Martin-Gutierrez S, Lambiotte R (2022) Variance and covariance of distributions on graphs. SIAM Rev 64(2):343–359
61. Ehrlich J, Potter J (2023) Estimating the effect of attendance on home advantage in the national basketball association. Appl Econ Lett 30(11):1471–1482
62. Pollard R, Armatas V (2017) Factors affecting home advantage in football world cup qualification. Int J Perform Anal Sport 17(1–2):121–135

Publisher's Note