Erasmus University Rotterdam

Erasmus School of Economics

Bachelor Thesis BSc^2 econometrics/economics

Corruption and relative home advantage in the European FIFA World Cup qualifiers: a data and machine learning analysis

Name student: Mati Listeş

Student ID number: 533494

Supervisor: Prof. dr. ir. Jan (JC) van Ours

Second assessor: ?

Date final version: 9 June 2023

Keywords - Corruption, relative home advantage, UEFA, FIFA

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam

Abstract

This paper extends previous research on relative home advantage in professional football by Peeters and Van Ours (2021). This extension examines the effect of endemic corruption on relative home advantage enjoyed by European national football teams in the European FIFA World Cup qualifiers. Data on the corruption perceptions index, the UEFA association club coefficient, results over 7 editions of European qualifiers, demographic and economic data are used. The quantitative analysis is done using ordinary least squares, instrumental variable regression and Ridge regression on the processed data using a slightly adapted version of the methodology by Peeters and Van Ours (2021). The main findings show that more corrupt countries enjoy a larger relative home advantage. The findings on the effects of corruption channeled through relative home attendance and through investments in football are inconclusive. Finally, statistical significance and robustness of the determined relationships are often questionable.

1 Introduction

Of the vast variety of modern sports, the game of football is arguably the most well-known. After all, football has become a way for two football factions to nonviolently resolve their rivalry unlike through warfare in the past. The beautiful game, as it is dubbed by many media, has attained the power to unite a population behind attempting to obtain an ultimate legacy in sports. When it comes to football at a national level, this ultimate legacy is given by winning the famous World Cup tournament. This tournament is organized by the world-governing body of football, known as the Fédération Internationale de Football Association or FIFA in short. Winning this tournament is generally seen as the highest prestige for any national football team and in terms of national pride. A recent example is the Argentine victory in the FIFA 2022 World Cup in Qatar, which took large masses of Argentines into the streets to celebrate Lionel Messi's first World Cup victory, which was also the first such victory since the 1986 World Cup in which the famous Diego Maradona featured. This popularity of international football has led the FIFA World Cup to become a largely lucrative business through among others large sponsorship deals and television rights. This gives FIFA and the European governing body of football, known as the Union of European Football Associations or UEFA in short, an unprecedented level of power and influence worldwide. Unfortunately, this power inevitably leads to abuse and corruption by highly powerful people in this sports business. Take for example the Swiss corruption investigation of 2015, leading to several high-ranking FIFA officials being arrested on charges of bribery in the awarding of the hosting rights of the FIFA World Cup to Russia in 2018 and to Qatar in 2022 (Rollin, 2023). Considering this level corruption is clearly present, it is not unreasonable to think that some national football associations might have also bribed for instance referees or other officials to give their national football team an unfair home advantage in the World Cup qualifiers, which every continental football confederation organizes to determine what countries can take part in the tournament on behalf of the continent in question. Corruption might however also work the other way around: home supporters might stay away from games of their national football team at home if the game there has a reputation to be dirty, which might disadvantage the home team through less home support.

This paper thus intends to answer the following research question:

"Does corruption relevantly impact the relative home advantage of European national football teams in the European FIFA World Cup qualifiers?"

This paper aims to answer the research question above by answering three sub questions by extending previous research on relative seasonal home advantage in English professional football by Peeters and Van Ours (2021). The first sub question is defined as follows: *"Is the corruption perceptions index a relevant determinant in the relative home advantage of European national football teams in the European FIFA World Cup qualifiers?"* Here the corruption perceptions index (CPI) is an indicator by Transparency International (2023) that provides the extent to which corruption is perceived as endemic in a certain country by relevant experts in the field, like businesspersons. The larger this index is, the less corruption is perceived as a relevant problem (Independent Commission Against Corruption, 2023). This index will be used as a proxy for actual corruption figures since those are part of the underground economy and are thus generally hard to find. Arguments can be brought for this effect to be both positive and negative. Relatively larger levels of endemic corruption may on the one hand positively affect relative home advantage through corrupt officials rigging games in favor of the home nation. On the other hand, this effect may be negative through for instance boycotting of the football system by investors through reduced trust or through boycotting of matches by the home attendance.

The second sub question is defined as follows: *"Is the impact of the CPI on relative home advantage of European national football teams in the European FIFA World Cup qualifiers channeled through relative home attendance?"* This question aims to answer whether indeed the endemic nationwide corruption indicator impacts relative home advantage European national football teams enjoy in the qualifiers through a reduced home attendance. After all, if the game is perceived to be dirty home supporters might stay away more, which by the findings of Peeters and Van Ours (2021) would mean a reduced relative home advantage and thus corruption would disadvantage the home team.

The third and final sub question is defined as follows: *"Is the impact of the CPI on relative home advantage of European national football teams in the European FIFA World Cup qualifiers channeled through the UEFA association club coefficient?"* The UEFA association club coefficient is a coefficient indicative of recent club-level results in European international club competitions for a particular football association (UEFA, 2023b). The intuition behind its potential relationship with relative home advantage is that the performance of national football teams is likely to be influenced to some extent by this coefficient. After all, the higher the coefficient the more indicative it is to larger prize money from European international club competitions, which can then be used for larger investments in national competitions which may lead to better performing national squads being bred in the process, which would positively impact relative home advantage as defined by Peeters and Van Ours (2021). Thus, this coefficient will be used as a proxy for investments in football. Once again arguments can be brought for both a positive and negative effect. On the one hand, relatively larger levels of endemic corruption may reasonably be expected to give the home nation more illegal funding through UEFA which would put it at an unfair advantage relative to its opposition. On the other hand, it could indicate reduced development of the sport through lower trust in the national football system in question, resulting lower investments and thus squads of lower strength.

This research is motivated by the fact that crimes related to corruption, both in football and beyond, are clearly part of the underground economy. This makes obtaining data on said crimes, and thus conducting research on their potential impact on professional football, complicated. It is necessary to attempt it regardless because of the destructive consequences corruption, both in football and beyond, can have on society and current literature on identifying a causal relationship between corruption and relative home advantage in national football is lacking for aforementioned reasons. While the setup of this paper will clearly use the CPI as a proxy for corruption because data on actual crimes like bribery and money laundering are scarce, it will at least make a first step in expanding existing literature and starting what could be a broader investigation into corruption in the FIFA World Cup qualifiers and football in general.

This research is particularly relevant for UEFA and FIFA, because it is UEFA and FIFA that are responsible for organizing the European qualifiers for the World Cup. If corruption indeed gives certain European national football teams a larger relative home advantage, or the other way around a disadvantage, during the qualifying stage of the FIFA World Cup, this means an illegal and unfair advantage or disadvantage is enjoyed by some teams but not by others. Since theoretically UEFA and FIFA should try to reduce inequalities in football, they then might fight harder to tackle corruption present in football to increase fairness. Furthermore, prosecutors and policy makers implicated in fighting corruption in football might also be interested in the results of this paper. It is those people who are the main assets in reducing corruption and the destructive inequality and unfairness it brings into football and beyond. This paper might give these people a reliable indication of the extent to which corruption in the European World Cup qualifiers is a problem such that appropriate measures can be undertaken against it.

The results of this paper show that more corrupt countries on average tend to enjoy a larger relative home advantage than dirtier countries. This paper mostly fails to detect a statistically significant and robust relationship between the two, however. The findings on the effect of corruption on relative home advantage through the UEFA association club coefficient are inconclusive. What is certain, though, is that less corrupt countries generally tend to enjoy a larger UEFA association club coefficient. This may be because less corruption tends to attract more investments into the sport. Finally, the findings on the effect channeled through relative home attendance are also inconclusive, with both a positive and negative overall effect being detected depending on the model estimated. There is consensus that cleaner countries tend to attract larger home crowds since then the game is cleaner and thus more attractive to watch in person. However, depending on the model relative home attendance subsequently either affects the relative home advantage positively or negatively.

The remainder of this paper will be structured as follows. Section 2 will elaborate on the relevant literature, while sections 3 and 4 will elaborate on the used data and methodology respectively. Furthermore, section 5 will provide and discuss the obtained results. Section 6 concludes the paper.

2 Literature

Table 1 below represents several related quantitative studies, the dependent and independent variables employed, and the results found, such that this paper and its contribution can be put into perspective. In the text below, Table 1 will be further augmented by less quantitative and more theoretical literature.

Buraimo et al. (2015), for instance, made a regression analysis on how consumer behavior responded following the Calciopoli corruption scandal in Italy. Buraimo et al. (2015) determined that home attendances dropped substantially for teams who were found guilty of and punished for corruption relative to teams that were acquitted. This economically makes sense as one would expect the demand of tickets for home games, and thus ticket revenues, to be positively affected by fairness and cleanliness of the game since supporters may be more eager to show up if they have a fair game to attend. This paper attempts to build on these findings by determining whether an increase in corruption affects relative home advantage negatively through a decreased home attendance, resulting in less support from the crowd and lost ticket revenues that may otherwise have been invested into better footballing infrastructure. The above is closely related to Peeters and Van Ours (2021), who determined a statistically significant positive effect of relative home attendance on relative home advantage. This paper attempts to determine whether their findings apply to the European FIFA World Cup qualifiers as well and, if so, how corruption plays a role in this.

Moreover, Hung Mo (2001) in his research determined that corruption decreases economic growth through various channels, one of them being private investments. After all, a more corrupt environment may discourage investors from depositing their assets in the respective country out of fear that such assets ultimately go into the pockets of corrupt officials, which means fewer assets are available for promoting economic growth. This is in accordance with the findings of Aidt (2009), who finds that the costs of corruption largely outweigh the potential gains of more speedy transactions and more efficient decision-making that would otherwise take longer. Also in line with these earlier findings are the findings by Amenta and Di Betta (2021), who found the Calciopoli scandal in Italy negatively impacted the national football industry in the long run. This paper attempts to determine whether corruption in football also decreases investments into the sport, leading to a smaller relative home advantage and thus reduced growth of the reputation in international football. Beyond the scope of this paper, this may also lead to more unemployment due to a smaller footballing sector and thus less demand for labor.

Pouliopoulos and Georgiadis (2022) examined corruption and scandals in Greek football, which they say were frequent phenomena. Pouliopoulos and Georgiadis (2022) claimed there seems to be a correlation between football and the social and political contexts in Greece and that FIFA threatened with sanctions when the Greek government tried to push through reforms. Indeed, Pouliopoulos and Georgiadis (2022) concluded that FIFA and UEFA quietly tolerated the highly institutionalized corruption present in Greek football and that only external pressure could help push through the highly necessary reforms. Considering that this paper implicitly suspects a link between the social and political contexts and national football teams, this paper is a relevant complement.

There is also a business economics part to this field. Referees are, besides employees, also humans subject to biases and incentives. Rocha et al. (2013), for instance, determined referees tend to award less extra time when the home team is ahead, while Boeri and Severgnini (2013) concluded that career concerns, such as promotions, are rather important for the corrupting of referees. Boeri and Severgnini. (2013) determined that better monitoring and adapted pay incentives for referees may contribute to reducing the risk of match fixing, while Rocha et al. (2013) argue that extra monitoring through, for example, television might help in reducing home bias. Brooks et al. (2012) find that combating fraud it is important to establish clear procedures and rules to this end once fraud is determined. If this paper indeed finds a link between corruption and relative home advantage this might be an incentive for organizations in charge, particularly FIFA and UEFA, to think about anti-corruption measures introduced previously to increase fairness of the game for the better.

There is also a political economy aspect to corruption. According to Krieckhaus et al. (2006), the impact of corruption on economic growth is partly determined by whether a country is a democracy. Krieckhaus et al. (2006) found economic growth in democracies is not significantly impacted by corruption whereas this effect is negative and significant in non-democratic regimes, which has to do with corrupt officials in positions of power eventually being ousted by the electorate in democratic countries. While distinguishing between democracies and non-democracies is beyond the scope of this paper, some European countries investigated are indeed more democratic than others, which thus might contribute to how severely corruption affects football in these countries given that professional football is also part of the economy.

Corruption can also be a source of inequality between individual players in terms of post-World Cup pay. Simmons and Deutscher (2012) found in a study conducted on the German Bundesliga that players enjoy an increase in pay resulting from a World Cup participation. In theory, if corruption enhances the relative home advantage countries enjoy in the European qualifiers, and thus their probability of participating in the next World Cup, the respective players might enjoy a larger income after the World Cup relative to players from clean countries which did not enjoy the same enhanced relative home advantage during the qualifying phase, the result ultimately being decreased equity. While this paper will not go in depth on relative wage like Peeters and Van Ours (2021) did, the results might be a starting point for future research in this field to determine whether the named mechanism indeed plays a role.

Finally, similarly to Peeters and Van Ours (2021), Pollard and Armatas (2017) investigated home advantage in the FIFA qualifiers and the effect of the FIFA ranking, home attendance and the confederation in question on this home advantage. Pollard and Armatas (2017) found a positive impact of the difference in FIFA rankings and home attendance on home advantage, with UEFA enjoying this the least. This sets an interesting precedent to the similar research that will be conducted in this paper.

Table 1 The related quantitative literature found, the dependent and independent variables used and the results established

Paper	Dependent variable	Independent variables	Results
Rocha et al. (2013)	Extra time awarded	Score difference, degree of television monitoring	Referees tend to award less extra time if the home team is ahead and this home bias is reduced through extra monitoring, e.g. television
Buraimo et al. (2015)	Attendance	Verdict in Calciopoli scandal	Home attendance dropped substantially for teams found guilty in the Calciopoli scandal
Aidt, T. (2009)	Economic growth rate	Corruption perceptions index by Transparency International	Potential positive effect of corruption through more speedy production of output largely disproven, instead finding substantial costs of corruption
Peeters and Van Ours (2021)	Relative home advantage	Type of pitch, relative home attendance	Some English football clubs tend to enjoy a larger relative home advantage than others, which depends on among others the type of pitch and relative home attendance
Amenta and Di Betta (2021)	Home attendance in Serie A	Relegation of Juventus to Serie B in the aftermath of the Calciopoli scandal	The Calciopoli scandal in the long run had a negative impact on the football industry in Italy as a whole
Hung Mo (2001)	Economic growth rate	Corruption perceptions index by Transparency International, human capital, private investments	Increasing the corruption level decreases economic growth. The channels through which corruption affects growth are among others the degree of human capital and private investments
Krieckhaus et al. (2006)	Economic growth rate	International Country Risk Guide index of corruption, Polity IV data and Freedom House measure for democracy	Economic growth in democracies is not significantly impacted by corruption, with growth in non-democratic regimes experiencing a negative and significant impact
Simmons and Deutscher (2012)	Player salary post-World Cup	Player appearance in a World Cup	Players enjoy substantial enlargement of their salaries from participation in the FIFA World Cup

Pollard and Armatas (2017)	Home advantage in the FIFA qualifiers	FIFA ranking, home attendance, confederation	The difference in FIFA rankings and home attendance have a significant positive impact on home advantage in the FIFA qualifiers, with UEFA exhibiting this advantage the least

3 Data

To conduct the research in this paper, several types of data are used. These data will be further elaborated on below. For the recording and pre-processing of the data, the statistical program Excel is used.

3.1 European FIFA World Cup qualifiers

Firstly, data are collected on 7 different editions of European FIFA World Cup qualifiers ranging from 1996 to 2021. Each edition is played over two years ahead of the FIFA World Cup in question. In other words: data are gathered for the 1996-1997, 2000-2001, 2004-2005, 2008-2009, 2012-2013, 2016-2017 and 2021-2022 editions of qualifiers. While data on European FIFA World Cup qualifiers go back further than this it was opted to only consider data for these 7 editions since 1995 is the earliest year for which data on the CPI is available (Transparency International, 2023). Since the purpose of this paper is to investigate the effect of corruption on relative home advantage, gathering data on editions for which data on the CPI are not available would not have a lot of added value. These data provide information on the number of wins, draws and losses both at home and away for each nation as well as the number of goals scored and conceded both away and at home in each of the 7 editions and on the group each nation is in. Thus, these data also provide information on the number of points obtained at home and away, the goal difference at home and away and the total goal difference, that is the sum between the goal difference at home and away for each nation in each edition. The previously introduced variables will be denoted by $HW_{ijg,}$ HD_{ijg}, HL_{ijg}, AW_{ijg}, AD_{ijg}, AL_{ijg}, GHS_{ijg}, GHC_{ijg}, GAS_{ijg}, GAC_{ijg}, HP_{ijg}, AP_{ijg}, GDH_{ijg}, GDA_{jg}, TGD_{ijg} for nation i in edition j in group g, with the specific definitions of these abbreviations given in the appendix in Table A5. These data are obtained from data base worldfootball.net (2023abcdefg), which contains information and numerical data on a vast variety of football tournaments.

Furthermore, data are collected on average home attendance, the names of the venues each team played its games in and the type of pitch each team played its games on for each edition of qualifiers. Like in Peeters and Van Ours (2021), the type of pitch of a venue is defined as a binary variable taking on the value 0 if the pitch was artificial and 1 if the pitch was natural grass. Hybrid grass, which is nowadays used by some venues, is assumed to count as natural grass. Note that in this paper this value is computed as the average over all venues that the nation in question played on since some nations played their games in more than one stadium and thus potentially on both artificial and natural grass. From now on this variable will be denoted by *ART_{ijg},* namely the proportion of games played on an artificial turf for nation i in edition j in group g. The names of the venues are recorded simply to determine the type of pitch that venue employed. Finally, the average home attendance is defined as the sum of attendances for each home game a nation played in a particular edition divided by the total number of home games that nation played. This variable will from now on be denoted by *$AVHATT_{ijg}$,* namely the average home attendance for nation i in edition j in group g. The above data will be obtained from Wikipedia (2023abcdefg).

3.2 Corruption perceptions index

Furthermore, data will be obtained on the CPI over 7 years from its founder Transparency International (2023), which is a nonprofit organization that aims to combat corruption worldwide. The 7 years comprise the starting years of each of the 7 editions of qualifiers investigated. Thus, the CPI is obtained for the years 1996, 2000, 2004, 2008, 2012, 2016 and 2021. The CPI is defined as an integer between 0 and 100, with large values indicating societies that are clean of corruption and low values indicating corruption is highly present and endemic in society. An exception to this is the CPI for the year 1996, where it is defined as a one decimal number between 0 and 100. This index is likely to be correlated with the level of corruption in football within a given country and will thus be used as an explanatory variable. From now on this variable will be denoted by *CPI_{ijg}* for nation i in edition j in group g. It will be used as a proxy for other corruption related crimes, like bribery and money laundering, since data on these crimes are generally hard if at all possible to find since they concern the underground economy. For instance, Lazic (2023) says that the estimated sum of money laundered every year ranges from 800 million United States Dollars (USD) to 2 billion USD, which is a rather large interval, of which the vast majority remains unnoticed.

3.3 UEFA association club coefficient

Furthermore, data will be collected on the UEFA association club coefficient from the official UEFA website (2023a) and another data base by Kassies (2023). The UEFA association club coefficient is defined as a three decimal number and is obtained over the 7 starting years of each of the 7 editions of qualifiers investigated. Thus, the UEFA association club coefficient is obtained for 1996, 2000, 2004, 2008, 2012, 2016 and 2021. The UEFA association club coefficient is a coefficient that is calculated based on the club-level results of each member national football association over the past five seasons in European international club competitions. The resulting ranking is used to decide the number of spots each national association gets in future European international club competitions (UEFA, 2023b). Like mentioned previously, due to the likely larger average prize money from European international club competitions associated with a larger UEFA association club coefficient and thus potentially the breeding of more and better-quality players for the national team in question, this coefficient will be used as a proxy for the funding of football. Henceforth this variable will be denoted by $UEFA_{ijg}$ for nation i in edition j in group g.

3.4 Demographic and economic data

Moreover, data were collected on the percentage of males aged 20-24 and 25-29 from the total male population and on the annual inflation percentage from the World Bank (2023a). These data will once again be collected over the 7 starting years of each of the 7 editions of qualifiers investigated. These thus again comprise the years 1996, 2000, 2004, 2008, 2012, 2016 and 2021. These variables will be denoted by $MALES\_20\_24_{ijg}$, $MALES\_25\_29_{ijg}$ and $INFL_{ijg}$ respectively for nation i in edition j in group g. These data will be used for several purposes. On the one hand these data are useful in the estimation of a benchmark model. The *MALES_20_24* and *MALES_25_29* variables, for instance, may well impact relative home advantage since it is these pools of the population where footballers must be selected from for the national team. If a nation possesses a relatively larger percentage of men aged between 20 and 29 than some other nation, then it demographically makes sense that that nation may have a larger pool of potential footballers at its disposal than the other. By the law of large numbers this would in expectation mean a relatively larger pool of high-quality footballers that may then positively impact the relative home advantage of that nation. On the other hand, these variables can serve as instruments for instrumental variable regression to deal with potential problems of endogeneity of some regressors. *INFL,* for instance, can well be expected to affect *UEFA* of the nation in question negatively. After all, large inflation makes it more expensive to invest and fund football in that nation, thus decreasing the relative home advantage through lower quality of the football system. *INFL* is in contrast not expected to directly affect other variables like *ART, CPI* and most importantly the dependent variable of relative home advantage, making it suitable as a potential instrument.

3.5 Descriptive statistics

To conclude this section some descriptive statistics will be provided to provide some first insights on the gathered data. The section below will mainly discuss scatter plots and correlations, which are obtained through the statistical program Excel, that are obtained from the data described above. Additional numerical descriptive statistics for the determined variables can be found in the appendix in Table A1, which will also be obtained through Excel.

Figure 1 below for example depicts the scatter plot between the CPI and the goal difference at home, the latter positively affecting relative home advantage as per Peeters and Van Ours (2021), over 7 editions of European FIFA World Cup qualifiers from 1996 until 2021 for which these data were available. There seems to be a weak yet present positive correlation between the two variables. Based on 310 observations this correlation is approximately 0.178. Thus, the first impression is that countries with lower levels of endemic corruption slightly tend towards having a larger goal difference at home and thus a potentially larger relative home advantage. However, correlation is not necessarily equal to causation. Momentarily though it will be presumed that the positive effect is more likely to work from the CPI to the goal difference at home than the other way around. In other words: less corruption is more likely to increase goal difference at home through for instance larger crowd support than an increased goal difference at home is to decrease corruption. However, it should be noted that relative home advantage is not just affected by the goal difference at home, that computation will also be done based on point difference as per Peeters and Van Ours (2021) and that more robust conclusions will be drawn in sections 5 and 6.

Furthermore, Figure 2 below depicts the same scatter plot but then between the UEFA association club coefficient and the goal difference at home over the same 7 editions of qualifiers investigated. Between these two variables there seems to be a reasonably strong positive correlation, an almost logarithmic shape even, between the two variables. The logarithmic shape is interesting in the context of decreasing marginal returns, that is when the investment level reaches some critical saturation point the goal difference at home does not increase any further or it increases negligibly. Based on 361 observations the correlation is approximately equal to 0.576, which is indeed rather strong. This could in theory be in accordance with earlier intuition that larger investments in football positively affect the relative home advantage in terms of goal difference because of extra opportunities for development of the sport. Reverse causality may be a relevant factor here however, because better performing teams may in return receive even more investments as a result. This paper attempts to deal with this problem through instrumental variable regression in section 5. This method will further be elaborated on in section 4.3 as well.

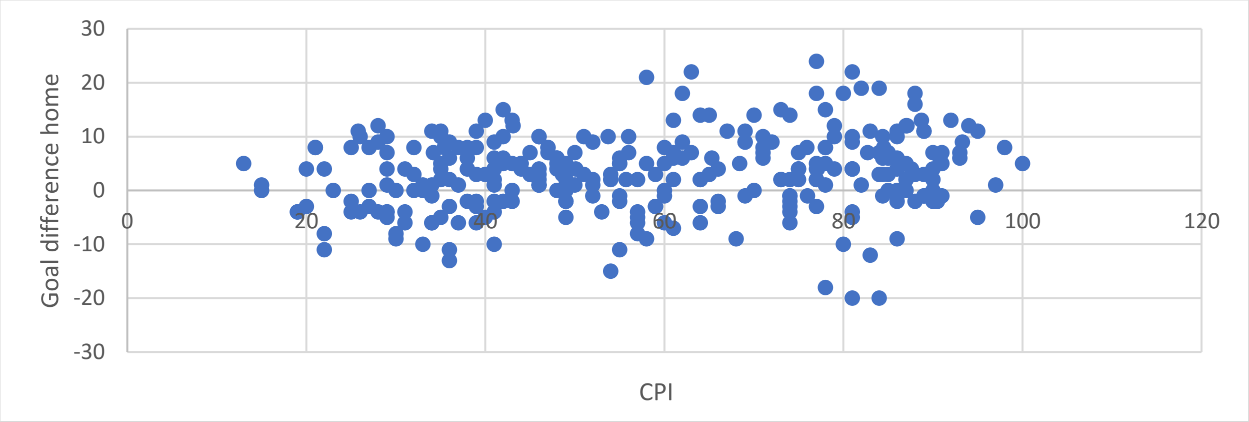

Figure 1 *Scatter plot of CPI against goal difference at home for seven editions of European FIFA World Cup qualifiers from 1996 to 2021*

Adapted sources: Transparency International (2023), Worldfootball.net (2023a), Worldfootball.net (2023b), Worldfootball.net (2023c), Worldfootball.net (2023d), Worldfootball.net (2023e), Worldfootball.net (2023f), Worldfootball.net (2023g)

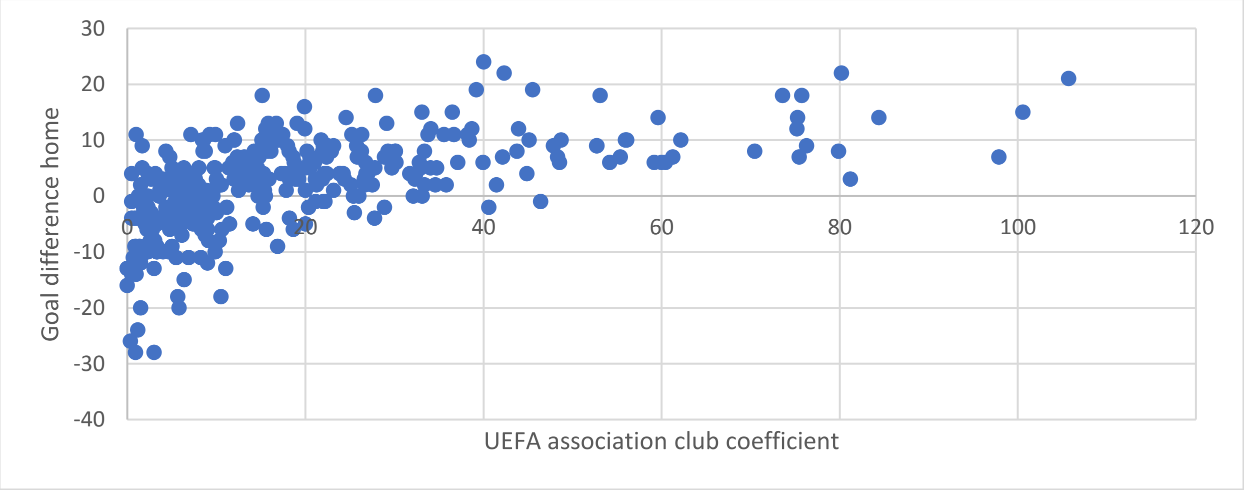

Figure 2 *Scatter plot of the UEFA association club coefficient against goal difference at home for seven editions of European FIFA World Cup qualifiers from 1996 to 2021*

Adapted sources: Kassies (2023), UEFA (2023a), Worldfootball.net (2023a), Worldfootball.net (2023b), Worldfootball.net (2023c), Worldfootball.net (2023d), Worldfootball.net (2023e), Worldfootball.net (2023f), Worldfootball.net (2023g)

Furthermore, Figure 3 below shows the scatter plot between the UEFA association club coefficient and CPI over the seven gathered editions of European FIFA World Cup qualifiers from 1996 to 2021. There one observes a weak yet somewhat positive correlation between the two variables. Based on 310 observations this correlation is about 0.129. This could theoretically imply that the larger CPI is the larger the UEFA association club coefficient becomes, which could imply that less corrupt countries are more likely to obtain funding for their football system due to for instance larger trust in the investment climate of that country. Obviously, one must again be aware that correlation does not imply causation. However, here for the moment the assumption will be made that it is more plausible for the positive effect to be present from CPI to the UEFA association club coefficient than the other way around. Put differently; less corruption is more likely to increase investment in football than an increase in the investment of football is to decrease corruption. On the contrary, increased investments in football may even be plausible to increase corruption since where there is more money present the more corruption is likely to be present as well. If collinearity between these two variables may indeed be an issue in the end, however, one would expect Ridge regression to deal with this issue. This method will be further elaborated on in section 4.4.

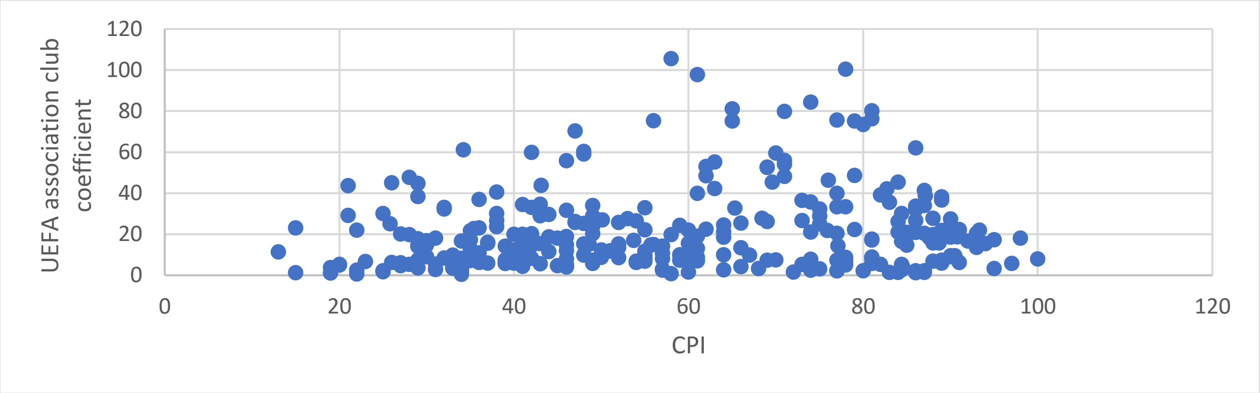

Figure 3 *Scatter plot of CPI against the UEFA association club coefficient for seven editions of European FIFA World Cup qualifiers from 1996 to 2021*

Adapted sources: Kassies (2023), Transparency International (2023), UEFA (2023a)

Finally, Figure 4 below shows the same scatter plot between CPI and the average home attendance. Again, a relatively moderate and positive correlation is visible between the two variables, which seems to be more pronounced than the one in Figure 3. Based on 308 observations this correlation is approximately equal to 0.319, which is indeed substantially larger than 0.130 for Figure 3. Put differently: this scatter plot may imply that indeed spectators tend to show up to support their home country more as the level of endemic corruption decreases and tend to stay away more if there is more corruption present. Once again, one must bear in mind that correlation does not imply causation. Once more though the assumption will momentarily be made that the positive effect is more plausible from CPI to average home attendance than vice versa. In other words: a dirty game is more likely to keep supporters away from the stadium than it is for a small attendance to suddenly corrupt the game.

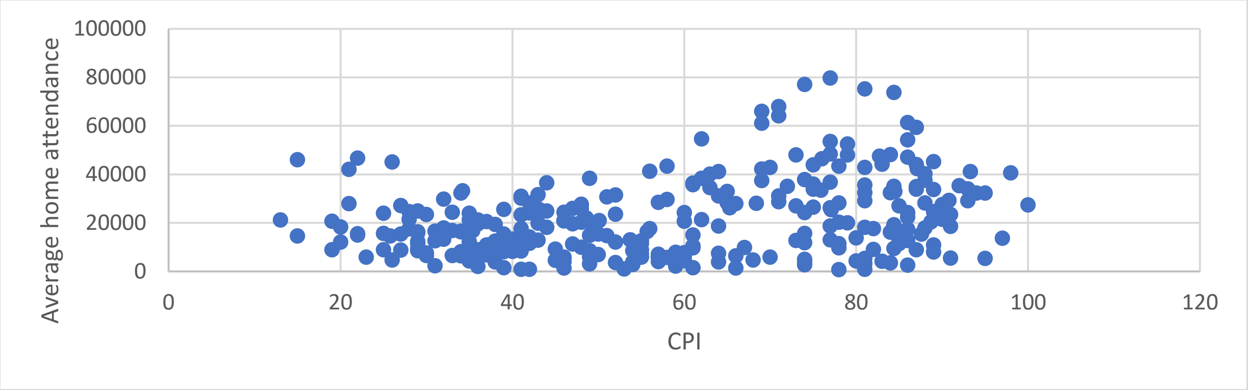

Figure 4 *Scatter plot of CPI against average home attendance for seven editions of European FIFA World Cup qualifiers from 1996 to 2021*

Adapted sources: Transparency International (2023), Wikipedia (2023a), Wikipedia (2023b), Wikipedia (2023c), Wikipedia (2023d), Wikipedia (2023e), Wikipedia (2023f), Wikipedia (2023g)

4 Methodology

This section will elaborate on the methodology used for conducting this research. Descriptive numerical statistics on all the variables introduced in sections 4.1 and 4.2 will be obtained using the statistical programming language R and can be found in appendix Table A2. Furthermore, the methods discussed below will also be implemented using the statistical programming language R, with a list of used R packages and their purpose displayed in appendix Table A3.

4.1 Computation of absolute and relative home advantage

Firstly, the methods employed by Peeters and Van Ours (2021) for the determination of absolute and relative home advantage will be used in this paper. In their paper Peeters and Van Ours (2021) explained a simple model for quantitatively expressing the absolute home advantage of teams participating in a round-robin competition, which in their case were the top four divisions in English professional football. Peeters and Van Ours (2021) ultimately derived that the absolute home advantage could be expressed in terms of the home and away point or goal difference in any given season and that this implied average seasonal performance depended on team quality and on absolute home advantage relative to the average home advantage enjoyed in the league in question that season. Since it is relative home advantage that served as the dependent variable in the paper by Peeters and Van Ours (2021) and since that will be the case in this paper as well, their methodology will be replicated to some extent in this paper.

Note that in this paper an adapted version of said measures is used since Peeters and Van Ours (2021) analyzed data that, apart from a small number of promoting and relegating teams each season, implied more or less homogeneity of involved clubs over time. Put differently: more or less the same clubs played each other over the seasons Peeters and Van Ours (2021) analyzed. In the European FIFA World Cup qualifiers, on the other hand, the participating nations are randomly divided into several groups each edition, thus randomization of which nation encounters which other nations occurs over time. To account for these groups the variables for each nation in each edition will be computed solely using the data for the group that nation belonged to in that particular edition. This also makes sense intuitively. After all, it would not be relevant to consider opponents from other groups that the nation in question did not play. Peeters and Van Ours (2021) also based their definition of relative home advantage as the absolute home advantage minus the average absolute home advantage enjoyed in either of the 4 English football leagues investigated. Thus, Peeters and Van Ours (2021) intuitively also only defined the opponents played by the team in question as relevant for the computation of the relative home advantage. Note that this paper also makes the distinction between these methods when computed using the point difference method and the goal difference method like Peeters and Van Ours (2021) did as well.

Peeters and Van Ours (2021) proceeded as follows in calculating the absolute home advantage using the goal difference method. Let N_{jg} be the number of nations in group g in edition j. Furthermore, for any j and g define the following variable:

$$GROUPGDH_{jg} = \frac{1}{N_{jg} - 1} \sum_{i=1}^{N_{jg}} GDH_{ijg} \tag{1}$$

as the scaled average of the aggregate home goal difference of all nations belonging to group g in edition j, which is what Peeters and Van Ours (2021) used. Ultimately, similarly to Peeters and Van Ours (2021), the absolute home advantage of nation i in edition j in group g using the goal difference method, henceforth defined as $AHAG_{ijg}$, can be defined as follows:

$$\frac{GDH_{ijg} - GDA_{ijg} - GROUPGDH_{jg}}{N_{jg} - 2} \tag{2}$$

Similarly to Peeters and Van Ours (2021) one can then define the average absolute home advantage in edition j of group g using the goal difference method, henceforth denoted by $GROUPAHAG_{jg}$, as follows:

$$\frac{1}{N_{jg}} \sum_{i=1}^{N_{jg}} AHAG_{ijg} \tag{3}$$

Using this notation, one then, similarly to Peeters and Van Ours (2021), finally arrives at the relative home advantage of nation i in edition j in group g using the goal difference method, henceforth denoted by *RELAHAG$_{ijg}$,* as follows:

$$AHAG_{ijg} - GROUPAHAG_{jg} \tag{4}$$

The calculation using the point difference method by Peeters and Van Ours (2021) goes similarly. Here, however, one must be careful about how the point difference, both at home and away, of nation i in edition j in group g is calculated. Similarly to Peeters and Van Ours (2021), this variable, from now on denoted by *PDH$_{ijg}$*, is calculated as follows:

$$3HW_{ijg} - 3HL_{ijg} \tag{5}$$

Note that *HD$_{ijg}$* immediately drops out of formula (5) since in the case of a draw the point difference would be 0. The computation of the away point difference, which will be denoted by *PDA$_{ijg}$,* is similar to formula (5). Similarly to *TGD$_{ijg}$,* the total point difference of nation i in edition j in group g, from now on denoted by *TPD$_{ijg}$,* is obtained by summing up *PDH$_{ijg}$* and *PDA$_{ijg}$*. The scaled average aggregate home point difference for group g in edition g, denoted by *GROUPPDH$_{jg}$,* can similarly to equation (1) be defined as follows:

$$\frac{1}{N_{jg} - 1} \sum_{i=1}^{N_{jg}} PDH_{ijg} \tag{6}$$

This leads to the absolute home advantage of nation i in edition j in group g using the point difference method, denoted by *AHAP$_{ijg}$,* similarly to formula (2) being defined as follows:

$$\frac{PDH_{ijg} - PDA_{ijg} - GROUPPDH_{jg}}{N_{jg} - 2} \tag{7}$$

Likewise to the computation in formula (3) the average absolute home advantage in edition j of group g using the point difference method, denoted by *GROUPAHAP$_{jg}$,* is then given by the following:

$$\frac{1}{N_{jg}} \sum_{i=1}^{N_{jg}} AHAP_{ijg} \tag{8}$$

Finally, one then in a comparable fashion to formula (4) gets the relative home advantage of nation i in edition j in group g using the point difference method, defined as *RELAHAP$_{ijg}$,* as follows:

$$AHAP_{ijg} - GROUPAHAP_{jg} \tag{9}$$

In the end, comparable to Peeters and Van Ours (2021), the variables computed by formulas (4) and (9) will be used as the dependent variables for the analysis in this paper. A final remark is that all abbreviations previously introduced and their exact definitions can be found in the appendix in Table A5.

4.2 Transformation of the regressors

This section will explain what regressors will be used as the independent variables in the main analyses and how to obtain them. In general, yet with one exception, this paper will work with relative regressors comparably to Peeters and Van Ours (2021). Put differently; all but one regressor will be defined as the value for nation i in edition j in group g relative to its opponents in edition j in group g.

The first regressor that will be used is ART_{ijg} as was also done by Peeters and Van Ours (2021). This variable will mainly be used for constructing a benchmark model and for robustness checks. This variable will be used as defined in section 4 and thus no further transformation of this variable is required.

The second regressor that will be used is the relative home attendance, which Peeters and Van Ours (2021) also used in their analysis. This variable will be one of the pivotal regressors used in this analysis. Similarly to Peeters and Van Ours (2021), this variable, from now on denoted by $RELAVHATT_{ijg}$ for nation i in edition j in group g, is given by the following formula:

$$\log\left(\frac{AVHATT_{ijg}}{GROUPAVHATT_{jg}}\right) \tag{10}$$

Here $GROUPAVHATT_{jg}$ is defined as the average of *AVHATT* for group g in edition j. Note that in this paper thus a similar logarithmic transformation as in Peeters and Van Ours (2021) will be used, which helps reduce the large scale of home attendance figures.

The third regressor will be the relative corruption perceptions index, which will be one of the main independent variables in the empirical analysis. It will be denoted by $RELCPI_{ijg}$ for nation i in edition j in group g and will, again using a logarithmic transformation and with $GROUPCPI_{jg}$ being defined as the average of *CPI* for group g in edition j, be defined as follows:

$$\log\left(\frac{CPI_{ijg}}{GROUPCPI_{jg}}\right) \tag{11}$$

Similarly, the relative UEFA association club coefficient, which will be another main independent variable in the empirical analysis, is denoted by *RELUEFA*$_{ijg}$ for nation i in edition j in group g and is given by the following formula:

$$\log\left(\frac{UEFA_{ijg}}{GROUPUEFA_{jg}}\right) \tag{12}$$

Finally, the relative inflation percentage, relative percentage of males aged 20-24 and the relative percentage of males aged 25-29 from the male population will be denoted by *RELINFL*$_{ijg}$*, RELMALES_20_24*$_{ijg}$ *and RELMALES_25_29*$_{ijg}$ respectively for nation i in edition j in group g. Those will be used for benchmarking and robustness purposes, but also as instrumental variables for the IV-regressions. They are obtained by transformations (13), (14) and (15) below respectively using a similar logic as in formulas (10), (11) and (12):

$$\log\left(\frac{INFL_{ijg}}{GROUPINFL_{jg}}\right) \tag{13}$$

$$\log\left(\frac{MALES\_20\_24_{ijg}}{GROUPMALES\_20\_24_{jg}}\right) \tag{14}$$

$$\log\left(\frac{MALES\_25\_29_{ijg}}{GROUPMALES\_25\_29_{jg}}\right) \tag{15}$$

Two final general remarks are that all logarithmic transformations above use the natural logarithm and that all abbreviations introduced and their exact definitions can be retrieved from the appendix in Table A5.

4.3 Ordinary least squares and instrumental variable regression

To conduct this research several econometric techniques will be used. The first technique is classical ordinary least squares (OLS) regression. This method will be used to replicate the research by Peeters and Van Ours (2021) since this was also the method that they used in their research to investigate whether some English football clubs enjoyed a relatively larger home advantage than others. This way replicability of the original research is tested. In total 14 different OLS models will be estimated in the usual linear form with intercept:

$$y_i = \alpha + \sum_{j=1}^{k} x_{ij}\beta_j + \varepsilon_i \tag{16}$$

The first 7 OLS models will set the dependent variable *y_i* to be the relative home advantage using the goal difference method. The first model will include relative inflation, the relative proportion of males aged 20 to 24, the relative proportion of males aged 25 to 29 and the artificial pitch indicator as regressors and will be used as a benchmark model. The second model only includes the relative CPI as a regressor, while the third model augments the second model using the benchmark variables from the first model as a robustness check. These two models will help determine the relationship between relative corruption and the relative home advantage. The fourth model includes the relative CPI and the relative home attendance as regressors, with the fifth model again augmenting the previous model using the previously introduced benchmark variables as a robustness check. These two models will help determine whether there is an effect of relative corruption on relative home advantage that is channeled through relative home attendance. Finally, the sixth model will include the relative CPI and the relative UEFA association club coefficient as regressors, with the seventh model again augmenting the previous model for robustness purposes. These final two models will help determine whether there is an effect of relative corruption on relative home advantage that is channeled through the relative UEFA association club coefficient. The final 7 models are essentially the same, with the only difference being that the dependent variable is the relative home advantage using the point difference method instead.

Moreover, instrumental variable (IV) regression will be employed to address potential issues of endogeneity with some regressors. IV-regression essentially extends equation (16) by using instruments on regressors that are either suspected to be endogenous, thus correlated with the error term, or be involved in reverse causality, that is the regressor affects the dependent variable which in return affects the regressor. Peeters and Van Ours (2021) for example claimed reverse causality between relative home attendance and relative home advantage might, although unlikely, be a concern. Furthermore, the relative CPI might well be endogenous too. This is because the relative CPI is likely to affect consumer confidence; more corrupt countries are plausible to enjoy less consumer confidence. Less consumer confidence generally means people are more reluctant to spend money, including on football merchandise. Put differently; less consumer confidence might negatively affect revenues, which ultimately may affect the relative home advantage negatively since there will be fewer financial means available to improve the football infrastructure. The main message of this intuition is that relative CPI might affect the relative home advantage through consumer confidence, which means relative CPI is suspected to be correlated with the error term and thus endogenous since consumer confidence is an omitted variable. Finally, the relative UEFA association club coefficient might be involved in reverse causality. More investments in the football system can be expected to lead to a larger relative home advantage, which in return might attract more investments. This paper will thus go more in depth on this by identifying and using plausible instruments. Such instruments must fulfill three basic conditions. Firstly, an instrument may itself not be correlated with the error term. Secondly, the instrument may not affect the dependent variable directly. Finally, the instrument must be correlated with the potentially endogenous regressor. A significant drawback of using IV-estimation is that there is no truly reliable way to test whether the aforementioned conditions hold, which means the chosen instruments and suspicion of endogeneity or reverse causality will largely be based on intuition.

4.4 Ridge regression

Finally, the machine learning technique of Ridge regression will be used. There are two reasons for this choice. Firstly, this method is a form of penalized regression, which has been developed to deal among others with potential collinearity between two or more regressors in the regression. In this case this collinearity is plausible. Think for instance about the UEFA association club coefficient and attendance; a larger coefficient might be an indication of better-quality football and more attention and financing of the sport, which might attract larger stadium crowds to cheer their teams on. Another example is potential collinearity between the CPI and the UEFA association club coefficient. After all, since one of the goals of this paper is to investigate whether corruption in the European qualifiers channels through the UEFA association club coefficient we implicitly suspect some collinearity between these two variables. Ridge regression addresses this by shrinking a subset of the corresponding variable coefficients in the regression towards 0 to reduce variance at the expense of interpretability, whilst never actually imposing sparsity on any of the coefficients. This also summarizes the second reason for the use of Ridge regression; this method besides variance reduction also, to some extent, applies variable selection to the model. If corruption, whether by its own, through UEFA or both, indeed plays a role in the relative home advantage of certain nations, one would expect the machine learning method to keep the CPI, UEFA association club coefficient or both reasonably large in magnitude compared to other regressors.

The general notation of the loss function associated with Ridge regression is an extension of the classical sum of squared residuals on which OLS estimation is based. It is given by the following loss function (Statistics How To, 2023):

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\,\beta_j)^2 + \lambda \sum_{j}\left|\beta_j\right|^2 \tag{17}$$

Clearly this loss function is composed of two terms. The first term is the classical sum of squared residuals associated with OLS estimation. The second term consists of the squared norm of the coefficient associated with the jth regressor multiplied by some nonnegative tuning parameter λ. Note that if λ = 0 one has classical OLS regression, which in practice means λ will be positive. The squared norm ensures shrinkage towards 0 of certain coefficients occurs, whilst never actually imposing sparsity on any given coefficient. Looking at formula (17) this makes sense since this function is clearly monotonically increasing in the second term. Since the goal is to minimize the loss function coefficients that are lower in magnitude are always preferred due to this penalization term, while maintaining the tradeoff ensuing jointly with the first term in formula (17). The nonnegative tuning parameter λ determines how harshly the coefficients are penalized. This intuitively again makes sense when looking at formula (17); the larger λ is the better it is to set the coefficients lower in magnitude in general, while maintaining the tradeoff jointly determined together with the first term in formula (17).

An obvious question that arises then is how to set the tuning parameter λ. To this end a procedure called k-fold cross-validation will be used. This procedure implies splitting the entire sample of data into a training and a test sample. K-fold cross-validation then implies splitting the training sample into k random subsamples. K-1 of those subsamples will be used to fit the model on while the remaining subsample will be used for evaluating the fitted model. This procedure is repeated k times and results in a prediction score being obtained at the end for a given value of λ (Kangralkar, 2021). We repeat this procedure for a randomly generated grid of 100 values for λ and ultimately choose the value for λ that results in the best prediction score (Malfait, 2021).

5 Results

The section below will present and elaborate on the results obtained. These results were obtained using the statistical programming language R, for which the used packages and their purposes are given in the appendix in Table A3. During the transformation of the data as described in sections 4.1 and 4.2, two important obstacles came to attention. Firstly, in some instances values for all the variable categories used to compute the regressors, except for *MALES_25_29,* were missing. For those variables in those instances, the means in the denominators used for the logarithmic transformation were computed using the remaining values that were available. The observations for which values were missing were ultimately cleaned from the available observations, reducing the sample size from 365 observations to 286 observations. Thus 79 observations were discarded to avoid complications with missing data. Secondly, in some editions certain countries experienced a negative inflation rate, with the mean of the inflation rate in the groups those countries belonged to still being positive since negative inflation is rather rare in general. This led to some nonreal values emerging when computing *RELINFL* using formula (13) since the logarithm of a negative value is nonreal. Thus, observations for which such nonreal values for *RELINFL* were present were discarded as well, contributing to the reduction in sample size from 365 observations to 286 observations, which is still a reasonable sample size.

5.1 OLS regression

Table 2 below shows the OLS regression results for a total of 7 models with *RELAHAG* as the dependent variable. To have some insight into the variance explained by each model, both the R^2 and adjusted R^2 are reported. The reason for reporting the latter is that the regular R^2 never decreases as more variables are added to the model. To account for this when estimating augmented models for robustness checks the adjusted R^2 is rather useful since this measure penalizes the regular R^2 in the number of regressors.

The first model is a naïve benchmark model intended for comparison purposes containing all regressors save for the three main regressors, namely *RELCPI, RELAVHATT* and *RELUEFA.* The variation this model explains, both in terms of the R^2 and adjusted R^2, is low, and none of the coefficients are statistically significant. This is the case for the other six models as well, which also never perform better in terms of the R^2 and adjusted R^2 compared to the benchmark model. Some interesting insights to note nonetheless is the quite large positive coefficient for the proportion of males aged 20-24 for the benchmark model, which suggests a larger such proportion may positively affect the relative home advantage through a larger pool of potential young talents for the national football team. Surprisingly, this is not the case for the proportion of males aged 25-29. An explanation for this is unknown. Finally, in line with the findings of Peeters and Van Ours (2021), playing on an artificial pitch seems to increase the relative home advantage.

Looking at the second model, one observes a negative though statistically insignificant relationship between *RELCPI* and *RELAHAG*. Put differently; more corruption increases the relative home advantage computed using the goal difference method. This finding disputes the first impression discussed in section 3.5. Apparently, corrupt officials rigging matches dominates the negative effect corruption may have on the size of the home crowd supporting their squad. This relationship is robust since the sign remains the same when including the benchmark variables in the third model, with the standard error of *RELCPI* staying about the same and the adjusted R^2 of the second model being even slightly larger. However, the coefficient of *RELCPI* largely decreases in magnitude. Note that the signs and standard errors of the benchmark variables also stay largely the same in the third model compared to the first model but also their magnitudes.

In the fourth model one observes a negative coefficient for *RELCPI* that is slightly smaller in magnitude and a negative coefficient for *RELAVHATT*. This implies, considering the results in the second model, that there is some positive effect of the relative level of corruption on the relative home attendance with the overall effect of relative corruption channeled through relative home attendance thus being negative. In other words: less corruption is expected to increase relative home attendance, which in turn is expected to decrease the relative home advantage. This is partially in line with the first impression discussed in section 3.5 and is even more interesting considering Peeters and Van Ours (2021) found the effect of relative home attendance on relative home advantage to be positive and oftentimes statistically significant. Perhaps some less strong national football teams still experience large crowds through for instance sentiments of nationalism, which in turn puts excessive pressure on those teams to perform, leading to a smaller instead of a larger relative home advantage. That less corruption seems to increase the crowds does not seem surprising though as it is plausible that supporters are more eager to show up if the game is clean. The determined relationship is once again robust, with all magnitudes, signs and standard errors staying roughly the same in the fifth model except for again a decreased magnitude of the coefficient of *RELCPI.* Also, the adjusted R^2 is slightly larger for the fourth model compared to the fifth model.

Finally, when looking at the sixth model one observes, relative to the second model, a negative coefficient for *RELCPI* that is slightly larger in magnitude and a positive coefficient for *RELUEFA.* This implies some positive effect of the relative CPI on the relative UEFA coefficient with the overall effect channeled through to the relative home advantage being positive. Thus, less corruption is expected to increase investments in the national football system, which in turn is expected to increase relative home advantage. This is completely in line with section 3.5. Logically, investors like to invest in sports in countries that are relatively clean of corruption as otherwise they risk losing their investments in the pockets of corrupt officials. More investments are expected to breed more high-quality players, improving relative home advantage. The relationship is no longer robust though looking at the seventh model and noticing the largely increased coefficient for *RELCPI* and the increased coefficient for *RELUEFA.* Also, the adjusted R^2 is somewhat smaller for the sixth model compared to the seventh model.

Table 2 OLS regression results with *RELAHAG* as the dependent variable

Variable	RELAHAG						
Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)
RELCPI	-	-0.17	-0.03	-0.15	-0.02	-0.19	-0.02
		(0.16)	(0.21)	(0.17)	(0.21)	(0.17)	(0.21)
RELAVHATT	-	-	-	-0.05	-0.03	-	-
				(0.09)	(0.09)		
RELUEFA	-	-	-	-	-	0.06	0.09
						(0.06)	(0.07)
RELINFL	0.06	-	0.05	-	0.05	-	0.05
	(0.07)		(0.07)		(0.07)		(0.07)
RELMALES_20_24	0.85	-	0.82	-	0.81	-	1.02
	(0.68)		(0.71)		(0.71)		(0.72)
RELMALES_25_29	-0.71	-	-0.72	-	-0.75	-	-0.75
	(0.77)		(0.78)		(0.78)		(0.78)
ART	0.40	-	0.40	-	0.38	-	0.44
	(0.37)		(0.39)		(0.39)		(0.39)
R^2	0.016	0.004	0.016	0.005	0.017	0.007	0.022
Adjusted R^2	0.002	0.000	-0.001	-0.002	-0.004	-0.000	0.001

Note. The intercept, although included in the estimation, is not reported; standard errors are in parentheses; * indicates significance at the 10% level; ** indicates significance at the 5% level; *** indicates significance at the 1% level; all results are reported in two decimal places except for the (adjusted) R^2.

Adapted sources: Kassies (2023), The World Bank (2023a), The World Bank (2023b), The World Bank (2023c), Transparency International (2023), UEFA (2023a), Wikipedia (2023a), Wikipedia (2023b), Wikipedia (2023c), Wikipedia (2023d), Wikipedia (2023e), Wikipedia (2023f), Wikipedia (2023g), Worldfootball.net (2023a), Worldfootball.net (2023b), Worldfootball.net (2023c), Worldfootball.net (2023d), Worldfootball.net (2023e), Worldfootball.net (2023f), Worldfootball.net (2023g)

Table 3 below shows the same OLS models with *RELAHAP* as the dependent variable. The results for the first, second, fourth and sixth models stay roughly the same. A notable difference is that this time the relationships between the relative level of corruption and the relative home advantage are no longer robust since the coefficients of *RELCPI* are now positive in the third, fifth and seventh models. Moreover, the adjusted R^2 values are always lower for the initial models than for the robustness check models, further confirming absence of robustness. Finally, the coefficient for *RELINFL* is now statistically significant at the 5% level. An explanation for this is unknown, though the positive coefficient may indicate for instance more revenues through increased ticket prices which can then be used to improve footballing infrastructure to increase the relative home advantage for the better.

Table 3 OLS regression results with *RELAHAP* as the dependent variable

Variable	RELAHAP						
Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)
RELCPI	-	-0.17 (0.20)	0.25 (0.26)	-0.16 (0.21)	0.24 (0.26)	-0.18 (0.21)	0.25 (0.25)
RELAVHATT	-	-	-	-0.03 (0.11)	0.02 (0.11)	-	-
RELUEFA	-	-	-	-	-	0.03 (0.08)	0.06 (0.08)
RELINFL	0.21** (0.08)	-	0.24*** (0.09)	-	0.24*** (0.09)	-	0.24*** (0.09)
RELMALES_20_24	0.75 (0.83)	-	0.99 (0.87)	-	1.00 (0.87)	-	1.14 (0.89)
RELMALES_25_29	-0.82 (0.95)	-	-0.76 (0.95)	-	-0.75 (0.96)	-	-0.78 (0.95)
ART	0.18 (0.47)	-	0.16 (0.47)	-	0.17 (0.48)	-	0.20 (0.48)
R^2	0.033	0.002	0.036	0.003	0.036	0.003	0.038
Adjusted R^2	0.018	-0.001	0.019	-0.004	0.015	-0.004	0.017

Note. The intercept, although included in the estimation, is not reported; standard errors are in parentheses; * indicates significance at the 10% level; ** indicates significance at the 5% level; *** indicates significance at the 1% level; all results are reported in two decimal places except for the (adjusted) R^2.

Adapted sources: Kassies (2023), The World Bank (2023a), The World Bank (2023b), The World Bank (2023c), Transparency International (2023), UEFA (2023a), Wikipedia (2023a), Wikipedia (2023b), Wikipedia (2023c), Wikipedia (2023d), Wikipedia (2023e), Wikipedia (2023f), Wikipedia (2023g), Worldfootball.net (2023a), Worldfootball.net (2023b), Worldfootball.net (2023c), Worldfootball.net (2023d), Worldfootball.net (2023e), Worldfootball.net (2023f), Worldfootball.net (2023g)

5.2 Ridge regression

Table 4 below shows the Ridge regression results for two models estimated, one with *RELAHAG* as the dependent variable and the other with *RELAHAP* as the dependent variable. The specific settings used for the K-fold cross validation and the resulting tuning parameter λ used for both models are given in the appendix in Table A4. Important to note is that the K-fold cross validation was done using a randomly generated grid of 100 values for λ, each between 0 and 100. For the Ridge regressions all regressors were included in the model, essentially letting the machine learning technique decide what coefficients would be penalized and how severely and thus what variables would matter more for model accuracy.

Looking at the results below it becomes clear that these are largely in line with the relationships determined in Tables 2. This is because the signs of all regressors remained the same. An exception to this is *RELMALES_25_29*, which is now positive though severely shrunken in magnitude. In both models the most weight is left on *RELMALES_20_24* and *ART*, with the remaining coefficients largely shrunken. This is in line with the fact that almost none of the coefficients established in Tables 2 and 3 are statistically significant and thus that the OLS models have little explanatory power.

Table 4 Ridge regression results for both dependent variables

Variable	RELAHAG	RELAHAP
RELCPI	-0.02	-0.01
RELAVHATT	-0.01	-0.01
RELUEFA	0.01	0.01
RELINFL	0.01	0.06
RELMALES_20_24	0.07	0.21
RELMALES_25_29	0.01	0.03
ART	0.06	0.09

Note. The intercept, although included in the estimation, is not reported; standard errors and the R^2 are not reported since those are more complex to obtain for Ridge regression due to the maximum likelihood nature of the estimation; all results are reported in two decimal places.

Adapted sources: Kassies (2023), The World Bank (2023a), The World Bank (2023b), The World Bank (2023c), Transparency International (2023), UEFA (2023a), Wikipedia (2023a), Wikipedia (2023b), Wikipedia (2023c), Wikipedia (2023d), Wikipedia (2023e), Wikipedia (2023f), Wikipedia (2023g), Worldfootball.net (2023a), Worldfootball.net (2023b), Worldfootball.net (2023c), Worldfootball.net (2023d), Worldfootball.net (2023e), Worldfootball.net (2023f), Worldfootball.net (2023g)

5.3 IV-regression

Finally, Table 5 below shows the IV regression results for three models estimated using both *RELAHAG* and *RELAHAP.* Those models are clearly analogous to the second, fourth and sixth models estimated in Tables 2 and 3. Firstly, it was opted to choose *RELINFL* as the instrument for *RELCPI.* This is because this variable is thought to be correlated with *RELCPI*; severe inflation is typical in countries with more corruption. It is furthermore plausible that this variable does not directly affect the other regressors *ART, MALES_20_24, MALES_25_29* and *RELAVHATT* or either dependent variable directly. A potential problem might arise with the regressor *RELUEFA*, as investments are generally inversely correlated with inflation. However, due to the nature of this variable as a proxy variable rather than actual investment figures the assumption will be made that this will not pose a too serious problem.

Secondly, *MALES_25_29* was chosen as the instrument for *RELAVHATT.* It is plausible that this instrument is correlated with this regressor since a larger population of men aged 25-29 can be expected to increase home attendance since it is generally young people who enjoy attending matches in person. This variable is furthermore unlikely to be correlated with *MALES_20_24*, since a potential effect is more likely to work the other way around; the ‘past’ generation is more likely to affect the ‘future’ generation.

Any correlation with the remaining regressors is unlikely and this instrument is unlikely to have a direct effect on either of the dependent variables.

Thirdly, *ART* was used as the instrument for *RELUEFA.* After all, changing from an artificial to a natural pitch and vice versa is indicative of some investment that was made to implement this change. Thus, the case is made for a correlation between the two here. While *ART* may directly affect the relative home advantage as Peeters and Van Ours (2021) statistically significantly determined, for the moment this concern will be omitted since in the previous tables the determined coefficients for this variable were not statistically significant. Finally, it is not plausible that this instrument is correlated with any remaining regressors.

The sign of *RELCPI* remained negative in both IV regressions and thus the results found using OLS and Ridge regression are confirmed. However, now the coefficients are statistically significant twice when using the point difference method, though not when using the goal difference method. The Hausman statistics for testing endogeneity are highly statistically significant for the point difference method and not statistically significant at all for the goal difference method. This may thus explain why for the point difference method relative corruption is statistically significant but not for the goal difference method since apparently instruments are appropriate for the former but not for the latter. For both methods the null hypothesis of *RELINFL* being a weak instrument is overwhelmingly rejected indicating this is a reasonable instrument to use.

Interestingly, for both dependent variables this time the sign of *RELAVHATT* is positive though still statistically insignificant. Thus, this time the findings would be in accordance with Peeters and Van Ours (2021), who determined relative home attendance positively affects relative home advantage, and with earlier intuition from section 3.5. The instruments used once again seem reasonably strong and endogeneity is overwhelmingly accepted for the point difference method and rejected for the goal difference method. These results contradict earlier findings determined using OLS and Ridge regression, since this time the overall effect of the relative CPI through relative home attendance would be positive. Still in line with earlier findings is that less corruption seems to attract larger home crowds.

Finally, the coefficient for *RELUEFA* is now negative although still statistically insignificant. Once again, the instrument used seems reasonably strong and endogeneity is overwhelmingly accepted for the point difference method yet again rejected for the goal difference method. This result rejects the earlier findings using OLS and Ridge regression since the overall effect of relative corruption on relative home attendance channeled through the relative UEFA association club coefficient is now negative. In line with earlier results though is that less corruption still seems to increase investments in football.

Table 5 IV regression results for both dependent variables

Variable	RELAHAG			RELAHAP		
Model	(1)	(2)	(3)	(1)	(2)	(3)
RELCPI	-0.44	-1.06	-0.12	-1.16***	-2.95	-1.02*
	(0.33)	(1.22)	(0.54)	(0.42)	(2.17)	(0.48)
RELAVHATT	-	0.78	-	-	2.25	-
		(1.22)			(2.22)	
RELUEFA	-	-	-0.71	-	-	-0.31
			(0.80)			(0.86)
Weak instruments (RELINFL)	97.60***	70.97***	48.63***	97.60***	70.97***	48.63***
Weak instruments (RELMALES_25_29)	-	13.52***	-	-	13.52***	-
Weak instruments (ART)	-	-	2.57*	-	-	2.57*
Hausman	0.88	0.44	1.20	8.33***	4.11**	4.30**

Note. The intercept, although included in the estimation, is not reported; standard errors are in parentheses; * indicates significance at the 10% level; ** indicates significance at the 5% level; *** indicates significance at the 1% level; all results are reported in two decimal places.

Adapted sources: Kassies (2023), The World Bank (2023a), The World Bank (2023b), The World Bank (2023c), Transparency International (2023), UEFA (2023a), Wikipedia (2023a), Wikipedia (2023b), Wikipedia (2023c), Wikipedia (2023d), Wikipedia (2023e), Wikipedia (2023f), Wikipedia (2023g), Worldfootball.net (2023a), Worldfootball.net (2023b), Worldfootball.net (2023c), Worldfootball.net (2023d), Worldfootball.net (2023e), Worldfootball.net (2023f), Worldfootball.net (2023g)

6 Conclusion and discussion

The goal of this paper was to quantitatively investigate whether corruption is a relevant determinant of the relative home advantage nations enjoy in the UEFA qualifiers for the FIFA World Cup. The first sub question formulated to answer this main research question was whether the CPI is a relevant determinant of said relative home advantage. The conclusion of this paper concerning this sub question is that this relationship is generally negative yet, except in three situations, never statistically significant with robustness also being questionable sometimes. Put differently; more corrupt countries tend, though insignificantly, to have a larger relative home advantage than cleaner countries. A definitive explanation of this negative effect remains unknown, though a plausible explanation may be that more corrupt countries tend to have officials involved in organizing or refereeing games more tightly in their grip and may thus be more likely to obtain an illicit home advantage over their opponent than cleaner countries are. Apparently, this effect dominates the intuitively larger home crowd showing up to support their national team and thereby increasing the relative home advantage, though more robust findings on the latter effect were conflicting in this research.

The second sub question formulated was whether the effect of the CPI on relative home advantage is in some way channeled through the relative home attendance. As far as the effect from the CPI to relative home attendance itself is concerned, the general conclusion is that cleaner countries tend to attract relatively larger home crowds. Intuitively this makes sense since people are expected to show up more when the game of their squad is perceived to be fair and entertaining rather than when it is plagued by unfairness. When it comes to the subsequent effect of relative home attendance on the relative home advantage the conclusion remains undetermined. Depending on the model investigated, this effect is either found to be positive or negative and never statistically significant. Were this effect to be positive then this would mean cleaner countries tend to attract larger home crowds which would in turn increase the relative home advantage, which intuitively would make sense since home crowds are generally considered to be the 'twelfth' player. Were this effect to be negative, on the other hand, this would mean cleaner countries still tend to attract larger home crowds, which would then negatively affect the relative home advantage in turn. A definitive explanation for this remains unknown, though it may be because of relatively small football nations still attracting larger home crowds through a sense of mere nationalism, which in turn puts excessive pressure on the players to perform, which may cause them to perform worse.

The third sub question was whether the effect of the CPI on relative home advantage is in some way channeled through the UEFA association club coefficient. The conclusion of this effect is inconclusive, though there is consensus that less corrupt countries tend to enjoy a larger relative UEFA association club coefficient. Given its role as a proxy variable for investments in football this is logical since less corrupt countries generally tend to attract more investors financing the football system.

The overall conclusion of this paper is thus that more corrupt countries tend to enjoy a larger relative home advantage than cleaner countries, that the effect of corruption on relative home advantage may have multiple sides to it and that the effects are not statistically significant or robust. It is ethical and appropriate to also discuss the limitations of this research. First and foremost, corruption is part of the underground economy, making actual figures on corruption difficult if not impossible to obtain. The CPI, which was used as a proxy, has limitations of its own. Like the name says, this index provides merely the perceptions of relevant experts in the field of business on whether corruption is a problem in the country in question or not, not actual corruption figures. Those perceptions may in turn be affected by, for instance, biases or the extent of freedom of speech and the press in the country in question. The CPI also provides the impression of corruption in the country in question in general, not specifically in football, though the two are likely to be positively correlated as corruption is oftentimes embedded in a multitude of layers in society. The second limitation is that plausible instruments are often hard if at all possible to determine since there is practically no truly trustworthy way to test whether an instrument is valid. While the weak instrument test results were given, this test merely says something on whether an instrument is strong enough rather than actually valid. Thus, the instruments selected were largely based on intuition, making the IV regression results somewhat questionable. Furthermore, it was implicitly assumed that corruption only affected the home team, which may be questionable in practice since it is away teams that may also appeal to obtaining illegal advantages since they have the disadvantage of playing away. Finally, the sample size was substantially reduced in magnitude compared to Peeters and Van Ours (2021). While the seven editions of qualifiers investigated initially resulted in 365 observations, 79 of those had to be discarded due to missing or nonreal values, leaving the final sample size at 286 observations. While this is still reasonably large, more observations would have been preferable regardless.

Lastly, some suggestions will be given for future research. For instance, actual figures on investments made in national football systems could be obtained. Their relationship to the CPI could then be robustly analyzed, leading to potentially more robust conclusions on whether the perception of corruption as a societal problem actually affects investments in football and how those investments then actually affect the relative home advantage. While the CPI as mentioned before will still have the mentioned limitations, it would be a substantial improvement regardless. Secondly, this analysis may be performed on qualifiers from other confederations which may have more data and thus more observations available to perform the quantitative analysis. This may lead to more robust conclusions. Moreover, the same setup may be investigated, not implicitly assuming that corruption only affects the home team. Furthermore, some regression technique that is not entirely dependent on determining plausible instruments may be employed, such as a logistic regression of the probability of enjoying a larger relative home advantage on several relevant types of data. These may be but are not limited to the data used in this paper. Finally, research could be done whether corruption increases the post-World Cup pay of players for the countries in question relative to cleaner countries that missed qualification for the respective World Cup, thereby increasing income inequality amongst players. What is certain is that this is a topic that is relevant to explore further, for if one wants to create a cleaner football world with equal opportunities for all, corruption should be addressed wherever possible and necessary.

References

Aidt, T. (2009). *Corruption, Institutions and Economic Development.* Retrieved from https://api.repository.cam.ac.uk/server/api/core/bitstreams/6b699908-4e0f-4992-8abc-3f3eab03fa70/content

Amenta, C. and Di Betta, P. (2021). *The impact of corruption on sport demand.* Retrieved from https://doi.org/10.1108/IJSMS-01-2020-0004

Boeri, T. and Severgnini, B. (2013). *Changing the way referees are paid would be an important step towards preventing match fixing in European football.* Retrieved from https://blogs.lse.ac.uk/europpblog/2013/02/27/match-fixing-european-football-champions-league-corruption-calciopoli/

Bonesteel, M. (2022). *Sepp Blatter, Michel Platini acquitted on FIFA corruption charge.* Retrieved from https://www.washingtonpost.com/sports/2022/07/08/sepp-blatter-michel-platini-acquitted/

Brooks, G., Lee, J. and Kim, H. (2012). *Match-Fixing in Korean Football: Corruption in the K-League and the Importance of Maintaining Sporting Integrity.* Retrieved from https://dx.doi.org/10.5392/IJoC.2012.8.2.082

Buraimo, B., Migali, G. and Simmons, R. (2015). *An Analysis of Consumer Response to Corruption: Italy's Calciopoli Scandal.* Retrieved from https://doi.org/10.1111/obes.12094

Hung Mo, P. Corruption and Economic Growth. *Journal of Comparative Economics,* 29(1): 66-79, 2001

Independent Commission Against Corruption (2023). *Corruption Perceptions Index.* Retrieved from https://www.icac.org.hk/en/intl-persp/ranking-and-research/corruption-perceptions-index/index.html

Kangralkar, S. (2021). *Regularization and Cross-Validation – How to choose the penalty value (lambda).* Retrieved from https://medium.com/analytics-vidhya/regularization-and-cross-validation-how-to-choose-the-penalty-value-lambda-1217fa4351e5

Kassies, B. (2023). *UEFA European Cup Coefficients Database.* Retrieved from https://kassiesa.net/uefa/data/

Krieckhaus, J., Cooper Drury, A. and Lusztig, M. Corruption, Democracy and Economic Growth. *International political science review,* 27(2): 121-136, 2006

Lazic, M. (2023). *27 Informative Money Laundering Statistics in 2023.* Retrieved from https://legaljobs.io/blog/money-laundering-statistics/#:~:text=The%20United%20Nations%20estimates%20that,this%20amount%20remains%20undetected%20today.

Malfait, M. (2021). *Lab 3: Penalized regression techniques for high-dimensional data.* Retrieved from https://statomics.github.io/HDDA/Lab3-Penalized-Regression.html

Peeters, T. and Van Ours, J. (2021). *Seasonal Home Advantage in English Professional Football; 1974-2018.* Retrieved from https://doi.org/10.1007/s10645-020-09372-z

Pollard, R. and Armatas, V. (2017). *Factors affecting home advantage in football World Cup qualification.* Retrieved from http://dx.doi.org/10.1080/24748668.2017.1304031

Pouliopoulos, T. and Georgiadis, K. (2022). *The Problematic Institutional Context of Greek Football and the Role of FIFA and UEFA.* Retrieved from https://doi.org/10.2478/pcssr-2022-0008

Rocha, B, Sanches, F., Souza, I. and Carlos, J. (2013). *Does monitoring affect corruption? Career concerns and home bias in football refereeing.* Retrieved from https://doi.org/10.1080/13504851.2012.736938

Rollin, J. (2023). *2015 FIFA corruption scandal.* Retrieved from https://www.britannica.com/event/2015-FIFA-corruption-scandal

Simmons, R. and Deutscher, C. (2012). *The Economics of the World Cup.* Retrieved from https://doi.org/10.1093/oxfordhb/9780195387773.013.0023

Statistics How To (2023). *Ridge Regression: Simple Definition.* Retrieved from https://www.statisticshowto.com/ridge-regression/

The World Bank (2023a). *Inflation, consumer prices (annual %).* Retrieved from https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG?view=chart

The World Bank (2023b). *Population ages 20-24, male (% of male population).* Retrieved from https://data.worldbank.org/indicator/SP.POP.2024.MA.5Y

The World Bank (2023c). *Population ages 25-29, male (% of male population).* Retrieved from https://data.worldbank.org/indicator/SP.POP.2529.MA.5Y

Transparency International (2023). *Corruption Perceptions Index.* Retrieved from https://www.transparency.org/en/cpi/2000

UEFA (2023a). *Country coefficients.* Retrieved from https://www.uefa.com/nationalassociations/uefarankings/country/#/yr/2020

UEFA (2023b). *How association club coefficients are calculated.* Retrieved from https://www.uefa.com/nationalassociations/uefarankings/country/about/

Wikipedia (2023a). *1998 FIFA World Cup qualification (UEFA).* Retrieved from https://en.wikipedia.org/wiki/1998_FIFA_World_Cup_qualification_(UEFA)

Wikipedia (2023b). *2002 FIFA World Cup qualification (UEFA).* Retrieved from https://en.wikipedia.org/wiki/2002_FIFA_World_Cup_qualification_(UEFA)

Wikipedia (2023c). *2006 FIFA World Cup qualification (UEFA).* Retrieved from https://en.wikipedia.org/wiki/2006_FIFA_World_Cup_qualification_(UEFA)

Wikipedia (2023d). *2010 FIFA World Cup qualification (UEFA).* Retrieved from https://en.wikipedia.org/wiki/2010_FIFA_World_Cup_qualification_(UEFA)

Wikipedia (2023e). *2014 FIFA World Cup qualification (UEFA).* Retrieved from https://en.wikipedia.org/wiki/2014_FIFA_World_Cup_qualification_(UEFA)

Wikipedia (2023f). *2018 FIFA World Cup qualification (UEFA).* Retrieved from https://en.wikipedia.org/wiki/2018_FIFA_World_Cup_qualification_(UEFA)

Wikipedia (2023g). *2022 FIFA World Cup qualification (UEFA).* Retrieved from https://en.wikipedia.org/wiki/2022_FIFA_World_Cup_qualification_(UEFA)

Worldfootball.net (2023a). *WC Qualifiers Europe 1996/1997.* Retrieved from https://www.worldfootball.net/schedule/wm-quali-europa-1998-gruppe-9/0/

Worldfootball.net (2023b). *WC Qualifiers Europe 2000/2001.* Retrieved from https://www.worldfootball.net/schedule/wm-quali-europa-2002-gruppe-9/0/

Worldfootball.net (2023c). *WC Qualifiers Europe 2004/2005.* Retrieved from https://www.worldfootball.net/schedule/wm-quali-europa-2006-gruppe-1/0/

Worldfootball.net (2023d). *WC Qualifiers Europe 2008/2009.* Retrieved from https://www.worldfootball.net/schedule/wm-quali-europa-2010-gruppe-1/0/

Worldfootball.net (2023e). *WC Qualifiers Europe 2012/2013.* Retrieved from https://www.worldfootball.net/schedule/wm-quali-europa-2012-2013-gruppe-a/0/

Worldfootball.net (2023f). *WC Qualifiers Europe 2016/2017.* Retrieved from https://www.worldfootball.net/schedule/wm-quali-europa-2016-2017-gruppe-a/0/

Worldfootball.net (2023g). *WC Qualifiers Europe 2021/2022.* Retrieved from https://www.worldfootball.net/schedule/wm-quali-europa-2021-2022-playoffs-finale/0/

Appendix

Table A1 Descriptive statistics on the variables introduced in section 4

Variable	Mean	SD	Median	Max	Min	Observations
HW	2.17	1.45	2.00	6.00	0.00	365
HD	1.60	0.92	1.00	4.00	0.00	365
HL	1.63	1.49	1.00	6.00	0.00	365
GHS	7.79	4.76	7.00	26.00	0.00	365
GHC	6.04	4.73	5.00	30.00	0.00	365
GDH	1.76	8.20	2.00	24.00	-28.00	365
AW	1.62	1.33	2.00	5.00	0.00	365
AD	1.04	0.94	1.00	3.00	0.00	365
AL	2.18	1.60	2.00	6.00	0.00	365
GHS	6.02	3.96	6.00	22.00	0.00	365
GAC	7.80	5.35	7.00	33.00	0.00	365
GDA	-1.78	8.15	-1.00	21.00	-33.00	365
TGD	-0.02	15.56	2.00	39.00	-53.00	365
HP	7.555	4.22	8.00	18.00	0.00	365
AP	5.91	4.07	6.00	16.00	0.00	365
AVHATT	19534.06	15672.23	16033.13	79753.60	703.60	362
CPI	58.52	21.92	57.50	100.00	13.00	310
UEFA	18.89	18.98	13.13	105.71	0.00	361
ART	0.06	0.22	0.00	1.00	0.00	365
MALES_20_24	7.10	1.21	6.81	10.60	4.63	365
MALES_25_29	7.33	0.98	7.22	10.20	4.77	365
INFL	7.21	15.99	2.88	168.62	-1.58	334

Note. SD stands for the sample standard deviation; Max stands for maximum; Min stands for minimum; all results are reported in two decimal places except for the number of observations.

Adapted sources: Kassies (2023), The World Bank (2023a), The World Bank (2023b), The World Bank (2023c), Transparency International (2023), UEFA (2023a), Wikipedia (2023a), Wikipedia (2023b), Wikipedia(2023c), Wikipedia (2023d), Wikipedia (2023e), Wikipedia (2023f), Wikipedia (2023g), Worldfootball.net(2023a), Worldfootball.net(2023b), Worldfootball.net(2023c), Worldfootball.net(2023d), Worldfootball.net(2023e), Worldfootball(2023f), Worldfootball(2023g)

Table A2 Descriptive statistics on the variables introduced in section 5

Variable	Mean	SD	Median	Max	Min	Observations
AHAG	0.37	1.14	0.35	3.60	-2.65	365
AHAP	0.34	1.32	0.30	5.50	-3.15	365
RELAHAG	-0.01	1.08	0.00	2.53	-2.87	286
RELAHAP	0.01	1.33	0.00	5.20	-3.00	286
GROUPGDH	7.13	1.68	2.20	6.20	-1.20	62
GROUPPDH	2.14	2.10	1.80	6.60	-3.75	62
GROUPAHAG	0.37	0.31	0.34	1.16	-0.22	62
GROUPAHAP	0.34	0.39	0.30	1.35	-0.75	62
PDH	1.61	8.29	3.00	18.00	-18.00	365
PDA	-1.65	8.27	0.00	15.00	-18.00	365
TPD	-0.03	15.51	3.00	30.00	-36.00	365
RELAVHATT	-0.12	0.74	0.04	1.16	-2.44	286
RELCPI	-0.06	0.39	-0.02	0.94	-1.39	286
RELUEFA	-0.27	0.99	-0.05	1.44	-3.58	286
RELINFL	-0.33	1.03	-0.20	2.08	-5.44	286
RELMALES_20_24	0.00	0.14	0.01	0.38	-0.38	286
RELMALES_25_29	0.01	0.12	0.00	0.33	-0.25	286
GROUPAVHATT	19730.90	7257.26	19445.83	34424.35	5901.03	62
GROUPCPI	59.08	10.19	58.25	83.70	37.00	62
GROUPUEFA	19.24	5.70	18.84	36.32	10.44	62
GROUPMALES_20_24	7.08	0.66	7.17	8.32	5.46	62
GROUPMALES_25_29	7.32	0.46	7.36	8.26	6.24	62
GROUPINFL	7.32	7.83	3.83	36.98	-0.34	62

Note. SD stands for the sample standard deviation; Max stands for maximum; Min stands for minimum; all results are reported in two decimal places except for the number of observations.

Adapted sources: Kassies (2023), The World Bank (2023a), The World Bank (2023b), The World Bank (2023c), Transparency International (2023), UEFA (2023a), Wikipedia (2023a), Wikipedia (2023b), Wikipedia(2023c), Wikipedia (2023d), Wikipedia (2023e), Wikipedia (2023f), Wikipedia (2023g), Worldfootball.net(2023a), Worldfootball.net(2023b), Worldfootball.net(2023c), Worldfootball.net(2023d), Worldfootball.net(2023e), Worldfootball(2023f), Worldfootball(2023g)

Table A3 R packages used and their purpose

Package	Purpose
readxl	Reading Excel files into the R script
glmnet	Performing Ridge regression and K-fold cross-validation
ivreg	Performing IV-regressions

Table A4 Number of folds K and the tuning parameter λ employed for both Ridge regressions

Variable	Relativehomeadvantagegoal	Relativehomeadvantagepoint
K	26	26
λ	7.70	3.23

Table A5 Used abbreviations and their definitions

Variable	Definition
HW	Number of wins at home
HD	Number of draws at home
HL	Number of losses at home
AW	Number of wins away
AD	Number of draws away
AL	Number of losses away
GHS	Goals scored at home
GAS	Goals scored away
GHC	Goals conceded at home
GAC	Goals conceded away
HP	Points obtained at home
AP	Points obtained away
GDH	Goal difference at home
GDA	Goal difference away
TGD	Total goal difference (GDH + GDA)
ART	Proportion of games played on an artificial pitch
AVHATT	Average home attendance
CPI	Corruption perceptions index
UEFA	UEFA association club coefficient
AHAG	Absolute home advantage computed using the goal difference method
GROUPGDH	The scaled group average goal difference at home
GROUPAHAG	The group average absolute home advantage computed using the goal difference method
RELAHAG	Relative home advantage computed using the goal difference method
PDH	Point difference at home
PDA	Point difference away
TPD	Total point difference (PDH + PDA)
GROUPPDH	The scaled average point difference at home
AHAP	Absolute home advantage computed using the point difference method
GROUPAHAP	The group average absolute home advantage computed using the point difference method
RELAHAP	Relative home advantage computed using the point difference method
GROUPAVHATT	The group average of the average home attendances

RELAVHATT	The relative average home attendance
GROUPCPI	The group average of the corruption perceptions index
RELCPI	The relative corruption perceptions index
GROUPUEFA	The group average of the UEFA association club coefficients
RELUEFA	The relative UEFA association club coefficient
INFL	Inflation percentage
GROUPINFL	The group average inflation percentage
RELINFL	The relative inflation percentage
MALES_20_24	Proportion of males aged 20-24 from the total male population
GROUPMALES_20_24	The group average of the proportion of males aged 20-24
RELMALES_20_24	The relative proportion of males aged 20-24
MALES_25_29	Proportion of males aged 25-29 from the total male population
GROUPMALES_25_29	The group average of the proportion of males aged 25-29
RELMALES_25_29	The relative proportion of males aged 25-29