Erasmus University Rotterdam

Erasmus School of Economics

Bachelor Thesis International Bachelor Economics and Business Economics

Testing the Uncertainty of Outcome Hypothesis in the Top Five European Leagues

Student Name: Moustafa Elshawaf

Student ID Number: 525932

Supervisor: Schelte Beltman

Second Assessor: Arie Barendregt

Date Final Version: October 2, 2023

TABLE OF CONTENTS

1. INTRODUCTION

1.1 INTRODUCTION

The term ‘uncertainty of outcome’ (UO) was first described in a sports context by Rottenberg (1956), he states that uncertainty of outcome is necessary for a customer to be willing to pay for admission. The ‘uncertainty of outcome hypothesis’(UOH) was posited by Neale (1964), who claimed that “the appeal of the seat depends mostly on the uncertainty of the outcome and on the weather”. Uncertainty of outcome is often confused with ‘competitive balance’ although they are not the same. Forrest and Simmons (2002) define competitive balance as when the league structure has relatively equal playing strength between members. In the context of association football, competitive balance would mean that the differences in points tallies between clubs would be minimal in a competitively balanced league. Whereas uncertainty of outcome would refer to a situation in which any given game in a league is evenly matched and by extension no predetermined winner is present in the league (Forrest & Simmons, 2002). The uncertainty of outcome hypothesis is based on the assumption that spectators receive more utility from watching a game in which the outcome is more unpredictable, thus as the UO increases, so will the attendance demand (Knowles et al., 1992).

This paper aims to shed new light on the UOH using a larger dataset than has ever been used in the literature as the author aims to research the top five European leagues (English Premier League, Spanish La Liga, German Bundesliga, Italian Serie A and French Ligue 1). As well use data from the 2015/16 season up to the season just passed 2022/23 excluding games played behind closed and games where attendance was limited due to the covid-19 pandemic. This research will take a quantitative statistical approach to assess the correlation between match uncertainty of outcome and stadium attendance. The data is collected from online sources including betting odds, match attendances and other metrics that explain the quality of the contest for match going supporters. The analysis used will include descriptive statistics to summarize the datasets, followed by a linear regression to assess the correlation between match uncertainty of outcome and stadium attendance. These analyses will be explained and

discussed. This research is highly relevant for the board members of the leagues that will be analyzed, as from the correlation, one can decide whether measures to increase the competitiveness of the division will increase attendances at games across the league. Even in the age of TV and online football viewership, attendances remain an important part of a club's revenue, and research into what determines attendance can help football clubs understand how to increase the number of match-goin fans.

The thesis is structured following this introduction with a literature review detailing the previous studies into the topic and setting the theoretical framework for this research. This is followed by the data and methodology section where the author details the data collection process, displays the descriptive statistics of the data and presents the data analysis framework. The author then presents and analyzes the results from the statistical analysis, followed by a discussion on the limitations, the validity and the future implications of the study, as well recommendations for future research.

1.2 RESEARCH OBJECTIVE

In this paper the authors aims to test the uncertainty of outcome hypothesis using new data from the past eight years in Europe's top five leagues, therefore the authors formulates the following central research question:

RQ: How has uncertainty of outcome correlated with attendance demand in Europe's top five leagues in the last eight years?*

Note: Top five leagues denotes English Premier League, Spanish La Liga, German Bundesliga, Italian Serie A and French Ligue 1

As the author uses moderating and mediating variables to get a better estimate for the correlation between UO and attendance demand, the authors formulates the following sub-questions:

SQ1: What is the effect of a game being a derby on attendance demand?

SQ2: What is the effect of the home team's probability of a victory on attendance demand?

SQ3: What is the effect of the importance of the game for the home and away team on attendance demand?

2. LITERATURE REVIEW

2.1 LITERATURE REVIEW

Uncertainty of outcome has been measured in three separate dimensions, short-term UO is measured at the match-level, matches played between evenly matched teams would have a high match-UO (Cairns et al., 1986). Mid-term UO is measured in the context of sub competitions within a league structure, such as races for continental qualification and relegation battles, seasons where there are many teams that can viably challenge for a European spot would rank high for mid-term UO. Long-term UO is based on whether a league is dominated by a single team for years on end. The German Bundesliga is an example of a league with a low long-term uncertainty of outcome as Bayern Munich has won the league 11 times consecutively. In this paper the author will investigate how match uncertainty of outcome affects the attendance demand of the 'top five' European leagues (English Premier League, Spanish La Liga, German Bundesliga, Italian Serie A and French Ligue 1) in the last eight years.

Match UO has been measured using various methods in the literature. Hart et al. (1975) measured UO as the logarithmic absolute difference of opponents in league standings and found no significant effect of the measure on match attendance in the English First Division. Peel and Thomas (1988) used betting odds to calculate the home team's a priori probability of winning and studied its impact on match attendance in the English First Division. The correlation was significantly positive, suggesting a rejection of the UOH. The measure of uncertainty introduced by Theil (1967) allows for all three possible outcomes in a football match to be taken into account when measuring uncertainty. Theil's measure is increasing with UO, the utilization of the measure in the literature usually leads to a significant negative correlation, suggesting a rejection of the UOH (Buraimo & Simmons, 2008; Jespersen & Pedersen, 2018). A significant negative measure in the Theil measure suggests that either that the home team's supporters

prefer to watch a game against inferior opposition (Buraimo & Simmons, 2008), or when the opposition is a far better team with a strong reputation (Pawlowski & Anders, 2012). Buraimo and Simmons (2008) analyzed English league games and concluded that the negative value in the Theil measure is explained by supporters of the home team valuing a victory higher than an equal contest with high quality opposition.

Studies into the effect of match-UO on stadium attendance have mostly focused on English divisions (Pawlowski, 2013). Using betting odds to calculate uncertainty of outcome it has been found that supporters favor an uncertainty of outcome (Forrest & Simmons, 2002). They argue that despite these results, better quality of teams in the league may still yield a fall in aggregate attendance due to the degree to which playing at home creates an unequal competition when two teams of comparable strength face each other (Forrest & Simmons, 2002). Forrest et al. (2005) studied the first, second and third divisions of English football and found a U-shaped relationship between their uncertainty measure and the logarithm of match attendance. The uncertainty measure in question is the ratio of the probability of a home win to the probability of an away win. The U-shaped relationship means that games with the lowest and highest measures of UO featured larger attendances, while those in the middle had lower attendances.

There have also been a number of studies into the UOH in the European mainland. A study into the French first division found no significant relationship between UO (measured as difference in points per game) and the logarithm of match attendance (Falter et al., 2008). A study of the German first division used various measures of uncertainty to ascertain whether differences in the resulting relationship arise and found that they indeed did (Benz et al., 2009). The study used the following measures of uncertainty: (a) absolute difference in league standings; (b) difference in points per game; (c) Theil index/betting odds; (d) home team winning probability/betting odds. It is worth noting that these measures, other than the Theil measure, are decreasing with uncertainty of outcome. The study found the resulting relationships between each measure and the logarithm of match attendance: (a) negative; (b) negative); (c) not significant; (d) inverse U-shaped relationship (Benz et al., 2009). Controlling for season-ticket holders, the researchers found support for the UOH for all measures of

uncertainty other than the Theil index based on betting odds (Benz et al., 2009). Another study into the German first division did however find a significant negative relationship between the Theil measure and the logarithm of match attendance, suggesting a rejection of the UOH (Pawlowski & Anders, 2012). Another study using the Theil measure however found no significant result in the Swiss and Austrian first divisions (Pawlowski & Nalbantis, 2015)

There have also been a number of studies into the uncertainty of outcome hypothesis for other sports. The author argues that the external validity of these studies into football are high as the common denominator are the fans who will have a lot of the same motivations when it comes to supporting their favorite sports team. However it must be noted that there are significant cultural differences between the sports played as well as the countries in which they are played.

There have been a number of studies on the topic based on Major League Baseball (MLB) of the United States with mixed results. Lee and Fort (2008) found no significant results when using the winning percentage distribution as the uncertainty measure against average attendance for over a hundred years of data. The home team win probability, measured through betting odds, was found to have an inverse-U relationship with the logarithm of match attendance for the 2007 of the MLB, regarded as support for the UOH (Lemke et al., 2010). Using the same uncertainty measure however, Coates et al. (2014) found a U-shaped relationship for the 2005-10 years of the MLB, suggesting a rejection of the UOH.

A study into a season of the British rugby league with the handicap spread as the uncertainty measure found a negative relationship with match attendance suggesting a rejection of the null hypothesis (Peel & Thomas, 1997). Coates and Humphreys (2010) rejected the UOH for the National Football League of the United States when measuring uncertainty as both the winning percentages of home and visiting teams and absolute point spread for the 1985-2008 seasons. Research into the 1950-2000 seasons of the National Basketball Association of the United States used distribution of winning percentages as the uncertainty measure and found a positive relationship with average yearly attendance, suggesting support for the UOH (Mills & Fort, 2014).

Overall, the author finds that the uncertainty of outcome hypothesis has seen mixed results in the literature with numerous studies both showing support as well as contradiction. Despite this, studies have leaned more towards support for the UOH, suggesting that uncertainty of outcome does indeed have a positive relationship with stadium attendance in football. The author observes that there have been many different proxies for uncertainty of outcome, therefore the author cannot be sure that the results might have been different if a different proxy was chosen. There have also been slight differences in the methods selected, which the author will dig deeper into in the data method and selection chapter.

Following from the results analyzed in the literature, the null and alternative hypotheses the author will use to answer the central question are as follows:

H_0: 'There is no correlation between the Theil measure and attendance'.

H_1: 'There is a positive and significant correlation between the Theil measure and attendance'.

2.2 DATA & METHOD SELECTION

In order to research the correlation between uncertainty of outcome and attendance demand, the author must first find a proxy for each variable. Throughout the literature, raw attendance values are used as a proxy for attendance demand due to the scarcity of ticket sales data and the accessibility of attendance values (Knowles et al., 1992; Forrest & Simmons, 2002; Buraimo & Simmons, 2008; Jespersen & Pedersen, 2018). The author finds however that uncertainty of outcome has been measured using a variety of methods in the literature as mentioned earlier in the thesis. Betting odds are the basis of the uncertainty model as they are a readily available data source that can quite accurately predict the result of any given football match. Using this as our basis, the author will use the Theil measure as a proxy for uncertainty of outcome as has plenty in the literature before (Peel & Thomas, 1992; Czarnitzki & Stadtmann, 2002; Buraimo & Simmons, 2008). The author uses this measure as it allows for the incorporation of the probabilities of all three outcomes of a game (Jespersen & Pedersen, 2018). It must be stated however that betting odds have been shown to be biased (Forrest & Simmons,

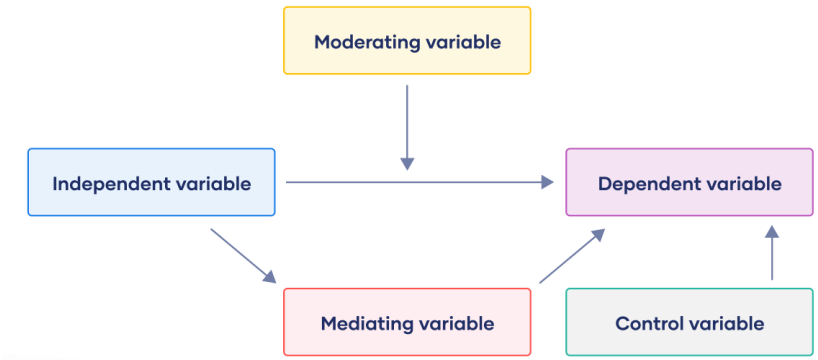
2002), however the author will not take steps to reduce this bias level to do time and budgetary constraints.

Furthermore there is the question of the statistical method the author uses to determine the level of correlation between the two main variables. The tobit model has been used extensively in the literature, based on the assumption that attendances are right-censored when used as a proxy for attendance demand due to the capacity limitations of the stadium (Kuypers, 1996; Welki & Zlatoper, 1994). Forrest and Simmons (2002), however state that the tobit model assumes that 'true' demand is observed at events where the capacity of a stadium is not reached but that this is not the case due to the bundling of tickets. In order to be sure of a seat at a game with more demand, one must also buy a ticket for a game which is not highly demanded. This is known as season tickets and are a staple of the ticketing systems for all European clubs. Therefore, true demand is not observed at games with a lower demand as attendances are inflated with those who purchased bundle tickets and would not have purchased a standalone ticket otherwise (Forrest & Simmons, 2002). As a result of this, the hypothesis will be tested with ordinary least squares multiple regressions rather than the tobit model.

2.3 CONCEPTUAL FRAMEWORK

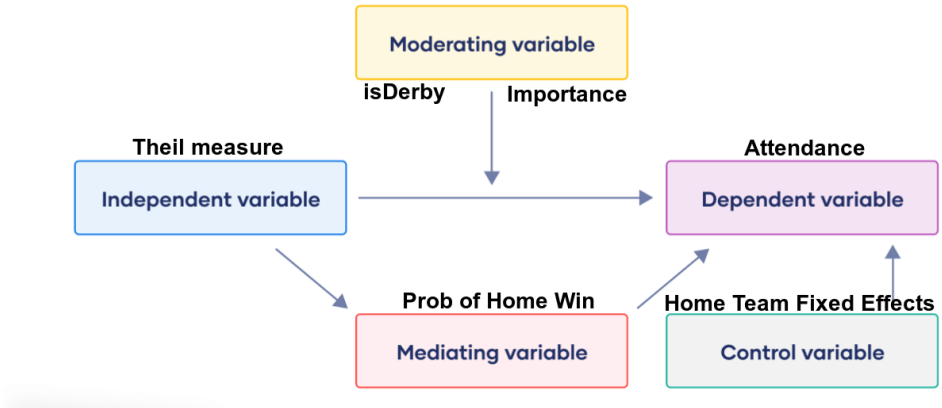
The author uses a conceptual framework to visualize the relationship between the variables in the model. Through the use of a conceptual framework, the author will more easily define the objectives of this research and map out the research process. The conceptual framework the author will use is shown in Figure 1. The framework distinguishes between the types of variables the author observes, in total defining five types of variables. The independent and dependent variables respectively are the cause and effect variables for which the research is mainly investigating the relationship. The mediating variable is a variable which is correlated to the independent variable but also has an effect on the dependent variable, while the moderating variable is not correlated to the independent variable but also has an effect on the dependent variable. The control variables are variables for which are held constant and aim to increase the accuracy of the relationship between the cause and effect variables (Swaen, 2022).

Figure 1: Conceptual Framework Example



As this study chiefly aims to research the relationship between UO and attendances in the top five leagues, these variables will be the independent and dependent variables respectively. The Theil measure which has frequently been used as a proxy for UO will be the proxy in this study. Furthermore, the author will use home team fixed effects to control for differences in stadium capacity across teams and this will be the control variable. Mediating the relationship between the cause and effect variables is the probability of a home team victory, while the importance of the game and whether it is a derby will moderate the relationship. This conceptual framework is fully visualized in Figure 2.

Figure 2: Conceptual Framework with Variables Used in this Study



3. DATA & METHODOLOGY

3.1 DATA COLLECTION

The author aims to investigate the correlation between the match uncertainty of outcome and attendance demand for football matches in the top five leagues in Europe; English Premier League, Spanish La Liga, German Bundesliga, Italian Serie A and French Ligue 1. These specific leagues are selected as they are the most reputable leagues in the world. There is a high interest in these leagues so readers are more likely to understand the match-going culture of fans in these leagues than in other, more obscure leagues. Furthermore, it will be of interest to examine the differences between the leagues as there are different perceptions of what drives fans to support the team at their stadium, for example Germans are viewed as fans that fill the stadium no matter the opposition or the weather.

Attendance data, which the author uses as a proxy for attendance demand, is collected from Football Reference[1] which is a widely-used source for football statistics. Attendance data is available for the top five leagues from the 2015/16 season so this is where the database begins. Due to the COVID-19 pandemic, stadiums across Europe were forced to close and/or limited to a certain capacity. For this reason, the author omits data from the season 2020/21 for games played after March 8, 2021 as well as from games in 2021/22 where stadiums had limited capacity. The author includes all games played in the 2022/23 season. The author takes the log of the attendance variable as it is more appropriate to investigate how variables affect the percentage change in attendance rather than the total change. This is due to stadiums across Europe being vastly different in capacity.

Following Forrest and Simmons (2002), the author uses betting odds to calculate the probability of the home team winning and Theil's measure, our measure of UO. Reasoning behind the selection of the Theil measure can be found in Chapter 2.2. Betting odds for all seasons were collected from football-data.co.uk[2] in the form of fixed odds, the author collects

[1]Football Reference can be accessed at the following link using an example of 2014/15 La Liga statistics: https://fbref.com/en/comps/12/2014-2015/schedule/2014-2015-La-Liga-Scores-and-Fixtures

[2] football-data.co.uk contains csv files for over 20 different leagues going back over 20 years with pre-game fixed odds

the odds for a home team win, draw and away team win. The website includes betting odds from various bookmakers, the author uses odds from Bet365 as odds from different bookmakers are heavily correlated and which one is used is irrelevant (Forrest & Simmons, 2002). In order to transform the betting data into the home team's probability of winning, the author takes the inverse of the odds and adjusts for the bookmaker's profit margin by turning the probabilities into probability ratios (Jespersen & Pedersen, 2018). Furthermore, the author calculates the measure of uncertainty introduced by Theil (1967) which is defined as:

(1) $$Theil = \sum_{i=1}^{3} \frac{p_i}{\sum_{i=1}^{3} p_i} \ln\left(\frac{\sum_{i=1}^{3} p_i}{p_i}\right)$$

where p_i denotes the probability of a home win, draw, and away win respectively. As the author removes the bookmaker's profit margin from the probabilities, the sum of p_i will take on the value of 1. This method of measuring uncertainty of outcome has been used numerous times in the literature (Peel & Thomas, 1992; Czarnitzki & Stadtmann, 2002; Buraimo & Simmons, 2008). Given the mixed results obtained when using the Theil measure, the authors have no clear indication of its sign or significance of the variable (Jespersen & Pedersen, 2018).

The author includes the moderating variables of whether the game is a derby and the importance of the game with regards to championship, European qualification or relegation in order to increase the statistical power of the model and alleviate omitted variable bias. The relevance of derbies to football attendance is shown within the literature to be sizable (Buraimo & Simmons, 2008). The author includes the variable to capture attendance increases due to the historical magnitude of a particular fixture. The correlation between the importance of a game and its attendance have not been researched much as it has been difficult to quantify the importance of any fixture. With the help of AI in recent years, it is now possible to quantify a game's importance without human biases. Previously in the literature, the month in which a game was played has been used as a proxy for game importance as games in April or May tend to be more important as the season draws to a close (Forrest & Simmons, 2002). However, there are many games in the latter stages of a season that are worth nothing to the team's

qualification or relegation as gaps can become too wide to overcome, with popular culture referring to this phenomenon as players being 'on the beach'. For this reason the author shies away from using month dummies as a proxy for importance.

A list of football derbies for each league was collected from footballderbies.com. The website contains rivalries ranging from city rivalries such as Tottenham vs Chelsea to historical competitive rivalries such as Arsenal vs Manchester United. Derby is a dummy variable which takes the value of 0 when the game is not a derby and 1 when it is. The author expects the derby variable to take a positive and significant coefficient as rivalries tend to attract more spectators as the literature suggests (Buraimo & Simmons, 2008; Martins & Cró, 2018).

Finally, the author specifies the variables importance1 and importance2, where importance1 denotes the significance of the game with regards to title races, European qualification, and relegation battles for the home team. Importance2 denotes the same but for the away team. The importance variables are collected from FiveThirtyEight[3] which is a subsidiary of ABC News which focuses on opinion polling in the United States as well as making predictions for sports games. The importance variables are available from the 2017/18 season onwards, however there are whole gameweeks for which the importance variables are unavailable.

3.2 DESCRIPTIVE STATISTICS

[3] The data, along with an explanation for the methodology of the calculation of the importance variables can be found at the following link: https://github.com/fivethirtyeight/data/tree/master/soccer-spi

Table 1: Descriptive statistics

Variable	Mean	Std. Dev.	Minimum	Maximum	Observations
ln(Attendance)					
England	10.450	0.447	9.208	11.329	2947
Spain	10.010	0.676	8.182	11.500	2805
Germany	10.464	0.640	7.821	11.307	2346
Italy	9.886	0.627	7.742	11.279	2901
France	9.830	0.605	7.876	11.167	2912
Probability of Home Win					
England	0.445	0.199	0.042	0.914	2947
Spain	0.455	0.190	0.036	0.920	2805
Germany	0.446	0.181	0.041	0.926	2346
Italy	0.443	0.191	0.056	0.908	2901
France	0.443	0.164	0.046	0.916	2912
Result					
England	0.849	0.860	0	2	2947
Spain	0.817	0.846	0	2	2805
Germany	0.928	0.878	0	2	2346
Italy	0.882	0.856	0	2	2901
France	0.954	0.873	0	2	2912
Theil Measure					
England	0.962	0.153	0.346	1.098	2947
Spain	0.968	0.161	0.326	1.098	2805
Germany	0.979	0.146	0.312	1.098	2346
Italy	0.974	0.135	0.365	1.098	2906
France	0.998	0.130	0.342	1.098	2912
Derby					
England	0.050	0.218	0	1	2947
Spain	0.053	0.224	0	1	2805
Germany	0.055	0.229	0	1	2346
Italy	0.063	0.244	0	1	2901
France	0.056	0.229	0	1	2912
Importance for Home Team					
England	36.501	26.325	0	100	2151
Spain	34.176	24.360	0	100	1263
Germany	38.964	24.892	0	100	1104
Italy	35.132	26.004	0	100	1364
France	31.941	23.349	0	100	1253
Importance for Away Team					
England	35.450	25.931	0	100	2151
Spain	33.547	23.934	0	100	1263
Germany	38.143	24.149	0	100	1104
Italy	34.656	25.889	0	100	1364
France	30.864	23.078	0	100	1253

Descriptive Statistics are shown in Table 1. All the leagues play 380 games a season, except for the German Bundesliga which plays 306 games per season, this explains the lower number of observations present in the table. The author observes that average attendances are highest in the German Bundesliga followed closely by the English Premier League, and they are the lowest in the French Ligue 1. ProbH depicts the probability of a home team victory based on betting odds. Result is an interval variable which denotes the outcome of the game, taking value 0 in the case of a home win, 1 in the case of a draw and 2 for an away win. The author finds that bookmakers give Spanish home teams the highest average probabilities of a win, which is referred to as home bias. This is justified as the result variable shows that home bias is strongest in Spain, as it has the lowest mean value for result.

The Theil variable captures UO and is increasing with uncertainty. The author finds that there is an upper limit of uncertainty at 1.098. The mean of the Theil measure is highest in France and lowest in England, which could be explained by the difference in team strength between the top six in England and the rest of the league. In Ligue 1, the league is more uncertain in that the teams that occupy European qualification spots change more often than in England. It is worth noting that France has the highest mean value for Theil but also has the least mean attendance of all leagues, while the English league has the lowest mean value for Theil and the second highest attendance. Of course, the author cannot come to any conclusions from this as cultural and stadium capacity differences greatly affect the mean attendances between leagues.

The author finds that in all leagues the prevalence of derbies centers around 5.5%, the Italian Serie A hosting the most derbies per game and England the least. Due to the importance variables containing many missing values, the author notices fewer observations across all leagues. The Premier League has by far the most observations, which could be a result of its greater popularity incentivizing FiveThirtyEight to focus on it more.

3.3 ANALYSIS FRAMEWORK

The hypothesis will be tested with ordinary least squares (OLS) multiple regressions with robust standard errors and fixed effects of the home team for each league separately. Ordinary least squares regression is a widely accepted and employed technique in econometrics for four distinct reasons. Firstly, it is an intuitive method for examining correlations between variables. This is achieved through creating a linear equation that minimizes the sum of the squared differences between the observed and predicted values, this is a simple process that is not difficult to understand. Secondly, multiple linear regression allows the researcher to add more variables that are likely to explain the deviance in the dependent variable which in this case is stadium attendance. With every execution of OLS regression, the author is provided with valuable measures to assess the goodness of fit and statistical significance of the model, through the R-squared and p-values respectively. These measures help ensure the robustness and reliability of the results. Finally, OLS regression allows for the identification of both the direction and magnitude of the relationship, providing meaningful insights for regulators and club directors to take action upon.

The data analysis is conducted in Stata as it provides a simple user-interface yet conducts powerful and fast analysis with a high number of maximum observations. The author's previous experience with Stata is also a factor in the decision to use the program. A key assumption of the linear regression model is homoscedasticity, which means that the residuals or the error term are constant across different values for the predicting variables. A violation of the homoscedasticity assumption may lead to unreliable results, so to correct for this one must use robust standard errors. The author uses robust standard errors as evidence of heteroscedasticity is found in every league after conducting the Breusch-Pagan/Cook-Weisberg test in StataMP 17 (Breusch & Pagan, 1979). Furthermore, the author conducted a variance inflation factor to test for multicollinearity between the variables and found no concrete evidence for multicollinearity as the highest VIF value found was 2.7 for ProbH in Serie A (Gujarati & Porter, 2009). The OLS regressions and relevant tests are completed in StataMP 17.

The sign and the significance of the coefficient will indicate whether the author can or cannot reject the null hypothesis. The null hypothesis is 'H_0: no correlation between uncertainty

of outcome and attendance demand'. The alternative hypothesis is 'H_1: There is a significant positive correlation between outcome uncertainty and attendance demand'. To test the null hypothesis the author will compute two OLS regressions, the first excluding the importance variables and the second including them. The following regressions will be used to test the null hypothesis:

$$(2) \quad Y_i = \beta_0 + \beta_1 * Theil_i + \beta_2 * ProbH_i + \beta_3 * Derby_i + \gamma_1 * D_{1i} + \gamma_2 * D_{2i} + ... + \gamma_{\square} * D_{\square i} + \varepsilon_i$$

$$(3) \quad Y_i = \beta_0 + \beta_1 * Theil_i + \beta_2 * ProbH_i + \beta_3 * Derby_i + \beta_4 * Importance1_i + \beta_5 * Importance2_i + \gamma_1 * D_{1i} + \gamma_2 * D_{2i} + ... + \gamma_{\square} * D_{\square i} + \varepsilon_i$$

Where Y_i denotes the dependent variable ln(Attendance) which the author uses as a proxy for attendance demand, β_0 denotes the constant term and ε_i denotes the error term. As the dependent variable is the natural logarithm of attendance, the β coefficients indicate the percentage change in attendance given a unit change in the explanatory variables. For example, the β_1 coefficient depicts the percentage change in attendance due to a one unit increase in the Theil measure. The significance of the coefficients will be determined through the 95% confidence interval. For the testing of the hypothesis the parameter β_1 is the coefficient of interest. A significant fluctuation from zero does not however mean that there is a causal effect of the Theil measure on attendance. A causal effect could only be inferred with the presence of a control and treatment group, however given time and budget constraints this will not be possible in this study. Therefore the author will only analyze whether there is a strong statistical correlation between the parameters. The author recommends that future research takes a deeper look into the likelihood of a causal relationship.

The γ coefficients indicate the percentage change in attendance given the dummy variable (D) is equal to one. Each dummy variable (D) takes a value of one if the observation belongs to the specific home team and zero otherwise. The author uses the home team's probability of winning, whether the game is a derby and the importance variables for the home

and away team as control variables in this model. Each pair of tests are conducted for each of the leagues separately to investigate international differences in the coefficients.

4. RESULTS

4.1 RESULTS EXCLUDING IMPORTANCE VARIABLES

Table 2 depicts the regression results excluding the importance variable where the natural logarithm of attendance is the dependent variable. The author finds that in all leagues except the German Bundesliga, the sign of the coefficient of the Theil measure is negative at the 1% significance level which leads to a rejection of the null hypothesis for these leagues. Outcome uncertainty is most negatively correlated with attendance in the French Ligue 1, where a 0.756 increase in the Theil measure, the highest possible given the minimum and maximum values in Table 1, results in a 0.457% decrease in attendance *ceteris paribus*.

Interestingly, the author finds that in the same four leagues the sign and coefficient of ProbH is negative at the 1% significance level. This means that as the home team's probability of a win increases, attendances fall. It is important to mention that in European football, the vast majority of spectators at a stadium are supporters of the home team and thus changes in attendance are mostly reflected by the utility the home team's supporters put on the spectacle. Therefore, attendances increasing with a lower home team probability of a victory could be due to the "David versus Goliath'' effect postulated by Buraimo and Simmons (2008) in which home team supporters desire to be present in the low likelihood that an upset victory takes place. It is also important to note that the Theil measure and ProbH are not directly correlated. This can be observed through example games with likelihoods 0.2, 0.3, 0.5 and likelihoods 0.5, 0.3, 0.2 where the first value is the ProbH, second is ProbD and third is ProbA. Both games have the same Theil measure but the second features a higher ProbH. The results indicate that in the four aforementioned leagues, the first game in which the home team is the 'underdog', is more appealing to home team supporters.

Furthermore, the author finds that the coefficient of derby is positive and significant at the 1% significance level in Spain, Italy and France as expected. It can thus be concluded that games which are derbies attract more spectators in these leagues.

The R squared is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by the independent variables in a regression model. The author finds that the R squared is the lowest in Germany, which could be due to the higher average attendances in Germany, a country that is well known for avid supporters who attend each game no matter the opposition. This could also explain why the author receives no significant coefficients in the German league.

Table 2: Ordinary least squares results without importance variables

Predictor	**England**	**Spain**	**Germany**	**Italy**	**France**
Theil	-0.074***	-0.335***	-0.116	-0.392***	-0.604***
	(0.014)	(0.032)	(0.102)	(0.065)	(0.064)
ProbH	-0.031***	-0.359***	-0.041	-0.599***	-0.494***
	(0.012)	(0.024)	(0.087)	(0.049)	(0.049)
Derby	0.013	0.137***	0.053	0.129***	0.175***
	(0.010)	(0.022)	(0.044)	(0.030)	(0.026)
Constant	11.083***	10.199***	10.330***	10.350***	9.665***
	(0.012)	(0.038)	(0.120)	(0.074)	(0.073)
Observations	2947	2805	2346	2901	2912
Fit (R^2)	0.952	0.922	0.453	0.694	0.804

*Notes: Dependent variable is ln(attendance). The coefficient is reported with three decimal places and the robust standard errors are depicted in parentheses. The significance level is depicted by the number of stars: * = p<0.1, ** = p<0.05, *** = p<0.01.*

4.2 RESULTS INCLUDING IMPORTANCE VARIABLES

Table 3 depicts the regression results with the importance variables for the home and away teams included. The author finds that after adding the importance variables, the sign and significance of the Theil measure, ProbH and derby remain the same as in the first model other

than Theil in England which became insignificant at even the 10% level. The importance variables have very small coefficients and little significance across the board. The author finds only in Spain where both the importance for the home team and away are significant at the 1% level albeit with quite negligible coefficients. The author finds that with increasing importance comes higher attendance in the Italian Serie A, at the 5% significance level, also to a very miniscule degree.

Although the results depicted in Table 3 contain more explanatory variables, the overall statistical power of the model seems to have decreased given the much lower number of total observations.

Table 3: Ordinary least squares results with importance variables

Predictor	England	Spain	Germany	Italy	France
Theil	-0.019	-0.269***	0.042	-0.310***	-0.519***
	(0.014)	(0.041)	(0.185)	(0.087)	(0.109)
ProbH	-0.046***	-0.348***	0.011	-0.506***	-0.578***
	(0.010)	(0.039)	(0.183)	(0.081)	(0.075)
Derby	0.008	0.104***	0.050	0.152***	0.147***
	(0.009)	(0.021)	(0.096)	(0.044)	(0.047)
Importance1	0.000	0.002***	0.000	0.001**	0.001
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
Importance2	0.000**	0.001**	-0.000	0.001	-0.000
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
Constant	11.031***	10.044***	10.116***	10.179***	9.968***
	(0.013)	(0.052)	(0.246)	(0.108)	(0.126)
Observations	2151	1263	1104	1364	1253
Fit (R^2)	0.974	0.940	0.467	0.722	0.832

*Notes: Dependent variable is ln(attendance). The coefficient is reported with three decimal places and the robust standard errors are depicted in parentheses. The significance level is depicted by the number of stars: * = p<0.1, ** = p<0.05, *** = p<0.01.*

5. DISCUSSION AND CONCLUSION

5.1 DISCUSSION

The central research question for this thesis was the following:

How has uncertainty of outcome correlated with attendance demand in Europe's top five leagues in the last eight years?

Following the results found in the previous chapter, it can definitively be said that uncertainty of outcome is negatively and significantly correlated with attendance in The English, Italian, French and Spanish first division. The correlation between uncertainty of outcome and attendance in the German Bundesliga remains unclear and conflicting signs were observed along with no significant coefficients for the uncertainty measure.

This thesis aimed to provide new insights into the uncertainty of outcome hypothesis by researching the correlation between uncertainty of outcome and demand. The literature review the author conducted at the opening of this paper gave mixed results in terms of the relationship between the two variables, with results ranging from support, contradiction and no evidence in either direction. The author also observed a great host of different proxies for uncertainty of outcome in the research, Benz et al. (2009) even found that using different proxies could lead to vastly different results. The results from this study are in line with those found by Buraimo and Simmons (2008), where the Theil measure was found to have a negative and significant correlation with attendances in the English Premier League. A study into a single season of the German first division also found a negative and significant correlation with attendance, conflicting our results for the German first division (Pawlowski & Anders, 2012). A further study into the Swiss and Austrian first divisions found no significant result between the Theil measure and attendance (Pawlowski and Nalbantis, 2015).

The author chose to study the top five European leagues due to their popularity not only in their own country but around the world. The years the author chose (15/16-22/23) was based on data availability, as the author could not affordably source complete attendance data for previous years. One of the main objectives in this research was to make sure the scope was as great as possible without impinging on the quality of the thesis, this is a goal the author

believes has been achieved. Research into this many different leagues with this many seasons has rarely been completed in the literature with only one possible exception (Jespersen & Pedersen, 2018).

With regards to the data and method selection, the author settled on the Theil measure as the proxy for uncertainty of outcome and measured its correlation with attendance demand through the use of multiple OLS regression. The Theil measure seemed particularly promising as a proxy for uncertainty of outcome as the literature showed both support and contradiction to the UOH with its usage, therefore the author wanted to help narrow down the correlation of the measure. The author decided against using the tobit model which has been used extensively in the research as the large negative aspects of the model seemed to outweigh the positive benefits. The benefits of the model are mainly its increased statistical power as it compiles 'censored' data points, which are within a certain threshold of the maximum stadium capacity, usually this threshold is set at 95% and is hit quite often with the Premier League boasting 98.7% percent attendance of capacity in 2022-23 (D'Urso & Alexander, 2023). A disadvantage of the model is that it requires data on the stadium capacities of stadiums, and with over a hundred different stadiums and with clubs playing in as many as three different stadiums during the study, this data would be very time and research draining. Furthermore, the assumptions made in the tobit model are stronger than in the linear OLS regression model that the author chose, these assumptions are discussed in detail in the data and method selection section of the literature review.

The author used derbies, game importance for the home and away team, and the home team's probability of a victory as explanatory variables to increase the goodness of fit of the model. The results computed lead us to reject the UOH in all five leagues, as for the UOH to hold, the authors would need positive and significant coefficients for the Theil measure. The results indeed show the opposite is true, that attendance demand decreases with increasing uncertainty of outcome in the English Premier League, the Spanish La Liga, the Italian Serie A and the French Ligue 1. The author encountered no significant coefficients for the German Bundesliga which could be a result of very modest variance in attendance values. The coefficients of home team probability of a victory followed the same coefficient signs and

significance. Furthermore, the author found that the derby variable is positively and significantly correlated with attendance in Spain, France and Italy, with a positive yet insignificant result in the other two leagues. The importance variables proved to be irrelevant as the coefficients were very negligible even when the author encountered a significant level of the p-value.

It must be noted that the results do not signify any sort of causal relationship between the variables as the author did not conduct an experiment with both a control and treatment group and for the differences between them to be examined. Given the type of research the author conducts this was never an option, however future research might be able to research this effect albeit with large government and/or corporate funding. The research offers new insights into the UOH as the samples are some of the largest collected and the scope captures all of the top five leagues.

5.2 LIMITATIONS OF THE STUDY

Omitted variable bias is likely to be present in the regression model as the author excluded many variables that could further explain variances in attendance demand. For example, the distances between teams was not taken into account although it has been shown to have a negative and significant effect on attendances in the literature as away fans incur more travel and time costs when the home team is further away (Jespersen & Pedersen, 2018). Furthermore, dummy variables for the month in which the game took place were not included, these could also explain differences in attendance through the weather, alternative seasonal attractions and varying interest. It has been shown that games played in December when Christmas shopping is an alternative form of entertainment and the weather is harsh have lower attendances than April and May, when the weather is warm and games are more important and the season draws to an end (Forrest & Simmons, 2002).

A limitation to the data collection technique are the biases in betting odds not being corrected. Betting odds have been shown to be biased through three biases; home-away, short odds-long odds and different levels of club support (Forrest & Simmons, 2002). These biases were not corrected in the sample due to time constraints and could have a negative impact on

the statistical power of the model as these odds were used to calculate two of the explanatory variables (ProbH and Theil measure). Furthermore, the data on derbies does not capture the strength of the derby, how important fans view derbies could affect their desire to watch the game live. For example Fulham vs Chelsea and Tottenham vs Chelsea are both classed as derbies, even though Chelsea fans would admit that their rivalry against Tottenham is more anticipated and thus could lead to more attendance demand. Not capturing this difference could lead to a dilution of the coefficient of derby as weaker rivalries are still classed as derbies. The importance variables might also be inconsistent through time and leagues as stated on the FiveThirtyEight website: "...for some leagues, the outlook only considers winning the league, while other leagues incorporate the possibility of being promoted or relegated, or qualifying for the Champions League"(FiveThirtyEight, 2020).

5.3 INTERNAL AND EXTERNAL VALIDITY

As for the internal validity of the results, given the high omitted variable bias expected, the author cannot be too sure that the results are valid. Another way of measuring how UO affects attendance demand could be through a survey of match-going fans in the leagues studied. Further research as conducted by Pawlowski (2013) into how supporters perceive UO and its effect on their desire to watch a match live is recommended. Given the varying results between leagues, it would be unwise to make generalizations to claim that the results here are applicable to other leagues such as the Dutch Eredivisie or the Egyptian Premier League for example. There are a myriad of cultural, societal and economic factors that affect the preferences of football supporters from different parts of the world with regards to football games. Further research into the specific leagues in question would be the only way to conclusively say whether UO is correlated with attendance demand in these places.

5.4 CONCLUSION

As the author rejected the null hypotheses for the four leagues excluding the German league, the author suggests that directors of these leagues do not use regulatory measures such as salary caps to enhance UO for better attendance numbers. It must be noted that an

extremely large proportion of a club in the top five leagues' annual turnover comes from broadcasting revenues and sponsorship income (Deloitte, 2022). As gate receipts are becoming less relevant to a club's overall financial health, research should be undertaken to find how best to increase income broadcasting revenues.

This research has several implications and applications for various stakeholders, such as board members of competitive leagues, football clubs or regulators. The results found in this study suggest that stakeholders who want to increase attendance should not expect policies that increase the competition of the league to have a positive effect on attendances given. An example of a policy aimed at decreasing differences in quality between teams are wage ceilings which are commonly used in American sports. This research suggests that in the context of European football, these measures should not be taken in an effort to increase attendance.

This research is valuable for future researchers as it has advanced the knowledge of the uncertainty of outcome hypothesis with concrete results in a large scope of study. Furthermore, the data collection and analysis in this thesis will allow for a better understanding of the process of analysis in the topic. The study has employed a novel parameter, being the importance variables, which were shown to have a near irrelevant effect on attendance, perhaps with a different measure of a game's importance a more concrete coefficient might be observed.

All in all, the correlations captured in this research regarding uncertainty of outcome and attendance demand call for further investigations on this relationship. The negative correlation between the Theil measure and attendance demand, while controlling for the relevant variables, suggests that increasing uncertainty of outcome is not a measure suitable for increasing stadium attendances. This is significant for leagues attempting to increase their attendances.

6. REFERENCES

Benz, M-A., Brandes, L., Franck, E. (2009). Do soccer associations really spend on a good thing? Empirical evidence on heterogeneity in the consumer response to match uncertainty of outcome. Contemporary Economic Policy, 27(2), 216-235. Borland, J. (1987). The demand for Australian Rules football. Economic Record, 63(3), 220-230.

Borland, J., & Macdonald, R. L. (2003). Demand for sport. *Oxford Review of Economic Policy*, *19*(4), 478–502. https://doi.org/10.1093/oxrep/19.4.478

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. Econometrica, 47(5), 1287-1294.

Buraimo, B., & Simmons, R. (2008). Do Sports Fans Really Value Uncertainty of Outcome? Evidence from the English Premier League. *International Journal of Sport Finance*, *3*(3), 146–155. http://clok.uclan.ac.uk/4676/

Buraimo, B., Simmons, R. (2015). Uncertainty of outcome or star quality? Television audience demand for English Premier League football. International Journal of the Economics of Business, 22(3), 449-469.

Cairns, J., Jennett, N. and Sloane, P.J. (1986), "The Economics of Professional Team Sports: A Survey of Theory and Evidence", Journal of Economic Studies, Vol. 13 No. 1, pp. 3-80. https://doi.org/10.1108/eb002618

Coates, D., Humphreys, B. R. (2010). Week to week attendance and competitive balance in the National Football League. International Journal of Sport Finance, 5(4), 239-252.

Coates, D., Humphreys, B. R., Zhou, L. (2014). Reference- Dependent Preferences, Loss Aversion, and Live Game Attendance. Economic Inquiry, 52(3), 959-973.

Deloitte. *Annual Review of Football Finance 2022*. (2022). Deloitte. https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/sports-business-group/deloitte-uk-annual-review-of-football-finance-2022.pdf

D'Urso, J., & Alexander, D. (2023, June 15). *Revealed: How every premier league club's attendances compare historically*. The Athletic. https://theathletic.com/4593912/2023/06/15/premier-league-attendances-compared/

Falter, J. M., Pérignon, C., Vercruysse, O. (2008). Impact of Overwhelming Joy on Consumer Demand: The Case of a Soccer World Cup Victory. Journal of Sports Economics, 9(1), 20-42.

FiveThirtyEight. (2020, July 2). *How our club soccer predictions work*. https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work/

Forrest, D., & Simmons, R. (2002). Outcome Uncertainty and Attendance Demand in Sport: The Case of English Soccer. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *51*(2), 229–241. http://www.jstor.org/stable/3650322

Forrest, D., Beaumont, J., Goddard, J., Simmons, R. (2005). Home advantage and the debate about competitive balance in professional sports leagues. Journal of Sports Sciences, 23(4), 439-445

Gujarati, D. N., & Porter, D. C. (2009). Basic Econometrics (5th ed.). McGraw-Hill Education.

Hart, R. G., Hutton, J. C., & Sharot, T. (1975). Correction: A statistical analysis of association football attendances. *Applied Statistics*. https://doi.org/10.2307/2347091

Knowles, G., Sherony, K., & Haupert, M. (1992). The Demand for Major League Baseball: A Test of the Uncertainty of Outcome Hypothesis. *The American Economist*, *36*(2), 72–80. http://www.jstor.org/stable/25603930

Kuypers, T. (1996) The beautiful game?: an econometric study of why people watch English football. Discussion Paper 96-01. Department of Economics, University College London, London

Lemke, R., Leonard, M., Tlhokwane, K. (2010). Estimating attendance at Major League Baseball games for the 2007 season. Journal of Sports Economics, 11(3), 316-348.

Martins, A. A., & Cró, S. (2016). The demand for football in Portugal. *Journal of Sports Economics*, *19*(4), 473–497. https://doi.org/10.1177/1527002516661602

Mills, B., Fort, R. (2014). League- Level Attendance and Outcome Uncertainty in US Pro Sports Leagues. Economic Inquiry, 52(1), 205-218.

Neale, W. C. (1964). The Peculiar Economics of Professional Sports: A Contribution to the Theory of the Firm in Sporting Competition and in Market Competition. *The Quarterly Journal of Economics*, *78*(1), 1–14. https://doi.org/10.2307/1880543

Pawlowski, T. (2013). Testing the uncertainty of outcome hypothesis in European professional football. *Journal of Sports Economics*, *14*(4), 341–367. https://doi.org/10.1177/1527002513496011

Pawlowski, T., Anders, C. (2012). Stadium attendance in German professional football - the (un)importance of uncertainty of outcome reconsidered. Applied Economics Letters, 19(16), 1553- 1556.

Pawlowski, T., Nalbantis, G. (2015). Competition format, championship uncertainty and stadium attendance in European football – a small league perspective. Applied Economics, 47(38), 4128-4139

Peel, D., & Thomas, D. (1988). OUTCOME UNCERTAINTY and THE DEMAND FOR FOOTBALL: AN ANALYSIS OF MATCH ATTENDANCES IN THE ENGLISH FOOTBALL LEAGUE. *Scottish Journal of Political Economy*, *35*(3), 242–249. https://doi.org/10.1111/j.1467-9485.1988.tb01049.x

Peel, D., Thomas, D. (1992). The demand for football: Some evidence on outcome uncertainty. Empirical Economics, 17(2), 323-331.

Peel, D., Thomas, D. (1997). Handicaps, outcome uncertainty and attendance demand. Applied Economics Letters, 4(9), 567-570.

Rottenberg, S. (1956). The Baseball Players' Labor Market. *Journal of Political Economy*, *64*(3), 242–258. http://www.jstor.org/stable/1825886

Schmidt, M. B., & Berri, D. J. (2001). Competitive balance and attendance. *Journal of Sports Economics*, 2(2), 145–167. https://doi.org/10.1177/152700250100200204

Swaen, B. (2022). What is a conceptual framework? | Tips & examples. *Scribbr*. https://www.scribbr.com/methodology/conceptual-framework/

Szymanski, S. (2003). The economic design of sporting contests. *Journal of Economic Literature*, 41(4), 1137–1187. https://doi.org/10.1257/jel.41.4.1137

Theil, H. (1967) Economics and Information Theory. North-Holland Publishing Company, Amsterdam.

Torres, O. (2007). *Linear Regression using Stata* [Slide show]. Princeton.edu. https://www.princeton.edu/~otorres/Regression101.pdf

Welki, A. and Zlatoper, T. (1994) U.S professional football: the demand for game-day attendance in 1991. Mang. Decsn Econ., 15, 489-495.