NOVA IMS Information Management School

MGI

Master Degree Program in
Information Management

Estimating the Determinants of Spanish La Liga Teams' Performance

Bruno Sousa Esteves

Master Thesis

presented as partial requirement for obtaining the Master Degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Estimating the Determinants of Spanish La Liga Teams' Performance

by

Bruno Sousa Esteves

Dissertation presented as requirement for obtaining the Master's degree in Information Management, with a specialization in Business Intelligence

Advisor: Professor Bruno Miguel Pinto Damásio

July 2024

Statement of Integrity

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, July 2024

Abstract

This study investigates the determinants of football teams' performance from both club level and match level perspectives, using panel data analysis. At the club level, variables such as squad size, average age of the players, percentage of foreign players, market value of teams, payroll, net transfers, season attendance, cup games and European games played were analyzed. In the analysis, linear regression methods were used and, although the variables did not show statistical significance individually, the overall model proved that the variables together explain a substantial amount of the variation in team performance. This finding highlights how important an exhaustive approach is to a football club's strategic planning. At the match level, the analysis revealed that gaining an advantage in the first half significantly increases the likelihood of a favorable full-time result. Additionally, home and away shots on target, fouls committed by the home team, corners and red cards were significant factors influencing match outcomes.

Keywords

Football, La Liga; Teams' Performance; Data Analysis; Panel Data; Fixed Effects Regression

Index

Tables Index

Introduction

Football is arguably the world´s most popular sport of all time and it serves as a showcase of tactics, abilities and organizational strengths making it a compelling topic for study. In this regard, the Spanish La Liga has considerable value for its technical expertise, devoted supporters and the sustained supremacy of its high level teams.

The goal of this study is to investigate the elements impacting league performance from two separate perspectives, which excels a general examination of team performance in the Spanish La Liga. Similar to the macro viewpoint, the first perspective is a detailed investigation of the causes affecting football clubs throughout the course of the preceding five seasons. This method offers a comprehensive standpoint of the more general organizational and structural elements of a club's management, that lead to a team's success in the league.

The other perspective focuses on an examination of how teams secure match victories, as in a micro viewpoint. This analysis looks at the specifics of each match played over the past three seasons, analyzing the match statistics and player performances. By researching each game in detail, the objective is to discover the factors that influence wins and impact a teams' performance throughout a season.

The primary objective of this thesis is to analyze and understand the determinants influencing the performance of Spanish La Liga teams. In pursuit of this purpose, the following research question guide this study, for the club level analysis perspective:

What are the most important organizational and financial factors that influence the competitive performance of football clubs in La Liga over the past five seasons?

Then, the match level analysis perspective focuses on matches factors:

What are the most important match statistics and tactical variables that significantly influence the outcome of La Liga matches over the past three seasons?

The research holds a great significance within the context of sports management, business intelligence and football analytics and understanding the factors influencing La

Liga teams' performance can greatly help in making decisions within football organizations. Analyzing both perspectives, the research attempts to fulfill the gap between the strategic planning of the club management variables, with the game by game nature of match winning performance, which brings value to this study.

To investigate these research questions, this study uses the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. CRISP-DM gives a structured framework for guiding the data analysis process and therefore adjust with the business intelligence approach adopted in this research. First, a Literature Review is performed, in order to analyze previous studies related to the topic, especially in the context of the Spanish La Liga. Following is the Business Understanding and Data Understanding and Preparation, where it is mentioned the importance of La Liga and why it was chosen, as well as an overview of the data and the respective descriptive statistics.

The modeling section explains the models that are going to be used, and in the Results and Discussion part, as the name suggests, the results from the models applied are presented and analyzed. In the Conclusion, a summary of the findings is presented, as well as the limitations that this research might encounter. Finally, the References section is included where the sources from the research are shown, followed by the Annexes providing complementary resources.

Literature Review

The determinants of team performance in football have been widely studied before, giving insights into what leads to the success of the clubs. A study by Szymanski and Kuypers (1999), where English football clubs were analyzed, discovered that the presence of strong financial capabilities, specifically club revenue and wage expenditure, has a major impact on the performance of the team. Other studies also backed this, as Barajas and Rodríguez (2010) demonstrated the relationship between financial resources and achieving success in La Liga, highlighting the critical role financial resources play in reaching better league positions.

Squad composition is also an important determinant of performance, as highlighted by Carmichael, Thomas and Ward (2000) who showed the significance of player quality, in this case measured by international experience and transfer fees, in determining team success. This goes along with the research conducted by Lago-Peñas and Sampaio (2015) on La Liga matches, where they found that player capacities and abilities and team strategies were important predicting factors of match outcomes. Their study revealed that possession, passing accuracy and shots on target are important performance indicators in Spanish football.

The influence of managerial decisions has also been studied, with Tena and Forrest (2007) investigating the repercussions of coaching changes in La Liga, where their findings indicated that maintaining manager stability typically results in improved team performance. Their study revealed that frequent changes in management can disturb team cohesion and have an impact on results, highlighting the importance of continuity for sustained success.

Match level determinants, such as home advantage, have been examined by Pollard (2006), who found that home teams generally perform better due to familiar conditions, crowd support and reduced travel fatigue levels. Additionally Lago-Peñas, Gómez and Pollard (2017) explored how situational variables impact match performance, with halftime scores proving to have a considerable effect on the result.

Team performance can also be influenced by discipline and player behavior, according to a study by Reilly and Gilbourne (2003), where the impact of red cards was analyzed and discovered that teams which are forced to play with fewer players as a result of red card sanctions are usually at great disadvantage. Additional studies, as the one by García-Rubio et al. (2020) which focused on La Liga matches, also discovered that fouls and red cards can have a negative effect on a team's performance, reinforcing the significance of discipline on the pitch.

While prior research on football teams' performance provides valuable insights into tactics and game performance, there is still a gap in the literature addressing an integrated analysis from the perspectives of successful seasons and club management. Previous research has examined quantitative measures, tactical plans and forecast models, but not many have focused on the overall success of football teams, which includes both on-field success and management practices, that could include determinants like recruitment strategies, financial planning and youth training initiatives.

This study is particularly valuable as it incorporates the most recent data from the latest seasons, ensuring that the analysis reflects the current dynamics in Spanish La Liga football. By using up-to-date information, this research provides relevant insights into the factors influencing team performance, making it highly pertinent for club management, coaches and analysts from professional football, since the use of recent data enhances the accuracy and usability of the findings.

Business Understanding

La liga was founded in 1929, and currently the best 20 Spanish teams participate in it, following a system of promotion and relegation with Spain's second division, La Liga 2. Each season, teams compete with each other in a home and away matches format and the team with the highest number of points at the end of the season is crowned the champion, while the three bottom ranked teams are relegated to the LaLiga 2 and replaced by the best three teams from that division.

This study focuses on La Liga as the field of research, for its unique features and impact on football both domestically and internationally. La Liga offers a wide range of data on individual statistics, team dynamics and match outcomes, which is essential for extensive analysis, providing valuable insights and powerful econometric models. The recurrent success of Spanish teams in international competitions demonstrates the skills and strategic management that La Liga has, and similar success could be obtained from other teams and leagues, trying to obtain identical achievements.

La Liga is known for being very competitive, especially in the spots aiming for the european qualification and the spots avoiding relegation. The top spots are usually more predictable, with the favorite teams in Spain generally reaching the top positions. Nevertheless, the champions are constantly changing, and regularly some less powerful teams enter into the top three positions. This rivalry provides a valuable atmosphere for examining how different elements affect a team's performance under specific conditions, and studying these dynamics can reveal how teams modify their strategies to achieve their objectives. The transfer market in the Spanish league is usually very busy, with great players moving in, out and within the league. This dynamic transfer scene presents a research opportunity and studying how transfers affect team performance and league competitiveness elucidates on how clubs manage the challenges of assembling and sustaining a competitive squad.

The detailed match statistics available for the league games offer a rich source of information for analyzing various aspects of the game, including team tactics, player contributions and match outcomes. Using this information enables researchers to inquire into how various factors such as shots on target, ball possession and defensive plays relate to on field achievements, and therefore gaining important insights in football analytics. The league is also very famous for its tactics and different playing styles such as possession based football or counter attacking strategies. This variety in tactics makes La Liga a great case study for understanding how different game plans affect match results and studying these aspects can improve the perception into the factors that lead to success in football.

Last but not least, the passionate supporters of La Liga football teams add an additional viewpoint to the study, showing how crowd involvement and support may affect team performance. Examining the supporters' influence enables us to investigate the ways in which their participation and encouragement influence the league's atmosphere and spirit of competition.

These elements make La Liga an interesting and valuable topic for examination, providing chances to understand the complex mechanisms of professional football and establish a solid foundation for effectively addressing the research questions this study attempts to answer.

Data Understanding and Preparation

Club Level Analysis

Data Collection and Preparation

The estimation of the determinants influencing performance at the club level is based on data collected from the seasons spanning 2018-2019 to 2022-2023. This five year period provides an overview of team performances allowing for the identification of trends and patterns within the league.

The dataset comprises records from 26 teams that participated in La Liga during the referred seasons. Given the league's promotion and relegation system, the dataset forms an unbalanced panel, which means that some teams have missing data entries for one or more seasons within the study timeframe.

Most of the data originates from *Transfermarkt*, a database known for its insights into team squads, player statistics, match outcomes and financial information like transfer spending and revenues. *Transfermarkt* is renowned for its accurate coverage of football related data. Additionally payroll details were obtained from *Capology*, a website specializing in sports finance that offers verified information on player salaries and team payrolls.

Preparing the data is crucial before diving into analysis to guarantee that the data is organized, clean and ready for applying the econometric models. To better understand team performance across seasons, team and season were setted as the index of the dataset. This indexing structure helps in arranging data and making it easier to access and manage when conducting time series or panel data analyses.

Due to the COVID-19 restrictions, the season attendance for the 2020-2021 season was severely changed from the normal attendance values. Therefore, in order for these values to not alter the analysis of this variable, the mean from the other seasons was calculated, and imputed on the 2020-2021 season.

Lastly, the dependent variables League Position and Point Percentage were converted into Log League Position and Log Point Percentage, respectively. This way, the volatility between the teams is not that noticeable.

Descriptive Statistics

Table 1 - Descriptive statistics for the club level perspective

Variables	Description	Mean	Standard Deviation	Maximum	Minimum
Number Players in Squad	The total number of players registered in the team squad	36.52	4.19	47	28
Average Age (Years)	The average age of all players in the squad	25.84	0.98	28.3	23.6
Number of Foreigners	The number of players in the squad who are from countries other than the one in which the team is based	13.42	5.41	24	1
Percentage of Foreigners in Squad (%)	The proportion of foreign players in the squad, expressed as a percentage of the total number of players	36.45	13.86	63.33	3.23
Total Market Value (M €)	The cumulative estimated market value of all players in the squad, measured in millions of euros	271.89	257.63	1016	45.65
Payroll (M €)	The total annual wages paid to all players in the squad, measured in millions euros	57.78	86.70	357	15.29
Net Transfers (M €)	The net expenditure on player transfers, calculated as the total amount spent on incoming transfers minus the total amount received from outgoing transfers, measured in millions of euros	-5.35	40.16	108.2	-218.7
Season Attendance (%)	The average percentage of stadium capacity filled by spectators during the season	69.91	10.82	91.4	40.3
Number of Cup Games	The total number of matches played in domestic cup competitions during the season	4.28	2.07	10	1
Number of European Games	The total number of matches played in European competitions (UEFA Champions League, Europa League or Conference League) during the season	3.72	5.34	16	0
Points	The total number of points earned in the league during the season, based on wins (3 points), draws (1 point) and losses (0)	51.76	15.53	88	25

Log Point Percentage (%)	The percentage of the maximum possible points that the team has earned in the league during the season, in log	45.4	13.63	77.19	21.93
Log League Position	The team's final standing in the league table at the end of the season, in log	-	-	-	-

Squad Composition and Demographics

Most teams usually have around 33 to 39 players on their squad with an average of 36 players and minimal variation, showing that team sizes are generally similar.

The typical age of players hovers around 26 years old, with an age range indicating that most teams tend to have players in their mid 20s. In fact, the majority of teams have player averages ranging from 25 to 26 years old, with a deviation of less than one year and it appears that teams maintain a very similar average age among their players.

There is significant variation in the number of foreign players across teams, highlighting different strategies regarding the recruitment of international talent. Most teams have between 10 to 17 foreign players and a mean of around 13 players. Consequently, the proportion of foreign players in squads also varies widely, with some teams having predominantly local players while others rely heavily on foreign talent. Foreign players constitute about 36% of the squads on average, with a significant variability.

It is important to note that the teams with lower numbers of foreigners (Athletic Bilbao, Club Atlético Osasuna and Real Sociedad), are teams from the Basque Country (autonomous community) and have a long tradition of being cautious with the foreign recruitment of players, even across other regions of Spain.

Financial Metrics

The total market value of teams has a high variation, with some teams valued at as low as €45.65 million and others exceeding €1 billion. The high variability reflects disparity in the financial strength and resources of the clubs. Most of the teams are valued between €96 million and €319 million, suggesting that only a few of the teams are valued highly and those are the ones that usually win the championship and get better results.

Similarly to the total market value of teams, payroll expenditures also show significant variability (higher than the mean itself), reflecting once again the different financial capabilities and investment levels in player salaries among the clubs.

The negative mean in the net transfers suggests that, on average, teams spend more on acquiring players than they receive from selling the players. There is a wide range, indicating various transfer strategies between the clubs and the variability is also high.

Attendance percentages vary moderately, which could be influenced by stadium capacity, supporters base size and team performance. However, most of the teams have attendances between 63.75% and 78.59%.

Performance Metrics

Most teams play between 3 to 5 cup games, and an average of about 4 cup games per season, with a relatively narrow distribution around this mean, which is expected since the spanish cup (Copa del Rey) is a variety of matches between the spanish first division and other spanish lower divisions, and usually the winner is from the first division.

As for the European journey, the participation varies greatly, with the majority of the teams not qualifying for European tournaments while others have extensive runs. However, the teams that qualify for European competitions usually do well in the different competitions.

There is a broad range in points accumulated, reflecting the competitive nature of the league. Teams generally earn between 41 to 60 points, with an average around 52. The percentage of points won out of possible points shows variability in team performance across seasons. On average, teams achieve about 45% of possible points, indicating moderate success.

Match Level Analysis

Data Collection and Preparation

The dataset gathered for the match-level perspective contains data about 25 different teams across the three most recent seasons, from 2020-2021 to 2022-2023 seasons. Again, due to the system of promotion and relegation, the dataset forms an unbalanced panel.

The data was retrieved from the website *Football-Data.co.uk*, which is a website that offers statistics on football games in various leagues and competitions The website provides an array of data such as match results, team performance metrics, player stats and historical information, meaning that this data has great value for conducting rigorous analyses and developing econometric models to explore the various determinants affecting match results.

After retrieving the data, it needed to be prepared, so, the columns of ‘Date’, ‘Division’ and ‘Time’ were dropped from the dataset as they are not directly relevant to the analysis - the focus is on the match results and performance metrics rather than the specific timing details.

Additionally, a column was created with the name ‘Season’ to assure that each match can be accurately placed within its corresponding season. Then, similarly to the club-level approach, a multi index was setted as ‘Team Pair’ (home and away teams) and ‘Season’. Again, this indexing structure allows for detailed analysis of match outcomes and team interactions across different seasons.

Finally, in order to facilitate the analysis of match outcomes, a label encoder for the full-time and half-time results was created. This encoder transforms the categorical outcomes (win, draw or loss) into numerical codes (0 for an away win, 1 for a draw and 2 for a home win). These encoded variables, named ‘Full_time_result_encoded’ and ‘Half_time_result_encoded’, are then used in the regression models to quantify the match outcomes.

Descriptive Statistics

Table 2 - Descriptive statistics for the match level perspective

Variables	Description	Mean	Standard Deviation	Maximum	Minimum
Full-time Home Goals	The number of goals scored by the home team by the end of the match	1.41	1.22	6	0
Full-time Away Goals	The number of goals scored by the away team by the end of the match	1.09	1.06	6	0
Half-time Home Goals	The number of goals scored by the home team by the end of the first half	0.64	0.81	4	0
Half-time Away Goals	The number of goals scored by the away team by the end of the first half	0.48	0.71	4	0
Home Shots	The total number of shots taken by the home team during the match	12.80	4.91	36	2
Away Shots	The total number of shots taken by the away team during the match	10.38	4.36	35	0
Home Shots on Target	The number of shots taken by the home team that were on target	4.40	2.42	17	0
Away Shots on Target	The number of shots taken by the away team that were on target	3.52	2.1	15	0
Home Fouls	The number of fouls committed by the home team during the match	13.31	4	29	2
Away Fouls	The number of fouls committed by the away team during the match	12.97	4.16	30	1
Home Corners	The number of corner kicks awarded to the home team during the match	5.06	2.85	19	0
Away Corners	The number of corner kicks awarded to the away team during the match	4.17	2.56	15	0
Home Yellow Cards	The number of yellow cards received by the home team during the match	2.35	1.53	9	0
Away Yellow Cards	The number of yellow cards received by the away team during the match	2.52	1.58	8	0

Home Red Cards	The number of red cards received by the home team during the match	0.13	0.36	2	0
Away Red Cards	The number of red cards received by the away team during the match	0.13	0.36	2	0
Full-time Result Encoded	The result of the match encoded numerically (0 for an away win, 1 for a draw and 2 for a home win)	1.16	0.84	-	-
Half-time Result Encoded	The result of the match at half-time encoded numerically (0 for an away win, 1 for a draw and 2 for a home win)	1.11	0.73	-	-

Goals Scored

Home teams typically score around 1 to 2 goals per match, with some variability, while Away teams generally score fewer goals than home teams, with most matches seeing 0 to 2 away goals. As for half-time goals, both home and away goals are generally low, with most matches having 0 to 1 goal and with the mean for half-time away goals being slightly lower than half-time home goals mean.

Shots made

Home teams usually take around 9 to 16 shots per match, with an average of about 13 shots. Of those shots, around 3 to 6 shots are on target per match, with an average of about 4. Away teams take fewer shots than home teams, with a typical range of 7 to 13 shots, and generally away teams have 2 to 5 shots on target per match.

Discipline

The number of fouls that both teams commit are very similar, with home teams committing around 11 to 16 fouls per match and away teams committing around 10 to 16. Regarding the number of cards received, both teams also receive a similar number of yellow cards (1 to 3 yellow cards per match) and red cards, with most matches having no red cards received.

Corners awarded

Home teams usually get around 3 to 7 corners per match, with a mean of 5 corners and high variability. As for away teams, generally they get fewer corners than home teams, typically 2 to 6 per match, with a mean of 4 corners and also high variability.

Result

The encoded full-time result and half-time result indicates that the outcomes are evenly spread among the possible results (home win, draw, away win), although it can be noticed a slight deviation towards home win, since the mean is slightly greater than 1 (again, 0 is encoded for away win, 1 for draw and 2 for home win).

Modeling

In this study's methodology, for both perspectives of club level and match level analysis, the same approach was used: three distinct econometric models were employed to understand the determinants of the dependent variables in both cases - Pooled Ordinary Least Squares (OLS), Fixed Effects and Random Effects.

The initial step of this analysis is to perform a Pooled OLS model, where it is assumed that no team/match specific factors can affect the dependent variables studied, that are not captured by the independent variables. The model also assumes that exists homogeneity, meaning that all teams and matches observed are identical in their behavior. Obviously, these are very strong assumptions and not very realistic. Previous studies usually begin with the Pooled OLS model as a guideline, as Wooldridge (2010) suggests, since it provides a straightforward and intuitive starting point for further analysis.

The Fixed Effects is useful when the unobserved teams/match effects are correlated with the independent variables. The assumptions of the Fixed Effects model include controlling for all time-invariant characteristics of the teams/matches, assuming that effects of teams/matches are correlated with the independent variables and that the error term is not serially correlated. To determine which model to use between the Fixed Effects and the Pooled OLS, a statistical test called F-test is performed to check if the teams/matches effects are significantly different from zero, and rejecting the null hypothesis means that Fixed Effects is a better fit compared to Pooled OLS.

The Random Effects model suggests that the cross-sectional units' specific effects of teams/matches are spread randomly, have no correlation with the independent variables and the model addresses the time-varying variability within and between teams/matches. The Random Effects model relies on two main assumptions: efficiency (i.e., if the assumptions are met, Random Effects is more efficient than Fixed Effects because it takes into account both within-unit and between-unit variations) and random matches/teams effects (i.e., random effects that are uncorrelated with the regressors). Typically, a Breusch-Pagan Lagrange Multiplier test is used to compare the Random Effects and Pooled OLS models and rejecting the null hypothesis indicates that the Random Effects model is preferred.

The final decision in which model to use is certainly between Fixed Effects or Random Effects and to choose between them, a Hausman test is conducted. This test checks if the unique errors (teams/matches effects) are correlated with the regressors. The null hypothesis of the Hausman test assures that the preferred model is Random

Effects, while the alternative hypothesis shows that the Fixed Effects is the model to use, as it is more appropriate since the teams/match effects are correlated with the regressors.

Lastly, one important factor to note when working with panel data is to guarantee no heteroskedasticity. To account for this, clustered standard errors are used in the models as they provide more robust standard errors, enhancing the reliability of statistical analysis. These standard errors assist in handling heteroskedasticity (non-constant variance of errors) but also autocorrelation (correlation of errors across time within the same cluster) which could result in wrong estimates and biased conclusions if not properly accounted for.

Results and discussion

Club Level Analysis

For the club level perspective, two dependent variables were chosen: *Log League Position* and *Log Point Percentage*. The two variables are, obviously, strongly correlated since the team with the highest point percentage will have the best league position and be crowned the champions. The reason behind the choice of the two variables instead of choosing just one is the comprehensive analysis that can be retrieved from using both variables, by cross validating the findings and ensuring robustness in the research. Also, although related, the variables show slightly different angles - *Log League Position* focuses on the team's rank relative to others, while *Log Points Percentage* provides a more direct measure of the team's success in terms of points. For both analyses, the independent variables of *Number of Foreigners* and *Points* were not considered for the modeling section, due to the very high correlation with the variables *Percentage of Foreigners in Squad (%)* and *Point Percentage (%)*, respectively.

Log League Position

After running the respective tests (F- test, Breusch-Pagan Lagrange Multiplier test and Hausman test), it was concluded that the Fixed Effects model was the most appropriate for the dependent variable of Log League Position. However, the results were in part disappointing, since no variable was considered to be statistically significant, according to the model.

Although the overall model is statistically significant and explains a substantial portion of the variability in league positions, both within and between teams, the analysis revealed that only the log point percentage variable (the other dependent variable that will be studied afterwards, here as an independent variable) is highly significant predictor of log league position, which is rather obvious, confirming that teams with better point percentages tend to have better league positions.

Other factors such as the number of players, average age, percentage of foreigners, total market value, payroll, net transfers, season attendance, number of cup games, and number of European games did not show statistically significant effects on league position. This suggests that these variables do not have a strong direct impact on league standings when accounting for team-specific characteristics.

Table 3 - Fixed Effects Model Results for the dep. variable Log League Position

Independent Variable	Parameter	Std. Err.	T-stat	P-value
Number Players in Squad	-0.0076	0.0069	-1.0977	0.2764
Average Age (Years)	-0.0136	0.0329	-0.4127	0.6812
Percentage of Foreigners in Squad (%)	0.0044	0.0041	1.0678	0.2896
Total Market Value (M €)	-0.0002	0.0005	-0.4364	0.6640
Payroll (M €)	-5.13E-07	-5.20E-07	-0.9869	0.3274
Net Transfers (M €)	0.0008	0.0006	1.3946	0.1679
Season Attendance (%)	-0.0034	0.0025	-1.3907	0.1691
Number of Cup Games	-0.0202	0.0119	-1.7001	0.0940
Number of European Games	-0.0079	0.0086	-0.9198	0.3611
Log Point Percentage (%)	-1.9048	0.1805	-10.554	0.0000

Log Point Percentage (%)

Using the Log Point Percentage as a dependent variable, did not bring any different results than using the previous variable. Once again, neither one of the variables showed statistical significance in determining the performance of football teams, at significance levels of 1%, 5% and 10%.

These results from both dependent variables could be explained for different reasons. One is the fact that, although five seasons may seem enough time for a research with these variables, it may not be able to capture enough relationships with the performance variables studied in this thesis. Therefore, if more seasons were added to study the results could be different, and the independent variables here studied could have shown some impact in the dependent variables. Another reason could simply be that these variables, alone, are not significant at determining the performance of teams in the Spanish La Liga.

Table 4 - Fixed Effects Model Results for the dep. variable Log Point Percentage (%)

Independent Variable	Parameter	Std. Err.	T-stat	P-value
Number Players in Squad	-0.0056	0.0034	-1.6181	0.1106
Average Age (Years)	0.0262	0.0189	1.3884	0.1698
Percentage of Foreigners in Squad (%)	-0.0005	0.0017	-0.2644	0.7923
Total Market Value (M €)	0.0002	0.0002	0.9429	0.3493
Payroll (M €)	-2.29E-07	1.64E-07	-1.3985	0.1668
Net Transfers (M €)	0.0003	0.0002	1.3054	0.1964
Season Attendance (%)	-0.0020	0.0014	-1.4151	0.1619
Number of Cup Games	-0.0048	0.0046	-1.0365	0.3039
Number of European Games	-0.0047	0.0044	-1.0666	0.2902
Log League Position	-0.3497	0.0373	-9.3628	1.32E-13

Match Level Analysis

In the match level perspective, the previously mentioned tests were again conducted to determine which model should be used. With a p-value of less than 0.05, obtained through the Hausman test, the most appropriate model was again the Fixed Effects model. In this perspective, it was possible to obtain some important inferences for the validation of this study.

The results shown suggest that the half time score is very significant for the outcome of the match, i.e. a team that gains advantage in the half time, usually gets awarded with points (either a win or a draw). The shots on target (both home and away, respectively for each team) are also very important in determining the outcome of the football matches, where more shots on target helps achieve successful outcomes. Conversely, a team receiving one or more red cards can also significantly impact the outcome of the matches and here the home teams seem to be more impacted than the away teams.

Finally, the model analysis shows that home fouls and home corners can impact the match outcome. The number of home team fouls seem to slightly increase the chances of the home team winning points, indicating that an increase in home fouls slightly increases the likelihood of a favorable full-time result for the home team, which may be justified from the home team engaging in tactical fouls or aggressive play in the attempt to regain possession of the ball, contributing to a better home performance.

Conversely, the number of home corners has a negative effect on the home team, meaning that an increase in home corners slightly decreases the likelihood of a good full-time result for the home team. This could mean that even though the home team is getting more corners, they are not taking full advantage of these set-piece opportunities, or it could mean that the opposition is playing a more defensive style.

Table 5 - Fixed Effects Model Results for the dep. variable Full_time_result_encoded

Independent Variable	Parameter	Std. Err.	T-stat	P-value
Half_time_home_goals	0.0315	0.0532	0.5923	0.5539
Half_time_away_goals	0.0247	0.0560	0.4409	0.6595
Half_time_result_encoded	0.4567	0.0755	6.0491	0.0000
Home_shots	-0.0055	0.0085	-0.6447	0.5194
Away_shots	0.0033	0.0087	0.3732	0.7092
Home_shots_target	0.0856	0.0152	5.6175	0.0000
Away_shots_target	-0.0887	0.0175	-5.0777	0.0000
Home_fouls	0.0169	0.0071	2.3861	0.0174
Away_fouls	-0.0072	0.0064	-1.1233	0.2618
Home_corners	-0.0276	0.0110	-2.5167	0.0121
Away_corners	0.0153	0.0120	1.2782	0.2017
Home_yellow_cards	-0.0078	0.0177	-0.4404	0.6599
Away_yellow_cards	0.0025	0.0162	0.1545	0.8773
Home_red_cards	-0.1654	0.0689	-2.3995	0.0167
Away_red_cards	0.0970	0.0701	1.3834	0.1671

Conclusion

In conclusion, it is worth noting that while the individual variables, at the club level, did not show significance on their own, the overall model revealed that these factors collectively play a role in determining performance outcomes. This suggests that club performance is influenced by an interplay of elements, in a complex mechanism. Surprisingly these results deviated from the expected outcomes of this study and diverged from the findings of previous research especially regarding variables like market value of players and payroll, which have consistently been shown to impact team performance.

Conversely, the findings from the models studied in the match level perspective align closely with previous research outlined in the Literature Review. The results indicate that when a team gains an advantage in the first half, it significantly boosts their chances of achieving a positive outcome by full time, highlighting the importance of securing leads in the first half for winning matches. Additionally accurate shots on target emerged as a factor, underscoring the need for teams to enhance their shooting precision and defensive tactics to limit opponents scoring opportunities. Other factors like fouls, corners and red cards also displayed significance in influencing match results as discussed earlier. This perspective showed important insights for teams to take advantage in order to get better performances.

Although this study provides some important conclusions regarding the performance of the Spanish teams, it is essential to keep in mind that the research may contain some considerable limitations. The first limitation is about the detail and completeness of the datasets, because even though they have important factors, some other equally important factors might be missing. For example, factors that are not so easily accessed like players fitness levels or detailed players injury reports, could be very important to analyze. Other factors like managerial and tactical aspects of the teams or socio economic factors of the clubs could also be significant. Therefore, a professional dataset would give the research another level of accuracy. Also, the dataset might cover a “limited” time period, which could be affecting the ability to capture long-term trends, but would also need more computational power.

A longer time frame research and a more extensive dataset would justify more advanced methodologies, such as machine learning algorithms, as these sophisticated methods can help in understanding the variables that impact team performance but also in predicting future outcomes more precisely. Machine learning models excel at processing large amounts of data and uncovering patterns that conventional statistical methods might overlook, thus enhancing decision making and performance optimization in the realm of professional football.

Lastly, information gathered from websites may contain discrepancies or measurement errors, which can skew the outcomes and lead to inaccurate conclusions. While statistical data and models offer insights into team performance, it is crucial to acknowledge that they do not encompass all aspects of football. Factors such as team dynamics, relationships on and off the field, leadership influence and psychological elements are not reflected in the data. Recognizing these limitations is essential when interpreting the findings presented in this thesis and planning future research endeavors.

References

Barajas, A., & Rodríguez, P. (2010). Spanish Football Clubs' Finances: Crisis and Player Salaries. International Journal of Sport Finance, 5(1), 52-66.

Carling, C., Wright, C., Nelson, L.J. and Bradley, P.S. (2014) Comment on 'Performance analysis in football: A critical review and implications for future research'. Journal of sports sciences 32, 2-7.

Carmichael, F., Thomas, D., & Ward, R. (2000). Team performance: The case of English Premiership football. Managerial and Decision Economics, 21(1), 31-45.

Garcia-del-Barrio, P., & Pujol, F. (2007). Hidden monopsony rents in winner-take-all markets–Sport and economic contribution of Spanish soccer players. Managerial and Decision Economics, 28 (1), 57–70.

García-Rubio, J., Gómez, M. A., Lago-Peñas, C., & Ibáñez, S. J. (2020). Effect of match location, team quality and match status on possession and set plays in professional football. Journal of Sports Sciences, 28(12), 1393-1399.

Hausman, J. A. (1978) Specification tests in econometrics, Econometrica, 46, 236–255.

Herold, M., Goes, F., Nopp, S., Bauer, P., Thompson, C. and Meyer, T. (2019) Machine learning in men's professional football: Current applications and future directions for improving attacking play. International Journal of Sports Science & Coaching. 1747954119879350

Lago-Peñas, C., Gómez, M. A., & Pollard, R. (2017). Home advantage in football: Examining the effect of scoring first on match outcome in the five major European leagues. International Journal of Performance Analysis in Sport, 17(1-2), 244-253.

Lago-Peñas, C., & Sampaio, J. (2015). Just how important is a good season start? Overall team performance and financial budget implications. Journal of Sports Economics, 16(2), 140-151.

Pollard, R. (2006). Home advantage in soccer: A retrospective analysis. Journal of Sports Sciences, 24(3), 231-240.

Reilly, T., & Gilbourne, D. (2003). Science and football: A review of applied research in the football codes. Journal of Sports Sciences, 21(9), 693-705.

Sarmento, H., Anguera, M. T., Pereira, A., & Araújo, D. (2014). Talent Identification and Development in Male Football: A Systematic Review. Sports Medicine, 44(7), 873-893.

Sloane, P. (2015). The economics of professional football revisited. Scottish Journal of Political Economy, 62(1), 1. https://doi.org/10.1111/sjpe.12063

Szymanski, S., & Kuypers, T. (1999). Winners and Losers: The Business Strategy of Football. Viking.

Tena, J. D., & Forrest, D. (2007). Within-season dismissal of football coaches: Statistical analysis of causes and consequences. European Journal of Operational Research, 181(1), 362-373.

Wooldridge, J. M. (2010) Econometric Analysis of Cross Section and Panel Data, The MIT Press, Cambridge/ London

Annexes

Panel Data Python Script (Match Level Example)

```
#Pooled OLS
# Define the independent variables
X = data[['Half_time_home_goals', 'Half_time_away_goals',
'Half_time_result_encoded', 'Home_shots', 'Away_shots',
'Home_shots_target', 'Away_shots_target', 'Home_fouls', 'Away_fouls',
'Home_corners', 'Away_corners', 'Home_yellow_cards', 'Away_yellow_cards',
'Home_red_cards', 'Away_red_cards']]

# Define the dependent variable
Y = data['Full_time_result_encoded']

# Add a constant term to the independent variables
X = sm.add_constant(X)

# Fit the linear regression model
model = PooledOLS(Y, X).fit()

# Print the model summary
print(model.summary)
```

```
#Fixed Effects
# Independent variables
X = data[['Half_time_home_goals', 'Half_time_away_goals',
'Half_time_result_encoded','Home_shots', 'Away_shots',
'Home_shots_target', 'Away_shots_target', 'Home_fouls', 'Away_fouls',
'Home_corners', 'Away_corners','Home_yellow_cards', 'Away_yellow_cards',
'Home_red_cards', 'Away_red_cards']]

# Dependent variable
Y = data['Full_time_result_encoded']

# Add a constant term to the independent variables
X = sm.add_constant(X)
```

```
# Fit the fixed effects model
model = PanelOLS(Y, X, entity_effects=True)
fe_res = model.fit(cov_type='clustered', cluster_entity=True)

# Print the summary of the model
print(fe_res.summary)
```

```
#Random Effects
# Independent variables
x = data[['Half_time_home_goals',
'Half_time_away_goals','Half_time_result_encoded', 'Home_shots',
'Away_shots','Home_shots_target', 'Away_shots_target', 'Home_fouls',
'Away_fouls','Home_corners', 'Away_corners', 'Home_yellow_cards',
'Away_yellow_cards', 'Home_red_cards', 'Away_red_cards']]

# Dependent variable
Y = data['Full_time_result_encoded']

# Add a constant term to the independent variables
X = sm.add_constant(X)

# Run the random effects model
model = RandomEffects(Y, X)
re_res = model.fit()

# Print the results
print(re_res.summary)
```

```
# Perform the Hausman test
comparison = compare({'Fixed Effects': fe_res, 'Random Effects': re_res})

# Print the results
print(comparison)

# Extracting the parameters and covariance matrices
b_fixed = fe_res.params
b_random = re_res.params

cov_fixed = fe_res.cov
cov_random = re_res.cov
```

```
# Calculate the Hausman test statistic
b_diff = b_fixed - b_random
cov_diff = cov_fixed - cov_random
hausman_stat = np.dot(np.dot(b_diff.T, np.linalg.inv(cov_diff)), b_diff)

# Degrees of freedom
df = b_fixed.shape[0]

# Calculate the p-value
p_value = 1 - chi2.cdf(hausman_stat, df)

print(f"Chi-square statistic: {hausman_stat}")
print(f"P-value: {p_value}")
```

Models Results for Panel OLS and Random Effects

Table 6 - Panel OLS and Random Effects models results for dep. variable Log League Position

Independent Variable	Parameter	P-value	Parameter	P-value
Number Players in Squad	-0.0081	0.1329	-0.0081	0.1329
Average Age (Years)	-0.0595	0.0054	-0.0595	0.0054
Percentage of Foreigners in Squad (%)	0.0006	0.6745	0.0006	0.6745
Total Market Value (M €)	-0.0011	0.0016	-0.0011	0.0016
Payroll (M €)	-3.53E-07	0.6628	-3.53E-07	0.6628
Net Transfers (M €)	0.0007	0.0786	0.0007	0.0786
Season Attendance (%)	0.0010	0.4473	0.0010	0.4473
Number of Cup Games	-0.0089	0.3998	-0.0089	0.3998
Number of European Games	0.0005	0.9586	0.0005	0.9586
Log Point Percentage (%)	-1.7457	0.0000	-1.7457	8.88E-16

Table 7 - Panel OLS and Random Effects models results for dep. variable Log League Position

Independent Variable	Parameter	P-value	Parameter	P-value
Number Players in Squad	1.31E-17	0.0420	-0.0045	0.0604
Average Age (Years)	0.0000	1.0000	-0.0130	0.2336
Percentage of Foreigners in Squad (%)	1.25E-18	0.3616	-0.0003	0.7105
Total Market Value (M €)	-9.26E-19	0.0498	-0.0001	0.1153
Payroll (M €)	1.21E-21	0.1826	-1.90E-07	0.4649
Net Transfers (M €)	2.89E-19	0.4587	0.0003	0.0156
Season Attendance (%)	-2.55E-20	0.9805	0.0010	0.1256
Number of Cup Games	-6.07E-18	0.6208	0.0011	0.7879
Number of European Games	-1.17E-17	0.0892	0.0022	0.6345
Log League Position	1.0000	0.0000	-0.3871	4.44E-16

Table 8 - Panel OLS and Random Effects models results for dep. variable Full_time_result_encoded

Independent Variable	Parameter	P-value	Parameter	P-value
Half_time_home_goals	0.0630	0.1456	0.0588	0.1713
Half_time_away_goals	-0.0169	0.7203	-0.0158	0.7366
Half_time_result_encoded	0.4547	0.0000	0.4553	1.95E-14
Home_shots	-0.0004	0.9403	-0.0010	0.8579
Away_shots	0.0053	0.4024	0.0044	0.4850
Home_shots_target	0.0872	0.0000	0.0880	1.11E-15
Away_shots_target	-0.1215	0.0000	-0.1170	0.0000
Home_fouls	0.0050	0.3195	0.0065	0.1935
Away_foults	-0.0042	0.3744	-0.0045	0.3382
Home_corners	-0.0224	0.0043	-0.0227	0.0039
Away_corners	0.0074	0.3878	0.0085	0.3213
Home_yellow_cards	-0.0003	0.9810	-0.0014	0.9192
Away_yellow_cards	-0.0082	0.5229	-0.0067	0.5979
Home_red_cards	-0.1777	0.0008	-0.1726	0.0011
Away_red_cards	0.0463	0.3727	0.0565	0.2733

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa