Does the Gravity Model of Trade apply to the European Football Transfer Market?

Ryan Gamble
Professional Economist BSc and Apprenticeship Level 6
School of Economics
University of Kent, July 2023

Abstract

The football transfer market has become a multi-billion pound industry with the price of players increasing by 9% annually over the last decade. This is attributed to numerous factors, including significant increases in commercial sponsorship and broadcast royalties. Richer clubs are seeing their wealth increase with greater audiences a reward for team performance and individual player star quality. With the richest clubs becoming wealthier causing increased competition in the transfer market for the stars who bring success and kudos. For the player, a transfer affects individual utility through 'wage' and location and future performance prospects. In this paper, the Gravity Model of Trade, implemented using a Fixed effects Poisson pseudo-maximum likelihood (PML) estimator, applied to the football transfer market. The impact of distance and team rivalry is modelled on both the transfer fee and volume, as alternative measures of trade. Annual wage, individual club size and combined club size have been used test robustness. The coefficient signs are consistent across all four models which supports their robustness, with distance and rivalry reducing transfers and club size increasing transfers. In the models with each individual club size, the distance coefficients are -0.341 and -0.633, presenting a 1% increase in distance reduces transfer fees by €34.1k and volume by 0.00633. The estimates for club size with transfer fee as the dependent variable are 0.398 and 0.482, and with the import and export club unable to be differentiated due to data quality issues, these results are interpreted as that when holding all else constant (including the other club's size) a 1% increase in club size increases transfer fees spent by €39.8k and €48.2k. In the transfer volume model these estimates are 0.079 and 0.141, presenting a 1% increase in club size increases the volume of transfers by 0.00079 and 0.00141. Therefore, the analysis supports the hypothesis that distance and club size have the same sign of effect on football transfers as international trade. The Gravity Model consequently can be used in football market analysis, however caveats must be considered including that only Europe's 'top 5' leagues have been analysed.

Acknowledgements

Firstly, I'd like to thank my dissertation adviser Lorraine, whose extensive knowledge on the Gravity Model and advice has been invaluable during this process, always being accommodating and providing extremely useful responses to my queries. Also, my family have been very supportive during my studies, giving me guidance that I feel very lucky to have had all the way from Primary School to Undergraduate studies, and which I wouldn't have succeeded without. Lastly, I'd like to thank my employer, the Ministry of Justice, for continually providing support and flexibility to accommodate my studies and allowing me to achieve the professional and educational development that I sought when joining this apprenticeship.

Introduction

A football transfer is a transaction where a football player moves from one club to another. If the player is under contract with a club, then a fee is paid to this club as compensation, known as the 'transfer fee'. Transfers for players with a contract take place during what are known as 'transfer windows', which are periods of time where transfers are permitted to take place. Carmichael and Thomas (1993), renowned researchers of the football transfer market, argue that the primary purpose of the transfer market is twofold: to facilitate the exchange of players with the aim of improving team performances, and to facilitate the movement of players in their search for better opportunities, earnings or job satisfaction. The transfer fees paid to clubs consider a variety of factors, from the perceived ability of a player and their contract length to a player's age and their on-field position.

The transfer market has been breaking records year on year, with football's governing body FIFA announcing a record $1.57bn was spent in the January transfer window alone (FIFA, 2023). The big clubs are continuing to get richer, with the Premier League (the richest league in the world) signing a record breaking three-year £10.4bn television deal with broadcasters (BBC, 2017). It's important to note that there are well-known inefficiencies in the transfer market, with Kuper and Szymanski's book 'Soccernomics' (2014c) highlighting many, including that stars from global tournaments will be overvalued. Benfica, for example, signed Enzo Fernandez in July 2022 for €44.25m, and after impressing at the FIFA World Cup he was sold just six months later for €121m (Transfermarkt. 2023).

To compete on-field it's important for teams to acquire players that improve their performance. With big clubs continuing to become richer, smaller teams are often unable to compete financially in the transfer market, and therefore it's important for the teams to identify market inefficiencies so that they can maximise their return on investment and compete in areas of the market away from their richer competitors. Analysing different factors that impact transfer fees and volumes may be one way to achieve this, such as by analysing the impact of club size and distance.

Literature Review

Theory

The Gravity Model of Trade dates back to 1962, where Jan Tinbergen applied Isaac Netwon's Theory of Gravity to trade. Newton stated that gravity is directly proportional to the mass of an object and inversely proportional to the square distance between two objects. The Gravity Model of Trade, therefore, is whereby the trade flow between two countries is expected to be proportional to the countries' size and inversely proportional to the distance between them. In even simpler terms, "gravity says that we expect larger country pairs to trade more, but we expect countries that are further apart to trade less" (Shepherd, 2019). With the impact of country size having a proportional impact on trade, trading with an economy that is ten times greater than another country should result in the trade flow being 10 times greater than with the other country (ignoring the effect of distance, ceterius paribus). However, distance's impact on trade is non-linear, such as in Newton's theory where the squared distance is used, and consequently trade decreases at an accelerating rate as distance between trade partners

increases. Therefore, distance can offset the impact of trading with larger economies (Estrin, Cote and Shapiro, 2018).

At the point of Tinbergen's introduction of the model in 1962, the model had great empirical success, with Anderson (2011) stating that "gravity has long been one of the most successful empirical models in economics". However, Tinbergen failed to provide theoretical reasoning behind the model's success. It wasn't until Anderson's contribution in 1979 where the theoretical foundations of the model were extended, with further contributions notably being added by Bergstrand (1985, 1989, 1990), Deardorff (1998) and Anderson and Wincoop (2003), along with others (Kabir, Salim and Al-Mawali, 2017). Many theories have been provided on the cause of bilateral trade falling with an increase in distance, with Silva and Tenreyro (2006) stating "endowment and technological differences, increasing returns to scale, and Armington demands, all predict a gravity relationship for trade flows". Transport costs are also a possible cause of the distance effect, however this effect has persisted over time despite such costs reducing, including sea freight costs being 78% lower in 2005 than in 1930 (Our World In Data, 2015), and consequently research continues into the cause of distance reducing bilateral trade. Leamer (2017) remarks that the distance effect on international commerce is "possibly the only important finding that has fully withstood the scrutiny of time and the onslaught of economic technique", and therefore despite some of the theoretical foundation of the Gravity Model being debated, the empirical success of the model remains, resulting in the model still being widely used by researchers.

Distance is still expected to have a negative effect on football transfers, despite the causes differing from on international trade. Kuper and Szymanski (2014a) highlight the difficulties faced by players when moving location, including loneliness and the struggles the player and their family face adapting to the weather and food. Such difficulties reduce the chances of a player succeeding when moving to a club in an unfamiliar location. As a result, football clubs historically have prioritised signing players from the same league, therefore being closer in distance, to improve the likelihood of the player improving their performance. Relocation consultants that help a player to adjust to a new club have, however, become increasingly common in football, such as Manchester City having their own 'player care department'.

Findings

The application of the Gravity Model has become widely popular, with Baier and Bergstrand (2014) stating it has "dominated the international trade literature as the main econometric approach". The Gravity Model is said to provide a "standard starting point for much empirical work in international trade, and for that reason is of particular interest to applied policy researchers" (Shepherd, 2019), which is used as a tool to analyse the impacts of a large variety of factors on trade by accounting for the impact of distance and economy size.

Anderson and Wincoop (2003), for example, used the Gravity Model to control for distance and economy size to isolate the impact of borders on trade between the U.S and Canada. In this analysis, they found that "borders reduce trade between the United States and Canada. . . by 44%. Trade among ROW countries is reduced by 29%". They explain that the difference in border impact on trade can be attributed to the theory that border effects are greater for bigger economies.

Frankel, Stein and Wei (1995), in their analysis of bilateral trade patterns throughout the world, found that all four standard Gravity Model variables were highly significant, supporting the

Gravity Model's empirical success. The 1990 coefficient on the log of distance was about -0.6 when including a dummy variable for countries that are adjacent, meaning that when distance increases between non-adjacent countries by 1% then trade falls by 0.6%. However, this research failed to examine the presence of heteroskedasticity and therefore these estimates may be bias (Kabir, Salim and Al-Mawali, 2017). Disdier and Head (2008) also found a negative effect of distance on trade. Following constructing a database of 1467 papers from 103 papers, the mean effect of distance was 0.9, meaning that on average a 1% increase in distance reduced bilateral trade by 0.9%. They conclude that the persistence of the distance effect, which "holds up in a very wide range of samples and methodologies", provides challenge to those who think that technologies change has caused distance to have less effect on the world economy.

Frankel, Stein and Wei (1995) also supported the Gravity Model by estimating a positive relationship between country size and trade. Gross National Product (GNP) was used as the measure of economy size, rather than the commonly used 'GDP', which includes the value of all goods and services owned by a country's citizens rather than the value of all goods and services produced. The estimated coefficient on the product of per capita GNPs varied from 0.26-0.4 from 1965 to 1980, which supports the Gravity Model prediction that richer countries do trade more, however the failure to account for heteroskedasticity means these estimates are likely bias.

Other factors are also found to have an impact on bilateral trade, such as Havrylyshyn and Pritchett (1994) stating that language has an impact on the Gravity Model with Portuguese, Spanish and English found to have a significant impact on model estimates. A country's own regulations are also shown to have an impact by Shepherd, Doytchinova and Kravchenko (2019), estimating that a one point increase in ETCR score (a measure of regulation in energy, transport and communications), meaning a more restrictive regulatory environment, was associated with a 36% or 37% decrease in trade. Therefore, variables other than country size and distance are also important in Gravity Model analyses.

Methodology

Following there being a 17-year gap between Tinbergen's release of the initial intuitive Gravity Model and Anderson's first notable extension of the model's theoretical foundation, there has been great debate regarding the correct methodology to use. The Gravity Model, in its most basic form, can be written as:

$$logExports = c + \beta\ logGDP + \beta\ logGDP + \beta\ logDistance + e\ ij\ 1\ i\ 2\ j\ 3\ ij\ ij$$

Whereby each country's GDP is included as a variable alongside the distance between them.

This basic model faces some theoretical problems, including that the model doesn't account for changes in trade creation/diversion. If trade costs change, such as countries entering a preferential trade agreement, then the lowered trade costs on one trade route impacts the relative trade costs on other routes, which is unaccounted for in this model (Shepherd, 2019). This is supported by Anderson and Wincoop (2003), who notably stated the model isn't correctly specified as it doesn't account for multilateral resistance (factors that reduce trade between two countries, such as tariffs).

Anderson and Wincoop adjusted the Gravity Model formula, which "has become one of the standard formulations used in applied work" (Shepherd, 2019), finding that such omitted

variable bias can have a substantial impact on Gravity Model analysis. Highlighting that the omitted variable bias in McCallum's analysis (1995) caused McCallum to incorrectly conclude that the US-Canada border caused trade between provinces to increase by 2,200%, rather than by a factor of 6, presents the importance of multilateral resistance variables.

Fixed Effects Models consequently largely dominates Gravity Model literature (Shepherd, 2019). In a Fixed Effects Model, intercepts vary across cross sectional units, but each individual cross-sectional unit is time invariant. Adding dummy variables for each trading country accounts for all sources of unobserved heterogeneity that are constant for a country across other trading partners, and therefore tackles the omitted variable bias discussed by Anderson and Wincoop. However, the use of a Fixed Effects Model does introduce a major restriction to the model, whereby to not break the OLS assumption that 'none of the explanatory variables is a linear combination of other explanatory variables', variables that vary only in the same dimension as the fixed effect cannot be included (Shepherd, 2019). This includes the variable of distance, as this variable remains the same for each trade partnership thus causing perfect collinearity.

The use of a Random Effects Model is another method, which assumes cross section intercepts are random instead of being fixed and therefore the model can include variables that change through time unlike in FEM. However, there has been wide criticism of the use of REM in Gravity Models, with Mundlak (1978) arguing the model overlooks the possible correlation between individual effects and the regressors, while Wooldridge (2009) states that if there's any correlation between the error term and some regressors then the estimates aren't valid. Furthermore, Shepherd (2019) states that REM requires multilateral resistance variables to be normally distributed, despite there being no theoretical support for this. This leads Shepherd to conclude that "all gravity model research should now include appropriate dimensions of fixed effects".

There's a long history of using log-linear OLS models in Gravity Model literature, but problems with values of zero and heteroskedasticity have turned researchers to using different estimation methods. In Newton's research of gravity, gravity can be small, but never zero. This isn't the case for trade data, where countries can have a trade value of zero between them. Due to the logarithm of zero being undefined, log-linear OLS models automatically remove the values of zero, causing concern regarding sample selection bias. When Westerlund and Wilhelmson (2011) examined the effects of zero trade on the estimation of the gravity model using both simulated and observed panel data, they demonstrated that log- linear estimation can lead to highly misleading results due to the zero observations skewing the data. In addition, Silva and Tenreyro (2006) found log-linear OLS methods to be "clearly inadequate because, despite its low dispersion, it is often badly bias". With trade values of zero being common in bilateral trade data, shown by Helpman (2008) reporting that around half of the bilateral trade matrix is filled with zeroes, using log-linear OLS estimation is not recommended in recent literature to avoid using skewed data.

In addition, heteroskedasticity is commonly found in Gravity Model data due to the large amounts of cross-sectional data being spread over time. The problem has persisted in a lot of literature, including one of the most cited recent pieces by Frankel (1995) which fails to examine the presence of heteroskedasticity. According to Silva and Tenreyro (2006), under heteroskedasticity the estimated parameters of log-linearised models by OLS lead to biased estimates, going on to find "overwhelming evidence that the error terms in the usual log linear specification of the gravity equation are heteroskedastic". Consequently, both Silva and

Tenreyro (2006) and Westerlund and Wilhelmson (2009) propose not using a log-linear OLS model in Gravity Model analysis.

They instead propose the use of the Poisson pseudo-maximum likelihood estimator (PML), which can naturally include observations of zero value, unlike the OLS log-linear specification, and therefore reduces the risk of sampling bias, while it also better deals with heteroskedasticity. Further desirable properties include that the estimator is consistent in the presence of fixed effects, which as previously stated, is important to include to account for multilateral resistance. Therefore, the use of the Poisson PML estimator takes care of the issues regarding heteroskedasticity and zero trade values, and importantly allows the use of fixed effects.

The PML estimator, under weak assumptions such as that the correct explanatory variables are included in the model, provides consistent estimates to the original OLS non-linear model (Shepherd, 2019). The estimator gives the same weight to all observations, assuming that all observations provide the same information on the parameters. Silva and Tenreyro (2006) used Monte Carlo simulations, a method that estimates the possible outcomes of uncertain events, to compare the performance of the PML estimator and the OLS log-linear specification. The results for the PML estimator were "very encouraging", with the "Poisson- based PML estimator [being] relatively robust to this form of measurement error of the dependent variable", and strikingly that "it is clear that apart from the Poisson PML method, all estimators will be very mis-leading". In addition, Shepherd (2019) highlights the estimator's desirable properties that has led the estimator to become a "workhouse estimator for gravity".

Hypothesis

Size and distance are hypothesised to have the same effect on the football transfer market as on international trade:

i. H0: Club size has no effect on transfers.

   H1: Greater club size will increase transfers.

ii. H0 Distance between clubs has no effect on transfers.

   H1: Greater distance will decrease transfers.

Increased club size is expected to increase transfers for two main reasons: firstly, that bigger clubs are expected to have more money and so can spend more on transfers, and that bigger clubs will transfer more between them as they have a similar standard of players in their squads.

With players facing more challenges when moving to a club in an unfamiliar location, such as adjusting to a new playing style and their family adapting to a new location, purchases are often considered to be higher risk when the player's club is further in distance. Clubs are known to therefore prioritise signing players from clubs within the same league, resulting in distance being hypothesised to have a negative effect on football transfers, similarly to international trade.

Rivalries between clubs can be both historical and geographical and reduce the likelihood of transfers between clubs due to the high probability of backlash from fans towards the clubs and player. Clubs with rivalry between them are consequently expected to have less transfer activity between them.

Data

Sources

Transfer Data:

The transfer data, which will be used as the dependent variable, is sourced from Transfermarkt, a website renowned in the football industry for its player and club data. The database isn't available to be downloaded, and therefore web scraping (the use of tools to extract information from a website) has been required to obtain the transfer data. An online user has carried out this web scraping and made the data freely available online in accordance to Transfermarkt's terms of use.

The data includes every male football player's club movement in the top nine European leagues from 1992 to 2022, containing the following variables:

- Club names
- Player name
- Age
- Position
- Fee (€m)
- Transfer Window
- League name
- Year

Club Size (Annual Wage):

For football clubs there isn't a widely agreed measure for club size, unlike in economics where GDP is used for economy size. Club size can be considered from a financial aspect, such as using annual revenue, or on-pitch performance by league position. Some attempts to value clubs have been made, however these differ between sources, for example in 2021 KPMG estimated Liverpool F.C. to be valued at £2.07billion, whereas Forbes estimated a value of £2.9bn, a significant difference (Slater, 2022). Annual revenue is arguably the most similar measure to GDP; however, this data isn't available for all clubs included in the analysis. In addition, annual revenue may not be considered the best measure of club size as revenue won't necessarily be invested back into the club. Annual wage has instead been used as the measure of club size. Higher wages are expected to be paid by richer clubs, and thus the measure considers the financial aspect of club size. The measure also accounts for on- pitch performance with wages found to have a significant impact on performance. In Kuper's and Szymanski's book 'Soccernomics' (2014b), they found for English Premier League clubs between 2003-2012 that "the size of their wage bill explained a massive 92% of variation in their league positions", attributing wage's impact on performance to the ability to attract better performing players.

The annual wage data is sourced from Capology, a website that provides the wages of football players. These figures are either verified by network insiders, or where the values are unknown are calculated using an algorithm. The data set includes the weekly and annual wages of each team in Europe's top 5 leagues (English, Spanish, German, Italian and French) from 2013-2021.

Stadium Address Data:

The addresses of each club's stadium have needed to be web scraped, however due to time constraints the open AI 'ChatGPT' has been used. The list of clubs that the addresses needed to be produced for was provided to ChatGPT, and spot checks have been carried out on the addresses, ensuring that the information produced is correct and that there are no missing observations. These stadium addresses are used in calculating the distance between clubs, which is discussed in the 'Data Manipulation' section.

Rivalry Data:

Due to there being no available data on football rivalries, a data set has been created specifically for this analysis. Extensive online research has been carried out to produce a binary variable, with a value of '1' being used for teams with a rivalry.

Data Manipulation

Extensive data manipulation has been required to allow the calculation of the required variables and merging of the different data sets. This manipulation was carried out in the coding language 'R', largely using the data manipulation package of 'dplyr'.

Changing the club names in each data set was required so that they are consistent across data sets to allow merging. Fuzzy matching was attempted, which is the use of technology to match similar but not identical character strings, but this wasn't possible due to some clubs' nicknames being used which are too dissimilar to their full name. The club names were therefore manually changed.

Transfer Data:

Due to the analysis focusing on transfers between clubs, other types of player movements have been filtered out of the data, including free transfers, career breaks and players retiring, while players ending their loan and returning to their home clubs are also removed as this isn't a transfer.

The data includes the youth teams of each club, which in this analysis shouldn't be differentiated from the adult team as the club itself is the same. Youth team names were consequently changed to match the club's name, and the movement of players from the youth to the adult's team were filtered out as this is a promotion rather than transfer.

Only five leagues out of the nine available have been used due to data availability, with the annual wage data only being available for, what have been consistently ranked by UEFA to be, the top five leagues in Europe:

- English Premier League

- Spanish La Liga
- German Bundesliga
- Italian Serie A
- French Ligue 1

To create a 'trade matrix' that is used in Gravity Model analysis, where a club is paired with each club that is available to trade with, transfers with clubs outside of these leagues have been removed. A function has then been created that produces a table with all possible club combinations in each season. Not only is this practical, but it prevents the skewing of data from including clubs that are relegated and are likely to sell more players to overcome the financial difficulties of relegation (receiving less revenue in lower leagues). The data has also been filtered to only include transfers from the 2013/14 to the 2021/22 season, as wage data is only available for this time period.

Transfers can be measured in two ways: transfer fee and the volume of transfers between two clubs. With the data set including each player transfer as an observation, the transfer fee and volume between each club is required to be calculated. The transfer fee of all transfers isn't known, and therefore the transfer fee calculations include less observations than the transfer volume calculations, meaning analysis of both measures should be carried out. Doing so will also check the robustness of the model and theory, allowing a comparison of the results following using different data.

Data quality issues arose with some transfers being recorded for one club but not the other club involved. Therefore, when calculating the transfer fee and volume, one club would have inaccurate data. Data manipulation has ensured each club has the correct transfer records, however the manipulation used has caused information on who the import and export club are to be removed.

Following the data manipulation, the summary statistics are as provided below:

Transfer Data

	Fee (€m)	Volume
Observations	41906	41906
Mean	0.597	0.099
Median	0	0
Maximum	222	9
Minimum	0	0
Skewness	16.080	5.494

Rounded to 3d.p.

In Gravity Model analysis of trade there are few countries with zero trade between them, whereas in football transfers it's very common for teams to not transfer between them because of their small squad sizes. This is the cause of the medians for both transfer fee and volume being 0, and both exhibiting a great positive skewness (of 16.080 for fee and 5.494 for volume). The positive skewness supports the use of a Poisson distributed model, with Poisson distributions always skewing to the right due to probability not being able to be less than zero.

Wage Data:

The numerical wage data has been cleaned. The wage data column included the wage value in both Euros and Pound Sterling, and code has been used to separate these values, keeping only the Euros value and converting to be in 'millions' to match the transfer data.

Annual Wage (€m)	
Observations	874
Mean	51.531
Median	31.789
Maximum	432.700
Minimum	4.590
Skewness	2.562

Rounded to 3d.p.

Distance:

Following obtaining each club's stadium addresses, the distance between them has been calculated. Using Google Maps' Geocoding API and the 'ggmap' package in 'R', the longitude and latitude co-ordinates of each stadium has been produced. The distance between the co-ordinates of each stadium is calculated using the 'geosphere' package, once this data is merged with the transfer and wage data.

Distance (km)	
Mean	934.411
Median	901.149
Maximum	3601.587
Minimum	0.000
Skewness	0.684

Rounded to 3d.p.

Merging:

Merging of the data sets is performed in a particular order, starting with the transfer data containing each possible transfer combination and the fee/volume between the clubs, before then merging in the wage, rivalry and co-ordinates data. Doing so in this order means that the wage, rivalry and co-ordinates data aren't required to contain all the possible club combinations, saving on calculation time. The distances between each stadium are then calculated as previously discussed. The annual wage data didn't contain observations for 'CD Leganes' and 'Hannover 96', and therefore they've been excluded from the analysis.

In Gravity Model analysis of trade, the countries included each year in the analysis are the same. In this analysis, however, due to only the 'top 5' leagues being included and teams able to be relegated from these leagues, the team combinations vary each year.

Methodology

Following the evidence and recommendations by Silva and Tenreyro (2006), supported by Westerlund and Wilhelmson (2009), a semi-log Poisson-Maximum Likelihood Estimator (PML) has been used. The PML estimator has been used to evaluate the effect and significance of distance and club size on football transfers to test the hypothesis that these effects are as expected by the Gravity Model of Trade.

Fixed Effects have been deployed, with dummy variables being added for when each club is included in a transfer, to account for unobserved heterogeneity and allowing the effect of distance to be evaluated.

As transfers can be measured by both fee and volume, two different models have been run with each variable being used as the dependent variable. The binary *RIVALRY* variable has been included to account for geographical and historical rivalries between clubs. Using the coding language 'R', the model equations are:

$$\text{Transfer Fee}_{it} = \alpha + \beta_1 \ln(Distance)_{it} + \beta_2 \ln(Club\ Size\ 1)_{it} + \beta_3 \ln(Club\ Size\ 2)_{it} + \beta_4 Rivalry_{it} + \Sigma D_{it}$$

$$\text{Transfer Volume}_{it} = \alpha + \beta_1 \ln(Distance)_{it} + \beta_2 \ln(Club\ Size\ 1)_{it} + \beta_3 \ln(Club\ Size\ 2)_{it} + \beta_4 Rivalry_{it} + \Sigma D_{it}$$

Here 'D' represents the use of binary dummy variables for each club, with a value of '1' when the club is involved in the transfer.

The combined size of the two trading clubs may have a different effect on transfers than each individual club size, and therefore models including an explanatory variable of the club sizes multiplied together have also been run, implying an exponential relationship:

$$\text{Transfer Fee}_{it} = \alpha + \beta_1 \ln(Distance)_{it} + \beta_2 \ln(Total\ Club\ Size)_{it} + \beta_4 Rivalry_{it} + \Sigma D_{it}$$

$$\text{Transfer Volume}_{it} = \alpha + \beta_1 \ln(Distance)_{it} + \beta_2 \ln(Total\ Club\ Size)_{it} + \beta_4 Rivalry_{it} + \Sigma D_{it}$$

Although the PML has a natural method of dealing with zero trade values, the use of a semi-log model causes implications with the *DISTANCE* variable. In analyses of international trade, distance is never zero as countries cannot be located in the same place. However, some football clubs share a stadium, meaning their location is the same and distance is zero. As the logarithm of zero is undefined, the distance between clubs sharing a stadium has been set as a small figure of 0.0000001km to allow these observations to be included in the analysis.

Data & Methodology Limitations

Both the transfer and wage data face data quality issues. Not all transfer fees are made publicly available, and therefore transfer fees are likely to differ between sources. This is a common problem in football transfer analysis, with Sky Sports stating that out of the approximate 700 transfers in the 2012/13 season in England that less than 9% of the fees were known (Ruijg and

van Ophem, 2015), and so the fees reported in the data set are unlikely to be completely accurate. Only the USA's 'Major League Soccer' publishes official wage data, and so the European leagues' wage figures used in this analysis are unofficial and thus can also vary between sources (Lynch, 2022). Even when the salaries are verified by network insiders, these values cannot be guaranteed to be completely accurate.

Sample selection bias is also an issue in this data. In Gravity Model analyses of trade, a set number of countries are included, such as Silva and Tenreyro (2006) including 136 countries, and data for these countries are commonly available. However, in football there are leagues in each country with multiple divisions, making it near impossible to have data available for each club. To create a 'trade matrix' as previously discussed, a decision must be made on what set teams will be included, which means excluding many clubs in the world. The wage data only includes the 'top 5' European leagues, resulting in only transfers between these leagues being included and consequently transfers with clubs from other leagues are excluded, which could have a significant impact on results.

Sample selection bias is also apparent in the club size variables. As explained previously, data manipulation has been carried out to address data quality issues which has caused the loss of information regarding who the importing and exporting clubs are. This process has required ordering the clubs in alphabetical order between the club size variables, with the first club size variable including teams that precede the other club alphabetically. Not only does this cause sample selection bias regarding what club is included in each variable, it also limits the interpretation of the model, with no interpretation being possible into the different impacts on importing and exporting clubs.

In Gravity Model analysis, each transaction (in this case, a player transfer) isn't included as an observation but rather included in the calculation of trade value between two countries (clubs). This results in individual player characteristics being unable to be included in the analysis, such as player age, nationality and minutes played which are likely to impact transfer fees. Consequently, the model may suffer from omitted variable bias.

Results

As discussed in the Methodology section, models with both transfer fee and volume as the dependent variable have been used, as well as models using a total of the club size and individual club size. This will help check the robustness of the model and theory, using different models and data to test the hypothesis.

In all these models, tests for superfluous variables have been carried out, identifying if variables are deemed to have no explanatory variable and therefore don't statistically improve the fit of the model. Variables can be superfluous for different reasons, including that it's unable to explain the variance of the dependent variable or that it's highly correlated to other variables. Including such variables can cause coefficient estimates to become less precise. A Wald Test is carried out to test if the explanatory variables are different from zero, with a null hypothesis stating that some of the variables are all equal to zero. If this null is unable to be rejected, then a variable can be removed from the model because it doesn't statistically improve the fit of the model, thus improving the precision of the model.

Tests for autocorrelation have also been used, whereby there's a correlation between past and present residuals of the same variable. A relationship between such values can cause the residuals to not have minimum variance, resulting in estimates being bias and therefore incorrect inferences being made. The Durbin-Watson Test tests for First Order Autocorrelation, which is correlation between successive residuals. If such autocorrelation is found then Robust Standard Errors can be used, which accounts not just for autocorrelation but also heteroskedasticity in the calculation of standard errors.

Transfer Fee Models

Econometric Tests

The Durbin-Watson Test for autocorrelation has been carried out for each of the models using individual club size and total club size. The tests on each of the models produced positive test statistics, implying a positive autocorrelation where the residuals are positively correlated with past values. The P-values are both below 0.05, meaning at the 5% significance level the null hypothesis can be rejected. With the null hypothesis stating that there's no autocorrelation, positive autocorrelation is found in both models. Consequently, Robust Standard Errors have been used in these models so that autocorrelation is accounted for in the calculation of the standard errors.

The Wald Test for superfluous variables has been carried out on both the individual variables and also all the variables together to test individual and joint significance. All the tests produced a P-value lower than 0.05, again allowing the null to be rejected at the 5% significance level, thus showing that all variables in both models improve the statistical fit of the model. Therefore, all variables have been kept in the model.

The Wald Test was also carried out on the dummy variables used to account for unobserved heterogeneity, which produced a very low P-value for each model. As the dummy variables are shown to improve the statistical fit of the model, unobserved heterogeneity does appear to be present and consequently the variables have been kept in the model.

Model Estimates: Individual Club Sizes

Fee: Model 1

	Coefficient	Z-value	P-value	Signficance
ln(Distance)	-0.341	-7.482	7.34E-14	***
ln(Club Size 1)	0.398	6.009	1.86E-09	***
ln(Club Size 2)	0.482	6.343	2.25E-10	***
Rivalry	-0.626	-2.013	0.044	*

*Signif. Codes: 0'***', 0.001'**', 0.01'*', 0.05'.'*

Both *DISTANCE* and each of the *CLUB SIZE* variables are found to be highly significant by displaying low P-values, however the *RIVALRY* variable is only significant at the 10% significance level. The *DISTANCE* coefficient estimate shows a negative relationship, where clubs further apart have a lower value of transfers between them. With an estimate of -0.341, for every 1% increase in distance the value of transfers between clubs decreases by €0.00341m (€34.1k).

Contrastingly, each *CLUB SIZE* coefficient displays a positive relationship. As explained in the 'data & methodology limitations' section, the import and export club is unable to be differentiated due to data quality issues, and thus we cannot interpret the results in terms of imports and exports. Therefore, with the estimates of 0.398 and 0.482, the estimates are interpreted as that holding all else constant, including the other club's size, a 1% increase in the annual wage bill will increase the transfer value between the clubs by €0.00398m and €0.00482m (€39.8k and €48.2).

The *RIVALRY* variable does display a negative relationship, implying clubs that are rivals will have a lower value of transfer activity between them, however due to the lack of significance it's uncertain whether the parameter is statistically different from 0.

Model Estimates: Total Club Size (Club Size 1 X Club Size 2)

Fee: Model 2

	Coefficient	Z-value	P-value	Signficance
ln(Distance)	-0.339	-7.422	1.15E-13	***
ln(Total Club Size)	0.442	6.86	6.89E-12	***
Rivalry	-0.608	-1.958	0.051	

*Signif. Codes: 0'***', 0.001'**', 0.01'*', 0.05'.'*

The model using *TOTAL CLUB SIZE* rather than each individual club size provides similar results, also providing highly significant results for the *DISTANCE* and *TOTAL CLUB SIZE* variable and not for the *RIVALRY* variable, consequently supporting the robustness of this model. *DISTANCE* again has a negative relationship and has a similar magnitude, being - 0.339 compared to the previous model's value of -0.341.

TOTAL CLUB SIZE is also of a similar magnitude to the previous model's club size estimates, with the total wage bill (each annual wage bill multiplied together) being 0.442 and thus implying a 1% increase in the total wage bill will increase the value of transfers by €0.00442m (€44.2k).

RIVALRY again is shown to have a negative impact on transfer activity with a value of - 0.608, but the estimate also lacks significance.

Interpretation: Transfer Fee Models

The Gravity Model of Trade predicts an increase in distance will decrease trade, and an increase in size will increase trade. Using transfer fees as the measure of transfer activity therefore produces results that are supported by the Gravity Model. The sign of the coefficients match what is found in the literature, such as Frankel, Stein and Wei (1995) also providing a negative coefficient for distance (-0.6). Consequently, these results support the hypothesis that size and distance have the same impact on the football transfer market as they do on international trade. *RIVALRY* estimates lack significance and so fail to support the hypothesis that rivalry reduces transfer activity.

Transfer Volume Models

Econometric Tests

Using the transfer volume between clubs as the dependent variable, the Durbin-Watson Test for autocorrelation again produced positive test results implying positive autocorrelation, and with P-values below 0.05 the null is rejected and thus autocorrelation is present in both models. Robust Standard Errors are therefore used in the transfer volume models also.

The Wald Test for superfluous variables is run again for all variables together and each variable individually. In both models the variables are found to jointly improve the statistical fit of the model, with the Wald Test of all the variables producing P-values of 0.0. In both models the *DISTANCE* variable fails to reject the null hypothesis that the variable is different from zero, with a P-values of 0.4 and 0.41 being produced. This could be due to rivalries often being caused by being close in distance, and therefore the *RIVALRY* variable could be accounting for some of the variation that the *DISTANCE* variable would otherwise account for. This result would imply that the *DISTANCE* variable is irrelevant and should be removed. However, distance is an important variable in Gravity Model analysis (as presented in the Literature Review), and therefore the *DISTANCE* variable has been kept in both models. The Wald Test is again carried out on the dummy variables of each model, which rejects the null hypothesis, showing that the variables improve the fit of the model and thus are kept in the models.

Model Estimates: Individual Club Sizes

Volume: Model 1

	Coefficient	Z-value	P-value	Signficance
ln(Distance)	-0.633	-30.006	<2.2E-16	***
ln(Club Size 1)	0.079	2.099	0.036	*
ln(Club Size 2)	0.141	3.935	8.31E-05	***
Rivalry	-1.907	-10.046	<2.2E-16	***

*Signif. Codes: 0'***', 0.001'**', 0.01'*', 0.05'.'*

DISTANCE, *RIVALRY* and the second *CLUB SIZE* variable are highly significant from 0, however the first *CLUB SIZE* variable is less significant with a P-value of 0.036, reducing confidence that the estimate is statistically different from zero. The *DISTANCE* estimate again produces a negative relationship, with its estimate of -0.633 presenting that an increase in distance by 1% will decrease the volume of transfers between two clubs by 0.00633.

Both *CLUB SIZE* estimates have a positive relationship, with estimates of 0.079 and 0.141 showing a 1% increase in the annual wage bill of a club will decrease transfer volume by 0.00079 and 0.00141. However, with the first *CLUB SIZ*E variable being less significant, it's not certain that this value is statistically different from zero.

RIVALRY is estimated to have a substantially negative impact on transfer volume, with its magnitude being greater than for *DISTANCE* and *CLUB SIZE* (-1.907).

Model Estimates: Total Club Size (Club Size 1 X Club Size 2)

Volume: Model 2

	Coefficient	Z-value	P-value	Signficance
ln(Distance)	-0.632	-29.975	<2.2E-16	***
ln(Total Club Size)	0.110	3.395	0.000	***
Rivalry	-1.899	-10.028	<2.2E-16	***

*Signif. Codes: 0'***', 0.001'**', 0.01'*', 0.05'.'*

All of the variables are found to be highly significant, with low P-values of <2.2e-16 for *DISTANCE* and *RIVALRY* and 0.0001 for *TOTAL CLUB SIZE*. The *DISTANCE* coefficient is -0.632, very similar to the previous model's estimates of -0.633, and thus presents that distance increasing by 1% decreases transfer volume by 0.00632. The *TOTAL CLUB SIZE* variable has a positive coefficient, matching the previous model's coefficient sign like *DISTANCE*, with an estimate of 0.110 showing the total annual wage bill increasing by 1% will increase transfer volume by 0.0011. Rivalry again is shown to have a substantial negative impact on transfer volume with an estimate of -1.899.

Interpretation: Transfer Volume Models

Despite changing the dependent variable from transfer fee to volume, both models still have the same sign of coefficients and similar magnitudes, which supports the robustness of the model. Again, the estimates match what is expected by the Gravity Model, with distance and size causing a decrease and increase in transfers, and therefore the hypothesis is again supported. *RIVALRY*, while supporting the hypothesis that it reduces transfer activity, has a greater impact on transfer volume than transfer fees, with the estimates of -1.907 and -1.899 being greater than -0.626 and -0.608 from the transfer fee models.

Conclusion

The conclusion of this research is that the Gravity Model of Trade is applicable to the European football transfer market. As hypothesised, distance is shown across all models to have a negative relationship with transfers, while club size has a positive relationship. Literature for international trade analysis has argued varied reasons for the persistence of the distance effect, including transport costs and time, however the causes in the football transfer market differ. Clubs are more likely to prioritise signing players from the same league so that the player faces less challenges to adjust, increasing the player's chances of improving the team's performance, thus causing distance to have a negative effect.

Rivalry also has a negative impact on football transfers across the models, however it is only in the models with 'transfer volume' as the dependent variable where the variable is statistically significant, while also having a higher magnitude than in the 'transfer fee' models. This could be attributed to rivalry reducing the likelihood of transfers between two clubs due to fear of fan backlash, but when they do transfer there's little effect on the fee.

Football clubs should therefore consider the impacts of club size, distance and rivalry when deciding what areas of the transfer market to target. With distance having a negative impact on

transfer activity, a club will face less competition from teams close geographically by trading with clubs further away. This lower competition for players could allow a club to pay a player lower wages due to fewer clubs offering the player a contract, saving costs. Consequently, clubs should increase their transfer activity with clubs further in distance. The review of literature highlighted the risk of signing players from further locations due to these players facing greater issues with adjustment, however these can be reduced by increased investment in 'relocation consultants' that help players to adjust to their new club.

There are caveats to be considered with these results. Only transfers between the 'top 5' leagues in Europe have been included, and consequently the effect of distance and size have only been analysed for a select few leagues. Policy recommendations therefore cannot be applied to wider leagues. In addition, the inclusion of only club characteristics omits player characteristics which are likely to have a substantial impact on transfers. These results may as a result suffer from omitted variable bias and lack accuracy.

References

Transfermarkt. (2023) Available at: https://www.transfermarkt.co.uk/enzo-fernandez/transfers/spieler/648195/transfer_id/3943263

Anderson, J.E. (2011) 'The Gravity Model', *Annual Review of Economics,* 3(1), pp. 133-160 Available at: https://doi.org/10.1146/annurev-economics-111809-125114.

Anderson, J.E. and van Wincoop, E. (2003) 'Gravity with Gravitas: A Solution to the Border Puzzle', *The American Economic Review,* 93(1), pp. 170-192 Available at: http://www.jstor.org/stable/3132167.

Baier, S., Bergstrand, J. and Feng, M. (2014) 'Economic integration agreements and the margins of international trade', *Journal of International Economics,* 93(2), pp. 339-350 Available at: https://EconPapers.repec.org/RePEc:eee:inecon:v:93:y:2014:i:2:p:339-350.

BBC (2017) *How does a football transfer work?* Available at: https://www.bbc.com/worklife/article/20170829-how-does-a-football-transfer-work

Disdier, A. and Head, K. (2008) 'The Puzzling Persistence of the Distance Effect on Bilateral Trade', *The review of economics and statistics,* 90(1), pp. 37-48 Available at: http://www.jstor.org/stable/40043123.

Estrin, S., Cote, C. and Shapiro, D. (2018) *Can Brexit defy Gravity?* Available at: https://blogs.lse.ac.uk/management/2018/11/09/can-brexit-defy-gravity-it-is-still-much- cheaper-to-trade-with-neighbouring-countries/

FIFA (2023) *International Transfer Snapshot (January 2023).* Available at: https://www.fifa.com/legal/media-releases/fifa-publishes-international-transfer-snapshot- january-2023-new-all-time-highs

Frankel, J., Stein, E. and Wei, S. (1995) 'Trading blocs and the Americas: The natural, the unnatural, and the super-natural', *Journal of Development Economics,* 47(1), pp. 61-95 Available at: https://EconPapers.repec.org/RePEc:eee:deveco:v:47:y:1995:i:1:p:61-95.

Havrylyshyn, O. and Wissels, R. (1994) 'Reviving Trade Amongst the Newly Independent States', *Economic Policy,* 9(19), pp. 172-190 Available at: https://doi.org/10.2307/1344606.

Helpman, E., Melitz, M. and Rubinstein, Y. (2008) 'Estimating Trade Flows: Trading Partners and Trading Volumes', *The Quarterly Journal of Economics,* 123(2), pp. 441-487 Available at: https://EconPapers.repec.org/RePEc:oup:qjecon:v:123:y:2008:i:2:p:441-487.

Kabir, M., Salim, R. and Al-Mawali, N. (2017) 'The gravity model and trade flows: Recent developments in econometric modeling and empirical evidence', *Economic Analysis and Policy,* 56, pp. 60-71 Available at: https://doi.org/10.1016/j.eap.2017.08.005.

Kuper, S. and Szymanski, S. (2014a) *Soccernomics* HarperSport, pp. 40.

Kuper, S. and Szymanski, S. (2014b) *Soccernomics* HarperSport, pp. 14.

Kuper, S. and Szymanski, S. (2014c) *Soccernomics* HarperSport, pp. 22.

Leamer, E. and Stern, R. (2017) 'Quantitative International Economics', Available at: https://www.taylorfrancis.com/books/mono/10.4324/9781315127897/quantitative- international-economics-edward-leamer-robert-stern.

Lynch, M. (2022) *Capology Wage Data Explainer.* Available at: https://www.sports-

reference.com/blog/2022/08/capology-wage-data-explainer/
Mundlak, Y. (1978) 'On the Pooling of Time Series and Cross Section Data', *Econometrica,* 46(1), pp. 69-85 Available at: https://doi.org/10.2307/1913646.
Our World In Data (2015) *The decline of transport and communication costs relative to 1930.* Available at: https://ourworldindata.org/grapher/real-transport-and-communication-costs
Poli, R., Ravenel, L. and Besson, R. (2023) *Inflation in the football players' transfer market.* Available at: https://www.football-observatory.com/IMG/sites/mr/mr82/en/#:~:text=Generally%20speaking%2C%20all%20other%20things,three%20seasons%20following%20the%20pandemic.
Ruijg, J. and van Ophem, H. (2015) 'Determinants of football transfers', *Applied Economics Letters,* 22(1), pp. 12-19 Available at: https://doi.org/10.1080/13504851.2014.892192.
Santos Silva, J. and Tenreyro, S. (2006) 'The Log of Gravity', *The Log of Gravity,* 88, pp. 641- 658 Available at: https://doi.org/10.1162/rest.88.4.641.
Shepherd, B. (2019) 'The Gravity Model of International Trade: R Version', Available at: https://www.unescap.org/sites/default/files/Gravity-model-in-R_1.pdf.
Slater, M. (2022) *How Do You Value A Football Club?* Available at: https://theathletic.com/3068619/2022/11/22/how-do-you-value-a-football-club/
Westerlund, J. and Wilhelmsson, F. (2011) 'Estimating the gravity model without gravity using panel data', *Applied Economics,* 43(6), pp. 641-649 Available at: https://doi.org/10.1080/00036840802599784.
Wooldridge, J. (2009) 'On estimating firm-level production functions using proxy variables to control for unobservables', 104 Available at: https://ideas.repec.org/a/eee/ecolet/v104y2009i3p112-114.html.