Unequal Migration[*]

Rebecca Freeman
Bank of England & CEP/LSE

John Lewis
Bank of England

J.M.C. Santos Silva
University of Surrey

Stefan Szymanski
University of Michigan

Silvana Tenreyro
LSE, CFM & CEPR

February 2, 2024

Preliminary and Incomplete. Please do not cite without authors' permission.

Abstract

We study the role of skills and their interaction with barriers to migration in the decision to move. These interactions are key to determine the skill composition of the pool of immigrants in different countries. To do so, we use a novel database that documents international and intra-national moves for subjects with different skill levels. The database is unique in the migration literature in its degree of granularity, which allows us to assess the role of skills in amplifying or mitigating existing barriers. Consistent with previous studies, we find that the various barriers to migration, including measures of distance, tend to hamper mobility. But we identify and quantify a novel interaction between those barriers and the level of skills, such that the impact of barriers on migration decreases as the level of skill increases. At highter skill levels, barriers appear to pose little obstacle to the migration of superstars.

Keywords: Migration, inequality, skill differentials, skill drain, geographic labour mobility, gravity equation.

JEL Classification: F22, J24, J40, J61, Z22.

[*] The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees. We are grateful to Francesco Caselli and seminar participants at various universities. We are especially thankful to Elliot Moore for superb research assistance.

1 Introduction

Migration has always been an important public policy issue. The academic literature has made significant progress in understanding the economic impact of immigration, with several articles stressing the critical role played by the human capital or skills of migrants in shaping the labour market outcomes of native workers. (See for example Borjas, 2003, Borjas and Katz, 2007, Ottaviano and Peri, 2012, Dustmann et al., 2012, and the references therein.)

However, less is known about the role that those skills played in determining migration in the first place and many questions remain unanswered. Do the underlying forces driving migration differ across levels of skills or human capital? Do standard barriers to migration have a differential impact on the ability or willingness to move depending on the skill or human capital of the potential migrant? And, if so, how do those barriers shape the skill composition of the migrant labour force in different countries?

Existing studies of the determinants of migration have typically relied on aggregate gravity models using national-level datasets, the mainstays of the migration literature (e.g., Beine et al., 2015, Adserà and Pytliková, 2015, Bertoli and Fernández-Huertas Moraga, 2013, Mayda, 2010). A perennial shortcoming, however, is that those datasets record migration flows at a relatively high level: typically, a headcount of all migrants from a given origin country or region to a given destination in a year. They do not record the skill levels (or any other characteristics) of those migrating—just the raw number. Notable exceptions include Artuc et al. (2015), who use data that records the educational status and gender of migrants.

This limitation of the data typically available leaves traditional models silent about the skills of migrants and on how the determinants of migration might differ with those skill levels. Therefore, such models do not answer many questions that are particularly important in countries with large employment concentration in sectors that act as international hubs.

In this paper we address the question of how highly-skilled migration (that is, migration of superstars) differs from migration in other segments of the skill distribution, by drawing on a novel source of granular data which is unique in the migration literature—the movements of male football (soccer) players drawn from the Transfermarkt website (`www.transfermarkt.co.uk`).

This comprehensive database spans low skill levels to world-class workers; once cleaned the dataset contains over 762,000 individual observations. Once this is converted into an origin-destination-time gravity dataset at the club level, this gives us over 100 million observations (as opposed to typical national-level datasets in the tens of thousands).

Aside from the sheer number of observations, our database has three key advantages over databases typically used in the country-level gravity literature. First and foremost, it records the human capital or measurable skill of workers, as proxied by the player's estimated market value, allowing us to study the differential impact of barriers by skill level. Second, individual employers (clubs) are recorded at the sub-national level, enabling us to directly distinguish between no migration (staying at the same club), internal migration (moving between two clubs within the same country), and international migration (moving from a club in one country to a club in a foreign country), rather than using a proxy for internal migration flows. Third, because we know the exact co-ordinates of the stadiums at which teams play, the distance in any given move is measured very precisely. Thus, we do not have to rely on approximations—such as distance between capitals or distance between major cities weighted by population—which are typically employed in conventional country-level gravity datasets.

By using these data we can address a number of live questions in the policy and academic debate that have, to date, been virtually unexplored in the literature on migration. How do border and distance frictions to migration flows vary with skill levels? Is there a skill level at which migration becomes effectively "borderless"? The distinction between internal and international flows is recognized in the recent literature as crucial in order to properly identify the role of external frictions (e.g., Yotov, 2012 and Yotov, 2022). After allowing for a traditional border effect, does the role of distance in migration flows differ between internal and external migration?

We frame the empirical estimation within a simple model of migration, which leads to a standard gravity equation in which bilateral flows depend on broadly construed measures of distance and size. We then estimate the model for different levels of skill using the Poisson pseudo maximum likelihood (PPML) estimator of Gourieroux et al. (1984) which, as noted by Santos Silva and Tenreyro (2006), is particularly well suited to the estimation of gravity equations and, crucially in our context, is not affected by the existence of zero flows (in the main sample, 99.7% of origin-destination flows are equal to zero, but this proportion is even higher in some of the samples we use).

The following findings stand out from the analysis. First, geographic distance plays a greater role for moves to a different country; this effect is over and above the standard border effect, meaning that the effect of distance as a deterrent to migration is larger, conditional on crossing the border. Second, the role of distance becomes much smaller as the skill level increases; the effect of internal distance even becomes positive at the very top end of the skill distribution. Third, controlling for distance, crossing a country border is an obstacle to mobility that decreases with the level of skills, and becomes largely insignificant at the top of the skill distribution. For the top 1% of players, the effect even becomes negative and significant, meaning that superstars are more likely to cross a national border. Language differences, whether official or ethnic, tend to reduce mobility, but, as with other barriers, less so for those at the top of the skill distribution. Finally, the results suggest strong assortative matching, with players more likely to move between teams with similar average skill levels.

Our paper contributes to three distinct literatures. The principal contribution is to the literature on gravity models of migration. The sub-national nature of our dataset allows us to address questions that until now have been largely ignored. Our findings also have relevance for the body of work on migration and skills. While most of that work has focused on the impact of migrant skills on the labour market of natives, our findings address the question of how those skills influence the decision to migrate. Finally, our paper also adds to the literature on labor markets in sports. Sports data has long been recognized as an almost ideal "laboratory" for the study of labor markets (e.g., Kahn, 2000), and recent work has emphasized the value of sports data for economic research (Palacios-Huerta, 2023). However, one limitation of labor market studies in sports is that they have tended to focus only on the top end of the market, that is, the superstars.[1] This focus on the superstars has led to a neglect of the broader labour market.[2] We thus extend and complement this literature by focusing on the entire skill distribution.

The rest of the paper is organized as follows. Section 2 describes the dataset we use and related summary statistics. Section 3 presents a simple conceptual model to frame the estimation and describes the approach we take to estimate the models of migration. Section 4 describes our findings and links them to the relevant literature. Section 5 presents checks on the robustness of our results to changes in the set of clubs included in the sample and to the stratification of the sample by the age of players. Finally, Section 6 concludes.

2 The Database

In what follows, we describe the main dataset used in our analysis, highlighting why it is particularly well suited to study the role of skills in migration flows. We then present key summary statistics and detail how the data was processed and converted to the level of aggregation relevant for our study.

2.1 Transfermarkt Data

The data on players and clubs used in this paper is taken from the website Transfermarkt (TM). The website was created in 2000 by Matthias Seidel, and was subsequently acquired by the multinational media company Axel Springer. Although German in origin, the website now has twenty different country sites. Its original purpose

[1] For example, Kleven et al. (2013) consider taxation effects on superstars in European football, Buraimo et al. (2015) consider the impact of contract length on a sample around 1000 professional players in the Germany's top league, the Bundesliga. Lucifora and Simmons (2003) examine wage determination of 533 "superstars" in the top two Italian football leagues. Bryson et al. (2014) likewise use data on 906 players in Italy's top two divisions to examine the impact of migration on wages.

[2] In the football market, this amounts, globally, to more than 100,000 players across more than 300 leagues at any one point in time.

was to document the career histories of football players, and has since expanded to include managers and other staff.

One of the most interesting aspects of the TM data is the player valuations, which started to be published on the site from 2004. Player values are estimates compiled by forums following individual clubs, for which fans sign up. Herm et al. (2014) describe the process involved in arriving at player values. Each club forum has a moderator who coordinates discussion. Each member of the forum is encouraged to suggest a valuation for an individual player, backed up by a rationale. The job of the moderator is to decide on the consensus value. The aim is to update the valuation of each player on a team at least twice a year.[3] In 2021, TM had 80 full time employees and over 1,000 volunteers involved in managing the process.[4] According to a post on its website about the 20 years of TM, in 2020 it had more than 600,000 registered users, each of whom could participate in the valuation process.[5]

The TM valuations have generated a significant amount of interest, including among academic researchers. The valuations are frequently cited as examples of the "wisdom of the crowd" (see, e.g., Herm et al., 2014, Müller et al., 2017, Peeters, 2018, Prockl and Frick, 2018).[6]

As explained also in a TM website post,[7] the valuation process attempts to arrive at a "market value" reflecting all aspects of a player's history. A natural interpretation of the TM values—and the use to which they are put in this paper and others—is as a cardinal ranking of player abilities, i.e., skills. For example, the ratio of TM aggregate squad values is a reliable predictor of match results, closely comparable in accuracy to bookmaker odds (see, e.g., Reade and del Barrio, 2023). Pre-season TM values are accurate predictors of

[3] See https://theathletic.com/3085749/2022/01/27/premier-league-how-do-you-value-a-player.

[4] "The Wisdom of the Crowd," New York Times, Aug 12, 2021, see https://www.nytimes.com/2021/08/12/sports/soccer/soccer-football-transfermarkt.html.

[5] See https://www.transfermarkt.co.uk/20-years-transfermarkt-from-non-league-football-to-the-champions-league/view/news/363452.

[6] The reliability of these estimates are widely commented on the in football press, and it is said that clubs and agents often use TM valuations as a baseline for negotiating contracts, both in relation to player wages and to transfer fees. See, e.g., www.theguardian.com/football/2020/dec/19/top-football-clubs-relying-on-transfer-valuations-made-by-volunteers.

[7] See https://www.transfermarkt.co.uk/market-value-definition/thread/forum/357/thread_id/3433.

end of season league rank (Szymanski, 2022) and national team performance (Prockl and Frick, 2018). They are also closely correlated with player salaries (Prockl and Frick, 2018)—which should reflect differences in productivity/skill levels. Finally, we note that there is also a reasonable number of players (39.4%) for whom there is no TM value. This is often the case for players at low-levels who have not had sufficient exposure in football to warrant a valuation, or for players at clubs with very small fan-bases (and thus limited fans to provide valuations).

It should be stressed that TM values are not, and are not intended to be, estimates of "transfer fees" paid by a club that acquires the services of a player currently under contract with another club. In theory, a transfer fee is a payment in compensation for losing the services of the player for the remainder of their contract. For our purposes, TM values have several important advantages over transfer fees. First, exact sums paid are not always officially reported, although clubs and player agents often leak a figure that is then reported in the press, generally for high profile transfers. Second, the transfer fee depends in part on how long the player has remaining on their contract; in the limit, if a player is out of contract, he is a free agent and no transfer fee is payable. Third, transfer deals may include other contingent payments (e.g., for goals, appearances, or a percentage of future transfer fees) or player swaps which are not captured in the headline fee. Fourth, by definition, transfer fees only exist for players who move between clubs. In this paper we consider players who stay at clubs, as well those who move, in order to estimate frictions associated with moves between clubs, leagues and countries.

The raw TM data comes in two forms: "stocks" of players and "flows" of player movements over time. Both datasets cover the time period 2004-2019. In addition to identifying information on individual football players (including their name and date of birth), in the stocks data we also observe players' nationalities, TM value, the team for which they play in each season, as well as the league (including the league division) and country in which the team is located. The flows data comprises the same identifying information on football players, in addition to a list of all player moves across seasons, including whether the move was a true transfer or a player loan. Specifically, for each season, if there was a transfer, we observe the club left and the club joined.

To construct a comprehensive dataset on players and their movements across time, we combine the information on stocks and flows, excluding youth players because of their non-comparability with professional players in the dataset. Importantly, combining the stocks and flows allows us to identify both "stayers" (players whose club in time t is the same as in time $t-1$) as well as "movers" (players whose club in time t is different than in time $t-1$). Movers can be either domestic, moving to a club in the same country in time t as in time $t-1$ or international, moving to a club in a different country in time t as in time $t-1$.

Armed with this information, we then use data on the geo-coordinates of the stadium in which each team plays, which can be found on the TM website for most clubs,[8] to construct a precise measure of bilateral distance between origin and destination associated with each move, domestic or international. Note that the distance variable for stayers is necessarily equal to zero.[9] Further, we construct additional variables for the continent associated with player nationality and the continent associated with a club's location.[10]

[8] When the information is not available on the TM site, we used the clubs' own websites and Wikipedia to find the exact location. We performed extensive checks and manual corrections to ensure the information is correct.

[9] It is also possible that the distance variable for movers is zero if the club to which a player moves plays at the same stadium as their previous club. In our dataset, this is the case for a trivial number of observations (1,421, or 0.19% of total observations).

[10] We map countries (both for clubs' locations and players' nationalities) to continents using the United Nations (UN) Geoscheme, a classification which provides a correspondence between 248 countries and territories into six continental regions (Africa, Asia, Europe, North America, Oceania, and South America). Note that the there are potential differences with the FIFA country-continent classification, used for defining football federations. Using the FIFA country-continent classification instead of the UN Geoscheme does not materially impact our analysis, however, as the differences between the two classification systems are generally small. UEFA, the European federation, closely maps to the UN definition of Europe; CAF, the African federation, closely maps to the UN definition of Africa, and so on. The one significant difference is between the UN definition of North America (US, Canada and Mexico) and the regional federation CONCACAF, which includes not only those three countries, but also Central America and the Caribbean islands). However, the fraction of players in the dataset from these regions is relatively small.

2.2 Descriptive Statistics

Combining data on stocks and flows over the period 2004-2019, we identify origin and destination clubs for 762,741 distinct cases, involving 681,050 player-season observations (some players move more than once per season). This is made up of data on 200,295 individual players and 7,104 clubs across 315 leagues in 93 different countries. The TM data also identifies the primary nationality of each player, of which 212 nations/territories are represented.

As mentioned above, if a player's origin and destination club are the same, this constitutes a stay, meaning that the player remained in employment at the same club during a given season. Football players who are stayers throughout the entire observation period are the minority. Indeed, many players move several times over their careers, and in our data only 83,799 players (41.9%) are not associated with any moves. This likely overstates the share of players who never move in their career, however, since the data does not cover the entire career history of every player. Moreover, roughly 72% of observations consist of cases where the country of the destination club is the same as the primary nationality of the player.[11] The average age of players in the data is 24.2 years. Movers are, on average, almost half a year younger than stayers (24.0 years old versus 24.4 years old, respectively).

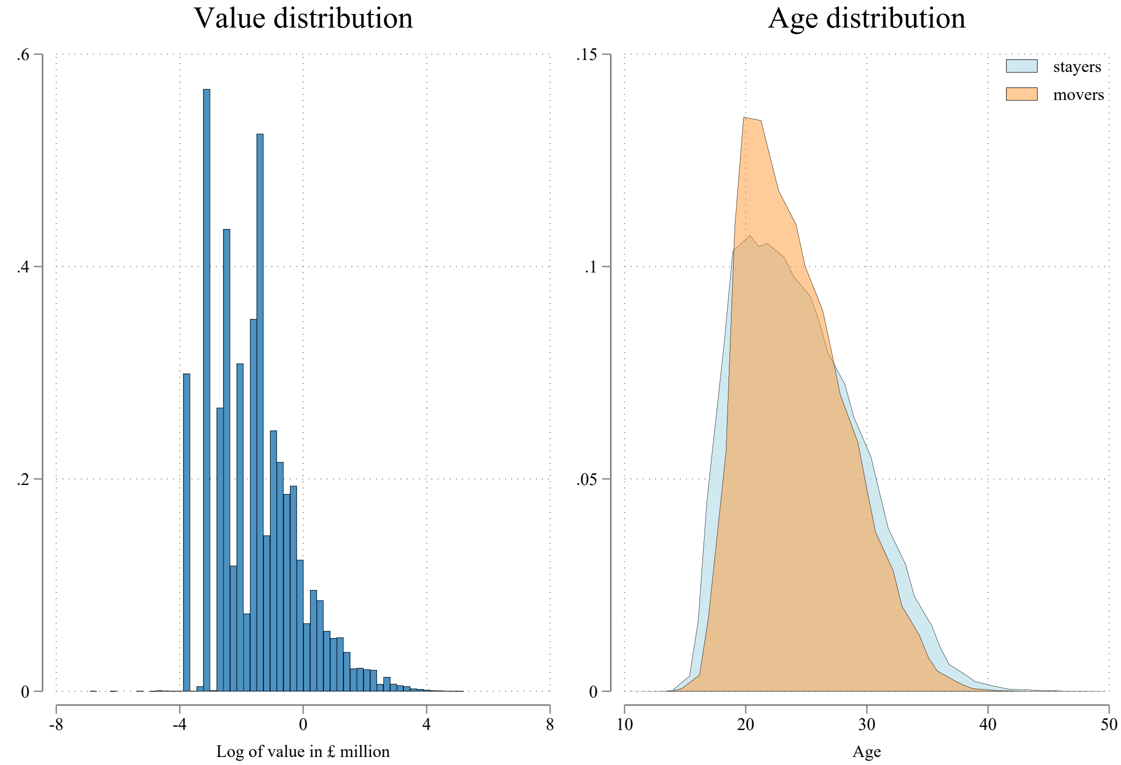
In addition to player characteristics and movements, we have unique insight into their relative ability through their TM value. As one should expect, these values are highly skewed. As noted before, 39.4% of players in the data have no TM value. For the purposes of the summary statistics presented below, we treat missing values associated with observations for these players as zeros, as they are associated with players at low-levels who have not had sufficient exposure in football to warrant a valuation, or players at clubs with very small fan-bases (and thus limited fans to provide valuations).[12] Having filled in the missings with zeros, the median TM value is £68,000. The 75th percentile

[11] This could be driven by: (i) players who move within their home country, and/or (ii) players who return to their home country having played abroad.

[12] In our regression analysis we do not fill missing values with zeros. Instead, these players enter into a seperate category of players identified by having "no value".

value is £270,000, the 90th percentile value is £765,000, and the 95th percentile value is £1.8 million. The highest recorded value (Kylian Mbappé in 2018) was £180 million. Figure 1 below illustrates the log distribution of player values (excluding the zeros), as well as player ages (by mover and stayer status).[13]

Figure 1: Value and age distribution of TM data



Source: Authors' illustration based on TM data. Notes: The left panel plots the (log) value distribution in £ million, excluding the zeros. The right panel plots the the age distribution, classifying players as movers or stayers.

In terms of geographic distribution, it is well known that the wealthiest clubs are located in Europe, and that Europe has been a major magnet in recent decades for the migration of players, especially from South America and Africa. This income divide is spotlighted in Table 1. Specifically, the first two columns show the number of times and share, respectively, that players appear in a given continent. The third column shows the average TM value (in £ millions) associated with these continent-based observations. In contrast, the fourth and fifth columns show the number of times and share, respectively,

[13] Without filling in the missings with zeros, the median TM value is £180,000; the 75th percentile value is £450,000, the 90th percentile value is £1.3 million, and the 95th percentile value is £2.7 million.

that players whose primary nationality maps to a given continent appear in the data. The final column shows the average TM value (in £ millions) associated with the players these nationality-based observations.

The main takeaway from Table 1 is that the wealthiest clubs by far are in Europe; however, players from South America and Africa have the highest valuations. The average valuation of South American players, for example, is £985,000, compared to £557,000 for Europeans. The average valuation of African players is £626,000. Players based in South America and Africa, however, have significantly lower valuations (£431,000 and £161,000, respectively). These lower valuations are reflected in the migration patterns of African and South American nationals to Europe, as shown in Figure 2.

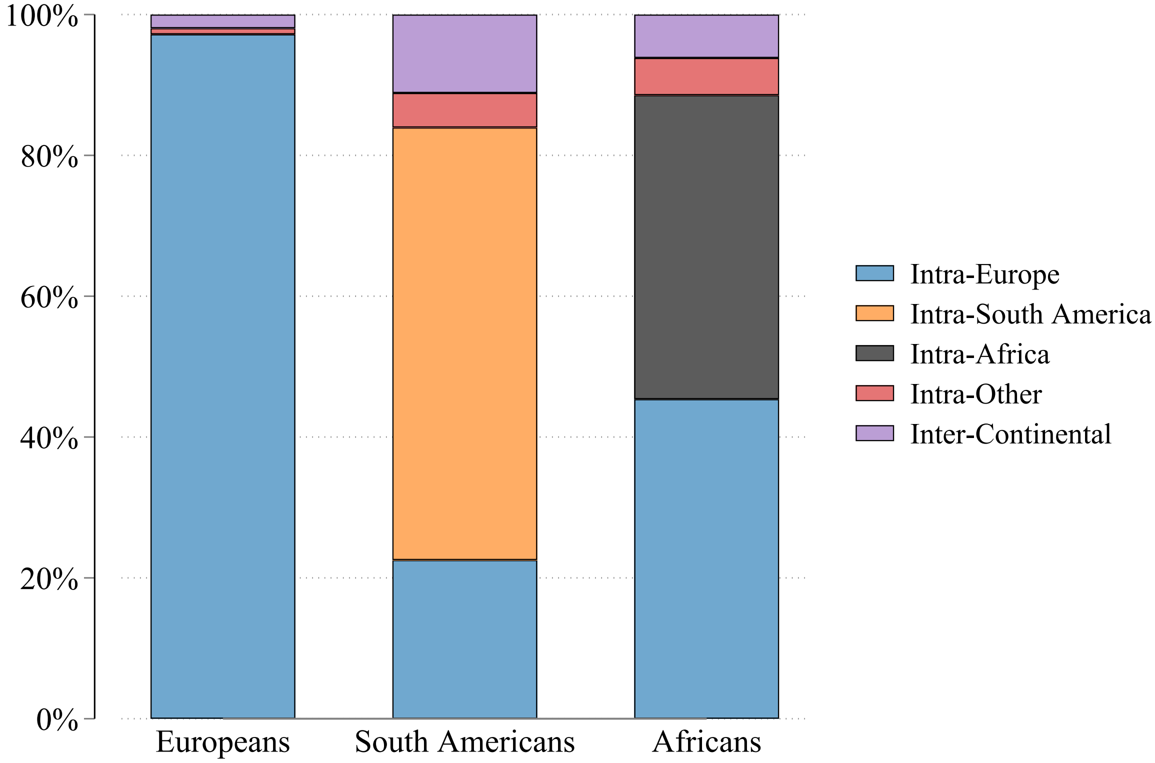
Table 1: Player observations and average TM value

	Continent of teams			Player nationality		
	Obs. (#)	Share (%)	Avg. value (£ mil.)	Obs. (#)	Share (%)	Avg. value (£ mil.)
Africa	24,254	3.18	0.161	53,789	7.05	0.626
Asia	106,973	14.02	0.261	101,290	13.28	0.214
Europe	573,442	75.18	0.543	522,420	68.5	0.452
North America	8,878	1.16	0.346	12,255	1.61	0.557
Oceania	4,672	0.61	0.201	5,526	0.72	0.323
South America	44,567	5.84	0.431	67,421	8.84	0.985
Total	762,741	100	0.481	762,701	100	0.481

Notes: This table shows the the number of times and share that players appear in the data, by team continent and player nationality, as well as associated average TM values. The total number of observations in the second and fifth column differ slightly because we do not observe player nationalities for 30 players, some of whom appear more than once in the data across seasons.

Player migration is complex. Although there is significant migration from other continents (notably South America and Africa) to Europe, these movements are relatively small in relation to the overall frequency of player movements within Europe. This point is illustrated in Figure 2, which plots intra- and inter-continental moves for players with African, South American and European nationalities. For each nationality, the blue bars tell us the share of player moves within Europe. The orange bar in the second column tells us the share of moves within South America by South Americans and the black bar in the third column tells us the share of moves within Africa by Africans. The red bars tell us the share of intra-continental moves of African, South American and European players already located in other continents (namely, North America, Asia, and Oceania), and lastly the purple bars tell us the share of African, South American and European players who move across continents.

Figure 2: Share of intra- and intercontinental moves, by player nationality

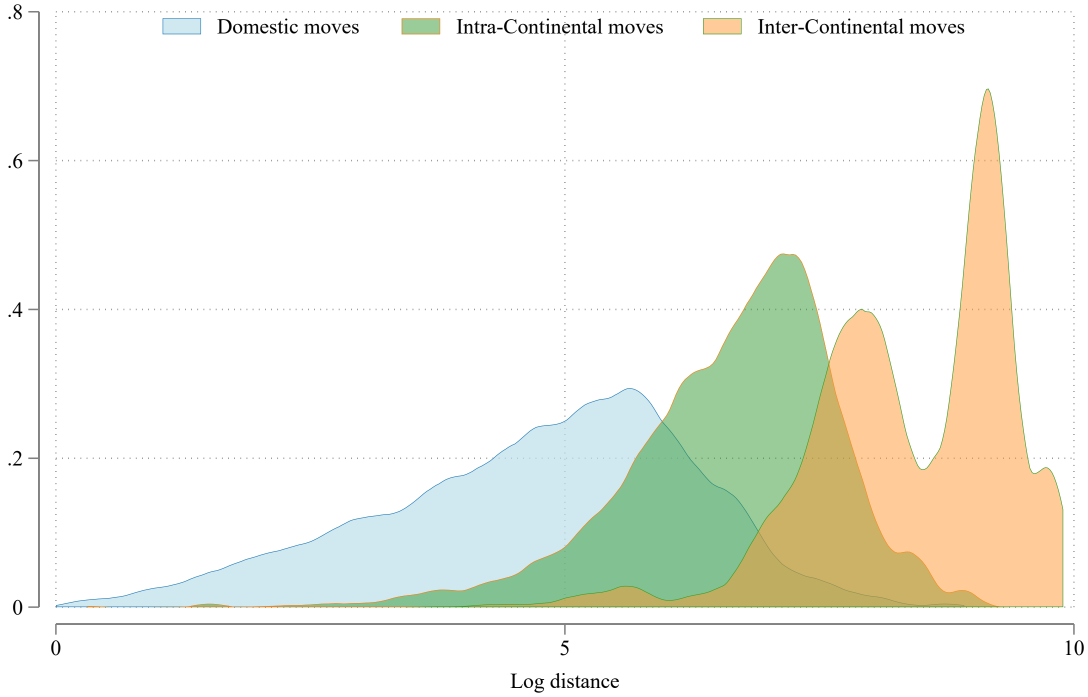


Source: Authors' illustration based on TM data. Notes: This figure shows the share of intra- and intercontinental player movements for players of European, South American and African nationality.

The takeaways are threefold. First, for European nationals the vast majority of moves (97%) are within Europe. Second, and perhaps more interestingly, a non-negligible share of moves by South American and African players are also intra-European (23% and 45%, respectively). This is likely connected to the fact that players are often recruited by clubs at a very early age. It is not uncommon for a player to move to a big club several years before they can play as a professional (at age 17). Thus, many players of non-European origin may have moved within Europe during their career, while crossing a continental boundary only once. Third, South American and African players also have a large share of intra-continental moves (61% and 43%, respectively) and these numbers dwarf inter-continental moves.

The range of distances travelled when players move clubs also varies substantially depending on the type of move. In Figure 3 we partition moves into three categories (domestic moves, intra-continental moves, and inter-continental moves) and plot the (log) distance distribution of each type of move.

Figure 3: Distance distribution, by type of move



Source: Authors' illustration based on TM data. Notes: This figure plots the density of the (log) distance of moves (in kilometers) by type of move (domestic, intra-continental, and inter-continental). For presentational purposes, we truncate the distribution at zero to show moves of more than 1 kilometer only. Moves of less than 1 kilometer are typically associated with moves by players to clubs which play at the same (or extremely nearby) stadiums, and account for only 0.22% of total observations.

The first interesting point to notice is that domestic moves cover a very wide range of distances (from essentially zero kilometers to just over 7,500 kilometers). While there is a critical mass just to the right of the middle of the distribution, there are also long tails on both ends, and in fact some domestic moves are as far or further than inter-continental moves. This is reflective of the fact that some countries themselves span several thousand kilometers.

The second interesting point relates to inter-continental moves. We observe "twin peaks" in the distribution for this type of move, with the first peak at a density of around 0.4 and the second peak at a desity of around 0.7. This is related to the fact that there is a smaller mass of player movements between nearby continents than there is between continents that are further away from one another. Otherwise stated, we observe relatively more players who cross long distances when moving continents, indicating that moves between, say, South America and Europe are more prevalent than moves between North Africa and Europe. Finally, and unsurprisingly, intra-continental moves span, on average, a larger distance than domestic moves and a shorter distance than inter-continental moves.

2.3 Club-level Flows

So far, we have presented summary statistics which are reflective of the merged stocks and flows TM data at the most disaggregate, i.e., player, level. However, for our empirical analysis we are interested in club-level moves, as well as cases in which a bilateral move could have, but has not, taken place. In order to shape the data as required, we collapse the raw TM data to the club level for each time period, yielding a dataset which records the total number of flows between each origin and destination club in each time period. Critically, the total number of stayers at the bilateral club level are reflected in cases where the origin club and destination club are the same, while the total number of movers are reflected in cases where the origin club and destination club are different.

Next, we construct a 'skeleton' dataset, which consists of all possible origin club-destination club-time triads present in the data in a given year. Otherwise put, this skeleton provides us with the full set of possibilities for player moves in any given time period. We then match all player moves in our raw dataset to the appropriate origin-destination-time observation. If there are no moves between a given origin club-destination club pair in a given year, this is then recorded in our dataset as a zero flow. As with other gravity applications, these zero flows are important in their own right and help us identify the determinants of migration.

We are also interested in the relative skill level of each club, as this provides useful information about transfer propensities (for example, when a worker moves to a new firm, is it likely that the average skill level at that firm will be similar to the worker's?). Hence, as we collapse the data to the club level, we also create a time-varying, bilateral "skill gap" variable which measures the (absolute value of the) difference between the median player TM value for each club pair. A small skill gap value thus indicates that player valuations are similar across clubs while a large value is indicative of larger differences in player valuations across clubs. Finally, we construct binary variables for (i) whether origin and destination clubs are in the same country and (ii) are the same. The latter, that is, the "same team" dummy variable, is analogous to the "same country" dummy variable typically used in the gravity literature, but at a finer level of granularity to match the level of disaggregation of our dataset.

As a last step, we match information on the countries in which clubs' are located with standard gravity variables on contiguity, official language, and ethnic language from the CEPII GeoDist database (Conte et al., 2022). Our final dataset at this level of aggregation consists of 105,524,512 origin club-destination club-time triads.[14]

3 A Model of Migration

In this section we present a simple framework to describe a player's decision to stay at his current club or to move to another club, which can be located in the same or in a different country. There are $o = 1, 2, 3, \ldots, O$ origin clubs and $d = 1, 2, 3, \ldots, D$ possible destination clubs, that include the club the player is currently at. In each period $t = 1, 2, 3, \ldots, T$, players maximize their utility across the full set of destinations, and move if the maximum utility is obtained at a club different from their current one.

As in Case 2 of Guimarães et al. (2003), we assume that players with level of skill/human capital h currently in club o are homogeneous, and write the (indirect) utility

[14] All regressions whose results are reported in Section 4 are based on samples of this size; the reported number of observations is smaller because it excludes observations that are singletons or dropped due to separation.

that a player with skill level h currently in origin club o would obtain in club d at time t as

$$u_{hodt} = z'_{hodt}\beta_h + \varepsilon_{hodt},$$

where z_{hodt} is a vector of observable characteristics of the origin and destination teams at time t, as well as of the cost of moving to d; ε_{hodt} is the unobserved component of the utility; and β_h is a vector of parameters that may vary with the skill level.

As is usual in this literature, we assume that ε_{hodt} follows an *iid* extreme value distribution,[15] and consequently the probability that a player with skill h moves from club o to club d in period t is given by a logit model of the form (see McFadden, 1974):[16]

$$\Pr\left[u_{hodt} = \max_k u_{hokt}\right] = p_{hodt} = \frac{\exp[z'_{hodt}\beta_h]}{\sum_{d=1}^{D}\exp[z'_{hodt}\beta_h]}. \tag{1}$$

Assuming that players decide independently and letting N_{hot} denote the number of players of level h in club o at time t, it follows that the expected number of players of level h moving from o to d at time t is given by the following exponential model:

$$\mathrm{E}\left(N_{hodt}|z_{hodt}, N_{hot}\right) = p_{hodt}N_{hot} = \exp[z'_{hodt}\beta_h + \ln N_{hot} + \ln R_{hot}], \tag{2}$$

where N_{hodt} denotes the flow of players of skill level h from origin club o to destination club d at time t, and $R_{hot} = \sum_{d=1}^{D}\exp[z'_{hodt}\beta_h]$.

To complete the specification of the model, we need to spell out the set of variables in z_{hodt}. The specific set of variables comprised in z_{hodt} is detailed in the next section, and varies based on the empirical exercise in question. However, as mentioned above, z_{hodt} includes i) characteristics of the players of skill level h in team o at time t, subsumed in the

[15] Note that the assumption that the errors follow an extreme value distribution is restrictive, but not as much as it may seem because any model can be approximated by as a logit. Suppose that $p_{hodt} = f\left(z_{hodt}\right)$, where $f\left(\cdot\right)$ is an unknown function. By definition $\sum_{d=1}^{D} f\left(z_{hodt}\right) = 1$, and therefore we can write

$$p_{hodt} = \frac{e^{\ln f\left(z_{hodt}\right)}}{\sum_{d=1}^{D} e^{\ln f\left(z_{hodt}\right)}},$$

which is a standard logit if $f\left(\cdot\right)$ is also a logit, or can be approximated by a logit with a sufficiently flexible specification.

[16] Similar logit mode can also be motivated by a matching model; see Galichon and Salanié (2022).

skill-origin-time fixed effects ϕ_{hot} which also capture $\ln N_{hot}$ and $\ln R_{hot}$; ii) characteristics of the destination team at time t, which are allowed to vary by skill level and are captured by skill-destination-time fixed effects ψ_{hdt}; and iii) the geographic distance as well as other measures of the cost of moving from o to d at time t, denoted τ_{odt}, whose parameters we allow to vary by h. Note that ψ_{hdt} accounts for the size of the destination club and therefore (2) can be interpreted as a gravity equation in which the flows depend on the sizes of the origin and destination, as well as on the distance between them.

The results in Guimarães et al. (2003) imply that estimates obtained from the multinomial logit model defined by (1) are identical to the PPML estimates of (2), and that is the approach we will follow. Therefore, we estimate by PPML two-way gravity models of the form:

$$\mathrm{E}\left(N_{hodt}|\tau_{hodt},\phi_{hot},\psi_{hdt}\right)=\exp[\tau'_{hodt}\beta_h+\phi_{hot}+\psi_{hdt}].$$

4 Empirical Results

This section presents our main results and is divided into two parts. We start by presenting the estimates obtained for the sample of all players, and then present the estimates obtained for players in different skill levels.

4.1 Determinants of Migration: Aggregate Results

Table 2 reports our baseline estimation results when the data are pooled across different skill levels. The models in columns (1) and (2) are standard gravity equations, augmented with variables that are specific to our data. The construction of the bilateral distance variable is described in Sub-section 2.1. Here, we also distinguish whether the bilateral distance associated with player movements is within country (i.e., internal) or crosses an international border (i.e., external). The skill gap variable (which is continuous), binary same team variable (equal to one if the origin and destination team is the same, i.e., for players who did not move), and binary same country variable (equal to one if the origin and destination country is the same), are described in Sub-section 2.3. Finally, the variables for contiguous countries, and common ethnic and official language are binary variables equal to one if countries share a border, official, and ethnic language, respectively.

As alluded to above, the distinctive feature of the Table 2 results is the inclusion of measures of bilateral distance (log kilometers) that distinguish moves that are wholly within national borders (internal) and moves across national borders (external). In column (1), distance is undifferentiated and the estimated elasticity is -0.365. In column (2) we split the distance variable into internal and external components. Here, we observe that the internal distance coefficient is roughly 25% smaller than the external distance coefficient, indicating that distance has a significantly stronger negative effect on international compared to domestic moves.[17]

The regressions also include variables that capture the effect of the "the cliff at the border" (Bertoli and Fernández-Huertas Moraga, 2015). These are the "same country" and "contiguous countries" dummies. The size of these coefficients suggest that there is a large pull for players to remain in their same country should they move, or when international moves occur, to move to a neighboring country.

Whilst these effects seem large, the estimates of the "same team" effect suggest that there is nearly as much of a cliff at the border of the team as there is at the border of the country. While there is an additional cost to crossing the border, comparing the point estimates on the same team and same country variables in columns (1) and (2) suggest that half to two thirds of the cost of moving is incurred at the club's "border." [18]

The positive coefficient on "common ethnic language" suggests additional cultural barriers to mobility, as is standard in the literature. (Ethnic and official language dummies

[17] The difference between the two parameters is statistically significant at any conventional level.

[18] The inclusion of the same club dummy is motivated by the recent structural gravity literature (see for example, Yotov, 2012; Yotov, 2022; Anderson and Yotov, 2022) who note that to evaluate cross border flows *relative to domestic flows*, one must explicitly include a measure of domestic flows in the equation (i.e., trade/migration flows within the country). At a club level dataset, this same logic implies that to study the relative costs of moving between clubs (as opposed to not moving), one must therefore include players who remain at the same club (origin club and destination club are the same), and pick up this effect with a dummy variable.

are highly correlated, so including the latter does not add explanatory power.) It is likely that the difference in the internal and external distance effects also accounts for some of the cultural barriers not captured by our other variables.

Table 2: Baseline results: All transfers

	(1)	(2)
Distance: internal		-0.356***
		(0.010)
Distance: external		-0.474***
		(0.025)
Clubs' skill gap	-0.466***	-0.466***
	(0.011)	(0.011)
Same team	2.280***	2.318***
	(0.038)	(0.038)
Same country	4.221***	3.344***
	(0.043)	(0.185)
Contiguous countries	0.523***	0.355***
	(0.046)	(0.052)
Common ethnic language	0.730***	0.779***
	(0.058)	(0.058)
Common official language	0.056	-0.002
	(0.065)	(0.067)
Distance	-0.365***	
	(0.010)	
Observations	99,046,476	99,046,476

Notes: Standard errors clustered by club of origin and destination in parenthesis; , the numbers of clusters are 5,935 and 6,199, respectively. Stars indicate statistical significance at 0.01 (***), 0.05 (**), and 0.1 (*). Models include origin-time and destination-time fixed effects.

In terms of the "skill gap" variable, the negative coefficient suggests that all else equal, players are more likely to move between clubs of similar skill levels. This result is rather intuitive and suggests assortative matching; in Sub-section 5.2 we discuss how this result varies by player age.

The inclusion of variables not typically found in gravity models of migration (internal distance, same club, and skill gap) all capture frictions that affect mobility above and beyond national borders. They suggest, unsurprisingly, that there exist barriers and disincentives to mobility within a region as well as between regions. Taking these into account, the cliff at the border does not look as high. The results suggest that, for this type of worker at least, there is a certain "stickiness" in employment, that when an employee does move, it tends to be to a similar firm, and that it will likely be close to home. All that said, national borders still represent significant barriers to mobility, either directly or indirectly through the external distance variable.

4.2 Determinants of Migration: Results by Skill Level

We now turn to the central results of our study, the results obtained when the model is estimated by skill level. To do this, we group players by their TM value, which is our measure of skill, and estimate separate equations for players within each skill group. To create the groups, in each year, we split players by percentiles of their value, and construct origin club-destination club player flows for each skill group.[19] As with most skilled occupations, the earnings/value distribution is characterized by a large mass concentrated in ranges from the very lowest to relatively high values, and a long tail of "superstars" with extremely high values. Superstars tend to draw the most attention, but we are interested in characterizing the entire distribution, and contrasting the different economic forces operating at different points along the distribution. To this end we estimate gravity equations for a number of different ranges in the skills distribution.

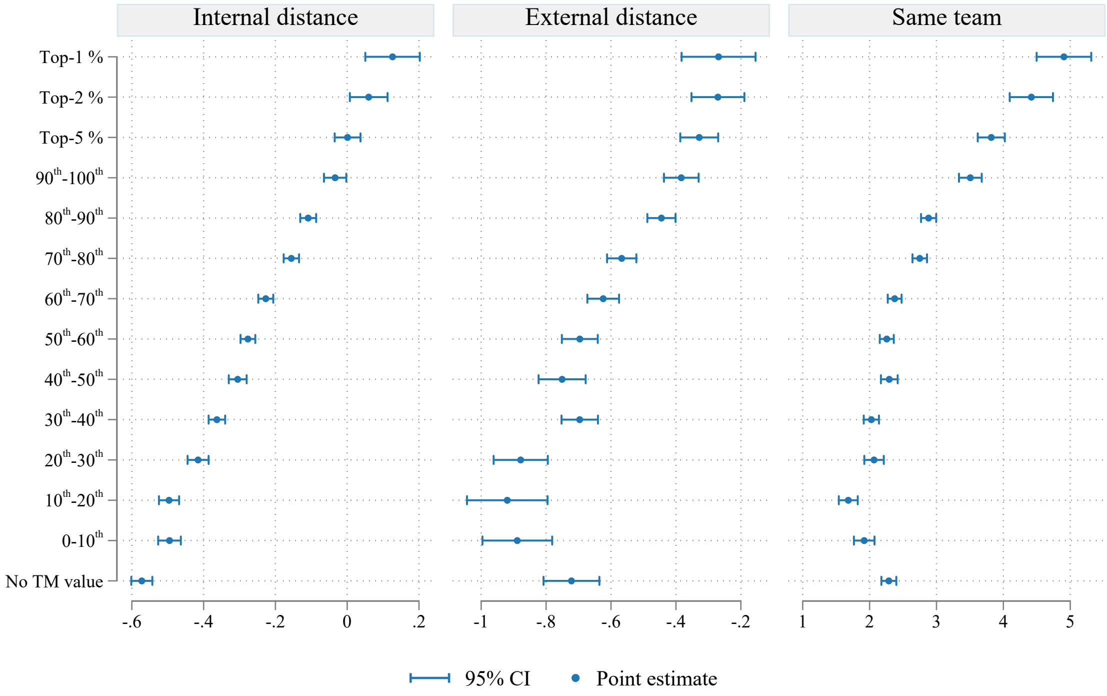
For the analysis by skill level, we divide the sample into deciles by TM value, with another bucket for those players with no TM value, and then in addition we identified the top 5%, top 2% and top 1% percentiles. This provides a set of estimates that represent players from the lowest level (likely semi-professional), journeymen, high skill and superstars.

The results of interest can be reported graphically, as we do in Figures 4 and 5. Corresponding estimates for the ten deciles are reported in Table 3.

[19] Because we rank player values within each year, our approach allows us to make meaningful comparisons over an extended period of time without the need to use a price index to deflate players' values, and ensures that the share of observations from a given season is approximately the same in each decile. Over time, the number of players reported on the TM website has grown while the average value has fallen a little, reflecting the expansion of the website's coverage, but as we will show, the results are robust to restricting the sample to club pairs that are present for a fixed number of years.

Figure 4 illustrates the three main distance-related effects that are the focus of this paper—internal distance, external distance and "same team." For each graph, the dot represents the point estimates, and the line represents the 95% confidence interval.

Figure 4: Distance elasticity estimates, by TM value



This figure plots point estimates for internal distance, external distance and same team variables from our baseline regression run separately when splitting the sample by skill level. As presented, we split the full sample into: players with no TM value; deciles by TM value; and the top 5%, top 2%, and top 1% percentiles. All regressions include origin-time and destination-time fixed effects.

These results confirm the earlier finding that external distance imposes more of a constraint than internal distance. However, the more striking feature of this figure is that both distances are less and less of a barrier as one moves up the value distribution. Indeed, the drag of internal distance decreases smoothly as we move up the skill ladder. For external distance, we find a less smooth pattern, but the main message is identical: distance matters less for the more skilled. These results are broadly consistent with those of Artuc et al. (2015), who found significantly lower distance elasticities for college-educated workers.

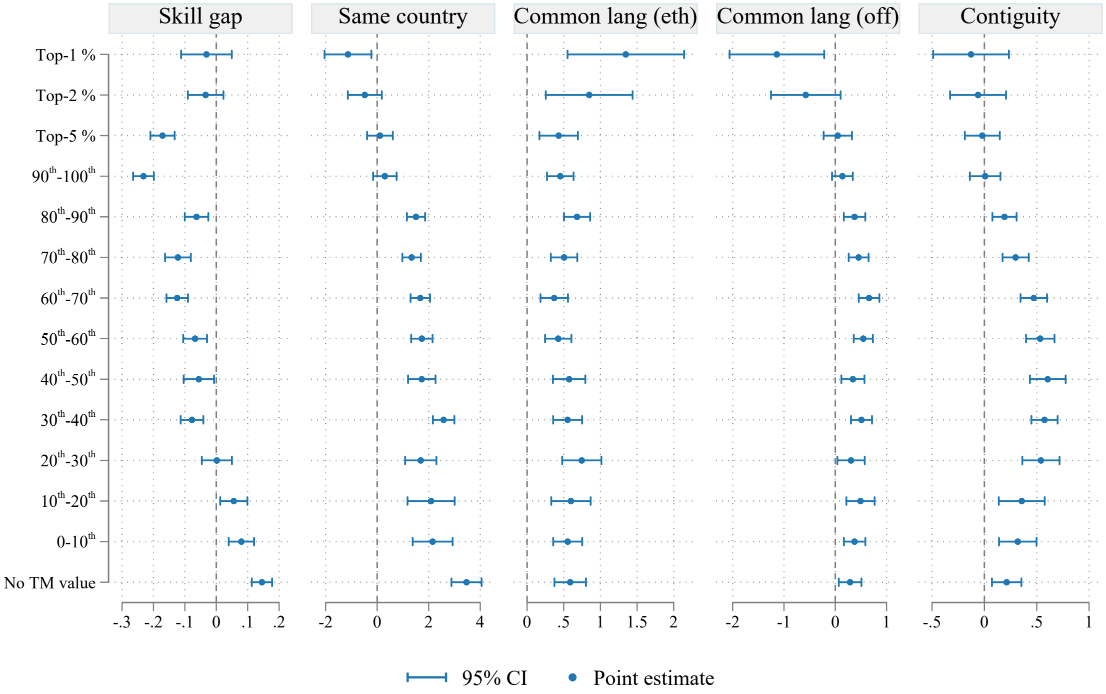
Another point that Figure 4 makes clear is that the difference between the roles of internal and external distances changes with the skill level. At the lower end of the decile distribution (0-40th percentiles) external distance coefficients are roughly twice the magnitude of internal distance coefficients. However, this gap significantly widens as skill level increases, and in the top decile the magnitude of the external distance coefficient is over 11 times that of the internal distance coefficient.

While external distance still exerts a negative influence on mobility at the top end of the value distribution, internal distance actually exerts a positive influence; high value players in the same country are more likely to move to a club that is further away than one that is close. This result is likely to be specific to the particular type of worker we are considering and reflect the fact that top players may prefer to move to a more distant team to moving to a local team with which there is an intense rivalry. While the mechanism may not be identical, the same phenomenon may happen in other professions where non-compete clauses may prevent moves to close rivals. What this result uncovers is that there might be additional barriers to mobility that could be relevant for particular high-end jobs that are not picked up by standard gravity equations at higher levels of aggregation.

Interestingly, the "same team" variable displays the opposite trend, indicating that players' propensity to remain at the same club is higher as they move up the value distribution. Looking at the distance and same team results together, we conclude that superstars are significantly less likely to move than the average player, but when they do move, distance is largely irrelevant. This helps to explain the highly nonlinear distribution of earnings, consistent with Rosen's (1981) well known account of superstar earnings. Because low skill players are poor substitutes for high skill players, and because the media allows the performance of players to be viewed by a global audience, the scarce talents of high skill players attract very large salaries. Those who employ high skill players cannot easily replace them, and hence they create endogenous economic barriers to their mobility once the exogenous barriers become less binding.

Figure 5 illustrates the differences across the value distribution for the effects of the other variables in our regressions (with corresponding estimates for the ten deciles reported in Table 3).

Figure 5: Additional regressors, by TM value



This figure plots point estimates for skill gap, same country, common ethnic language, common official language and contiguity variables, from our baseline regression run separately when splitting the sample by skill level. As presented, we split the full sample into: players with no TM value; deciles by TM value; and the top 5%, top 2%, and top 1% percentiles. All regressions include origin-time and destination-time fixed effects.

The "skill gap" effects are interesting. Recall that at the aggregate level, this effect was negative, suggesting that players tend to move to clubs that have similar levels of skill on average. However, in Figure 5 it is apparent that at the bottom end of the distribution the effect is positive; low value players tend to move to clubs where the skill gap is large. From the middle to the top end of the distribution they tend to move to clubs with similar skill levels, while at the top end the skill gap is less relevant. This likely reflects the career progress of players.[20] The TM value of players at the beginning of their career will be low, will tend to rise as they become established (there is a selection issue here—low value players will tend to disappear as they age), and those who emerge as potential stars will move to bigger clubs. Once established, players will spend most of their career playing at

[20] Note that, because some of the explanatory variables vary by skill level, the estimates obtained with the full sample are not necessarily a convex combination of the estimates by decile; see Breinlich et al. (2022) for details.

the same level, and hence the sign of the skill gap effect will reverse. For the top 2% and top 1% quantiles, the estimates are still negative, but estimated with little precision.[21]

The "same country" effects are positive and significant at the bottom end of the distribution and then reverse sign at the very top end of the distribution, becoming negative and significant for the for the top percentile. This is consistent with the distance effects discussed above: as one progresses up the value distribution, national borders become less relevant. The negative coefficient for the top one percentile offsets in part the persistent negative effect of external distance. The negative "same country" effect for superstars is likely driven by the same reluctance of top clubs to trade big stars to their local rivals.

"Contiguity" is similar, in that it exerts a positive effect at lower levels of the distribution while having no influence at the very top end.

The language effects are less clear. Common ethnic language appears to have a larger effect as one moves up the distribution, while common official language has a positive effect only at the lower part of the distribution, reversing sign at the top end. As mentioned before, these two measures of language are highly correlated and therefore it is difficult to separately identify their effects.

[21] Note that although the number of observations in each subsample is the same, those corresponding to the top quantiles have a larger proportion of zeros, which reduces the precision of the estimates.

Table 3: Baseline results by TM value decile distribution

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	0-10th	10-20th	20-30th	30-40th	40-50th	50-60th	60-70th	70-80th	80-90th	90-100th
Distance: internal	-0.495***	-0.496***	-0.415***	-0.363***	-0.305***	-0.277***	-0.227***	-0.156***	-0.109***	-0.034**
	(0.016)	(0.014)	(0.015)	(0.012)	(0.013)	(0.010)	(0.011)	(0.011)	(0.011)	(0.016)
Distance: external	-0.889***	-0.920***	-0.878***	-0.696***	-0.751***	-0.696***	-0.624***	-0.567***	-0.445***	-0.384***
	(0.055)	(0.063)	(0.043)	(0.029)	(0.037)	(0.028)	(0.025)	(0.023)	(0.022)	(0.027)
Clubs' skill gap	0.080***	0.056**	0.002	-0.077***	-0.055**	-0.067***	-0.124***	-0.122***	-0.063***	-0.232***
	(0.021)	(0.022)	(0.024)	(0.018)	(0.025)	(0.019)	(0.017)	(0.021)	(0.019)	(0.017)
Same team	1.920***	1.682***	2.067***	2.027***	2.294***	2.257***	2.375***	2.750***	2.883***	3.507***
	(0.078)	(0.072)	(0.074)	(0.058)	(0.064)	(0.053)	(0.053)	(0.055)	(0.058)	(0.087)
Same country	2.153***	2.093***	1.692***	2.580***	1.732***	1.735***	1.674***	1.336***	1.507***	0.299
	(0.396)	(0.467)	(0.309)	(0.213)	(0.272)	(0.211)	(0.192)	(0.183)	(0.180)	(0.233)
Contiguous countries	0.319***	0.357***	0.540***	0.574***	0.605***	0.533***	0.472***	0.298***	0.192***	0.007
	(0.091)	(0.112)	(0.091)	(0.064)	(0.087)	(0.069)	(0.065)	(0.064)	(0.059)	(0.074)
Common ethnic language	0.554***	0.598***	0.746***	0.554***	0.574***	0.425***	0.370***	0.504***	0.682***	0.453***
	(0.101)	(0.137)	(0.136)	(0.101)	(0.113)	(0.092)	(0.096)	(0.092)	(0.091)	(0.092)
Common official language	0.376***	0.492***	0.307**	0.510***	0.343***	0.546***	0.658***	0.455***	0.375***	0.139
	(0.107)	(0.140)	(0.136)	(0.105)	(0.115)	(0.095)	(0.103)	(0.099)	(0.107)	(0.103)
Observations	23,024,064	21,648,681	12,987,201	28,766,930	12,064,735	22,704,406	21,974,653	15,824,372	12,056,507	4,717,792

Notes: This table presents estimates from our baseline regression, run separately for each decile of the TM value distribution. Standard errors clustered by club of origin and destination in parenthesis. Stars indicate statistical significance at 0.01 (***), 0.05 (**), and 0.1 (*). All regressions include origin-time and destination-time fixed effects.

5 Robustness

To test the reliability of our main findings we conducted two robustness checks whose results are summarised in this section.

5.1 Sample Composition by Club

Our first robustness exercise aims to check whether our results are robust to using a sample with an equal number of observations for each season. This is important as the TM database contains fewer observations for earlier years; as the platform expanded so did its coverage of players and clubs. As such, we restrict the sample to clubs which have been in the sample for all of the past k seasons, where $k = 1, \ldots 15$, and run our regression for each value of k. So, for example, when $k = 1$, our sample consists only of observations for 2019, the last year in our sample. When $k = 2$ the sample consists of observations for 2018 and 2019 of clubs which are present both years, and so on. Naturally, as k increases, the sample includes fewer clubs but more observations for each one.

Table 4 presents the results for selected values of k. The effect of increasing k is comparable to progressing from lower to high value percentiles, i.e., the distance effects exert less influence and the "same team" coefficient gets larger. This can be explained by the fact that clubs that more persistently appear in the data tend to be bigger clubs which employ players in the higher value percentiles. Overall, however, the main results change little when the sample is restricted in this way.

Table 4: Robustness: Baseline results by TM value decile distribution

	(1) $k = 1$	(2) $k = 3$	(3) $k = 5$	(4) $k = 7$	(5) $k = 9$	(6) $k = 11$	(7) $k = 13$	(8) $k = 15$
Distance: internal	-0.358***	-0.352***	-0.350***	-0.348***	-0.295***	-0.280***	-0.210***	-0.194***
	(0.011)	(0.012)	(0.013)	(0.016)	(0.017)	(0.019)	(0.019)	(0.017)
Distance: external	-0.538***	-0.500***	-0.499***	-0.494***	-0.488***	-0.476***	-0.425***	-0.342***
	(0.029)	(0.026)	(0.029)	(0.032)	(0.032)	(0.035)	(0.039)	(0.045)
Clubs' skill gap	-0.545***	-0.511***	-0.473***	-0.451***	-0.440***	-0.414***	-0.367***	-0.340***
	(0.016)	(0.014)	(0.014)	(0.014)	(0.014)	(0.015)	(0.015)	(0.014)
Same team	2.358***	2.433***	2.454***	2.487***	2.624***	2.662***	2.947***	3.055***
	(0.049)	(0.050)	(0.056)	(0.067)	(0.075)	(0.084)	(0.083)	(0.077)
Same country	2.959***	2.980***	2.796***	2.632***	2.353***	2.209***	2.084***	2.395***
	(0.224)	(0.204)	(0.221)	(0.242)	(0.246)	(0.257)	(0.283)	(0.327)
Contiguous countries	0.212***	0.203***	0.212***	0.263***	0.361***	0.336***	0.376***	0.399***
	(0.070)	(0.062)	(0.066)	(0.073)	(0.077)	(0.078)	(0.083)	(0.082)
Common ethnic language	0.745***	0.871***	0.693***	0.633***	0.725***	0.706***	0.518***	0.517***
	(0.109)	(0.084)	(0.082)	(0.084)	(0.085)	(0.093)	(0.110)	(0.123)
Common official language	0.178	0.035	-0.055	-0.227**	-0.225**	-0.162	0.035	-0.011
	(0.116)	(0.091)	(0.089)	(0.096)	(0.096)	(0.106)	(0.126)	(0.143)
Observations	13,461,652	22,561,345	22,820,308	18,894,943	14,686,440	10,882,526	7,468,817	5,294,539

Notes: This table presents results of our baseline specification, run separately when restricting the sample to clubs which have been in the sample for all of the past k seasons. For example, when $k = 3$ the sample consists of observations for 2017, 2018 and 2019, of clubs which were present in all three years. Standard errors clustered by club of origin and destination in parenthesis. Stars indicate statistical significance at 0.01 (***), 0.05 (**), and 0.1 (*). All regressions include origin-time and destination-time fixed effects.

5.2 Age Effects

Football player careers exhibit characteristics that are comparable to most other professions in that productivity at first rises with age and then declines. The difference is that this happens within a much shorter timespan than in most other professions, with few players remaining active beyond the age of 40.[22] In this section we check the robustness of the results presented before to the consideration of player's age.

To do so, we split the sample already partitioned by skill into various different age brackets and re-run our baseline specification for each group. We consider age brackets of three years as well as players under 21 and over 32.

Two things stand out in relation to age. First, as can be seen in Table 5, the size of the distance coefficients tends to be U-shaped. For the youngest and oldest age group the coefficients are highest, while for the those in mid-career, the coefficients are lowest.

Table 5: Robustness: External and internal distance effects by age group and TM value decile distribution

	Distance: External				Distance: Internal			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All values	0-10th	40-50th	90-100th	All values	0-10th	40-50th	90-100th
Under 20	-0.708***	-0.941***	-0.742***	-0.410***	-0.524***	-0.651***	-0.450***	-0.092***
	(0.043)	(0.068)	(0.108)	(0.066)	(0.011)	(0.017)	(0.024)	(0.033)
Ages 21-23	-0.580***	-1.104***	-0.746***	-0.419***	-0.358***	-0.501***	-0.350***	-0.048**
	(0.032)	(0.072)	(0.064)	(0.041)	(0.011)	(0.022)	(0.020)	(0.021)
Ages 24-26	-0.490***	-0.993***	-0.766***	-0.357***	-0.277***	-0.381***	-0.270***	-0.006
	(0.026)	(0.103)	(0.073)	(0.032)	(0.011)	(0.026)	(0.022)	(0.020)
Ages 27-29	-0.415***	-0.945***	-1.089***	-0.375***	-0.255***	-0.410***	-0.289***	-0.004
	(0.024)	(0.137)	(0.095)	(0.029)	(0.011)	(0.045)	(0.030)	(0.023)
Ages 30-32	-0.445***	-1.592***	-1.000***	-0.343***	-0.280***	-0.439***	-0.378***	-0.018
	(0.027)	(0.253)	(0.125)	(0.039)	(0.012)	(0.048)	(0.043)	(0.024)
Ages over 32	-0.560***	-1.619***	-0.766**	-0.302***	-0.403***	-0.517***	-0.509***	-0.052
	(0.045)	(0.569)	(0.314)	(0.079)	(0.015)	(0.049)	(0.063)	(0.050)

This table presents estimates of the external and internal distance variables included in our baseline regression, run separately on slices of the data where we restrict the sample to players in different age brackets (under 20, 21-23, 24-26, 27-29, 30-32, and over 32) and by skill level. Standard errors clustered by club of origin and destination in parenthesis. Stars indicate statistical significance at 0.01 (***), 0.05 (**), and 0.1 (*). All regressions include origin-time and destination-time fixed effects.

[22] One significant difference is that the registration of a player that can be traded; each player must be registered with their national association when hired by a club, and ownership of the registration resides with the club as long as the player is under contract. Thus, clubs have the potential to profit from the movement of players between clubs. There is no other profession where trade of this nature is legally permitted.

Second, the results in Table 6 show that the skill gap coefficient is generally negative and significant across the age groups, but is consistently positive in the youngest age group (under 20), and is positive for some higher value deciles in the 21 to 23 age group. This suggests that younger players whose talent may not yet have been fully recognized are moving to clubs with higher skill levels, while by the age of 24 most players are a "known quantity," and hence unlikely to move to clubs where the average skill level is significantly different. There is no evidence that this is occurring in the oldest age group when player skills are deteriorating due to age—in this case players may prefer to retire rather move to a lower level of competition.

Table 6: Robustness: Skill gap effects by age group and TM value decile distribution

	(1) All values	(2) 0-10th	(3) 40-50th	(4) 90-100th
Under 20	0.119***	0.459***	0.713***	0.295***
	(0.017)	(0.032)	(0.051)	(0.042)
Ages 21-23	-0.371***	0.092***	0.260***	-0.034
	(0.011)	(0.030)	(0.046)	(0.024)
Ages 24-26	-0.848***	-0.109**	-0.235***	-0.379***
	(0.012)	(0.046)	(0.056)	(0.027)
Ages 27-29	-1.029***	-0.404***	-0.484***	-0.492***
	(0.014)	(0.076)	(0.088)	(0.028)
Ages 30-32	-1.063***	-0.330***	-0.822***	-0.405***
	(0.017)	(0.098)	(0.114)	(0.039)
Ages over 32	-1.003***	-0.513***	-1.029***	-0.343***
	(0.024)	(0.129)	(0.171)	(0.067)

This table presents estimates of the skill gap variable included in our baseline regression, run separately on slices of the data where we restrict the sample to players in different age brackets (under 20, 21-23, 24-26, 27-29, 30-32, and over 32) and by skill level. Standard errors clustered by club of origin and destination in parenthesis. Stars indicate statistical significance at 0.01 (***), 0.05 (**), and 0.1 (*). All regressions include origin-time and destination-time fixed effects.

6 Concluding remarks

We study the effect of skill on the determinants of migration in a gravity model. To do so, we use a novel database that documents international and intra-national moves for subjects with different levels of skills. The database is unique in the migration literature in its degree of granularity, which allows us to assess the role of skills in amplifying or mitigating existing mobility barriers. Consistent with previous studies, we find that the various barriers to migration, including measures of distance, tend to hamper mobility. But we find that the magnitude of these barriers varies strongly by skill level, such that the impact of distance, borders or linguistic differences on migration decreases as the level of skill increases. At the very top end of the skill distribution, border frictions almost disappear, and in some cases even reverse, implying that national boundaries, while constituting a substantial barrier to others, appear to pose little obstacle to the migration of superstars.

References

Adserà, A. and M. Pytliková (2015). The role of language in shaping international migration. *Economic Journal 125*(Feature Issue), F49–F81.

Anderson, J. E. and Y. V. Yotov (2022). Estimating gravity from the short to the long run: A simple solution to the 'International Elasticity Puzzle'. NBER Working Papers 30809, National Bureau of Economic Research, Inc.

Artuc, E., F. Docquier, Özden, and C. Parsons (2015). A global assessment of human capital mobility: The role of non-OECD destinations. *World Development 65*(C), 6–26.

Beine, M., S. Bertoli, and J. Fernández-Huertas Moraga (2015). A practitioners' guide to gravity models of international migration. *The World Economy 39*(4), 496–512.

Bertoli, S. and J. Fernández-Huertas Moraga (2013). Multilateral resistance to migration. *Journal of Development Economics 102*, 79–100.

Bertoli, S. and J. Fernández-Huertas Moraga (2015). The size of the cliff at the border. *Regional Science and Urban Economics 51*, 1–6.

Borjas, G. J. (2003). The labor demand curve is downward sloping: Reexamining the impact of immigration on the labor market. *The Quarterly Journal of Economics 118*(4), 1335–1374.

Borjas, G. J. and L. F. Katz (2007). The evolution of the Mexican-born workforce in the united states. In *Mexican immigration to the United States*, pp. 13–56. University of Chicago Press.

Breinlich, H., D. Novy, and J. M. C. Santos Silva (2022). Trade, gravity and aggregation. *The Review of Economics and Statistics*, 1–29.

Bryson, A., G. Rossi, and R. Simmons (2014). The migrant wage premium in professional football: A superstar effect? *Kyklos 67*(1), 12–28.

Buraimo, B., B. Frick, M. Hickfang, and R. Simmons (2015). The economics of long-term contracts in the footballers' labour market. *Scottish Journal of Political Economy 62*(1), 8–24.

Conte, M., P. Cotterlaz, and T. Mayer (2022, July). The CEPII gravity database. Working Papers 2022-05, CEPII research center.

Dustmann, C., T. Frattini, and I. P. Preston (2012). The effect of immigration along the distribution of wages. *The Review of Economic Studies 80*(1), 145–173.

Galichon, A. and B. Salanié (2022, 12). Cupid's invisible hand: Social surplus and identification in matching models. *The Review of Economic Studies 89*(5), 2600–2629.

Gourieroux, C., A. Monfort, and A. Trognon (1984). Pseudo maximum likelihood methods: Applications to Poisson models. *Econometrica 52*(3), 701–720.

Guimarães, P., O. Figueirdo, and D. Woodward (2003). A tractable approach to the firm location decision problem. *The Review of Economics and Statistics 85*(1), 201–204.

Herm, S., H.-M. Callsen-Bracker, and H. Kreis (2014). When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community. *Sport Management Review 17*(4), 484–492.

Kahn, L. M. (2000). The sports business as a labor market laboratory. *Journal of economic perspectives 14*(3), 75–94.

Kleven, H. J., C. Landais, and E. Saez (2013). Taxation and international migration of superstars: Evidence from the European football market. *American Economic Review 103*(5), 1892–1924.

Lucifora, C. and R. Simmons (2003). Superstar effects in sport: Evidence from Italian soccer. *Journal of Sports Economics 4*(1), 35–55.

Mayda, A. M. (2010). International migration: A panel data analysis of the determinants of bilateral flows. *Journal of Population Economics 23*, 1249–1274.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Fontiers in Econometrics*, pp. 105–142. New York: Academic press.

Müller, O., A. Simons, and M. Weinmann (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. *European Journal of Operational Research 263*(2), 611–624.

Ottaviano, G. I. P. and G. Peri (2012). Rethinking the effect of immigration on wages. *Journal of the European Economic Association 10*(1), 152–197.

Palacios-Huerta, I. (2023). The beautiful dataset. Technical report. Available at SSRN: https://ssrn.com/abstract=4665889 or http://dx.doi.org/10.2139/ssrn.4665889.

Peeters, T. (2018). Testing the wisdom of crowds in the field: Transfermarkt valuations and international soccer results. *International Journal of Forecasting 34*(1), 17–29.

Prockl, F. and B. Frick (2018). Information precision in online communities: Player valuations on www.transfermarkt.de. *International Journal of Sport Finance 13*, 319–335.

Reade, J. and P. G. del Barrio (2023). A forecasting test for the reliability of salary data. Available at SSRN: https://ssrn.com/abstract=4399518 or http://dx.doi.org/10.2139/ssrn.4399518.

Rosen, S. (1981). The economics of superstars. *The American Economic Review 71*(5), 845–858.

Santos Silva, J. M. C. and S. Tenreyro (2006). The log of gravity. *Review of Economic Statistics 88*(4), 641–658.

Szymanski, S. (2022). Are footballers rewarded for luck? A surprise test. Available at SSRN: https://ssrn.com/abstract=4262288 or http://dx.doi.org/10.2139/ssrn.4262288.

Yotov, Y. V. (2012). A simple solution to the distance puzzle in international trade. *Economics Letters 117*(3), 794–798.

Yotov, Y. V. (2022). On the role of domestic trade flows for estimating the gravity model of trade. *Contemporary Economic Policy 40*(3), 526–540.