# Predicting Spanish Emigration and Immigration

**Jesús Fernández-Huertas Moraga and Gonzalo López Molina**

*AIReF Working Paper Series*

*WP/2018/*

The mission of AIReF, the Independent Authority for Fiscal Responsibility, is to ensure strict compliance with the principles of budgetary stability and financial sustainability contained in article 135 of the Spanish Constitution.

AIReF: José Abascal, 2, 2nd floor. 28003 Madrid. Tel. +34 910 100 599

E-mail: Info@airef.es

Website: www.airef.es

# Predicting Spanish Emigration and Immigration

Jesús Fernández-Huertas Moraga[a] and Gonzalo López Molina[b]

[a] *Universidad Carlos III de Madrid*[*]

[b] *Universidad Complutense de Madrid*[†] *and AIReF*

October 4, 2018

**Abstract**

What is the future of international migration flows? The growing availability of bilateral international migration data has resulted in an improved understanding of the determinants of migration flows through the estimation of theory-based gravity models. However, the use of these models as a prediction tool has remained a mostly unexplored research area. This paper estimates simple gravity models of bilateral migration flows for the whole world and projects these models into the future. We focus on a particularly hard-to-predict country, Spain, which evolved from an emigrant-sending to an immigrant-receiving country, and compare the estimates from internationally available data to those coming from local sources. Our results show that the projections of migration flows depend more heavily on the span of the baseline dataset used to estimate the model than on the particular functional form and variables chosen for the prediction although we confirm the relevance of origin-country demographic factors.

**Keywords**: international migration; prediction; gravity model.

**JEL classification codes**: F22, J11, J61, O15.

[*] Madrid, 126, 28903, Getafe (Madrid), Spain; email: jesferna@eco.uc3m.es (corresponding author).

[†] Campus de Somosaguas, 28223 Pozuelo de Alarcón (Madrid), Spain; email: gonzalolopezmolina@ucm.es.

# 1 Introduction

The immigration flows into and emigration flows out of a given country depend on the evolution of economic and political conditions both in that particular country and in all countries in the world. The reason is that the rest of the world is a potential destination for the inhabitants of any given country.

Despite being aware of this reality, those performing immigration projections tend to use notably rough assumptions. National and international institutions regularly update population projections, needed to predict the future of macroeconomic variables and different policy scenarios. These projections are based on different assumptions on the path of fertility, mortality and, also, immigration and emigration.

In the case of Spain, migration flows accounted for 34 per cent of the Spanish population growth between 1960 and 2015. In more recent years, from 1998 to 2015, this figure went up to 85 per cent.[1] Hence, if the past is any indication, any model of the evolution of the population of Spain needs to carefully examine the role of immigration. However, Spain provides a clear example of the roughness of the assumptions on immigration scenarios that several institutions have established. The United Nations (2015) predict net migration to Spain to remain stable at what they call "current" levels until 2050. These current levels refer to the level of net migration in Spain in the year 2010, which was a net inflow of 100,000 immigrants. Later, United Nations (2015) foresees a slow reduction of this net intake of 50 per cent by the year 2100. EUROSTAT (Lanzieri, 2017) makes a more sophisticated prediction. They estimate an ARIMA model for net migration to Spain. The result is basically a projection of the 1996-2015 trend until 2050. Then, Lanzieri (2017) goes on to assume that all similarly calculated trends for European countries will converge towards common immigration and emigration rates from 2050 to 2100. Finally, the official Spanish statistics office, INE (2016), fixes gross immigration flows to Spain to the last available level, the one for 2015: 343,614 immigrants. They then calculate average emigration rates by province, age and gender from 2011-2015 and use them to project gross emigration into the future.

The contribution of this paper is to use all of the available data and recent methodological innovations to predict the evolution of emigration out of Spain and immigration into Spain in the XXIst century, thus escaping these rough assumptions. On the data front, we contribute

[1] Own calculations on data from World Bank (2017) and INE (2017b).

on two aspects:

- Longer time series. We will be using 1960-2000 decennial data from the World Bank (Özden, Parsons, Schiff, and Walmsley, 2011) combined with United Nations data (United Nations, 2017) for 2010 and 2017 to measure immigration and emigration flows.[2] We will also use 1988-2016 yearly data from the Spanish *Estadística de Variaciones Residenciales* (INE, 2017a) to get a more precise image of the distribution of Spanish migration flows by age and gender.

- Wider cross-sectional data. We will be using dyadic datasets with information on origins and destinations. This means 231 × 231 dyads in the World Bank dataset and more than 200 origin countries in the Spanish data. We will treat both immigration and emigration at the same level.

The main advantage from using these datasets is that they will allow us to take advantage from newly developed dyadic estimation strategies (Beine, Bertoli, and Fernández-Huertas Moraga, 2016). We will make explicit the part of the history the practitioner can choose to forecast migration flows and will add explanatory variables of three types to the model. First, we will consider the demographic structure of both origin and destination countries. Many authors have emphasized the role of demography in shaping international migration flows.[3] Young countries have larger migration propensities because migration is an investment decision (Sjaastad, 1962) whose rewards are typically easier to reap when the migratory move takes place at younger ages. As far as destination countries are concerned, the demographic structure may imply that the labor market competition is stronger if there are larger young cohorts at destination and less fierce otherwise (Hanson and McIntosh, 2016).

Secondly, economic conditions both at origin and at destination will be included in the model, allowing for different elasticities, as a microfounded gravity model will typically imply that the elasticity of migration flows with respect to economic conditions at destination should be larger than with respect to economic conditions at origin. The reason for this is that destinations are more interchangeable among themselves than origins from the point of view of the potential migrant (Beine, Bertoli, and Fernández-Huertas Moraga, 2016).

[2] Throughout the paper, we will treat the 2017 observations as if they corresponded to 2020 to keep the decennial nature of the data.

[3] Recent examples are Hanson and McIntosh (2010) and Hanson and McIntosh (2012).

Finally, a third element that will be added is the role of networks, also called diasporas (Beine, Docquier, and Özden, 2011). The stocks of immigrants from the same origin present at a destination facilitate migration movements both due to the possibility of reducing migration costs (Mckenzie and Rapoport, 2007) and the possibility of increasing the earnings potential of immigrants at destination (Munshi, 2003).

The main results of the paper can be summarized as follows. When using these large datasets, dyadic fixed effects absorb most of the explanatory power of the models, leaving little room for demography, networks or economic activity to play a significant role on migration flows. These dyadic fixed effects absorb time-invariant variables such as physical distance, linguistic distance, cultural distance, contiguity, common colonial history, etc. but they also encompass average effects of the other variables. Unfortunately, in the absence of an appropriate instrument, we cannot disentangle these effects. We can only say that there is not enough variation in the data to identify a meaningful role for demography, networks or economic activity. However, in the long run, even small variations in the trajectories of these variables compound over time through their feedback in population, leading to a huge variety of estimates of migration flows to Spain and in the rest of the world beyond the year 2050. Specifically, the demographic structure of the countries of origin of the emigrants explain most of the variation in the projections of migration flows into the future.

Many of the main determinants of international migration flows are not easy to project into the future (Beine, Bertoli, and Fernández-Huertas Moraga, 2016). The main example would be policies, which have been shown to be quantitatively very relevant in explaining international migration in general (Bertoli and Fernández-Huertas Moraga, 2015) and more particularly in the case of Spain (Bertoli and Fernández-Huertas Moraga, 2013). The visa policy of the European Union regarding Turkey might become the most influential factor explaining the arrival of immigrants into Europe within the next 50 years. This is clearly a limitation of this paper and of other existing projections but it is still useful to understand the forces over whose future evolution we do have some information, such as demography. At least in the medium run, we can be reasonably certain about the demographic pressures that different countries will experience.

Beyond the omission of potentially relevant variables, a fundamental assumption of the model will be the null effect of immigration and emigration on the standard of living both in Spain and in emigrant-sending countries. Despite the enduring controversies about the labor market effects of emigration,[4] a common characteristic of most of the studies is that these effects, be they slightly positive or slightly negative, tend to cluster around zero (Docquier, Özden, and Peri, 2014), so that this is not as outlandish an assumption as it may appear at first sight. Fundamentally for our purposes, this assumption allows us to treat economic fundamentals as an exogenous variable in the model. It would even allow us to close the model if we did not need the future demographic predictions to generate our forecasts of immigration and emigration flows. When we use Spanish data only, we will be using these demographic predictions as exogenous variables as well. The implied assumption there will be that Spain is small enough in the world so as not to significantly affect the population stocks of any given country. Only when we use the World Bank-UN data we will actually be able to close the model and predict immigration and emigration for every country in the world, thus updating the population estimates from United Nations (2015).

We are of course not the first ones that have tried to predict immigration and emigration flows in the academic literature, setting aside the examples in international institutions mentioned above. Hanson and McIntosh (2016) is a recent effort in this sense that is very closely related to our work. However, there are two key differences in their approach. First, they estimate their model only on 2000-2010 data on net migration to the OECD. This is particularly problematic for the case of Spain since they are projecting that the Spanish immigration boom, by which Spain experienced the largest and fastest immigration growth in the OECD, will repeat every decade into the future. Second, their estimation assumptions are much more restrictive than ours. Specifically, we take the critique by Bertoli (2017) seriously and do not assume that the migration rate of a given cohort-gender cell only depends on the size of these cells at origin and at destination. The rationale for this was that there is economic competition within cells but not across cells, as in Borjas (2003). We actually allow for cross-elasticities across cohort-cells and estimate these cross-elasticities although there is not enough variation in the data to actually identify them separately. Also, differently from Hanson and McIntosh (2016), we do not assume that origin variables have the same elasticity as destination variables. We will estimate distinct coefficients for origin and destination variables. For example, this gives us the possibility of considering non-linear origin income effects on emigration (Clemens, 2014). Campos (2017) follows the same approach as Hanson and McIntosh (2016) but he extends the dataset to a longer period,

[4] Borjas (2003) and Ottaviano and Peri (2012) could be considered the more classic references.

using the same data from Özden, Parsons, Schiff, and Walmsley (2011) that we use in one part of our analysis at the cost of not being able to differentiate by cohorts and gender.

An even more recent and more comprehensive effort than the one undertaken here is the one by Dao, Docquier, Maurel, and Schaus (2017). They create and calibrate a general equilibrium model to forecast immigration and emigration until 2100. In their model they distinguish between two different types of workers (college educated and less educated), which we are not able to do. They can do it because they only use 2010 stocks from DIOC-E to parameterize their model. They use the parameterized model to make projections both into the future and into the past.

The paper proceeds as follows. Section 2 explains the basic random utility maximization model and how its aggregation results in a classical gravity equation. We next briefly develop the characteristics of the datasets we employ in section 3. In section 4, we describe our methodology to show the correspondence between the data and the model. Next, in section 5, we present our results. We offer some concluding remarks in the last section of the paper, section 6.

## 2 The Model

In economics, migration decisions are typically modeled as the result of a choice of destination by utility-maximizing individuals (Beine, Bertoli, and Fernández-Huertas Moraga, 2016). Given that the choice of destination is a discrete one, the workhorse model for migration studies is the random utility maximization model, developed by McFadden (1974). Once the individual decisions from the random utility maximization model are aggregated, we end up with a classical gravity equation. Next, we briefly develop how to derive the gravity framework.

We start by letting $pop_{ot}$ represent the stock of the population residing in country $o$ at time $t$. We can then write the scale $m_{odt}$ of the migration flow from country $o$ to country $d$ at time $t$ as:

$$m_{odt} = p_{odt} pop_{ot-1} \tag{1}$$

The term $p_{odt}$ is the probability that an individual from country $o$ migrates to country $d$ at time $t$, also known as the emigration rate.

If we specialize the model to the Spanish case, net migration to Spain ($nm_{SPt}$) in a given year can be obtained as:

$$nm_{SPt} = \sum_{o \neq SP} m_{oSPt} - \sum_{d \neq SP} m_{SPdt} \tag{2}$$

Hence, net migration flows to Spain are composed by as many gross flows as twice the number of countries in the world.

The random utility maximization model is used to estimate these emigration rates by obtaining the expected value of $p_{odt}$.

The utility that one individual $i$ who was located in country $o$ at time $t-1$ derives from opting for country $d$ at time $t$ is:

$$U_{iodt} = \beta' x_{odt} + \epsilon_{iodt} \tag{3}$$

The vector $x_{odt}$ includes all deterministic components of utility while $\epsilon_{iodt}$ is an individual-specific stochastic component.

The distributional assumptions on $\epsilon_{iodt}$ determine the expected probability $E(p_{odt})$ that opting for country $d$ represents the utility-maximizing choice (McFadden, 1974). The use of very general distributional assumptions for $\epsilon_{iodt}$ leads to this type of expression:

$$\ln\left(\frac{p_{odt}}{p_{oot}}\right) = \frac{1}{\tau}\beta' x_{odt} - \beta' x_{oot} + MRM_{odt} \tag{4}$$

$MRM_{odt}$ is the multilateral resistance to migration term (Bertoli and Fernández-Huertas Moraga, 2013). It reflects the effect of alternative destinations on bilateral migration rates. The parameter $\tau$ is the dissimilarity parameter, which is related to the inverse of the correlation in $\epsilon_{iodt}$ across alternative destinations. When the independence of irrelevant alternatives holds, then $\tau = 1$ and $MRM_{odt} = 0$. In that case, $\epsilon_{iodt}$ is modeled as following an iid extreme value type I distribution.

From the above theoretical expression, what we actually take to the data is the following:

$$\ln\left(\frac{m_{odt}}{m_{oot}}\right) = \frac{1}{\tau}\beta' x_{odt} - \beta' x_{oot} + MRM_{odt} + \xi_{odt} \tag{5}$$

The vector $x_{odt}$ contains the independent variables in the model. As mentioned in the introduction, in our case these are:

- Demographic structure. We will include the log of the size of different cohorts. Subscripts $o$ and $d$ can be trivially extended to include country-cohort-gender groups in some specifications although these have not been finally included in the paper.
- Economic conditions. We will proxy them by the log of the GDP per capita.
- Networks. This will be the stock of co-nationals from country $o$ already residing in the country of destination $d$.

If our interest laid on the effect of each of these variables on migration flows, we would need to concern ourselves with the potential endogeneity of these variables. However, since our interest is to predict, this is not such a pressing issue. Still, Hanson and McIntosh (2016) argue that it is reasonable to assume that the demographic structures of the countries are exogenous to current immigration flows at least in a generation's horizon. Also, we will assume that the effect of immigration and emigration on economic activity is close enough to zero as to consider it negligible, justified by the fact that many estimates of the effects of immigration and emigration actually cluster around zero.[5] We would only need to worry about the endogeneity of the network variable, which has often been instrumented by past settlements or by proxies for these past settlements (Beine, Docquier, and Özden, 2011).

# 3 Data

This paper uses two different datasets to build the dependent variable in equation (5). On the one hand, we will have the data collected by Özden, Parsons, Schiff, and Walmsley (2011) from censuses around the world between 1960 and 2000, complemented with similar data for 2010 and 2017 from United Nations (2017). On the other hand, we will also build estimates based on data from INE (2017a). Their main characteristics as well as their advantages and disadvantages are detailed below.

## 3.1 World Bank data

Özden, Parsons, Schiff, and Walmsley (2011) collected the number of foreign-born individuals from each country in the world residing in every other country every ten years. Most countries

[5] Docquier, Özden, and Peri (2014) is a good example in this respect. Most effects of immigration and emigration on wages are between -1 and 1 per cent over a decade.

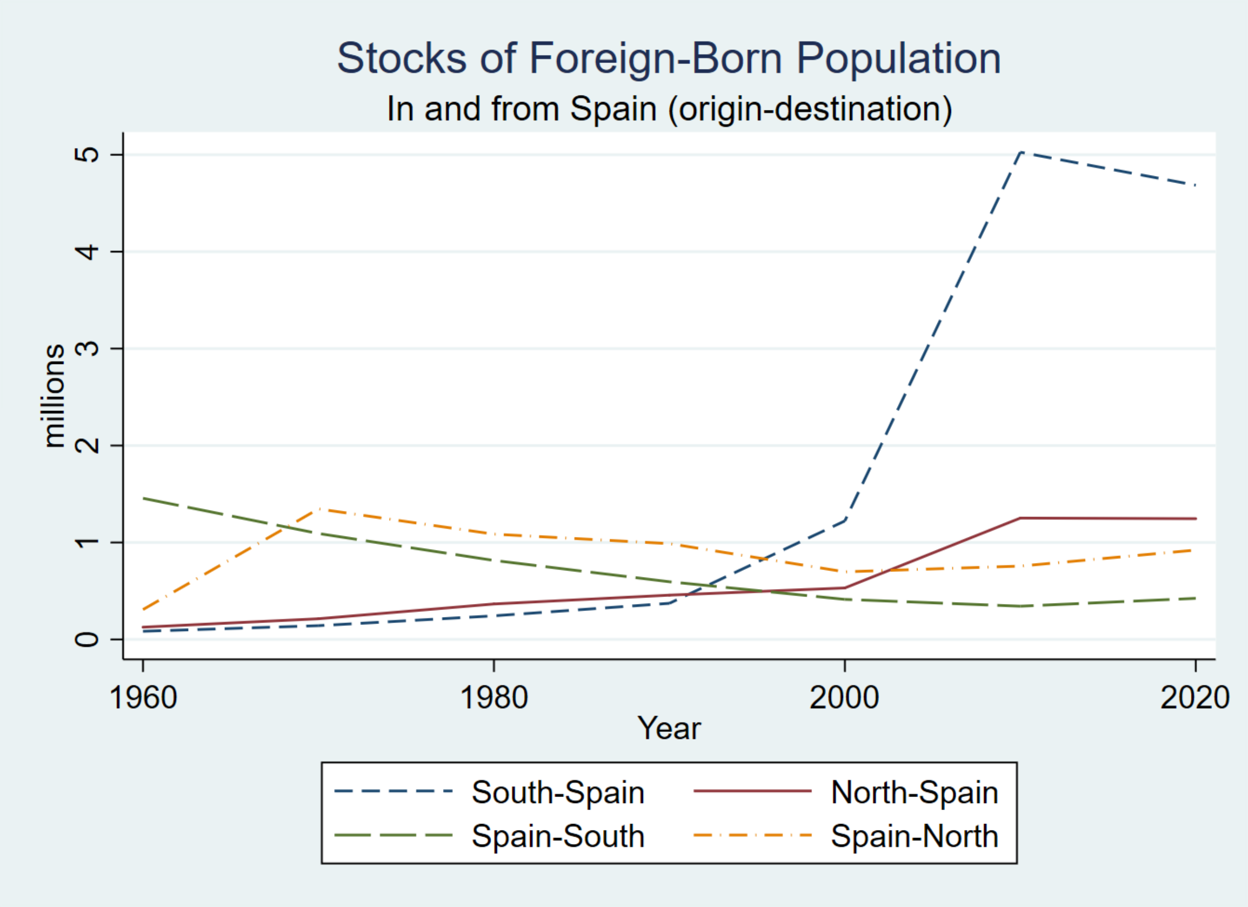
undertake decennial census counts of their populations and these are the ones that Özden, Parsons, Schiff, and Walmsley (2011) exploit. Since not all countries have censuses every ten years, they have interpolated some of their data points, as described in their paper. Overall, they present a matrix of 231 × 231 entries for five different periods: 1960, 1970, 1980, 1990 and 2000. The census years in each country do not necessarily coincide but they have been grouped. For example, the 2000 census for Spain was in fact conducted in 2001. We complement the data from Özden, Parsons, Schiff, and Walmsley (2011) with data from United Nations (2017) for the year 2010 and 2017 to get a more recent picture of migration flows over the 21st century.

The main advantage of this combined World Bank-UN dataset is how comprehensive it is. It spans the whole world for 67 years. This is the largest dataset available in terms of time series and cross-sectionally. The cost is the lack of details about the composition of the flows. We cannot distinguish by age. Furthermore, the random utility maximization model described above implies that $m_{odt}$ should be constructed with data on gross flows. However, the data will only provide us with net flows that we will have to use as a proxy for gross flows, as in Bertoli and Fernández-Huertas Moraga (2015), for example.

Figure 1 shows the evolution of the migration stocks related to Spain in the dataset. The figure reflects four series: stocks of Spanish emigrants in the North, stocks of Spanish emigrants in the South, stocks of immigrants from the North in Spain and stocks of immigrants from the South in Spain. Following Özden, Parsons, Schiff, and Walmsley (2011), the North is defined as Western Europe plus the US, Canada, Japan, Australia and New Zealand, while the South is the rest of the world.

There are two fundamental observations that stand out from looking at figure 1. The first one is how Spain changed from being an emigration country between 1960 and 1990 to an immigration country between 2000 and 2010 and an emigration country again between 2010 and 2017. In 1960, almost 2 million Spanish individuals lived in the South, mostly Latin America. In 1970, the preferred destination for Spanish emigrants became the North, specifically Western Europe. This changed in 2000 and 2010, when Spain became the host of millions of immigrants from the South, and went back in the following years, when Spain became an emigration country. The second observation that stands out is the huge magnitude of the Spanish immigration boom in 2000 and 2010. By 2010, more than 5 million immigrants from the South and one additional million from the North had chosen Spain

Figure 1: Bilateral migration stocks to and from Spain (1960-2020)

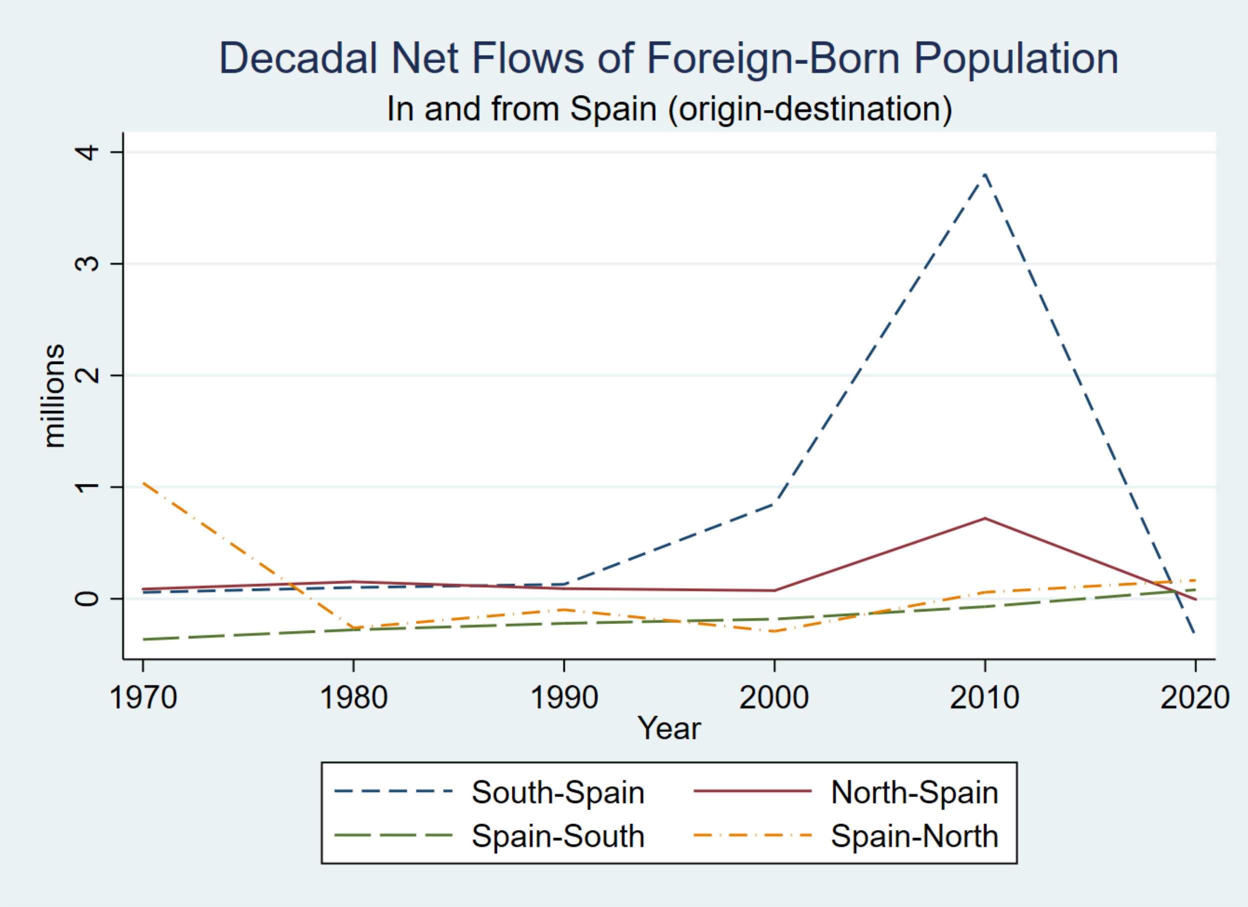


Source: own elaboration on data from Özden, Parsons, Schiff, and Walmsley (2011) and United Nations (2017). 2017 data represented as if corresponding to 2020.

as their destination. The main countries of origin in this stock from the South were Latin American countries (Ecuador, Colombia, etc.), Eastern Europeans (mainly Romania) and Northern Africa (Morocco).

Figure 2 shows the same information as figure 1 but in terms of net flows rather than stocks. For example, the 4 million mark for the South-Spain migration corridor comes from substracting the 2010 data point from the 2000 one in figure 1. It thus means that immigrants from Southern countries in Spain increased by 4 million between 2000 and 2010. These net flows will be used as a proxy for gross flows when estimating equation (5).

Figure 2: Net migration flows to and from Spain (1970-2020)



Source: own elaboration on data from Özden, Parsons, Schiff, and Walmsley (2011) and United Nations (2017). 2017 data represented as if corresponding to 2020.

### 3.2 INE data

The second data source that we will use will come from the Spanish statistics office. INE (2017a) contains every change recorded in the Spanish population registry between 1988 and 2016. The inscriptions and cancellations into this Spanish population registry can be exploited to build series of gross migration flows from every country in the world into Spain, as in Bertoli and Fernández-Huertas Moraga (2013), and from Spain into every country in the world yearly. In addition, the registry provides the microdata so that it is easy to divide the data by age and gender groups. As shown by Bertoli and Fernández-Huertas Moraga (2013), the coverage of immigration since 2000 is particularly good, even in the case of undocumented immigrants, as they had strong incentives to register in order to enjoy health and education services from the municipalities where they resided.

Unfortunately, the data from INE (2017a) presents also some drawbacks. The most notable one is its deficient coverage of emigration. There are no incentives for cancellations. Most cancellations correspond to foreigners who must renew their inscription every two years and these are only recorded since 2002. As a result, the emigration of Spanish nationals, naturalized immigrants and European Union citizens is severely underestimated (González-Ferrer, 2013).

With these caveats in mind, figure 3 presents the evolution of migration flows within the dataset in INE (2017a). It offers a great description of the development of the Spanish immigration boom until its peak in 2007 and its aftermath. After 2007, inflows go down every year until 2013 while outflows go up after they started being recorded in 2002 and they stabilize around 2010.

### 3.3 Additional data

The data for our independent variables comes from two additional sources. For the demographic structure, we consider the population by age and gender for each country from United Nations (2015). For the economic conditions, we take the series on the expenditure-side real GDP at chained PPPs to generate GDP per capita from the Penn World Tables (Feenstra, Inklaar, and Timmer, 2015).

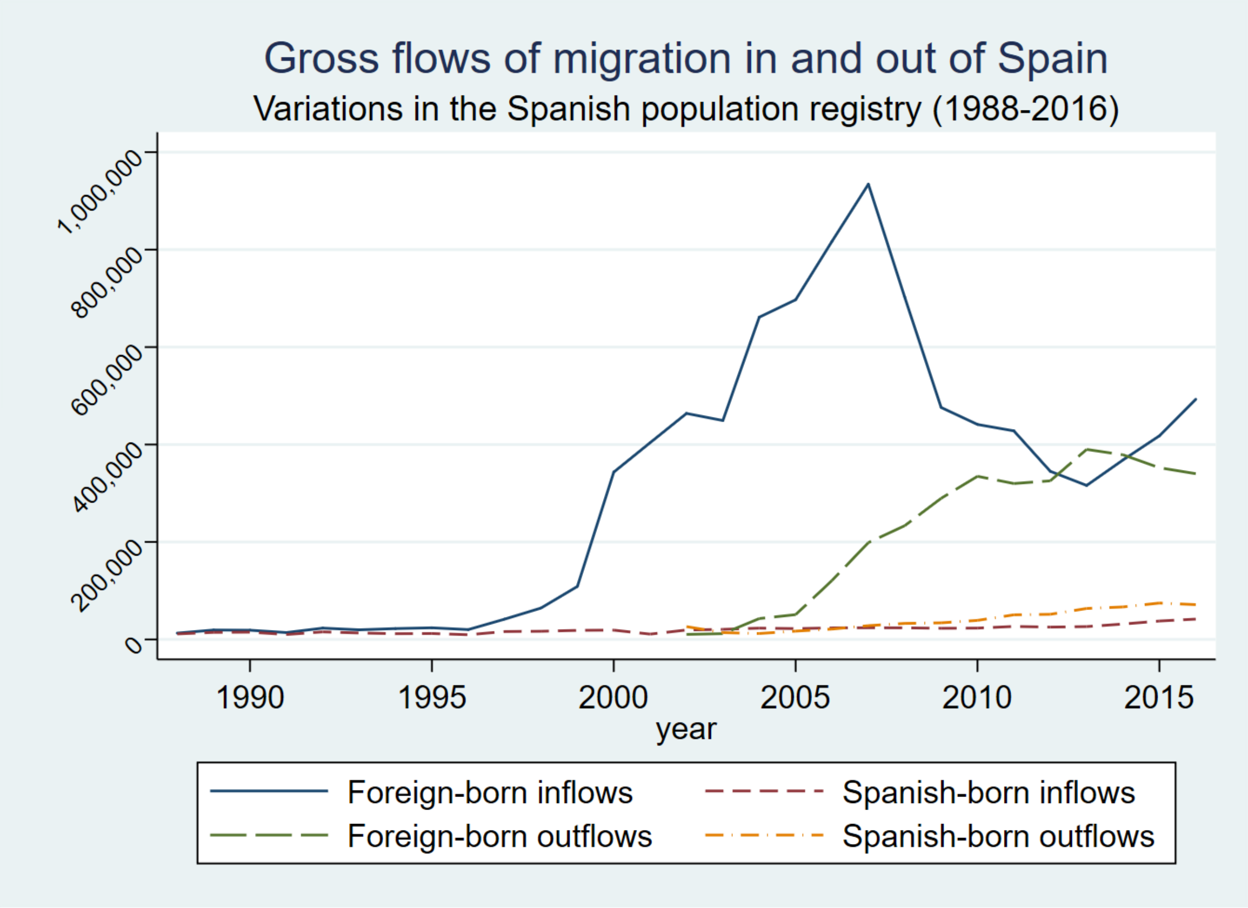
## 4 Methodology

This section describes the methodology that we are using for the projection of net immigration flows to Spain between 2017 and 2100, that is, how we actually treat the datasets described above so as to estimate equation (5).

### 4.1 World Bank-UN data

The World Bank-UN data tell us the number of migrants born in an origin country that were residing in a particular destination between 1960 and 2017. These are decennial data until 2010. To keep the exercise as simple and transparent as possible, we simply assume

Figure 3: Gross migration flows to and from Spain (1988-2016)



Source: own elaboration on data from INE (2017a).

that 2017 data correspond to 2020. Notice that this is what Özden, Parsons, Schiff, and Walmsley (2011) would have done if they had had data from 2017 censuses. We are interested in estimating decennial emigration rates that we can then use to project net emigration to Spain based on the population projections from United Nations (2015).

The basic expression that we will use is:

$$nm_{odt} \equiv M_{odt} - M_{odt-10} = p_{odt} pop_{ot-10} \tag{6}$$

Here $nm_{odt}$ is the net emigration between origin $o$ and destination $d$ between year $t - 10$ and year $t$. $M_{odt}$ denotes the total number of migrants born in country $o$ who were residing in country $d$ in census year $t$. As before, $p_{odt}$ refers to the emigration rate and $pop_{ot-10}$ is the total population of country $o$ in the previous decade.

We can obtain the emigration rate as:

$$p_{odt} = \frac{M_{odt} - M_{odt-10}}{pop_{ot-10}} \tag{7}$$

### 4.1.1 Immigration to Spain

In order to compute immigration rates to Spain, we fix the destination so that $d = SP$. We use four different $p_{oSPt}$ calculations based on the estimation of equation (5) for the immigration rate to Spain from every origin between 1970 and 2020, between 1980 and 2020, between 1990 and 2020 and finally between 2000 and 2020. We then apply these four immigration rates to the eight variants of population projections provided by the United Nations. Hence, we have 32 different projections, denoted by $p$. For each of these projections, we have fitted values from equation (5):

$$\widehat{y_{oSPt}} \equiv \ln \widehat{\left(\frac{m_{oSPt}}{m_{oot}}\right)} = \left(\frac{\widehat{\beta}}{\tau}\right)' x_{oSPt} - \widehat{\beta}' x_{oot} \tag{8}$$

We need to recall that we use $nm_{oSPt}$ as a proxy for $m_{oSPt}$ since gross migration flows are not available with the World Bank data. The estimated gross migration flow into Spain for each of these 32 projections is calculated as:

$$\widehat{m_{oSPt}} = \frac{\widehat{y_{oSPt}}}{1 + \sum_d \widehat{y_{odt}}} \widehat{pop_{ot-10}} \tag{9}$$

We explain below different assumptions under which we can recover estimates of the relevant population for the projection period: $\widehat{pop_{ot-10}}$.

### 4.1.2 Emigration out of Spain

In this case, we fix Spain as an origin ($o = SP$) and similarly calculate 32 different values of estimated gross emigration flows out of Spain to all destinations in the world: $\widehat{m_{SPdt}}$. These projections are calculated in the same way as above.

### 4.1.3 Population adjustment

In purity, the migration data we are using refer to individuals born in each of the origins while the population data include both the origin-born population and the migrants from the rest of the world residing in the country in that particular year. We should not expect this to create great biases in our projections since most of the variation will come from the rates rather than from the population stocks but, still, it is interesting to make sure that this is the case.

To this end, we calculate native populations for each of the origins of Spanish immigrants as well as adjusting the population of Spain by subtracting the immigrants residing in the country when we project emigration out of Spain. The native populations are calculated as:

$$pop_{ot}^{N} = pop_{ot} - \sum_{r \neq o} M_{rot} \tag{10}$$

We can make a second population adjustment by actually adding the total number of individuals born in a particular origin that are emigrants somewhere else. This would also make sense because these emigrants (say Bolivians living in Chile) could also show up in the following census as natives living in a different destination country. The fully adjusted native population would be:

$$pop_{ot}^{F} = pop_{ot} - \sum_{r \neq o} M_{rot} + \sum_{d} M_{odt} \tag{11}$$

In the calculation of the estimated migration flows, we just need to substitute the population for the adjusted population data:

$$\widehat{m_{odt}} = \frac{\widehat{y_{odt}}}{1 + \sum_{e} \widehat{y_{oet}}} \widehat{pop_{ot-10}^{F}} \tag{12}$$

The adjusted population shows as estimated because it is also a projection into the future. The different variants of the population projections from United Nations (2015) include their own migration scenarios. In order to be completely consistent, we subtract their migration data and replace it with the migration data generated with our own projections.[6] Hence, even though we are only running a prediction exercise for immigration to Spain and emigration out of Spain, we do need to estimate the whole matrix of bilateral gross migration flows

[6] Ideally, we would like to subtract United Nations stocks of immigrants and add stocks of emigrants but we can only subtract net migrants from the past 10 years. To avoid double counting, we substitute net migration flows from the United Nations estimates with our own net migration flows rather than net migration stocks.

and project it into the future. This is something that we cannot do with the data from INE (2017a), since we do not have the same type of data for the rest of the countries in the world.

## 4.2 INE data

The difference between the INE data and the World Bank data is that the former refers to gross flows whereas the latter deals with net flows. In order to understand how comparable both methodologies are for the case of Spain, we will first generate immigration and emigration rates based on net flows and then use all of the available INE information to work on gross flows.

### 4.2.1 Net flows

We need to redefine the migration variable in terms of the definitions that we have from the INE. For immigration, we would have:

$$m_{oSPt} = INF_{ot} - OUT_{ot} \qquad \forall o \neq SP \tag{13}$$

Now $o$ refers to country of birth in the INE data. $INF_{ot}$ are inflows from country of birth $o$ in year $t$ and $OUT_{ot}$ are outflows from country of birth $o$ in year $t$.

For the emigration of Spanish-born individuals, we would use:

$$m_{SPdt} = OUT^{SP}_{dt} - INF^{SP}_{dt} \tag{14}$$

Here $OUT^{SP}_{dt}$ refers to the emigration of Spanish-born individuals going to destination $d$ in year $t$ and $INF^{SP}_{dt}$ refers to the return migration of Spanish-born individuals coming from country $d$.

We can then check the correlation between the series $m_{SPdt}$ and $m_{oSPt}$ from both datasets. The INE data span the period 1988-2016 so we can make this comparison for the 1990-2000, the 2000-2010 decade and the 2000-2016 years by summing up all the INE flows during the corresponding period. The overall correlation between the two series is very high for immigration into Spain: 0.98. Unfortunately, the series are not comparable for emigration out of Spain: the correlation is barely 0.47 in this case. This is likely to be due to the problems of the INE data with respect to emigration that were pointed out above. As a result, the emigration predictions based on INE data should be taken with caution as many emigration flows are likely to be severely underestimated (González-Ferrer, 2013).

### 4.2.2 Gross flows

Next, we can take advantage of all the data possibilities from the INE for our exercise. First, we denote by $g$ the subindex for different age-gender groups. We work with six groups: 0-15 year-old men, 0-15 year-old women, 15-65 year-old men, 15-65 year-old women and 65+ year-old men and women. It is straightforward to further disaggregate into smaller age groups but it has the cost of creating many zero-value cells.

We will have up to four different models for emigration rates:

1. *Inflows of foreign-born individuals into Spain.* The emigration rate is calculated as:

$$\frac{m_{goSPt}}{m_{goot}} = \frac{INF_{got}}{pop_{got}} \qquad \forall o \neq SP \tag{15}$$

   $INF_{got}$ are the inflows in year $t$ of individuals from group $g$ born in country $o$. The population of country $o$ in year $t$ from group $g$ ($pop_{got}$) is used to proxy for the non-migrating individuals during the year. The appropriate weight for this regression is $pop_{got-1}$.

2. *Inflows of Spanish-born individuals into Spain.* We need to introduce a new subindex to denote the country where Spanish-born individuals were residing: $r$.

$$\frac{m^{SP}_{grSPt}}{m^{SP}_{grrt}} = \frac{INF^{SP}_{grt}}{pop^{SP}_{grt+1}} \qquad \forall r \neq SP \tag{16}$$

   $INF^{SP}_{grt}$ are the inflows of Spanish-born individuals from group $g$ coming from country of last residence $r$ during year $t$. For the denominator, we can use $pop^{SP}_{grt+1}$, the population of Spanish-born individuals living in country $r$ at the beginning of year $t+1$. These data are available only since 2009 in INE (2017c), which limits severely the number of available observations. The appropriate weight for this regression is $pop^{SP}_{grt}$.

3. *Outflows of foreign-born individuals out of Spain.* The emigration rate is calculated as:

$$\frac{m_{gSPdt}}{m_{gdSPt}} = \frac{OUT_{gdt}}{pop_{gdSPt+1}} \qquad \forall d \neq SP \tag{17}$$

$OUT_{gdt}$ are the outflows in year $t$ of individuals from group $g$ born in country $d$. The denominator $pop_{gdSPt+1}$ is the population born in country $d$ from group $g$ that still lived in Spain at the beginning of year $t+1$. The appropriate weight for this regression is $pop_{gdSPt}$. The population data are available in INE (2017b) since 1996 although the outflows only began to be counted in 2002.

4. *Outflows of Spanish-born individuals out of Spain.* The emigration rate is calculated as:

$$\frac{m_{gSPdt}^{SP}}{m_{gSPt}^{SP}} = \frac{OUT_{gdt}^{SP}}{pop_{gSPt+1}^{SP}} \tag{18}$$

$OUT_{gdt}^{SP}$ are the outflows in year $t$ of individuals from group $g$ born in Spain towards country $d$. All Spanish-born individuals with an unknown destination (approximately one fifth of the total) should are redistributed in proportion to the known-destination numbers. The denominator $pop_{gSPt+1}$ is the population born in Spain from group $g$ that still lived in Spain at the beginning of year $t+1$. The appropriate weight for this regression is $pop_{gSPt}$.

We estimate equation (5) for each of these four series and six age-gender groups and then proceed to forecast emigration rates as described above with the World Bank data. The only difference is that for future years we only update the reference populations for Spain and we keep the variants in United Nations (2015) as the reference for the populations of the rest of the countries in the world.

# 5 Results

This section summarizes the main results of the paper. We first estimate different versions of equation (5) on the World Bank-UN data by sequentially adding additional explanatory variables. We then repeat the exercise with the INE data and finally we describe our predictions.

## 5.1 Estimates with World Bank-UN data

There are several challenges that taking an equation like (5) to the data must face. The main ones are listed by Beine, Bertoli, and Fernández-Huertas Moraga (2016) and we explain now whether we address them and how.

The first one refers to the origin of the migrant. The data we use refers to country of birth although Özden, Parsons, Schiff, and Walmsley (2011) needed to often use citizenship as a proxy for country of birth. The second one is the empirical counterpart of the log odds of migrating. As mentioned above, we proxy the gross migration flows implied by the model by net migration flows between an origin and a destination. The third challenge is what to do about multilateral resistance to migration. As long as our objective is to predict migration flows, this is not a concern for us aside from the fact that the lack of a proxy for the term $MRM_{odt}$ reduces the explanatory power of our model. We will be unable to obtain unbiased estimates of $\beta$ and $\tau$, though. This reflects our inability to deal with the fourth challenge. It will not be possible for us to recover the structural parameters of the random utility maximization model. The fifth challenge refers to the choice of estimating equation (5) in logs, as the theory delivers, or in levels as authors such as Grogger and Hanson (2011) have sometimes preferred. In our case, the fit of the model improves drastically when we estimate it in logs. A related challenge is the presence of zeros in the data. In our case, since we are using net flows to proxy gross flows, we have both zeros (38 per cent of the observations), missing values (30 per cent of our observations, mostly for the UN part of the data) and negative values (10 per cent of the observations). When we also subtract missing population values, our preferred estimates presented below only keep 19 per cent of the potential number of observations. Again, this introduces large biases in our estimated parameters but it does not significantly affect our predictions. A seventh and last challenge is the problem of the endogeneity of some of the variables but we already discussed this point above. We disregard potential endogeneity problems in what follows.

Table 1 introduces our first set of estimates. Column 1 only includes dyadic fixed effects. Given the R$^2$, this means that 76 per cent of the variability in the data can be attributed to time-invariant dyadic factors such as physical distance, linguistic distance, cultural distance, contiguity, common colonial past, etc. Column 2 adds origin-year fixed effects and destination-year fixed effects. This would control for time-varying origin-specific factors and destination-specific factors such as inequality, the level of economic activity, the welfare state,

Table 1: World Bank-UN data (1970-2020)

| Variables | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Networks | | | | 0.21*** | 0.11*** | 0.11*** |
| | | | | (0.04) | (0.04) | (0.04) |
| Pop. <15 origin | | | -0.68*** | | -0.74*** | -0.92*** |
| | | | (0.19) | | (0.23) | (0.22) |
| Pop. 15-65 origin | | | 0.36 | | 0.41 | 0.08 |
| | | | (0.29) | | (0.33) | (0.31) |
| Pop. >65 origin | | | 0.17 | | 0.28 | 0.15 |
| | | | (0.26) | | (0.28) | (0.23) |
| Pop. <15 dest | | | -0.87*** | | -0.35 | -0.34 |
| | | | (0.17) | | (0.24) | (0.24) |
| Pop. 15-65 dest | | | 0.94*** | | 0.08 | 0.06 |
| | | | (0.21) | | (0.32) | (0.32) |
| Pop. >65 dest | | | -0.53*** | | -0.03 | -0.01 |
| | | | (0.20) | | (0.25) | (0.25) |
| GDP pc origin | | | | | -0.04 | 1.97*** |
| | | | | | (0.06) | (0.67) |
| GDP pc origin$^2$ | | | | | | -0.13*** |
| | | | | | | (0.04) |
| GDP pc dest | | | | | 0.07 | 0.07 |
| | | | | | (0.09) | (0.09) |
| Observations | 62,042 | 62,042 | 62,042 | 62,042 | 44,961 | 44,961 |
| Adjusted R-squared | 0.76 | 0.92 | 0.81 | 0.93 | 0.82 | 0.82 |
| Dyad FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Year FE | No | Yes | Yes | Yes | Yes | Yes |
| Origin-year FE | No | Yes | No | Yes | No | No |
| Dest-year FE | No | Yes | No | Yes | No | No |

*** p<0.01, ** p<0.05, * p<0.1. Standard errors clustered by origin-year in parentheses. The dependent variable is the log odds of migrating. All independent variables are in logs (+1 added to the network variable to keep zeros) and lagged 10 years. Regressions weighted by the fully adjusted population lagged 10 years.

the political environment, general migration policies, etc. These controls increase the explanatory power of the model by 16 percentage points. In column 3, we drop these origin and destination-time fixed effects and add the demographic structure of origin and destination countries, which would be collinear with them. We almost go back to the same explanatory power as in column 1. Despite the emphasis on demographic indicators in recent studies, they are only able to explain 5 additional percentage points of the variability in migration rates between 1970 and 2020. In column 4, we try to similarly gauge the contribution of the network variable to explaining the variability across migration rates in the world in that period. When comparing columns 2 and 4, we find that the network variable does not add any explanatory power to the model, barely 1 percentage point. In columns 5 and 6, we add our proxy for economic activity at the origin and at the destination: GDP per capita.[7] These variables increase the explanatory power of the model up to 82 per cent. It is noteworthy that in column 5 no GDP variable appears significant. Only when we consider the quadratic relationship between the economic activity at origin and the emigration rate (Clemens, 2014) we find that GDP per capita at origin becomes significant in equation 6. We find the expected inverted-U relationship, which means that migration rates are first increasing and then decreasing in the development level of an origin country. Our estimated turning point, however, is around $2,000 per capita, much lower than the one estimated by Clemens (2014) without controls but very close to the range that Dao, Docquier, Parsons, and Peri (2018) refer as relevant for the existence of wealth and credit constraints affecting migration decisions.

The main conclusion that we can extract from table 1 is that we should not expect great leverage from the inclusion of additional independent variables into the model. Most of the action is contained in the dyadic fixed effects. As a result, the main difference in projections will come from the span of the data that is used to calculate these fixed effects.

Except for column 1, none of the models estimated in table 1 is useful for prediction purposes. The reason is that they all contain some type of time fixed effect that cannot be projected into the future. The actual prediction models that we will be using in the next section are shown in table 2. Model 1 includes only the demographic structure of the origin and destination country lagged 10 years with respect to the migration rate. The actual

[7] Notice that we lose many observations with missing GDP data in Feenstra, Inklaar, and Timmer (2015) when we do this.

Table 2: Prediction models: World Bank-UN data (1970-2020)

| Variables | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Networks | | 0.08* | 0.07* | 0.08* |
| | | (0.04) | (0.04) | (0.04) |
| Pop. <15 origin | 0.53 | 0.58 | -0.61 | -0.86 |
| | (0.70) | (0.71) | (0.78) | (0.75) |
| Pop. 15-65 origin | -0.97 | -1.08 | 0.24 | -0.16 |
| | (1.03) | (1.02) | (1.14) | (1.26) |
| Pop. >65 origin | 1.15 | 1.08 | 1.19 | 1.43 |
| | (0.75) | (0.77) | (0.90) | (0.99) |
| Pop. <15 dest | -0.12 | 0.05 | 0.27 | 0.28 |
| | (0.21) | (0.21) | (0.21) | (0.22) |
| Pop. 15-65 dest | 0.45** | 0.25 | -0.34 | -0.36 |
| | (0.20) | (0.20) | (0.30) | (0.29) |
| Pop. >65 dest | -0.49* | -0.45* | -0.05 | 0.16 |
| | (0.26) | (0.27) | (0.35) | (0.29) |
| GDP pc origin | | | -0.77*** | 2.05 |
| | | | (0.25) | (2.08) |
| GDP pc origin$^2$ | | | | -0.18 |
| | | | | (0.13) |
| GDP pc dest | | | -0.03 | 0.04 |
| | | | (0.12) | (0.11) |
| Observations | 62,042 | 62,042 | 44,961 | 44,961 |
| Adjusted R-squared | 0.77 | 0.77 | 0.78 | 0.78 |
| Dyad FE | Yes | Yes | Yes | Yes |
| F-test indep. var. | 2.31 | 8.63 | 5.00 | 5.99 |
| p-value | 0.03 | 0.00 | 0.00 | 0.00 |

*** p<0.01, ** p<0.05, * p<0.1. Standard errors clustered by origin-year in parentheses. The dependent variable is the log odds of migrating. All independent variables are in logs (+1 added to the network variable to keep zeros) and lagged 10 years. Regressions weighted by the fully adjusted population lagged 10 years. F-test on the joint significance of the independent variables presented.

variables added to the model are the sizes of the populations aged less than 15, between 15 and 65 and more than 65 years old. Only the size of the working age population at destination and the size of the population above 65 years old at the destination appear significant but none of the two variables keep a consistent sign across specifications. Still, the collinearity of many of these variables and the lack of control for multilateral resistance to migration demands caution for the interpretation of individual coefficients.

The same caution is required for the interpretation of the low and barely significant coefficient on the network variable. Despite the fact that the network variable is the only one that is consistently significant across specifications, its magnitude is much lower than typically reported in the literature, much closer to 1 (Beine, Docquier, and Özden, 2011).

As far as economic conditions are concerned, the coefficient on the GDP per capita at destination is a relatively precisely estimated zero in all specifications in tables 1 and 2. Beyond long-run differences in income, there does not seem to be a role for medium-run fluctuations in economic conditions at destination in explaining migration flows. When we drop dyadic fixed effects however, the correlation becomes statistically significant and the same happens if we estimate the model on data until 2000.[8] The GDP per capita at origin, on the contrary, shows the expected negative relationship with the probability of migrating in column 3 of table 2 with an elasticity of 0.77, close to the classical range in the literature (Beine, Bertoli, and Fernández-Huertas Moraga, 2016). The quadratic specification is no longer significant in table 2 but the size of the coefficients is similar to the one reported in table 1, where year fixed effects were included.

It must be emphasized that none of the models in table 2 adds more than 2 percentage points in explanatory power to the dyadic fixed effects model from column 1 in table 1. Together with the lack of significance of most of the coefficients, this raises the legitimate question of whether the variables included have any meaning at all. To this end, the last two rows in table 2 present a test of the joint significance of all the independent variables. The null hypothesis of no significance is always rejected at conventional levels both when all variables are considered jointly and by blocks, that is, the demographic variables are jointly significant by themselves and the GDP variables are jointly significant by themselves. Hence we can still trust that these variables add something meaningful to the model.

[8] Results available from the authors upon request.

## 5.2 Estimates with INE data

The estimation of equation (5) with Spanish data from the National Statistics Office (INE) turns out to be much more unstable than the estimates based on World Bank-UN data. If we were to show the results of a table like table 1 or table 2, with different models corresponding to a different set of explanatory variables, it would be difficult to ascertain clear patterns. This instability in the coefficients comes from the multicollinearity of many of the variables included in the model and the low number of observations in certain datasets. This problem becomes particularly serious when we estimate the model by age and gender groups. In terms of predictions, the models based on data from the INE have a tendency either to explode or to implode after a few years.[9]

As an illustration of the data that are available, we present in table 3 the estimates equivalent to model 4 in table 2 for the four series of inflows and outflows of foreign-born and Spanish-born individuals. As the dataset becomes smaller, the coefficients get wilder. Despite a very good fit of the models, which leads to quite precise predictions in the very short run, collinear variables inflate coefficients, leading to cyclical variations in the predictions of migration rates.

As a result, we proceed in the rest of the paper with the much more stable results coming form the World Bank-UN data.

## 5.3 Predictions

We next present the predictions associated to 2020-2100 forecasts based on the four models estimated in table 2. To perform these predictions, we also need to predict the independent variables from the table. We take at face value the eight variants that United Nations (2015) provides for the demographic structure of the population. Next, the network variable is updated every ten years based on our own estimates. We proceed similarly with respect to our baseline population, which is both our weight for the regression and the reference to which we apply our fitted emigration rates, as in equation (9). As far as economic activity is concerned, we follow Hanson and McIntosh (2016) and allow per capita GDP to change over time since 2014 based on IMF forecasts of annual GDP growth for 2015-2023 (IMF, 2017) minus population growth from United Nations (2015). After 2023, we assume that the

[9] Results available from the authors upon request.

Table 3: Prediction models: INE data (1988-2016)

| Variables | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Foreign-Born | | Spanish-Born | |
| | Inflows | Outflows | Inflows | Outflows |
| Networks | 0.11*** | 6.49 | -103.74*** | -0.07 |
| | (0.02) | (6.13) | (31.72) | (0.19) |
| Pop. <15 origin | -2.45*** | -26.81*** | -2.30 | -48.65*** |
| | (0.53) | (6.35) | (1.93) | (13.61) |
| Pop. 15-65 origin | 1.82** | 32.37*** | 2.23* | 19.95** |
| | (0.89) | (5.20) | (1.15) | (9.93) |
| Pop. >65 origin | -0.51 | 19.26*** | 0.54 | 21.92*** |
| | (0.70) | (3.43) | (1.13) | (6.60) |
| Pop. <15 dest | -0.80 | 1.33 | -18.12 | -0.14 |
| | (0.66) | (1.37) | (12.60) | (1.62) |
| Pop. 15-65 dest | 5.17*** | -8.28** | 18.14*** | -3.58** |
| | (1.65) | (3.99) | (4.62) | (1.48) |
| Pop. >65 dest | -0.00 | 1.02 | 30.39*** | 2.00** |
| | (1.36) | (0.82) | (3.44) | (0.77) |
| GDP pc origin | 5.87*** | 0.24 | 4.07 | -9.65** |
| | (1.25) | (0.50) | (3.46) | (3.95) |
| GDP pc origin$^2$ | -0.39*** | | -0.22 | |
| | (0.08) | | (0.17) | |
| GDP pc dest | 2.49*** | -0.74** | -7.07*** | 1.08** |
| | (0.41) | (0.36) | (2.22) | (0.52) |
| Observations | 3,742 | 2,033 | 296 | 288 |
| Adjusted R-squared | 0.94 | 0.82 | 0.98 | 0.96 |
| Dyad FE | Yes | Yes | Yes | Yes |
| Data | 1988-2016 | 2003-2016 | 2009-2015 | 2002-2015 |

*** p<0.01, ** p<0.05, * p<0.1. Standard errors clustered by origin-year in parentheses. The dependent variable is the log odds of migrating. All independent variables are in logs (+1 added to the network variable to keep zeros) and lagged 1 year. Regressions weighted by the reference population reported in the main text (section 4.2.2).

growth rate of GDP is constant and equal to the average growth rate between 2000 and 2023 minus the population growth rate from United Nations (2015). We force outliers in terms of GDP growth, countries below the 5th and above the 95th percentile of average growth rates, to grow at least as the 5th percentile and at most as the 95th percentile.

### 5.3.1 World

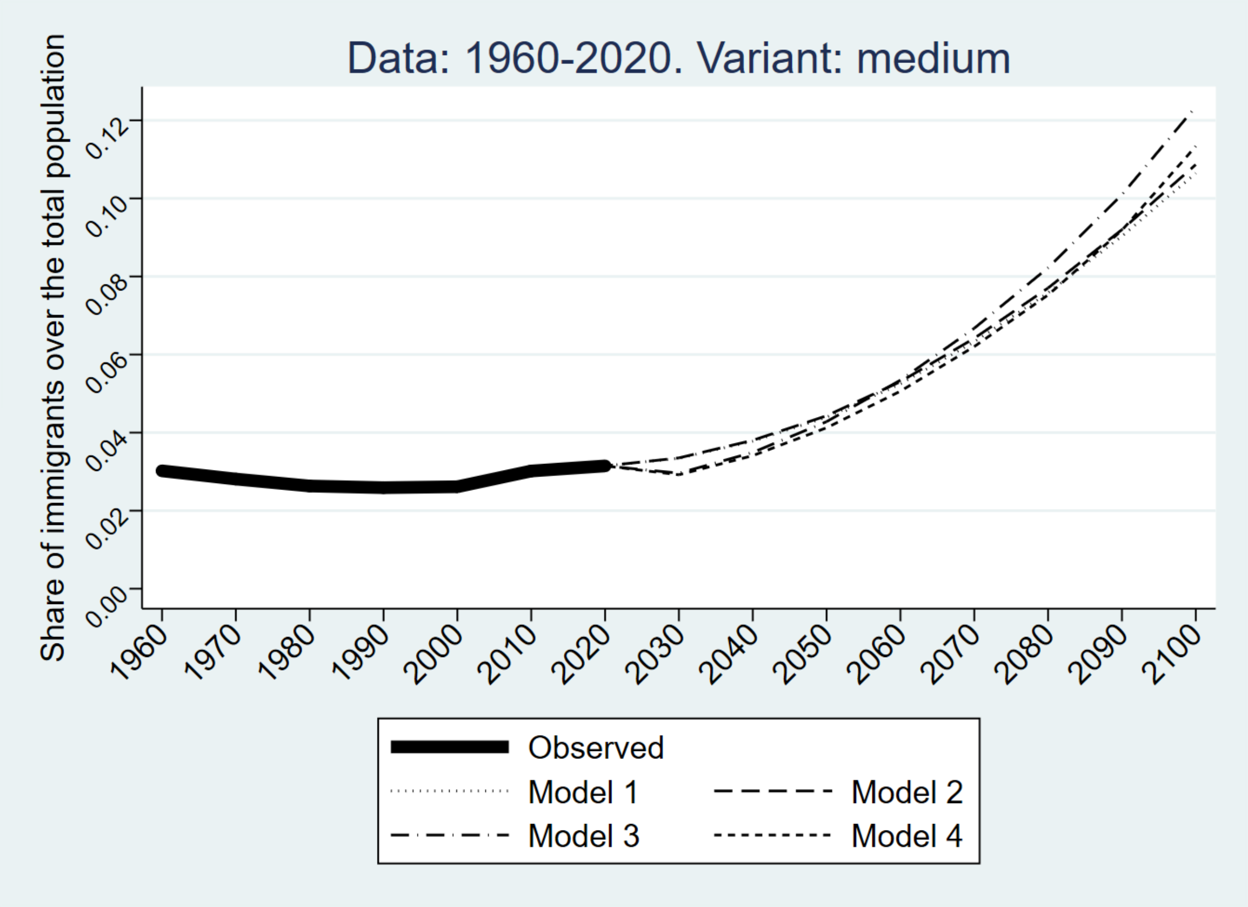
We first show what happens when we estimate our four models on the 1970-2020 World Bank-UN data. Figure 4 represents the actual and predicted evolution of the share of international migrants over the total population between 1960 and 2100. The four models provide a pretty consistent picture in terms of the evolution of the share of immigrants over the world population. From a level of 3.1 per cent at the beginning of the prediction period, the four models go over 4 per cent between 2040 and 2050. They then accelerate to 6 per cent between 2060 and 2070. Ten years later, by 2080, all models are above 7.5 per cent, and ten more years make them go past the 9 per cent mark in 2090. By 2100, the estimates range between 10.6 per cent for model 1 and 12.4 per cent in model 3 of table 2. Model 4, the most complete one, features 11.3 per cent of the world population as international migrants. Most of this growth in the immigrant population can be explained by the evolution of the demographic structure, as model 1 is already quite similar to model 4. If we drop destination-country demographic variables out of model 1, we still predict a similar 12.7 per cent migration share by 2100.[10] Hence, as other papers like Hanson and McIntosh (2016) or Dao, Docquier, Maurel, and Schaus (2017) have already emphasized, demographic conditions in origin countries, notably Africa, will generate large migration pressures over the coming years.

In terms of flows, figure 4 implies moving from the 3.3 millions of immigrants per year between 2010 and 2017 to between 5.7 and 7.8 millions per year between 2020 and 2030. By 2050, total flows would amount to between 11.6 and 13.5 millions per year. At the end of the period, the range would be between the 31.6 millions of model 2 and the 38.6 millions of yearly immigrants from model 3.

The choice of the population variant from United Nations (2015) matters a lot for the predictions. To see how, figure 5 reproduces the evolution of the share of international migrants over the world population according to model 4 from table 2 and the 8 population

[10] Results available from the authors upon request.

Figure 4: Actual and predicted share of world migrants by model (1960-2100)

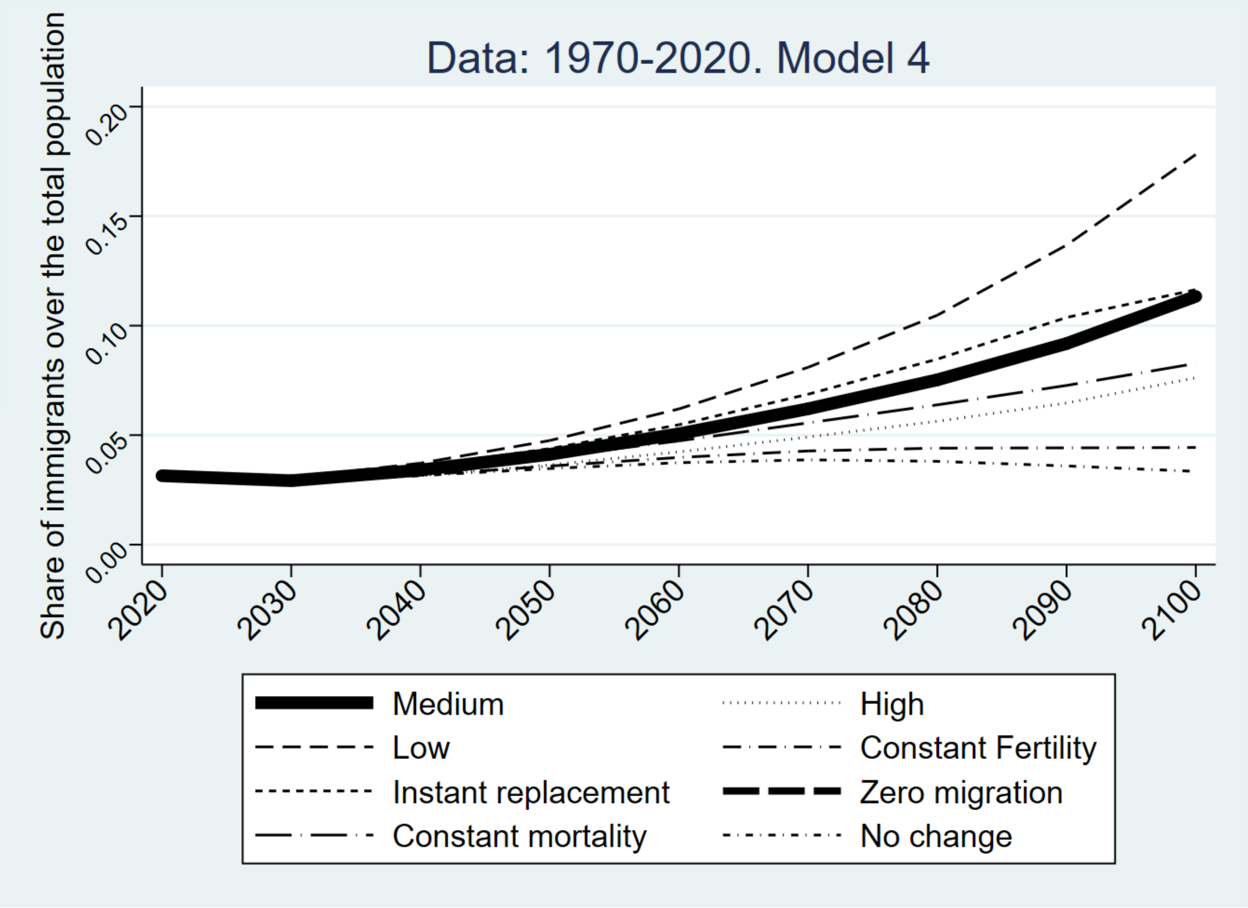


Source: own elaboration on data from Özden, Parsons, Schiff, and Walmsley (2011) and United Nations (2017) and predictions out of table 2. We use the medium population variant of United Nations (2015) to forecast the demographic structure. GDP per capita predictions as described in the text.

variants in United Nations (2015). While the projections appear fairly similar until 2050, the range opens quickly afterwards and by the end of the century different variants predict results as diverse as a share of 3.3 per cent to 17.8 per cent.

Even more important than the population variant is the range of the data over which the model is disciplined. Both table 1 and 2 presented results from the full dataset: 1970-2020. However, the models can be computed in subsets of this full dataset. For example, it might appear reasonable to use only XXIst century data to forecast XXIst century migration flows.

Figure 5: Predicted share of world migrants by population variant (2020-2100)

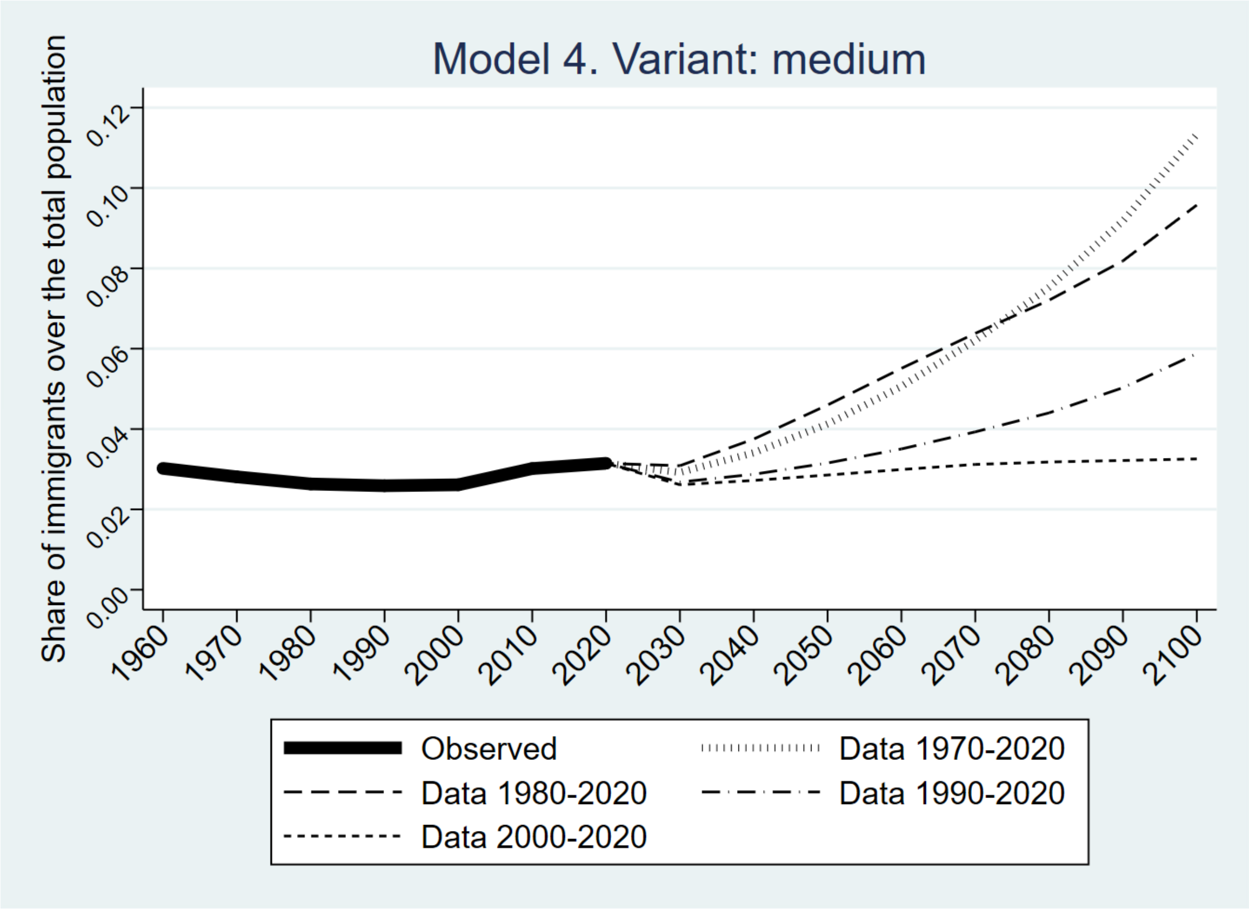


Source: own elaboration on data from Özden, Parsons, Schiff, and Walmsley (2011) and United Nations (2017) and predictions out of column 4 table 2. We use all the population variants of United Nations (2015) to forecast the demographic structure. GDP per capita predictions as described in the text.

This turns out to be a capital choice. Figure 6 draws four such choices.[11] Each line drops the oldest decade of data at a time from 1960-1970 to 1990-2000. As older data get dropped, the predictions of the model become smaller for future migration flows. The range of forecasts is still reasonably narrow by 2050, between 4.6 per cent world migrant share using 1980-2020 data and 2.9 per cent using 2000-2020 data. Nevertheless, the range widens notably again after that point so that by 2100 the full dataset predicts the already known 11.3 share of migrants over the world population while the 2000-2020 restricts this number to 3.3 per cent

[11] The full results from these regressions are available upon request.

Figure 6: Actual and predicted share of world migrants by estimation dataset (1960-2100)



Source: own elaboration on data from Özden, Parsons, Schiff, and Walmsley (2011) and United Nations (2017) and predictions out of the model in column 4 from table 2 for the ranges of data reported. We use the medium population variant of United Nations (2015) to forecast the demographic structure. GDP per capita predictions as described in the text.

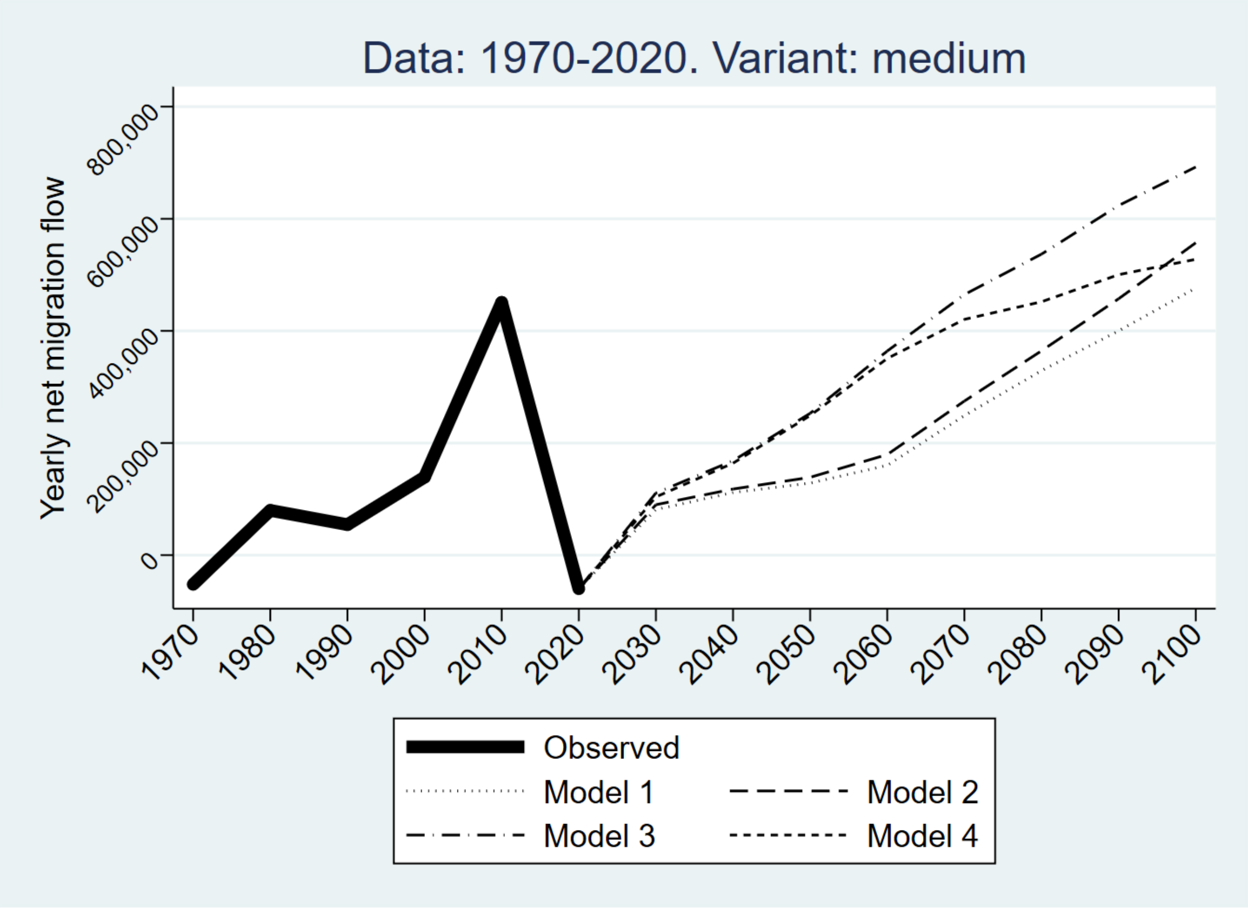
of the world population.

### 5.3.2 Spain

Obviously, these general figures mask a large variety across origin and destination countries. We focus on the Spanish case to provide a specific example of this variety.

The results in terms of net flows for Spain are presented in figure 7 for the medium population variant of United Nations (2015). There are very small differences for the predictions of net flows across specifications in the first decade: 2020-2030. They range between

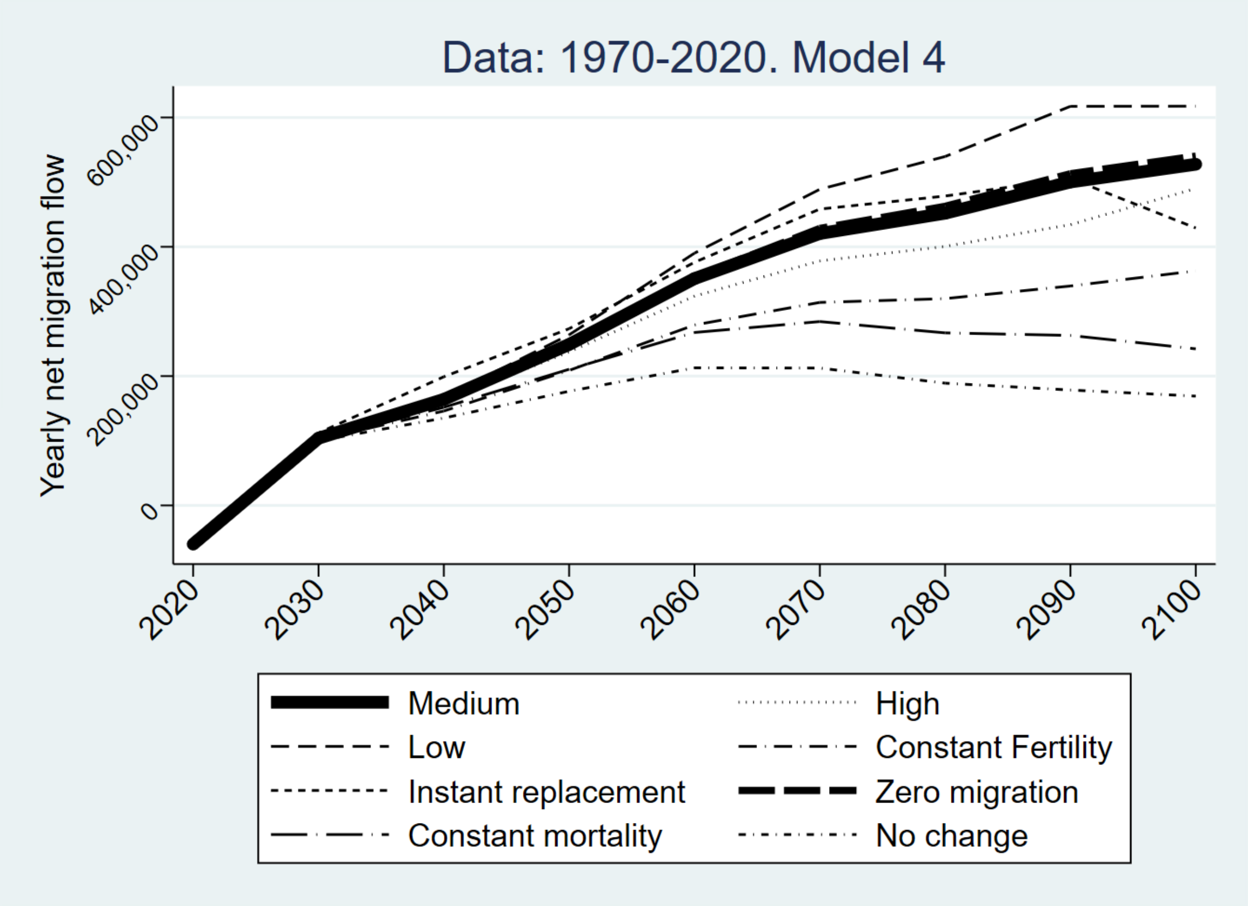
Figure 7: Actual and predicted net migration flows to Spain (1970-2100)



Source: own elaboration on data from Özden, Parsons, Schiff, and Walmsley (2011) and United Nations (2015) and predictions out of table 2. We use the medium population variant of United Nations (2015) to forecast the demographic structure. GDP per capita predictions as described in the text. Net flows calculated as total inflows of foreign-born to Spain minus total outflows of Spanish-born out of Spain.

81,000 and 111,000 net migrants received per year. From that point on, the models without economic variables, columns 1 and 2 in table 2, tend to predict higher yearly inflows than the models that include GDP per capita variables at origin and at destination, columns 3 and 4 in table 2. By 2050, the former predict between 128,000 and 139,000 net immigrants to Spain while the models with economic factors range between 249,000 and 253,000 immigrants. The divergence in the series stops by 2060 and by 2070 model 4 starts to converge with the two non-economic models. This happens because some African countries enter the decreasing phase of the mobility transition. Until 2070, as their GDP per capita increases, their emigration rates increase according to model 4. By 2070, they reach the turning point and their emigration rates start to decline as their economy grows. In 2100, model 3 is an outlier forecasting 693,000 net migrants to Spain per year. Models 1 and 2 range between 476,000 and 557,000. The full model, model 4 from table 2 stays in a middle prediction of 528,000 net immigrants per year during the last decade of the XXIst century.

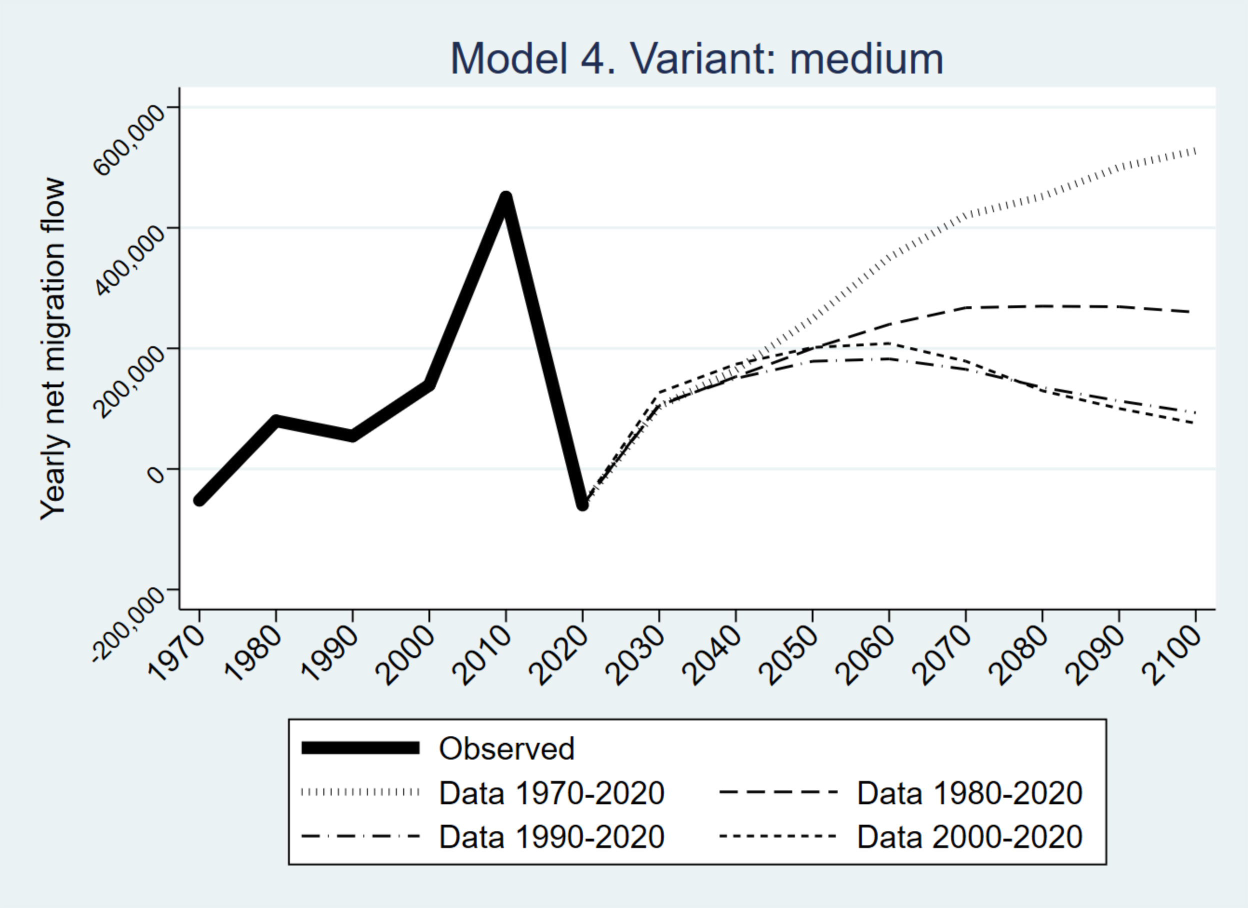
Figure 8: Predicted net migration flows to Spain by population variant (2020-2100)



Source: own elaboration on data from Özden, Parsons, Schiff, and Walmsley (2011) and United Nations (2017) and predictions out of column 4 table 2. We use all the population variants of United Nations (2015) to forecast the demographic structure. GDP per capita predictions as described in the text. Net flows calculated as total inflows of foreign-born to Spain minus total outflows of Spanish-born out of Spain.

It is also interesting to look at the variation depending on the population variant and on the range of data used to generate the predictions, as we did when looking at the evolution of the world share of migrants. This is done in figures 8 and 9. We focus in both cases on the full specification from column 4 in table 2.

Figure 9: Actual and predicted net migration flows to Spain by estimation dataset (1960-2100)



Source: own elaboration on data from Özden, Parsons, Schiff, and Walmsley (2011) and United Nations (2017) and predictions out of the model in column 4 from table 2 for the ranges of data reported. We use the medium population variant of United Nations (2015) to forecast the demographic structure. GDP per capita predictions as described in the text. Net flows calculated as total inflows of foreign-born to Spain minus total outflows of Spanish-born out of Spain.

Figure 8 projects the estimates from model 4 in table 2 between 2030 and 2100 by using the 8 population variants in United Nations (2017). As we saw for the case of the total migrant share in the world, the different variants give rise to quite dissimilar predictions, particularly after 2050. In 2050, all the variants' projections were contained between 177,000 and 273,000 net migrants per year. However, by 2100 the range had become much wider: between 170,000 and 617,000 immigrants per year.

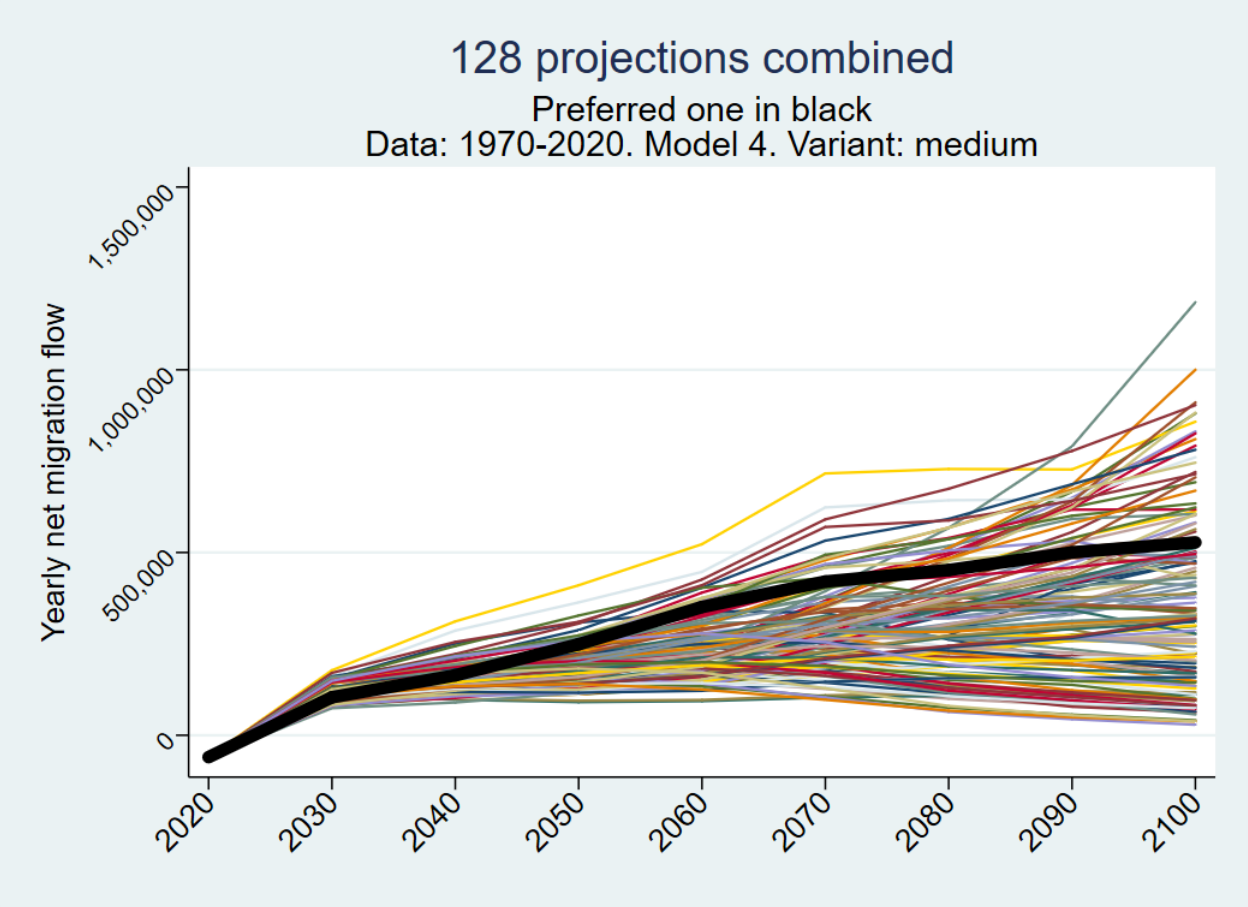
Next, figure 9 shows actual and predicted net flows of migrants to Spain between 1970 and 2100 by estimating model 4 from table 2 on different datasets. As before, these dataset progressively drop their oldest decade so that we go from the full dataset in table 2 (1970-2020) to a restricted version estimated only on XXIst century data (2000-2020). Again, there is very little variation across data ranges as long as we are close enough in time. For the first decade, the predictions range between 104,000 and 127,000 net immigrants per year. By 2040-2050, the range is smaller than in figure 7: it goes from 179,000 to 249,000. Finally, by the end of the period, the predictions range between 76,000 net immigrants per year when forecasting based on the newest data from 2000-2020 to our known 528,000 net immigrants per year when forecasting based on the whole dataset from 1970-2020.

Overall, we have calculated 128 different migration trajectories based on 8 population variants, 4 ranges of data to calculate elasticities and 4 specifications including alternative sets of regressors. Out of these, our preferred specification is the one that uses all of the available data and the most complete model, column 4 in table 2. Figure 10 emphasizes our preferred projection among the 128 that we have presented. Our preferred option could be considered an average prediction. In general, we confirm the fact that most predictions are relatively similar over the medium run while they vary wildly over the long run.

Finally, we break down our preferred prediction by inflows and outflows in figure 11. This figure makes it even clearer than the previous ones that our models consider the Spanish immigration boom, mostly fitting in the 2000-2010 decade, as an anomaly. Visually, it looks like inflows are projected to come back to the path they were on by year 2000, with a bit of an acceleration until the years 2060-2070. Only then the levels of net flows of 2000-2010 would be repeated. On the other hand, outflows are projected to grow until around 2040 when they would start to decline from a maximum just short of 30,000 emigrants per year.

The translation of these flows or net flows into shares can be useful to compare the evolution of Spanish immigration to the evolution of world migration. According to model 4 in table 2, Spain would rise from a 12 per cent foreign born share in its population in 2020 to 13 per cent in 2030. It would go above 15 per cent in 2040 and 19 per cent in 2050. It

Figure 10: Predicted net migration flows to Spain by population variant, model and data range (2020-2100)

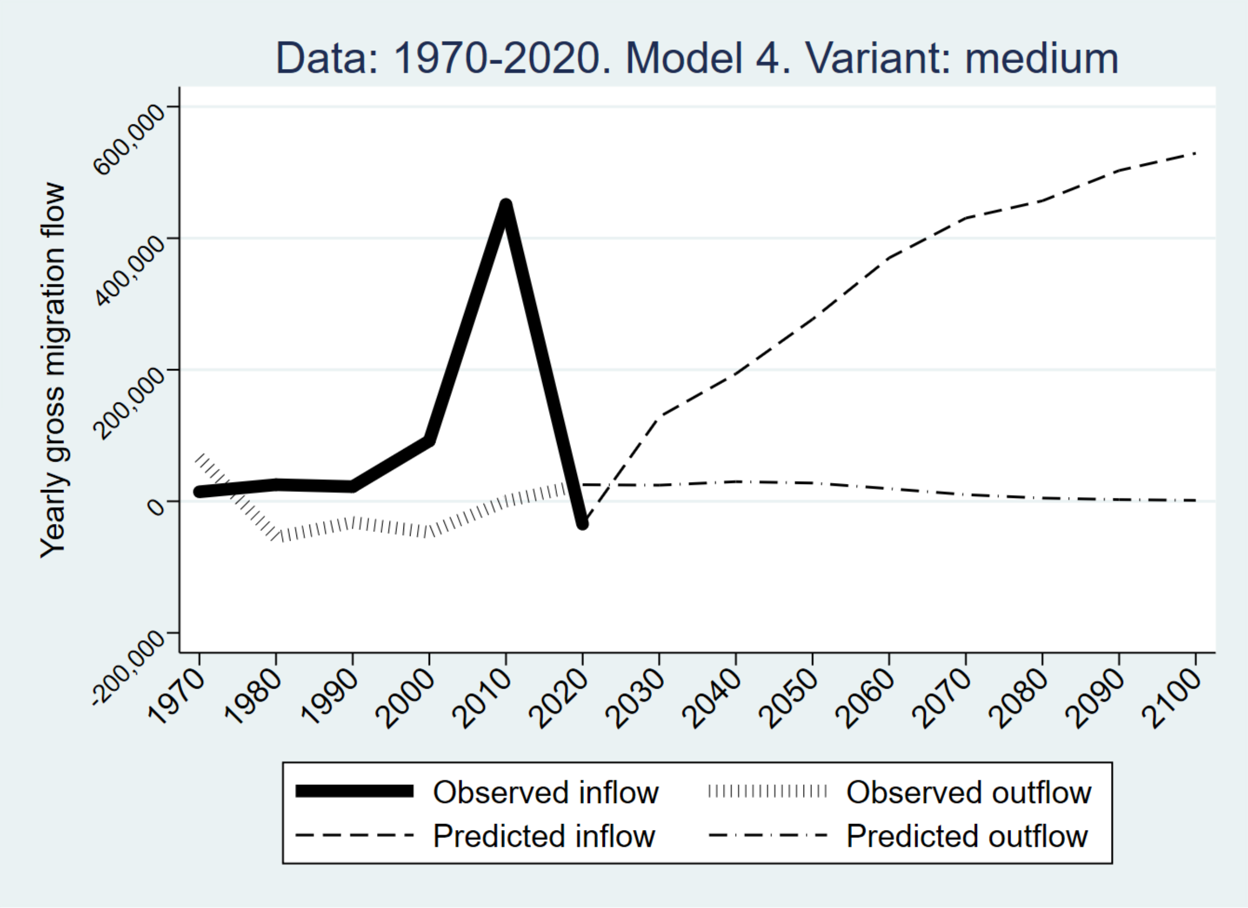


Source: own elaboration on data from Özden, Parsons, Schiff, and Walmsley (2011) and United Nations (2017) and predictions out of the models in table 2 over different data ranges: 1970-2020, 1980-2020, 1990-2020 and 2000-2020. We use all the population variants of United Nations (2015) to forecast the demographic structure. GDP per capita predictions as described in the text. Net flows calculated as total inflows of foreign-born to Spain minus total outflows of Spanish-born out of Spain.

would then continue to grow until reaching 40 per cent in 2100.

As far as the composition of the flows is concerned, Latin America and Africa would be dominant, with a growing relevance of the latter. In the first decade of the projection, Morocco would be the main origin country, followed by Venezuela and Ecuador. By 2050, the top three would be the same but with Venezuela on top and followed closely by Colombia

Figure 11: Actual and predicted inflows and outflows to Spain (1970-2100)



Source: own elaboration on data from Özden, Parsons, Schiff, and Walmsley (2011) and United Nations (2017) and predictions out of the model in column 4 from table 2. We use the medium population variant of United Nations (2015) to forecast the demographic structure. GDP per capita predictions as described in the text.

and Algeria. At the end of the period, Senegal, Gambia and Guinea would join Venezuela and Ecuador in the top five of origin countries in terms of flows.

### 5.3.3 Uncertainty

The comparison of the 128 migration scenarios that we have mentioned gives us an idea of the wide variety of results that can be expected in terms of migration flows. However, it should not be forgotten that none of these scenarios is taking into account the classical prediction error associated to each of the models. This subsection provides an approximation to this concept. We focus only on our preferred set of estimates: model 4 in table 2 estimated on the full dataset 1970-2020.

Model 4, a version of equation (5), is linear. However, as shown in section 4, the predictions of flows out of the log odds of migrating are non-linear. Furthermore, these flows have to be aggregated over origins or destinations, hence complicating the obtention of analytical standard errors. As a consequence, we have decided to resort to bootstrapping in order to generate standard errors and a confidence interval for our predictions. We present our results below in terms of the standard error associated to the forecast of migration flows to and from Spain according to our model but the exercise has been done for every dyadic migration flow between an origin and a destination.

Our bootstrapped standard errors come from resampling observations with replacement out of our full dataset. We have repeated this procedure 1,000 times to obtain the results presented below.

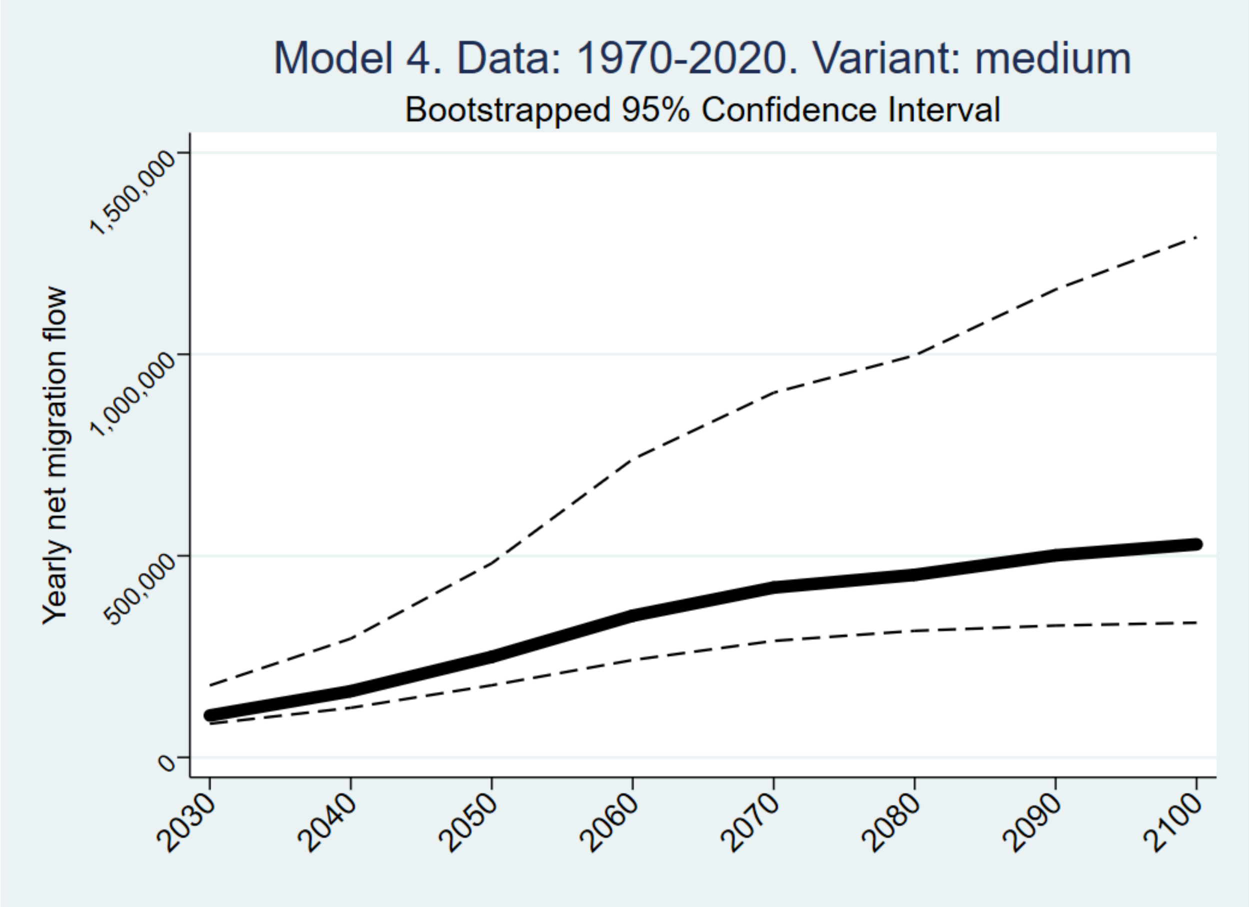
Consistently with the evidence from the migration scenarios above, the uncertainty associated to one particular projection is also increasing over time. The standard error on projected immigration flows multiplies by more than 10 throughout the projection period, from 27,000 in the first decade to more than 300,000 in the last. Similarly, while migration stocks multiply by almost 4, from 6.5 million in 2030 to 24.7 million in 2100, standard errors associated to immigration stocks multiply by 34, from around 300,000 in 2030 to 10.2 million in 2100.

We also performed the same exercise for emigration flows and stocks of emigrants out of Spain. In this case, the numbers are smaller and declining after 2040 for flows and after 2050 for emigrant stocks. Although the standard errors also go down with the size of the flows, we end up the prediction period with a standard error larger than the predicted outflow of Spanish emigrants.

Beyond standard errors, perhaps the degree of uncertainty involved in our preferred prediction is best described by figure 12. The figure represents the net migration flows to Spain from our preferred specification, the one of model 4 in table 2, together with the bootstrapped 95% confidence interval. Hence the thick black line in figure 12 is obtained by subtracting emigration flows from immigration flows and we already explained above how we computed the 1,000 replications that gave rise to our confidence interval.

We can see in figure 12 that the confidence interval is much wider on the upper than on

Figure 12: Predicted net migration flows to Spain. Preferred specification and bootstrapped 95% confidence interval (2030-2100)



Source: own elaboration on data from Özden, Parsons, Schiff, and Walmsley (2011) and United Nations (2017) and predictions out model 4 in table 2. Bootstrapped confidence interval from 1,000 replications resampling the data with replacement. We use the medium population variant of United Nations (2015) to forecast the demographic structure. GDP per capita predictions as described in the text. Net flows calculated as total inflows of foreign-born to Spain minus total outflows of Spanish-born out of Spain.

the lower part of our main prediction. There is a mechanical reason for this, since predicted inflows cannot be less than or equal to zero by definition and they dominate the size and variability of migration outflows, which enter in the calculation of net flows with a negative sign. As a result, we end up with an asymmetric confidence interval that goes as far up as almost 1.3 million immigrants per year by 2100, while the lower bound barely goes above

0.3 in our central prediction of 0.5 million net immigrants per year in the last decade of the XXIst century. In the first decade, the range is much narrower. The fifth percentile lies at 84,000 while the ninety-fifth reaches 179,000. Curiously, the ninety-fifth percentile of 2030 becomes the fifth percentile in 2050, while the upper bound goes up to 482,000 net immigrants per year.

# 6 Conclusion

This paper has provided a rich variety of predictions of Spanish immigration and emigration flows for the XXIst century. Two main conclusions stand out from the paper.

Firstly, the main determinants of migration flows seem to be dyadic and quite time-invariant. Dyadic fixed effects absorb more than 75 per cent of the historical variation in migration flows across origins, destinations and time. This could mean that time-invariant dyadic variables such as distance are the main determinants of migration flows or else that other relevant determinants have long-standing average effects that do not vary much across the three dimensions, thus preventing identification. Despite this, the prominent role of demography at origin emphasized by the recent literature is also confirmed by our estimates when we use our model to predict future migration flows.

Secondly, the projections of emigration flows for Spain tend to paint a very consistent picture in the medium run, until around 2050, and an erratic one afterwards. In the medium run, most models and datasets forecast Spain to receive large quantities of immigrants, around 200,000 in net terms per year. In the longer run, predictions are all over the place, from close to zero per year to about 1 million. This huge variation limits the usefulness of the exercise in the very long run. Furthermore, many of the assumptions needed to generate the results are reasonable within a generation but become quite difficult to sustain further into the future.

# References

Beine, M., S. Bertoli, and J. Fernández-Huertas Moraga (2016): “A Practitioners Guide to Gravity Models of International Migration,” *The World Economy*, 39(4), 496–512.

Beine, M., F. Docquier, and c. Özden (2011): “Diasporas,” *Journal of Development Economics*, 95(1), 30–41.

Bertoli, S. (2017): “Is the Mediterranean the New Rio Grande? A Comment,” *Italian Economic Journal: A Continuation of Rivista Italiana degli Economisti and Giornale degli Economisti*, 3(2), 255–259.

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

——— (2015): “The size of the cliff at the border,” *Regional Science and Urban Economics*, 51(C), 1–6.

Borjas, G. J. (2003): “The Labor Demand Curve is Downward Sloping: Reexamining the Impact of Immigration in the Labor Market,” *Quarterly Journal of Economics*, 118(4), 1335–1374.

Campos, R. G. (2017): “International migration pressures in the long run,” Working Papers 1734, Banco de España.

Clemens, M. A. (2014): “Does development reduce migration?,” in *International Handbook on Migration and Economic Development*, Chapters, chap. 6, pp. 152–185. Edward Elgar Publishing.

Dao, T., F. Docquier, M. Maurel, and P. Schaus (2017): “Global Migration in the 20th and 21st Centuries: the Unstoppable Force of Demography,” .

Dao, T., F. Docquier, C. Parsons, and G. Peri (2018): “Migration and development: Dissecting the anatomy of the mobility transition,” *Journal of Development Economics*, 132(C), 88–101.

Docquier, F., c. Özden, and G. Peri (2014): “The Labour Market Effects of Immigration and Emigration in OECD Countries,” *The Economic Journal*, 124(579), 1106–1145.

Feenstra, R. C., R. Inklaar, and M. P. Timmer (2015): “The Next Generation of the Penn World Table,” *American Economic Review*, 105(10), 3150–3182.

González-Ferrer, A. (2013): "La nueva emigración española. Lo que sabemos y lo que no," Zoom Político 18, Fundación Alternativas.

Grogger, J., and G. H. Hanson (2011): "Income maximization and the selection and sorting of international migrants," *Journal of Development Economics*, 95(1), 42–57.

Hanson, G. H., and C. McIntosh (2010): "The Great Mexican Emigration," *The Review of Economics and Statistics*, 92(4), 798–810.

——— (2012): "Birth Rates and Border Crossings: Latin American Migration to the US, Canada, Spain and the UK," *Economic Journal*, 122(561), 707–726.

——— (2016): "Is the Mediterranean the New Rio Grande? US and EU Immigration Pressures in the Long Run," *Journal of Economic Perspectives*, 30(4), 57–82.

IMF (2017): *IMF Datamapper*. `http://www.imf.org/external/datamapper/index.php`.

INE (2016): *Proyecciones de Población de España 2016-2066. Metodología*. `http://www.ine.es/inebaseDYN/propob30278/docs/meto_propob_2016_2066.pdf`.

——— (2017a): *Estadística de Variaciones Residenciales*. `http://www.ine.es`.

——— (2017b): *Estadística del Padrón Continuo*. `http://www.ine.es`.

——— (2017c): *Estadística del Padrón de españoles residentes en el extranjero*. `http://www.ine.es`.

Lanzieri, G. (2017): *Summary methodology of the 2015-based population projections*. EUROSTAT. `http://ec.europa.eu/eurostat/cache/metadata/Annexes/proj_esms_an1.pdf`.

McFadden, D. (1974): "Conditional logit analysis of qualitative choice behavior," in *Frontiers in Econometrics*, ed. by P. Zarembka, pp. 105–142. New York: Academic Press.

Mckenzie, D., and H. Rapoport (2007): "Network effects and the dynamics of migration and inequality: Theory and evidence from Mexico," *Journal of Development Economics*, 84(1), 1–24.

Munshi, K. (2003): "Networks in the Modern Economy: Mexican Migrants in the U. S. Labor Market," *The Quarterly Journal of Economics*, 118(2), 549–599.

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.
ΩÖzden, Parsons, Schiff, and Walmsley

Özden, c., C. R. Parsons, M. Schiff, and T. L. Walmsley (2011): "Where on Earth is Everybody? The Evolution of Global Bilateral Migration 1960-2000," *World Bank Economic Review*, 25(1), 12–56.

Sjaastad, L. A. (1962): "The Costs and Returns of Human Migration," *Journal of Political Economy*, 70(5, Part 2), 80–93.

United Nations (2015): "World Population Prospects: The 2015 Revision, Methodology of the United Nations Population Estimates and Projections," Department of Economic and Social Affairs, Population Division Working Papers ESA/P/WP.242.

——— (2017): "Trends in International Migrant Stock: The 2017 Revision," Department of Economic and Social Affairs, Population Division. United Nations database POP/DB/MIG/Stock/Rev.2017.

World Bank (2017): *World Development Indicators*. http://databank.worldbank.org.