Networks and Selection in International Migration to Spain*

Nina Neubecker‡ Marcel Smolka§ Anne Steinbacher¶

Berlin Institute for Population and Development

Aarhus University

This version: November 2016; first version: May 2012

Abstract

We offer fresh evidence on the effect of migrant networks on two essential aspects of migration: (i) the total scale of migration, and (ii) the skill composition of migration. Our analysis is for the remarkable case of Spain, which experienced a full-blown immigration boom from the mid 1990s up to the Global Financial Crisis. To accommodate flexible substitution patterns across alternative migrant destinations, we use a three-level nested multinomial logit model. We find a strong positive network effect on the scale of migration and a strong negative effect on the ratio of high-skilled to low-skilled migrants. Simplifying restrictions on the structure of cross-destination substitutability are rejected by the data.

JEL Codes: F22, J61

Keywords: international migration · migrant networks · nested multinomial logit model · skill composition of migration · Spain

*This paper is a revised version of Aarhus University Economics Working Paper No. 2015-03. We have benefitted from comments by Wilhelm Kohler, Udo Kreickemeier, Peter Eppinger, Gordon Hanson, Johannes Pfeifer, Melissa Siegel, two anonymous referees, and seminar and conference participants at various universities. Research assistants at the Universities of Tübingen and Aarhus, especially Boris Georgiev, have provided excellent support. Marcel Smolka gratefully acknowledges financial support from the Tuborg Foundation, as well as from the Volkswagen Foundation under the project "Europe's Global Linkages and the Impact of the Financial Crisis". Part of the work on this paper was done while Marcel Smolka was a visiting PhD student at University College London (UCL). The hospitality of the Department of Economics and the Centre for Research and Analysis of Migration (CReAM) at UCL is gratefully acknowledged.

‡Berlin Institute for Population and Development, Schillerstraße 59, 10627 Berlin, Germany.

§**Corresponding author:** Department of Economics and Business Economics, Aarhus University, Fuglesangs Allé 4, Building 2632, 8210 Aarhus V, Denmark. E-mail: msmolka@econ.au.dk, Phone: +4 8716 4974.

¶No longer affiliated with an academic institution.

1 Introduction

Societies care a lot about the number and type of immigrants they receive. A major concern are the economic consequences of immigration, e.g. the effects on wages and employment of already resident workers. But there are also other concerns, unrelated to economic factors. A careful analysis of attitudes in the UK, for example, finds that racial and cultural prejudice plays a considerable role in explaining hostility towards immigration (Dustmann & Preston, 2007).

This paper contributes to the empirical literature on the *causes* of migration. We focus on one specific cause of migration: migrant networks. According to a popular definition by Massey (1988, 396), migrant networks are "sets of interpersonal ties that link migrants, former migrants, and nonmigrants in origin and destination areas through the bonds of kinship, friendship, and shared community origin." Following a large literature, both in sociology and economics, we assume that migrant networks reduce the migration costs of those left behind in their countries of origin. One important component of migration costs, for instance, is the job search process in the destination country. Empirical evidence shows that informal job referrals through the migrant network can greatly facilitate this process (Munshi, 2003). Many other channels through which migrant networks can reduce migration costs are conceivable, e.g. the provision of credit, protection, and companionship, as well as of information on border crossing, education, housing and the like. From this perspective, the notion of migrant networks is instrumental to understanding observed patterns of migration and ultimately predicting future migration.[1]

We use the remarkable case of Spain to offer fresh evidence on the effect of migrant networks on two essential aspects of migration: (i) the total scale of migration, and (ii) the skill composition of migration. Our estimates of network effects are important for several reasons. First, they can help predict the magnitude and distribution of future migration based on the settlement pattern of previous migrants. Importantly, network effects imply that those destinations that already host a lot of migrants can expect to receive large inflows of migrants, and this may happen even if the income gap between origin and destination is shrinking. Although not the focus of our paper, this has, for example, implications for the classical question whether economic development in poor countries will cause less South-North migration, or more. Secondly, how migrants are selected in terms of their education and skills is a matter of considerable public and academic interest. By reducing the migration costs of those left behind, migrant networks can influence this selection, because they encourage those individuals

[1]Note that migrant networks involve a positive externality: migrants do not take into account the favorable effect they have on the migration costs of those left behind. Hence, in a dynamic model of migration, network effects indicate a welfare loss in the laissez-faire transition path equilibrium, and the optimal policy response is to accelerate the speed of migration (Carrington et al., 1996).

to migrate who have initially higher migration costs (typically the less skilled individuals). Our paper is one of the few to offer causal empirical evidence on this important mechanism. Finally, while our estimates come from a static cross-sectional approach (as will become evident later), we believe our analysis of network effects also speaks to the dynamics and timing of migration. In particular, network effects imply that, in the aggregate, migration is a process that takes place gradually over time, and that the first migrants to arrive are drawn from a different part of the origin's skill distribution than later migrants (Carrington et al., 1996). In our view, these insights can be used to expand the perspective on the labor market effects of immigration, which have mostly been analyzed, not in a dynamic, but in a static framework.

The focus of our analysis is on Spain. The country has a long tradition of emigration, but developed into one of the world's most attractive migrant destinations due to its strong economic growth ahead of the Global Financial Crisis. According to the Spanish Instituto Nacional de Estadística (INE), the country received roughly six million new migrants from 1997 to 2009. These are drawn from a diverse ethnic background: 13.6% are Romanians, followed by Moroccans (11.1%), Ecuadorians (8.2%), Colombians (6.1%), Britons (5.3%), and Bolivians (4.7%). The foreign-born share among the total population increased quite spectacularly over just a few years, from 4.9% in 2000 to 14.1% in 2008 (OECD, 2010, 240).

Our empirical analysis exploits *aggregate* migration data, i.e. stocks and flows of migrants. To guide our analysis and obtain estimates that have a clear and useful structural interpretation, we apply a random utility framework as a modelling device. A central feature of our model—a three-level nested multinomial logit (NMNL) model—is its hierarchical structure (countries, regions, and provinces) which can accommodate rich substitution patterns across alternative destinations (McFadden, 1984, 1422-1428). More specifically, our model suggests that migrants are more likely to substitute destinations *within* rather than *across* countries and regions because they are "similar" (to varying degrees, which we model through heterogeneous similarity parameters): they share the same legal and political system; they have a common cultural background; they engage in similar economic activities and so on. This means that people with strong idiosyncratic preferences for, say, Barcelona tend to find other places in Cataluña (and Spain) attractive, too (e.g. due to their language or climatic preferences). The opposite is also true. People who dislike Barcelona tend to dislike Cataluña (and Spain) as a whole. This idea contrasts sharply with the model usually employed in the literature—the standard multinomial logit (MNL) model—which assumes a uniform degree of cross-destination substitutability. It also poses additional challenges in the estimation, as well as the interpretation of the obtained estimates, as we will discuss in more detail later.

The Spanish case is well-suited for an empirical study of network effects in migration. The migration data available from INE are of exceptional quality and coverage. They allow us to exploit variation across a large number of countries of origin, as well as across all regions and provinces of destination in Spain. By slicing our data along different dimensions of origin and destination, we can thus use a rich set of fixed effects in order to (i) provide estimates that are fully consistent with the hierarchical structure of our NMNL migration model, and (ii) tackle endogeneity of migrant networks due to unobserved heterogeneity in migration costs. For example, if migrants from Latin America are generally more welcome in, say, Barcelona, then we can control for this in the estimation. Also, if migrants from France have particularly strong preferences for the provinces in nearby Cataluña, then this is absorbed into our fixed effects. We are not aware of any other study on migrant networks that employs a similarly extensive set of fixed effects to address the issue of endogeneity bias. To further strengthen our analysis, we instrument migrant networks by historical internal migration flows in Spain.

Our estimates reveal robustly positive network effects on the scale of migration. The effects are of considerable size, and overall similar to those reported in the literature. Our estimates also attest to strong negative effects of migrant networks on the skill composition of migration, defined as the ratio of high-skilled to low-skilled migrants. This finding is consistent with migration costs being higher for low-skilled individuals (as argued by Chiswick, 1999) and suggests that it is the low-skilled who benefit the most from the presence of migrant networks. Our estimates strongly reject a uniform degree of substitutability across alternative destinations, thus indicating the relevance of the NMNL model in our application to the Spanish case. We find pronounced heterogeneity in the estimated network coefficients (reflecting heterogeneous similarity parameters across regions), an observation that has received no attention so far in the literature. We use the structural interpretation of our network coefficients in order to exploit this heterogeneity and compute elasticity values for the network effect. We find the lowest network elasticity for the region of Extremadura, slightly exceeding a value of 0.1, and the highest network elasticity for the region of Cataluña, lying in the vicinity of 0.55. This indicates that the different provinces in Cataluña are relatively similar and thus easier to substitute for one another—a plausible result given the region's high degree of political and cultural autonomy.

Our paper is related to several studies that use aggregate migration data to estimate network effects in migration. Beine et al. (2011) investigate the determinants of the scale and skill composition of migration between the years 1990 and 2000 to 30 OECD countries. They find that economies with larger migrant networks attract more new migrants as well as a larger share of low-skilled migrants.[2]

[2]See also Grogger & Hanson (2011, 53) for complementary evidence. McKenzie & Rapoport (2010) find that positive self-selection on education from Mexican migrants to the U.S. is more likely, the larger the number of return migrants in

Similar results are obtained by Beine & Salomone (2013) who study potential gender differences in network effects. The paper by Beine et al. (2015) disentangles what the authors call local and national network externalities, arguing that local networks facilitate the assimilation into the host society, while nation-wide networks help overcome the legal entry barriers to migration. However, all of these papers are based on a standard MNL model which assumes a uniform degree of cross-destination substitutability.[3]

Our paper is also related to a number of macro-level studies that are more generally concerned with the determinants of international migration.[4] In this literature, migrant networks robustly rank among the most important factors shaping migration, but the estimated migration functions often lack an explicit micro-foundation (Clark et al., 2007; Lewer & Van den Berg, 2008; Pedersen et al., 2008; Mayda, 2010). Exceptions are Bertoli & Fernández-Huertas Moraga (2013), who use the same Spanish data source as we do in this paper[5], and Ortega & Peri (2013), but these papers do not identify the effects of migrant networks on the scale and skill composition of migration, as we do in this paper.

The rest of the paper is organized as follows. Section 2 presents our three-level NMNL migration model. We use this model to derive estimation equations for both the scale and the skill composition of migration. In Section 3 we introduce the migration data we employ in the estimation. In Section 4 we present our estimation strategy and the results, and we provide a structural interpretation of these results in terms of our NMNL migration model. Section 5 concludes.

2 The Model

In this section we develop a three-level NMNL migration model with many countries of origin and many provinces of destination within countries.

the origin community. Bertoli (2010) finds a positive interaction between the number of migrants abroad and the extent of negative self-selection, using individual-level data on Ecuadorian emigrants.

[3]This is also true for a recent paper by Llull (2016). An exception are Bertoli & Fernández-Huertas Moraga (2015) who use the same migration data as Beine et al. (2011) in order to estimate network effects in migration. The most general version of the model they estimate is a two-level NMNL model with a homogeneous similarity parameter for all "nests" (countries and regions in our paper). Hence, the pattern of cross-destination substitutability permitted by their model is more restrictive than the one in our model.

[4]For the location choice of migrants within borders, see Bartel (1989), Zavodny (1997, 1999), Chiswick & Miller (2004), Card & Lewis (2007), and Jayet et al. (2010). Selected survey-based studies on migration decisions at the micro-level include Åslund (2005), Baghdadi (2005), Bauer et al. (2005, 2009), and Dolfin & Genicot (2010).

[5]Fernández-Huertas Moraga et al. (2015) use the Spanish data to investigate the location choices of natives in response to the Spanish immigration boom.

2.1 Basic Setup

We assume that decision making follows a hierarchical structure in which provinces (the final migration destinations) are grouped into higher-level territorial entities (nests). Individuals "eliminate" nests until a single province remains. Decision making can be described in a hierarchical manner[6]: first to which country to migrate (including the country of origin); second which region to move to within the chosen country; and third which province to pick within the preferred region. We index the countries of origin by $i = 1, \ldots, I$; the countries of destination (the primary nests) by z or $y = 1, \ldots, Z$; the regions of destination (the secondary nests) by r or $\ell = 1, \ldots, R$; and the provinces of destination by j or $k = 1, \ldots, J$. Let the country of origin i be one element in each of the sets $\{1, \ldots, Z\}$, $\{1, \ldots, R\}$, and $\{1, \ldots, J\}$, thus representing a degenerate nest with a single final migration destination. Define A_{zr} as the set of provinces in region r of country z, and A_z as the set of regions in country z.

We write the utility of individual o who migrates from country i to province j as:

$$U_{ij}^{o} = Y_j - C_{ij} + e_{ij}^{o}, \tag{1}$$

where $o = 1, \ldots, m_i$ indexes individuals from country i, the terms Y_j and C_{ij} are sub-utility functions for moving from country i to province j, and the term e_{ij}^{o} is a random utility variable with idiosyncratic realizations for each province $j = 1, \ldots, J$. This last variable reflects unobserved individual heterogeneity relevant for the migration decision (age, productivity, family status, occupation etc.). The function Y_j summarizes characteristics of province j such as the wage rate, the state of the housing market, or the climate.[7] The function C_{ij} captures the costs of moving and assimilation, henceforth called migration costs. Similarly to Beine et al. (2011, 33-34), we hypothesize that these costs are decreasing convexly in the migrant network, M_{ij}, defined as the number of co-national migrants already resident in province j. A convenient specification of migration costs that incorporates positive but diminishing returns to the migrant network uses the log of M_{ij}:

$$C_{ij} = c_{iz} + c_{ir} + c_{ij} - \theta \ln(1 + M_{ij}), \quad j \in A_{zr}, r \in A_z, \tag{2}$$

where the parameter $\theta > 0$ measures the strength of the network effect, and where we add one to the variable M_{ij} before taking logs in order to rule out infinitely large migration costs. The other

[6] We assume that each decision in this hierarchy is made conditional on both the fixed preceding decisions and the optimal succeeding decisions. Hence, individuals decide on all aspects of their migration moves simultaneously (cf. Domencich & McFadden, 1975, 33-46).

[7] The focus of our model and the empirical analysis is on how migrants choose their first destination. In reality, migrants often move in stages, giving rise to internal migration within the country of destination, or onward migration to a third country. If some destinations serve as particularly attractive "transit migration hubs", then this is captured by the function Y_j.

cost components will be described in more detail below. Suffice it to say here that, for a given country of origin i, they vary across either countries of destination (c_{iz}), regions of destination (c_{ir}), or provinces of destination (c_{ij}). For expositional convenience, we define $U_{ij} \equiv U_{ij}^o - e_{ij}^o = Y_j - C_{ij}$ and $\xi_{ij} \equiv Y_j - c_{ij} + \theta \ln(1 + M_{ij})$.

Individuals are assumed to maximize utility. Let $j^o \in \{1, \ldots, J\}$ denote the destination that offers the highest utility to individual o. The probability that this individual migrates from country i to province j thus reads as:

$$\begin{aligned} P_i^o(j^o = j) &= \Pr(U_{ij}^o > U_{ik}^o \ \forall\, k \in \{1, \ldots, J\} : k \neq j) \\ &= \Pr(e_{ik}^o - e_{ij}^o < U_{ij} - U_{ik}; \\ &\quad \forall\, k \in \{1, \ldots, J\} : k \neq j). \end{aligned} \tag{3}$$

By the laws of conditional probability, we can express this probability as a product of transition probabilities:

$$P_i^o(j^o = j) = P_i^o(j^o = j | j^o \in A_{zr}) P_i^o(j^o \in A_{zr} | r \in A_z) P_i^o(r \in A_z), \quad j \in A_{zr}, r \in A_z. \tag{4}$$

These probabilities depend on the distribution assumed for the random utility variables, $e_{i1}^o, \ldots, e_{iJ}^o$. It can be shown that under certain assumptions about the function H_i the function

$$F_i(e_{i1}^o, \ldots, e_{iJ}^o) = \exp\left[-H_i\left(\exp[-e_{i1}^o], \ldots, \exp[-e_{1J}^o]\right)\right] \tag{5}$$

is a multivariate extreme value distribution; see McFadden (1978, 80-81; 1981, 226-230). As is standard in the literature, we assume that $(e_{i1}^o, \ldots, e_{iJ}^o)$ is distributed F_i. However, we depart from the literature in that we introduce a function H_i that generates the response probabilities of a three-level NMNL model. This model allows for the random utilities associated with provinces in the same region (or the same country) to be mutually correlated, whereas the random utilities associated with provinces in different countries are independent. This means that an individual with strong preferences for a certain destination j is likely to also have stronger preferences for other destinations in the same region (or country) as destination j. The strength of this effect depends on how "homogeneous" the region/country is (i.e. how similar the provinces are that belong to this region/country).

Let λ_z and κ_r with $0 < \kappa_r, \lambda_z \leq 1$ measure the similarity of the provinces in country z and region r, respectively. These parameters govern the degree of substitutability across alternative migration destinations. High parameter values indicate little similarity among provinces (and weak correlations

among the random utilities), low parameter values indicate high similarity (and strong correlations). For $\kappa_r = \lambda_z = 1 \ \forall r, z$ our model collapses to the standard MNL model, which would rule out any correlation among the random utilities. We will very briefly return to this case below.

Our assumptions about the distribution of the random utility variables and about H_i are crucial for the model. We show in Appendix A what precise specification of H_i implies the three-level NMNL model, and how to use the function H_i in order to derive the following transition probabilities:

$$P_i^o(r \in A_z) = \exp[\Omega_{iz}\lambda_z - c_{iz} - \Psi_i], \tag{6}$$

$$P_i^o(j^o \in A_{zr} | r \in A_z) = \exp[\Phi_{ir}\kappa_r - c_{ir}/\lambda_z - \Omega_{iz}], \tag{7}$$

$$P_i^o(j^o = j | j^o \in A_{zr}) = \exp[\xi_{ij}/(\lambda_z\kappa_r) - \Phi_{ir}], \tag{8}$$

where Φ_{ir}, Ω_{iz}, and Ψ_i are "inclusive values" defined as:

$$\Phi_{ir} \equiv \ln \sum_{k \in A_{zr}} \exp[\xi_{ik}/(\lambda_z\kappa_r)], \tag{9}$$

$$\Omega_{iz} \equiv \ln \sum_{\ell \in A_z} \exp[\Phi_{i\ell}\kappa_\ell - c_{i\ell}/\lambda_z], \tag{10}$$

$$\Psi_i \equiv \ln \sum_z \exp[\Omega_{iz}\lambda_z - c_{iz}]. \tag{11}$$

The inclusive values Φ_{ir}, Ω_{iz}, and Ψ_i summarize the characteristics of all provinces in region r, all provinces in country z, and all provinces in the complete set of final migration destinations, respectively. Using equation (4) along with equations (6) through (11) and aggregating over all individuals from country i, we can write the expected rate of migration from country i to province j as:

$$\frac{m_{ij}}{m_i} = \frac{\exp[\xi_{ij}/(\lambda_z\kappa_r) - c_{ir}/\lambda_z - c_{iz}]}{\exp[\Psi_i + (1-\kappa_r)\Phi_{ir} + (1-\lambda_z)\Omega_{iz}]}, \tag{12}$$

where m_{ij} is the number of individuals migrating from i to j, and m_i is the initial population in country i. This migration rate depends on the inclusive values Φ_{ir}, Ω_{iz}, and Ψ_i, and is therefore responsive to the attractiveness of all provinces $k = 1, \ldots, J$, whether in the same region r (or the same country z) as province j or not. For example, consider the elasticity with respect to Y_k, the characteristics of province k, where $j \in A_{zr}, r \in A_z$, and $k \in A_{y\ell}, \ell \in A_y$. Straightforward though

cumbersome differentiation yields:

$$\begin{aligned}\frac{\partial \ln(m_{ij}/m_i)}{\partial \ln(Y_k)} &= Y_k \left[\frac{I(j,k)}{\lambda_z \kappa_r} - \left(\frac{m_{ik}}{m_i} \right) \right. \\ &\quad \left. - \frac{I(\ell,r)}{\lambda_z \kappa_r}(1-\kappa_r) \left(\frac{m_{ik}}{m_{ir}} \right) - \frac{I(y,z)}{\lambda_z}(1-\lambda_z) \left(\frac{m_{ik}}{m_{iz}} \right) \right], \end{aligned} \tag{13}$$

where $m_{ir} = \sum_{j \in A_{zr}} m_{ij}$, $m_{iz} = \sum_{r \in A_z} m_{ir}$, and $I(a,b) = 1$ if $a = b$ and zero otherwise. Given that $0 < \kappa_r, \lambda_z \leq 1$, this elasticity is positive for $k = j$ and negative for all other provinces $k \neq j$.

Any change in the conditions in some province $k \neq j$ induces *non-uniform* effects on the ij-specific migration rate, depending on whether this province belongs to the same country or region as province j. In particular, the elasticity in (13) is largest (in absolute value) for any change in the conditions in other destinations within the same region, $I(\ell,r) = I(y,z) = 1$. The fact that these substitution effects are strongest within regions and weakest across countries is due to the similarity of provinces in the same region (and in the same country). Contrast this result with the much simpler pattern of cross-elasticities implied by the standard MNL model where $\lambda_z = \kappa_r = 1 \ \forall r, z$: for $k \neq j$, the elasticity in (13) collapses to $\partial \ln(m_{ij}/m_i)/\partial \ln(Y_k) = -Y_k m_{ik}/m_i$ (independently of whether or not the provinces j and k share the same region or country). The migration rate in (12) similarly reduces to $\frac{m_{ij}}{m_i} = \frac{\exp[\xi_{ij} - c_{ir} - c_{iz}]}{\exp[\Psi_i]} = \frac{\exp[U_{ij}]}{\sum_k \exp[U_{ik}]}$, where the inclusive values Φ_{ir} and Ω_{iz} (but not Ψ_i) disappear.

2.2 Scale of Migration

Substituting ξ_{ij} in (12), taking logs, and rearranging terms yields the following migration function for $j \in A_{zr}$, $r \in A_z$:

$$\begin{aligned}\ln(m_{ij}) &= \frac{\theta}{\lambda_z \kappa_r} \ln(1 + M_{ij}) + \ln(m_i) + \frac{1}{\lambda_z \kappa_r} Y_j - c_{iz} - \frac{1}{\lambda_z} c_{ir} - \frac{1}{\lambda_z \kappa_r} c_{ij}, \\ &\quad - \Psi_i - (1-\lambda_z)\Omega_{iz} - (1-\kappa_r)\Phi_{ir}. \end{aligned} \tag{14}$$

Identification of the network effect is thus complicated by the presence of both the different cost components and the inclusive values. Moreover, the network coefficient (defined as $\eta_{zr} \equiv \eta(\lambda_z, \kappa_r) = \frac{\theta}{\lambda_z \kappa_r}$) is decreasing in both λ_z and κ_r. This means that more similar provinces are associated with larger network coefficients. Because migrants are more likely to substitute provinces for lower values of λ_z and κ_r, a given increase in the migrant network in province $k \in A_{zr}$ will, ceteris paribus, draw more migrants away from other provinces in country z, and in particular from other provinces in region $r \in A_z$.

2.3 Skill Composition of Migration

We now distinguish between high-skilled and low-skilled individuals, denoted by h and l, respectively. We augment the utility function by an inverse measure of ability to handle migration costs: $\gamma^s > 0, s \in \{h, l\}$:

$$U_{ij}^o = Y_j - \gamma^s C_{ij} + e_{ij}^o. \tag{15}$$

We assume $\gamma^h < \gamma^l$, so that the high-skilled individuals have lower effective migration costs than the low-skilled individuals. This assumption is in line with Chiswick (1999), who argues that the high-skilled can handle their migration process more efficiently than the low-skilled. We can thus derive one migration function for each skill group by complete analogy to equation (14). Combining the two migration functions yields:

$$\begin{aligned} \ln\left(\frac{m_{ij}^h}{m_{ij}^l}\right) &= \frac{\theta\gamma^*}{\lambda_z\kappa_r}\ln(1+M_{ij}) + \ln\left(\frac{m_i^h}{m_i^l}\right) - \gamma^* c_{iz} - \frac{\gamma^*}{\lambda_z}c_{ir} - \frac{\gamma^*}{\lambda_z\kappa_r}c_{ij} \\ &\quad -\Psi_i^* - (1-\lambda)\Omega_{iz}^* - (1-\kappa_r)\Phi_{ir}^*, \end{aligned} \tag{16}$$

where an asterisk (*) indicates differences between the corresponding parameters (or variables) for high-skilled and low-skilled individuals. Since $\gamma^* < 0$, the ratio of new high-skilled to low-skilled migrants is decreasing in the migrant network. This is due to the fact that individuals differ in their effective costs of migration, and that this difference is less important for low levels of migration costs. Hence, it is the low-skilled individuals who benefit the most from a reduction in migration costs through a larger migrant network.

3 Migration Data

Our empirical analysis investigates whether those destinations that have larger migrant networks to start with receive more migrants as well as a larger share of low-skilled migrants in subsequent years. Hence, the variation in the data we try to explain is cross-sectional: *across* different countries of origin, as well as *across* different provinces/regions of destination. This approach is in line with virtually all of the recent literature that tries to identify network effects in aggregate migration data; see Beine et al. (2011, 2015), Grogger & Hanson (2011), Beine & Salomone (2013), Neubecker & Smolka (2013), and Bertoli & Fernández-Huertas Moraga (2015).

The model for the scale of migration is estimated at the level of provinces in Spain, whereas due to reasons of data availability the model for the skill composition is estimated at the level of regions.[8]

[8]Spain consists of 52 provinces and 19 regions. We exclude the provinces of Ceuta and Melilla due to their specific geographical location, and thus end up with 50 provinces nested in 17 regions. See `http://www.ine.es/daco/`

The baseline sample we use for the scale model comprises the 55 most important countries of origin (all countries with at least 630 migrants in Spain in the year 1996). We choose this sample to cover those countries that are responsible for the lion's share of Spanish immigration, and to make sure we have sufficient cross-sectional variation. For the skill model we use, in principle, the same set of countries as for the scale model, but the actual estimation is carried out on 28 countries due to insufficient data for the dependent variable (the skill composition of migration). All migration data come from the Spanish Instituto Nacional de Estadística (INE); see Tables B.1 and B.2 in Appendix B for the list of countries and the different data sources, respectively.

3.1 Scale of Migration

The dependent variable in the scale model is the log of the province-level migration flow aggregated from the beginning of 1997 until the end of 2006.[9] We use a period of ten years in order to make our estimates comparable to the ones reported in the literature (e.g. Beine et al., 2011, 2015; Grogger & Hanson, 2011; and Beine & Parsons, 2015). The migrant network, M_{ij}, is given by the number of migrants included in the Spanish Municipal Register as of May 1, 1996.

From the year 2000 onwards, our data are likely to include both documented and undocumented migrants due to the incentives deriving from the *"Law on the Rights and Freedoms of Aliens in Spain and their Social Integration"* (*Ley Orgánica 4/2000, artículo 12*). This law became effective in 2000 and entitled all registered foreigners to free medical care under the same conditions as Spanish nationals, regardless of their legal status.[10] Each registrant must provide his or her name, surname, sex, usual domicile, nationality, passport number, as well as the place and date of birth.[11] As this information is confidential and must not be communicated to other administrative units, the probability of forced repatriation is plausibly independent of registration.

3.2 Skill Composition of Migration

Information on the skill composition of migration derives from the National Immigrant Survey 2007 (NIS). The survey gathers information on 15,465 migrants through field interviews conducted between

daco42/codmun/cod_provincia.htm and http://www.ine.es/daco/daco42/codmun/cod_ccaa.htm (both accessed on 04/17/2012) for a list of provinces and regions, respectively.

[9] Migrants are defined as individuals whose last country of residence (other than Spain) corresponds to their country of birth and nationality.

[10] As part of its austerity measures in 2012, the Spanish government has restricted this access to health care for undocumented migrants from September 2012 onwards. Exceptions are made for pregnant women and minors, as well as for cases of emergency care. (http://www.presseurop.eu/en/content/news-brief/2614611-no-more-free-treatment-undocumented-migrants based on http://elpais.com/elpais/2012/08/29/opinion/1346265472_538020.html, accessed on 08/31/2012).

[11] See INE at http://www.ine.es/en/metodologia/t20/t203024566_en.htm, accessed on 08/19/2011.

November 2006 and February 2007; see Reher & Requena (2009, 255-261).[12] Migrants report, *inter alia*, their year of arrival in Spain, their first destination in Spain, and their highest level of education they completed before migrating.[13] We aggregate the number of migrants by country of birth and region of destination, distinguishing between individuals with completed tertiary education before migrating (high-skilled) and all other individuals (low-skilled) and applying the provided population weights. Although the data can be considered representative of migrants who arrived shortly before the survey was taken, the numbers for earlier cohorts are less reliable due to the lack of information on migrants who died, returned, or migrated onward. We deal with the trade-off between a large number of individuals and data representativeness by considering only migrants who arrived between January 1, 2002, and December 31, 2006.[14]

A potential problem with the data is nevertheless the relatively small sample size of the NIS, which implies that we cannot compute the migrant skill ratio for a considerable share of country-region pairs. Most of the variation in the data comes from countries in South America: Argentina, Colombia, Bolivia, Ecuador, Cuba, Brazil, Venezuela, and Peru all have at least 11 regions in Spain for which data are available; see Section A of the Online Addendum to this paper. Other important countries are, for example, Romania (13 regions), Poland (9), and Morocco (9). Not surprisingly, these are also the countries that rank high in the overall incidence of migration to Spain over the period considered. The two most important regions with sufficient data are Cataluña (27 countries of origin) and Madrid (21). Overall, we believe, therefore, that the cross-sectional coverage of our data (in terms of both countries and regions) is rich enough to be meaningfully exploited in the type of regression analysis we consider here, but we will conduct a more formal analysis of this issue later.

4 Estimation and Results

We start with a descriptive look at the relationship between migrant networks and the scale and skill composition of migration. Figure 1(a) plots the migration flow between 1997 and 2006 against the migrant network in 1996, where each dot represents a different country-province pair. We observe a positive correlation between the two variables. Figure 1(b) plots the skill composition of migration between 2002 and 2006 against the migrant network at the beginning of 2002, where now each dot

[12]The sample was obtained through a relatively complex three-stage sampling scheme designed to offer reliable and representative data to policy makers and researchers. More detailed information on the sampling can be found in Reher & Requena (2009) as well as in INE (2007).

[13]They are defined as individuals aged 16 years or older who were born abroad and have lived in Spain for more than a year, or at least intended to stay for more than a year at the time the survey was conducted. Foreign-born individuals with Spanish nationality from birth who migrated to Spain within two years after birth are not considered as migrants. As this definition is independent of the individual's legal status, the data again include both documented and undocumented migrants.

[14]The migrant network in this model is measured as of January 1, 2002.

represents a different country–region pair. The figure suggests a weak negative correlation between the two variables. In the following we test whether these correlations reflect a causal relationship running from migrant networks to the scale and skill composition of migration, and we provide a structural interpretation of our estimation results in terms of our NMNL migration model. We also offer a robustness analysis.

<<Figures 1(a) and 1(b) about here>>

4.1 Results for the Scale of Migration

We first estimate the scale model in equation (14) with an *average* network coefficient that abstracts from heterogeneity in the similarity parameter κ_r. Table 1 shows the results from different fixed effects (FE) specifications. In columns (a) and (b), we control for both province fixed effects and country fixed effects (via an adequate within-transformation of the data). In columns (c) and (d) we also control for country-and-region fixed effects by modifying the within-transformation accordingly. And in columns (e) and (f) we additionally control for world region-and-province fixed effects. We discuss the different models and results in turn.

The country fixed effects in columns (a) and (b) control for all terms with subscripts i and iz (since our data refer to a single country of destination z). Hence, it controls for the population size of country i, the inclusive values Ψ_i and Ω_{iz}, and the cost term c_{iz} (which includes, for example, the impact of country-specific migration policies and the geographical and cultural distance between country i and Spain). The province fixed effects absorb the impact of the pull factors included in Y_j, as well as the province-specific migration costs included in c_{ij} (for example native attitudes toward migrants in different provinces). The number of observations in these specifications is equal to 2,593, which is the result of having 55 countries, 50 provinces, and 157 undefined values for the dependent variable due to zero migrant flows ($55 \times 50 - 157 = 2,593$). In column (a), the estimated network coefficient is equal to 0.689. The coefficient is statistically significant at the 1% level and estimated with very high precision (heteroskedasticity-robust standard error, clustered by countries, equal to 0.029). In column (b), we augment the model to include bilateral trade and capital flows. Both variables could be part of the cost term c_{ij}. Trade is not only facilitated by, but is also conducive to a good infrastructure for traveling and transportation. Capital invested by foreign firms (FDI) could create demand for specific types of labor, especially foreign labor.[15] The coefficient of the FDI variable is positive and significant, but the point estimate is rather small (0.012). Trade relations do not seem to have a significant effect.

[15]Trade data is available at the province level, while FDI data is only available at the regional level. We measure trade by the sum of exports and imports in 1996, and capital flows by gross FDI inflows in 1997 (earlier data are not available); see Table B.2 in Appendix B for the data sources.

The estimates of the network coefficient are virtually unchanged in this version of the model.

In columns (c) and (d), the country-and-region fixed effects control for all terms with subscript ir.[16] These include, first, the inclusive value Φ_{ir}, so that this estimation is fully compatible with our three-level NMNL model; and secondly, they include the cost term c_{ir} representing the geographical and cultural distance between country i and region j. Important elements of this distance derive from a cultural, political, and historical context. For example, the different regions in Spain feature considerable heterogeneity in terms of native languages; the Basque Autonomous Community and Navarre both have strong cultural ties with the Northern Basque Country, which is part of French national territory; the region of Galicia has long been suffering from a chronic growth weakness leading to mass emigration in the 19th and 20th century, in particular to Latin American countries. Relative to the estimates in columns (a) and (b), in this more demanding specification we see a drop in the estimated network coefficient down to 0.54, which corresponds to a decrease by roughly 20%.

The specification in columns (e) and (f) controls for additional sources of migration costs, namely those that are specific to both the province of destination and the world region of origin (grouping countries of origin). An example would be that individuals from Ecuador feel attracted not only by a network of migrants from Ecuador but also by a network of migrants from other Latin American countries; see Neubecker & Smolka (2013). This additional effect, a "cross-national" network externality, would reduce the migration costs for potential migrants from Ecuador and thus lead to a higher incidence of migration. Relative to columns (c) and (d), we see a reduction of the estimated network coefficient by more than 10% when we control for these effects.[17]

Unobserved heterogeneity in our model has two sources: first, the inclusive values, and secondly, the different cost components. Failing to account for the inclusive values leads to downward-biased estimates of the network coefficient due to a positive covariance between the migrant network and the terms Ψ_i, Ω_{iz}, and Φ_{ir}, respectively. Failing to account for the different cost components, in contrast, leads to upward-biased estimates due to a negative covariance between the migrant network and the terms c_{iz}, c_{ir}, and c_{ij}, respectively. Because our estimation results point towards a sizeable upward bias in the estimation of the network coefficient in specifications (a)-(d), the second source of unobserved heterogeneity clearly "dominates" the first one.

<<Tables 1 and 2 about here>>

In case we omit ij-specific variables that are correlated with both m_{ij} and M_{ij}, the migrant network

[16] This approach excludes all country-and-region pairs that have no within variation (for example due to regions that consist of one province), and thus reduces the number of observations to 2,200.

[17] In terms of world regions, we distinguish between East Asia & Pacific; Eastern Europe & Central Asia; Latin America & Caribbean; Middle East & North Africa; North America, Australia & New Zealand; South & South-East Asia; Sub-Saharan Africa; and Western Europe.

is endogenous to the subsequent migrant flow. In view of our most ambitious FE specification, it is not easy to think of any such omitted variable. However, suppose there is a province-specific labor demand for workers from a certain nationality, such as the demand for German engineers in SEAT's car production in Barcelona. Then, the FE model may produce biased and inconsistent estimates. Consistent estimation would call for an instrument that is uncorrelated with the structural error term but correlated with the endogenous regressor. We instrument the migrant network with historical internal migration flows in Spain, defined as the log of the number of people holding country i's nationality and migrating from province j to any other province $k \neq j$ in Spain in 1988 (henceforth simply called internal migration).[18]

Because it indicates a large historical network, internal migration can be expected to correlate positively with the migrant network in 1996.[19] Our first-stage regressions attest to a statistically significant positive (partial) correlation. Its significance is also reflected in relatively high values for the first-stage F statistics. For internal migration to be a valid instrument, it must be uncorrelated with the structural error term.[20] This assumption could be violated if a large internal migration observed for a certain province reflects and signals a poor matching quality (for example in terms of jobs) between this province and the corresponding migrants, thus leading to a lower incidence of migration today. However, this signaling effect does not necessarily render our instruments endogenous. One reason is that most, if not all, of the variation in the matching quality across countries and provinces is absorbed into our fixed effects. Another, probably more important, reason is that the signaling effect should be captured by the (observable) migrant network itself, given that this network is a function of all past migration flows. We use internal migration in 1989 as a second excluded instrument. This allows us to perform tests on overidentifying restrictions and check for instrument exogeneity.

The 2SLS FE estimations in Table 2 strengthen our interpretation of a quantitatively important causal effect of migrant networks on the scale of migration. They suggest a somewhat larger role for the network effect, with a coefficient ranging between 0.718 and 0.955. The difference between the FE estimates and the 2SLS FE estimates could be due to stochastic measurement errors in the migrant network, which would result in downward-biased estimates of the network coefficient when applying the FE estimator; see Hausman (2001). As in the FE estimations, the network coefficient is smallest when controlling for country-and-region effects as well as for world region-and-province effects. The loss in precision from using the 2SLS FE approach is fairly small when interpreted against the FE

[18]The year 1988 is the first year for which this information is available. It is well before the start of the Spanish migration boom. We add one to the number of people before taking logs in order to keep observations with zero migration flows.

[19]It follows from its definition, however, that internal migration also reduces the size of the historical network.

[20]Therefore, the focus on *internal* migration is on purpose because it excludes return migrants who could shape future migration in one way or the other.

model. The effects of both trade and FDI are essentially zero.

Next we allow for differences in the similarity parameter κ_r, which implies region-specific network coefficients, η_{zr}. The specification employed is equivalent to the one in column (f) of Table 1, except that we now interact the migrant network with dummy variables for the different regions. Table 3 reveals substantial heterogeneity in the estimated network coefficient. It is largest for Cataluña (0.795) and smallest for Extremadura (0.155).[21] Hence, the provinces in Cataluña (Barcelona, Girona, Lleida, and Tarragona) are easier to substitute for one another than the ones in Extremadura (Badajoz and Cáceres). This is consistent with the idea that a region that puts a lot of emphasis on the independence of its political and cultural life is perceived as rather homogenous when compared to other regions. It is thus not surprising that two other regions with a second official language, Comunitat Valenciana and Galicia, rank next to Cataluña in terms of the size of the estimated network coefficient. For the Basque country, however, we find a surprisingly low network coefficient (equal to 0.287), which indicates that migrants view the provinces in this region as rather heterogeneous. At any rate, the large and significant cross-regional differences in the estimated network coefficient suggest that the assumption of a uniform degree of cross-destination substitutability featured in the standard MNL model is too restrictive to be plausible in the Spanish case.

<<Table 3 about here>>

The estimated network coefficients can be used to compute both the network elasticity of migration as well as the cross-elasticities of the network defined as:

$$\begin{aligned}\frac{\partial \ln(m_{ij})}{\partial \ln(1+M_{ik})} &= \theta\left[\frac{I(j,k)}{\lambda_z\kappa_r}-\left(\frac{m_{ik}}{m_i}\right)\right.\\ &\left.-\frac{I(\ell,r)}{\lambda_z\kappa_r}(1-\kappa_r)\left(\frac{m_{ik}}{m_{ir}}\right)-\frac{I(y,z)}{\lambda_z}(1-\lambda_z)\left(\frac{m_{ik}}{m_{iz}}\right)\right]. \qquad (17)\end{aligned}$$

The network elasticity ($j=k$) is a function of the network parameter θ, the similarity parameters κ_r and λ_z, and the relative attractiveness of province j (reflected by the shares m_{ij}/m_i, m_{ij}/m_{ir}, and m_{ij}/m_{iz}). Neither κ_r nor λ_z can be estimated directly due to the use of aggregate data. This implies an uncertainty about the true network elasticity.[22] However, we can compute estimates of the upper and lower bounds for this elasticity, separately for each region. For this purpose, we use the results in

[21] In the estimation, Cataluña serves as the reference region. The differences between the network coefficients of Cataluña and either of the other regions (except for Comunitat Valenciana and Canarias) are statistically significant at least at the 10% level according to t-tests.

[22] Schmidheiny & Brülhart (2011) discuss a related type of uncertainty in a two-level NMNL model. They show that the Poisson model and the standard MNL model are the polar cases of a two-level NMNL model with two nests, one being a degenerate nest with a single alternative, and the other one featuring many alternatives with a single similarity parameter $\lambda \in (0,1)$. When λ is unknown, the elasticities of the Poisson model and of the standard MNL model can thus serve as boundary values for the true elasticities.

Table 3 to compute estimates of the ratio $\kappa_r/\kappa_\ell = \eta_{z\ell}/\eta_{zr}, \forall r, \ell \in A_z$. Since the region of Extremadura features the lowest estimated network coefficient, its similarity parameter κ_r can take on any value between zero and one, while the similarity parameters for all other regions must be strictly lower than one. For example, the range of permissible similarity parameter values for the region of Cataluña runs from zero to 0.195 (= 0.155/0.795).

Figure 2(a) shows counterfactual network elasticities as a function of κ_r (Extremadura). The exact value of κ_r is unknown, but fixing this parameter also fixes the similarity parameters of all other regions. In order to focus on the heterogeneity in the network elasticity stemming from differences in the similarity parameters, we have imposed the following assumptions: first, there are 200 countries of destination outside the country of origin i; second, all countries consist of 51 provinces that are uniformly distributed across 17 regions; and third, all foreign provinces are equally attractive destinations, with an overall share of migrants equal to three percent, $\sum_{j\neq i} m_{ij}/m_i = 0.03$. These assumptions imply: $m_{ij}/m_i = 1/340,000$, $m_{ij}/m_{ir} = 1/3$, and $m_{ij}/m_{iz} = 1/51$. For Extremadura, we thus find a network elasticity slightly above 0.1. For Cataluña, the elasticity lies in the vicinity of 0.55.[23] These are quite large differences. For any given region, the difference between the upper and the lower bound of the network elasticity is roughly equal to 0.05, so the uncertainty is a minor issue here. Importantly, the figure also incorporates the uncertainty about the country-specific similarity parameter λ_z (namely through the thickness of the upward-sloping lines). This uncertainty, however, turns out to be almost irrelevant for the computation of the network elasticity.

<<Figures 2(a) and 2(b) about here>>

We have also computed the cross-elasticities of the network based on (17). Figure 2(b) depicts cross-elasticities for destinations in the same region. For Extremadura, we find avalue between 0.0 and −0.05, for Cataluña between −0.22 and −0.27. In Section B of the Online Addendum to this paper, we also depict the cross-elasticities for destinations in different regions and countries, respectively. These are smaller (in absolute value) than the cross-elasticities depicted in Figure 2(b), and they are characterized by a higher uncertainty about their true values.

[23] The configuration underlying these elasticities is of course quite unrealistic given the pull of the advanced economies versus developing economies. However, the elasticities are pretty robust to alternative configurations. They hardly change at all when we, for example, drastically increase the relative attractiveness of Spain by setting $m_{ij}/m_i = 1/1,000$ and leave the other parameters unchanged. More generally, from equation (17) we see that the elasticity is larger for less attractive provinces. The upper bound is found by setting m_{ij} equal to zero and is thus simply the estimated network coefficient (equal to 0.155 for Extremadura and 0.795 for Cataluña).

Robustness Analysis

We now describe a series of robustness checks, but relegate more detailed results to Section C of the Online Addendum to this paper. A first issue is that the migration rate m_{ij}/m_i which we use to derive the estimating equation (14) is an *expected* migration rate, with the *actual* migration rate depending on the realizations of the random utility variables for all individuals $(e^o_{i1}, \ldots, e^o_{iJ})$. In the presence of heteroskedasticity in the stochastic deviations of the expected migration rate from the true migration rate, our FE estimator yields inconsistent estimates. To handle the same problem in the gravity model of international trade, Santos Silva & Tenreyro (2006) propose to estimate the equation in levels, using the Poisson pseudo-maximum-likelihood (PPML) estimator. The estimator is consistent even in the presence of heteroskedastic errors in the level equation and has the additional advantage that zero migration flows can be included in the estimation. We find in regressions based on the PPML model that our baseline estimations lead to a certain underestimation of the average network coefficient. This corroborates our results from the 2SLS FE estimator and indicates that the FE estimates should be interpreted as a lower bound for the true size of the network effect.

We obtain further evidence in this direction by considering different time periods in our analysis. As argued above, we have chosen to aggregate the migration flow over a ten-year period in our baseline estimations, in order to make our estimates comparable to those provided in the literature. The underlying assumption is that changes in the relative attractiveness of migration destinations over the period considered are not material for the estimation. What happens if we shorten or extend the period of aggregation? Intuitively, the initial allocation of migrants should be most relevant for the early movers, and gradually fade out as we move on in time. This is exactly what we find in the data when we start with a model that includes the migration flow in the year 1997 as the dependent variable and incrementally extend the period of aggregation by one year. The results indicate an almost monotonic decline in the estimated network coefficient up to the year 2007 as we move away in time from the initial network allocation in 1997.[24]

In a further robustness check, we apply alternative sample selection criteria in order to see whether our results suffer from endogenous sample selection. In particular, we consider all observations (country-province pairs) with a migrant network of more than either 10, 20, or 50 migrants in the year 1996. Applying these criteria results in unbalanced samples of 98, 90, or 74 countries, respectively. Again, the results we obtain (not reported) indicate a slightly larger average network coefficient than do our baseline estimates.

A further concern might be a potential estimation bias due to non-stochastic measurement errors

[24] We find a similar pattern when we change the starting dates of our analysis and measure the migrant network in years after 1997 (and shorten the period to compute the migration flow accordingly).

in our migration data. In our baseline estimations we cover the period 1997-2006. To the extent that undocumented migrants arrived in or before 1996 and registered in later years (especially due to the *Ley Orgánica 4/2000* in 2000), we understate the true size of the migrant network in 1996 and overstate the true size of the migrant flow over the period 1997-2006. We show in Appendix C that our extended FE specification is entirely immune to both types of measurement errors under a relatively mild assumption, namely that the ratio of "mismeasured" to observed migrants is constant within country-region pairs.

4.2 Results for the Skill Composition of Migration

Reliable information about the skill composition of migration is only available at the level of regions, not provinces. We deal with this issue in two different ways. First, we simplify the structure of our model to a two-level NMNL model in which the regions rather than the provinces serve as the final migration destinations. This approach is straightforward and the one we emphasize in the following. Secondly, we develop an estimation strategy at the regional level that is fully consistent with the three-level NMNL model presented above. This approach is offered in our robustness analysis.

Defining regions (indexed here by j) as the final migration destination, we re-write equation (16) as:

$$\ln\left(\frac{m_{ij}^h}{m_{ij}^l}\right) = \frac{\theta\gamma^*}{\lambda_z}\ln(1+M_{ij}) + \ln\left(\frac{m_i^h}{m_i^l}\right) - \gamma^* c_{iz} - \frac{\gamma^*}{\lambda_z}c_{ij} - \Psi_i^* - (1-\lambda_z)\Omega_{iz}^*. \tag{18}$$

Table 4 reports the results from FE estimations of variants of this equation. The full data matrix would contain 935 pairs of 55 countries and 17 regions, but, as we have explained above, the migrant skill ratio is missing for a considerable share of observations. Moreover, due to the inclusion of country fixed effects in all specifications, identification requires that each country has at least two regions with non-missing values for the migrant skill ratio. For these reasons, the total number of observations in the baseline estimations is 234 in all columns (a) through (f).

Because we exploit variation in the data across regions within countries of origin, we cannot control for country-and-region fixed effects. Instead, we augment the model by observable variables that are likely to influence the migration costs. In addition to trade and FDI flows[25], we control for the geographical distance between country i and region j using the STATA module GEODIST by Picard (2010) in combining geographical data from Mayer & Zignago (2006) and the Spanish Wikipedia/GeoHack webpage. We also control for a common language through a dummy variable that equals one if at least 80% of region j's total population are native speakers of a language spoken

[25]Regional-level trade data refer to the year 2001. Gross FDI inflows are aggregated from the beginning of 1998 until the end of 2001.

by at least 20% of the people living in country i. The information about native languages in Spain is taken from a number of recent survey studies; see Table B.2 in Appendix B. Language information about the countries of origin comes from Mayer & Zignago (2006). The influence of all terms indexed j is absorbed by a set of dummy variables for the different regions. The complete specification of our model furthermore controls for world region-and-region fixed effects.

In all the specifications employed in Table 4, we find a robustly significant negative effect of migrant networks on the skill composition of migration. The estimated coefficient varies between -0.506 and -0.637, so the differences across specifications are rather small. Neither trade nor FDI turns out to be statistically significant. Maybe surprisingly, the effects of a common language and geographical proximity are often estimated to be zero and have an unexpected sign, but one should keep in mind here that identification comes only from variation across regions within countries of origin.

<<Tables 4 and 5 about here>>

We also apply the instrumental variables approach to this model. In particular, we instrument the migrant network with the log of the number of people holding country i's nationality and migrating from region j to any other region $k \neq j$ in Spain in 1988. As before, we use the corresponding migration flow in 1989 as a second excluded instrument. The results reported in Table 5 suggest a causal interpretation of the network effect on the skill composition of migration.[26] In all the specifications considered, the estimated coefficient of the migrant network is negative and statistically significant at the 5% level. The point estimates range from -0.534 to -1.105 and are thus smaller (in absolute value) than those obtained from the FE estimations. In the full specification of the model in column (f), the migrant network is the only structural explanatory variable that has a statistically significant effect.

In order to interpret our results in terms of elasticities, we compute:

$$\frac{\partial \ln(m_{ij}^h/m_{ij}^l)}{\partial \ln(1+M_{ij})} = \theta\gamma^* \left[\frac{1}{\lambda_z} - \left(\frac{m_{ij}}{m_i}\right) - \frac{1-\lambda_z}{\lambda_z}\left(\frac{m_{ij}}{m_{iz}}\right)\right], \tag{19}$$

where we have assumed, for simplicity, that $m_{ij}/m_i = m_{ij}^h/m_i^h = m_{ij}^l/m_i^l$ and $m_{ij}/m_{iz} = m_{ij}^h/m_{iz}^h = m_{ij}^l/m_{iz}^l$. We assume, as before, that there are 200 countries of destination outside the country of origin i; that each of these countries consists of 17 regions; and that all regions abroad are equally attractive destinations, with an overall share of migrants equal to three percent.[27] Then, because the similarity parameter λ_z can take on any value between zero and one, an estimated coefficient of the

[26] The first-stage F test suggests that our instruments are relevant in specifications (a), (b), and (c), but that they might be weak in specifications (d), (e), and (f).

[27] This implies that $m_{ij}/m_i = 3/340,000$ and $m_{ij}/m_{iz} = 1/17$.

migrant network equal to -0.621 (as in column (f) of Table 4) implies that the corresponding elasticity lies between -0.621 and -0.584.

Robustness Analysis

We have checked the robustness of these results and the validity of some underlying assumptions in various ways. A first concern is measurement error in the dependent variable due to the small sample size of the underlying survey data. Section A of the Online Addendum to this paper shows that many observations in our baseline sample come from country-region pairs for which the survey records few respondents. We must therefore expect the skill ratio to be mismeasured for a sizeable fraction of observations. However, while stochastic measurement error in the dependent variable leads to less precisely estimated coefficients, it does not lead to inconsistent estimates; see Hausman (2001). Hence, we believe that the small sample size of the underlying survey data does not per se invalidate our estimates.

Yet, we perform additional regressions on restricted estimation samples. For example, when we restrict the sample to observations for which the skill ratio is constructed on the basis of at least 15 migrants in the survey, the sample reduces to 75 observations, but the estimates continue to reflect a highly significant negative effect of the migrant network on the skill composition of migration. In this particular example, we obtain an estimated coefficient of -0.519 (with a 95% confidence interval of $[-.232; -.772]$) for a specification that resembles column (f) in Table 4. This estimate is not distinguishable (in a statistical sense) from the estimate we find based on the unrestricted sample. We obtain similar results when we employ a threshold of 10 migrants in the survey (rather than 15 migrants), which results in a sample of 105 observations.

A related issue is a potential sample selection bias that could be due to the large number of missing values for the migrant skill ratio. In order to investigate this issue, we develop a Heckman (1976)-style procedure similar to the one proposed by Wooldridge (1995, 123-124). We describe this procedure in detail in Appendix D. We find no evidence for sample selection bias in our analysis.

Next, we estimate the model with the PPML estimator rather than the FE estimator, and find strong evidence for a negative and significant skill effect of migrant networks also with this alternative estimator; see Table D.1 in the Online Addendum to this paper.

Moreover, following the methodology proposed by Grogger & Hanson (2011, 53-54), we exclude the possibility that individuals group regions of destination into another layer of nests at the sub-country level. To do so, we repeatedly estimate the scale model as given by equation (14), using regional data instead of province-level data and each time excluding the observations for one region. The estimated

network coefficient is very stable across regressions, ranging from 0.665 to 0.719.

Finally, we estimate a migration function that describes migration into regions of destination but derives from the three-level NMNL model featuring provinces as the final migration destinations. The starting point is to use equations (6) and (7) in order to compute the probability $P_i^o(j^o \in A_{zr}) = P_i^o(j^o \in A_{zr}|r \in A_z)P_i^o(r \in A_z)$, separately for each skill group. It is easy to show that this alternative migration function depends, among other things, on the number of provinces in each region and on the within-region distribution of migrant networks across provinces. This last argument is part of a highly non-linear term, which collapses to zero if we look at regions that consist of a single province. Hence, we have estimated the model excluding all regions that consist of more than one province. In spite of the reduced number of observations, our estimates continue to reflect a negative and statistically significant impact of migrant networks on the skill composition of migration.[28]

5 Conclusion

We have studied empirically how the settlement pattern of previous migrants influences the size and skill composition of subsequent migration flows. Using detailed and comprehensive data on a recent migration boom to Spain, we have found strong positive effects of the number of already resident migrants on the scale of migration, and a strong negative effect on the ratio of high-skilled to low-skilled migrants. Both effects prove robust to the use of different estimators, estimation samples, and sets of control variables.

We conclude the paper with two observations. The first is methodological: The random utility framework we use in our estimation postulates that individuals are more likely to substitute destinations within, rather than across, countries and regions. In addition, it allows for some countries and regions to be more homogeneous than others, a feature that is important in the Spanish context, where some regions have a very specific cultural and political background. While we do not offer an independent test of the structure we impose on the model, the hierarchy in terms of countries, regions, and provinces seems very plausible and opens up an interesting perspective on how to interpret estimates of migration models based on aggregate data. Our approach finds support in significant cross-regional differences in the estimated network elasticity, and commands a careful specification of substitution effects in future work.

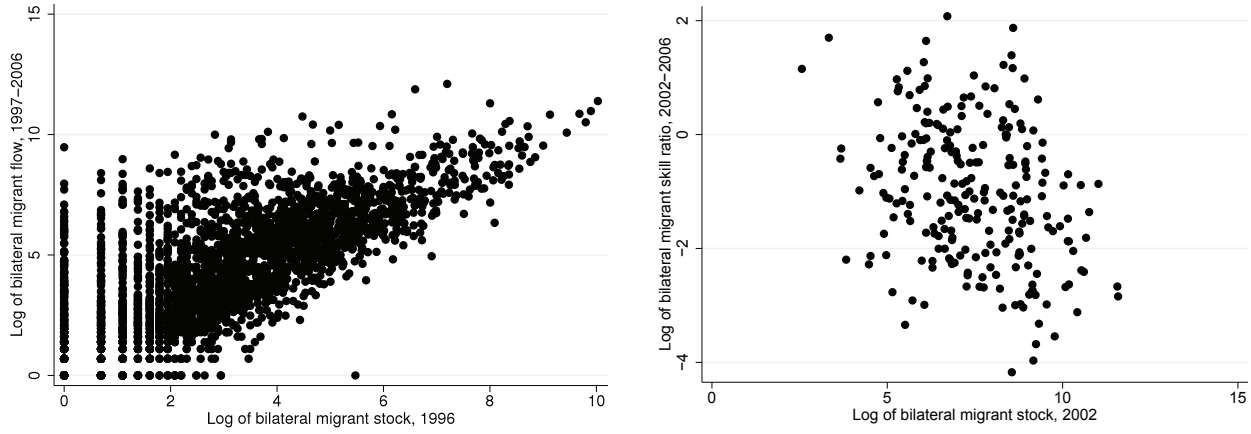
The second observation concerns the phenomenon of origin-specific "clustering" (i.e. a strong ethnic concentration of migrants). Network effects have the potential to generate two types of clus-

[28] We have also experimented with two alternative estimation approaches following Quigley (1976) and Lerman (1976). Both include the full set of regions in Spain and are summarized in McFadden (1978, 91-94). Again, we have obtained a robustly significant, negative impact of migrant networks on the skill composition of migration.

tering. The first type is *spatial* clustering, because later migrants follow the locational decisions of previous migrants. This is consistent with the evidence in Spain. Of the migrants from Ecuador who arrived between 1997 and 2009, for example, a remarkable 42.5% chose to reside in Madrid, while the same number for migrants from Morocco is just 12%[29] (the share of native residents in Madrid was about 20% at the end of the 1990s). The second type of clustering is *sectoral* and *occupational* clustering, meaning that migrants end up in the same sectors of activity and/or occupations as their co-ethnic peers. This is also consistent with the evidence in Spain. For example, in 2007 40.9% of male migrants from Latin America were employed within the construction sector, and 10.4% within accommodation and catering. For migrants from Asia the numbers are strikingly different (16.1% and 29%, respectively), as are the ones for natives (18.1% and 4.7%, respectively); see Table 5 in del Río & Alonso-Villar (2012). As previously noted by Patel & Vella (2013) in the context of immigration to the U.S., the extent to which network effects generate clustering, as well as the exact channels through which this happens and the wider economic implications of this phenomenon, are interesting areas for future research.

Figures and Tables

Figure 1: Migrant Networks and the Scale and Skill Composition of Migration

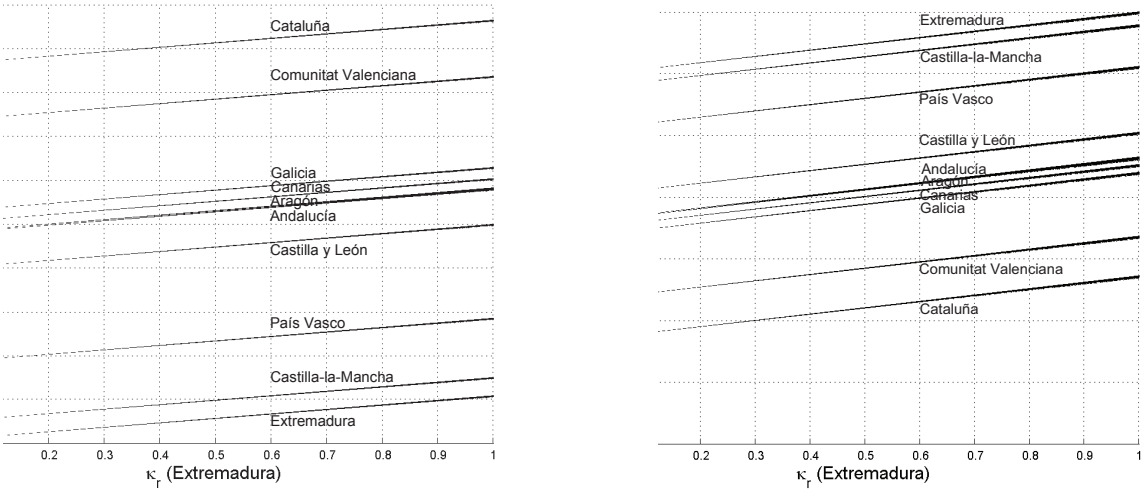


(a) $\ln(m_{ij})$ plotted against $\ln(1+M_{ij})$, provincial level

(b) $\ln(m^h_{ij}/m^l_{ij})$ plotted against $\ln(1+M_{ij})$, regional level

[29]The primary destination of migrants from Morocco was Barcelona instead.

Figure 2: Counterfactual Network Elasticities and Cross-elasticities



(a) Network Elasticities

(b) Cross-elasticities for $j, k \in A_{zr}$

Table 1: Scale of Migration – FE Model†

	Dependent Variable: Migration Flow (Province-Level 1997-2006)					
	(a)	(b)	(c)	(d)	(e)	(f)
Stock of Migrants	0.689***	0.683***	0.540***	0.541***	0.470***	0.470***
(Province-Level 1996)	(0.029)	(0.029)	(0.029)	(0.029)	(0.033)	(0.033)
FDI Flow		0.012**				
(Region-Level 1997)		(0.005)				
Trade Flow		0.005		0.003		0.008
(Province-Level 1996)		(0.007)		(0.007)		(0.007)
Province Effects	Yes	Yes	Yes	Yes	Nested	Nested
Country Effects	Yes	Yes	Nested	Nested	Nested	Nested
Country-and-Region Effects	No	No	Yes	Yes	Yes	Yes
World Region-and-Province Effects	No	No	No	No	Yes	Yes
Observations	2,593	2,593	2,200	2,200	2,200	2,200
Within R2	0.791	0.792	0.669	0.669	0.763	0.764

†All variables are in natural logs. Heteroskedasticity-robust standard errors (clustered by countries or country-region pairs) are given in parentheses. *,**,*** denote significance at the 10%, 5%, 1% levels, respectively. The regressions include all countries with at least 630 nationals residing in Spain in 1996 (55 countries). See Section 3 for a detailed description of all variables.

Table 2: Scale of Migration – 2SLS FE Model†

	Dependent Variable: Migration Flow (Province-Level 1997-2006)					
	(a)	(b)	(c)	(d)	(e)	(f)
Stock of Migrants	0.955***	0.952***	0.823***	0.825***	0.718***	0.721***
(Province-Level 1996)	(0.069)	(0.070)	(0.080)	(0.080)	(0.101)	(0.101)
FDI Flow		0.004				
(Region-Level 1997)		(0.005)				
Trade Flow		0.004		0.005		0.008
(Province-Level 1996)		(0.007)		(0.008)		(0.007)
Province Effects	Yes	Yes	Yes	Yes	Nested	Nested
Country Effects	Yes	Yes	Nested	Nested	Nested	Nested
Country-and-Region Effects	No	No	Yes	Yes	Yes	Yes
World Region-and-Province Effects	No	No	No	No	Yes	Yes
Observations	2,593	2,593	2,200	2,200	2,200	2,200
Within R2	0.770	0.770	0.635	0.635	0.745	0.745
Robust first stage F	31.44	30.73	18.57	18.51	12.82	12.79
Hansen J test	0.0128	0.0194	0.420	0.384		
Hansen J test p-value	0.910	0.889	0.517	0.535		
Endogeneity test	14.52	14.21	10.86	10.93		
Endogeneity test p-value	0.000	0.000	0.001	0.001		
Kleibergen-Paap LM test	20.23	20.15	24.32	24.30	20.35	20.33
Kleibergen-Paap LM p-value	0.000	0.000	0.000	0.000	0.000	0.000

†All variables are in natural logs. Heteroskedasticity-robust standard errors (clustered by countries or country-region pairs) are given in parentheses. *,**,*** denote significance at the 10%, 5%, 1% levels, respectively. The regressions include all countries with at least 630 nationals residing in Spain in 1996 (55 countries). The (log) stock of migrants in 1996 is instrumented with the (log) migration flows of foreign nationals within Spain in 1988 and in 1989. See Section 3 for a detailed description of all variables.

Table 3: Region-specific Network Coefficients†

Region r	Estimate of η_{zr}	Region r	Estimate of η_{zr}
Cataluña	0.795	Andalucía	0.507
Comunitat Valenciana	0.699	Castilla y León	0.447
Galicia	0.544	País Vasco	0.287
Canarias	0.525	Castilla-La Mancha	0.186
Aragón	0.509	Extremadura	0.155

†This table reports region-specific estimates of the network coefficient, η_r. The specification employed is equivalent to the one reported in column (f) of Table 1, except that we interact the migrant network with dummy variables for the different regions of destination. F tests reveal that each of the above-reported network coefficients—with the exception of the one for Extremadura—is significant at least at the 5% level. The number of observations is 2,200 and the within R^2 is 0.771.

Table 4: Skill Composition of Migration – FE Model†

	Dependent Variable: Migrant Skill Ratio (Region-Level 2002-2006)					
	(a)	(b)	(c)	(d)	(e)	(f)
Stock of Migrants	-0.513***	-0.510***	-0.506***	-0.626***	-0.637***	-0.621***
(Region-Level 2002)	(0.085)	(0.084)	(0.087)	(0.094)	(0.090)	(0.098)
FDI Flow			-0.006			-0.012
(Region-Level 1998-2001)			(0.019)			(0.015)
Trade Flow			-0.001			0.080
(Region-Level 2001)			(0.079)			(0.095)
Language		0.248	0.246		0.463***	0.559***
(Region-Level)		(0.209)	(0.210)		(0.149)	(0.131)
Distance		-0.636*	-0.657*		-1.450	-1.388
(Region-Level)		(0.373)	(0.369)		(1.159)	(1.148)
Region Effects	Yes	Yes	Yes	Nested	Nested	Nested
Country Effects	Yes	Yes	Yes	Yes	Yes	Yes
World Region-and-Region Effects	No	No	No	Yes	Yes	Yes
Observations	234	234	234	234	234	234
Within R2	0.245	0.261	0.261	0.466	0.477	0.481

†All variables except for the language dummy are in natural logs. Heteroskedasticity-robust standard errors (clustered by countries) are given in parentheses. *,**,*** denote significance at the 10%, 5%, 1% levels, respectively. See Section 3 for a detailed description of all variables.

Table 5: Skill Composition of Migration – 2SLS FE Model†

	Dependent Variable: Migrant Skill Ratio (Region-Level 2002-2006)					
	(a)	(b)	(c)	(d)	(e)	(f)
Stock of Migrants	-0.534**	-0.550***	-0.549***	-1.002***	-1.089***	-1.105***
(Region-Level 2002)	(0.210)	(0.210)	(0.210)	(0.380)	(0.404)	(0.422)
FDI Flow			-0.005			0.009
(Region-Level 1998-2001)			(0.021)			(0.029)
Trade Flow			0.004			0.077
(Region-Level 2001)			(0.080)			(0.101)
Language		0.244	0.243		0.325*	0.344
(Region-Level)		(0.205)	(0.206)		(0.178)	(0.215)
Distance		-0.637*	-0.649*		-1.877	-1.795
(Region-Level)		(0.371)	(0.366)		(1.192)	(1.167)
Region Effects	Yes	Yes	Yes	Nested	Nested	Nested
Country Effects	Yes	Yes	Yes	Yes	Yes	Yes
World Region-and-Region Effects	No	No	No	Yes	Yes	Yes
Observations	234	234	234	234	234	234
Within R2	0.245	0.260	0.260	0.419	0.411	0.409
Robust first stage F	13.93	12.40	11.62	5.961	6.119	5.210
Hansen J test	0.863	0.613	0.673			
Hansen J test p-value	0.353	0.434	0.412			
Endogeneity test	0.110	0.146	0.177			
Endogeneity test p-value	0.740	0.702	0.674			
Kleibergen-Paap LM test	11.41	10.85	10.42	8.258	8.520	7.860
Kleibergen-Paap LM p-value	0.003	0.004	0.005	0.016	0.014	0.020

†All variables except for the language dummy are in natural logs. Heteroskedasticity-robust standard errors (clustered by countries) are given in parentheses. *,**,*** denote significance at the 10%, 5%, 1% levels, respectively. The (log) stock of migrants in 2002 is instrumented with the (log) migration flows of foreign nationals within Spain in 1988 and in 1989. See Section 3 for a detailed description of all variables.

References

[1] Åslund, Olof, "Now and forever? Initial and Subsequent Location Choices of Immigrants," *Regional Science and Urban Economics*, 35:2 (2005), 141–165.

[2] Baghdadi, Leila, "Mexico-U.S. Migration: Do Spatial Networks Matter?," Universite Paris I. mimeograph (2005).

[3] Bartel, Ann P., "Where Do the New U.S. Immigrants Live?," *Journal of Labor Economics*, 7:4 (1989), 371–391.

[4] Bauer, Thomas, Gil S. Epstein, and Ira N. Gang, "Enclaves, Language, and the Location Choice of Migrants," *Journal of Population Economics*, 18:4 (2005), 649–662.

[5] Bauer, Thomas, Gil S. Epstein, and Ira N. Gang, "Measuring Ethnic Linkages among Migrants," *International Journal of Manpower*, 30:1/2 (2009), 56-69.

[6] Beine, Michel, Frédéric Docquier, and Çağlar Özden, "Diasporas," *Journal of Development Economics*, 95:1 (2011), 30–41.

[7] Beine, Michel, Frédéric Docquier, and Çağlar Özden, "Dissecting Network Externalities in International Migration," *Journal of Demographic Economics*, 81:4 (2015), 379–408.

[8] Beine, Michel, and Christopher Parsons, "Climatic Factors as Determinants of International Migration," *Scandinavian Journal of Economics*, 117:2 (2015), 723–767.

[9] Beine, Michel, and Sara Salomone, "Network Effects in International Migration: Education versus Gender," *Scandinavian Journal of Economics*, 115:2 (2013), 354–380.

[10] Bertoli, Simone, "Networks, Sorting and Self-selection of Ecuadorian Migrants," *Annals of Economics and Statistics*, 2010:97/98 (2010), 261–288.

[11] Bertoli, Simone, and Jesús Fernández-Huertas Moraga, "Multilateral Resistance to Migration," *Journal of Development Economics*, 102 (2013), 79–100.

[12] Bertoli, Simone, and Jesús Fernández-Huertas Moraga, "The Size of the Cliff at the Border," *Regional Science and Urban Economics*, 51 (2015), 1-6.

[13] Card, David and Ethan G. Lewis, "The Diffusion of Mexican Immigrants During the 1990s: Explanations and Impacts," in George J. Borjas (Ed.), *Mexican Immigration to the United States* (Chicago: University of Chicago Press, 2007), chapter 6, 193–227.

[14] Carrington, William J., Enrica Detragiache, and Tara Vishwanath, "Migration with Endogenous Moving Costs," *American Economic Review*, 86:4 (1996), 909–930.

[15] Chiswick, Barry R., "Are Immigrants Favorably Self-Selected?," *American Economic Review: Papers and Proceedings*, 89:2 (1999), 181–185.

[16] Chiswick, Barry R., and Paul W. Miller, "Where Immigrants Settle in the United States," *Journal of Comparative Policy Analysis*, 6:2 (2004), 185–197.

[17] Clark, Ximena, Timothy J. Hatton, and Jeffrey G. Williamson, "Explaining U.S. Immigration, 1971- 1998," *Review of Economics and Statistics*, 89:2 (2007), 359–373.

[18] del Río, Coral, and Olga Alonso-Villar, "Occupational Segregation of Immigrant Women in Spain," *Feminist Economics*, 18:2 (2012), 91–123.

[19] Dolfin, Sarah, and Garance Genicot, "What Do Networks Do? The Role of Networks on Migration and 'Coyote' Use," *Review of Development Economics*, 14:2 (2010), 343–359.

[20] Domencich, Thomas A., and Daniel L. McFadden, *Urban Travel Demand: A Behavioral Analysis* (North Holland: Amsterdam, 1975. Reprinted by The Blackstone Company: Mount Pleasant, MI, 1996.)

[21] Dustmann, Christian, and Ian P. Preston, "Racial and Economic Factors in Attitudes to Immigration," *The B.E. Journal of Economic Analysis & Policy*, 7:1 (Advances), Article 62 (2007).

[22] Fernández-Huertas Moraga, Jesús, Ada Ferrer, and Albert Saiz, "Immigrant Locations and Native Residential Preferences in Spain: New Ghettos?," Universidad Autónoma de Madrid mimeograph (2015).

[23] Grogger, Jeffrey, and Gordon H. Hanson, "Income Maximization and the Selection and Sorting of International Migrants," *Journal of Development Economics*, 95:1 (2011), 42–57.

[24] Hanson, Gordon H., "International Migration and the Developing World," in D. Rodrik and M. Rosenzweig (Eds.), *Handbook of Development Economics, Vol. 5* (Amsterdam: North-Holland, 2010), 4363–4414.

[25] Hausman, Jerry, "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left," *Journal of Economic Perspectives*, 15:4 (2001), 57–67.

[26] Heckman, James J., "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5:4 (1976), 475–492.

[27] INE, "National Immigrant Survey 2007. Methodology," (2007).

[28] Jayet, Hubert, Nadiya Ukrayinchuk, and Giuseppe De Arcangelis, "The Location of Immigrants in Italy: Disentangling Networks and Local Effects," *Annals of Economics and Statistics*, 2010:97/98 (2010), 329–350.

[29] Lerman, Steven R., "Location, Housing, Automobile Ownership, and Mode to Work: A Joint Choice Model," *Transportation Research Record*, 610 (1976), 6–11.

[30] Lewer, Joshua J., and Hendrik Van den Berg, "A Gravity Model of Immigration," *Economics Letters*, 99:1 (2008), 164–167.

[31] Llull, Joan, "Understanding International Migration: Evidence from a New Dataset of Bilateral Stocks (1960–2000)," *SERIEs*. 7:2 (2016), 221-255.

[32] Massey, Douglas S., "Economic Development and International Migration in Comparative Perspective," *Population and Development Review*. 14:3 (1988), 383-413.

[33] Mayda, Anna M., "International Migration: A Panel Data Analysis of the Determinants of Bilateral Flows," *Journal of Population Economics*, 23:4 (2010), 1249–1274.

[34] Mayer, Thierry, and Soledad Zignago, "Notes on CEPII's Distances Measures," (2006).

[35] McFadden, Daniel L., "Modelling the Choice of Residential Location," in Anders Karlqvist, Lars Lundqvist, Folke Snickars, and Jörgen W. Weibull (Eds.), *Spatial Interaction Theory and Planning Models* (Amsterdam: North-Holland, 1978), 75–96.

[36] McFadden, Daniel L., "Econometric Model of Probabilistic Choice," in Charles F. Manski and Daniel L. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications* (Cambridge, MA: MIT Press, 1981), chapter 5, 198-272.

[37] McFadden, Daniel L., "Econometric Analysis of Qualitative Response Models," in Zvi Griliches and Michael D. Intriligator (Eds.), *Handbook of Econometrics, Vol. II* (Amsterdam: Elsevier Science Publishers, 1984), chapter 24, 1396–1457.

[38] McKenzie, David, and Hillel Rapoport, "Self-Selection Patterns in Mexico-U.S. Migration: The Role of Migration Networks," *Review of Economics and Statistics*, 92:4 (2010), 811–821.

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

[40] Neubecker, Nina, and Marcel Smolka, "Co-national and Cross-national Pulls in International Migration to Spain," *International Review of Economics & Finance*, 28 (2013), 51–61.

[41] OECD, *International Migration Outlook: SOPEMI 2010,* OECD (2010), Paris.

[42] Ortega, Francesc, and Giovanni Peri, "The Effect of Income and Immigration Policies on International Migration," *Migration Studies*, 1:1 (2013), 47-74.

[43] Patel, Krishna, and Francis Vella, "Immigrant Networks and Their Implications for Occupational Choice and Wages," *Review of Economics and Statistics*, 95:4 (2013), 1249–1277.

[44] Pedersen, Peder J., Mariola Pytlikova, and Nina Smith, "Selection and Network Effects – Migration Flows into OECD Countries 1990-2000," *European Economic Review*, 52:7 (2008), 1160–1186.

[45] Picard, Robert, "GEODIST: Stata Module to Compute Geodetic Distances," Statistical Software Components, Boston College Department of Economics (2010).

[46] Quigley, John M., "Housing Demand in the Short Run: An Analysis of Polytomous Choice," in NBER (Ed.), *Explorations in Economic Research, Vol. 3, No. 1* (NBER, 1976), chapter 3, 76–102.

[47] Reher, David, and Miguel Requena, "The National Immigrant Survey of Spain: A New Data Source for Migration Studies," *Demographic Research*, 20:12 (2009), 253–278.

[48] Santos Silva, João M.C., and Silvana Tenreyro, "The Log of Gravity," *Review of Economics and Statistics*, 88:4 (2006), 641–658.

[49] Schmidheiny, Kurt, and Marius Brülhart, "On the Equivalence of Location Choice Models: Conditional Logit, Nested Logit and Poisson," *Journal of Urban Economics*, 69:2 (2011), 214–222.

[50] Stock, James H., Jonathan H. Wright, and Motohiro Yogo, "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business and Economic Statistics*, 20:4 (2002), 518–529.

[51] Wooldridge, Jeffrey M., "Selection Corrections for Panel Data Models Under Conditional Mean Independence Assumptions," *Journal of Econometrics*, 68:1 (1995), 115–132.

[52] Zavodny, Madeline, "Welfare and the Locational Choices of New Immigrants," *Federal Reserve Bank of Dallas Economic Review*, (1997), 2–10.

[53] Zavodny, Madeline, "Determinants of Recent Immigrants' Locational Choices," *International Migration Review*, 33:4 (1999), 1014–1030.

A Detailed derivations of the three-level NMNL migration model

We start with the general assumptions about the function H_i that guarantee that $F_i(e^o_{i1},\ldots,e^o_{iJ}) = \exp\left[-H_i\left(\exp[-e^o_{i1}],\ldots,\exp[-e^o_{1J}]\right)\right]$ is a multivariate extreme value distribution. Let $\mathbf{g}_i = (g_{i1},\ldots,g_{iJ})$ be a $(1\times J)$ row vector with non-negative entries, and let H_i be a non-negative function of $\mathbf{g}_i$ with:

$$\lim_{g_{ij}\to\infty} H_i(\mathbf{g}_i) = +\infty \quad \text{for } j = 1,\ldots,J. \tag{A.1}$$

Furthermore, assume that H_i is homogeneous of degree one in $\mathbf{g}_i$, and let H_i have mixed partial derivatives of all orders, with non-positive even and non-negative odd mixed derivatives.

These assumptions also imply that, if $(e^o_{i1},\ldots,e^o_{iJ})$ is distributed F_i, equation (3) in the main text can be written as:

$$\begin{aligned} P^o_i(j^o = j) &= \frac{\exp[U_{ij}]}{H_i(\exp[U_{i1}],\ldots,\exp[U_{iJ}])}\frac{\partial H_i(\exp[U_{i1}],\ldots,\exp[U_{iJ}])}{\partial \exp[U_{ij}]} \\ &= \frac{\partial \ln H_i(\exp[U_{i1}],\ldots,\exp[U_{iJ}])}{\partial U_{ij}}; \end{aligned} \tag{A.2}$$

see McFadden (1978, 80-81; 1981, 226-230). The function H_i that generates the response probabilities of our three-level NMNL model reads as:

$$\begin{aligned} H_i(\exp[U_{i1}],\ldots,\exp[U_{iJ}]) &= \sum_z \left(\sum_{r\in A_z}\left(\sum_{j\in A_{zr}} \exp[U_{ij}/(\kappa_r\lambda_z)]\right)^{\kappa_r}\right)^{\lambda_z} \\ &= \sum_z \exp[-c_{iz}]\left(\sum_{r\in A_z}\exp[-c_{ir}/\lambda_z]\left(\sum_{j\in A_{zr}}\exp[\xi_{ij}/(\kappa_r\lambda_z)]\right)^{\kappa_r}\right)^{\lambda_z} \end{aligned} \tag{A.3}$$

In order to derive the transition probability $P^o_i(r\in A_z)$ (equation (6) in the main text), one has to compute $\partial \ln H_i(\cdot)/\partial(-c_{iz})$, and similarly for the other transition probabilities (equations (7) and (8)). In the following, we show how to compute $P^o_i(j^o = j) = \partial \ln H_i(\cdot)/\partial U_{ij}$. Since

$$\ln H_i(\cdot) = \ln \sum_z \left(\sum_{r\in A_z}\left(\sum_{j\in A_{zr}} \exp[U_{ij}/(\kappa_r\lambda_z)]\right)^{\kappa_r}\right)^{\lambda_z} \tag{A.4}$$

we have

$$\frac{\partial \ln H_i(\cdot)}{\partial U_{ij}} = H_i(\cdot)^{-1}\exp[U_{ij}/(\kappa_r\lambda_z)]QX, \tag{A.5}$$

where

$$\begin{aligned} Q &= \left(\sum_{j\in A_{zr}} \exp[U_{ij}/(\kappa_r\lambda_z)]\right)^{\kappa_r-1} \\ &= (\exp[(-c_{iz}-c_{ir})/(\kappa_r\lambda_z)])^{\kappa_r-1}\left(\sum_{j\in A_{zr}} \exp[\xi_{ij}/(\kappa_r\lambda_z)]\right)^{\kappa_r-1} \end{aligned} \tag{A.6}$$

and

$$\begin{aligned} X &= \left(\sum_{r\in A_z}\left(\sum_{j\in A_{zr}} \exp[U_{ij}/(\kappa_r\lambda_z)]\right)^{\kappa_r}\right)^{\lambda_z-1} \\ &= (\exp[-c_{iz}/\lambda_z])^{\lambda_z-1}\left(\sum_{r\in A_i}(\exp[-c_{ir}/\lambda_z])\left(\sum_{j\in A_{zr}} \exp[\xi_{ij}/(\kappa_r\lambda_z)]\right)^{\kappa_r}\right)^{\lambda_z-1}. \end{aligned} \tag{A.7}$$

By defining $\Phi_{ir} = \ln\sum_{k\in A_{zr}} \exp[\xi_{ik}/(\kappa_r\lambda_z)]$, $\Omega_{iz} = \ln\sum_{\ell\in A_z} \exp[\Phi_{i\ell}\kappa_\ell - c_{i\ell}/\lambda_z]$ and $\Psi_i = \ln\sum_z \exp[\Omega_{iz}\lambda_z - c_{iz}]$, equation (A.5) can be written as:

$$\begin{aligned} \frac{\partial \ln H_i(\cdot)}{\partial U_{ij}} &= \frac{\exp[\xi_{ij}/(\kappa_r\lambda_z) - c_{ir}/\lambda_z - c_{iz}]}{H_i(\cdot)\exp[(1-\kappa_r)\Phi_{ir} + (1-\lambda_z)\Omega_{iz}]} \\ &= \frac{\exp[\xi_{ij}/(\kappa_r\lambda_z) - c_{ir}/\lambda_z - c_{iz}]}{\exp[\Psi_i + (1-\kappa_r)\Phi_{ir} + (1-\lambda_z)\Omega_{iz}]}, \end{aligned} \tag{A.8}$$

which gives $P_i^o(j^o = j)$, where $j \in A_{zr}, r \in A_z$; see equation (A.2) above and equation (12) in the main text.

B Data Appendix

Table B.1: List of Countries by World Region†

EAST ASIA & PACIFIC	Cuba*	NORTH AMERICA,	WESTERN EUROPE
China*	Dominican Republic*	AUSTRALIA	Austria
Japan	Ecuador*	& NEW ZEALAND	Belgium*
Korea	El Salvador	Australia	Denmark
Philippines	Honduras	Canada	Finland
	Mexico*	United States*	France*
EASTERN EUROPE	Peru*		Germany*
& CENTRAL ASIA	Uruguay*	SOUTH	Ireland
Bosnia and Herzegowina	Venezuela*	& SOUTHEAST ASIA	Italy*
Bulgaria*		India	Netherlands*
Poland*	MIDDLE EAST	Pakistan	Norway*
Romania*	& NORTH AFRICA		Portugal*
Russia*	Algeria*	SUB-SAHARAN	Sweden
	Egypt	AFRICA	Switzerland
LATIN AMERICA	Iran	Angola	United Kingdom*
& CARIBBEAN	Lebanon	Cape Verde	
Argentina*	Morocco*	Equatorial Guinea	
Bolivia*	Syria	Gambia	
Brazil*		Guinea	
Chile*		Mauritania	
Colombia*		Senegal	

† The baseline estimation sample for the scale model includes all countries with at least 630 migrants residing in Spain in the year 1996. These are the 55 countries listed above. The corresponding sample for the skill model includes all of the above countries that have sufficient data for the dependent variable (i.e. the skill composition of migration). These are the 28 countries marked with an asterisk.

Table B.2: Data Sources

Variable	Definition	Data Sources
Migrant Flow m_{ij}	Migrants who registered at municipalities in Spain between January 1, 1997, and December 31, 2006 (or other years depending on the regression), by province of destination (or region of destination) and by country of origin. Migrants are defined as individuals whose last country of residence (other than Spain) corresponds to their country of birth and nationality.	Spanish Residential Variation Statistics, INE, http://www.ine.es/en/prodyser/micro_varires_en.htm, accessed on 10/05/2010 (as well as on 11/24/2014 for the revision)
Migrant Skill Ratio m^h_{ij}/m^l_{ij}	Ratio of new high-skilled migrants over new low-skilled migrants, aggregated from 2002 to 2006, by region of destination in Spain and by country of birth. Migrants are individuals aged 16 years or older who were born abroad and have lived in Spain for more than a year, or at least intended to stay for more than a year at the time the survey was conducted.	National Immigrant Survey 2007, INE, http://www.ine.es/prodyser/micro_inmigra.htm, accessed on 10/05/2010
Migrant Network M_{ij}	Number of settled migrants as of May 1, 1996, by province of destination (or region of destination) in Spain and by nationality.	Population by Nationality, Autonomous Communities and Provinces, Sex and Year, Municipal Register, Main Population Series since 1998, INE, http://www.ine.es/jaxi/menu.do?type=pcaxis&path=%2Ft20%2Fe245&file=inebase&L=0, accessed on 10/07/2010
Trade Flow	Sum of exports and imports, by province (or region) in Spain and by country of destination/origin.	DataComex Statistics on Spanish Foreign Trade, Spanish Government, Ministry of Industry, Tourism and Trade, http://datacomex.comercio.es/principal_comex_es.aspx, accessed on 10/20/2010
FDI Flow	Gross FDI flow in Euros, by region in Spain and by country of the last owner.	DataInvex Statistics on Foreign Investments in Spain, Spanish Government, Ministry of Industry, Tourism and Trade, http://datainvex.comercio.es/principal_invex.aspx, accessed on 10/20/2010
Historical Internal Migrant Flow	People moving from one province (or region) to another province (or region) in Spain in 1988 and 1989, by province (or region) in Spain and by nationality.	Spanish Residential Variation Statistics, INE, http://www.ine.es/en/prodyser/micro_varires_en.htm, accessed on 10/05/2010
Geographical Distance	Distances are constructed on the basis of latitudinal and longitudinal data for regions in Spain and countries of origin and using the STATA module GEODIST by Picard (2010).	SpanishWikipedia/GeoHack, http://es.wikipedia.org, accessed on 09/05/2011; Mayer & Zignago (2006)

Table B.2 *continued*

Variable	Definition	Data Sources
Indicator for Common Lanuage	This variable is equal to one if at least 80% of a region's population in Spain are native speakers of a language spoken by at least 20% of the people in the country of origin; it is zero otherwise.	*Cataluña*: Generalitat de Catalunya, Institut d'Estadística de Catalunya (2008). Enquesta d'usos lingüístics de la població 2008. *Comunidad Foral de Navarra*: Instituto de Estadística de Navarra (2001). Censo 2001 de Población y Viviendas en Navarra. *Comunitat Valenciana*: Universidad de Salamanca (2007). Estudio CIS No. 2.667. La identitad nacional en España. *Galicia*: Instituto Galego de Estatística (2008). Enquisa de condicións de vida das familias. Coñecemento e uso do galego. Edición 2008. *Illes Balears*: Villaverde i Vidal, J. A. (2003). L'Enquesta Sociolingüística 2003. Principals Resultats. *País Vasco*: Universidad de Salamanca (2007). Estudio CIS No. 2.667. La identitad nacional en España. *Countries of origin*: Mayer & Zignago (2006).

C Measurement Error

We argue that the potential non-stochastic measurement errors discussed at the end of Section 4.1 are unlikely to result in biased estimates. Let $\tilde{m}_{ij} < m_{ij}$ and $\tilde{M}_{ij} > M_{ij}$ denote the unobserved true size of the migrant flow and the migrant network, respectively. Let the relationship between the migrant flow and the migrant network be given by the following equation:

$$\ln(\tilde{m}_{ij}) = \eta_{zr} \ln(\tilde{M}_{ij}). \tag{C.1}$$

Let y_{ij} denote the ratio of unobserved (i.e. "excess") migrants to observed migrants in the flow, and let x_{ij} denote the ratio of unobserved (i.e. unregistered) migrants to observed migrants in the network. Hence, $\tilde{m}_{ij} = (1 - y_{ij})m_{ij}$ and $\tilde{M}_{ij} = (1 + x_{ij})M_{ij}$ and thus:

$$\ln((1 - y_{ij})m_{ij}) = \eta_{zr} \ln((1 + x_{ij})M_{ij}), \tag{C.2}$$

which can be rewritten as:

$$\ln(m_{ij}) = \eta_{zr} \ln(M_{ij}) + \eta_{zr} \ln(1 + x_{ij}) - \ln(1 - y_{ij}). \tag{C.3}$$

The last two terms in equation (C.3), if not controlled for, may introduce a bias in the estimation of the network coefficient η_{zr}. Obviously, a sufficient condition for our FE model controlling for country-and-region fixed effects to deliver unbiased estimates is:

$$v_{ij} = v_{ir}, \quad v = \{x, y\}. \tag{C.4}$$

Hence, the type of mismeasurement potentially present in our migration data is not a problem *per se* for the estimation. For example, suppose that migrants are possibly measured with error, so that $x_{ij} \leq 0$ and $y_{ij} \leq 0$ for all provinces in Spain. Furthermore suppose that these errors are large for some regions of destination but small for others, and that they are large for some countries of origin but small for others. Then, a mild but sufficient condition for our estimates to be unbiased is: $x_{ij} = x_{ik}$ and $y_{ij} = y_{ik}$, where $j \neq k$ and $j, k \in A_{zr}$.

D Testing for Sample Selection Bias

We briefly present our procedure for identifying a potential sample selection bias in the model for the skill composition of migration. It is a slight modification of Wooldridge (1995, 123-124), who proposes

a method for testing for sample selection bias in panel data. It will become evident below that we impose very strong assumptions on the selection equation and the mechanism governing selection. These assumptions would often be inappropriate if we were to derive *corrections* for a sample selection bias in models with fixed effects. It turns out, however, that they do not pose a threat to the correct *testing* for a sample selection bias. For further details on this, the reader is referred to Wooldridge (1995).

We start by rewriting the model for the skill composition of migration as:

$$y_{ij} = \mu_i + \mathbf{x}_{ij}\boldsymbol{\beta} + u_{ij}, \qquad j = 1, \ldots, J, \tag{D.1}$$

where y_{ij} is the ij-specific log of the ratio of high-skilled to low-skilled migrants, μ_i is an unobserved country fixed effect, $\mathbf{x}_{ij}$ is a $1 \times K$ vector of explanatory variables (including region dummies and interactions between region dummies and world region dummies), $\boldsymbol{\beta}$ is a $K \times 1$ vector of parameters to be estimated, and u_{ij} is an independent and identically distributed error term. We explicitly allow for $E(\mu_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iJ}) \neq E(\mu_i)$. Since J is fixed, the asymptotic analysis is valid for $I \to \infty$. Now suppose that $(y_{ij}, \mathbf{x}_{ij})$ is sometimes unobserved, and that $\mathbf{s}_{ij} = (s_{i1}, \ldots, s_{iJ})'$ is a vector of selection indicators with $s_{ij} = 1$ if $(y_{ij}, \mathbf{x}_{ij})$ is observed and zero otherwise. Define $\mathbf{x}_i \equiv (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iJ})$ and $\mathbf{s}_i \equiv (\mathbf{s}_{i1}, \ldots, \mathbf{s}_{iJ})$ and suppose that $E(u_{ij}|\mu_i, \mathbf{x}_i, \mathbf{s}_i) = 0 \ \forall j$, which implies that the selection process is strictly exogenous conditional on μ_i and $\mathbf{x}_i$. Then, our FE estimator employed in the main text is consistent and asymptotically normal even when selection arbitrarily depends on $(\mu_i, \mathbf{x}_i)$ (Wooldridge 1995, 118).

In our application, the explanatory variables $\mathbf{x}_{ij}$ are observed for all regions $j = 1, \ldots, J$. The variable y_{ij} is observed if $s_{ij} = 1$, but not otherwise. For each $j = 1, \ldots, J$, define an unobserved latent variable

$$h^*_{ij} = \boldsymbol{\delta}_{j0} + \mathbf{x}_{i1}\boldsymbol{\delta}_{j1} + \cdots + \mathbf{x}_{iJ}\boldsymbol{\delta}_{jJ} + v_{ij}, \tag{D.2}$$

where v_{ij} is a stochastic term independent of $(\mu_i, \mathbf{x}_i)$, and $\boldsymbol{\delta}_{jp}$ is a $(K+1) \times 1$ vector of unknown parameters, $p = 1, 2, \ldots, J$.[30] The binary selection indicator is defined as $s_{ij} \equiv 1[h^*_{ij} > 0]$. Since $\mathbf{s}_i$ is a function of $(\mathbf{x}_i, \mathbf{v}_i)$, where $\mathbf{v}_i \equiv (v_{i1}, \ldots, v_{iJ})'$, a sufficient condition for the selection process to be

[30] In the following, $\mathbf{x}_{ij}$ includes one element more than in equation (D.1), despite the fact that we use the same notation for convenience. We thus assume that there is exactly one exclusion restriction in equation (D.1). In the estimation, we use the log of the number of people holding country i's nationality and migrating from region j in Spain to any other region $k \neq j$ within or outside Spain over the period from January 1, 2006, to December 31, 2007, as an exclusion restriction.

strictly exogenous conditional on μ_i and $\mathbf{x}_i$ is:

$$E(u_{ij}|\mu_i, \mathbf{x}_i, \mathbf{v}_i) = 0, \qquad j = 1, \ldots, J. \tag{D.3}$$

Under (D.3), there is no sample selection bias. An alternative that implies sample selection bias is:

$$E(u_{ij}|\mu_i, \mathbf{x}_i, \mathbf{v}_i) = E(u_{ij}|v_{ij}) = \rho v_{ij}, \qquad j = 1, \ldots, J, \tag{D.4}$$

where $\rho \neq 0$ is some unknown scalar. Under the alternative (D.4) we have:

$$E(y_{ij}|\mu_i, \mathbf{x}_i, \mathbf{s}_i) = \mu_i + \mathbf{x}_{ij}\boldsymbol{\beta} + \rho E(v_{ij}|\mu_i, \mathbf{x}_i, \mathbf{s}_i) = \mu_i + \mathbf{x}_{ij}\boldsymbol{\beta} + \rho E(v_{ij}|\mathbf{x}_i, \mathbf{s}_i). \tag{D.5}$$

Let $E(v_{ij}|\mathbf{x}_i, \mathbf{s}_i) = E(v_{ij}|\mathbf{x}_i, s_{ij})$ and assume a standard uniform distribution for v_{ij}. Then,

$$E(v_{ij}|\mathbf{x}_i, s_{ij} = 1) = E(v_{ij}|\mathbf{x}_i, v_{ij} > -\mathbf{x}_i\boldsymbol{\delta}_j) = (1 + \mathbf{x}_i\boldsymbol{\delta}_j)/2. \tag{D.6}$$

and

$$E(y_{ij}|\mu_i, \mathbf{x}_i, s_{ij} = 1) = \rho^* + \mu_i + \mathbf{x}_{ij}\boldsymbol{\beta} + \rho^*\mathbf{x}_i\boldsymbol{\delta}_j, \tag{D.7}$$

where $\rho^* \equiv \rho/2$ and $\mathbf{x}_i$ now includes unity as its first element. The procedure to test for sample selection bias is as follows. We first obtain estimates of $\mathbf{x}_i\boldsymbol{\delta}_j$ by estimating region-specific selection equations (where s_{ij} is the dependent variable) derived from equation (D.2), using linear probability models for the full data matrix. We then estimate equation (D.7) in an FE framework (within-transformed data), using only observations with $s_{ij} = 1$. We finally test $H_0 : \rho = 0$, using the t-statistic for ρ^*.