

Neubecker, Nina; Smolka, Marcel; Steinbacher, Anne

**Working Paper**

## Networks and selection in international migration to Spain

DIW Discussion Papers, No. 1306

**Provided in Cooperation with:**

German Institute for Economic Research (DIW Berlin)

*Suggested Citation:* Neubecker, Nina; Smolka, Marcel; Steinbacher, Anne (2013) : Networks and selection in international migration to Spain, DIW Discussion Papers, No. 1306, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:

<https://hdl.handle.net/10419/77090>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

# 1306

## Discussion Papers

Deutsches Institut für Wirtschaftsforschung

2013

# Networks and Selection in International Migration to Spain

Nina Neubecker, Marcel Smolka and Anne Steinbacher

Opinions expressed in this paper are those of the author(s) and do not necessarily reflect views of the institute.

#### IMPRESSUM

© DIW Berlin, 2013

DIW Berlin  
German Institute for Economic Research  
Mohrenstr. 58  
10117 Berlin

Tel. +49 (30) 897 89-0  
Fax +49 (30) 897 89-200  
<http://www.diw.de>

ISSN print edition 1433-0210  
ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:  
<http://www.diw.de/discussionpapers>

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN:  
<http://ideas.repec.org/s/diw/diwwpp.html>  
<http://www.ssrn.com/link/DIW-Berlin-German-Inst-Econ-Res.html>

# Networks and Selection

## in International Migration to Spain\*

Nina Neubecker<sup>‡</sup>                      Marcel Smolka<sup>§</sup>                      Anne Steinbacher<sup>¶</sup>  
University of Tübingen              University of Tübingen              University of Tübingen

This version: May 2013; first version: May 2012

### Abstract

This paper provides new evidence on migrant networks as determinants of the total size (scale) and skill structure of migration, using aggregate data from a recent migration boom to Spain. We draw upon McFadden (1984, 1422-1428) in order to develop and apply a three-level nested multinomial logit migration model. Our model accommodates varying degrees of similarity of destinations located in the same region (or the same country), allowing for a rich structure of substitutability across alternative destinations. We find strong positive network effects on the scale of migration and a strong negative effect on the ratio of high-skilled to low-skilled migrants. Simplifying restrictions on substitutability across destinations are rejected by the data.

**JEL Codes:** F22, J61

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

---

\*This paper is part of the authors' research for their doctoral theses (unpublished to date). An earlier draft of this paper was published in May 2012 (University of Tübingen Working Papers in Economics and Finance No. 35 & IAW Discussion Paper No. 83). The paper has benefitted from valuable suggestions by Wilhelm Kohler and Udo Kreickemeier. We also thank Peter Eppinger, Gordon Hanson, Johannes Pfeifer, Melissa Siegel, two anonymous referees, and participants of the 14th Göttinger Workshop "Internationale Wirtschaftsbeziehungen" (University of Göttingen), the Migration Seminar at the University of Tübingen, the XIII Conference on International Economics (University of Granada), a project workshop at the Sino-German Center for Research Promotion, Beijing, and the ifo Lunchtime Seminar at the Ifo Institute in Munich for their comments. Our research assistants at the Chairs of International Economics at the University of Tübingen have provided excellent support. Marcel Smolka gratefully acknowledges financial support from the Volkswagen Foundation under the project "Europe's Global Linkages and the Impact of the Financial Crisis: Policies for Sustainable Trade, Capital Flows, and Migration".

<sup>‡</sup>Present address: German Institute for Economic Research, DIW Berlin, Mohrenstraße 58, 10117 Berlin, Germany.

<sup>§</sup>Corresponding author, University of Tübingen, Mohlstraße 36, 72074 Tübingen, Germany. E-mail: marcel.smolka@uni-tuebingen.de, Phone: +49 (0) 7071 2978183.

<sup>¶</sup>University of Tübingen, Mohlstraße 36, 72074 Tübingen, Germany.



# 1 Introduction

An established body of literature argues that already settled migrants, often simply called a migrant network, alleviate the burden of migration for prospective newcomers, for example through informal job referrals among co-national peers (Munshi, 2003).<sup>1</sup> In this paper, we provide new evidence on migrant networks as determinants of the total size (scale) and skill structure of migration, drawing on aggregate data from a recent migration boom to Spain. Spain is an interesting case to look at. The country has become one of the world’s most attractive destinations for migrants due to its strong economic growth ahead of the Global Financial Crisis. From 1997 to 2009, Spain received roughly six million new migrants.<sup>2</sup> The foreign-born share among the total population has increased dramatically over the past few years, starting out from 4.9% in 2000 and approaching 14.1% in 2008 (OECD, 2010, 240).

In order to identify network effects in migration to Spain, we develop and apply a three-level nested multinomial logit (NMNL) migration model along the lines of McFadden (1984, 1422-1428). The model generalizes the standard multinomial logit (MNL) model described in McFadden (1984, 1411-1415), which assumes that, conditional on observables, any two migration destinations are equally substitutable for one another. This assumption is largely at odds with the fact that destinations located in the same *territorial entity* (e.g. a sovereign state or a country subdivision with independent legislative authority) are similar in many respects that are difficult or impossible to observe. They share the same legal and political framework; they have a common cultural background; and they engage in similar economic activities. Our NMNL framework allows for such similarities in the multi-level hierarchy of territorial entities, featuring the highest degree of substitutability across destinations that are located in the same region of a given country, and the lowest one across destinations that are located in different countries. In doing so, our model introduces unobserved heterogeneity into the migration function that challenges previous identification strategies based on aggregate cross-sectional migration data.

Previous attempts to model cross-destination substitutability in migration are furthermore challenged by the so-called “Dispositive Principle”, an important feature of the Spanish political system. As part of the Spanish constitution, it grants regional authorities the right to define the extent of their legislative autonomy (Morales & Molés, 2002, 180). Hence, destinations in regions with a high demand for self-government are rendered more similar to each other than destinations in other re-

---

<sup>1</sup>Massey (1988, 396) defines migrant networks as “[...] 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.”

<sup>2</sup>Of these migrants, 13.6% are Romanians, followed by Moroccans (11.1%), Ecuadorians (8.2%), Colombians (6.1%), Britons (5.3%), and Bolivians (4.7%). Unless stated otherwise, all migration figures in this paper are own calculations based on data from the Spanish Instituto Nacional de Estadística (INE).

gions. Related arguments derive from the fact that some, but not all, regions have a second official language that is actively used by the population (in addition to *castellano*). Therefore, as a general rule, destinations in regions with a pronounced political and cultural autonomy should appear as close substitutes, relative to destinations in other regions. Our NMNL framework allows us to model this issue by introducing similarity parameters that are specific to the different regions of destination in Spain.<sup>3</sup> Although we cannot estimate these parameters directly, our model suggests that estimated network coefficients are not homogeneous across destinations, a possibility that we explore in detail and that requires a careful interpretation of the network effect.

Obtaining consistent and unbiased estimates of network effects in migration is not trivial. The main endogeneity concern is the two-way relationship between migration costs and migrant networks, defined as the number of migrants from a certain nationality that are already settled in a certain destination. On the one hand, the migrant network appears as an argument in the migration cost function determining future migration. On the other hand, the migrant network is the result of past migration, and is thus itself influenced by migration costs. Our data distinguish among both different countries of origin and different provinces of destination in Spain. This allows us to go beyond the existing literature in the way we control for unobserved heterogeneity in migration costs through fixed effects. By grouping countries of origin into world regions, we control for all migration costs specific to the world region of origin and the province of destination (e.g. Latin American people being especially well-received in the province of Murcia).<sup>4</sup> By grouping provinces of destination into regions, we control for all migration costs specific to the country of origin and the region of destination in Spain (e.g. the short distance between France and Cataluña). 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, although smaller than those reported in the received literature. Since individual migration moves are independent of the effect they have on others' migration decisions our results have important policy implications. In a dynamic model of labor migration, network effects indicate a welfare loss in the laissez-faire transition path equilibrium (Carrington et al., 1996; Chau, 1997). From the perspective of a social planner who wants to maximize world welfare, they call for migration subsidies that accelerate the speed of migration. Our estimates also attest to strong negative effects of migrant networks on the skill structure of migration, defined as the ratio of high-skilled to low-skilled

---

<sup>3</sup>To the best of our knowledge, no other random utility model that could be estimated with our data would allow us to do likewise. For example, the generalized nested logit (GNL) model by Wen & Koppelman (2001) could be used to closely approximate our three-level NMNL, but its estimation is not feasible with our data.

<sup>4</sup>This approach also controls for the fact that migrants are attracted to destinations hosting migrants from countries that are culturally and geographically close to their own country of origin (cf. Neubecker & Smolka, forthcoming).

migrants. This finding accords with the idea that high-skilled individuals have lower effective migration costs than low-skilled individuals (Chiswick, 1999). Intuitively, migrant networks are more important for low-skilled individuals than they are for high-skilled individuals, biasing the skill structure of migration toward the low-skilled individuals.

Our estimates strongly reject a uniform degree of substitutability across alternative destinations, working against the standard MNL model in our application to the Spanish case. We find pronounced heterogeneity in the estimated network coefficients across destinations, an observation that has (to the best of our knowledge) received no attention at all 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. The estimated elasticity is lowest for the destinations located in the region of Extremadura, slightly exceeding a value of 0.1. It is highest for the destinations located in the region of Cataluña, lying in the vicinity of 0.55. We can conclude from our results that the ease with which one destination can be substituted for another one is highest in the region of Cataluña, arguably the region with the highest degree of political and cultural autonomy in Spain.

Our paper is related to recent estimates of network effects based on aggregate migration data. Beine et al. (2011) investigate the determinants of the scale and skill structure of migration between the years 1990 and 2000 to 30 OECD countries. They find that economies that already host migrants from a given country attract both a larger number of new migrants as well as a larger fraction of low-skilled migrants from that country.<sup>5</sup> Similar results are obtained by Beine & Salomone (2013) who study potential gender differences in network effects. The paper by Beine et al. (2012) disentangles what the authors call local and national network externalities, saying that local migrant networks facilitate the assimilation of migrants in the host society, while nation-wide migrant networks help overcome the legal entry barriers to migration. However, all of these papers derive the estimated migration functions from a standard MNL model that assumes a uniform degree of cross-destination substitutability.<sup>6</sup>

Our paper is also related to a number of macro-level studies that are more generally concerned with the determinants of international migration.<sup>7</sup> In this literature, migrant networks robustly rank

---

<sup>5</sup>See also Grogger & Hanson (2011, 53) for complementary evidence. McKenzie & Rapoport (2010) find positive self-selection on education from Mexican migrants to the U.S. to be more likely, the larger the number of return migrants in 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.

<sup>6</sup>While revising this paper, we became aware of research by Bertoli & Fernández-Huertas Moraga (2012). They use the same migration data as Beine et al. (2011) in order to estimate network effects in migration, relaxing the assumption of a uniform degree of substitutability across alternative destinations. The most general version of their estimated model reduces to a two-level NMNL model with a single similarity parameter for all “nests” (territorial entities in our paper); see our online Addendum A for details.

<sup>7</sup>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).

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). Two recent papers, Bertoli & Fernández-Huertas Moraga (2013) and Ortega & Peri (2013), develop micro-founded random utility migration models in order to estimate the determinants of migration. In both papers, the standard MNL assumption of a uniform degree of cross-destination substitutability is relaxed. Bertoli & Fernández-Huertas Moraga (2013) use the same Spanish data source as we do in this paper. They argue that the Common Correlated Effects (CCE) estimator, a panel estimator proposed by Pesaran (2006), yields consistent estimates of the migration function under arbitrary specifications of the cross-nested logit (CNL) model due to Vovsha (1997). The CNL model allocates a “portion” of each destination to a set of “nests” (territorial entities in our paper), assuming, contrary to our model, that there is a single similarity parameter for all nests.<sup>8</sup> Ortega & Peri (2013) investigate the impact of income and immigration policies on migration to OECD countries, using panel data detailed by country of origin and country of destination.<sup>9</sup> Their model, best understood as a two-level NMNL model with a single similarity parameter for all nests, allows for a higher degree of substitutability across destinations that are located outside the individual’s country of origin. However, neither Bertoli & Fernández-Huertas Moraga (2013) nor Ortega & Peri (2013) identify the effects of migrant networks on the scale and skill structure of migration, as we do in this paper.

The remainder of this paper is organized as follows. Section 2 characterizes individual decision making in a three-level NMNL model. From this model, we derive estimable equations for the scale and skill structure of migration. In Section 3 we present our estimation strategy and introduce in detail the data that we employ in our econometric analysis. Section 4 presents our estimation results. 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 multi-country random utility framework with many countries of origin and many provinces of destination at the sub-country level.

---

<sup>8</sup>The CNL model is a special case of the GNL model. Unlike the GNL model, the CNL model cannot be used to approximate our three-level NMNL model (cf. Wen & Koppelman, 2001). Bertoli et al. (2013) employ the CNL model in order to study the effect of the recent economic crisis in Europe on migration to Germany.

<sup>9</sup>In Ortega & Peri (2009), a previous version of Ortega & Peri (2013), the authors also study the effects of migration on employment, investment, and productivity.

## 2.1 Basic Setup

We assume that the decision making process leading to migration follows a hierarchical structure in which provinces of destination (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<sup>10</sup>: 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.<sup>11</sup> Let  $i = 1, \dots, I$ , index countries of origin,  $j$  or  $k = 1, \dots, J$ , index provinces of destination,  $z$  or  $y = 1, \dots, Z$ , index the primary nests (countries of destination), and  $r$  or  $\ell = 1, \dots, R$ , index the secondary nests (regions of destination within countries), as perceived by individuals living in country  $i$ .<sup>12</sup> Let the country of origin  $i$  be one element in each of the sets  $\{1, \dots, Z\}$ ,  $\{1, \dots, R\}$ , and  $\{1, \dots, J\}$ ; it represents a degenerate nest with a single final migration destination. Define  $A_{zr}$  as the set of provinces belonging to region  $r$  in country  $z$ , and  $A_z$  as the set of regions belonging to country  $z$ .

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

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

where the index  $o = 1, \dots, m_i$ , identifies individuals originating from country  $i$ , the terms  $Y_j$  and  $C_{ij}$  are sub-utility functions relevant for moving from country  $i$  to province  $j$  and living in province  $j$ , and the term  $e_{ij}^o$  is a stochastic (random) utility variable with individual-specific realizations for each province  $j = 1, \dots, J$ . The function  $Y_j$  summarizes utility-relevant characteristics of province  $j$  such as the wage rate, the state of the housing market, and the climate. It is assumed to be independent of the individual’s country of origin. The function  $C_{ij}$  captures the costs of moving and assimilation, henceforth called migration costs. Similar to Beine et al. (2011, 33-34), we hypothesize that these costs are a decreasing and globally convex function of the migrant network,  $M_{ij}$ , defined as the number of co-national migrants already settled in province  $j$ . A convenient specification of migration costs that incorporates the idea of positive but diminishing returns to the migrant network uses the log of  $M_{ij}$ :

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

---

<sup>10</sup>We assume that each decision in this hierarchy is made conditional on both the fixed preceding decisions and the optimal succeeding decisions. Hence, one can think of individuals as deciding on all aspects of their migration moves simultaneously (cf. Domencich & McFadden, 1975, 33-46).

<sup>11</sup>In Ortega & Peri (2013), the first decision of individuals is between going abroad and staying at home. Our econometric implementation is compatible with this additional structure.

<sup>12</sup>Strictly speaking, the final migration destinations  $j$  and the nests  $r$  and  $z$  are  $i$ -specific. We omit this index in order to avoid notational clutter.

where the parameter  $\theta > 0$  is a measure for the strength of the network effect, and where we add one to the variable  $M_{ij}$  before taking logs in order to abstract from infinitely large migration costs. The other cost components not related to the migrant network will be described in more detail below. Suffice it to say here that, for a given country of origin  $i$ , they vary either across countries of destination ( $c_{iz}$ ), across regions of destination ( $c_{ir}$ ), or across 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 choose from the set of provinces the alternative from which they derive the highest utility:

$$j^o = \operatorname{argmax}(U_{i1}^o, \dots, U_{iJ}^o), \quad j^o \in \{1, \dots, J\}. \quad (3)$$

The probability that individual  $o$  from country  $i$  migrates to province  $j$  is equal to the probability that this individual associates the largest utility with moving to province  $j$ :

$$\begin{aligned} P_i^o(j^o = j) &= \Pr(U_{ij}^o > U_{ik}^o \quad \forall k \in \{1, \dots, J\} : k \neq j) \\ &= \Pr(e_{ik}^o - e_{ij}^o < U_{ij} - U_{ik}; \\ &\quad \forall k \in \{1, \dots, J\} : k \neq j). \end{aligned} \quad (4)$$

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

$$P_i^o(j^o = j) = P_i^o(j^o = j | j^o \in A_{zr}) P_i^o(j^o \in A_{zr} | r \in A_z) P_i^o(r \in A_z), \quad j \in A_{zr}, r \in A_z. \quad (5)$$

These probabilities depend on the distribution assumed for the random utility variables,  $e_{i1}^o, \dots, e_{iJ}^o$ . Let  $\mathbf{g}_i = (g_{i1}, \dots, g_{iJ})$  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:

$$\lim_{g_{ij} \rightarrow \infty} H_i(\mathbf{g}_i) = +\infty \quad \text{for } j = 1, \dots, J. \quad (6)$$

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. It can be shown that the function

$$F_i(e_{i1}^o, \dots, e_{iJ}^o) = \exp[-H_i(\exp[-e_{i1}^o], \dots, \exp[-e_{iJ}^o])] \quad (7)$$

is a multivariate extreme value distribution function, and that, if  $(e_{i1}^o, \dots, e_{iJ}^o)$  is distributed  $F_i$ , (4)

can be written as:

$$\begin{aligned} P_i^o(j^o = j) &= \frac{\exp[U_{ij}]}{H_i(\exp[U_{i1}], \dots, \exp[U_{iJ}])} \frac{\partial H_i(\exp[U_{i1}], \dots, \exp[U_{iJ}])}{\partial \exp[U_{ij}]} \\ &= \frac{\partial \ln H_i(\exp[U_{i1}], \dots, \exp[U_{iJ}])}{\partial U_{ij}}; \end{aligned} \quad (8)$$

see McFadden (1978, 80-81; 1981, 226-230).<sup>13</sup>

We depart from the received literature in that we introduce a function  $H_i$  that generates the response probabilities of a three-level NMNL model. It allows for the random utilities associated with provinces belonging to the same region (or the same country) to be mutually correlated, whereas the random utilities associated with provinces in different countries are independent. Define on the half-open unit interval two parameters,  $\lambda_z$  and  $\kappa_r$  ( $0 < \kappa_r, \lambda_z \leq 1$ ), measuring the similarity of the provinces located in country  $z$  and region  $r$ , respectively. These two parameters govern the degree of substitutability across alternative destinations; they are allowed to vary across countries and across regions, respectively. High parameter values indicate little similarity among provinces (and weak correlations among the random utilities), low parameter values indicate much similarity (and strong correlations). As we have argued in the introduction, cross-regional differences in the similarity parameter  $\kappa_r$  in Spain could derive, for example, from the constitutionally anchored “Dispositive Principle”, which allows for region-specific degrees of legislative autonomy. We assume:

$$\begin{aligned} H_i(\exp[U_{i1}], \dots, \exp[U_{iJ}]) &= \sum_z \left( \sum_{r \in A_z} \left( \sum_{j \in A_{zr}} \exp[U_{ij}/(\kappa_r \lambda_z)] \right)^{\kappa_r} \right)^{\lambda_z} \\ &= \sum_z \exp[-c_{iz}] \left( \sum_{r \in A_z} \exp[-c_{ir}/\lambda_z] \left( \sum_{j \in A_{zr}} \exp[\xi_{ij}/(\kappa_r \lambda_z)] \right)^{\kappa_r} \right)^{\lambda_z}. \end{aligned} \quad (9)$$

It is instructive to note that the function  $H_i(\cdot)$  nests the generating function for the response probabilities of the standard MNL model as a special case with  $\kappa_r = \lambda_z = 1 \ \forall r, z$ . This rules out any correlation among the random utilities. We shall return to this in more detail below. From equations (8) and (9) it follows that each transition probability in equation (5) has a closed-form analytical solution<sup>14</sup>:

$$P_i^o(r \in A_z) = \exp[\Omega_{iz} \lambda_z - c_{iz} - \Psi_i], \quad (10)$$

$$P_i^o(j^o \in A_{zr} | r \in A_z) = \exp[\Phi_{ir} \kappa_r - c_{ir}/\lambda_z - \Omega_{iz}], \quad (11)$$

<sup>13</sup>We show in our online Addendum B how to derive (8).

<sup>14</sup>For example, in order to derive  $P_i^o(r \in A_z)$ , one simply has to compute  $\partial \ln H_i(\cdot)/\partial(-c_{iz})$ , and similarly for the other transitional probabilities. We show in Appendix A how to compute  $P_i^o(j^o = j) = \partial \ln H_i(\cdot)/\partial U_{ij}$ .

$$P_i^o(j^o = j | j^o \in A_{zr}) = \exp[\xi_{ij}/(\lambda_z \kappa_r) - \Phi_{ir}], \quad (12)$$

where  $\Phi_{ir}$ ,  $\Omega_{iz}$ , and  $\Psi_i$  are “inclusive values” defined as:

$$\Phi_{ir} \equiv \ln \sum_{k \in A_{zr}} \exp[\xi_{ik}/(\lambda_z \kappa_r)], \quad (13)$$

$$\Omega_{iz} \equiv \ln \sum_{\ell \in A_z} \exp[\Phi_{i\ell} \kappa_\ell - c_{i\ell}/\lambda_z], \quad (14)$$

$$\Psi_i \equiv \ln \sum_z \exp[\Omega_{iz} \lambda_z - c_{iz}]. \quad (15)$$

The inclusive values  $\Phi_{ir}$ ,  $\Omega_{iz}$ , and  $\Psi_i$  summarize, respectively, the characteristics of all provinces belonging to region  $r$ , all provinces belonging to country  $z$ , and all provinces belonging to the complete set of final migration destinations. Using equation (5) together with equations (10) to (15) and aggregating over all individuals from country  $i$ , we can write the expected rate of migration from country  $i$  to province  $j$  as:

$$\frac{m_{ij}}{m_i} = \frac{\exp[\xi_{ij}/(\lambda_z \kappa_r) - c_{ir}/\lambda_z - c_{iz}]}{\exp[\Psi_i + (1 - \kappa_r)\Phi_{ir} + (1 - \lambda_z)\Omega_{iz}]}, \quad (16)$$

where  $m_{ij}$  is the number of individuals migrating from  $i$  to  $j$ , and  $m_i$  is the initial population size of country  $i$ . This  $ij$ -specific migration rate depends on the inclusive values  $\Phi_{ir}$ ,  $\Omega_{iz}$ , and  $\Psi_i$ . It is therefore responsive to the attractiveness of all provinces  $k = 1, \dots, J$ , whether in the same region  $r$  (or the same country  $z$ ) as province  $j$  or not. It is in this sense that we refer to the inclusive values as “multilateral resistance” terms.<sup>15</sup> For example, consider the elasticity of the  $ij$ -specific migration rate,  $m_{ij}/m_i$ , with respect to  $Y_k$ , the utility-relevant 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<sup>16</sup>:

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

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.<sup>17</sup> 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  $ij$ -

<sup>15</sup>Mayda (2010) speaks of “multilateral pull” effects. The idea of multilateral resistance here is similar to that in the gravity equation for international trade flows (cf. Anderson & van Wincoop, 2003). Anderson (2011) sketches a general equilibrium migration model with multilateral resistance. See also Hanson (2010, 4373-4375) for a discussion.

<sup>16</sup>We show in Appendix B how to compute this elasticity.

<sup>17</sup>Notice that  $I(j, k) = 1$  implies that  $I(\ell, r) = I(y, z) = 1$  but not the other way around.

specific migration rate, depending on whether this province belongs to the same country or region as province  $j$ . In particular, the elasticity in (17) is largest (in absolute value) for any change in the conditions in other provinces in 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 located in the same region (and in the same country). In the standard MNL model with  $\lambda_z = \kappa_r = 1 \forall r, z$ , the pattern of cross-elasticities becomes strikingly simple: for  $k \neq j$ , (17) 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  $k$  are located in the same region or in the same country.

The flexible structure of cross-destination substitutability in our NMNL model notwithstanding, the issue of multilateral resistance is not a special feature of the NMNL model. It is a key element of the standard MNL model as well. To see this, note that with  $\lambda_z = \kappa_r = 1 \forall r, z$ , the  $ij$ -specific migration rate reads as:

$$\left. \frac{m_{ij}}{m_i} \right|_{\lambda_z, \kappa_r=1} = \frac{\exp[\xi_{ij} - c_{ir} - c_{iz}]}{\exp[\Psi_i]} = \frac{\exp[U_{ij}]}{\sum_k \exp[U_{ik}]}, \quad (18)$$

which depends not only on the conditions in  $i$  and  $j$ , but also on the conditions in all other provinces through the multilateral resistance term  $\Psi_i$ . Based on the standard MNL model of equation (18), a common approach in the literature is to compute the  $ij$ -specific migration rate (namely, the fraction of the population in  $i$  who migrate to  $j$ ) relative to the  $i$ -specific stay rate (namely, the fraction of non-migrants of the population in  $i$ ):

$$\frac{m_{ij}}{m_{ii}} = \exp[U_{ij} - U_{ii}], \quad (19)$$

where the multilateral resistance term  $\Psi_i$  cancels out. In the standard MNL model, the odds ratio between any two provinces is thus independent of the number and characteristics of other provinces, a property known as the independence of irrelevant alternatives (IIA) assumption (McFadden 1974, 1978).<sup>18</sup> Thus, estimating a log-linearized version of equation (19) (instead of estimating a log-linearized version of equation (18)) has the advantage that no attention needs to be paid to the multilateral resistance term, provided that the IIA assumption is not violated. In our more general NMNL modeling framework, the relative odds become:

$$\frac{m_{ij}}{m_{ii}} = \frac{\exp[\xi_{ij}/(\lambda_z \kappa_r) - \xi_{ii} - c_{ir}/\lambda_z + c_{il} - c_{iz} + c_{iy}]}{\exp[(1 - \kappa_r)\Phi_{ir} + (1 - \lambda_z)\Omega_{iz}]}, \quad (20)$$

---

<sup>18</sup>Strictly speaking, the standard MNL model as such does not imply the IIA property. The IIA property would indeed be absent in the standard MNL model if  $U_{ij}$  was a function of any of the characteristics of province  $k \neq i, j$ .

where  $j \in A_{zr}$ ,  $r \in A_z$  and  $i \in A_{y\ell}$ ,  $\ell \in A_y$ , and where we have used the fact that the country of origin  $i$  represents a single final migration destination. It is thus easy to verify that the odds ratio between any two provinces belonging to two different regions is not independent of the number and characteristics of other provinces. This involves a partial relaxation of the IIA assumption. Hence, in our NMNL framework, the issue of multilateral resistance needs to be addressed explicitly, whether we estimate a log-linearized version of equation (16) or of equation (20).<sup>19</sup> Given that the variable  $m_i$  in equation (16) is exogenous, while the variable  $m_{ii}$  in equation (20) is endogenous and potentially difficult to observe, we use the  $ij$ -specific migration rate in equation (16) for our econometric implementation.

## 2.2 Scale of Migration

Substituting  $\xi_{ij}$  in equation (16), taking logs, and rearranging terms yields the following migration function for  $j \in A_{zr}$ ,  $r \in A_z$ :

$$\begin{aligned} \ln(m_{ij}) = & \frac{\theta}{\lambda_z \kappa_r} \ln(1 + M_{ij}) + \ln(m_i) + \frac{1}{\lambda_z \kappa_r} Y_j - c_{iz} - \frac{1}{\lambda_z} c_{ir} - \frac{1}{\lambda_z \kappa_r} c_{ij}, \\ & \underbrace{-\Psi_i - (1 - \lambda_z)\Omega_{iz} - (1 - \kappa_r)\Phi_{ir}}_{\text{Multilateral resistance}}. \end{aligned} \quad (21)$$

Identification of the network effect is thus complicated by the presence of both the different cost components and the multilateral resistance terms. Moreover, the network coefficient, defined as  $\eta_{zr} \equiv \eta(\lambda_z, \kappa_r) = \frac{\theta}{\lambda_z \kappa_r}$ , is a decreasing function of  $\lambda_z$  and  $\kappa_r$ ; it is larger the larger the similarities of provinces in country  $z$  and region  $r$ , respectively. For low values of  $\lambda_z$  and  $\kappa_r$ , it is easy to substitute one province for another one in the same country or region, respectively. In this case, a small increase in the migrant network in province  $k \in A_{zr}$ ,  $r \in A_z$ , leads a large number of individuals to substitute another province  $j \in A_{zr}$  by province  $k$ , other things held constant. We expect to find higher degrees of cross-destination substitutability (and thus larger network coefficients) in regions that put a lot of emphasis on their political and cultural autonomy.

## 2.3 Skill Structure of Migration

We now distinguish between high-skilled and low-skilled individuals, denoted by  $h$  and  $l$ , respectively. We augment the utility function by a parameter  $\gamma^s > 0$ ,  $s \in \{h, l\}$ , representing the ease with which individuals are able to cope with migration costs (decreasing with higher values):

$$U_{ij}^o = Y_j - \gamma^s C_{ij} + e_{ij}^o, \quad (22)$$

---

<sup>19</sup>The same applies to the CNL migration model estimated in Bertoli & Fernández-Huertas Moraga (2013).

where  $s = h$  if individual  $o$  is high-skilled and  $s = l$  otherwise. We assume that  $\gamma^h < \gamma^l$ , so high-skilled individuals have lower effective migration costs than 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 (21). Subtracting the equation for low-skilled migrants from the same equation for high-skilled migrants, we obtain:

$$\ln \left( \frac{m_{ij}^h}{m_{ij}^l} \right) = \frac{\theta \gamma^*}{\lambda_z \kappa_r} \ln(1 + M_{ij}) + \ln \left( \frac{m_i^h}{m_i^l} \right) - \gamma^* c_{iz} - \frac{\gamma^*}{\lambda_z} c_{ir} - \frac{\gamma^*}{\lambda_z \kappa_r} c_{ij} - \Psi_i^* - (1 - \lambda_z) \Omega_{iz}^* - (1 - \kappa_r) \Phi_{ir}^*, \quad (23)$$

where the variables with an asterisk (\*) are differences between the corresponding parameters (or variables) for high-skilled and low-skilled individuals. Since  $\gamma^* < 0$ , the ratio of new high-skilled to new low-skilled migrants is a decreasing function of the migrant network. This result 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.<sup>20</sup>

### 3 Estimation Strategy and Data

In this section we describe our estimation strategy and we present the different variables that we use in the estimation. We estimate different variants of the models given by equations (21) and (23), each augmented by a stochastic error term. We consider two different aggregation levels for final migration destinations in Spain. The model for the scale of migration is estimated at the level of provinces in Spain. Due to reasons of data availability, the model for the skill structure of migration is estimated at the level of regions in Spain.<sup>21</sup> For both models, our benchmark estimates are based on a sample comprising the 55 most important countries of origin listed in Table C.1 in Appendix C. These are all countries with at least 630 migrants in Spain in the year 1996. All migration data come from the Spanish Instituto Nacional de Estadística (INE). The full internet sources of our data are listed in Table C.2 in Appendix C.

<sup>20</sup>This is reflected in the following inequality:  $\partial U_{ij}(\gamma^l)/\partial M_{ij} > \partial U_{ij}(\gamma^h)/\partial M_{ij}$ . In this respect, our modeling approach is akin to the one in Beine et al. (2011).

<sup>21</sup>Spain is divided into 52 provinces which are nested in 19 regions. We exclude the provinces (enclaves) of Ceuta and Melilla due to their specific geographical location and thus we end up with 50 provinces nested in 17 regions. See <http://www.ine.es/daco/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.

### 3.1 Scale of Migration

The dependent variable is the log of the migration flow to provinces of destination in Spain, obtained from the Spanish Residential Variation Statistics and aggregated from the beginning of 1997 until the end of 2006.<sup>22</sup> This period covers Spain’s unprecedented migration boom, which was eventually attenuated by the global financial and economic crisis starting in 2007. The migrant network,  $M_{ij}$ , is measured by the number of already settled migrants in 1996, as reported by the Spanish Municipal Register. We rely on population figures disaggregated by nationalities and by provinces in Spain as of May 1, 1996.

From the year 2000 onwards, our migration 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, irrespective of their legal status.<sup>23</sup> Each registrant must provide his or her name, surname, sex, usual domicile, nationality, passport number, as well as the place and date of birth.<sup>24</sup> Since this information is confidential and must not be communicated to other administrative units, the probability of forced repatriation is plausibly independent of registration.

We identify the model from the within-cluster variation across provinces in the data. We start with a parsimonious fixed effects (FE) specification in which we define as clusters the different countries of origin, computing all variables in equation (21) as deviations from their country means (within-transformation).<sup>25</sup> This approach wipes out, first, all terms with subscript  $i$  and thus controls for the initial population size in the country of origin as well as for the multilateral resistance term  $\Psi_i$ ; and second, it wipes out all terms with subscript  $iz$  because our migration data refer to a single country of destination  $z$ . By eliminating  $c_{iz}$ , it thus controls, for example, for the impact of country-specific migration policies and the geographical and cultural distance between the country of origin and the country of destination. By eliminating  $\Omega_{iz}$ , it is compatible with a model in which the degree of cross-destination substitutability is larger within than across countries of destination.

---

<sup>22</sup>Migrants are defined as individuals whose last country of residence (other than Spain) corresponds to their country of birth and nationality. In their raw form, the migration flow data are observed for periods of less than a year. We aggregate the data over time because the model cannot deal with a time dimension in any convenient way, unless we assume that in every period individuals left in the home country draw new realizations of the random utility variables  $e_{i1}^o, \dots, e_{iJ}^o$ , an assumption too strong to be plausible.

<sup>23</sup>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 in case 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).

<sup>24</sup>See INE at [http://www.ine.es/en/metodologia/t20/t203024566\\_en.htm](http://www.ine.es/en/metodologia/t20/t203024566_en.htm), accessed on 08/19/2011.

<sup>25</sup>When zero values inflate the dependent variable, the FE estimator delivers inconsistent estimates (cf. Santos Silva & Tenreyro, 2006). In our sample we observe only a modest number of zero migration flows (5.75% of all country-province pairs) and therefore apply the FE estimator.

In more demanding specifications of our FE model, we define as clusters the different pairs of countries of origin and regions of destination, computing all variables as deviations from their country-and-region means. In addition to the above-described country effects, this approach wipes out all terms with subscript  $ir$ . These terms include, first, the multilateral resistance term  $\Phi_{ir}$ , so that this approach is fully compatible with our three-level NMNL model; and second, they include the cost term  $c_{ir}$  representing the geographical and cultural distance between the country of origin and the region of destination. Important elements of this distance derive from a cultural, political, and historical context. For example, the different regions in Spain feature substantial 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<sup>26</sup>; 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.

All other migration costs are summarized in the term  $c_{ij}$ . Some of these costs, for example those related to the attitudes of the native population toward migrants, may be specific to the province of destination  $j$  but independent of the country of origin  $i$ . We control for these province-specific migration costs by including a set of province fixed effects in the estimation. The province fixed effects also absorb the impact of province-specific pull factors summarized in the term  $Y_j$ . Some other migration costs may be 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 co-national migrants (i.e., migrants from Ecuador) but also by a network of migrants from other Latin American countries (cf. Neubecker & Smolka, forthcoming). This additional effect, a “cross-national” network externality, would lower the migration costs for potential migrants from Ecuador, leading to a higher incidence of migration. In more demanding specifications of our model, we therefore control for these other migration costs with a set of world region-and-province fixed effects.<sup>27</sup>

As further control variables, we include bilateral trade and capital flows where possible. 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 could create demand for specific types of labor, especially foreign labor. Data on both trade and foreign direct investment (FDI) are provided by the Spanish Ministry of Industry, Tourism and Trade. We measure

---

<sup>26</sup>The Basque Autonomous Community and Navarre form the Spanish part of the Basque Country (*País Vasco* in Spanish; *Euskal Herria* in Basque language).

<sup>27</sup>In terms of world regions, we distinguish among 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; as well as Western Europe. For a similar classification used by the IMF, see <http://www.imf.org/external/datamapper/region.htm>, accessed on 07/25/2012.

$ij$ -specific trade flows by the sum of exports and imports (in Euros) in the year 1996. These information are taken from DataComex Statistics on Spanish Foreign Trade. Ideally, we would like to use FDI stocks to measure inward investment but we only have information on gross FDI inflows (in Euros). These are available from DataInvex Statistics on Foreign Investments in Spain and detailed by the country of the last owner and by the region of destination in Spain.<sup>28</sup> Due to limited data availability, we have to use FDI flows for the year 1997. We think that endogeneity is unlikely, however, because the decision to engage in FDI is often made some time before the actual investments are carried out.

In case we omit  $ij$ -specific variables that are correlated with both  $m_{ij}$  and  $M_{ij}$ , the migrant network is endogenous to the subsequent migrant flow. In view of our extended FE specification, it is difficult 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 country  $i$ 's migrant network in province  $j$  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).<sup>29</sup>

Because it indicates a large historical network, internal migration can be expected to correlate positively with the migrant network in 1996.<sup>30</sup> 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.<sup>31</sup> 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 across 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

---

<sup>28</sup>Hence, the effect of FDI on migration is not identified in the model controlling for country-and-region fixed effects.

<sup>29</sup>The year 1988 is the first year for which these information are 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.

<sup>30</sup>It follows from its definition, however, that internal migration also reduces the size of the historical network.

<sup>31</sup>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.

instrument. This allows us to perform tests on overidentifying restrictions and check for instrument exogeneity.

### 3.2 Skill Structure of Migration

Aggregate migration data with reliable information on the skill structure of migration can only be constructed at the level of regions rather than at the level of provinces. We therefore simplify the structure of our model to a two-level NMNL model in which the regions of destination (indexed here by  $j$ ) are the final migration destinations within the primary nest of Spain. Equation (23) then becomes:

$$\ln \left( \frac{m_{ij}^h}{m_{ij}^l} \right) = \frac{\theta \gamma^*}{\lambda_z} \ln(1 + M_{ij}) + \ln \left( \frac{m_i^h}{m_i^l} \right) - \gamma^* c_{iz} - \frac{\gamma^*}{\lambda_z} c_{ij} - \Psi_i^* - (1 - \lambda_z) \Omega_{iz}^*. \quad (24)$$

The dependent variable measures the skill structure of migration. Skill-specific migration flows are obtained from the National Immigrant Survey 2007 (NIS). The survey gathers unique information on a total of 15,465 migrants through field interviews conducted between November 2006 and February 2007; see Reher & Requena (2009, 255-261) for this and the following information.<sup>32</sup> Migrants report, *inter alia*, their year of arrival in Spain, their first destination in Spain, as well as their highest level of education they completed before migrating. 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.<sup>33</sup> Importantly, this definition is independent of the individual's legal status, so the data again include documented and undocumented migrants. 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 in that we consider only migrants who arrived in Spain between January 1, 2002, and December 31, 2006.

The migrant network,  $M_{ij}$ , is measured by the number of settled migrants as of January 1, 2002. These data, detailed by country of origin and region of destination, are taken from the Spanish Municipal Register. The sum of import and export values in 2001 is collected at the level of regions.

---

<sup>32</sup>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).

<sup>33</sup>Foreign-born individuals with Spanish nationality from birth who migrated to Spain within two years after birth are not considered as migrants.

Investment stocks as of 2001 are approximated by gross FDI inflows from the beginning of 1998 until the end of 2001. Country-specific fixed effects, absorbing, among other things, the multilateral resistance terms  $\Psi_i^*$  and  $\Omega_{iz}^*$ , are wiped out by applying the corresponding within-transformation to the data. Hence, cross-regional differences in the migrant network of a given country of origin are used as identifying variation so that we cannot control for country-and-region fixed effects. We instead augment the model by observable variables that are likely to influence the migration costs. We control for the geographical distance between the country of origin  $i$  and the region of destination  $j$ , using the STATA module GEODIST by Picard (2010) in combining geographical data on the countries of origin from Mayer & Zignago (2006) and on the regions of destination from the Spanish Wikipedia/GeoHack webpage. We control for a common language through an indicator variable that is equal to one if at least 80% of the region's total population are native speakers of a language spoken by at least 20% of the people living in the country of origin, and zero otherwise. The information on native languages in Spain are taken from a number of recent survey studies.<sup>34</sup> Language information on the countries of origin come from Mayer & Zignago (2006). The influence of all terms indexed  $j$  is absorbed by a set of dummy variables for the different regions of destination. The complete specification of our model furthermore controls for world region-and-region fixed effects.

We also apply the instrumental variables approach to this model, by analogy to the model for the scale of migration. In particular, we instrument the migrant network in 2002,  $M_{ij}$ , with the log of the number of people holding country  $i$ 's nationality and migrating from region  $j$  in Spain to any other region  $k \neq j$  in Spain in 1988. As before, we use the corresponding migration flow in 1989 as a second excluded instrument.

## 4 Estimation Results

In this section we present and discuss our estimation results. We start with a descriptive look at the relationship between migrant networks and the scale and skill structure of migration to different destinations in Spain. Figure 1(a) is a scatter plot for migration between 1997 and 2006 versus migrant networks in 1996, where each dot represents a different pair of country of origin and province of destination. We observe a positive correlation between the two variables. Figure 1(b) is a scatter plot for the skill structure of migration between 2002 and 2006 versus migrant networks at the beginning of 2002, where now each dot represents a different pair of country of origin and region of destination. The figure suggests a weak negative correlation between the two variables. In what follows, we test whether these correlations reflect a causal relationship running from migrant networks to the scale

---

<sup>34</sup>See Table C.2 in Appendix C for a list of surveys.

and skill structure of migration, and we provide a structural interpretation of our estimation results in terms of our NMNL migration model. We also discuss the results of several robustness checks.<sup>35</sup>

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

#### 4.1 Results for the Scale of Migration

In this subsection we present the estimation results of the model for the scale of migration as specified in equation (21). We first estimate an *average* network coefficient, abstracting from potential differences in the parameter  $\kappa_r$  across regions. Tables 1 and 2 show the results from the FE model and the two stage least squares (2SLS) FE model, respectively. In columns (a) and (b) of both tables, we eliminate country fixed effects via an adequate within-transformation of the data. The number of observations is equal to 2,592, which is the result of having 55 countries of origin, 50 provinces of destination, and 158 undefined values for the dependent variable due to zero migrant flows ( $55 \times 50 - 158 = 2,592$ ). In columns (c) to (f), we eliminate country-and-region fixed effects by modifying the within-transformation accordingly. This excludes all regions consisting of a single province and thus reduces the number of observations to 2,209.<sup>36</sup>

In the most parsimonious specification of the FE model in column (a) of Table 1, the estimated network coefficient is equal to 0.688.<sup>37</sup> The coefficient is statistically significant at the 1% level and estimated with very high precision (heteroskedasticity-robust standard error, clustered by countries of origin, equal to 0.029). When we augment the model by FDI and trade flows in column (b), we find a positive and statistically significant coefficient of the FDI variable. Yet, the point estimate of this coefficient is equal to 0.012 and thus implies a moderate quantitative importance only. Trade relations, instead, do not seem to have a significant impact on the scale of migration. More importantly, the estimates of the network coefficient are virtually unchanged in this version of the model. However, once we control for country-and-region fixed effects in columns (c) and (d), we see a drop in the estimated network coefficient down to 0.539, which corresponds to a decrease by roughly 20%. We see a further reduction by more than 10% once we take out the variation that is constant for each pair of world regions of origin and provinces of destination via dummy variables.

Unobserved heterogeneity in our model has two sources: first, the multilateral resistance terms, and second, the different cost components. Failing to account for the multilateral resistance terms leads to downward-biased estimates of the network coefficient due to a positive covariance between the

<sup>35</sup>The detailed results of these robustness checks are available from the others upon request.

<sup>36</sup>Seven regions consist of a single province. Applying the within-transformation to such observations yields all zeros.

<sup>37</sup>This estimate of the average network coefficient is virtually identical to the local network externality estimated by Beine et al. (2012).

migrant network and the terms  $\Psi_i$ ,  $\Omega_{iz}$ , and  $\Phi_{ir}$ , respectively. Failing to account for the different cost components, in turn, leads to upward-biased estimates of the network coefficient due to a negative covariance between the migrant network and the terms  $c_{iz}$ ,  $c_{ir}$ , and  $c_{ij}$ , respectively. Given that 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>>

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 0.732 and 0.958. 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 (cf. Hausman, 2001). As in the FE estimations, the network coefficient is lowest when we control 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 if interpreted relative to the FE model. The effects of both trade and FDI on the scale of migration are essentially zero.

The 2SLS diagnostics are all encouraging. The first-stage  $F$  statistic for the joint significance of the excluded instruments is relatively high and thus points to the relevance and strength of the instruments. In all the specifications employed, it exceeds the critical value of 10, which is required for reliable inference in the case of a single endogenous regressor (Stock et al., 2002, 522). Wooldridge’s robust score  $\chi^2$  test of overidentifying restrictions checks for instrument exogeneity. The null hypothesis (exogeneity) of this test can never be rejected at any reasonable significance level. This suggests that our instruments are uncorrelated with the structural error term, and that our structural equation is correctly specified. We also report the results from an exogeneity test for the migrant network. The robust regression-based  $F$  test rejects the null hypothesis that the migrant network is exogenous at the 1% level. It should thus be treated as endogenous.

Our next specification allows for cross-regional differences in the similarity parameter  $\kappa_r$ , which implies region-specific network coefficients,  $\eta_{zr}$ . The specification employed is equivalent to the one reported in column (f) of Table 1, except for the fact that we now interact the migrant network with dummy variables for the different regions of destination. Table 3 reveals substantial heterogeneity in the estimated network coefficient across regions. It is largest for the region of Cataluña (0.795) and smallest for the region of Extremadura (0.155).<sup>38</sup> Hence, individuals seem to consider the provinces

---

<sup>38</sup>In the estimation, the region of Cataluña serves as the reference region. The differences between the network coefficients estimated for Cataluña and for either of the other regions (except for the regions of Comunitat Valenciana and Canarias) are statistically significant at least at the 10% level according to  $t$ -tests.

in the region of Cataluña (Barcelona, Girona, Lleida, and Tarragona) to be very similar to each other, relative to the provinces in the region of Extremadura (Badajoz and Cáceres). This result accords with the pronounced autonomy of Cataluña in terms of its political and cultural life. It is not surprising either 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. At any rate, the large and significant cross-regional differences in the estimated network coefficient show 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:

$$\begin{aligned} \frac{\partial \ln(m_{ij})}{\partial \ln(1 + M_{ik})} = & \theta \left[ \frac{I(j, k)}{\lambda_z \kappa_r} - \left( \frac{m_{ik}}{m_i} \right) \right. \\ & \left. - \frac{I(\ell, r)}{\lambda_z \kappa_r} (1 - \kappa_r) \left( \frac{m_{ik}}{m_{ir}} \right) - \frac{I(y, z)}{\lambda_z} (1 - \lambda_z) \left( \frac{m_{ik}}{m_{iz}} \right) \right]. \end{aligned} \quad (25)$$

The network elasticity ( $j = k$ ) is a function of the network parameter  $\theta$ , the similarity parameters  $\kappa_r$  and  $\lambda_z$ , and the relative attractiveness of the province of destination  $j$  (reflected by the shares  $m_{ij}/m_i$ ,  $m_{ij}/m_{ir}$ , and  $m_{ij}/m_{iz}$ ). Neither  $\kappa_r$  nor  $\lambda_z$  can be estimated directly due to the use of aggregate migration data. This implies an uncertainty about the true network *elasticity*, which would prevail even if the true network *coefficient*,  $\eta_{zr}$ , was known with certainty.<sup>39</sup> However, we can compute estimates of the upper and lower bounds for this elasticity, separately for each region of destination. For this purpose, we use the results reported in Table 3 in order 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  $\kappa_r$  can take on any value between zero and one, while the similarity parameters for all other regions  $\kappa_\ell, \ell \neq r$ , 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 by region of destination as a function of the similarity parameter of the region of Extremadura,  $\kappa_r$ . The exact value of  $\kappa_r$  is unknown, but fixing this parameter also fixes the similarity parameters of all other regions. In order to focus on the

---

<sup>39</sup>Schmidheiny & Brühlhart (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  $\lambda$  is unknown, the elasticities of the Poisson model and of the standard MNL model can thus serve as boundary values for the true elasticities.

heterogeneity in the network elasticity that is due to differences in the similarity parameters across regions, we have imposed the following assumptions: first, there are 200 countries of destination outside the country of origin  $i$ ; second, each of these countries consists of 51 provinces that are uniformly distributed across 17 regions; and third, all provinces abroad are equally attractive destinations, with an overall fraction of migrants in the total population 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 the provinces in the region of Extremadura, we find a network elasticity that slightly exceeds a value of 0.1. For the provinces in the region of Cataluña, the same elasticity lies in the vicinity of 0.55. These are quite large differences. For any given region, the difference between the upper and the lower bound (i.e., the permissible range) of the network elasticity is roughly equal to 0.05, so the uncertainty about the network elasticity is a minor issue here. Importantly, the figure also incorporates the uncertainty about the country-specific similarity parameter  $\lambda_z$ , which can take on any value between zero and one. This uncertainty, which turns out to be almost irrelevant for the computation of the network elasticity, is reflected in the thickness of the upward-sloping lines.<sup>40</sup>

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

We have also computed the cross-elasticities of the network based on (25), by analogy to the network elasticity. Cross-elasticities for two provinces belonging to one of the regions listed in Table 3 are depicted in Figure 2(b). For the provinces in the region of Extremadura, we find an extremely low cross-elasticity that ranges between 0.0 and -0.05. For the provinces in the region of Cataluña, the same cross-elasticity lies between -0.22 and -0.27. In Figures D.1(a) and D.1(b) in Appendix D, we also depict the cross-elasticities when the two provinces  $j$  and  $k$  are located in different regions of the same country and when they are located in different countries, respectively. These cross-elasticities are not specific to any region of destination in Spain, they are lower (in absolute value) than the cross-elasticities depicted in Figure 2(b), and they are characterized by a higher uncertainty about their true values.

## Robustness Analysis

We have conducted two types of robustness checks. Both of them seem to indicate, if anything, a slightly larger average network coefficient than do our estimates in Tables 1 and 2. The first robustness check addresses a potential estimation bias due to non-stochastic measurement errors in our migration data. The migration data that we have considered above covers the period 1997-2006. To the extent

---

<sup>40</sup>Individual lines are upward-sloping because, for a given similarity parameter  $\lambda_z$  and a given estimate of the network coefficient  $\eta_{zr}$ , a larger similarity parameter  $\kappa_r$  for the region of Extremadura is only compatible with a larger network parameter  $\theta$ .

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 E 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 clusters. However, we have also employed the migrant network as of January 2002 along with the migrant flow from 2002 to 2006.<sup>41</sup>

In a second robustness check, we have applied alternative sample selection criteria in order to see whether our results suffer from endogenous sample selection. In particular, we have considered all observations (country-province pairs) with a migrant network of more than either 10, 20, or 50 migrants in the year 1996.<sup>42,43</sup> Applying these criteria results in unbalanced samples of 98, 90, or 74 countries, respectively.

## 4.2 Results for the Skill Structure of Migration

Table 4 reports the results from FE estimations of our model for the skill structure of migration as specified in equation (24). The full data matrix would contain 935 pairs of 55 countries of origin and 17 regions of destination. However, for some observations we lack the information on the migrant skill ratio (the dependent variable) due to the limited sample size of the NIS. The FE estimator is therefore applied to 241 observations with non-missing values for the migrant skill ratio. In all the specifications employed in Table 4, we find a robustly significant negative impact of migrant networks on the skill structure of migration, as suggested by theory. The estimated coefficient varies between -0.506 and -0.637, so the differences across specifications are rather small in magnitude. Neither the trade variable nor the FDI variable turns out to be statistically significant. This accords with the poorly suggestive evidence in favor of a positive effect of trade or FDI on the scale of migration. 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 within-cluster variation.

<<Tables 4 and 5 about here>>

Table 5 reports the results from the 2SLS FE estimations. They do not alter our causal interpretation in any significant way. As with the previous model for the scale of migration, the first-stage  $F$  test and the test on overidentifying restrictions suggest that our instruments are both relevant and

---

<sup>41</sup>For trade and FDI flows we have used the observations from 2001.

<sup>42</sup>Sample selection based on explanatory variables is a type of exogenous sample selection (cf. Wooldridge, 2009, 323).

<sup>43</sup>Identification requires, of course, that we have at least two observations within each cluster.

exogenous. 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 between -0.374 and -0.609 and are thus found to be slightly smaller than those obtained from the FE estimations. In the full specification of the model in columns (e) and (f), the migrant network is the only structural explanatory variable whose effect is statistically different from zero.

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

$$\frac{\partial \ln(m_{ij}^h/m_{ij}^l)}{\partial \ln(1 + M_{ij})} = \theta\gamma^* \left[ \frac{1}{\lambda_z} - \left( \frac{m_{ij}}{m_i} \right) - \frac{1 - \lambda_z}{\lambda_z} \left( \frac{m_{ij}}{m_{iz}} \right) \right], \quad (26)$$

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 fraction of migrants in the total population equal to three percent.<sup>44</sup> Then, because the similarity parameter  $\lambda_z$  can take on any value between zero and one, an estimated coefficient of the 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. First, we have tested for sample selection bias that could be due to the large number of missing values for the migrant skill ratio. We have found contrary evidence, using a Heckman (1976)-style procedure similar to the one proposed by Wooldridge (1995, 123-124).<sup>45</sup> This procedure is described in detail in Appendix F. Second, following the methodology proposed by Grogger & Hanson (2011, 53-54), we have excluded the possibility that individuals group regions of destination into nests at the sub-country level. To do so, we have repeatedly estimated the scale model as given by equation (21), using regional data instead of provincial 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. Third, we have restricted the sample to observations for which the dependent variable is constructed on the basis of at least ten migrants in the underlying survey data. The negative and significant effect of migrant networks on the skill structure of migration proves to be robust to this restriction, even though it reduces the sample size down to 110 observations.

<sup>44</sup>This implies that  $m_{ij}/m_i = 3/340,000$  and  $m_{ij}/m_{iz} = 1/17$ .

<sup>45</sup>Technically, the two-step Heckman procedure for testing and correcting for sample selection bias could be applied if the country fixed effects were not differenced out but, rather, if they were estimated by including a set of country dummy variables. However, this approach would result in inconsistent estimates due to the incidental parameters problem described in Neyman & Scott (1948).

Finally, we have estimated 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 (10) and (11) 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 regional nest and on the within-nest 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 structure of migration.<sup>46</sup>

## 5 Conclusion

In this paper, we have documented strong positive network effects on the scale of migration and a strong negative effect on the ratio of high-skilled to low-skilled migrants. Both types of effects are robust across alternative estimators, estimation samples, and sets of control variables. Our identification strategy is based on a three-level NMNL model that allows for varying degrees of substitutability across alternative migration destinations. The ease with which one destination in Spain can be substituted by another one depends on whether the two destinations are located in the same region or not; in case they are, it also depends on the degree of political and cultural autonomy of that region. Our approach is corroborated by the significant degree of heterogeneity in the estimated network elasticity across regions.

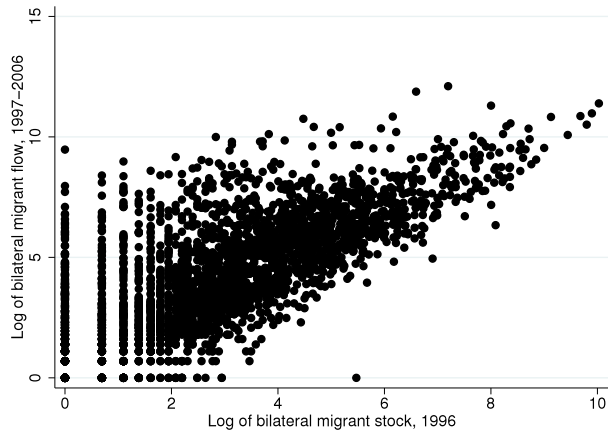
Our findings add to the understanding of the recent migration phenomenon in Spain. This migration has gained momentum through Spain's strong economic growth in the years before the Global Financial Crisis. It has changed the size and composition of the country's population and labor supply, with potentially important effects on a number of key macroeconomic variables such as wages, unemployment, and production, as well as on the national welfare state. The recent economic recession in Spain is reflected in a sharp decline in new migration and a significant amount of return migration in the very short run. The analysis of the structural relationships among past migration, future migration, and the labor market outcomes involves non-trivial dynamics. Attempts to study these dynamics seem to appear as a challenging yet promising avenue for future research.

---

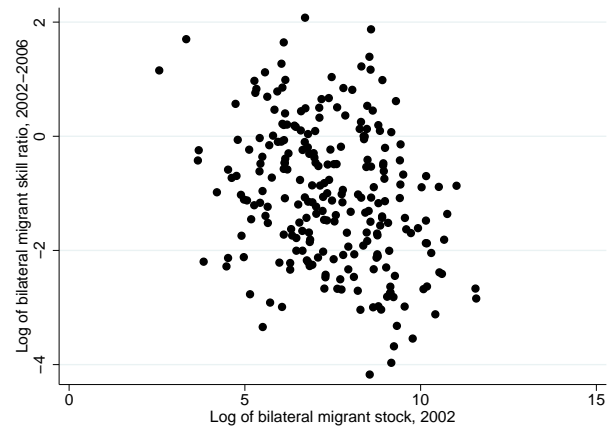
<sup>46</sup>We have also experimented with two alternative estimation approaches following Quigley (1976) and Lerman (1976). Both include the full set of regions in Spain and are summarized in McFadden (1978, 91-94). Again, we have obtained a robustly significant, negative impact of migrant networks on the skill structure of migration.

# Figures and Tables

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

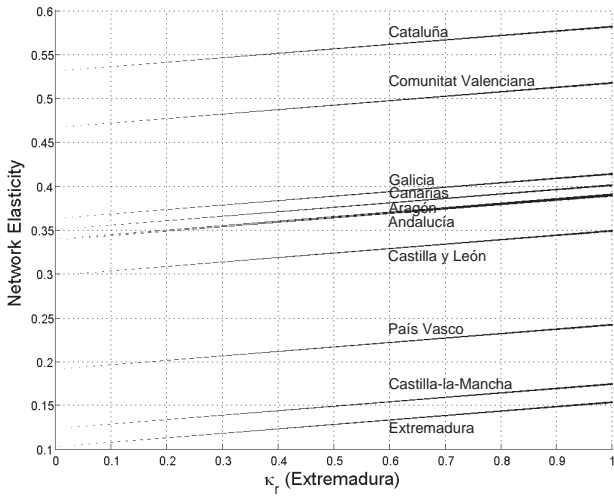


(a)  $\ln(m_{ij})$  plotted against  $\ln(1+M_{ij})$ , provincial level

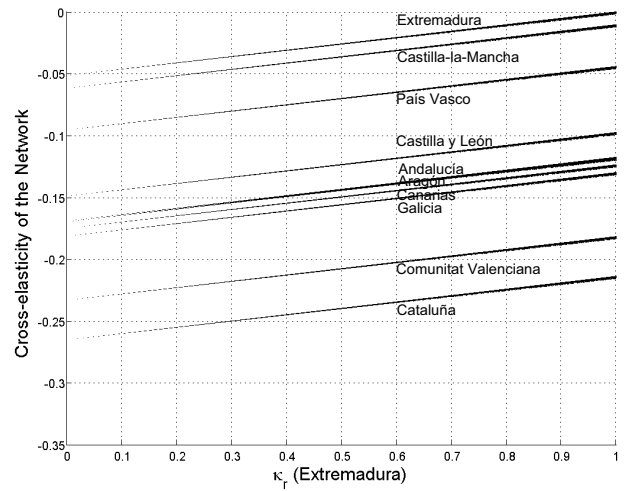


(b)  $\ln(m_{ij}^h/m_{ij}^l)$  plotted against  $\ln(1+M_{ij})$ , regional level

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<sup>†</sup>

	<i>Dependent Variable: Migration Flow (Province-Level 1997-2006)</i>					
	(a)	(b)	(c)	(d)	(e)	(f)
<i>Stock of Migrants</i> (Province-Level 1996)	0.688*** (0.029)	0.682*** (0.029)	0.539*** (0.029)	0.539*** (0.029)	0.469*** (0.035)	0.469*** (0.035)
<i>FDI Flow</i> (Region-Level 1997)		0.012** (0.005)				
<i>Trade Flow</i> (Province-Level 1996)		0.005 (0.007)		0.004 (0.007)		0.008 (0.007)
Constant	2.357*** (0.124)	2.215*** (0.171)	2.566*** (0.089)	2.619*** (0.139)	2.322*** (0.125)	2.313*** (0.162)
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 E.	No	No	No	No	Yes	Yes
Observations	2,592	2,592	2,209	2,209	2,209	2,209
Within $R^2$	0.791	0.792	0.670	0.670	0.764	0.764

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

Table 2: Scale of Migration – 2SLS FE Model<sup>†</sup>

	<i>Dependent Variable: Migration Flow (Province-Level 1997-2006)</i>					
	(a)	(b)	(c)	(d)	(e)	(f)
<i>Stock of Migrants</i> (Province-Level 1996)	0.958*** (0.068)	0.955*** (0.069)	0.826*** (0.078)	0.829*** (0.079)	0.732*** (0.096)	0.735*** (0.097)
<i>FDI Flow</i> (Region-Level 1997)		0.004 (0.005)				
<i>Trade Flow</i> (Province-Level 1996)		0.005 (0.007)		0.007 (0.008)		0.010 (0.007)
Constant	0.169 (0.117)	0.156 (0.120)	0.107 (0.097)	0.112 (0.098)	0.047 (0.103)	0.053 (0.103)
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 E.	No	No	No	No	Yes	Yes
Observations	2,592	2,592	2,209	2,209	2,209	2,209
Within $R^2$	0.769	0.769	0.632	0.631	0.740	0.740
Robust first-stage $F$ test	32.33	31.70	19.18	19.15	12.92	12.91
Test on Overidentifying R.						
Robust score $\chi^2$ test	0.014	0.022	0.467	0.416	0.308	0.243
- $p$ -value	0.905	0.881	0.494	0.519	0.579	0.622
Exogeneity Test						
Robust regression $F$ test	20.14	19.40	12.33	12.43	5.29	5.37
- $p$ -value	0	0	0.001	0.001	0.022	0.021

<sup>†</sup> All variables are in natural logs. Heteroskedasticity-robust standard errors (clustered by countries of origin or pairs of countries of origin and regions of destination) are given in parentheses. \*, \*\*, \*\*\* denote significance at the 10%, 5%, 1% levels, respectively. The regressions include all countries of origin with at least 630 nationals residing in Spain in 1996 (55 countries of origin). 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: Estimated Network Coefficients, by Region<sup>†</sup>

Region $r$	Estimate of $\eta_{zr}$	Region $r$	Estimate of $\eta_{zr}$
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

<sup>†</sup> This table reports region-specific estimates of the network coefficient,  $\eta_{zr}$ . The specification employed is equivalent to that 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 2,209, and the within  $R^2$  is 0.771.

Table 4: Skill Structure of Migration – FE Model<sup>†</sup>

	<i>Dependent Variable: Migrant Skill Ratio (Region-Level 2002-2006)</i>					
	(a)	(b)	(c)	(d)	(e)	(f)
<i>Stock of Migrants</i> (Region-Level 2002)	-0.513*** (0.090)	-0.510*** (0.089)	-0.506*** (0.093)	-0.626*** (0.110)	-0.637*** (0.106)	-0.621*** (0.115)
<i>FDI Flow</i> (Region-Level 1998-2001)			-0.006 (0.020)			-0.012 (0.018)
<i>Trade Flow</i> (Region-Level 2001)			-0.001 (0.084)			0.080 (0.112)
<i>Language</i> (Region-Level)		0.248 (0.221)	0.246 (0.223)		0.463** (0.175)	0.559*** (0.154)
<i>Distance</i> (Region-Level)		-0.636 (0.394)	-0.657 (0.392)		-1.450 (1.358)	-1.388 (1.353)
Constant	2.991*** (0.729)	8.216** (3.388)	8.443** (3.894)	3.733*** (0.857)	15.770 (11.275)	13.755 (11.692)
Region Effects	Yes	Yes	Yes	Nested	Nested	Nested
Country Effects	Yes	Yes	Yes	Yes	Yes	Yes
World Region-and-Region E.	No	No	No	Yes	Yes	Yes
Observations	241	241	241	241	241	241
Within $R^2$	0.245	0.261	0.261	0.466	0.477	0.481

<sup>†</sup> All variables except for the language dummy are in natural logs. Heteroskedasticity-robust standard errors (clustered by countries of origin) 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 Structure of Migration – 2SLS FE Model<sup>†</sup>

	<i>Dependent Variable: Migrant Skill Ratio (Region-Level 2002-2006)</i>					
	(a)	(b)	(c)	(d)	(e)	(f)
<i>Stock of Migrants</i> (Region-Level 2002)	-0.374*** (0.144)	-0.382*** (0.145)	-0.405** (0.169)	-0.506** (0.214)	-0.579** (0.238)	-0.609** (0.265)
<i>FDI Flow</i> (Region-Level 1998-2001)			0.005 (0.022)			-0.003 (0.022)
<i>Trade Flow</i> (Region-Level 2001)			0.063 (0.070)			0.094 (0.074)
<i>Language</i> (Region-Level)		0.134 (0.205)	0.158 (0.199)		0.010 (0.353)	0.084 (0.313)
<i>Distance</i> (Region-Level)		-0.649* (0.386)	-0.562 (0.380)		-0.927 (0.573)	-0.824 (0.552)
Constant	0.077 (0.177)	0.077 (0.183)	0.033 (0.173)	0.143 (0.206)	0.194 (0.226)	0.137 (0.214)
Region Effects	Yes	Yes	Yes	Nested	Nested	Nested
Country Effects	Yes	Yes	Yes	Yes	Yes	Yes
World Region-and-Region E.	No	No	No	Yes	Yes	Yes
Observations	241	241	241	241	241	241
Within $R^2$	0.208	0.220	0.225	0.412	0.417	0.419
Robust first-stage $F$ test	24.11	19.77	13.57	14.48	11.42	10.34
Test on Overidentifying R.						
Robust score $\chi^2$ test	1.070	0.769	0.909	0.310	0.284	0.430
- $p$ -value	0.301	0.381	0.340	0.577	0.594	0.512
Exogeneity Test						
Robust regression $F$ test	0.794	0.867	0.860	0.873	0.678	0.618
- $p$ -value	0.070	0.029	0.032	0.026	0.175	0.253

<sup>†</sup> All variables except for the language dummy are in natural logs. Heteroskedasticity-robust standard errors (clustered by countries of origin) 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] Anderson, James E., “The Gravity Model,” *Annual Review of Economics*, 3 (2011), 133–160.
- [2] Anderson, James E., and Eric van Wincoop, “Gravity with Gravitas: A Solution to the Border Puzzle,” *American Economic Review*, 93:1 (2003), 170–192.
- [3] Åslund, Olof, “Now and forever? Initial and Subsequent Location Choices of Immigrants,” *Regional Science and Urban Economics*, 35:2 (2005), 141–165.
- [4] Baghdadi, Leila, “Mexico-U.S. Migration: Do Spatial Networks Matter?,” Universite Paris I. mimeograph (2005).
- [5] Bartel, Ann P., “Where Do the New U.S. Immigrants Live?,” *Journal of Labor Economics*, 7:4 (1989), 371–391.
- [6] 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.
- [7] Bauer, Thomas, Gil S. Epstein, and Ira N. Gang, “Measuring Ethnic Linkages among Migrants,” *International Journal of Manpower*, 30:1/2 (2009), 56–69.
- [8] Beine, Michel, Frédéric Docquier, and Çağlar Özden, “Diasporas,” *Journal of Development Economics*, 95:1 (2011), 30–41.
- [9] Beine, Michel, Frédéric Docquier, and Çağlar Özden, “Dissecting Network Externalities in International Migration,” University of Luxembourg mimeograph (2012).
- [10] Beine, Michel, and Sara Salomone, “Network Effects in International Migration: Education versus Gender,” *Scandinavian Journal of Economics*, 115:2 (2013), 354–380.
- [11] Bertoli, Simone, “Networks, Sorting and Self-selection of Ecuadorian Migrants,” *Annals of Economics and Statistics*, 2010:97/98 (2010), 261–288.
- [12] Bertoli, Simone, and Jesús Fernández-Huertas Moraga, “Visa Policies, Networks and the Cliff at the Border,” IZA Discussion Paper No. 7094 (2012).
- [13] Bertoli, Simone, and Jesús Fernández-Huertas Moraga, “Multilateral Resistance to Migration,” *Journal of Development Economics*, 102 (2013), 79–100.
- [14] Bertoli, Simone, Herbert Brücker, and Jesús Fernández-Huertas Moraga, “The European Crisis and Migration to Germany: Expectations and the Diversion of Migration Flows,” IZA Discussion Paper No. 7170 (2013).

- [15] 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.
- [16] Carrington, William J., Enrica Detragiache, and Tara Vishwanath, “Migration with Endogenous Moving Costs,” *American Economic Review*, 86:4 (1996), 909–930.
- [17] Chau, Nancy, “The Pattern of Migration with Variable Migration Cost,” *Journal of Regional Science*, 37:1 (1997), 35–54.
- [18] Chiswick, Barry R., “Are Immigrants Favorably Self-Selected?,” *American Economic Review: Papers and Proceedings*, 89:2 (1999), 181–185.
- [19] Chiswick, Barry R., and Paul W. Miller, “Where Immigrants Settle in the United States,” *Journal of Comparative Policy Analysis*, 6:2 (2004), 185–197.
- [20] 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.
- [21] 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.
- [22] 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.)
- [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] Massey, Douglas S., “Economic Development and International Migration in Comparative Perspective,” *Population and Development Review*, 14:3 (1988), 383–413.
- [32] Mayda, Anna M., “International Migration: A Panel Data Analysis of the Determinants of Bilateral Flows,” *Journal of Population Economics*, 23:4 (2010), 1249–1274.
- [33] Mayer, Thierry, and Soledad Zignago, “Notes on CEPII’s Distances Measures,” (2006).
- [34] McFadden, Daniel L., “Conditional Logit Analysis of Qualitative Choice Behavior,” in Paul Zarembka (Ed.), *Frontiers in Econometrics* (New York: Academic Press, 1974), 105–142.
- [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] Morales, María J.G., and Eduard R. Molés, “The Spanish ‘Autonomic State’,” in Jürgen Rose and Johannes Ch. Traut (Eds.), *Federalism and Decentralization: Perspectives for the Transformation Process in Eastern and Central Europe*, George C. Marshall European Center for Security Studies 2 (Münster: Lit, 2002), chapter 10, 179–192.

- [40] Munshi, Kaivan, “Networks in the Modern Economy: Mexican Migrants in the U.S. Labor Market,” *Quarterly Journal of Economics*, 118:2 (2003), 549–599.
- [41] Neubecker, Nina, and Marcel Smolka, “Co-national and Cross-national Pulls in International Migration to Spain,” *International Review of Economics & Finance* (forthcoming).
- [42] Neyman, Jerzy, and Elizabeth L. Scott, “Consistent Estimates Based on Partially Consistent Observations,” *Econometrica*, 16:1 (1948), 1–32.
- [43] OECD, *International Migration Outlook: SOPEMI 2010*, OECD (2010), Paris.
- [44] Ortega, Francesc, and Giovanni Peri, “The Causes and Effects of International Migrations: Evidence from OECD Countries 1980-2005,” NBER Working Paper No. 14833 (2009).
- [45] Ortega, Francesc, and Giovanni Peri, “The Effect of Income and Immigration Policies on International Migration,” *Migration Studies*, 1:1 (2013), 47-74.
- [46] 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.
- [47] Pesaran, M. Hashem, “Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure,” *Econometrica*, 74:4 (2006), 967–1012.
- [48] Picard, Robert, “GEODIST: Stata Module to Compute Geodetic Distances,” Statistical Software Components, Boston College Department of Economics (2010).
- [49] 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.
- [50] 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.
- [51] Santos Silva, João M.C., and Silvana Tenreyro, “The Log of Gravity,” *Review of Economics and Statistics*, 88:4 (2006), 641–658.
- [52] 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.
- [53] 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.

- [54] Vovsha, Peter, “Application of Cross-nested Logit Model to Mode Choice in Tel Aviv, Israel, Metropolitan Area,” *Transportation Research Record*, 1607 (1997), 6–15.
- [55] Wen, Chieh-Hua, and Frank S. Koppelman, “The Generalized Nested Logit Model,” *Transportation Research Part B*, 35:7 (2001), 627–641.
- [56] Wooldridge, Jeffrey M., “Selection Corrections for Panel Data Models Under Conditional Mean Independence Assumptions,” *Journal of Econometrics*, 68:1 (1995), 115–132.
- [57] Wooldridge, Jeffrey M., *Introductory Econometrics. A Modern Approach*, Fourth Edition, International Student Edition (South-Western Cengage Learning, 2009).
- [58] Zavodny, Madeline, “Welfare and the Locational Choices of New Immigrants,” *Federal Reserve Bank of Dallas Economic Review*, (1997), 2–10.
- [59] Zavodny, Madeline, “Determinants of Recent Immigrants’ Locational Choices,” *International Migration Review*, 33:4 (1999), 1014–1030.

## A Derivation of $\partial \ln H_i(\cdot)/\partial U_{ij}$

Since

$$\ln H_i(\cdot) = \ln \sum_z \left( \sum_{r \in A_z} \left( \sum_{j \in A_{zr}} \exp[U_{ij}/(\kappa_r \lambda_z)] \right)^{\kappa_r} \right)^{\lambda_z} \quad (\text{A.1})$$

we have

$$\frac{\partial \ln H_i(\cdot)}{\partial U_{ij}} = H_i(\cdot)^{-1} \exp[U_{ij}/(\kappa_r \lambda_z)] Q X, \quad (\text{A.2})$$

where

$$\begin{aligned} Q &= \left( \sum_{j \in A_{zr}} \exp[U_{ij}/(\kappa_r \lambda_z)] \right)^{\kappa_r - 1} \\ &= (\exp[(-c_{iz} - c_{ir})/(\kappa_r \lambda_z)])^{\kappa_r - 1} \left( \sum_{j \in A_{zr}} \exp[\xi_{ij}/(\kappa_r \lambda_z)] \right)^{\kappa_r - 1} \end{aligned} \quad (\text{A.3})$$

and

$$\begin{aligned} X &= \left( \sum_{r \in A_z} \left( \sum_{j \in A_{zr}} \exp[U_{ij}/(\kappa_r \lambda_z)] \right)^{\kappa_r} \right)^{\lambda_z - 1} \\ &= (\exp[-c_{iz}/\lambda_z])^{\lambda_z - 1} \left( \sum_{r \in A_i} (\exp[-c_{ir}/\lambda_z]) \left( \sum_{j \in A_{zr}} \exp[\xi_{ij}/(\kappa_r \lambda_z)] \right)^{\kappa_r} \right)^{\lambda_z - 1}. \end{aligned} \quad (\text{A.4})$$

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.2) can be written as:

$$\begin{aligned} \frac{\partial \ln H_i(\cdot)}{\partial U_{ij}} &= \frac{\exp[\xi_{ij}/(\kappa_r \lambda_z) - c_{ir}/\lambda_z - c_{iz}]}{H_i(\cdot) \exp[(1 - \kappa_r)\Phi_{ir} + (1 - \lambda_z)\Omega_{iz}]} \\ &= \frac{\exp[\xi_{ij}/(\kappa_r \lambda_z) - c_{ir}/\lambda_z - c_{iz}]}{\exp[\Psi_i + (1 - \kappa_r)\Phi_{ir} + (1 - \lambda_z)\Omega_{iz}]}, \end{aligned} \quad (\text{A.5})$$

which gives  $P_i^o(j^o = j)$ , where  $j \in A_{zr}, r \in A_z$ ; see equations (8) and (16).

## B Derivation of $\partial \ln(m_{ij}/m_i)/\partial \ln Y_k$

In the following, we derive  $\partial \ln(m_{ij}/m_i)/\partial \ln Y_k$  for  $k = j \in A_{zr}, r \in A_z$ . The other (simpler) derivatives where  $k \neq j$  can be derived analogously. They depend on whether or not  $k \in A_{zr}$  and whether or not  $z = y$  if  $k \in A_{y\ell}, \ell \in A_y$ . Since

$$\ln \left( \frac{m_{ij}}{m_i} \right) = \xi_{ij}/(\lambda_z \kappa_r) - c_{ir}/\lambda_z - c_{iz} - \Psi_i - (1 - \kappa_r)\Phi_{ir} - (1 - \lambda_z)\Omega_{iz} \quad (\text{B.1})$$

we have

$$\begin{aligned}
\frac{\partial \ln(m_{ij}/m_i)}{\partial \ln Y_k} &= \frac{Y_k}{\lambda_z \kappa_r} - \frac{\exp[\Omega_{iz} \lambda_z - c_{iz}] \lambda_z}{\exp[\Psi_i]} \frac{\partial \Omega_{iz}}{\partial \ln Y_k} - (1 - \kappa_r) \frac{\partial \Phi_{ir}}{\partial \ln Y_k} \\
&\quad - (1 - \lambda_z) \frac{\partial \Omega_{iz}}{\partial \ln Y_k} \\
&= \frac{Y_k}{\lambda_z \kappa_r} - \frac{m_{iz} \lambda_z}{m_i} \frac{\partial \Omega_{iz}}{\partial \ln Y_k} - (1 - \kappa_r) \frac{\partial \Phi_{ir}}{\partial \ln Y_k} - (1 - \lambda_z) \frac{\partial \Omega_{iz}}{\partial \ln Y_k}.
\end{aligned} \tag{B.2}$$

Since

$$\frac{\partial \Phi_{ir}}{\partial \ln Y_k} = \frac{\exp[\xi_{ik}/(\lambda_z \kappa_r)]}{\sum_{k \in A_{zr}} \exp[\xi_{ik}/(\lambda_z \kappa_r)]} \frac{Y_k}{\lambda_z \kappa_r} = \frac{m_{ik}}{m_{ir}} \frac{Y_k}{\lambda_z \kappa_r} \tag{B.3}$$

and

$$\frac{\partial \Omega_{iz}}{\partial \ln Y_k} = \frac{\exp[\Phi_{ir} \kappa_r - c_{ir}/\lambda_z] \kappa_r}{\sum_{\ell \in A_z} \exp[\Phi_{i\ell} \kappa_\ell - c_{i\ell}/\lambda_z]} \frac{\partial \Phi_{ir}}{\partial \ln Y_k} = \frac{m_{ir} \kappa_r}{m_{iz}} \frac{\partial \Phi_{ir}}{\partial \ln Y_k} \tag{B.4}$$

equation (B.2) can be written as:

$$\frac{\partial \ln(m_{ij}/m_i)}{\partial \ln Y_k} = Y_k \left( \frac{1}{\lambda_z \kappa_r} - \frac{m_{ik}}{m_i} - \frac{(1 - \kappa_r)}{\lambda_z \kappa_r} \frac{m_{ik}}{m_{ir}} - \frac{(1 - \lambda_z)}{\lambda_z} \frac{m_{ik}}{m_{iz}} \right). \tag{B.5}$$

## C Data Appendix

Table C.1: List of the 55 Countries Considered in the Empirical Analysis, by World Region

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

Table C.2: Data Sources

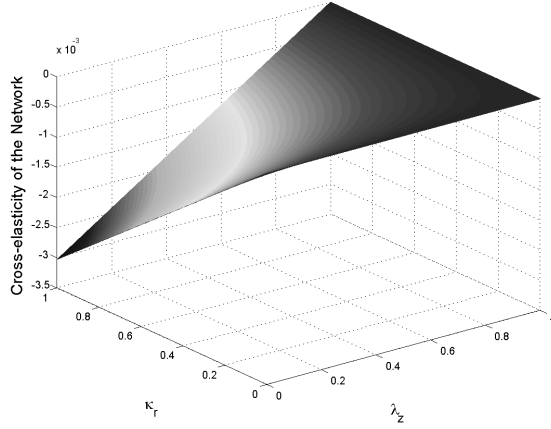
Variable	Definition	Data Sources
Migrant Flow $m_{ij}$	Migrants who registered at municipalities in Spain between January 1, 1997 (or January 1, 2002), and December 31, 2006, by province of destination (or region of destination) and by country of origin. Migrants are defined as individuals whose last country of residence (other than Spain) corresponds to their country of birth and nationality.	Spanish Residential Variation Statistics, INE, <a href="http://www.ine.es/en/prodyser/micro_varires_en.htm">http://www.ine.es/en/prodyser/micro_varires_en.htm</a> , accessed on 10/05/2010
Migrant Skill Ratio $m_{ij}^h/m_{ij}^l$	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, <a href="http://www.ine.es/prodyser/micro_inmigra.htm">http://www.ine.es/prodyser/micro_inmigra.htm</a> , accessed on 10/05/2010
Migrant Network $M_{ij}$	Number of settled migrants as of May 1, 1996 (or January 1, 2002), 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, <a href="http://www.ine.es/jaxi/menu.do?type=pcaxis&amp;path=%2Ft20%2Ft245&amp;file=inebase&amp;L=0">http://www.ine.es/jaxi/menu.do?type=pcaxis&amp;path=%2Ft20%2Ft245&amp;file=inebase&amp;L=0</a> , 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, <a href="http://datacomex.comercio.es/principal_comex.es.aspx">http://datacomex.comercio.es/principal_comex.es.aspx</a> , 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, <a href="http://datainvex.comercio.es/principal_invex.aspx">http://datainvex.comercio.es/principal_invex.aspx</a> , 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, <a href="http://www.ine.es/en/prodyser/micro_varires_en.htm">http://www.ine.es/en/prodyser/micro_varires_en.htm</a> , 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).	Spanish Wikipedia/GeoHack, <a href="http://es.wikipedia.org">http://es.wikipedia.org</a> , accessed on 09/05/2011; Mayer & Zignago (2006)

Table C.2 *continued*

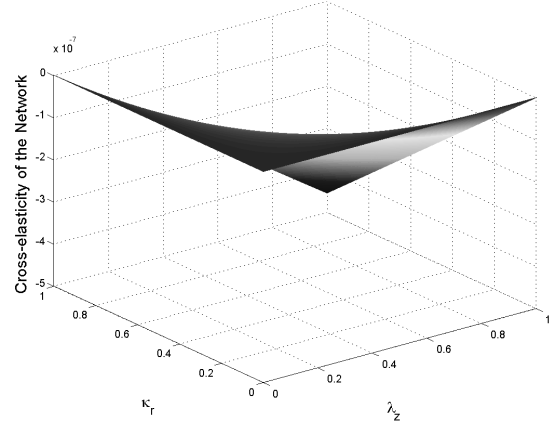
Variable	Definition	Data Sources
Indicator for Common Language	This variable is equal to one if at least 80% of a region's population in Spain are native speakers of a language spoken by at least 20% of the people in the country of origin; it is zero otherwise.	<p><i>Cataluña</i>: Generalitat de Catalunya,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

## D Counterfactual Cross-elasticities of the Migrant Network

Figure D.1: Counterfactual Cross-elasticities of the Migrant Network



(a) Cross-elasticities for  $j \in A_{zr}$  and  $k \in A_{z\ell}$ ,  $r \neq \ell$



(b) Cross-elasticities for  $j \in A_{zr}$  and  $k \in A_{y\ell}$ ,  $z \neq y$

## E 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:

$$\ln(\tilde{m}_{ij}) = \eta_{zr} \ln(\tilde{M}_{ij}). \quad (\text{E.1})$$

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:

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

which can be rewritten as:

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

The last two terms in equation (E.3), if not controlled for, may introduce a bias in the estimation of the network coefficient  $\eta_{zr}$ . Obviously, a sufficient condition for our FE model controlling for

country-and-region fixed effects to deliver unbiased estimates is:

$$v_{ij} = v_{ir}, \quad v = \{x, y\}. \quad (\text{E.4})$$

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  $j, k \in A_{zr}$ .

## F Testing for Sample Selection Bias

We briefly present our procedure for identifying a potential sample selection bias in the model for the skill structure 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 structure of migration as:

$$y_{ij} = \mu_i + \mathbf{x}_{ij}\boldsymbol{\beta} + u_{ij}, \quad j = 1, \dots, J, \quad (\text{F.1})$$

where  $y_{ij}$  is the  $ij$ -specific log of the ratio of high-skilled migrations to low-skilled migrants,  $\mu_i$  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  $E(\mu_i | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iJ}) \neq E(\mu_i)$ . Since  $J$  is fixed, the asymptotic analysis is valid for  $I \rightarrow \infty$ . Now suppose that  $(y_{ij}, \mathbf{x}_{ij})$  is sometimes unobserved, and that  $\mathbf{s}_{ij} = (s_{i1}, \dots, s_{iJ})'$  is a vector of selection indicators with  $s_{ij} = 1$  if  $(y_{ij}, \mathbf{x}_{ij})$  is observed and zero otherwise. Define  $\mathbf{x}_i \equiv (\mathbf{x}_{i1}, \dots, \mathbf{x}_{iJ})$  and  $\mathbf{s}_i \equiv (\mathbf{s}_{i1}, \dots, \mathbf{s}_{iJ})$  and suppose that  $E(u_{ij} | \mu_i, \mathbf{x}_i, \mathbf{s}_i) = 0 \quad \forall j$ , which implies that the selection process is strictly exogenous conditional on  $\mu_i$  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  $j = 1, \dots, J$ . The variable  $y_{ij}$  is observed if  $s_{ij} = 1$ , but not otherwise. For each  $j = 1, \dots, J$ , define an unobserved latent variable

$$h_{ij}^* = \delta_{j0} + \mathbf{x}_{i1}\delta_{j1} + \dots + \mathbf{x}_{iJ}\delta_{jJ} + v_{ij}, \quad (\text{F.2})$$

where  $v_{ij}$  is a stochastic term independent of  $(\mu_i, \mathbf{x}_i)$ , and  $\delta_{jp}$  is a  $(K+1) \times 1$  vector of unknown parameters,  $p = 1, 2, \dots, J$ .<sup>47</sup> 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  $\mathbf{v}_i \equiv (v_{i1}, \dots, v_{iJ})'$ , a sufficient condition for the selection process to be strictly exogenous conditional on  $\mu_i$  and  $\mathbf{x}_i$  is:

$$E(u_{ij}|\mu_i, \mathbf{x}_i, \mathbf{v}_i) = 0, \quad j = 1, \dots, J. \quad (\text{F.3})$$

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

$$E(u_{ij}|\mu_i, \mathbf{x}_i, \mathbf{v}_i) = E(u_{ij}|v_{ij}) = \rho v_{ij}, \quad j = 1, \dots, J, \quad (\text{F.4})$$

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

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

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,

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

and

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

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 (F.2), using linear probability models for the full data matrix. We then estimate equation (F.7) in an FE framework (within-

---

<sup>47</sup>In the following,  $\mathbf{x}_{ij}$  includes one element more than in equation (F.1), despite the fact that we use the same notation for convenience. We thus assume that there is exactly one exclusion restriction in equation (F.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.

transformed data), using only observations with  $s_{ij} = 1$ . We finally test  $H_0 : \rho = 0$ , using the  $t$ -statistic for  $\rho^*$ .