WPS5721

Policy Research Working Paper 5721

Diaspora Effects in International Migration

Key Questions and Methodological Issues

Michel Beine
Frédéric Docquier
Çağlar Özden

The World Bank
Development Research Group
Trade and Integration Team
June 2011

Policy Research Working Paper 5721

Abstract

This paper reviews the existing literature on the impact of migrants networks on the patterns of international migration. It covers the theoretical channels at stake in the global effect of the networks. It identifies the key issues, namely the impact on size, selection and concentration of the migration flows. The paper also reviews the empirical hurdles that the researchers face in assessing the importance of networks. The key issues concern the choice of micro vs a macro approach, the definition of a network, the access to suitable data and the adoption of econometric methods accounting for the main features of those data. Finally, the paper reports a set of estimation outcomes reflecting the main findings of the macro approach.

This paper is a product of the Trade and Integration Team, Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The author may be contacted at cozden@worldbank.org.

Diaspora effects in international migration: key questions and methodological issues.

Michel Beine[a], Frédéric Docquier[b] and Caglar Ozden[c]
[a]University of Luxemburg and CES-Ifo
[b]FNRS and IRES, Université Catholique de Louvain
[b]DECRG, The World Bank

1 Introduction

This paper investigates how existing diasporas or networks (i.e. stock of immigrants of own national background already resident in a given destination) impact the number, skill composition and geographical concentration of new migrants. The role of the diaspora on migration flows is well known and undisputed. The contribution of this paper is both methodological and empirical. Using global aggregate data, we show that diasporas not only influence the future flows but also their other characteristics, such as composition and concentration. We present various econometric and data challenges that are relevant in this literature, discuss how we try to resolve them and show how different empirical methods influence the results. In order to guide our thinking on the diaspora effects, we construct a unified yet simple theoretical framework based on various bilateral factors that influence migration costs. Among the main determinants of migration costs are distance, linguistic overlap, political bonds such as colonial links.

We argue that diaspora externalities influence migration patterns (flow size, skill composition and concentration) through their effect on bilateral migration costs. These diaspora externalities operate through two main channels. First, diasporas reduce assimilation and information costs for newcomers. They help them with jobs, housing, education and various cultural adjustment issues. Second, diasporas attract new migrants through family reunion programs and other venues that lower legal migration barriers. Assessing the combined effect of these two channels is key to understand the dynamics of the size and composition of migration flows.

Until recently, the analysis of diaspora externalities has been conducted at the micro level (see Boyd, 1989, Massey, 1993, Munshi, 2003, McKenzie and Rapoport, 2010). Individual or household level micro data have multiple advantages such as detailed information on demographic, human capital, social and economic parameters. However a major drawback of those micro approaches is that they can only focus on a limited number of migration corridors at time (e.g. the Mexican-US corridor) and can hardly be generalized to other country pairs. An exception is provided by Beine et al (2010) who use bilateral macrodata on migration flows and stocks. They took advantage of a recent data set on international migration by educational attainment to investigate the role of diaspora size on the educational structure of migration from 195 countries to the 30 OECD countries. Their paper shows that networks are by far the most important determinant of migration flows, explaining 71 percent of the observed variability of the size of migration flows, and 47% of the variability of the selection ratio in 2000.

In this paper, we extend the study of Beine et al. by looking at the diaspora effect on the geographical concentration or dispersion of new migrants, and by comparing results obtained with different estimation techniques. Indeed, a macro analysis of diaspora externalities raises multiple econometric problems. The main issues are the large number of empty corridors (due to truncation rules or true 'zeroes'), and the difficulty to identify causation (unobserved variables

are likely to affect the existing stock of migrants and the flows of newcomers). Several econometric techniques are available to address these issues; one of our goals is to evaluate the quantitative robustness of diaspora externalities to the choice of a particular method.

The remainder of this paper is organized as following. Section 2 depicts theory and key issues. Section 3 reviews the main empirical hurdles researchers face when studying the impact of diasporas. Section 4 discusses econometric issues while Section 5 presents some estimation results. Section 6 concludes.

2 Theory and key issues

We first consider a simplified version of the model use din Beine et. al. (2010) to describe how existing diasporas impact the size, skill composition, and geographic concentration of migration flows. An individual endowed with h units of human capital earns a wage $w_i h$ in country $i = 1, ..., I$, where w_i is the skill price in that country. The skill price is linked to labor productivity and the level of development of the country. This structure accommodates the assumptions that the main variation in wages within a country is due to differences in human capital levels (h) among workers whereas the main source variation across countries is due to the skill prices (w_i). The utility of a type-h individual working in his birth country (denoted o) is given by

$$u_{oo}(h) = w_o h + A_o + \varepsilon_o$$

where A_o is a variable capturing non-wage characteristics and amenities (such as climate) of the home country.

The utility obtained when the same person migrates to a destination country d is given by

$$u_{od}(h) = w_d h + A_d - C_{od}(h) + \varepsilon_d$$

where $C_{od}(h)$ denotes moving and assimilation costs that are borne by the migrant. Those costs depend on factors such as physical distance, destination and origin countries' social, cultural and linguistic characteristics. Assimilation costs are generally assumed to be decreasing with human capital ($\partial C_{od}/\partial h < 0$) since high-skilled migrants tend to have more adaptive and transferrable linguistic, technical and cultural skills. A second set of costs involve policy induced costs faced by the migrant to overcome legal hurdles created by the destination country d. These would include migration related fees, legal barriers to citizenship and other civil rights, which we refer to as visa costs with slight abuse of terminology. Again, it is generally the case that these visa costs are lower for high-skilled migrants, especially in the presence of selective migration programs that specifically target highly educated workers and give them special preferences and priorities.

Let N_o denote the size of the native population that is within migration age in country o. When the random term ε_i follows an iid extreme-value distribution

(see McFadden, 1974), the probability that a type-h individual born in country o will move to country d is given by:

$$\Pr\left[u_{od}(h) = \max_k u_{ok}(h)\right] = \frac{\exp\left[w_d h + A_d - C_{od}(h)\right]}{\sum_k \exp\left[w_k h + A_k - C_{ok}(h)\right]}$$

and the log-ratio of emigrants to country d to non-migrants is determined by

$$\ln\left[\frac{N_{od}(h)}{N_{oo}(h)}\right] = (w_d - w_o)\, h + (A_d - A_o) - C_{od}(h) \tag{1}$$

This model with a single skill dimension h is rich enough to depict some key patterns of international migration. Among these, the most important are the following:

- The size of bilateral migration flows N_{od} increases with the wage differential $(w_d - w_o)$, differences in country fixed effects or amenities $(A_d - A_o)$, and decreases with the level of overall migration costs (C_{od}).
- Migration rates are lower for low-skilled workers than for the high-skilled since the latter benefit more from wage differentials and have lower migration costs (positive selection).
- The proportion of high-skilled migrants is larger in countries with higher skill prices (positive sorting).

The key insight from the model is that it helps us to understand how existing diasporas affect the magnitude and structure of migration flows. In what follows, we denote the size of the diaspora from country o in country d by M_{od}. We show how diaspora externalities can be introduced into the equation (1) and how they can be empirically estimated using bilateral data on migration stocks and flows.

2.1 Impact on size

As mentioned above, existing social networks or diasporas reduce migration costs through two main channels. First, they lower information, assimilation and adaptation costs. For example, members of a diaspora can help new migrants find jobs, adjust to different social norms and navigate linguistic barriers. Second, family members who have migrated earlier and obtained certain legal rights in the destination country can lower the visa costs. This channel mainly operates by allowing migrants to benefit from family reunification programs for their legal entry into the destination country. For these reasons, the diaspora size are included in the determinants of migration costs and C_{od} becomes a function of M_{od} with $\partial C_{od}/\partial M_{od} < 0$.

The size-externality of diasporas can be tested by regressing skill-specific bilateral flows, $N_{od}(h)$, on the stock of existing migrants at the beginning of the period, M_{od}. Assuming a logarithmic functional form for the diaspora effect, equation (1) can now be rewritten as

$$\ln\left[N_{od}(h)\right] = \alpha_o^h + \alpha_d^h + \beta^h \ln\left[M_{od}\right] + \delta^h D_{od} + \eta_{od}^h \tag{2}$$

where $\alpha_o^h \equiv \ln[N_{oo}(h)] - w_0 h - A_0$ captures origin-country fixed effects and $\alpha_d^h \equiv w_d h + A_d$ denotes destination country fixed effects. D_{od} is a vector of other observable bilateral variables affecting migration decisions (such as distance, linguistic overlap, and historical/political connections) and η_{od}^h is the error term. A positive value for β^h is expected if existing networks reduce migration costs and, thus, increase migration flows.

2.2 Impact on skill selection

If the effect of existing networks varies by skill group, then diaspora size will also influence selection effects in terms of human capital levels. Indeed, as assimilation and information costs are sources of positive selection (because they decrease with human capital), any factor that lowers migration costs will favor low-skilled migrants. Second, when the diaspora size is bigger, the probability that a migrant relies on an economic migration program declines and the probability she/he will benefit from family reunion programs increases. In short, the advantages of being skilled are likely to be less important when a destination country already hosts a large diaspora from a given origin country. As a result, ceteris paribus, countries with larger diasporas will tend to attract a larger proportion of less skilled migrants.

Effect of diasporas on educational/skill composition of migrant flows can be indirectly evaluated by differentiating the β^h obtained from the skill-specific regressions in (2). A more direct way to capture this externality is to regress the log-ratio of high-skilled to low-skilled migrants on the overall diaspora size. Denoting by $\overline{h}$ and $\underline{h}$ the human capital levels of high-skilled and low-skilled individuals, the dependent variable can be written as $\ln[S_{od}] \equiv \ln\left[N_{od}(\overline{h})/N_{od}(\underline{h})\right]$. By subtracting equation (1) for low skill levels from the parallel equation for high skill levels, we obtain

$$\ln[S_{od}] = \alpha_o + \alpha_d + \beta \ln[M_{od}] + \delta D_{od} + \eta_{od} \tag{3}$$

where $\alpha_o \equiv \ln\left[N_{oo}(\overline{h})/N_{oo}(\underline{h})\right] - w_0(\overline{h} - \underline{h})$ is a origin-country fixed effect, $\alpha_d \equiv w_d(\overline{h} - \underline{h})$ is a destination country fixed effect, and η_{od} is the error term. We should note that these fixed effect parameters and error terms are not the same as the ones in (2); we are just using parallel notation A negative value for β is expected if existing networks reduce positive selection and larger diasporas lead to larger proportion of low skilled migrants.

2.3 Impact on relative concentration

The next question is about the impact of diasporas on the relative concentration of migration flows across education levels. Unskilled migrants from a given country will be more concentrated (relative to skilled migrants) if they go to fewer number of countries in larger numbers. The impact of diasporas on the concentration levels should be in line with the effect in terms of selection. In particular, if diasporas tends to reinforce a negative selection process, it should

increase the concentration of low-skilled migrants compared to the concentration of high-skilled migrants.

Our relative concentration measure is defined as the following:

$$RC_{od} = \left[N_{od}(\overline{h}) / \sum_{k} N_{ok}(\overline{h}) \right]^2 - \left[N_{od}(\underline{h}) / \sum_{k} N_{ok}(\underline{h}) \right]^2 \quad (4)$$

A nice property of this bilateral measure is that its sum across destination countries boils down to the difference between Herfindhal indices for high-skilled and low-skilled migrants. The equation to be estimated for relative concentration can be written as the following:

$$RC_{od} = \alpha_o + \alpha_d + \beta \ln [M_{od}] + \delta D_{od} + \eta_{od} \quad (5)$$

Similar to previous two equations, we have α_o and α_d as the origin and destination fixed effects, respectively, D_{od} is a vector of explanatory bilateral variables and η_{od} is the error term. A negative value for β is expected if existing networks increase the concentration of low-skilled migrants across destinations compared to the concentration of the high-skilled migrants.

3 Key empirical issues

Our main empirical goal is to quantify the diaspora effects on the size, skill composition and concentration of migration flows as well as evaluate the robustness of the elasticity based on econometric techniques used. We use the Docquier, Lowell and Marfouk (2009, referred to as DLM from now on) database. Based on census and register information on the size and structure of immigration in all OECD countries, DLM database provides the stock of migrants from any given country to any one of the 30 OECD countries by education level for 1990 and 2000. The dataset covers only the adult population aged 25 and over, and migration is defined on the basis of the country of birth rather than citizenship[1]. We should note that the DLM database does not fully capture undocumented migration for which systematic statistics by education level and country of origin are not available in most destination countries. US census is believed to count most undocumented migrants, however this is not the case in many other OECD countries. By disregarding undocumented migrants (which are disproportionately unskilled), the database probably underestimates bilateral migration stocks/flows and overestimates the average level of education of the immigrant populations in many destination OECD countries.

[1]Even though this is the standard definition of a migrant, especially in the economics literature, the dataset does not include second generation children who are born in the destination country even though they might constitute an important part of a diaspora in the sociological sense. This is simply due to absence of comprehensive administrative data in tracking of the migrants' children. However, we expect diaspora sizes inclusive and exclusive of second generation to be highly correlated.

The main strength of the DLM database is that it distinguishes between three levels of education for migrants. High-skilled migrants are those with post-secondary/tertiary education. Medium-skilled migrants are those with upper-secondary education completed. Low-skilled migrants are those with less than upper-secondary education, including those with lower-secondary and primary education or those who did not go to school. The main characteristics of the diaspora that we consider in this paper are the following:

- The bilateral migration flow for each skill group from origin country o to destination OECD country d is proxied by the change between 1990 and 2000 in the stock of migrants from o to d.
- The bilateral indicator of positive selection is proxied by the log-ratio of the number of high-skilled to low-skilled new migrants from o to d (we disregard medium-skilled migrants for this specification with no impact on the results)
- The bilateral indicator of relative concentration is the 'high-skill minus low-skill' difference in the squared proportions of migrants from a given origin country o to the 30 possible destinations, following equation (4).
- The size of the existing diaspora is measured as the immigrant population born in country o and living in the OECD country d ($\neq o$) in 1990.

3.1 Individual vs aggregate data

The use of aggregate macro data has many advantages but also introduces certain complications. In this section, we discuss the pros and cons of using this dataset, and the main econometric issues triggered by our approach. An important distinction in the empirical analysis of migration data concerns the use of a individual (micro) data as opposed to aggregate (macro) one. Micro data are collected at the household and/or individual level to study the impact of networks on the propensity to migrate and the educational composition of the migrants (Massey, 1986; Munshi, 2003, McKenzie and Rapoport, 2010). These datasets include different measures of individual economic, demographic and social characteristics, such as age, income, occupation and education. This contrasts with more approaches the employ aggregate international migration data. Both approaches have their advantages and drawbacks and should be seen as complementary strategies to address the key issues at stake here.

By focusing on individuals or households, micro data directly accounts for the role of individual characteristics of the migrants. For example, McKenzie and Rapoport (2010) confirm that the networks are more important for uneducated migrants than for educated ones in the case of Mexican migration to the United States. Another appealing feature of the micro datasets is that they can distinguish between different types of networks such as networks defined at the community (city or region) level or at the household (family) level. If such data were available, then we can identify what kind of assistance (such as financial

support or cultural assimilation) is provided by different network along the lines suggested by Massey (1986).

A final appealing feature is the possibility of finding suitable instruments for the network at destination. Since both the current migration flows and diasporas (i.e. past migration flows) are influenced by same factors, endogeneity and other statistical problems (see section on the reflection problem) arise in the estimation. Appropriate instruments should be strong predictors of the network but be uncorrelated with the size of the flows or their composition (i.e. the dependent variables). Munshi (2003) provides a good example for networks defined at the community level, again for the Mexico-US corridor. Rainfall in origin communities in Mexico are supposed to predict the rate of emigration of those migrants but are uncorrelated with labor market outcomes at destination (in the US) that are potentially affected by Mexican networks already present.

A major drawback of micro datasets is that they can consider only a limited number of corridors at a time. That is why a large number of the studies focus on the Mexican-US migration patterns since that is where the best datasets exist. Most other prominent corridors lack such detailed and high-quality data. Furthermore, destination selection effects are quite limited since the US is the destination for more than 99% of the Mexican migrants. This is unfortunately not the case for a majority of origin countries which send their migrants to a diversified set of destination countries. Even in origin countries where migrants have a limited number of choices, the patterns are likely to vary across destination countries. A good example is provided by Cape Verde, which sends a majority of its unskilled migrants to Portugal while sending the most skilled migrants to the US.

Pooling a large set of origin and destination countries in a macro dataset makes it possible to statistically assess the determinants of the various patterns in international migration which might not be easily captured in an analysis of a single corridor. Furthermore, a large number of cells in a migration matrix are filled with zeros (see the section below for a more detailed discussion of this critical issue.) The informational content of the empty cells (such as for the corridor between Cape Verde and Turkey) of migration flows or diasporas between country pairs is valuable. The presence of zeros reflects that the net gains of migration in those corridors are too low for potential migrants and/or . certain factors lead to high levels of migration costs. In other terms, while it creates additional statistical complications that need to be addressed, the inclusion of zero values in macro datasets tends to highlight and identify selection biases.

3.2 The widespread presence of zero observations

The Docquier, Lowell and Marfouk (DLM, 2009) dataset includes almost all origin countries in the world and 30 destination OECD countries. Many statistical properties of the migration flows and stocks can be easily constructed using DLM. The distribution of the migration flows turns out to be unimodal, highly left skewed with a large amount of zero values for both the (net) migration flows between 1990 and 2000 and the stocks in 1990 and 2000. For example, for the

flows, DLM includes 34% of pairs of countries with zero values.

What do these large numbers of zero values truly reflect? For a group of country pairs, the zero values might be the result of a statistical truncation process. For instance, for reasons of statistical confidentiality, national statistical agencies might prefer not to report some low number of migrants of country o in country d. This is reported to be the case for provincial data of international migrants in Canada (see Wagner et al., 2003). Under 5 recorded migrants, the statistical offices are expected to report a zero to preserve the anonymity of the migrants. Similarly, due to imperfect sampling, many smaller and positive migrant stock and flows might not be fully captured in censuses or labor force surveys. Also, it is possible that a number diplomats are not counted in the official stock of migrants following international conventions.

In majority of the cases, a large number of zero values in the migration datasets reflect true zeroes. Like in international trade, many bilateral migration corridors are not 'profitable' so that there are simply no migrants to observe and record. Ignoring such zero values would be highly detrimental to assess the relevance of the determinants of international migration patterns. Zero values imply that the costs of migration is too high for any potential migrant to move from country o to country d. Among those factors, The absence of a network at destination might be a leading factor that deters potential migrants from choosing that particular destination.[2] Therefore, it is important in the empirical investigation of the network effect to employ methods that properly account for those zero migration flows. For example, for the size estimation, possible methods include Poisson regressions, 2-step Heckman approach and Tobit. For the selection and relative concentration, however, Tobit and Poisson regression methods are not possible.

3.3 Stocks vs Flows

A critical choice in the investigations of the network/diaspora effect is the appropriate dependent variables. For instance, in their investigation of the determinants of international migration and in particular, the role of the bilateral wage differential, Grogger and Hanson (2010) use stock data (observed in 2000) which allows to focus on the long run effects. Not surprisingly, variables such as colonial links turn out to be strong predictors of stocks in the long run. Colonial links exert two separate effects. First, they allow people to move during colonial times and shortly after independence through special legal rights and arrangements. Part of those migrants are still included in the contemporaneous stock, depending on when the independence was acquired and how long the legal links were sustained. A second more indirect effect is that colonial links create a dependence path for future migrants through the assimilation and family reunification effects. The relationship between migration flows over a specific period and the size of the stocks at the starting point of that period allows to

[2]Santos Silva and Tenreyro, 2006 show that ignoring the zero trade country pairs lead to overestimation of other bilateral factors such as distance.

capture the (short or medium-run) network effects. Interestingly, when colonial links are included in such a stock-flow model, they turn out to be insignificant since their effect is absorbed by the existing network. From an economic point of view, the implications of those results is that recent migrants tend to come because they can rely on a network at destination, not because of past colonial links that offer current advantages.

The measurement of migration flows in destination countries is also a tricky issue. In most countries, we can rely on census data to provide the stock of migrants in a given year. For most national census rounds, a ten-year frequency is the rule. Therefore, the only way to measure migration flows over a ten-year period is to take the difference between stocks in two successive census rounds. This in turn raises several additional complications. First, the net migration flows are affected by the mortality rate of migrants present in the initial census. Second, there can be significant return migration which varies across origin and destination countries. For instance, using US data, Rosenzweig (2008) shows that the level of skill premia in the origin country is an important factor for the return migration rate of skilled migrants and students. Third issue arises due to regularization (legalization) programs implemented for undocumented migrants who are not recorded in many censuses (such as in most European countries). If regularization programs are implemented between the two censuses, the stocks of migrants in the second census and hence the size of the migration flows will increase in the data without actual movement of people. Another theoretical argument developed by Brücker (2006) suggests that using net migration flows instead of stocks might be misleading in the case of heterogeneous agents. This is especially important when it comes to estimating the impact of wage differentials on migration. In models like ours, we do not explicitly include the wage differentials for several reasons. One reason is the absence of reliable wage indices by skill level in most origin countries. A second reason is that wages are captured by country fixed effects in most of our estimations. To sum up, there are obviously negative and positive biases in measuring migrations flows through the changes in migrant stocks from census data. Whether this tends to underestimate or overestimate the true values is obviously difficult to know in advance.

3.4 Defining a network

The investigation of the network effect relies on a specific definition of diaspora which is the stock of nationals from country o living in destination d at a given time. This is a natural definition of the people who are supposed to provide assistance and help to the new arrivals. On the one hand, restricting the diaspora to people with the same nationality might be restrictive. Ethnic networks are also known to be efficient and do not necessarily correspond to national borders. Migrants speaking the same language can be also very useful for the assimilation of new migrants. On the other hand, defining the network at the national level might overestimate the number of people able to provide help. Obviously, people located in large countries such as the US can provide assistance mainly

to a restricted number of new migrants within certain geographic proximity. This is especially true if concentration of migrants in the destination country is not very high.

3.5 The reflection problem

As explained by Manski (1993), one issue in identifying and estimating endogeneous social effects like the network effect is the presence of unobservable correlated effects. In our framework, it could be the case that unobservable bilateral components will affect the size of the diaspora M_{ij} and the dependent variables. For instance, unobserved cultural proximity between country i and country j might affect simultaneously the stock of migrants, the current flows of new migrants and their selection. The cross-sectional nature of the data prevents us to estimate directly those unobservable components. Therefore, those effects will be included in the error term, which in turn leads to some kind of omitted variable bias and to some correlation between M_{ij} and the error term.

4 Econometric Methods

There are several alternative methods that can be used to estimate the impact of diasporas on migration flows, on their skill composition and on their relative concentration by education. A simple and easy way of estimation the models is OLS, but, high occurrence of zero observations is likely to lead to inconsistent estimates. The use of a log specification drops the zero observations from the sample which is likely to result in biased estimates of the impact of diasporas and other variables on the migration flows and their selection. For instance, it might be the case that there are no migrants from country i to country j because migration costs are too high. In turn, migration costs might be too high because distance is too high and there is no diaspora. In this case, the exclusion of those observations leads to underestimation of the impact of the variables affecting the migration costs such as distance, colonial links, linguistic similarities or diasporas.

One option is to use Heckman 2-step estimation methods to minimize the bias due to selection issues. In general, for all the features that we analyze (migration flows, skill ratios and relative concentration), the first step involves the estimation of a selection equation - the probability for a given country pair to have a positive migration flow[3]. The usual procedure implies the use of an instrument in the probit equation, i.e. a bilateral variable that influences the probability of observing a diaspora between the two countries but does not influence the size of this diaspora. It is obviously extremely difficult to find an

[3]To be more precise, for the analysis of migration stock, the probability that a given observation will be included in the regression is directly related to the probability of observing a diaspora (either regardless of the skill level, either for a particular skill level) for this country pair. For the migration flows, the probability is exactly the same since we have no case of zero migration flow with positive values of the stock in 1990 and 2000. For the analysis of selection, the probability is related to the existence of a diaspora or at least a skilled diaspora.

instrument that influences, in a sense, the arrival of the first migrant (i.e. the presence of a diaspora) but not the other migrants (i.e. the size of a diaspora). One possible candidate is diplomatic representation of the destination country in the origin country. Diplomatic representation might affect the probability of having at least one migrant by setting some kind of threshold on the visa costs faced by the initial migrant. In the absence of any diplomatic representation of country j in country i, the cost to get a visa can simply be too high so that nobody would consider to migrate to country j. The role of diplomatic representation in the migration process is to a certain extent analogous to the role played by a common religion for trade relationships. As argued by Helpman et al.(2007), a common religion (a proxy of costs of establishing business linkages) affects the extensive margin of trade (i.e. the probability of export) but not the intensive margin (i.e. trade volumes). In regressions (2-3), the use of a two-step Heckman approach yields intuitive results both for the flow and for the selection equation. In particular, for the selection equation, we find that diplomatic representation of country j in county i tends to positively affect the probability of observing a diaspora of country i in country j. Furthermore, the mills ratio turns out to be significant in the flow equation, suggesting that accounting for a selection bias is important[4].

An alternative is to use Poisson regression models that rely on pseudo maximum likelihood estimates, as advocated by Santos Silva and Tenreyro (2006) who show that the use of log linearization for gravity models leads to inconsistent estimates of the coefficients (such as the one relative to distance). One main cause of this problem, as mentioned before, is the exclusion of zero observations for the dependent variable. A second reason is that the expected value of the error will depend on the covariates of the model and hence will lead to estimation biases of the coefficient. The Poisson solution is nevertheless unfeasible for the selection and the concentration analyses. For the selection, the existence of zero values for $M_{i,j}(h)$ leads to undefined values for S_{ij}, which cannot be handled by the Poisson approach.

The above mentioned estimation methods do not address one solution proposed by Munshi (2003) which is to estimate the effects of M_{ij} by IV. For that purpose, one has to find instruments of M_{ij}, i.e. variables uncorrelated with the flows but that are good predictors of the stocks. Beine et al. (2010) use two instruments. The first is a dummy variable capturing whether the two countries were subject to a temporary guest worker agreement in the 60's and 70's. One can expect those guest worker agreements to exert a strong impact on the initial formation of a stock of migrants in the 60's and the 70's, hence influencing the stock in 1990. In contrast, it is unclear why those initial agreements (that

[4]Since the observed level of diaspora in 1990 is used as a regressor, the use of diplomatic representation leads to some colinearity problems in the selection equation. In order to mitigate the collinearity problems, it is possible to run Heckman two-step regressions without any additional instrument. As stressed by Wooldridge (2002), the use of an additional instrument in the probit equation is not strictly necessary. The drawback of not using an additional instrument is that the Mills ratio might become highly collinear with the explanatory variables of the flow equation, which in turn lowers the significance of the coefficients. This is not the case for most of our regressions.

are no longer valid) would influence the contemporaneous migration flows beyond the impact exerted by the diaspora itself. For instance, it turns out that guest worker agreements did not create any preferential treatment at the level of country pairs in the migration policy. Therefore, it is expected that these guest worker agreements are not themselves correlated with the bilateral unobservable components. Examples of such a process are illustrated for instance by the impact of the post-war guest worker agreements between Belgium and Italy or Spain.

The second instrument proposed by Beine et al. (2010) is a variable capturing the unobserved diaspora in the 1960's through a combination of variables representing some push factor in country i , size in country i, openness and size in country j and distance between i and j. We use four different measures. The basic measure is $ln(pop_i * immst_j/dist_{ij}) * armedconflict_i$ where pop_i is the population size in the 60's of country i, $immst_j$ is the immigrant stock of country j in the 60's, $dist_{ij}$ is the distance between i and j and $armedconflict_i$ is a dummy variable capturing the occurrence of armed conflicts in country i during the 60's. To capture push-factors leading to emigration in the 1950s and 1960s, we only consider conflicts observed between 1946 and 1960. We distinguish minor conflicts (number of battle-related deaths between 25 and 999) denoted CONFL1 and wars (at least 1,000 battle-related deaths in a given year) denoted CONFL2. The variables CONFL1 and CONFL2 sum up the number of annual conflicts over the period 1946-1960. IV estimation methods are suited to address the issues (size, selection and concentration) listed above. Nevertheless, like OLS, they are subject to issues related to the selection bias. A combination of Poisson regression models along with IV estimation is proposed by Tenreyro (2009) within the GMM framework. This is relevant only for the size issues but it is nevertheless beyond the scope of this paper.

5 Results

After listing all the potential problems with the data, estimation methods and identification issues, we finally turn to the estimation of the three main equations listed above (2, 3 and 5) which correspond to analysis of the impact of diasporas on the size, skill composition and concentration of migration flows, respectively.

Table 1 reports the results for five different estimation techniques for the estimation of the impact of diaspora on migration flows. The techniques used are OLS (using lows as the dependent variable), Heckman two stage method with and without an instrument for the selection, Maximum likelihood Poisson and IV regression (on the flows as well) using the two above mentioned instruments. The results illustrate the strong robustness of the estimation of the key elasticity parameter which ranges between 0.62 and 0.76. This means that a 1% increase in the size of the migrant network present in the destination country in 1990 tends to increase the subsequent migrant flow from a given origin country over the next ten years by around 0.7%. This result is in line with some of the previous results in the literature using the US data. For instance, focusing only

on family reunification programs, Jasso and Rosenzweig (1986, 1988) show that the multiplier associated with sponsored migration is about 1.2. If this were true for other countries, our results suggest that the multiplier associated with the pure network effect (assimilation effect) should be around 1.5 for the US. the coefficients of the other explanatory variables are also worth noting. The common language and (log) distance variables are significant in all cases with the expected sizes. A 1% increase in distance between a pair of countries reduces migration flows by around 0.3-0.5%. Similarly, if two countries share a common language, they experience between 30-60% higher migration flows. Unlike it is the case with linguistic overlap and distance, the effect of colonial links is not robust to the estimation method used, as we had mentioned earlier. Once we control for the diaspora size, the contemporaneous effect on the flow weakens considerably.

Table 2 looks at the selection issue, using the log of the skill ratio (log of the number of skilled migrants over unskilled migrants from o to d) as the dependent variable. Four different estimation results are reported. The first one uses OLS applied to the log of the ratio (observed in 2000). The second column reports the same estimate but with Heckman two stage method (without instrument). The third column does the same but on the change in the (log of) the skill ratio between 2000 and 1990. Finally, the last column also looks at the variation but using instrumental variable. The results shows that the networks exert important effect in terms of negative selection. In the first two cases, we see that diasporas significantly reduce the overall skill level of migrant stocks. More specifically, a 1% increase in the diaspora size reduces the skill ratio by around 0.2%. Linguistic overlap, distance and Schengen agreement, on the other hand, increase the skill composition while colonial links has no statistically significant effect. In the last two columns, the results show that the diaspora size also negatively influences the change in the skill ratio. In other words, if there is larger diaspora from country o in country d, the migrant flows become more unskilled more rapidly. The coefficients of the other explanatory variables also have the expected and significant signs.

Table 3 investigates the same analysis but on the relative concentration between skilled and unskilled migrants as explained earlier. We use three different estimation methods. In the first two (OLS, Heckman two stage method), the dependent variable is the level of relative concentration of skilled migrants as given in (5); the third estimation uses Heckman two stage method with the change in the relative concentration measure as the dependent variable. The results from the first two columns show that a 1% increase in the diaspora size tends to decreases the relative concentration of skilled migrants with respect to the unskilled ones by around 0.5%. Furthermore, larger diaspora size also negatively influences the change in the relative concentration of skilled migrants. These results are in line with and confirm the results above concerning the skill selection of the migrants.

6 Conclusion

This paper reviews the existing literature on the impact of migrant networks (diasporas) on the international migration patterns. In addition to size of the migration flows, we include the skill composition and concentration among these patterns we analyze and show that diasporas strongly influence all three. We first present a simple theoretical model that identifies the channels through which diasporas would influence migration patterns. These channels mainly operate through lowering of bilateral migration barriers via assimilation effects and family reunification programs. It identifies the key issues, namely the impact on size, selection and concentration of the migration flows. The paper also reviews the data and econometric hurdles that the researchers face in assessing the importance of networks. Among the key issues are the choice of individual micro vs aggregate macro approach, the definition of a network, the access to suitable data, and the adoption of econometric methods accounting for the main features of those data, such as wide prevalence of zeros.

The main results are illustrated with estimation results obtained using the Docquier-Lowell-Marfouk (DLM, 2009) data. Larger networks are shown to exert strong positive influence on the size of the international flows and lead to lower skill composition for a given corridor. We also show that diasporas also favour the concentration of the unskilled migrants with respect to the skilled ones. Destination and origin country specific factors are captured via fixed effects. All other bilateral variables, such as linguistic overlap, distance, colonial linkages have the expected signs and economically significant effects.

Naturally, there are many questions remain unanswered. One venue to consider is whether if these results hold for non-OECD destination countries and south-south migration. They require higher quality data that covers larger number of destination countries outside the OECD. Another key issue is separating the assimilation effect of diasporas from the visa effect which operates mainly through the family reunification programs. Such questions will require different types of dataset that combine aggregate data with household level data. In closing, diasporas are among the key determinants of migration patterns and we have only scratched the surface in identifying their effects.

Table 1. Determinants of migration flows

	(1) OLS	(2) Heck with	(3) Heck w/o	(4) Poisson	(5) IV
Lagged diasp	0.620	0.660	0.699	0.703	0.761
	(34.35)***	(47.97)***	(43.91)***	(16.20)***	(10.92)***
Col links	0.331	0.219	0.127	-0.312	-0.051
	(2.45)**	(2.03)**	(1.10)	(1.65)*	(0.26)
language	0.388	0.477	0.496	0.298	0.234
	(5.20)***	(6.71)***	(6.48)***	(2.53)**	(2.27)**
Log(dist)	-0.408	-0.501	-0.448	-0.337	-0.259
	(9.04)***	(12.04)**	(10.69)***	(3.28)***	(2.84)***
Schengen	0.168	0.257	0.277	0.061	0.160
	(1.19)	(2.00)	(2.02)**	(0.23)	(1.11)
Constant	3.750	2.785	2.365	3.461	2.365
	(6.92)***	(4.82)***	(4.02)***	(3.06)***	(2.69)
Observations	3608	5610	5760	5374	3486
Mills ratio	-	0.908	1.19	-	-
		(7.60)***	(9.35)***	-	-

Table 2. Impact of diaspora on selection (level and change in log high-skill/low-skill ratio)

	Log-skill ratio (OLS)	Log-skill ratio (Heck)	ΔLSR (Heck)	ΔLSR (IV)
Lagged diasp	-0.171	-0.194	-0.212	-0.215
	(16.19)***	(20.62)***	(17.62)***	(2.95)***
Col links	-0.042	-0.022	0.101	0.270
	(0.62)	(0.32)	(1.67)*	(1.77)*
language	0.466	0.460	0.176	0.235
	(9.38)***	(9.37)***	(4.17)***	(3.19)***
Log(dist)	0.096	0.090	0.086	0.019
	(3.35)***	(3.40)***	(3.78)***	(0.30)
Schengen	0.502	0.519	0.390	0.414
	(5.65)***	(6.26)***	(5.48)***	(6.08)***
Constant	-1.109	-0.734	-1.250	-0.481
	(1.16)	(1.32)	(2.54)**	(0.63)
Mills	-	(-0.380)	(-0.10)	-
	-	(6.86)***	(0.22)	-
F-stat First stage	-	-	-	30.07
Hansen J-test (p-value)	-	-	-	0.747
Observations	3486	5760	5760	3486

Absolute values of robust t statistics in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

Instrument sets for M_{ij} in all columns include a dummy for bilateral guest-worker agreements and a proxy for diaspora size in 1960. In column (1) , the proxy is computed as $ln(pop_i * immst_j/dist_{ij}) * Conf1_i$.In column (2) , the proxy is computed as $ln(pop_i * immst_j/dist_{ij}) * Conf2_i$;in column (3), the proxy is computed as $ln(pop_i * immst_j/dist_{ij}) * (conf1_i + Conf2_i)$.

Table 3. Explaining relative concentration between high-skill and low-skill and change in relative concentration

	Rel conc (OLS)	Rel conc (Heck)	ΔRC (Heck)
Lagged diasp	-0.502	-0.514	-0.008
	(5.87)***	(9.67)***	(16.05)***
Col. links	-4.635	-4.619	-0.040
	(4.68)***	(10.69)***	(9.93)***
Language	0.338	0.321	-0.004
	(0.84)	(1.09)	(1.58)
Log(dist)	0.266	0.269	0.006
	(1.24)	(1.69)*	(3.78)***
Schengen	-0.193	-0.180	0.002
	(0.50)	(0.36)	(0.49)
Constant	5.607	-3.240	-0.037
	(0.29)	(1.19)	(1.60)
Mills		-0.405	-0.873
		(1.07)	(2.44)**
Observations	3920	5730	5730

Robust t statistics in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

7 References

Bertolini, S. (2009), "Networks, Sorting and Self-selection of Ecuadorian Migrants", Paper presented at the second TOM Meeting, Louvain-La-Neuve, January.

Borjas, G (1987), "Self-selection and the earnings of migrants", American Economic Review, 77 (4), 531-53.

Borjas, G.J. (1994), "The economics of immigration", Journal of Economic Literature, 32, 1667-1717.

Borjas, G.J. (1995), "The economic benefits from immigration", Journal of Economic Perspectives, 9 (2), 3-22.

Borjas, G.J. (1999), Heaven's door: immigration policy and the American economy, Princeton University Press.

Carrington, W.J., E. Detragiache and T. Vishwanath (1996), "Migration with endogenous moving costs", American Economic Review, 86 (4), 909-30.

Chiquiar, D. and G.H. Hanson (2005), "International migration, self-selection, and the distribution of wages: evidence from Mexico and the United States", Journal of Political Economy, 113 (2), 239-81.

Clair, G., G. Gaullier, Th. Mayer and S. Zignago (2004), "A note on CEPII's distances measures", Explanatory note, CEPII, Paris.

Cohen, A. and A. Razin (2008), "Skill composition of migration and the generosity of the welfare state: free vs. policy-restricted migration", Mimeo., Tel-Aviv University.

Docquier, F. and E. Lodigiani (2009), "International migration and business networks", Open Economies Review, forthcoming.

Docquier, F., O. Lohest and A. Marfouk (2007), "Brain drain in developing countries", World Bank Economic Review, 21, 193-218.

Docquier, F. and A. Marfouk (2006), "International migration by educational attainment (1990-2000)", in C. Ozden and M. Schiff (eds). International Migration, Remittances and Development, Palgrave Macmillan: New York (2006), chapter 5.

Docquier, F., B.L. Lowell and A. Marfouk (2007), "A gendered assessment of highly skilled emigration", Population and Development Review, 35 (2), 297-321.

Friedberg, R.M. and J. Hunt (1995), "The impact of immigrants on the host country wages, employment and growth", Journal of Economic Perspectives, 9, 23-44.

Gao, T.(2003), "Ethnic Chinese Networks and International Investment: Evidence from Inward FDI in China", Journal of Asian Economics, 14, 611-629.

Gleditsch, P., M. Eriksson and M. Sollenberg (2002), "Armed Conflict 1946-2001: A New Dataset", Journal of Peace Research, 39 (5), 615-637.

Grogger, J and G.H. Hanson, 2008, "Income Maximisation and the selection and sorting of international Migrants, NBER Working Paper, No. 13821.

Harbom, L., E. Melander and P. Wallensteen (2007), "Dyadic Dimensions of Armed Conflict, 1946—2007", Journal of Peace Research, 45 (5), 697-710.

Helpman, E., M. Melitz and Y. Rubinstein (2007), "Estimating Trade Flows: Trading Partners and Trading Volumes", NBER Working Paper W12927.

Manski, C.F. (1993), "Identification of Endogeneous Social Effects: the Relection Problem", Review of Economic Studies, 60 (3), 531-42.

Massey, D.S., J. Arango, G. Hugo, A. Kouaouci, A. Pellegrino and J. E. Taylor (1993), "Theories of international migration: Review and Appraisal," Population and Development Review, 19 (3), 431-466.

McFadden, D. (1984), "Econometric analysis of qualitative response models", in: Z. Griliches and M. Intriligator, eds., Handbook of Econometrics, Volume 2, Amsterdam. Elsevier/North-Holland.

McKenzie, D. and H. Rapoport (2007), "Self-selection patterns in Mexico-US migration: the role of migration networks", Review of Economics and Statistics, forthcoming.

Munshi, K. (2003), "Networks in the modern economy: Mexican migrants in the US labor market", Quarterly Journal of Economics, 118 (2), 549-99.

Pedersen, P.J., M. Pytlikova and N. Smith (2008),. "Selection and network effects–Migration flows into OECD countries 1990-2000", European Economic Review, 52 (7), 1160-1186.

Rapoport, H. and M. Kugler (2006), "Skilled Emigration, Business Networks and Foreign Direct Investment", CESifo Working Paper Series No. 1455.

Rauch, J. (2003), "Diasporas and development: Theory, Evidence and Programmatic Implications", Department of Economics, University of California at San Diego.

Rauch, J. and A. Casella (1998), "Anonymous Market and Group ties in International Trade", Journal of International Economics, vol 58(1):19-47.

Rauch, J. and V. Trindade (2002), "Ethnic Chinese Networks In International Trade", The Review of Economics and Statistics, MIT Press, vol. 84(1):116-130.

Razin, A. and E. Sadka (2004), " Welfare migration: Is the net fiscal burden a good measure of its economic impact on the welfare of the native-born population?", NBER Working Paper 10682.

Rosenzweig, M (2008), The global Migration of Skill, Paper presented at the Migration and Development Workshop, Lille, June.

Roy, A.D. (1951), "Some thoughts on the distribution of earnings", Oxford Economic Papers, 3 (2), 135-46.

Santos Silva, J.M.C. and S. Tenreyro (2006), "The Log of Gravity", Review of Economics and Statistics, 88 (4): 641-658.

Williamson, J.G. (2006), "Global migration: Two centuries of mass migration offers insights into the future of global movements of people", Finance and Development, 43 (3).

Wagner, D., K. Head and J. Ries (2003), "Immigration and the Trade of Provinces, Scottish Journal of Political Economy, 49 (5), 507-525.

Wooldridge, J.M. (2002), Econometric Analysis of Cross Section and Panel Data, MIT Press.