Return migration: an empirical investigation

Roman Zakharenko

June 12, 2008

Abstract

Many people emigrating abroad eventually return home. Yet, little is known about the returnees: who are they and how do they compare to those who did not return? How does their decision to return depend on economic situation at home? In this paper, I empirically analyze the propensity of US immigrants to return. To identify return migration, I use the method developed by Van Hook et.al. (2006). The method is based the U.S. Current Population Survey (CPS) which interviews households for two consecutive years. About a quarter of foreign-born individuals drop out of the sample between the first and the second years, due to various causes including return migration. After eliminating all other causes of dropout, I estimate the propensity of immigrants to return, depending on personal and home country characteristics. The propensity to return increases with age for skilled immigrants and decreases with age for others, intersecting around the retirement age. Immigrants from OECD countries can be divided into two categories: those participating in the "brain circulation" come for short periods and then return; those who stayed more than five years in the US are extremely unlikely to return. This pattern in not the case for other immigrants.

Contents

List of Figures

List of Tables

1 Intro

Many emigrants eventually return home. Yet, little is known about the returnees. Are they more or less successful than those who stayed abroad? Does the return propensity increase or decrease with age? Are family ties significant for decisions to return? How do the return patterns depend on their home country culture and economic performance?

In this paper, I analyze empirically the factors affecting return migration. I use Current Population Survey (CPS) data collected by the U.S. Census. This database has three features that make it particularly useful for a study of return migration. First, its size: there are hundreds of thousands person observations available each year. Second, its information on nativity of respondents: the survey identifies immigrants from over 90 world countries and territories, which enables a cross-country analysis. Third, the sample design: each address is questioned several times during two consecutive years, which makes observations longitudinal. By observing respondents prematurely leaving the sample, we can estimate the fraction of immigrants leaving the US, as a function of individual and home country characteristics.

A methodology for estimating the return migration using the CPS data was developed in the demographic literature (Van Hook et.al. 2006) and, to my knowledge, has not been used in the field of economics.

When estimating the fraction of immigrants leaving the US, we cannot claim that they necessarily return home: part of them could go to third countries. Given limited data availability, it is not possible to estimate accurately how many foreign-born emigrants choose to return home, and how many migrate to third countries. However, a partial inference can be made using Integrated Public Use Microdata Series (IPUMS) which provides large (up to 10% of population) samples collected in several countries of the world. The IPUMS database has a particularly good coverage of the Latin countries: it has data from Mexico, Costa Rica, Colombia, Brazil, Argentina, and Chile, covering most of the Latin world.[1] Information from

[1]Samples from Ecuador and Venezuela are also available, but they lack data on previous migration experience which is vital for identification of return- and third-country migrants

Table 1: Return migration vs. third-country emigration

Home country	returned home from US	moved to 3rd country from US
Mexico	267,000	386
Costa-Rica	4,820	146
Colombia	22,370	555
Brazil	12,600	209
Argentina	4,310	424
Chile	5,550	204

Note: possible "third countries" are countries listed in first column (excluding home country), Spain, Portugal

another likely destinations of Latin foreign-born leaving the US – Spain and Portugal – is also available. Using this information, we can estimate the number of, say, Mexican-born individuals who migrated to the US and then either returned to Mexico (return migrants) or migrated to all other countries listed above (third-country migrants). The same exercise can be done for all other Latin countries listed above. The results are listed in table 1. Mexicans leaving the US rarely travel to third countries; among other countries, about 97% of those leaving the US return home and 3% go to other destinations (Argentina is a notable exception with 90/10 ratio). Given these results, we may conclude that migration to third countries is a rare phenomenon compared to return migration. Throughout the rest of the paper, I assume that all immigrants leaving the US return home, and the terms "emigration of the foreign-born" and "return migration" are used interchangeably.

2 Related literature

Overall, return migration is a scarcely studied topic due to lack of data. Data is usually collected within one country, and therefore migrants are not tracked as they move across borders.[2] Given this data limitation, the com-

[2]The only known exception is the dataset constructed by German *Institut fur Arbeitsmarks und Berufsforschung* (IAB) which contains information on Turkish migrants

mon approach to approximate return migration is to estimate the number of people which "disappear" from the host country over time. Historically, two methods have been used.

Panel data. With longitudinal/panel data, dropping out of the sample may be attributed to return migration (of course, one has to eliminate other causes of dropping out such as death). The popular sample is German Socio-Economic Panel (GSOEP) which has been used, among others, by Kirdar, Bellemare, Constant and Massey. The strength of GSOEP is that it follows individuals when they migrate within Germany,[3] thus greatly reducing the dropout rate and allowing to identify return migrants more accurately. A shortcoming of GSOEP is that it covers immigrants from relatively few countries and therefore not suitable for a cross-country study.

Repeated cross sections. With two repeated cross-section nationwide databases (such as decennial US Census), one can use the method developed by Warren and Peck (1980). In economic literature, a version of this method has been used by Borjas and Bratsberg (1996). According to this approach, the entire sample is divided into non-overlapping groups (e.g., immigrants by country of birth). The decrease in the number of migrants within a certain group can be attributed to return migration. Indeed, the researcher should exclude all new migrants, arriving between the two dates (and therefore must observe everyone's year of entry). One observation is thus not an individual, but a subsample of individuals (e.g. all immigrants from Kenya). This method is suitable for studying macroeconomic factors affecting return decisions, but not particularly useful for studying demographic characteristics of return migrants. Indeed, we can disaggregate immigrants by exogenous characteristics (gender, age, year of entry into the host country.[4]) But variable characteristics such as education cannot be controlled

returning home from Germany, both before and after their return migration. The study, however, focused only on individuals intending to return, and therefore cannot be used to compare returnees and non-returnees. Since it includes only one home and one host country, it cannot be used for cross-country analysis. Dustmann and Kirchkamp (2002) provide a study based on this dataset

[3] which is not the case in the American CPS

[4] the year of entry is exogenous in the sense that it cannot be changed after the person has immigrated

for, because individuals can make unobservable transitions from one educational group to another. For example, the number of low-skilled immigrants may decrease not only due to emigration or death, but also because some of them have acquired more skill.

One more problem with using repeated Census data is incomplete coverage of the population. In theory, the Census should cover *all* residents of the country. In reality, many groups of people, especially immigrants, are not fully covered. Moreover, the coverage is improving over time, causing a strong downward bias in return migration estimates. For example, the number of people born in country X who entered the US before 1990 must decrease between years 1990 and 2000, due to death and emigration. But due to improved coverage between dates 1990 and 2000, the estimated number of these people may actually increase, resulting in low or even negative return migration estimates.

To summarize, longitudinal data is suitable for a study of demographic characteristics of return migrants, while repeated cross-section data has proved more efficient in a study of macroeconomic factors affecting the decision to return. None of these, however, enabled researchers to study the interaction of demographic and macroeconomic factors affecting return migration. For example, are gender differences more significant for immigrants from less developed countries? Do senior immigrants from rich countries behave differently than those from poor countries? In this paper, I was able, perhaps for the first time, to conduct such tests.

[...]

3 The method

My method of return migration estimation is based on Van Hook et. al. (2006), who use Current Population Survey (CPS) to estimate the number of emigrating foreign-born. This sample is longitudinal: households are questioned for two consecutive years. It is important to note that the sample follows addresses, not individuals: if a family moves out and a new family moves in, the latter will be interviewed, and the former will not be followed.

Table 2: Observed year-2 outcomes, for respondents of age 18-70 (percentage points)

Year-2 observed outcome	natives	second generation	foreign-born
Person followed up	78.91	77.67	73.14
Person absent, same household	5.09	6.02	6.48
Address occupied by other household	5.64	5.87	8.44
Failed to conduct interview	5.47	5.75	5.82
Vacant address	4.89	4.68	6.11
Number of observations	272,294	22,521	47,450

Therefore, people may drop out of sample not only due to emigration, but also due to migration within the US, as well as another unknown factors. As a result, only about 73% of foreign-born respondents are found in the sample one year later; the rest drop out for various reasons, including return migration which is our estimation target. Table 2 summarizes observed year-2 follow-up outcomes, disaggregated by three groups of respondents:

- US-born individuals with US-born parents - labeled as *natives* throughout the paper;
- US-born individuals with at least one immigrant parent - labeled as second generation Americans, or simply *second generation*;
- Immigrants - labeled as *foreign-born*.

Generally, a foreign-born individual may drop out of the sample for the following reasons:

- death
- emigration: either return to home country, or migration to a third country
- moving to another US address
- other non-followup, e.g. refusal to continue participation, or not at home on the day of interview

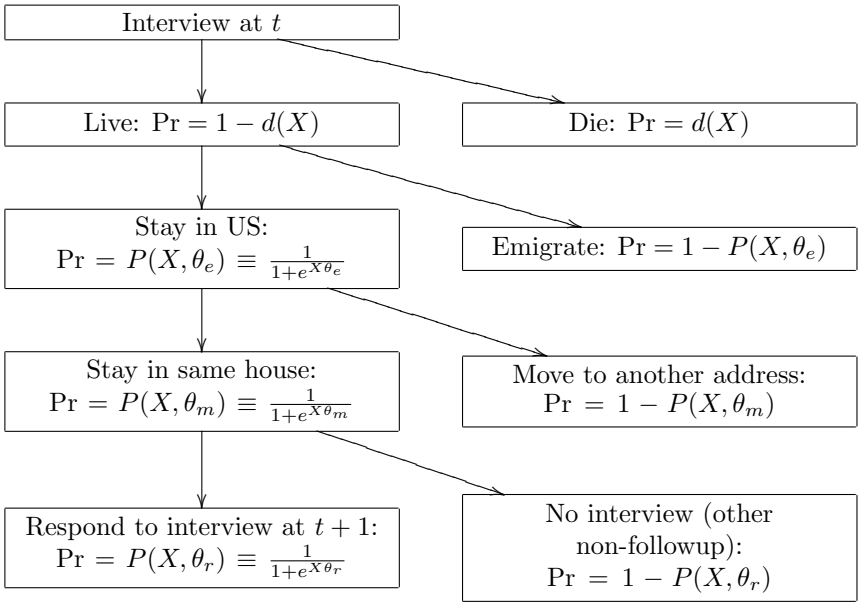

Figure 1: Nested logit model

I assume that the death incidence does not depend on individual migration decisions, the decision to emigrate is independent of mobility within the US, and the decision to move to another address does not depend on the CPS interview process. With these assumptions, we can build an outcome tree as described in figure 1. Probabilities of all outcomes except death are specified as logit probabilities. X is the vector of individual characteristics, θ_e is the propensity to emigrate, θ_m is the propensity to move within the US, and θ_r is the propensity to drop out of sample for other reasons.

Note that only one outcome (interview at $t+1$) is observed in the data, all other information (person absent, no interview, vacant address) cannot be clearly attributed to death, or emigration, or moving to another address, or other non-followup. Therefore, some additional restrictions have to be specified.

The probability of death may be estimated using readily available U.S. life tables. See section 4.5 for details.

Table 3: reported migration experience in the past year, for respondents of age 18-70 (percentage points). Based on year-1 interview

Migration status, 1 yr ago	natives	second generation	foreign born
Lived in same house	85.94	86.38	82.21
Other house in US	13.90	13.33	14.98
Abroad	0.15	0.29	2.81

The propensity to move within the US θ_m may be estimated using the information about recent migration experience of CPS dataset respondents: they report where they lived one year ago. The fraction of recent movers, among those who lived in the US one year ago, as a function of personal characteristics X, may serve as a proxy for θ_m.

Van Hook et.al. (2006) use the *year-2* interview data to estimate θ_m. Using the year-2 interview, however, may possible lead to a bias: non-movers are interviewed for the second time, while recent movers are interviewed for the first time (in the previous year, another residents at the same address were interviewed). It is possible that people are more likely to give an interview in the second year than in the first (because they got used to the survey and expect a follow-up interview), thus the estimates for θ_m may be biased downward if year-2 data is used. Indeed, using the CPS weights may help to remove part of that bias, but using the first (date t) interview asking about mobility experience between dates $t-1$ and t seems to be a better way to go. Table 3 reports the statistics of recent mobility experience.

One more problem arises from the timing of measurements. The logit model specified above estimates probabilities of various events within one year, as a function of person characteristics at the *beginning* of the year. The proposed method of θ_m estimation, however, relies on information collected at the *end* of the period. Clearly, some person characteristics might change during the year: age obviously increases by one, educational attainment might improve, citizenship status and employment status might change. All these changes, except age, are not observed and hence cannot be controlled for, creating a possible bias in the estimation of θ_m.

The employment status is the most likely cause of bias because it changes more frequently than educational attainment or marital status or citizenship status, and because it is unclear whether unemployment causes mobility or vice versa. When estimating θ_m, we observe the employment status *after* moving to a new address; the observed positive correlation between mobility and unemployment implies that unemployment may be a *consequence* of recent mobility. However, when estimating the main model described by figure 1, we assume that unemployment is the *cause* of mobility, whether internal or international. Therefore, applying the estimates of θ_m to the main model may produce spurious results. To prevent the problem, I do not use employment status in the regression.

Further, to estimate θ_m accurately, the data and the model should satisfy the following few properties:

- The choice of households for the interviews does not depend on recent mobility. This property is very likely to hold, because the CPS sample is random by design.

- The probability of rejecting the interview by respondents does not depend on their recent mobility experience. Again, given the effort of the US Census to make the sample representative, we can expect this property to hold. We can also test this property by comparing the weights assigned by the CPS to recent movers and non-movers. The average weight of the former is higher by only 3.5%, which means that movers are only slightly less likely to respond. This property, combined with the previous one, ensures that the estimates of the US population mobility are accurate.

- Independence of irrelevant alternatives: the possibility of emigration to other countries does not affect internal mobility. With this assumption, we only need the data on non-movers and internal movers to estimate the propensity to move θ_m; this propensity will not change when the emigration opportunity is added. Since I use the logit model, this property holds automatically.

The probability of "other non-followp" θ_r is the least straightforward to estimate, due to uncertain nature of this group. I adopt the method proposed by Van Hook et.al. (2006), who use second generation Americans (US-born individuals with foreign-born parents) as a comparison group. The following assumptions allow to identify θ_r:

- Second-generation Americans never emigrate. Docquier and Marfouk (2004) report that only 0.4% of US-born working-age population live in other OECD countries. Indeed, some US-born also live in countries other than OECD, but their number is probably even smaller, hence the total number of emigrants should be far below 1% of US adult population. Even if second-generation Americans have a somewhat higher propensity to live outside of the US,[5] these numbers are still incompatible with the estimate of 30% of first-generation immigrants eventually leaving the US (Warren and Peck 1980, Jasso and Rosenzweig 1988). With this assumption, we have enough data to identify θ_r for the second generation.

- The probability of "other non-followup" is the same for the foreign-born and the second generation. Since second-generation Americans are closer (or, at least, not more distant) to immigrants than any other comparison group, their non-followup propensity is the closest to that of immigrants. Indeed, they might still be unequal, but there is no information to identify the difference.

Overall, there are four sets of unknown parameters: internal migration propensity of foreign-born θ_m^f, internal migration propensity of second-generation Americans θ_m^s, propensity to reject the second interview θ_r, and propensity to emigrate θ_e which is our target. Four sets of conditions identify these parameters: recent internal migration experience of foreign born, similar experience for second generation Americans, foreign-born dropping out of sample, second-generation Americans dropping out of sample. The estimation is done using maximum likelihood method. The description of

[5] from table 3, their likelihood of living abroad is twice as high

the log-likelihood function and the optimization details are provided in appendix A.

The disadvantage of this method is that the return migration estimate is not guaranteed to be positive for *all* subsamples of the data. Suppose that some group of people drop out of sample with probability 10%. It may happen that the estimated probability of dropping out for reasons other than migration (that is, moving internally, death, or other non-followup) is actually higher than 10%, forcing the emigration probability to be negative. In the logit model, negative probability is impossible; in such cases, the algorithm tries to reduce θ_e down to negative infinity (making emigration probability equal to zero). As a result, computational time greatly increases, and standard error estimates become less accurate. To avoid the problem, I set lower bounds on θ_e parameters.

4 Person data

4.1 Current Population Survey

The Current Population Survey is a project administered by the US Census since early 1940s. Its main goal is to collect data on the US labor force characteristics. Currently, the CPS visits about 100,000 (65,000 before 2002) addresses across all of the US every month. Every month, one-eighth of all addresses are replaced by new randomly chosen addresses, thus each address is visited and interviewed exactly eight times.[6] The visiting pattern is as follows: every address is visited four consecutive months, then left out for eight months, and then visited for four more months. In the dataset, the interviews are numbered by the *month in sample* variable. For example, a household could be visited monthly from February to May 2004 (months in sample 1-4), and then again February to May 2005 (months in sample 5-8). The list of questions asked varies from month to month, but generally consistent across years.

In this study, I use the data collected in March of years 1998-2007. The

[6] except a small number of addresses which became non-residential between visits

March survey is the most commonly used by economists and demographers, because it contains the most comprehensive list of socioeconomic questions. Since the interviews are conducted for two consecutive years, each address that was visited in March of year t, must have been also visited either in year $t-1$ or in year $t+1$ (but not both). Consider an example given above: an address visited from February to May 2004, and then again February to May 2005. Since we use March samples only, we observe this household twice: March 2004 (when it was visited for the second time, month in sample = 2) and in March 2005 (month in sample = 6). By observing people living at this address at both dates, we can identify those who have left during the year for whatever reason.

4.2 Matching the data

To match person records across years, we have to conduct three steps: first, match addresses across years; second, identify whether an address is occupied by the same household; third, match person records for the same household across years.

4.2.1 Matching addresses

Each address is identified by *household identification number*, which is supposed to be unique for a given combination of sample year and month-in-sample. In practice, however, there are many occasions of duplicate ID's before the year 2005 when the identification methodology was improved. The number of duplicate ID's peaked in year 2002, when only about 60% of ID's were unique. To prevent potential erroneous matches, I dropped all addresses which non-unique ID's. Since the ID's were assigned by the CPS staff, most likely they were not correlated with household characteristics, and therefore dropping ambiguous records should not bias the results. After removing ambiguous records, addresses were matched across years; records without a match were dropped.

Table 4: Number of duplicate address ID's

year	unique ID	2 duplicate ID's	3+ duplicate ID's
1998	64,656	0	3
1999	65,327	38	12
2000	64,857	78	9
2001	64,246	102	14
2002	61,283	33,930	3,635
2003	93,390	6,320	276
2004	93,324	5,372	283
2005	98,664	0	0
2006	97,352	0	0
2007	98,015	0	0

4.2.2 Matching households

To identify whether the same household lives at an address one year later, the CPS dataset contains the *household number*. In theory, the household number is equal to one during the first interview; in subsequent interviews, it remains the same if the address is occupied by the same household, and increments by one otherwise. In practice, the household number sometimes *decreases* over time (about 0.2% of all addresses), which implies it could be recorded with an error. An erroneous household number could result in both erroneous match (two different households are treated as one) and erroneous mismatch (two records one the same household are treated as different households), causing noise in observations. To account for these errors, I conduct additional checks as described below.

4.2.3 Matching individuals

Usually there are several people living in a household; these people are differentiated by the *line number*. The line number is constant over time for the same person. When a person moves out, the line number is left blank in subsequent interviews. When a new person moves in, he/she is assigned a new (unique) line number. However, when the entire household moves out and is replaced by another household, the line number count starts over. Thus, if two different households were erroneously treated as

Table 5: Person record matching outcomes, respondents of age 18-70 (percentage points in parentheses)

number of matching additional characteristics	same household number	other household number
All three	248,236 (92.85)	555 (0.74)
Two	12,614 (4.72)	3,066 (4.09)
One	4,612 (1.73)	8,820 (11.77)
None	1,887 (0.71)	26,281 (35.08)
No person record in year-2	0	36,194 (48.31)
Total	267,349 (100.00)	74,916 (100.00)

the same household, the line numbers of two different people could match. To estimate the likelihood of a possible mismatch, I check for consistency of other information supplied by individuals at different dates.

4.3 Quality of matching

To check whether person records were matched correctly, I check the consistency across years of the following three additional characteristics:

- gender
- age: generally should increase by one. Since the interviews were conducted not *exactly* one year apart, age remaining the same and increasing by two were also accepted
- migration status: "place of residence one year ago" reported in the second year. Respondents should report that they lived in the same place at the time of the first-year interview

The results are presented in table 5.

Overall, 267,349 (78.11%) of all person records could be matched across years according to *household number* parameter.[7] Of them, 92.85% have consistent sex, age, and migration status; 4.72% have a mismatch in one of

[7] In theory, we should check the consistency of the household number for all household members simultaneously. But for computational speed, all persons were treated independently

those characteristics; remaining records have a mismatch in two or all three characteristics.

These results could be produced by erroneously recorded personal characteristics. To approximate the probability of an error in a certain characteristic, I calculate the probability of a mismatch in this characteristic, conditional on all other characteristics matching. For example, there are 464 observations in which gender doesn't match, while household number, age, and migration status do. Similarly, there are 7,945 (4,205) observations in which age (migration status) is the only mismatching characteristic. Table 5 indicates that there are 555 records with a similar mismatch in the household number: while age, gender, and migration status match (meaning that this is most likely the same person), the household number is different.

Apparently, the household number and gender are much higher quality observations – they are ten times less likely to be recorded with an error. Most likely, there are fewer errors because these characteristics are identified by the interviewer, while age and migration status are reported by the respondent. It is quite likely that the respondent does not remember the exact date of moving to the current residence, or misunderstood the question. It is also possible that the respondent has rounded up his/her age. Figure 2 reports the distribution of respondents' age; there are clearly visible spikes at years 25, 30, 35, etc., which implies that a good number of people are rounding up. It is quite likely that people with certain characteristics (e.g. low education, or foreign-born) are more likely to round up age than others. For example, among natives, 1.98% of all respondents have a mismatch in age (while other characteristics match), while among foreign-born individuals, this figure is 4.43% – more than twice as high! Thus, using age as one of the matching criteria may lead to biased results in the analysis of return migration.

Throughout the paper, I match person records using the household number only. To check for robustness of results, I use an alternative matching rule: person records are matched if at least three out of four characteristics (household number, gender, age, migration status) match.

Figure 2: Distribution of reported age

4.4 Birthplace of CPS respondents

The CPS asks respondents about their place of birth, which allows to identify immigrants, and about their parents' place of birth, allowing to identify second-generation immigrants. Overall, about 100 distinct home countries can be identified with that data.

Before use, several adjustments had to be made to the birthplace data. First, I exclude observations with too vague birthplace categories such as "other Central America" or "other Africa". Second, I correct information on birth countries which no longer exist. For example, there are people who describe their birthplace as "Czechoslovakia" and those who were born in "Czech Republic". The criteria of choosing between the two options are not clear; the CPS does not provide any instructions regarding this issue. It is very likely that the choice between the two options was dependent on individual characteristics, and therefore using the data *as is* may lead to estimation bias. Another problem with dissolved countries of birth is that recently collected home country characteristics (e.g., recent GDP per capita) are not applicable to those countries, and therefore cannot be used

as regressors.

To handle these problems, I merge immigrants from dissolved countries with immigrants from the most likely successor countries. People born in Czechoslovakia were attributed to Czech Republic, those from Soviet Union were attributed to Russia. Since different successors of the same dissolved country have similar characteristics, the resulting bias should be small.

4.5 Death rate data

The probability of death within a year, disaggregated by gender and age, is supplied by National Center for Health Statistics (NCHS). For individuals younger than 50, the probability of death within a year is negligible; for those of age 70, it equals approximately 2.5% for males and 1.5% for females. Indeed, the immigrant's death probability may also depend on his/her birth country characteristics (e.g. malaria incidence) and on the number of years spent in the US, but this information is not available in NCHS statistics. The death probability may also depend on personal health status, but this information is not provided by the CPS. Hence, the death probability in the model depends on age and gender only. To reduce possible bias arising from limited data on death probability, I exclude individuals older than 70 from the analysis. The death probabilities by age are provided in figure 3.

5 Home country data

To study the effect of home country characteristics on return migration, I use several sources of country-level data. Most characteristics are disaggregated not only by country but also by year: this allows to study the effects of not only *levels*, but also *changes* in home country characteristics. Also, with only about 100 home countries observed, one cannot include more than 8-10 country characteristics (some of which are highly correlated) because of regressor collinearity problem. Disaggregation of country data by year greatly increases the number of macro-level observations, allowing to include all observed characteristics into the regression.

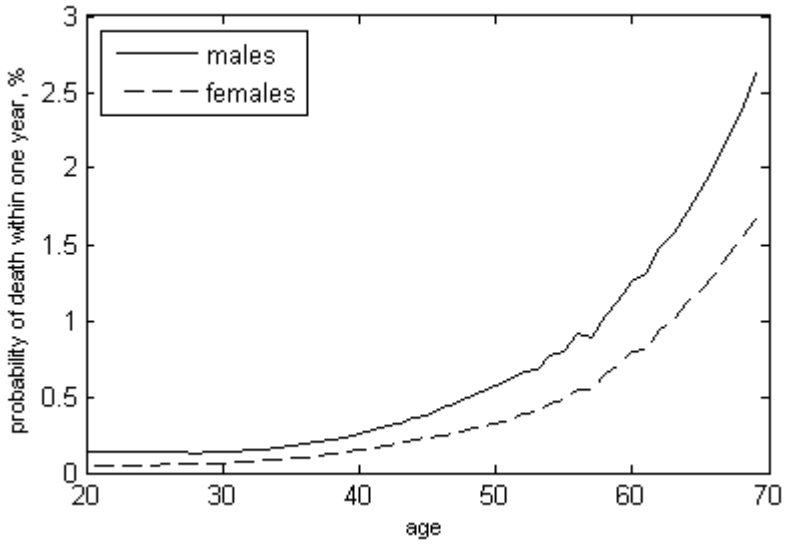

Figure 3: Death rates

The distance between the US and the home country was calculated using the EuGene software (Bennett Stam 2000); it is measured as the shortest distance between national capitals. The missing data was filled manually using Google Earth software.

The economic data was taken from World Economic Outlook database compiled by the International Monetary Fund (IMF data henceforth). The two statistics used were GDP per capita (based on Purchasing Power Parity (PPP), in current prices), and the "exchange rate" defined as the ratio of PPP-based GDP over nominal GDP.[8] This variable was created to verify Dustmann's (???) hypothesis that return migration may be caused by higher purchasing power of the US dollar at home.

Some countries and territories present in the CPS sample are missing in the IMF data (Bermuda, Cuba, Iraq, Puerto-Rico). Information on these countries was taken from PennWorld tables (PWT, Heston Summers Aten 2006). Since this data is available only until 2004, I extrapolated GDP per

[8] an alternative definition is nominal over PPP exchange rate

capita using information on GDP and population growth for these countries; to extrapolate the exchange rates, I simply extrapolated the last available observation.

Statistics on one more country – Myanmar – was taken from the CIA World Factbook, because neither IMF data nor PWT had reliable information on this country.

The data on the quality of institutions was taken from "Governance" dataset compiled by Kaufmann Kraay Mastruzzi (2006), which provides the following measures: "Voice and Accountability", "Political Stability", "Government Effectiveness", "Regulatory Quality", "Rule of Law", and "Control of Corruption". The data is measured biannually, with the last observation in 2004. The observations in 2001 and 2003 were imputed by interpolating the 2000,2002,and 2004 data; the 2005-7 observations are assumed to be equal to that of 2004.

In the Governance database, each country-year observation is made by interviewing a small number of experts that are familiar with the country. The observations are thus made with errors which are estimated by the dataset designers. Typically, the errors are higher in smaller and more remote countries, because fewer experts on these countries could be found. In theory, one has to account for these errors in regression analysis by giving a smaller weight to observations with higher error (the "total least squares" model). For the purpose of this research, however, these measurement errors were ignored: in the CPS data, there are usually fewer immigrants observed from smaller countries; therefore these smaller countries will receive a lower weight in the regression anyway.

Another problem in the Governance data is a very high correlation between some measures. For example, the "rule of law" and "control of corruption" measures have a correlation of almost 97%. Therefore, these measures cannot be used all at once.

Besides economic and political measures mentioned above, I include the following country dummies: English-speaking country (a measure of cultural similarity), an OECD country, a transition economy, a muslim country, a sub-saharan country, landlocked country, and a small island. The list of

English-speaking countries was taken from EuGene database, all others were borrowed from Docquier and Marfouk (2005) data.

6 Results

6.1 Basic model

The results of the model estimation are given in table 6. Both emigration propensity θ_e and propensity to move internally θ_m are reported. Two different models are presented, with different sets of home country regressors.

The propensity to move within the US θ_m has generally predictable patterns. Mobility decreases with age. Women are less mobile than men. Married people are less mobile if they live with their spouses, and more mobile if their spouse is away. Recent immigrants and non-citizens are more mobile than others.

An interesting finding is that immigrants from English speaking countries are more mobile within the US than others; this could be a result of their faster assimilation. This result is robust to changes in the model specification, see table 10 for comparison.

Our main estimation target, the propensity to emigrate, is presented in the first (and third) column of table 6. Since this parameter was basically computed as residual non-followup, many coefficients have a considerably higher standard error; nevertheless, most of them are significant.

The dependence of θ_e on personal characteristics has a pattern similar to that of θ_m. As one might expect, the effect of citizenship status on θ_e is much stronger than the effect on θ_m. It turns out that less educated immigrants are more likely to leave the US (the corresponding parameter in θ_e is negative), which could be not their own choice but the policy of the US government.

Immigrants from English-speaking countries and from OECD countries[9] return less often than others. This could be a result of faster assimilation, or more favorable attitude of the US government. From the basic model shown

[9] these groups are, indeed, not mutually exclusive

Table 6: Main results

regressor	Model 1: country groups emigrate (θ_e)	move within US (θ_m)	Model 2: country characteristics emigrate (θ_e)	move within US (θ_m)
constant	-2.417***	-0.001	-0.200	-0.140
	(0.219)	(0.063)	(2.078)	(0.604)
person characteristics				
age	-0.005*	-0.040***	-0.005*	-0.040***
	(0.003)	(0.001)	(0.003)	(0.001)
female	-0.202***	-0.116***	-0.200***	-0.119***
	(0.072)	(0.025)	(0.072)	(0.025)
married, spouse present	-0.780***	-0.283***	-0.742***	-0.283***
	(0.077)	(0.027)	(0.076)	(0.027)
married, spouse absent	0.393***	0.259***	0.400***	0.268***
	(0.131)	(0.064)	(0.132)	(0.064)
higher education	-0.219**	0.019	-0.221**	0.032
	(0.092)	(0.028)	(0.092)	(0.028)
immigrated 0-5 yrs ago	0.808***	0.487***	0.797***	0.497***
	(0.082)	(0.033)	(0.082)	(0.033)
non-citizen	0.706***	0.097***	0.678***	0.100***
	(0.102)	(0.029)	(0.101)	(0.029)
home country characteristics				
Mexico	0.321***	-0.023	0.417***	-0.026
	(0.089)	(0.034)	(0.101)	(0.037)
English-speaking country	-0.433***	0.086***		
	(0.139)	(0.033)		
OECD country	-0.412***	0.061*		
	(0.154)	(0.035)		
transition economy	-0.358	-0.103*		
	(0.224)	(0.062)		
muslim country	-1.196***	0.076		
	(0.449)	(0.057)		
Sub-Saharan Africa	0.645**	0.157		
	(0.328)	(0.104)		
small island country	0.067	-0.231***		
	(0.150)	(0.055)		
log(distance to US)			-0.196**	-0.007
			(0.079)	(0.021)
log(GDP per capita)			-0.258	0.067
			(0.838)	(0.244)
exchange rate			-0.127	0.027
			(0.117)	(0.032)
institutions			-0.055***	0.014**
			(0.021)	(0.006)
year fixed effects	yes	yes	yes	yes

* – significant at 90%, ** – at 95%, *** – at 99% level

in table 6, these two home country groups appear to have a very similar return migration pattern. However, when immigrants are disaggregated by length of stay in the US (next subsection), we can see that these two groups have very different migration patterns.

People living in muslim countries have very low rates of emigration to the West, including the United States (see Docquier and Marfouk 2005). From table 6, it follows that they are also far less likely to return – muslim immigrants appear to have made a firm choice not to go back, probably because of ideological differences. Borjas and Bratsberg (1996) made a similar finding about immigrants from Communist countries.[10] Home country ideology does matter in making a decision to return. Not surprisingly, modern immigrants from ex-communist countries (transition economy dummy) do not show any large differences from the rest of the sample – which changed ideology, migration patterns have also changed and became more "normal".

It is well known that residents of small island countries are far more likely to emigrate than others (Docquier and Marfouk 2005). Apparently, this is a one-way mobility: immigrants from small island countries are not any more likely to go back than other immigrants.

Immigrants from geographically closer countries are found to me more mobile than others; this is especially true for those from Mexico. Apparently, people from closer countries are more likely to travel back-and-forth than others. People who travel back and forth, obviously, stay in the US for shorter periods and thus more likely to fall in the "recent immigrant" category. If there are more back-and-forth migrants from Mexico, we might expect that the difference in θ_e between recently immigrated Mexicans and other Mexicans is greater than a similar difference among non-Mexicans. This hypothesis is verified below.

The negative effect of distance to home on return migration was pointed out by Borjas and Bratsberg (1996). However, their other finding, the positive effect of GDP on return migration, could not be confirmed. The exchange rate (the purchasing power of the US dollar in the home country) is also not significant. The effect of institutions, measured as the sum of all six

[10] they used data collected in the 1970s, at the peak of the cold war

Governance parameters, is of the "wrong" (negative) sign. Overall, we may conclude that economic and institutional characteristics of home countries cannot be used as powerful predictors of migration patterns; they were not used in the rest of the paper.

6.2 Emigration by group of interest

In demographic literature, it is common to treat males and females (especially immigrants) separately, because they are believed to follow very different patterns. My research, however, did not find vast differences between genders in their return migration pattern; table 7 reports the results. The only significant difference is that in the intercepts: males are more likely to emigrate than females. To account for this difference, it is sufficient to include a gender dummy into the regression.

We can also point out the large difference between muslim men and women: the coefficient for muslim women is -2.5, which means that they are $e^{2.5} \approx 12$ times less likely to emigrate than other women. This basically means that they never return; in such cases, the estimates have large standard errors which prevent us from making judgements about significance of these estimates.

Table 8 reports significant differences in emigration patterns between skilled and unskilled immigrants. The former are more likely to emigrate as they get older; the opposite is true for the latter. Given significantly different intercepts, this finding implies that differences between skilled and unskilled immigrants are very large when they are young, and become smaller as they get older. Young high skilled immigrants may experience a higher return to skill and/or expect to learn more while in the US; they are likely to postpone their decision to return until later. Young unskilled immigrants, on the other hand, are more likely to come to the US for short-period temporary jobs and leave thereafter. Figure 4 presents the probabilities of emigration by age, skill, and length of stay derived from table 8.

Skilled immigrants living with families are far more sedentary than their unskilled counterparts; the difference between them is significantly higher

Table 7: Emigration by gender

regressor	females: emigration	males: emigration	difference
constant	-2.941***	-2.121***	-0.820*
	(0.335)	(0.288)	(0.441)
person characteristics			
age	-0.000	-0.009**	0.009
	(0.004)	(0.004)	(0.006)
married, spouse present	-0.731***	-0.840***	0.109
	(0.111)	(0.109)	(0.155)
married, spouse absent	0.103	0.524***	-0.421
	(0.255)	(0.155)	(0.298)
higher education	-0.147	-0.285**	0.138
	(0.134)	(0.127)	(0.185)
immigrated 0-5 yrs ago	0.925***	0.708***	0.217
	(0.123)	(0.109)	(0.164)
non-citizen	0.789***	0.585***	0.204
	(0.151)	(0.135)	(0.203)
home country characteristics			
Mexico	0.285**	0.367***	-0.081
	(0.133)	(0.118)	(0.178)
English-speaking country	-0.476**	-0.413**	-0.063
	(0.199)	(0.194)	(0.278)
OECD country	-0.329	-0.481**	0.151
	(0.201)	(0.235)	(0.309)
transition economy	-0.421	-0.328	-0.093
	(0.324)	(0.314)	(0.452)
muslim country	-2.553	-0.976**	-1.577
	(2.289)	(0.487)	(2.340)
Sub-Saharan Africa	-0.001	0.910**	-0.911
	(0.812)	(0.355)	(0.886)
small island country	0.037	0.104	-0.067
	(0.209)	(0.218)	(0.302)
year fixed effects	yes	yes	

* – significant at 90%, ** – at 95%, *** – at 99% level

Table 8: Emigration by educational level

regressor	skilled: emigration	unskilled: emigration	difference
constant	-3.981***	-2.019***	-1.962***
	(0.503)	(0.244)	(0.559)
person characteristics			
age	0.014**	-0.011***	0.025***
	(0.007)	(0.003)	(0.008)
female	-0.087	-0.215***	0.128
	(0.160)	(0.080)	(0.179)
married, spouse present	-1.144***	-0.672***	-0.472**
	(0.181)	(0.084)	(0.200)
married, spouse absent	0.293	0.444***	-0.151
	(0.293)	(0.147)	(0.328)
immigrated 0-5 yrs ago	1.336***	0.625***	0.711***
	(0.182)	(0.093)	(0.205)
non-citizen	1.137***	0.560***	0.577**
	(0.252)	(0.113)	(0.276)
home country characteristics			
Mexico	0.512**	0.273***	0.240
	(0.216)	(0.096)	(0.236)
English-speaking country	-0.462**	-0.560***	0.097
	(0.214)	(0.200)	(0.293)
OECD country	-0.253	-0.563**	0.311
	(0.211)	(0.230)	(0.312)
transition economy	-0.665*	-0.069	-0.596
	(0.364)	(0.265)	(0.450)
muslim country	-1.712*	-1.417*	-0.295
	(0.892)	(0.808)	(1.203)
Sub-Saharan Africa	0.351	1.011**	-0.660
	(0.583)	(0.416)	(0.716)
small island country	-0.150	0.134	-0.284
	(0.375)	(0.168)	(0.411)
year fixed effects	yes	yes	

* – significant at 90%, ** – at 95%, *** – at 99% level

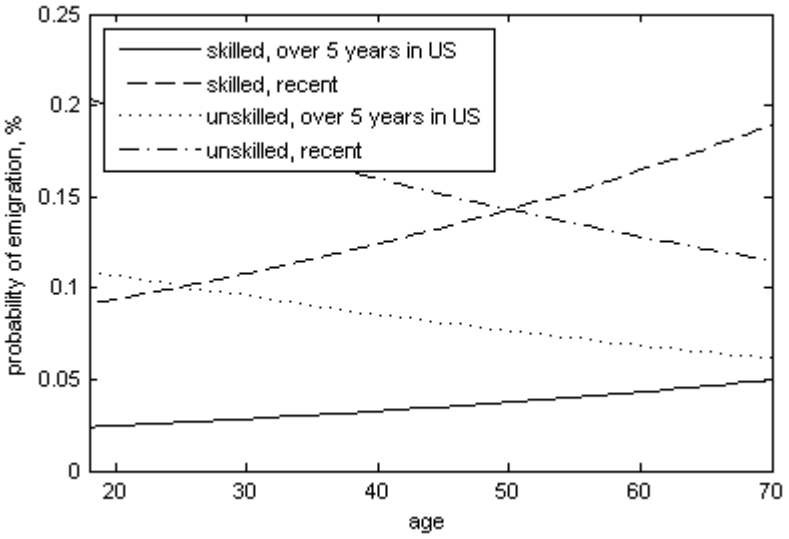

Figure 4: Emigration probability by age, skill, and length of stay in the US

than the difference between skilled and unskilled immigrants without families.

Table 9 reports differences between recent immigrants and others. Surprisingly, the difference in constants is not significant and is actually of the wrong sign: recent immigrants do not emigrate more often *other things being equal*. They emigrate more because other things are unequal, and because many of these other things have a very different impact on recent immigrants compared to others. In particular, we observe that recent individuals from OECD countries return home far more often than those who came to the US 5 or more years ago. The same is true for Mexicans.[11] Given that immigrants from OECD countries and Mexicans constitute more than half of all foreign-born population in the US, their behavior generates the positive sign for "recent immigrant" dummy in the benchmark model 6.

[11] In this research, Mexico was not included into the list of OECD countries, because it has very different migration patterns

Table 9: Emigration by length of stay

regressor	recent immigrants: emigration	other immigrants: emigration	difference
constant	-2.065***	-1.834***	-0.231
	(0.476)	(0.242)	(0.534)
person characteristics			
age	-0.004	-0.009***	0.005
	(0.006)	(0.004)	(0.007)
female	-0.166	-0.246***	0.081
	(0.127)	(0.086)	(0.153)
married, spouse present	-1.171***	-0.737***	-0.434**
	(0.157)	(0.090)	(0.181)
married, spouse absent	0.214	0.449***	-0.235
	(0.202)	(0.169)	(0.264)
higher education	0.011	-0.519***	0.531***
	(0.155)	(0.126)	(0.200)
non-citizen	0.840**	0.588***	0.252
	(0.338)	(0.107)	(0.355)
home country characteristics			
Mexico	0.548***	0.189*	0.359*
	(0.166)	(0.103)	(0.195)
English-speaking country	-0.501**	-0.402**	-0.099
	(0.217)	(0.173)	(0.277)
OECD country	0.684***	-1.028***	1.712***
	(0.199)	(0.280)	(0.344)
transition economy	-0.298	-0.484	0.185
	(0.290)	(0.367)	(0.468)
muslim country	-4.998	-1.030*	-3.968
	(20.451)	(0.533)	(20.458)
Sub-Saharan Africa	1.607***	0.184	1.423**
	(0.391)	(0.557)	(0.681)
small island country	0.429	-0.161	0.590*
	(0.291)	(0.183)	(0.343)
year fixed effects	yes	yes	

* – significant at 90%, ** – at 95%, *** – at 99% level

6.3 Robustness

Table 10 presents the results of an alternative method of estimation, which takes account for the possibility of "other non-followup". The resulting θ_e is generally similar to that of the benchmark model presented in table 6. The major differences are in the sign of "education" and "Sub-Saharan Africa" regressors. In the benchmark model, it may be the case that less educated people appear to be more mobile not because they actually emigrate more often, but because they drop out of CPS sample due to "other factors" more often than their more educated counterparts. The same may be true about immigrants from Sub-Saharan Africa. Using this alternative method for various groups of interest, however, did not prove efficient because it often produces near-negative-infinite estimates of return migration propensity: a small error in θ_m or θ_r may result in a very large error in θ_e.

7 Discussion and future work

This essay utilizes the American Current Population Survey (CPS) to estimate the return migration patterns of US foreign-born. The key feature of the CPS is that each household is interviewed eight times within two years. By using two of these eight interviews, made exactly one year apart, I infer which of the respondents have departed during this year. After adjusting for the probability of death and migration within the US, I estimate what factors affect immigrants' decision to return. Among expected results, I find that recent immigrants are more likely to leave; females are less mobile; married foreign-born emigrate less frequently, if their spouses live with them in the US, and more frequently otherwise.

The new results are that propensity to return increases with age for skilled migrants, and decreases for their unskilled counterparts, intersecting around the retirement age. Ideological differences matter: immigrants from muslim countries rarely return, while immigrants from ex-communist countries, who rarely returned in the 1970s, are not very different from others.

I also find a large difference between recent and non-recent immigrants

Table 10: Model with "other non-followup"

regressor	emigration	move, foreign-born	move, second generation	other non-followup
notation	θ_e	θ_m^f	θ_m^s	θ_r
constant	-4.982***	-0.148**	-0.325***	-0.854***
	(0.702)	(0.065)	(0.079)	(0.132)
person characteristics				
age	-0.001	-0.038***	-0.040***	-0.014***
	(0.006)	(0.001)	(0.002)	(0.002)
female	0.049	-0.099***	0.048	-0.244***
	(0.145)	(0.026)	(0.036)	(0.059)
married, spouse present	-0.278*	-0.243***	-0.159***	-0.729***
	(0.158)	(0.028)	(0.039)	(0.065)
married, spouse absent	0.738***	0.291***	0.524***	0.028
	(0.266)	(0.066)	(0.143)	(0.176)
higher education	0.406**	0.056*	0.035	-0.383***
	(0.165)	(0.029)	(0.037)	(0.064)
immigrated 0-5 yrs ago	1.415***	0.491***		
	(0.170)	(0.033)		
non-citizen	1.466***	0.115***		
	(0.456)	(0.029)		
home country characteristics				
Mexico	0.428**	-0.035	0.154***	0.065
	(0.192)	(0.035)	(0.056)	(0.079)
English-speaking country	-0.808***	0.071**	0.082*	-0.042
	(0.297)	(0.034)	(0.043)	(0.075)
OECD country	-0.046	0.075**	0.093**	-0.331***
	(0.235)	(0.036)	(0.046)	(0.080)
transition economy	-0.247	-0.080	-0.056	-0.298*
	(0.318)	(0.062)	(0.081)	(0.156)
muslim country	-1.369	0.035	0.062	-0.043
	(0.889)	(0.061)	(0.122)	(0.151)
Sub-Saharan Africa	-0.332	0.152	0.123	0.844***
	(1.515)	(0.113)	(0.305)	(0.211)
small island country	0.481	-0.183***	-0.135	-0.099
	(0.298)	(0.057)	(0.099)	(0.128)
year fixed effects	yes	yes	yes	yes

* – significant at 90%, ** – at 95%, *** – at 99% level

from OECD countries. The former emigrate more frequently than other recent immigrants, which could be a result of a brain circulation between the US and other developed countries. Those who spent more than five years in the US are extremely unlikely to return.

To improve the estimation methodology, the following things can be done. First, since the CPS data is collected errors, the algorithm of matching person records across years could be improved by creating a model of CPS data collection with explicitly defined error probabilities. These error probabilities can be estimated using the maximum likelihood method; given these estimates, it would be possible to estimate the probability of a match or mismatch of a given person record. The probability of mismatch in each observation could be subsequently used in the estimation of the main model parameters.

Another way to improve the results is to use not only March surveys, but also datasets collected in all other month. It would allow to increase the number of observations, and obtain eight records on each household instead of two. The latter improvement greatly increases the quality of matching of person records across time.

Finally, it is also possible to include family-level estimation errors into the model, to account for possible correlation of observation errors within the same family.

A The optimization problem

A.1 The main model

In the model presented in figure 1, the probability of a person i being followed up is

$$\hat{P}(X_i, \theta_e, \theta_m, \theta_r) = Pd(X_i)P(X_i, \theta_e)P(X_i, \theta_m)P(X_i, \theta_r)$$

The probability of not being followed up for whatever reason is, correspondingly, $1 - P$. If y_i is the indicator of followup and $\hat{P}(X_i, \theta_e, \theta_m, \theta_r) \equiv \hat{P}(X_i)$

for brevity, the log-likelihood function is

$$L = \sum_{y_i=1} \log \hat{P}(X_i) + \sum_{y_i=0} \log(1 - \hat{P}(X_i))$$

Its derivative with respect to any parameter θ_j is

$$\frac{\partial L}{\partial \theta_j} = \sum_{y_i=1} X_i'(-(1 - P(X_i, \theta_j))) + \sum_{y_i=0} X_i' \frac{\hat{P}(X_i)}{1 - \hat{P}(X_i)}(1 - P(X_i, \theta_j))$$

The expected second derivative, used to compute standard errors and speed up optimization, is

$$E\frac{\partial^2 L}{\partial \theta_{j1} \partial \theta_{j2}} = -\frac{\hat{P}(X_i)}{1 - \hat{P}(X_i)}(1 - P(X_i, \theta_{j1}))(1 - P(X_i, \theta_{j2}))X_i' X_i$$

[to be continued]

B Data

List of English-speaking countries

Australia, Bahamas, Barbados, Belize, Canada, Dominica, Fiji, Ghana, Grenada, Guyana, Hong Kong, India, Ireland, Jamaica, New Zealand, Nigeria, Philippines, Puerto Rico, Singapore, South Africa, Trinidad and Tobago, United Kingdom

List of OECD countries

Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Japan, South Korea, Netherlands, New Zealand, Norway, Poland, Portugal, Slovakia, Spain, Sweden, Switzerland, United Kingdom

Table 11: List of countries

country	# of records	country	# of records
Afghanistan	78	Israel	185
Argentina	166	Italy	589
Armenia	97	Jamaica	618
Australia	83	Japan	626
Austria	81	Jordan	70
Bahamas, The	21	Kenya	53
Bangladesh	113	Korea, South	996
Barbados	68	Laos	218
Belgium	63	Latvia	19
Belize	64	Lebanon	175
Bermuda	19	Lithuania	44
Bolivia	70	Malaysia	61
Brazil	318	Mexico	12886
Burma	41	Morocco	45
Cambodia	178	Netherlands	125
Canada	1330	New Zealand	38
Chile	124	Nicaragua	299
China	1233	Nigeria	155
Colombia	735	Norway	38
Costa Rica	86	Pakistan	269
Cuba	1288	Panama	115
Czech Republic	87	Peru	421
Denmark	43	Philippines	2174
Dominica	32	Poland	591
Dominican Republic	1018	Portugal	440
Ecuador	475	Puerto Rico	1799
Egypt	147	Romania	138
El Salvador	1443	Russia	649
Ethiopia	121	Saudi Arabia	30
Fiji	17	Serbia	187
Finland	20	Singapore	29
France	252	Slovakia	30
Germany	1480	South Africa	104
Ghana	115	Spain	151
Greece	223	Sweden	71
Grenada	37	Switzerland	53
Guatemala	671	Syria	66
Guyana	271	Taiwan	392
Haiti	539	Thailand	239
Honduras	446	Trinidad and Tobago	253
Hong Kong	234	Turkey	133
Hungary	116	Ukraine	212
India	1498	United Kingdom	1004
Indonesia	96	Uruguay	64
Iran	407	Venezuela	165
Iraq	135	Vietnam	1112
Ireland	227		

List of transition economies

Armenia, Czech Republic, Hungary, Latvia, Lithuania, Poland, Romania, Russia, Serbia, Slovakia, Ukraine

List of Sub-Saharan countries

Ethiopia, Ghana, Kenya, Nigeria, South Africa

List of muslim countries

Afghanistan, Bangladesh, Egypt, Guyana, Indonesia, Iran, Iraq, Jordan, Lebanon, Malaysia, Morocco, Nigeria, Pakistan, Saudi Arabia, Syria, Turkey

List of small island economies

Bahamas, Barbados, Belize, Bermuda, Dominica, Dominican Republic, Fiji, Grenada, Guyana, Haiti, Jamaica, Singapore, Trinidad and Tobago

References

[1] Kirdar, M. Labor Market Outcomes, Savings Accumulation and Return Migration: Evidence from Immigrants in Germany, mimeo

[2] Bellemare, C. (2004) A Life-Cycle Model of Outmigration and Economics Assimilation of Immigrants in Germany, CentER Discussion Paper, no. 29.

[3] Constant, A., D.S. Massey (2003). Self-selection, earnings, and outmigration: A longitudinal study of immigrants to Germany. Journal of Population Economics, vol. 16, no.4, pp. 631-653

[4] Heston, A., R. Summers and B. Aten, Penn World Table Version 6.2, Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania, September 2006.

[5] Cassarino, J.P (2004). Theorising Return Migration: the Conceptual Approach to Return Migrants Revisited. International Journal on Multicultural Societies, vol. 6, no.2, pp. 253-279. UNESCO.

[6] Van Hook, J., W. Zhang, F.D. Bean, J.S. Passel (2006). Foreign-Born Emigration: a New Approach and Estimates Based on Matched CPS Files. Demography, vol. 43, no. 2; p. 361

[7] Madrian, B. and L. J. Lefgren (1999). A Note on Longitudnally Matching Current Population Survey Respondents. NBER Working Paper T0247.

[8] Warren, R., J.M. Peck (1980). Foreign-Born Emigration from the United States: 1960 to 1970. Demography, vol.17, no.1, p.71

[9] Bennett, D. Scott, and Allan Stam (2000). EUGene: A Conceptual Manual. International Interactions 26: 179-204. Website: http://eugenesoftware.org.

[10] Jasso, G., M.R. Rosenzweig (1988). How well do U.S. immigrants do? Vintage effects, emigration selectivity, and occupational mobility. Research in population economics, vol.6: a research annual, edited by T. Paul Schultz. Greenwich, Connecticut, JAI Press, 1988. pp. 229-53.