INTERNATIONAL MIGRATION REVIEW

Determinants of International Migration Flows to and from Industrialized Countries: A Panel Data Approach Beyond Gravity[1]

Keuntae Kim
University of Wisconsin–Madison

Joel E. Cohen
Rockefeller University & Columbia University

We quantified determinants of international migratory inflows to 17 Western countries and outflows from 13 of these countries between 1950 and 2007 in 77,658 observations from multiple sources using panel-data analysis techniques. To construct a quantitative model that could be useful for demographic projection, we analyzed the logarithm of the number of migrants (inflows and outflows separately) as dependent variables in relation to demographic, geographic, and social independent variables. The independent variables most influential on log inflows were demographic [log population of origin and destination and log infant mortality rate (IMR) of origin and destination] and geographic (log distance between capitals and log land area of the destination). Social and historical determinants were less influential. For log outflows from the 13 countries, the most influential independent variables were log population of origin and destination, log IMR of destination, and log distance between capitals. A young age structure in the destination was associated with lower inflows while a young age structure in the origin was associated with higher inflows. Urbanization in destination and origin increased international migration. IMR affected inflows and outflows significantly but oppositely. Being landlocked, having a common border, having the same official language, sharing a minority language, and colonial links also had

[1] J. E. C. acknowledges with thanks the support of U.S. National Science Foundation grant DMS-0443803, the assistance of Priscilla K. Rogerson, and the hospitality of Mr. and Mrs. William T. Golden and their family during this work. This NSF grant also partially supported the work of K. K., who began this project as an intern at the UNPD in summer 2008. K. K. thanks Taeke Gjaltema and Patrick Gerland for encouragement and guidance.

DOI: 10.1111/j.1747-7379.2010.00830.x

statistically significant but quantitatively smaller effects on international migration. Comparisons of models with different assumed correlation structures of residuals indicated that independence was the best assumption, supporting the use of ordinary-least-squares estimation techniques to obtain point estimates of coefficients.

INTRODUCTION

The volume of immigrants to more developed nations has grown significantly over the last four decades. The end of the Cold War in the early 1990s ended some regimes that restricted migration (Massey, 1999). The annual number of immigrants to 17 selected Western countries increased after the mid-1990s, with a few exceptions such as Croatia and Germany.[2] In countries that experienced declines of fertility and rapid population aging, international migration became increasingly important. Net immigration accounts for roughly 40% of population growth in the United States and about 90% in the EU-15 countries (Howe and Jackson, 2006; Bijak, 2006). Immigrants or individuals of mixed origin could become a majority in these societies if immigration into more developed countries continues (Coleman, 2006). International migration affects demographics, economies, cultures, and politics around the world. The demand for reliable methods to project international migratory flows is greater than ever.

Fewer studies quantify the non-economic factors that influence international migration than investigate the consequences of international migration. This discrepancy may be due to a paucity of data on international migration streams (Vogler and Rotte, 2000; Mayda, 2005). Most studies that address determinants of international migration either neglect migration *from* wealthy nations to the rest of the world or treat these flows as subject to the same forces that influence immigration *to* rich countries. However, the determinants of immigration into affluent nations might be different from the determinants for emigration from affluent nations (Massey, 2006), and the determinants of migrant flows in both of these directions might differ from the determinants of ‘‘south–south’’ migration among developing countries. In contrast to rising immigration,

[2] More detailed figures of migration flows are provided in Figure S1 (inflows by destination countries) and Figure S2 (outflows by origin countries) of the supporting information.

annual emigration *from* the 17 specified Western countries *to* the rest of the world showed no clear upward or downward trend in most of the countries. Either different factors drove outflows or inflow factors exerted influence differently from outflow factors.

Most past studies on international migration treated a single destination country such as the United States (Isserman *et al.*, 1985; Greenwood and McDowell, 1999; Clark, Hatton, and Williamson, 2007), the United Kingdom (Hatton, 2005; Mitchell and Pain, 2003), and Germany (Vogler and Rotte, 2000) or a small conglomeration such as North American destinations (Greenwood and McDowell, 1991; Karemera, Oguledo, and Davis, 2000). Those countries are among the wealthiest nations and have similar characteristics. Today's international migration is not limited to those destinations. We need a more complete picture of international migration.

Fertig and Schmidt (2000) observed that research on the driving forces of international migration emphasized economic variables (*e.g.,* income and employment) and neglected demographic factors (*e.g.,* age structure, health, and life expectancy). Fertig and Schmidt argued that to predict economic variables is very difficult and that macro-economic conditions might be influenced by previous migration.

This paper investigates non-economic variables as predictors of international migration. Because economic and demographic factors are closely related, the present study leaves open the option of using demographic variables like life expectancy, infant mortality rate (IMR), and potential-support ratio (PSR) as proxies for economic or living conditions of countries. Because many demographic variables change more slowly (on a scale of quinquennia to generations) than many economic variables (on a scale of quarter-years to several years), this paper explores models of international migratory flows (not stocks) using only demographic, geographic, and very slowly changing social or unchanging historical variables in extensions of the gravity model. Determinants of immigration into affluent nations are compared to determinants of emigration from affluent nations. To test and extend the methods of Cohen *et al.* (2008), this paper employs panel-data analysis to investigate the correlations of residuals within a panel. Here, a panel is defined as a pair consisting of an origin country and a destination country. We use generalized estimating equations (GEE) for model specifications and quasi-likelihood under the independence model information criterion (QIC) for model selections (Hardin and Hilbe, 2003; Cui, 2007).

Section 2 of this paper surveys theoretical discussions on the determinants of international migration and results of empirical studies focused chiefly on gravity models. Section 3 reports this study's methods and empirical model. Section 4 reports the results. Section 5 discusses some limitations of the results. Section 6 draws conclusions.

BACKGROUND

Many theories of international migration have been proposed (Howe and Jackson, 2006). Massey *et al.* (1993) described six theoretical frameworks, with different strengths and weaknesses, that purport to explain international migration: neoclassical theory, new economics theory, dual (segmented) labor market theory, world system theory, social capital theory, and cumulative causation theory. Rogers (2006) reviewed four techniques for modeling migration: linear regression models, gravity models, Markov chain models, and matrix population models.

We chose a gravity model as our framework because it yielded results that were easy to interpret, and because recent developments in panel-data analysis enable estimation based on the model. The gravity model, in its simplest form, views migration as determined by the sizes of the populations of destination and origin and the distance between origin and destination:

$$M_{ij} = k \cdot \frac{P_i P_j}{d_{ij}}, i \neq j \tag{1}$$

where M_{ij} denotes the number of migrants from origin i to destination j, P_i denotes population of i, P_j denotes population of j, d_{ij} refers to distance between i and j, and k denotes a constant.

The gravity model is a phenomenological description. It predicts that, all other things being equal, countries with large populations send more emigrants to destinations than countries with small populations, and that countries with large populations attract more immigrants. The greater the distance between origin and destination, the smaller the migration predicted.

In the remainder of this section, we develop hypotheses about factors affecting international migration on the basis of prior empirical studies and simple arguments. We test these hypotheses later.

Empirically, the effect of distance between two countries is negative, significant, and robust across different model specifications (Greenwood and McDowell, 1982; Mayda, 2005). Increases in distance can be a proxy for increases in transportation cost and psychic cost (Greenwood, 1975). Persons tend to have less information about relatively distant places and are less likely to move to a locale about which they have little or no prior information.

This argument suggests that if two countries share a border, the cost of moving could be significantly lower than otherwise, while a relatively inaccessible destination, for example, a land-locked country, should have fewer immigrants than countries with oceans or seas as borders, due to the increased cost of over-land transportation (Mayda, 2005).

Language, culture, and shared history also affect international migration (Greenwood and McDowell, 1982; Karemera, Oguledo, and Davis, 2000; Mayda, 2005; Neumayer, 2005; Clark, Hatton, and Williamson, 2007). For example, Clark, Hatton, and Williamson (2007) found that having an English-speaking origin significantly and positively affected U.S.-bound immigration. Former colonial relationships appear to facilitate both trade and migration. The former colonial power's language is often spoken in the former colony, and the former colonial power may host many people from a former colony – people who can help migrants from the former colony find jobs and assistance in the new environment (Neumayer, 2005). Former colonial links consistently and significantly increased international migration in empirical studies (Karemera, Oguledo, and Davis, 2000; Mayda, 2005; Neumayer, 2005; Pedersen, Pytlikova, and Smith, 2008).

Neumayer (2005) suggested that people living in cities are likely to be better informed than rural inhabitants about international migration. Also, migrants go to cities in developing countries to get visas and documents for legal migration or make arrangements for illegal migration (Martin, 2003). Therefore, a higher percentage of an origin country's urban population is expected to become international migrants than the corresponding percentage of the origin's rural population. In a destination country, relatively large urban populations might indicate better job opportunities for newly arrived immigrants and a greater likelihood of getting help from people who came from the same origin. Furthermore, world system theory suggests that global cities in destination countries, such as New York, London, or Tokyo, concentrate wealth and a highly educated workforce and create strong demands for unskilled workers from overseas (Massey *et al.*, 1993). Frey (1996) observed that recent immigrants to the United States tended to stay in a small number of traditional port-of-entry cities, which are the largest metropolitan areas in the United States. If this observation holds true over time, large urban populations in the origin and the destination should be associated with large numbers of international migrants.

The age structure of a population may also affect international migration. For example, a low PSR, defined as the number of people aged 15–64 per person aged 65 or over, indicates population aging, and (depending on retirement ages and labor-force participation rates among the elderly) may indicate a shortage in the working-age population and a destination's economic demand for immigrants workers. Currently, most developed countries have a low PSR and sometimes express a need for a larger percentage of working-age people. Hence, if all other conditions are equal, an origin with a high PSR would be expected to send more migrants to wealthy destinations than would an origin with a low PSR. Also, all other things being equal, a destination with a low PSR would be expected to attract more immigrants than a destination with a high PSR.

Infant mortality rate and life expectancy at birth are demographic indices of quality of life for whole populations because factors affecting the health of an entire population have a significant impact on the mortality of infants (Reidpath and Allotey, 2003). For less developed countries, IMR or life expectancy might be the only available measures of health or quality of life. Thus, *ceteris paribus*, an origin with a high IMR or a low life expectancy might be expected to send more emigrants to a destination than an origin with a low IMR or a high life expectancy. And *ceteris paribus*, a destination having a high IMR would be expected to attract fewer immigrants than a destination having a low IMR.

METHODS

Data and Variables

Descriptive statistics and data sources for all variables in this analysis are presented in Table 1.[3] The source for numbers of migrants is "International Migration Flows to and from Selected Countries: The 2008 Revision," then unpublished and subsequently published as United Nations (2009b).

[3] The complete raw data are available on-line in two files, inflow.csv and outflow.csv, which are in plain text with comma-separated variables. The variables in those files are defined in the supporting information section Table S6, with further details here.

TABLE 1

Descriptive Statistics for Variables in the Inflow and Outflow Models and Data Sources. Log = $\log_{10}$

	Inflow					Outflow				
	N	Mean	SD	Min	Max	*N*	Mean	SD	Min	Max
Migrants	48832	1709.33	7549.92	1	269012	28826	1035.02	6144.66	1	220263
Log migrants	48832	2.07	1.10	0	5.43	28826	1.64	1.06	0	5.34
Log population (destination)	48832	7.28	0.72	5.39	8.48	28826	7.00	0.85	1.70	9.12
Log population (origin)	48832	6.87	0.90	1.70	9.12	28826	6.97	0.59	5.39	7.92
Log distance between capitals (km)	48832	3.75	0.38	1.91	4.29	28826	3.69	0.43	1.91	4.29
Log land area (destination)	48832	5.92	0.82	4.48	7.00	28826	5.29	0.99	0.70	7.23
Log land area (origin)	48832	5.17	1.06	0.70	7.23	28826	5.45	0.54	4.48	6.89
Log potential support ratio (destination)	48832	0.70	0.09	0.54	0.90	28082[a]	1.00	0.26	0.50	1.85
Log potential support ratio (origin)	46978[a]	1.02	0.25	0.50	1.85	28826	0.67	0.08	0.54	0.88
Log infant mortality rate (destination)	48832	−2.10	0.23	−2.52	−1.51	28082[a]	−1.53	0.47	−2.52	−0.58
Log infant mortality rate (origin)	46978[a]	−1.48	0.46	−2.52	−0.58	28826	−2.13	0.22	−2.52	−1.51
Log percentage of urban population (destination)	48832	1.89	0.05	1.74	1.99	28082[a]	1.68	0.25	0.34	2.00
Log percentage of urban population (origin)	46978[a]	1.67	0.25	0.34	2.00	28826	1.90	0.05	1.74	1.99
Landlocked (destination)	48832	1.06	0.74	1	10	28826	2.33	3.20	1	10
Landlocked (origin)	48832	2.38	3.24	1	10	28826	1.10	0.95	1	10
Border	48832	1.23	1.41	1	10	28826	1.30	1.61	1	10
Common official language	48832	2.75	3.56	1	10	28826	2.02	2.85	1	10
9% minority speak same language	48832	2.94	3.70	1	10	28826	2.01	2.84	1	10
Colonial link	48832	1.37	1.79	1	10	28826	1.30	1.61	1	10
Year – 1985	48832	4.84	11.95	−35	22	28826	5.81	11.27	−26	22
$(\text{Year} - 1985)^2$	48832	166.27	168.33	0	1225	28826	160.78	148.87	0	676

Notes: [a]Data for these variables are available only for countries with population size >100,000.

Number of migrants is from *International Migration Flows to and from Selected Countries: The 2008 Revision* (United Nations, 2009b). Population, land area, potential support ratio, infant mortality ratio, and percent of urban population are taken from *World Population Prospects: 2006 Revision* (United Nations, 2006a,b,c). Distance between capital cities, landlocked location, border, common official language, ethnic minority language, and colonial link are from Centre d'Etudes Prospectives et d'Informations Internationales (CEPII) <http://www.cepii.fr/>.

It contains time-series data on the flows of international migrants recorded by 17 countries (Australia, Belgium, Canada, Croatia, Denmark, Finland, France, Germany, Hungary, Iceland, Italy, New Zealand, Norway, Spain, Sweden, the United Kingdom, and the United States). These data concern only legal migration reported by each country's national agencies in charge of collecting migration data. Canada, France, Spain, and the United States do not provide information about emigration to other countries. Here, inflow refers to people coming into those 17 countries while outflow denotes people moving out of the 13 countries. Inflow may be from other developed countries, including the 17 sources of inflow data, and outflow may be to other developed countries, including the 13 sources of outflow data.

The inflow data represented 230 origin countries while the outflow data represented 216 destination countries. Although the United States has had inflow data since 1946, the earliest data point in this analysis is 1950 because only from this year forward are all other demographic variables available in the United Nations demographic data base (demobase). Not all countries reported migration information for the full time period, so the data set is not perfectly balanced in the sense of panel-data analysis. Whenever a country reported zero migrants, the observation was excluded. After the elimination of reports of zero migrants, there were 77,658 observations (48,832 for inflows and 28,826 for outflows).

Another major data source is the UNPD's data base called "demobase" that stores all estimates and projections for publication (United Nations, 2006a,b,c). Demobase is based on the medium variant of estimates and projections. For origins and destinations, demobase provided the total populations each year, the surface areas (in square kilometers), the PSRs, the life expectancy at birth, the IMRs, the proportions of populations aged 15–24, and the proportions of the populations considered urban.

From the Centre d'Etudes Prospectives et d'Informations Internationales (CEPII, or the French Research Center in International Economics)[4] came data on distances between geographical regions, official languages, colonial relationships, and proportions of a destination country's ethnic minorities who speak the origin country's language (Glick and Rose, 2002).

[4] The website is <http://www.cepii.fr/anglaisgraph/bdd/distances.htm> (accessed 15 October, 2010).

Dependent Variable. The dependent variable[5] of our models is the logarithm of the annual number, m_{ijt}, of migrants from origin country i to destination country j in year t. All logs here refer to base-10 logarithms. Normally, the year refers to the calendar year, but we noted an exception for U.S. data below.

We excluded migrant-related information involving geographical regions of multiple countries (*e.g.,* African Commonwealth).[6] Also, we excluded countries that, in the original data, lacked country codes. For instance, the study excludes Taiwan because the United Nations recognizes the island as a province of China. The term "migrants" here refers to foreign-born people who obtained a residence permit or a work permit from the destination. Hence, for example, we excluded Australian citizens who had settled abroad and later moved back to Australia. In addition, some countries such as Germany maintain separate migration-registration systems for foreigners and citizens. We excluded all data for in- and out-migration of countries' own citizens. Although demobase assigns country codes for Hong Kong and Macao and provides separate migration flows for these areas, we treated their migrants as Chinese migrants.

In the U.S. data, "year" refers to fiscal year. Until 1976, fiscal years ran from July 1 of a calendar year to June 30 of the following calendar year. In 1976, fiscal years were adjusted to run from October 1 of a calendar year to September 30 of the following calendar year. Hence, there were two migration reports in 1976, and we combined the two reports. Also, for the fiscal years 1989 through 1998, the U.N. data presented separate reports regarding persons legalized under the U.S. Immigration Reform and Control Act of 1986 (IRCA). Since those persons resided in the United States before the enactment of the IRCA, they cannot be

[5] Dependent variable and independent variable are terms frequently used in econometrics (*e.g.,* Wooldridge, 2006). Some users of non-experimental regression models prefer the equivalent terms outcome or response rather than dependent variable, and explanatory variables or regressors rather than independent variables. The choice of terms is a matter of taste.

[6] There is no such entity as an "African commonwealth." There is a British commonwealth, of which African countries are a part, and there is an African Union. The UNPD uses the term "African commonwealth" although it does not conform to United Nations practices. Because UNPD draws data from national statistical offices, *e.g.,* the Office for National Statistics in the United Kingdom, the original country nomenclature is maintained whenever standardization of the country code is not possible. Here, we followed UNPD's practice.

considered migrant flows that occurred during those years. Rather, these people constituted immigrant stocks in the United States. We excluded these people from the analysis. We also excluded countries, such as Czechoslovakia, the USSR, Yugoslavia, Serbia and Montenegro, and the German Democratic Republic, that no longer officially exist owing to separation or unification.

Independent Variables. We now list several independent variables. First are the population of the destination and the population of the origin.

Urbanization is the percentage of urban population, constructed by dividing the urban population in the given year by the total population of that year and multiplying by 100.

The PSR is 100 times the number of persons aged 15–64 divided by the number of persons 65 or older. Demobase furnishes only quinquennial estimates for the numerator and the denominator of PSR, and we linearly interpolated annual estimates by assigning one fifth of the 5-year change to each year.

The IMR is the probability (between 0 and 1) that a live birth died before 1 year of age for boys and girls combined. IMR is a proxy for overall living conditions and well-being.[7] Demobase provides only quinquennial IMR estimates for each country. We linearly interpolated annual estimates. In demobase, IMR is available only for countries with more than 100,000 inhabitants in 2007. As a result, IMR for small countries was not available and the number of observations of IMR was smaller than the numbers of observations of other demographic variables.

An official or national language is defined as a language spoken by at least 20% of the population of a country (Mayer and Zignago, 2006). If the destination and the origin had a common official language, the independent variable "common official language" is defined to equal 10; otherwise, the variable was 1. The values of 10 and 1 were chosen because $\log_{10}10 = 1$ and $\log_{10}1 = 0$, so the logarithms became a standard dummy variable with values 1 and 0. The independent variable called "common second language" is 10 if a specific language was spoken by at least 9% of the population in both the origin and the destination; 1 otherwise.

[7] Preliminary analysis suggested that IMR has better explanatory power than life expectancy at birth as a proxy for economic conditions.

Geographical distance is defined as the distance (in kilometers) between the two capital cities. Distances were calculated from the cities' longitude and latitude using the great circle formula (Mayer and Zignago, 2006).

A country is coded 10 if it is landlocked and, otherwise, 1. If two countries share a common border, the independent variable for having a common border is set to 10 and, otherwise, to 1.

When two countries have had a colonial or post-colonial relationship of colonizer to colonized for a relatively long period of time and when the (possibly former) colonizer substantially participated in the governance of the colonized country (Mayer and Zignago, 2006), the independent variable for colonial relations is set to 10 for colonial relations; and to 1 otherwise.

Chronological time is represented by continuous variables in all but one of the models we considered, and by dummy variables in one model. Time is usually represented by the sum of a linear variable, calendar year (in the Western calendar) – 1985, plus a quadratic variable, (calendar year – 1985)^2. To avoid ill-conditioning, 1985 is subtracted from year as an approximate centering. All other independent variables had mean values between −5 and +10 whereas if year and year^2 had been used without approximate centering, they would have had mean values 3–6 orders of magnitude larger. In one model only, each year is represented by a dummy variable. For example, the dummy variable for 1970 takes the value 1 when the year of the data is 1970 and takes the value zero otherwise. There were 57 dummy variables for years 1951–2007 in the inflow model 2 (M2) (explained below) and 48 dummy variables for years 1960–2007 in the outflow M2 (explained below).

As Cohen *et al.* (2008) observed in different data, destination population was highly correlated with destination area, and origin population was highly correlated with origin area. To check for multicollinearity among some independent variables, we calculated variance inflation factors (VIFs) for all the independent variables in the inflow model and the outflow model.[8] The mean VIF for variables in the inflow model was 2.40, and none of the VIFs for each variable exceeded 10. In the outflow model, the mean VIF for variables was 2.49, and none of the VIFs for each variable was greater than 10. Therefore, multicollinearity seems unlikely to be a concern in this study.

[8] We used collin routine in Stata (version 10.1). We did not include the linear or quadratic year variables in calculation of VIFs.

Empirical Model

The gravity model,[9] equation (1), is log-linear. A natural generalization estimates rather than assumes the exponents:

$$\log(m_{ijt}) = \beta_0 + \beta_1 \log(P_i) + \beta_2 \log(P_j) + \beta_3 \log(d_{ij}) + \varepsilon_{ijt} \qquad (2)$$

In equation (2), the gravity model suggests that $\beta_1 > 0$ and $\beta_2 > 0$ but $\beta_3 < 0$. We expanded the gravity model by adding to it more independent variables which might promote or deter migration:

$$\begin{aligned}\log(m_{ijt}) = {} & \beta_0+\beta_1\log(P_{it})+\beta_2\log(P_{jt})+\beta_3\log(PSR_{it})+\beta_4\log(PSR_{jt})\\ & +\beta_5\log(IMR_{it})+\beta_6\log(IMR_{jt})+\beta_7\log(urban_{it})+\beta_8\log(urban_{jt})\\ & +\beta_9\log(D_{ij})+\beta_{10}\log(LA_i)+\beta_{11}\log(LA_j)+\beta_{12}\log(LL_i)\\ & +\beta_{13}\log(LL_j)+\beta_{14}\log(LB_{ij})+\beta_{15}\log(OL_{ij})+\beta_{16}\log(EL_{ij})\\ & +\beta_{17}\log(COL_{ij})+\beta_{18}(Year-1985)+\beta_{19}[(Year-1985)^2]+\varepsilon_{ijt}\end{aligned} \qquad (3)$$

[9] The gravity model and the population potential model have such close conceptual and historical associations that they are almost indistinguishable (Isard, 1998). Duncan, Cuzzort, and Duncan (1963) defined the population potential $\mathrm{PP_i}$ for i^{th} areal unit in a universe of territory as $\sum_{j\neq i}^{n}(P_j/D_{ij})$, where P_{j} is the population of the *j*th area and D_{ij} is the distance of location *i* from location *j*. The primary purpose of including (generalized) population potential in a model is to control for the impact of other geographical units on local social processes. For example, church attendance rate in a county might be higher than expected because the county is surrounded by counties having very high rates of church attendance. Land and Deane (1992) proposed a two stage least squares (2SLS) estimation technique to accommodate large samples. Although 2SLS is computationally efficient compared to the maximum likelihood estimation, it would not be consistent if all the exogenous independent variables in the model are irrelevant (Lee, 2007). Multicollinearity problem can be pronounced when using 2SLS estimation (Wooldridge, 2006). Thus, Lee (2007) proposed using generalized method of moments (GMM) estimation when estimating spatial-effects model. To overcome the limitation of GMM (*see* more details in the Supporting Information), we used GEE. Therefore, our use of population potentials in the form of a generalized gravity model is sufficient to control for spatial effects in the data.

where the origin i and the destination j in year t are identified by subscripts, P_{it} and P_{jt} denote populations, PSR_{it} and PSR_{jt} denote the PSR, IMR_{it} and IMR_{jt} denote infant mortality, "urban" refers to percentage of total population that is urban, D_{ij} is the distance between the two capital cities, LA_i and LA_j denote land surface area of the origin and destination, *LL* stands for landlocked location, *LB* stands for shared border, *OL* stands for shared official language, *EL* refers to shared minority language, and *COL* stands for colonial relationship.

RESULTS

The percentage distributions of migrants for each period by the major regions of origin for inflow and by the major regions of destination for outflow indicated that the share of non-European immigrants to the 17 countries increased while those who emigrated from the 13 countries increasingly moved to non-European countries.[10] Countries varied greatly in mean numbers of immigrants and emigrants.

Table 2 for inflow data and Table 3 for outflow data present the results of pooled ordinary least square (OLS) regressions and other model specifications.

Equation (3) specifies model 1 (M1) in Tables 2 and 3. A plot of the residuals of M1 against predicted values suggested heteroscedasticity.[11] To test for homoscedasticity, we conducted the Breusch–Pagan/Cook–Weisberg test (Breush and Pagan 1979; Cook and Weisberg 1983).[12] The null hypothesis of the test was that the variance of residuals was homogeneous. The Breusch–Pagan chi-square statistic was 35.66 with 1 df ($p < 0.00005$) for inflow (M3 in Table 2) and 18.34 with 1 df ($p < 0.00005$) for outflow (M3 in Table 3), rejecting the null hypothesis of homoscedasticity at the levels shown.

Heteroscedasticity does not necessarily cause bias in the estimated coefficients, but may misleadingly deflate estimates of standard errors and, consequently, may exaggerate statistical significance (Frees, 2004). The on-line Appendix describes methods of estimation in the possible presence

[10] More detailed percentage distribution of inflows by origin and outflows by destination are provided in Tables S1 and S2, respectively, in the supporting information.

[11] Plots of residuals against fitted values for inflow (M1 in Table 2) and outflow (M1 in Table 3) are available in Figure S3 in the supporting information.

[12] We used Stata command hettest to test heteroscedasticity.

TABLE 2

Ordinary Least Square (OLS) and Generalized Estimating Equations (GEE) Regression Analysis of Inflows to 17 Selected Countries, 1950–2007. For Example, in Model 1 (M1), Log(migrants) into the 17 Developed Countries Varied in Proportion to 0.601 Times Log Population of the Destination, Where the Standard Error of the Estimated Coefficient 0.601 was 0.009

	Dependent variable: Log(Migrants)					
	M1	M2	M3	M4	M5	M6
	OLS	OLS	OLS (Beta)	GEE (ind)	GEE (exc)	GEE (ar1)
Demographic determinants						
Log population (destination)	0.601*** (0.009)	0.602*** (0.009)	0.391 (0.009)	0.601*** (0.035)	0.560*** (0.037)	0.721*** (0.029)
Log population (origin)	0.728*** (0.006)	0.728*** (0.006)	0.507 (0.006)	0.728*** (0.031)	1.028*** (0.052)	0.683*** (0.028)
Log potential support ratio (destination)	−0.811*** (0.069)	−0.806*** (0.071)	−0.066 (0.069)	−0.811*** (0.241)	−0.303 (0.236)	−0.901*** (0.240)
Log potential support ratio (origin)	0.045** (0.020)	0.043** (0.020)	0.010 (0.020)	0.045 (0.079)	−0.141 (0.116)	−0.253*** (0.079)
Log infant mortality rate (destination)	1.007*** (0.049)	1.018*** (0.052)	0.213 (0.049)	1.007*** (0.156)	−0.256** (0.123)	−0.568*** (0.132)
Log infant mortality rate (origin)	−0.466*** (0.013)	−0.465*** (0.013)	−0.197 (0.013)	−0.466*** (0.054)	0.396*** (0.071)	−0.304*** (0.052)
Log percentage of urban population (destination)	3.057*** (0.072)	3.067*** (0.073)	0.132 (0.072)	3.057*** (0.245)	3.387*** (0.473)	3.434*** (0.257)
Log percentage of urban population (origin)	0.332*** (0.017)	0.330*** (0.017)	0.077 (0.017)	0.332*** (0.078)	1.054*** (0.107)	0.449*** (0.075)
Geographic determinants						
Log distance between capitals	−0.819*** (0.011)	−0.822*** (0.011)	−0.286 (0.011)	−0.819*** (0.049)	−0.923*** (0.061)	−0.693*** (0.047)
Log land area (destination)	0.234*** (0.008)	0.234*** (0.008)	0.175 (0.008)	0.234*** (0.030)	0.323*** (0.034)	0.233*** (0.029)
Log land area (origin)	−0.047*** (0.005)	−0.047*** (0.005)	−0.039 (0.005)	−0.047* (0.026)	−0.286*** (0.038)	−0.019 (0.024)
Landlocked (destination)	−0.610*** (0.040)	−0.615*** (0.040)	−0.047 (0.040)	−0.610*** (0.136)	−0.019 (0.138)	−0.113 (0.126)
Landlocked (origin)	−0.170*** (0.009)	−0.169*** (0.009)	−0.057 (0.009)	−0.170*** (0.039)	−0.182*** (0.043)	−0.173*** (0.036)
Border	0.077*** (0.022)	0.076*** (0.022)	0.011 (0.022)	0.077 (0.100)	0.375*** (0.102)	0.237** (0.094)
Social and historical determinants						
Common official language	0.138*** (0.014)	0.138*** (0.014)	0.048 (0.014)	0.138* (0.077)	0.239*** (0.079)	0.233*** (0.076)
9% minority speak same language	0.266*** (0.014)	0.265*** (0.014)	0.096 (0.014)	0.266*** (0.073)	0.194*** (0.072)	0.281*** (0.071)

TABLE 2 (Continued)

Ordinary Least Square (OLS) and Generalized Estimating Equations (GEE) Regression Analysis of Inflows to 17 Selected Countries, 1950–2007. For Example, in Model 1 (M1), Log(migrants) into the 17 Developed Countries Varied in Proportion to 0.601 Times Log Population of the Destination, Where the Standard Error of the Estimated Coefficient 0.601 was 0.009

	Dependent variable: Log(Migrants)					
	M1	M2	M3	M4	M5	M6
	OLS	OLS	OLS (Beta)	GEE (ind)	GEE (exc)	GEE (ar1)
Colony	0.427*** (0.017)	0.427*** (0.017)	0.076 (0.017)	0.427*** (0.102)	0.475*** (0.098)	0.376*** (0.091)
Year – 1985	0.008*** (0.001)		0.088 (0.001)	0.008*** (0.003)	−0.001 (0.003)	−0.010*** (0.002)
(Year – 1985)^2	4E–04*** (2E–05)		0.058 (0.000)	4E–04*** (5E–05)	3E–04*** (4E–05)	0.001*** (7E–05)
Constant	−9.960*** (0.231)	−9.718*** (0.245)		−9.960*** (0.773)	−14.055*** (1.121)	−14.785*** (0.719)
Observations	46978	46978	46978	46978	46978	46921a
Adjusted R^2	0.635	0.636	0.635			
MSE	0.435	0.435	0.435			
AIC	94285	94251	94285			
BIC	94461	94908	94461			
Dispersion				0.435	0.537	0.469
QIC				21204	26396	22743

Notes: Standard errors in parenthesis.

MSE, Mean squared residual; AIC, Akaike's information criterion; BIC, Bayesian information criterion; QIC, quasi-likelihood information criterion; ind, independent error structure; exc: exchangeable error structure; ar1, first order autoregressive error structure.

[a]Panels having fewer than two consecutive years of observations are excluded.

*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$.

TABLE 3

Ordinary Least Square (OLS) and Generalized Estimating Equations (GEE) Regression Analysis of Outflows from 13 Selected Countries, 1960–2007

	Dependent variable: Log(Migrants)					
	M1	M2	M3	M4	M5	M6
	OLS	OLS	OLS (Beta)	GEE (ind)	GEE (exc)	GEE (ar1)
Demographic determinants						
Log population (destination)	0.372*** (0.008)	0.373*** (0.008)	0.257 (0.008)	0.372*** (0.036)	0.425*** (0.057)	0.389*** (0.032)
Log population (origin)	0.936*** (0.011)	0.948*** (0.011)	0.519 (0.011)	0.936*** (0.042)	0.740*** (0.039)	0.873*** (0.035)
Log potential support ratio (destination)	−0.052** (0.024)	−0.049** (0.024)	−0.013 (0.024)	−0.052 (0.100)	−0.591*** (0.141)	−0.065 (0.086)
Log potential support ratio (origin)	0.915*** (0.079)	0.994*** (0.080)	0.069 (0.079)	0.915*** (0.274)	0.704*** (0.210)	0.940*** (0.213)
Log infant mortality rate (destination)	−0.783*** (0.016)	−0.786*** (0.016)	−0.348 (0.016)	−0.783*** (0.063)	−0.086 (0.087)	−0.724*** (0.052)
Log infant mortality rate (origin)	0.359*** (0.054)	0.290*** (0.056)	0.076 (0.054)	0.359** (0.177)	−0.160 (0.137)	0.159 (0.117)
Log percentage of urban population (destination)	0.307*** (0.021)	0.306*** (0.021)	0.072 (0.021)	0.307*** (0.089)	0.853*** (0.133)	0.308*** (0.073)
Log percentage of urban population (origin)	2.578*** (0.077)	2.545*** (0.078)	0.133 (0.077)	2.578*** (0.277)	2.052*** (0.445)	2.805*** (0.256)
Geographic determinants						
Log distance between capitals	−0.660*** (0.012)	−0.660*** (0.012)	−0.267 (0.012)	−0.660*** (0.058)	−0.564*** (0.069)	−0.626*** (0.053)
Log land area (destination)	0.146*** (0.007)	0.146*** (0.007)	0.122 (0.007)	0.146*** (0.031)	0.055 (0.040)	0.129*** (0.028)
Log land area (origin)	0.030*** (0.009)	0.025*** (0.009)	0.016 (0.009)	0.030 (0.036)	0.150*** (0.039)	0.074** (0.033)
Landlocked (destination)	−0.086*** (0.011)	−0.085*** (0.011)	−0.029 (0.011)	−0.086* (0.044)	−0.120** (0.050)	−0.102** (0.041)
Landlocked (origin)	−1.043*** (0.038)	−1.023*** (0.038)	−0.106 (0.038)	−1.043*** (0.133)	−0.692*** (0.122)	−0.843*** (0.125)
Border	0.096*** (0.024)	0.094*** (0.024)	0.016 (0.024)	0.096 (0.107)	0.431*** (0.116)	0.215** (0.105)
Social and historical determinants						
Common official language	0.346*** (0.027)	0.345*** (0.027)	0.098 (0.027)	0.346** (0.143)	0.492*** (0.149)	0.402*** (0.138)
9% minority speak same language	0.003 (0.027)	0.005 (0.027)	0.001 (0.027)	0.003 (0.134)	0.011 (0.138)	0.001 (0.129)
Colony	0.747*** (0.023)	0.746*** (0.023)	0.119 (0.023)	0.747*** (0.136)	0.860*** (0.145)	0.757*** (0.138)
Year – 1985	−0.001 (0.001)		−0.011 (0.001)	−0.001 (0.003)	−0.000 (0.003)	−0.004** (0.002)
(Year – 1985)^2	−2E–04*** (3E–05)		−0.027 (0.000)	−2E–04*** (5E–05)	4E–05 (4E–05)	−1E–04** (5E–05)

TABLE 3 (Continued)
Ordinary Least Square (OLS) and Generalized Estimating Equations (GEE) Regression Analysis of Outflows from 13 Selected Countries, 1960–2007

	Dependent variable: Log(Migrants)					
	M1	M2	M3	M4	M5	M6
	OLS	OLS	OLS (Beta)	GEE (ind)	GEE (exc)	GEE (ar1)
Constant	−12.408*** (0.258)	−12.780*** (0.270)		−12.408*** (0.950)	−11.422*** (1.091)	−13.171*** (0.777)
Observations	28082	28082	28082	28082	28082	27989a
Adjusted R^2	0.664	0.665	0.664			
MSE	0.375	0.374	0.375			
AIC	52177	52158	52177			
BIC	52342	52702	52342			
Dispersion				0.375	0.446	0.380
QIC				11241	13575	11309

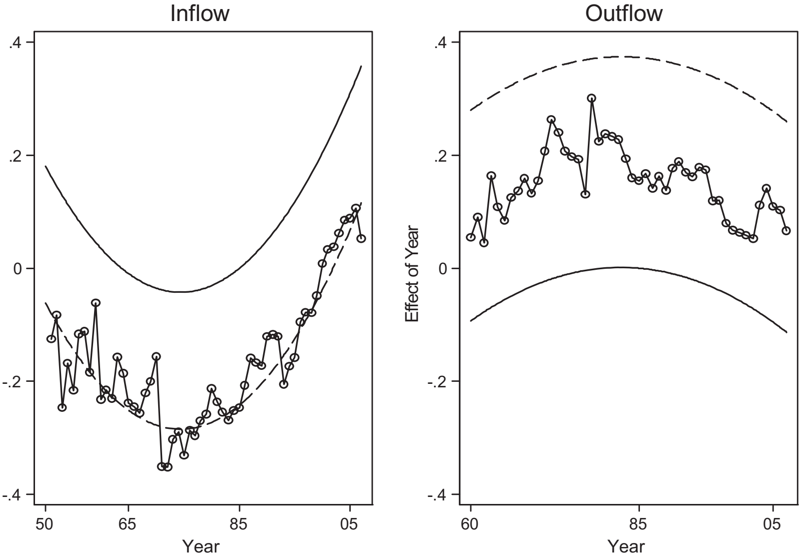
Notes: Standard errors in parenthesis.

MSE, mean square residual; AIC, Akaike's information criterion; BIC, Bayesian information criterion; QIC, quasi-likelihood information criterion; ind, independent error structure; exc, exchange error structure; ar1, first order autoregressive error structure.

[a]Panels having fewer than two consecutive years of observations are excluded.

$*p < 0.1$, $**p < 0.05$, $***p < 0.01$.

Figure I. **Effects of Year on Log Migrants Presented by Independent Dummy Variables for Each Year (Lines With Circles) and by a Continuous Quadratic Function of Year (Solid Lines) for Inflow and Outflow Models.**



Note: Coefficients for the quadratic function come from M1 in Tables 2 and 3. Dashed line adjusts the quadratic function for the difference between the constant terms of model 1 and model 2 in Tables 2 and 3.

of correlation and heteroscedasticity and explains the population-averaged GEE estimator, used here.

Following Wooldridge (2006), in models 2 in Tables 2 and 3, we used year dummy variables for OLS specifications to account for the possibility of a changing likelihood of international migration (as found in *e.g.,* Cohen *et al.*, 2008), conditional on all the other independent variables (Figure I).[13]

As Massey (1999) suggested, inflows to the 17 countries during the early 1970s to mid-1980s were significantly lower than those in 1950 while outflows during the early 1970s to mid-1980s were significantly higher than those in 1959. This result suggested that during the early 1970s to mid-1980s immigration to Western countries was suppressed while emigration from them was enhanced.

Although M2 with year dummy variables revealed interesting historical patterns in inflows and outflows, it was ill suited for projecting future

[13] The coefficients for all year dummy variables are presented in Table S3.

international migration as part of a population projection model because past years gave no guidance about the coefficients of future year dummy variables. All other models incorporated linear and quadratic terms in (year – 1985) as shown at the end of equation (3). Figure I compares the modeled effect of time on log migrants using year dummy coefficients in M2 (lines with small circles) and using linear and quadratic terms in (year – 1985) (solid line). The effects on log migrants were very similar in time course but the vertical location was different.

What accounts for the difference in vertical location? The estimated coefficients of M1 and M2 in Table 2 for inflows were nearly identical except for the constant term: constant (M1) = −9.960 while constant (M2) = −9.718. This difference reflected the presence of the scaling constant −1985 in the linear and quadratic terms for time in M1. When constant(M1) – constant(M2) = −0.242 was added to the solid curve (M1) in Figure I, the resulting dashed line passed through the estimated effects of the M2 year dummy variables, indicating that models M1 and M2 estimated practically coincident effects of time, conditional on all other variables. In the outflow model (Table 3), the differences in the estimated coefficients of M1 and M2 were larger and the year dummy variables varied more erratically. When constant(M1) − constant(M2) = −12.408–(−12.780) = +0.372 was added to the solid curve (M1) in Figure I, the resulting dashed line had the same temporal pattern as, but a different vertical location from, the M2 year dummy variables. For outflows, models M1 and M2 estimated somewhat different effects of time, conditional on all other independent variables, in part because of the differing relative importance of the other independent variables.

The statistical significance of the coefficient of the quadratic term (year – 1985)^2 for inflows and outflows differed from the lack of statistical significance of the coefficient of the quadratic term (year – 1985)^2 in the log-linear model of Cohen *et al.* (2008), which identified a significant increase in log migrants with time. That model did not distinguish inflows from outflows. It seems likely that the dip in inflows canceled the peak in outflows, leading to no significant curvature in log migrants.

In M1, variables that were expected to promote migration had positive coefficients while variables expected to deter migration had negative coefficients, except for IMR. For example, for both inflows and outflows, the coefficient of the log PSR of the destination was negative and significant whereas the coefficient of the log PSR of the origin was positive and significant. As expected, more working-age people as a fraction of the origin population were associated with an increased number of emigrants. More working-age people as a fraction of the destination population were associated with a decreased number of immigrants.

The coefficients of the IMR were more complex. For inflows, the coefficient of the IMR was positive for the destination and negative for the origin, while for outflows the coefficient of the IMR was negative for the destination and positive for the origin (M1 in Tables 2 and 3). This result was counterintuitive and is discussed below.

The percentages of urban population in destination and origin increased inflow and outflow significantly. But urbanization in the 17 countries was more influential than urbanization in the other countries to which migrants went or from which they came: for inflows in M1, the coefficient of log percentage urban in the destination was several times larger than the coefficient of log percentage urban in the origin, while for outflows in M1, the coefficient of log percentage urban in the destination was several times smaller than the coefficient of log percentage urban in the origin.

Among the geographic determinants, a greater distance between origin and destination decreased the predicted number of migrants, as expected from the gravity model. The coefficient of log distance was more negative for inflows (−0.819) than for outflows (−0.660), suggesting that distance posed a bigger obstacle to immigrants to these 17 countries than distance posed for emigrants from these 13 countries.

For inflows, larger land area in the destination facilitated migration while larger area in the origin hindered migration. For outflows, larger land area in both the destination and the origin increased migration significantly.

When either origin or destination was landlocked, inflows and outflows were reduced. For inflows to the 17 countries, a landlocked destination reduced inflows much more than a landlocked origin. For outflows from the 13 countries, a landlocked origin reduced outflows much more than a landlocked destination. Thus, whether one of the 17 countries was landlocked influenced inflows and outflows much more than whether the other country was landlocked. Among the 17 countries, only Hungary was landlocked, and Hungary differed from the other 16 countries in other respects as well. It remains to be seen whether these results remain true for a larger set of landlocked Western countries.

Sharing a border increased migration in both directions.

All coefficients of the social determinants were positive. All were significant except for the presence of ethnic minorities speaking a common language. Having a colonial link increased inflow about 2.7 times ($10^{0.427} = 2.67$) and increased outflow more than twice as much ($5.58 = 10^{0.747}$).

The directions of association (signs of coefficients) in the outflow M1 (Table 3) were generally but not always consistent with those in the inflow M1 (Table 2). Population size in the origin and the destination were positively associated with both inflow and outflow. Also, young age structure (high PSR) of the destination country decreased outflows by about 11% [that is, $100 \times (1–10^{-0.052})$] whereas young age structure of the origin country increased the outflows by a factor of 8.22 (that is, $10^{0.915}$). Notable differences between the outflow model and the inflow model were noted above.

To compare how much one standard deviation of change in each independent variable in the model influenced the dependent variable log migrants, we replaced each independent variable by a standardized variable with a mean zero and standard deviation one and we computed the regression coefficients, which are called beta coefficients (M3 in Tables 2 and 3).

For inflows (Table 2), only six of the beta coefficients in M3 had values that, when rounded to the nearest 0.1, exceeded 0.2 or were less than −0.2. These most positive or most negative beta coefficients identified the independent variables where a one standard deviation change had the greatest influence on log migrants. Four of these independent variables were demographic: log population of origin and destination and log IMR of origin and destination. Two of these independent variables were geographic: log distance between capitals and log land area of the destination. None of the social and historical determinants was as important as these six variables. Of these six, the three most important variables were, in decreasing order of importance (measured by the absolute value of the beta coefficient), log population of the origin, log population of the destination, and log distance between capitals, precisely the three variables identified in the gravity model.

For outflows (Table 3), only four of the beta coefficients in M3 had values that, when rounded to the nearest 0.1, exceeded 0.2 or were less than −0.2. Three of these independent variables were demographic: log population of origin and destination and log IMR of destination, and one of these independent variables was geographic: log distance between capitals. Thus, all four of these most important independent variables for outflows were among the six most important independent variables for inflows. (The two important independent variables for inflows that were not among the independent variables important for outflows were the log IMR of the origin and the log land area of the destination.)

The coefficients from inflow and outflow data largely conformed qualitatively to what existing theories suggested, but gave these theories quantitative specificity. However, the signs of the coefficients of log IMR in the inflow model were counterintuitive. They suggested that a higher IMR in the destination greatly increased inflows and a higher IMR in the origin decreased emigration from that origin to one of the 17 countries. The statistical significance of these coefficients may be due to mistakenly small standard errors resulting from serial correlation or autocorrelation. In the presence of serial correlation, OLS is not the best linear unbiased estimator and the usual OLS standard errors and test statistics are not valid (Wooldridge, 2006). We tested autocorrelation by following Drukker (2003).[14] Rejecting the null hypothesis that there was no autocorrelation, the test statistics were 623.027 ($p < 0.00005$) for inflow and 246.732 ($p < 0.00005$) for outflow. Thus, there was a significant autocorrelation within panels in both inflow and outflow models.

Following Cui (2007), QIC values were used to select among alternative models of correlation structure within panels. In both inflow and outflow models, the assumption of independence had the smallest QIC values and, therefore, was chosen as the preferred working correlation structure within panels, notwithstanding the significant autocorrelation within panels in both inflow and outflow models (reported in the previous paragraph). The second best option was autoregressive-1 [AR(1)] correlation rather than exchangeable correlation, which was sometimes selected in the international migration literature using GEE (*i.e.,* Neumayer, 2005; Pedersen, Pytlikova, and Smith, 2008). Based on this result, we identified the most parsimonious subset of covariates using QIC.[15] None of the models we considered accounts for autocorrelation between panels.

[14] We used xtserial routine in Stata (version 10.1).

[15] The first half under inflow of Table S4 in the supporting information presents QIC values with various correlation structures and the second half under inflow in the table indicates the most parsimonious model specification. Outflow of Table S4 follows the same order as inflows.

Models 4 through 6 in Tables 2 and 3 report the estimated coefficients resulting from GEE estimation for inflow and outflow, respectively, specifying independence, exchangeable, and AR(1) as the correlation structure within panels, including demographic, geographic and social independent variables. Both the dispersion and the QIC statistics were smallest for the GEE with independence, which yields estimates of the coefficients exactly the same as the estimates of the corresponding OLS models for inflow and outflow (Hardin and Hilbe, 2003). However, the standard errors in GEE (M4 in Tables 2 and 3) differ from OLS standard errors in that GEE uses semi-robust standard errors, a modified sandwich estimate of variance. The semi-robust standard errors tend to be greater than naïve standard errors, making it more difficult to reach conventional statistical significance given the same estimated coefficients. More important, semi-robust standard errors are robust to misspecification of the assumed correlation structure (Hardin and Hilbe, 2003:94). The dispersion of M4 in Tables 2 and 3 equaled the mean squared error of the OLS M1 because the predictive accuracy (difference between observed and predicted values) of M1 and M4 is identical, given that they had identical coefficients.

Models M5 and M6 for inflow and outflow with exchangeable and autoregressive correlation structure yielded coefficient values and signs that mostly do not differ substantially from those estimated assuming independence. In Table 2, the reversal of the signs of log IMR of origin and log IMR of destination between M4 (with independent residuals) and M5 (with exchangeable residuals) carries little meaning as the dispersion and QIC show that the assumption of exchangeable residuals yields a much worse description of the variation in log migrants. Similar remarks apply to the reversal of the signs of log IMR of destination between M4 (with independent residuals) and M6 [with AR(1) residuals] for inflows (Table 2) and to the reversal of the signs of log IMR of origin between M4, M5 and M6 for outflows (Table 3). These results suggest that our models are robust against different specifications and correlation structures, within this limited exploration.[16]

[16] Theoretically, there might be a trivial joint endogeneity between the dependent variable log migrants and two of the independent variables in year *t*, *i.e.,* the PSR and IMR. However, the GEE corrections for heteroscedastic and correlated errors handle any slight empirical occurrences of this possibility quite well. Table S5 in the supporting information presents the results of sensitivity analysis by excluding outliers based on DFITS statistics. Our models were robust with respect to outliers.

Quasi-likelihood information criterion values suggested that the most parsimonious specification for inflow, given the independence correlation structure, excluded PSR of origin, sharing a border, having a common official language, and the land area of the origin. The most parsimonious outflow model excluded the presence of an ethnic minority speaking the same language, year, PSR in the destination, land area of the origin, sharing a border, and the destination being landlocked.

LIMITATIONS

This study investigated determinants of international migration flows on the basis of a large panel-data set and identified differences between inflows and outflows. Caution should be exercised when interpreting the results.

The primary objective of the study was to develop a model of international migration that could be a useful component of a demographic projection model. Therefore, we selected explanatory variables whose future uncertainty was no greater than that of other demographic variables normally found in a demographic projection. We ignored the effects of policy changes. States and governments influence migration via their laws and regulations (Greenwood and McDowell, 1999; Vogler and Rotte, 2000), and several past empirical studies attempted to incorporate some form of policy measures (Mayda, 2005). However, data on this subject are sparse,[17] and predictive models of policy do not seem to be available.

Second, the present analysis is constrained by data availability: only 17 nations are considered for inflows and only 13 countries for outflows. No migration data for this study came from countries in the global south. The dynamics of migration between South Africa and Brazil, for instance, may differ significantly from the dynamics of the inflows and outflows described here. While migrations to and between developing countries may grow, the developed countries absorbed the vast majority (33 million out of 36 million) of all the increases in stocks of international migrants between 1990 and 2005 while migrant stocks in developing countries grew slowly during the same period (United Nations, 2009a).

[17] We considered the United Nations population-policy data, but the policy measures are available only at decades' mid-point (1975, 1985, and 1995). These measures are too far apart to use in an annual migration-change model since we cannot ascertain whether there were major policy changes during the decades.

Consequently, the concentration of the stock of international migrants in the more developed region increased. In 2005, about 60% of all international migrants in the world lived in the more developed regions: 23.3% in North America, 33.6% in Europe, and 2.6% in Oceania. Only 3.5% of international migrants lived in Latin America and the Caribbean region (United Nations, 2009a).[18] Therefore, our model and estimates apply to more than half of the world population despite the small sample size and its focus on developed countries. It would be highly desirable to develop a similar model for south–south migration.

Third, we focused on legal migration. Although United Nations does not provide information on illegal or unauthorized migrants, illegal migration may be large and heterogeneous in size across countries. The data to overcome this limitation do not exist although there are some indirect estimation techniques for illegal immigrant flows or stocks (*e.g.*, Jandl, 2004). Presumably illegal immigrants would be influenced by the determinants in our models but the dynamics of illegal flows is beyond the scope of this study.

Fourth, each country has its own definitions regarding international migrants. For example, Denmark considers a person who holds a residence permit or a work permit for at least 3 months to be a migrant whereas Finland defines a migrant as a person who has a residence permit and who intends to stay there for at least 1 year. The United States and Canada use the place of birth to classify migrants whereas European countries use previous residence or citizenship (Cohen *et al.*, 2008). Given the wide variations in defining migration and migrants, the numbers of migrants reported by the United Nations may include very different groups of people. Although we used the best available data, future research must take these problems into account to get more reliable estimates, and national statistical systems need to be harmonized to generate more comparable data (Poulain *et al.*, 2006). Internationally harmonized time-series estimates of migrant stocks by origin and destination are not presently available so migrant stocks are not considered in this analysis.

Fifth, though there is serial autocorrelation of residuals within panels, the QIC criterion demonstrated that it is better to assume independence within panels than to assume the alternative correlation structures

[18] These estimates of the number of international migrants at the global and regional levels are drawn from censuses of foreign-born or non-national population in each country (United Nations, 2009a).

such as autoregressive and exchangeable. However, our method of model fitting (GEE) does not deal with serial correlation between panels. Determining the extent of between-panel correlation and incorporating any such correlation in the modeling approach is a challenge for future work. Another challenge for the future is to model possible lagged effects on migration in the current of values of migration or independent variables in prior years.

CONCLUSION

This study examined determinants of international immigration to 17 wealthy nations – and international emigration from 13 of those 17 wealthy nations – between 1950 and 2007 with a panel-data approach. This study used only demographic, geographic, and social independent variables that are less time-sensitive and less uncertain than economic factors. This feature was important because the aim of the study was to build models suitable for predicting future international migration as a component of demographic projections. The overall results were consistent with, amplify, and quantify existing migration theories.

We employed panel-data analysis to correct for heteroscedasticity and autocorrelation within panels, the major threats to pooled OLS estimates, by modeling the correlations within panels across time. Although the results were mostly consistent across different models, some methods required large computing resources and time. Hence, we proposed a more efficient way to estimate by using GEE, an extension of generalized linear models (GLM) for panel data. To our knowledge, this is the first study of international migration using GEE to select among alternative models using QIC. The results suggested that independence of residuals within panels best fitted the inflow and outflow data. We obtained estimates broadly consistent with an independent correlation structure even after correcting for autocorrelations within panels. This study, therefore, confirmed and extended Cohen *et al.* (2008)'s suggestion that international migration can be effectively estimated by using time-invariant covariates and GLM methods. While the use of OLS gives the same point estimates of regression coefficients as GEE, the confidence intervals of the coefficients are smaller in OLS estimates than in GEE estimates.

The models identified the independent variables that were the most important predictors of log migrants. These variables, when standardized to have mean 0 and standard deviation 1, had coefficients that rounded to a value greater than 0.2 or less than −0.2. As predictors of outflows from the 13 countries, the four most important independent variables were demographic: log population of origin and destination, log IMR of destination, and log distance between capitals. The six most important independent variables for inflows to the 17 selected countries were the four variables above plus the log IMR of the origin and the log land area of the destination. Relative to the pure gravity model, the additional important predictor variables of international migration were the log IMR of the origin and destination and the land area of the destination. None of the social and historical determinants appeared among the most important predictors, and neither did calendar year in linear or quadratic form, although these independent variables had coefficients that differed significantly from zero and contributed materially to the goodness of fit of the final models.

According to M1 in Table 2, the number of immigrants to one of the 17 countries in a given year was proportional to the population of the destination raised to the power 0.601. Consequently, holding all else constant, a doubling in the population of the destination was predicted to increase the annual number of immigrants by a factor of $1.52 = 2^{0.601}$, or 52%. Similarly, holding all else constant, a doubling in the population of the origin was predicted to increase the annual inflow by a factor of $1.66 = 2^{0.728}$, or 66%. Doubling the distance between the capitals of an origin and a destination, holding all else constant, was predicted to multiply the annual inflow by a factor of $0.57 = 2^{-0.819}$, that is, to reduce the annual inflow by 43%.

A higher PSR of the origin, which indicated a young age structure, slightly facilitated inflows whereas a higher PSR in the destination countries substantially lowered inflows (Table 2, M1 or M4). By contrast, for outflows, a higher PSR of the origin substantially facilitated outflows whereas a higher PSR in the destination countries slightly lowered outflows (Table 3, M1 or M4). The signs of the coefficients of PSR remained the same for inflows and outflows, but for inflows the PSR of the destination was relatively more influential (and negatively so), whereas for outflows the PSR of the origin was relatively more influential (and positively so). To simplify, the younger the age structure of one of the 17 countries, the lower the migratory inflow and the higher the migratory outflow, all else being equal.

Urbanization of both destinations and origins significantly increased inflows. A 1% increase in the percentage urban of a destination's population (not an increase by 1 percentage point, but an increase by 1% of the baseline percentage urban, *e.g.,* from 50% to 50.5%) was predicted to increase inflows to that destination by a factor of $1.03 = 1.01^{3.057}$, or roughly 3%. Similarly, a 1% increase in the proportion urban of an origin's population was predicted to increase inflows from that origin by a factor of 1.003, or 0.3%.

Among other geographical determinants of inflows, landlocked location mattered both for origin and destination countries. If the origin was landlocked, the inflow decreased by roughly 32%. If the destination was landlocked, then inflow was predicted to decrease by 76%.

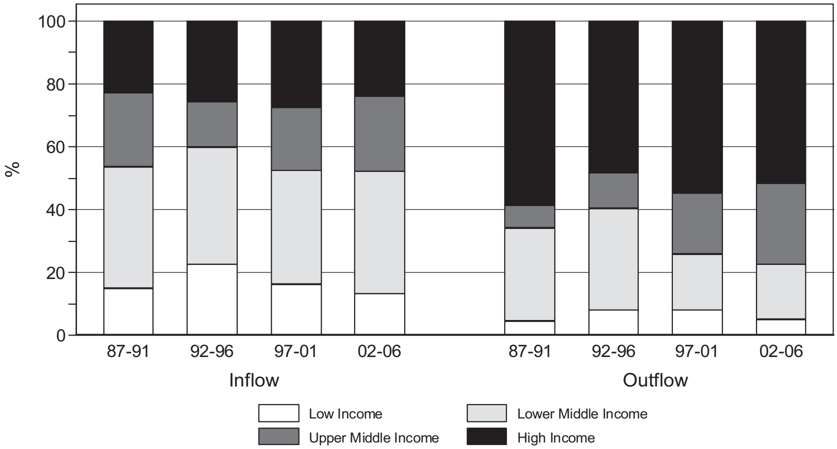
With respect to social and historical factors, inflows were larger when an origin and a destination had the same official language; and when at least 9% of minority in a host country spoke the same language as the migrants. Presence of colonial links between destination and origin increased the inflow by about 2.67 times. Having a 9% minority in the origin and destination who spoke the same language had an insignificantly positive effect on outflows.

The signs of outflow determinants differed for only a few variables from the signs of inflow determinants, according to M1 in Tables 2 and 3. Signs were reversed between the inflow model and the outflow model for these variables only: log IMR of the destination and of the origin, and log land area of the origin. The coefficient of year – 1985 was significantly positive for inflows and negative but not significantly different from 0 for outflows. The coefficient of $(\text{year} - 1985)^2$ was significantly positive for inflows and significantly negative for outflows.

Economic theories of international migration typically postulate that differences in economic factors such as income and employment drive international migration. If IMR can represent the general economic situation in a country and can be projected using demographic methods more accurately than economic factors such as income and employment, then we might be able to project international migration more reliably by incorporating IMR as a predictor.

When the annual inflows were classified by the income class of the origin (Figure II, left), about 40% of immigrants to rich countries came from "lower middle-income" countries while about 15–20% of immigrants came from the low-income countries. This finding is consistent with the theory of the "migration hump" (Olesen, 2002), which postulates that development and migration exhibit an inverted U-shape pattern over time. When annual outflows from the 13 selected affluent nations

Figure II. **Income Classifications of Origins of Inflow to the 17 Selected Countries and Destinations of Outflow from 13 of the 17 Selected Countries.**



Sources: Historical income (GNI per capita in US$) classifications 1987–2006 came from the World Bank and migrant flows came from United Nations (2009b).

were classified by the income class of the destination in that year (Figure II, right), about 50–60% of the migrants were heading to other wealthy nations while only 5% were heading to low-income countries. The outflows by destination-development levels exhibited a pronounced J shape. In sum, if countries with higher IMR are more likely to be economically less developed countries, then the significantly negative coefficient of the origin's IMR in the inflow model may indicate that people in countries with higher IMR may lack resources to migrate to wealthy nations. Similarly, the significantly positive coefficient of the destination's IMR in the inflow model may indicate that destination countries with the lowest IMR (presumably highly prosperous) are less likely to be receptive to immigrants than countries with higher IMR. These uncertain interpretations are *post hoc* and are offered to stimulate further empirical investigation.

Our description of inflows and outflows by separately estimated models (Tables 2 and 3) is equivalent to a unified model in which every independent variable of the original separate models interacts with an indicator variable that specifies whether each datum and each estimate are for inflows or outflows. Eventually such unified models would incorporate independent variables that describe why some flows are classified as inflows and other flows as outflows. Such a unified gravity-based model should make it possible to extrapolate from data on north–north, north–south, and south–north migration to south–south migration.

Remaining tasks are to test whether the extended gravity models developed here generate estimates and projections of net migration considered plausible by statistical agencies and users; and, if so, to embed these models into detailed deterministic and probabilistic cohort-component demographic projections. One reward for that difficult work is that use of migrant flows (not net migration) assures that the sum of net migration over all countries is zero, as it must be in the absence of interplanetary travel. Another reward is that the positive coefficients of log population of origin and of log population of destination assure that, all else being equal, as the population of an origin or destination declines toward zero, migration from or to that country also declines.

REFERENCES

Bijak, J.
2006 Forecasting International Migration: Selected Theories, Models, and Methods. Working Paper 4/2006. Central European Forum for Migration Research (CEFMR): Warsaw, Poland.

Breusch, T. S., and A. R. Pagan
1979 "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica* 47(5):1287–1294.

Clark, X., T. J. Hatton, and J. G. Williamson
2007 "Explaining U.S. immigration, 1971–1998." *The Review of Economics and Statistics* 89(2):359–373.

Cohen, J. E., M. Roig, D. C. Reuman, and C. GoGwilt
2008 "International Migration Beyond Gravity: A Statistical Model for Use in Population Projections." *Proceedings of National Academy of Science* 105(40):15269–15274.

Coleman, D.
2006 "Immigration and Ethnic Change in Low-Fertility Countries: A Third Demographic Transition." *Population and Development Review* 32(3):401–446.

Cook, R. D. and S. Weisberg
1983 "Diagnostics for Heteroscedasticity in Regression." *Biometrika* 70(1):1–10.

Cui, J.
2007 "QIC Program and Model Selection in GEE." *The Stata Journal* 7(2):209–220.

Drukker, D. M.
2003 "Testing for Serial Correlation in Linear Panel-Data Models." *The Stata Journal* 3(2):168–177.

Duncan, O. T., R. P. Cuzzort, and B. Duncan
1963 *Statistical Geography: Problems in Analyzing Areal Data.* New York: Free Press.

Fertig, M., and C. M. Schmidt
2000 "Aggregate-Level Migration Studies as a Tool for Forecasting Future Migration Streams." IZA Discussion Paper No. 183.

Frees, E. W.
2004 *Longitudinal and Panel Data: Analysis and Applications for the Social Sciences.* Cambridge: Cambridge University Press.

Frey, W. H.
1996 "Immigration, Domestic Migration, and Demographic Balkanization in America: New Evidence for the 1990s." *Population and Development Review* 22(4):741–763.

Glick, R., and A. K. Rose
2002 "Does Currency Union Affect Trade? The Time Series Evidence." *European Economic Review* 46:1125–1151.

Greenwood, M. J.
1975 "Research on Internal Migration in the United States: A Survey." *Journal of Economic Literature* 13(2):397–433.

———, and J. M. McDowell
1982 "The Supply of Immigrants to the United States." In *The Gateway: U.S. Immigration Policies and Issues.* Ed. B. R. Chiswick. Washington, DC: American Enterprise Institute, Pp. 54–85.

———, and ———
1991 "Differential Economic Opportunity, Transferability of Skills, and Immigration to the United States and Canada." *The Review of Economics and Statistics* 73(4):612–623.

———, and ———
1999 *Legal U.S. Immigration: Influences on Gender, Age, and Skill Composition.* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Hardin, J. W., and J. M. Hilbe
2003 *Generalized Estimating Equations.* Boca Raton, FL: Chapman & Hall/CRC.

Hatton, T. J.
2005 "Explaining Trends in UK Immigration." *Journal of Population Economics* 18:719–740.

Howe, N., and R. Jackson
2006 Long-term Immigration Projection Methods: Current Practice and How to Improve It. Center for Retirement Research, working paper 2006-3. Washington D.C.: Center for Strategic and International Studies (CSIS).

Isard, W.
1998 "Gravity and Spatial Interaction Models." In *Methods of Interregional and Regional Analysis.* Ed. W. Isard *et al.* Brookfield, Vermont: Ashgate Publishing Company. Pp. 243–279.

Isserman, A. M., D. A. Plane, P. A. Rogerson, and P. M. Beaumont
1985 "Forecasting Interstate Migration with Limited Data: A Demographic-Economic Approach." *Journal of the American Statistical Association* 80(330):277–285.

Jandl, M.
2004 "The Estimation of Illegal Migration in Europe." *Migration Studies XLI* 153:141–155.

Karemera, D., V. I. Oguledo, and B. Davis

2000 "A Gravity Model Analysis of International Migration to North America." *Applied Economics* 32:1745–1755.

Land, K. C., and G. Deane

1992 "On the Large-Sample Estimation of Regression Models with Spatial- Or Network Effects." *Sociological Methodology* 22:221–248.

Lee, L.

2007 "GMM and 2SLS Estimation of Mixed Regressive, Spatial Autoregressive Models." *Journal of Econometrics* 137:489–514.

Martin, P.

2003 "Economic Integration and Migration: The Mexico-US Case." UNU/WIDER. Discussion Paper No. 2003/35. Helsinki, Finland: United Nations University/ World Institute for Development and Economic Research.

Massey, D. S.

1999 "International Migration at the Dawn of the Twenty-First Century: The Role of the State." *Population and Development Review* 25(2):303–322.

———

2006 Building a Comprehensive Model of International Migration. Boston, MA: Center for Retirement Research at Boston College. Center for Strategic and International Studies Project on Long-term Immigration Projections. Annex to working paper wp 2006–3.

——— *et al.*

1993 "Theories of International Migration: A Review and Appraisal." *Population and Development Review* 19(3):431–466.

Mayda, A. M.

2005 "International Migration: A Panel Data Analysis of Economic and Non-Economic Determinants." IZA Discussion Paper No. 1590.

Mayer, T., and X. Zignago

2006 "Notes on CEPII's Distance Measures." <http://www.cepii.fr/distance/noticedist_en.pdf> (accessed 11 June, 2008).

Mitchell, J., and N. Pain

2003 *The Determinants of International Migration into the UK: A Panel based Modelling Approach.* London: National Institute of Economic and Social Research.

Neumayer, E.

2005 "Bogus Refugees? The Determinants of Asylum Migration to Western Europe." *International Studies Quarterly* 49:389–409.

Olesen, H.

2002 "Migration, Return, and Development: An Institutional Perspective." *International Migration* 40(5):125–149.

Pedersen, P. J., M. Pytlikova, and N. Smith

2008 "Selection and Network Effects - Migration Flows into OECD Countries 1990–2000." *European Economic Review* 52:1160–1186.

Poulain, M. *et al.*

2006 *THESIM – Towards Harmonised European Statistics on International Migration.* Louvain-la-Neuve, Presses Universitaires de Louvain.

Reidpath, D. D., and P. Allotey
2003 "Infant Mortality Rate as an Indicator of Population Health." *Journal of Epidemiology of Community Health* 57(5):344–346.

Rogers, A.
2006 *Demographic Modeling of the Geography of Migration and Population: A Multiregional Perspective*. Population Program Working Paper POP2007-02. Boulder, CO: University of Colorado.

United Nations
2006a *World Population Prospects: The 2006 Revision*. Population Database. <http://esa.un.org/unpp/> (accessed 10 June, 2008).

——
2006b *International Migration Report 2006*. New York: United Nations.

——
2006c *International Migration 2006 (Wall Chart)*. New York: United Nations.

——
2009a *International Migration Report 2006: A Global Assessment*. New York: United Nations.

——
2009b *International Migration Flows To and From Selected Countries: The 2008 Revision*. New York: United Nations.

Vogler, M., and R. Rotte
2000 "The Effects of Development on Migration: Theoretical Issues and New Empirical Evidence." *Journal of Population Economics* 13:485–508.

Wooldridge, J. M.
2006 *Introductory Econometrics: A Modern Approach*, 3rd edn. Cincinnati, OH: South Western College Publishing.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article.

Table S1. Percentage Distribution of Inflows by Region of Origin, for Each Destination and Period

Table S2. Percentage Distribution of Outflows by Region of Origin, for Each Destination and Period

Table S3. Coefficients for Year Dummy Variables in Models 2 in Tables 2 and 3

Table S4. Quasi-Likelihood Information Criterion Statistics for Model Selection under Normal Distribution of the Inflow and Outflow Data

Table S5. Robustness Checks: Inflow and Outflow Models Estimated After Exclusion of Outliers Based on DFITS Statistics

Table S6. Variables in the Supplementary Data Sets Inflow.csv and Outflow.csv

Figure S1. Total Annual Inflows by Destination

Figure S2. Total Annual Outflows by Origin

Figure S3. Regression Diagnostics for Inflow and Outflow Models

The raw data are given on-line in inflow.csv and outflow.csv, two files which are in plain text with comma-separated variables.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

APPENDIX

Determinants of International Migration Flows to and from Industrialized Countries: A Panel Data Approach Beyond Gravity

Keuntae Kim and Joel E. Cohen

This appendix describes methodological details and presents 6 supplementary tables and 3 supplementary figures.

Contents of Appendix

Methods of Estimation in the Possible Presence of Correlation and Heteroscedasticity

Panel data, also called longitudinal data, are a marriage of cross-sectional and time series and follow the same units of observation (e.g., individuals, families) across time (Wooldridge, 2006; Frees, 2004). In the present paper, a panel is defined as a pair consisting of an origin country and a destination country. The pooled ordinary least squares (OLS) method is the most basic estimation technique for panel-data sets. It does not address the panel structure of the data and treats observations as being serially uncorrelated for a given origin-destination pair. The pooled OLS assumes homoscedastic errors across origin-destination pairs and time periods.

A generalized method of moments (GMM) estimator allows for the possibility of heteroscedasticity and correlation of residuals. This estimator requires the use of instrumental variables in the form of lags. However, the GMM estimator is poorly suited to unbalanced panels, that is, different numbers of observation for different origin-destination pairs (Neumayer, 2005; Pedersen et al., 2008).

The population-averaged generalized estimating equation (GEE) estimator permits time-independent variables and allows the user to specify panels' within-group correlation structure (Liang and Zeger, 1986; Pedersen et al., 2008; Hardin and Hilbe, 2003). The GEE is equivalent to the random-effect estimator when the distribution of the dependent variable is Gaussian with an identity-link function and when the working correlation structure is exchangeable, but GEE allows the user to adjust standard errors for clustering (Hardin and Hilbe, 2003; Horton and Lipsitz, 1999; Pedersen et al., 2008; Neumayer, 2005).

An advantage of GEE is the gain in efficiency in parameter estimation that results from including a hypothesized structure of the within-panel correlation (Hardin and Hilbe, 2003). Hypothetical correlation structures include independence, exchangeable, autoregressive, stationary, non-stationary, and unstructured (Cui, 2007; Hardin and Hilbe, 2003). The independence structure is equivalent to OLS models because it assumes observations within a panel are independent. The exchangeable correlation structure hypothesizes that observations within a panel have some common correlation. When the variance is Gaussian with an identity link, the exchangeable correlation GEE estimates are equal to random effects linear regression (Hardin and Hilbe, 2003). Autoregressive structure assumes a time-dependence for the association in the repeated observations within the panels. Stationary structure hypothesizes that correlation exists only for small number of time units, and the non-stationary is the same as stationary except that it does not assume constant correlations down the diagonals. Finally, unstructured correlation imposes no assumptions on the correlation matrix.

GEE models do not provide conventional indices for comparing model fits such as log likelihood, AIC, or BIC because GEE is based, not on maximum likelihood theory (Pan, 2001; Cui, 2007), but rather on quasilikelihood theory. Pan (2001) proposed a “quasilikelihood under the independence model information criterion” (QIC), which is an extension of AIC for GEE models. Like AIC, the smaller the QIC value, the better the model fit. As Cui (2007) acknowledged, there are no rules of thumb similar to those Raftery (1995) suggested. Cui (2007) and Hardin and Hilbe (2003) proposed two steps for using the QIC measure for model selection. First, use the QIC measure to choose among competing correlation structures within panels. Second, given the best fitting correlation structure, select subsets of covariates by using QIC values.

Comparing the accuracy of economic and demographic forecasts

To compare the accuracy of demographic and economic projections, we reviewed studies by Bongaarts and Bulatao (2000), Groemling (2002), and Congressional Budget Office (2005). The two principal measures used in the comprehensive study of the accuracy of demographic projections by Bongaarts and Bulatao (2000) were proportional error and absolute proportional error. Groemling (2002) studied only projections of GDP in Germany from 1995 to 2001. None of his four measures of forecast accuracy (listed on his p. 247) was the same as either measure used by Bongaarts and Bulatao. The Congressional Budget Office (2005) comparison of the accuracy of its economic forecasts with those made by the Blue Chip consensus (an average of private-sector forecasters) and by the incumbent Administrations from 1976 through 2003 used two measures, mean absolute error and the root mean squared error, which also differed from the two measures used by Bongaarts and Bulatao (2000). We have not thus far found a demographic study and an economic study that used the same measure of forecast accuracy.

The two forecast time intervals used by the CBO were two years and five years, whereas the forecast time intervals used by Bongaarts and Bulatao (2000) were from 5 to 30 years into the future. These choices of forecast time intervals suggest the limitations of economic forecasting for long-term demographic projections.

Comparing the explanatory power of demographic, geographic, and social influences

To compare the explanatory power of demographic, geographic, and social influences on international migration, log migrants were regressed on the independent variables by adding blocks of demographic determinants, geographic determinants, and social determinants. Explanatory power measured by the adjusted R^2 increased progressively as blocks of geographical and social variables were added to demographic variables.

For inflows, the adjusted R^2 suggested that 54.1% of the variation in log migrants was attributable to the demographic variables alone, even after taking account of the number of observations. The addition of geographic variables significantly improved the fit, as the adjusted R^2 increased to 0.609, and both AIC and BIC suggested that the latter model fitted substantially better than the former. Adding social and historical independent variables to demographic and geographic variables further improved fit substantially.

For outflow from the 13 countries, demographic determinants alone explained 56.8% of the variation in the logged migrants. The model fit improved substantially after adding geographic determinants and social and historical determinants. The full model with demographic, geographic, and social-historical independent variables explained 66.4% of the variation in log migrants.

Outflow models performed (Table 3) better than inflow models (Table 2) with the same set of predictors. For these 17 countries, emigration (outflow) was more predictable given the determinants we measured than immigration (inflow). However, there were little more than half as many outflow observations as inflow observations and the outflow observations covered a time interval shorter by a decade. Cohen et al. (2008) found that the shorter the interval of observation, the higher the fraction of variation in log migrants that could be explained by a log-linear model. It is not clear that emigrants were more predictable than immigrants, as the available observations of them were more restricted.

Robustness Checks

To investigate how robust our estimations were given the heterogeneity of our sample, we conducted several sensitivity analyses. First, we excluded data from Croatia and Hungary

(Table S5, model (1)). Those Eastern European countries were less wealthy than the remaining 17, and they might have experienced very different migration history from Western Europe and North America. All but one of the coefficients retained the same direction and magnitude compared with those of M1 in Table 2 and Table 3. The exceptional coefficient for Border fell by half in the inflow model and by two-thirds in the outflow model after excluding Hungary and Croatia. When we excluded the United Kingdom because its migration data were rounded to the nearest 100, in addition to excluding Croatia and Hungary, the results were not materially affected (Table S5, model (2)).Second, to exclude outliers, we used DFITS (Table S5, models (3)-(6), using data from all countries). DFITS identifies observations that have high leverage and high residuals (Baum, 2006). For each observation, DFITS is the difference between the fitted values calculated with and without that observation. If the absolute value of DFITS was greater than $2\sqrt{k/n}$, where k was the number of predictors in the model and n was the number of observations, we excluded that observation. For the inflow data, this criterion resulted in the exclusion of 2534 data points for models 1, 4, and 5. For model 6, 2603 data points were excluded because panels having fewer than 2 observations were also excluded. For the outflow data, the numbers of observations excluded were 1594 (M1, M4, M5) and 1691 (M6), respectively. In both inflow and outflow data, the fraction of data considered outliers by this criterion was approximately 6%. Then we estimated models 1, 4, 5, and 6 for inflow and outflow (Table S5). No signs of coefficients changed and there were no very large changes in values of the coefficients. Our models were robust with respect to outliers.

Table S1. Percentage distribution of inflows by region of origin, for each destination and period

		Regions of Origin					
Destination	Period	Africa	Latin America & the Caribbean	Northern America	Europe	Oceania	Asia
Australia	60-64	1.9	0.0	1.8	90.9	1.6	3.8
	65-69	1.5	0.0	2.7	84.6	3.4	7.7
	70-74	2.4	1.8	5.8	70.8	4.4	14.8
	75-79	4.5	4.6	3.4	45.5	15.9	26.2
	80-84	4.1	1.8	2.9	44.1	12.7	34.5
	85-89	5.2	3.1	2.6	29.3	17.1	42.8
	90-94	4.8	2.4	2.9	23.1	12.7	54.0
	95-99	8.3	0.9	2.5	22.2	23.9	42.1
	00-04	12.0	1.0	2.0	19.7	20.5	44.7
Belgium	60-64	2.9	0.0	5.4	80.3	0.0	11.4
	65-69	2.0	0.0	9.4	81.9	0.0	6.8
	70-74	2.0	0.0	11.1	79.1	0.0	7.9
	75-79	11.8	0.8	10.3	67.3	0.0	9.8
	80-84	15.5	1.5	10.6	64.5	0.0	7.8
	85-89	13.5	1.5	10.6	66.5	0.0	7.8
	90-94	13.5	1.3	7.7	65.0	0.3	12.3
	95-99	13.8	1.5	7.3	64.5	0.4	12.5
	00-04	19.2	1.8	5.2	58.8	0.3	14.6
Canada	60-64	3.2	2.1	11.3	76.4	2.0	5.0
	65-69	3.0	0.5	10.1	72.6	2.4	11.4
	70-74	3.1	6.6	16.3	47.1	2.0	24.8
	75-79	5.2	18.3	9.8	33.7	1.9	31.2
	80-84	4.1	13.9	6.7	28.4	1.3	45.6
	85-89	5.9	17.1	4.5	22.7	0.8	49.0
	90-94	7.4	14.3	2.5	16.7	1.0	58.2
	95-99	8.2	9.1	2.4	16.4	0.6	63.3
	00-04	10.1	8.7	2.5	15.3	0.7	62.6
Denmark	80-84	4.5	2.5	12.5	56.0	1.4	23.2
	85-89	4.8	1.9	9.3	39.0	1.3	43.6

	90-94	9.4	2.9	8.5	44.0	1.7	33.6
	95-99	9.5	2.3	5.6	58.6	1.2	22.7
	00-04	6.2	2.6	6.5	54.8	1.4	28.5
Finland	80-84	1.1	0.6	3.5	91.0	1.1	2.7
	85-89	2.3	1.0	5.0	83.9	1.2	6.7
	90-94	8.3	1.1	4.0	71.6	0.8	14.1
	95-99	4.5	1.1	4.5	75.6	0.9	13.3
	00-04	4.4	1.6	5.1	72.0	0.9	15.9
France	95-99	35.8	3.2	4.8	40.9	0.3	15.0
	00-04	47.8	4.0	3.5	26.3	0.3	18.2
Germany	65-69	1.2	0.7	3.7	76.7	0.3	17.4
	70-74	2.0	0.8	3.5	60.5	0.4	32.8
	75-79	2.9	1.3	4.3	50.1	0.4	41.0
	80-84	4.1	1.7	4.6	55.9	0.4	33.3
	85-89	4.1	1.5	4.0	61.6	0.3	28.5
	90-94	5.3	1.3	2.9	67.6	0.3	22.6
	95-99	5.2	2.1	3.1	62.6	0.3	26.7
	00-04	5.3	2.7	3.2	62.1	0.4	26.2
Iceland	85-89	1.0	0.5	12.0	81.6	1.9	3.0
	90-94	1.3	0.8	10.5	80.9	1.7	4.7
	95-99	1.6	1.4	8.1	82.1	0.8	6.1
	00-04	1.6	1.5	7.6	79.5	0.7	9.0
Italy	80-84	1.1	7.7	13.2	75.3	2.0	0.7
	85-89	6.4	14.1	11.3	65.6	2.0	0.7
	90-94	30.8	16.9	7.1	41.3	1.3	2.6
	95-99	24.4	11.1	2.8	44.8	0.4	16.4
	00-04	20.3	10.3	2.3	50.9	0.2	16.0
New Zealand	80-84	1.1	7.7	13.2	75.3	2.0	0.7
	85-89	6.4	14.1	11.3	65.6	2.0	0.7
	90-94	30.8	16.9	7.1	41.3	1.3	2.6
	95-99	24.4	11.1	2.8	44.8	0.4	16.4
	00-04	20.3	10.3	2.3	50.9	0.2	16.0
Norway	80-84	4.7	2.8	13.2	59.5	1.0	18.8

	85-89	6.1	5.6	11.6	52.0	0.8	24.0
	90-94	6.6	2.9	9.4	60.2	0.8	20.1
	95-99	5.3	1.9	9.1	64.3	1.0	18.3
	00-04	8.3	2.0	6.2	56.1	0.9	26.5
Spain	80-84	0.1	15.7	3.6	78.9	1.7	0.0
	85-89	3.0	30.4	5.9	58.6	1.6	0.5
	90-94	12.9	26.7	3.8	50.2	1.2	5.1
	95-99	16.2	31.1	3.0	46.0	0.5	3.3
	00-04	14.1	50.8	1.1	30.6	0.1	3.3
Sweden	60-64	0.7	1.0	7.5	89.1	0.5	1.2
	65-69	1.5	0.9	4.7	89.0	0.6	3.3
	70-74	2.1	2.0	5.0	83.6	1.0	6.2
	75-79	2.8	6.6	3.8	73.2	0.9	12.7
	80-84	3.6	8.7	5.2	60.5	0.9	21.1
	85-89	5.2	9.9	4.3	48.4	0.9	31.3
	90-94	8.5	4.7	3.8	54.7	0.9	27.4
	95-99	6.2	4.9	6.7	53.5	1.4	27.4
	00-04	6.0	4.3	6.0	53.3	1.3	29.1
USA	50-54	0.2	12.7	12.9	70.3	0.2	3.7
	55-59	0.5	26.7	10.6	54.6	0.3	7.3
	60-64	0.6	34.7	12.3	43.5	0.4	8.5
	65-69	1.0	42.0	7.9	33.8	0.5	14.8
	70-74	1.7	41.0	2.9	22.9	0.7	30.7
	75-79	2.1	42.0	2.6	13.6	0.8	38.9
	80-84	2.6	36.5	2.2	9.9	0.7	48.1
	85-89	3.0	41.0	2.0	9.8	0.7	43.5
	90-94	3.3	36.5	2.0	14.5	0.6	43.0
	95-99	6.0	43.5	1.6	13.7	0.6	34.6
	00-04	6.0	43.9	1.9	13.9	0.6	33.7

Notes. Croatia, Hungary, and United Kingdom inflows are excluded because migration data are available since 1990s and there are too many missing values.

Table S2. Percentage distribution of outflows by region of origin, for each destination and period

		Regions of Destination					
Origin	Period	Africa	Latin America & the Caribbean	Northern America	Europe	Oceania	Asia
Australia	60-64	1.0	0.0	5.0	84.0	6.2	3.8
	65-69	0.9	0.0	5.1	86.4	5.4	2.3
	70-74	1.0	0.2	6.7	81.1	8.7	2.3
	75-79	1.1	1.0	5.4	77.3	11.6	3.6
	80-84	1.6	1.1	4.4	55.3	32.1	5.4
	85-89	1.1	0.9	6.0	41.7	42.6	7.7
	90-94	1.0	1.6	5.4	37.4	40.2	14.4
	95-99	1.2	0.4	5.7	33.2	34.5	25.0
	00-04	1.9	0.5	4.7	26.6	28.4	38.0
Belgium	60-64	2.8	0.0	7.0	87.2	0.0	3.1
	65-69	3.6	0.0	10.3	79.1	0.0	7.0
	70-74	3.5	0.0	13.3	80.6	0.0	2.6
	75-79	7.4	0.6	13.1	74.7	0.0	4.2
	80-84	12.1	1.0	10.4	71.2	0.0	5.2
	85-89	11.9	1.0	12.9	67.6	0.0	6.6
	90-94	6.1	1.4	13.8	71.2	0.4	7.1
	95-99	3.7	1.2	13.1	73.6	0.5	7.8
	00-04	3.0	1.2	12.1	74.7	0.6	8.4
Denmark	80-84	3.8	1.4	16.3	64.1	2.1	12.2
	85-89	4.0	1.8	19.6	60.2	2.7	11.6
	90-94	3.9	2.3	15.9	61.1	3.0	13.8
	95-99	4.8	2.2	11.6	67.5	2.4	11.5
	00-04	5.5	2.0	9.9	68.7	1.9	12.0
Finland	80-84	1.0	0.6	5.2	89.0	2.2	2.0
	85-89	0.8	0.5	6.0	89.0	1.8	1.9
	90-94	1.2	0.8	7.4	85.7	1.5	3.4
	95-99	1.1	0.8	8.1	84.0	1.3	4.8
	00-04	1.7	0.9	9.5	80.1	1.5	6.3

Germany	65-69	1.2	0.7	4.4	82.0	0.5	11.2
	70-74	1.6	0.8	4.1	71.9	0.5	21.0
	75-79	2.1	1.0	4.0	60.0	0.4	32.5
	80-84	3.1	1.3	4.2	55.5	0.5	35.4
	85-89	3.8	1.4	5.9	63.7	0.5	24.7
	90-94	4.2	1.2	4.1	73.1	0.4	17.0
	95-99	4.4	1.5	4.7	70.0	0.5	19.0
	00-04	4.5	2.1	4.3	68.2	0.5	20.5
Iceland	85-89	0.5	0.3	11.4	85.2	1.9	0.7
	90-94	1.0	0.6	11.9	82.6	2.0	1.8
	95-99	0.8	0.5	9.3	86.5	1.1	1.8
	00-04	0.8	0.8	9.4	86.0	0.7	2.3
Italy	80-84	1.2	4.8	11.3	80.0	2.2	0.5
	85-89	2.0	5.2	12.1	77.7	2.6	0.4
	90-94	2.4	5.6	10.0	78.6	1.9	1.5
	95-99	6.1	10.9	9.6	67.1	1.2	5.2
	00-04	7.6	12.1	9.2	63.7	0.5	6.9
New Zealand	80-84	0.8	0.4	6.3	20.0	66.7	5.8
	85-89	0.4	0.3	5.0	22.5	67.4	4.3
	90-94	0.7	0.7	7.6	30.5	48.9	11.5
	95-99	0.6	0.7	6.4	28.7	52.1	11.5
	00-04	0.7	0.7	6.2	26.1	51.3	15.1
Norway	80-84	4.8	2.1	17.3	64.5	1.5	9.8
	85-89	3.3	1.7	12.3	75.6	1.1	6.0
	90-94	3.8	2.5	14.4	67.9	1.3	10.2
	95-99	2.2	1.5	14.0	72.3	1.7	8.4
	00-04	1.7	0.9	8.6	80.8	1.1	6.8
Sweden	60-64	1.7	1.4	12.4	82.0	1.1	1.3
	65-69	2.1	1.5	10.6	81.4	2.4	2.0
	70-74	1.8	1.3	6.4	86.2	2.2	2.2
	75-79	2.1	1.9	7.8	83.1	1.5	3.6
	80-84	1.6	2.4	7.8	82.9	1.6	3.7
	85-89	1.2	3.0	8.7	81.0	2.0	4.1

90-94	2.0	4.3	9.6	74.4	2.2	7.4
95-99	2.3	3.7	12.4	70.0	2.1	9.5
00-04	7.1	5.1	5.7	61.2	1.7	19.1

Notes. Croatia, Hungary, and United Kingdom inflows are excluded because migration data are available since 1990s and there are too many missing values; Canada, France, Spain, USA do not report outflows and are excluded.

Table S3. Coefficients for year dummy variables in models 2 in Table 2 and Table 3

	Inflow		Outflow	
Year	Coefficient	SE	Coefficient	SE
1951	-0.126	0.140	---	---
1952	-0.083	0.139	---	---
1953	-0.247*	0.133	---	---
1954	-0.168	0.133	---	---
1955	-0.216	0.132	---	---
1956	-0.116	0.132	---	---
1957	-0.112	0.131	---	---
1958	-0.184	0.130	---	---
1959	-0.062	0.119	---	---
1960	-0.232**	0.114	0.054	0.110
1961	-0.215*	0.113	0.090	0.113
1962	-0.230**	0.113	0.045	0.113
1963	-0.157	0.112	0.164	0.111
1964	-0.186*	0.111	0.108	0.107
1965	-0.239**	0.109	0.084	0.100
1966	-0.245**	0.109	0.125	0.100
1967	-0.256**	0.109	0.137	0.099
1968	-0.221**	0.109	0.159	0.100
1969	-0.201*	0.109	0.132	0.100
1970	-0.156	0.109	0.154	0.100
1971	-0.351***	0.108	0.207**	0.100
1972	-0.352***	0.108	0.263***	0.100
1973	-0.303***	0.108	0.240**	0.100
1974	-0.290***	0.107	0.207**	0.100
1975	-0.331***	0.107	0.197**	0.100
1976	-0.287***	0.107	0.192*	0.100
1977	-0.296***	0.107	0.130	0.100
1978	-0.270**	0.107	0.301***	0.098
1979	-0.258**	0.107	0.224**	0.098
1980	-0.213**	0.106	0.238**	0.096
1981	-0.237**	0.106	0.233**	0.096
1982	-0.254**	0.107	0.228**	0.096
1983	-0.269**	0.107	0.193**	0.097
1984	-0.252**	0.107	0.160*	0.097
1985	-0.247**	0.107	0.154	0.097
1986	-0.207*	0.107	0.167*	0.097
1987	-0.159	0.107	0.141	0.097
1988	-0.167	0.107	0.163*	0.097
1989	-0.173	0.107	0.138	0.098
1990	-0.121	0.107	0.177*	0.097
1991	-0.117	0.107	0.189*	0.097
1992	-0.121	0.107	0.169*	0.097
1993	-0.205*	0.108	0.162*	0.098
1994	-0.174	0.108	0.178*	0.098
1995	-0.158	0.108	0.174*	0.098

1996	-0.095		0.108		0.119		0.098	
1997	-0.078		0.108		0.120		0.099	
1998	-0.079		0.109		0.079		0.099	
1999	-0.049		0.109		0.067		0.099	
2000	0.008		0.109		0.063		0.100	
2001	0.033		0.109		0.058		0.100	
2002	0.038		0.109		0.052		0.100	
2003	0.062		0.110		0.112		0.100	
2004	0.085		0.110		0.141		0.101	
2005	0.089		0.110		0.109		0.101	
2006	0.106		0.110		0.102		0.101	
2007	0.052		0.113		0.066		0.103	
Summary statistics for coefficients	Mean	SD	Min	Max	Mean	SD	Min	Max
	-0.163	0.115	-0.352	0.106	0.149	0.060	0.045	0.031

Notes: Outflow data are available from 1960 because four countries do not have outflows. The year against which other years are compared is 1950 for inflows and 1959 for outflows. SE = standard error. SD = standard deviation.

Table S4. QIC statistics for model selection under normal distribution of the inflow and outflow data. Model 4 (M4) is defined in Table 2 and Table 3. AR *n* = *n*th-order autoregressive correlation structure.

Inflow			
Correlation	Variable	p	QIC
Independent	M4[a]	20	**21203.84**
Exchangeable	M4	20	26396.26
AR 1	M4	20	22743.20
AR 2	M4	20	23100.35
AR 3	M4	20	22982.44
AR 4	M4	20	22756.11
AR 5	M4	20	22648.69
Independent	A: M4 - PSR (origin)	19	21176.34
Independent	B: A - Border	18	21141.15
Independent	C: B - Common official language	17	21130.84
Independent	D: C - Land area (origin)	16	**21111.27**
Independent	E: D - (Year-1985)	15	21141.33

Outflow			
Correlation	Variable	p	QIC
Independent	M4[a]	20	**11241.02**
Exchangeable	M4	20	13575.19
AR 1	M4	20	11308.81
AR 2	M4	20	11444.97
AR 3	M4	20	11486.20
AR 4	M4	20	11532.12
AR 5	M4	20	11518.12
Independent	A: M6 - 9% minority speak same language	19	11190.76
Independent	B: A - (Year - 1985)	18	11169.58
Independent	C: B - PSR (destination)	17	11139.00
Independent	D: C - Land area (origin)	16	11107.54
Independent	E: D - Border	15	11072.86
Independent	F: E - Landlocked (destination)	14	**11062.74**

Independent	G: F - IMR (origin)	13	11161.37

Note: Values in boldface indicate the smallest QIC value. p denotes number of parameters in the model. Although we tested stationary, non-stationary, and unstructured correlation structures, we could not achieve convergence.

Table S5. Robustness checks: for inflow and outflow models estimated after exclusion of Hungary and Croatia (1), exclusion of Hungary, Croatia, and U.K. (2), and exclusion of outliers based on DFITS statistics (3)-(6).

	Inflow						Outflow					
	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	OLS	OLS	GEE (ind)	GEE (exc)	GEE (ar1)	OLS	OLS	OLS	GEE (ind)	GEE (exc)	GEE (ar1)
Demographic determinants												
Log population (destination)	0.613***	0.607***	0.656***	0.656***	0.631***	0.741***	0.375***	0.379***	0.419***	0.419***	0.460***	0.419***
	(0.009)	(0.009)	(0.008)	(0.030)	(0.031)	(0.025)	(0.008)	(0.008)	(0.007)	(0.029)	(0.044)	(0.027)
Log population (origin)	0.733***	0.738***	0.784***	0.784***	0.993***	0.741***	0.911***	0.904***	0.948***	0.948***	0.794***	0.904***
	(0.006)	(0.006)	(0.006)	(0.025)	(0.041)	(0.023)	(0.011)	(0.011)	(0.010)	(0.037)	(0.033)	(0.031)
Log potential support ratio (destination)	-0.779***	-0.739***	-0.874***	-0.874***	-0.196	-0.770***	-0.058**	-0.050**	-0.001	-0.001	-0.460***	0.014
	(0.070)	(0.070)	(0.062)	(0.210)	(0.212)	(0.194)	(0.024)	(0.024)	(0.022)	(0.080)	(0.121)	(0.071)
Log potential support ratio (origin)	0.045**	0.051**	0.078***	0.078	-0.159	-0.183***	0.823***	0.854***	0.644***	0.644***	0.662***	0.771***
	(0.020)	(0.020)	(0.019)	(0.067)	(0.102)	(0.065)	(0.079)	(0.079)	(0.070)	(0.241)	(0.188)	(0.189)
Log infant mortality rate (destination)	0.927***	0.936***	0.902***	0.902***	-0.080	-0.297***	-0.801***	-0.798***	-0.846***	-0.846***	-0.224***	-0.804***
	(0.051)	(0.051)	(0.044)	(0.130)	(0.107)	(0.110)	(0.016)	(0.016)	(0.014)	(0.051)	(0.075)	(0.045)
Log infant mortality rate (origin)	-0.471***	-0.464***	-0.488***	-0.488***	0.266***	-0.353***	0.531***	0.560***	0.408***	0.408***	-0.165	0.202**
	(0.013)	(0.013)	(0.012)	(0.046)	(0.062)	(0.043)	(0.057)	(0.057)	(0.049)	(0.148)	(0.124)	(0.101)
Log proportion of urban population (destination)	3.193***	3.146***	3.042***	3.042***	3.105***	3.249***	0.302***	0.285***	0.297***	0.297***	0.762***	0.294***
	(0.075)	(0.075)	(0.064)	(0.216)	(0.363)	(0.213)	(0.021)	(0.021)	(0.019)	(0.067)	(0.107)	(0.060)
Log proportion of urban population (origin)	0.329***	0.324***	0.365***	0.365***	1.014***	0.432***	2.367***	2.398***	2.395***	2.395***	2.433***	2.615***
	(0.017)	(0.017)	(0.016)	(0.063)	(0.090)	(0.059)	(0.080)	(0.080)	(0.069)	(0.244)	(0.339)	(0.228)
Geographic determinants												
Log distance between capitals	-0.825***	-0.843***	-0.801***	-0.801***	-0.858***	-0.707***	-0.674***	-0.693***	-0.656***	-0.656***	-0.593***	-0.641***
	(0.011)	(0.011)	(0.010)	(0.040)	(0.051)	(0.039)	(0.012)	(0.012)	(0.011)	(0.046)	(0.056)	(0.043)
Log land area (destination)	0.233***	0.241***	0.227***	0.227***	0.274***	0.217***	0.147***	0.144***	0.145***	0.145***	0.048	0.136***
	(0.008)	(0.008)	(0.007)	(0.026)	(0.028)	(0.025)	(0.007)	(0.007)	(0.006)	(0.025)	(0.032)	(0.022)
Log land area (origin)	-0.050***	-0.056***	-0.072***	-0.072***	-0.261***	-0.047**	0.033***	0.043***	0.054***	0.054*	0.155***	0.086***
	(0.005)	(0.005)	(0.005)	(0.021)	(0.030)	(0.020)	(0.009)	(0.009)	(0.008)	(0.031)	(0.033)	(0.028)
Landlocked (destination)	n/a	n/a	-0.636***	-0.636***	-0.155*	-0.272***	-0.084***	-0.086***	-0.035***	-0.035	-0.092**	-0.056
	n/a	n/a	(0.045)	(0.063)	(0.087)	(0.060)	(0.011)	(0.011)	(0.010)	(0.036)	(0.043)	(0.034)

Landlocked (origin)	-0.173***	-0.172***	-0.133***	-0.133***	-0.165***	-0.147***	n/a	n/a	-1.132***	-1.132***	-0.724***	-0.981***
	(0.009)	(0.009)	(0.008)	(0.032)	(0.037)	(0.031)	n/a	n/a	(0.043)	(0.069)	(0.072)	(0.064)
Border	0.034	0.040*	0.035*	0.035	0.289***	0.159**	0.035	0.049*	0.075***	0.075	0.374***	0.174**
	(0.023)	(0.023)	(0.021)	(0.081)	(0.082)	(0.074)	(0.025)	(0.025)	(0.023)	(0.080)	(0.089)	(0.077)
Social and historical determinants												
Common official language	0.143***	0.136***	0.046***	0.046	0.157***	0.123**	0.361***	0.355***	0.215***	0.215***	0.298***	0.259***
	(0.014)	(0.014)	(0.014)	(0.054)	(0.057)	(0.053)	(0.028)	(0.028)	(0.028)	(0.059)	(0.064)	(0.054)
9 % minority speak same language	0.257***	0.252***	0.360***	0.360***	0.255***	0.359***	-0.021	-0.038	0.054**	0.054	0.078	0.052
	(0.014)	(0.014)	(0.013)	(0.051)	(0.054)	(0.050)	(0.027)	(0.027)	(0.027)	(0.051)	(0.061)	(0.048)
Colony	0.434***	0.368***	0.385***	0.385***	0.481***	0.382***	0.775***	0.682***	0.795***	0.795***	0.894***	0.819***
	(0.017)	(0.018)	(0.017)	(0.062)	(0.059)	(0.054)	(0.024)	(0.028)	(0.025)	(0.083)	(0.097)	(0.083)
Year - 1985	0.007***	0.007***	0.006***	0.006***	0.001	-0.008***	0.002	0.003**	-0.001	-0.001	-0.003	-0.004**
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.003)	(0.002)	(0.002)
(Year - 1985)^2	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	-0.000***	-0.000***	-0.000***	-0.000***	0.000	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	-10.476***	-10.321***	-10.925***	-10.925***	-13.691***	-14.322***	-11.416***	-11.357***	-12.473***	-12.473***	-12.777***	-13.248***
	(0.247)	(0.247)	(0.207)	(0.627)	(0.879)	(0.589)	(0.280)	(0.280)	(0.237)	(0.786)	(0.909)	(0.662)
Observations	46458	46216	44444	44444	44444	44375[a]	27576	27334	26488	26488	26488	26391[a]
R-squared	0.639	0.635	0.7086				0.67	0.656	0.7309			
MSE	0.436	0.433	0.329				0.371	0.371	0.283			
AIC	94285	92534	76690				50946	50514	41785			
BIC	94461	92700	76864				51102	50670	41949			
Dispersion				0.328	0.393	0.345				0.283	0.336	0.286
QIC				21148	24666	21994				11102	12670	11054

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
MSE: Mean square residual
ind: independent error structure, exc: exchangeable error structure, ar1: first order autoregressive error structure
a: Panels having fewer than 2 years of observations are excluded.
n/a: not available

Table S6. Variables in the supplementary data sets inflow.csv and outflow.csv

Name of variable	
id	Identification number
year	Calendar year
dest	Numeric country code (destination)
dest2	Country name (destination)
origin	Numeric country code (origin)
origin2	Country name (origin)
mig	Number of migrants
areaorig	Land area (origin)
areadest	Land area (destination)
ppndest	Population (destination)
ppnorig	Population (origin)
distcap	Distance between origin and destination (km)
d_landlocked	Landlocked location (destination)
o_landlocked	Landlocked location (origin)
contig	Sharing a border
comlang_off	Common official language
comlang_ethno	9% minority speak same language
colony	Colonial link
o_pcturban100	Percentage of urban population (origin)
d_pcturban100	Percentage of urban population (destination)
d_imr3	Infant mortality rate (destination)
o_imr3	Infant mortality rate (origin)
d_psr2	Potential support ratio (destination)
o_psr2	Potential support ratio (origin)

Figure S1. Total annual inflows by destination

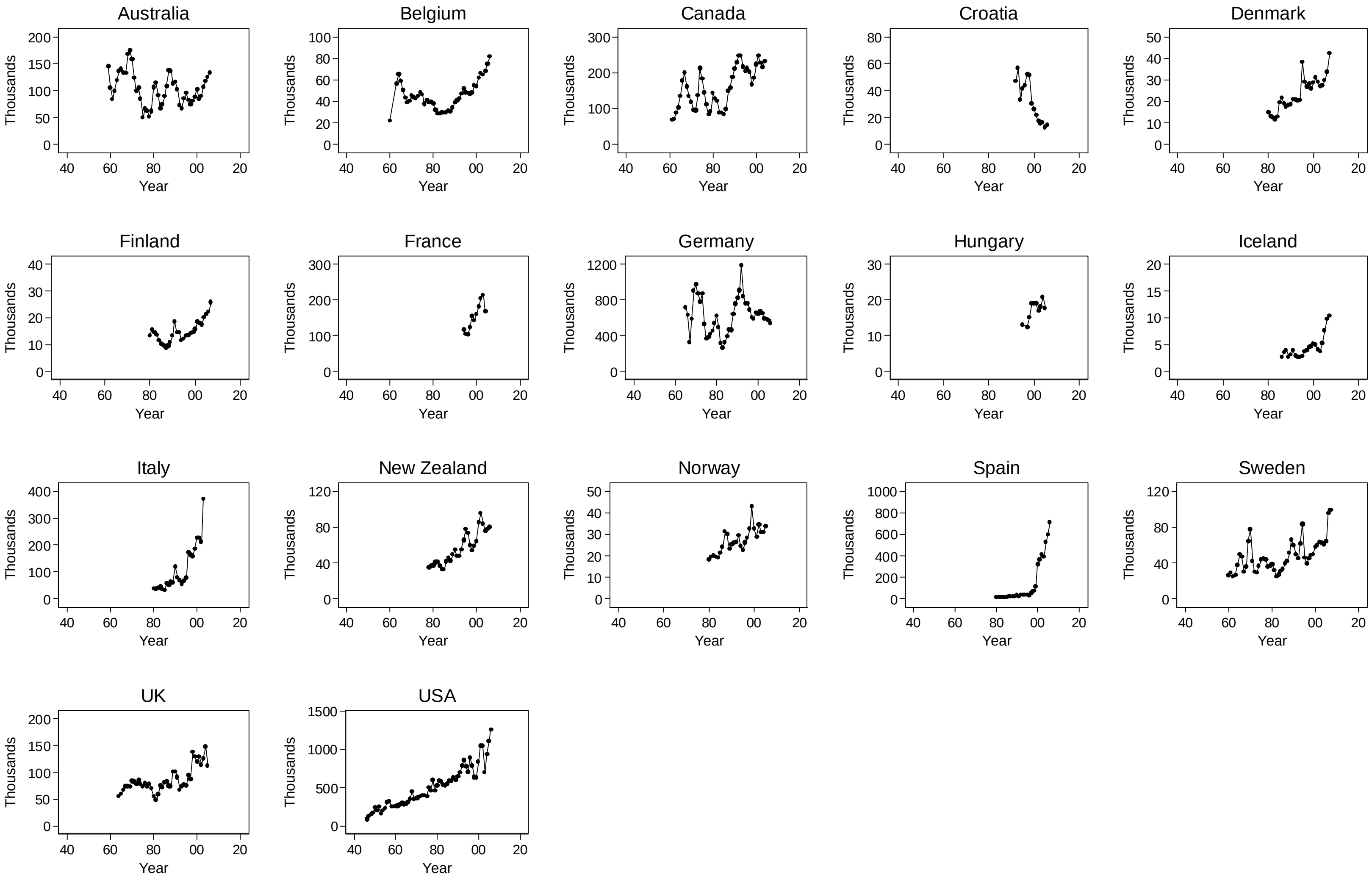


Figure S2. Total annual outflows by origin

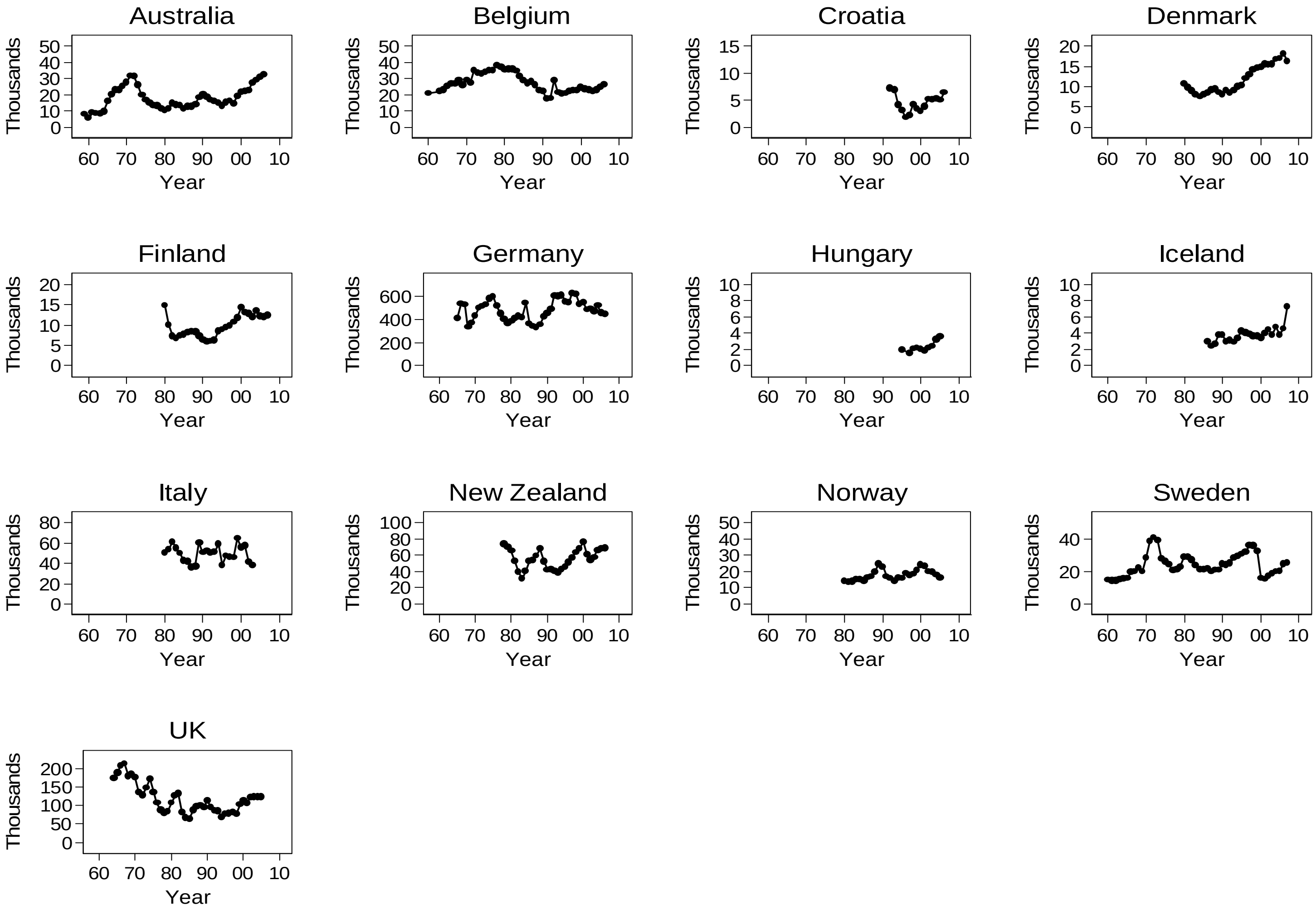
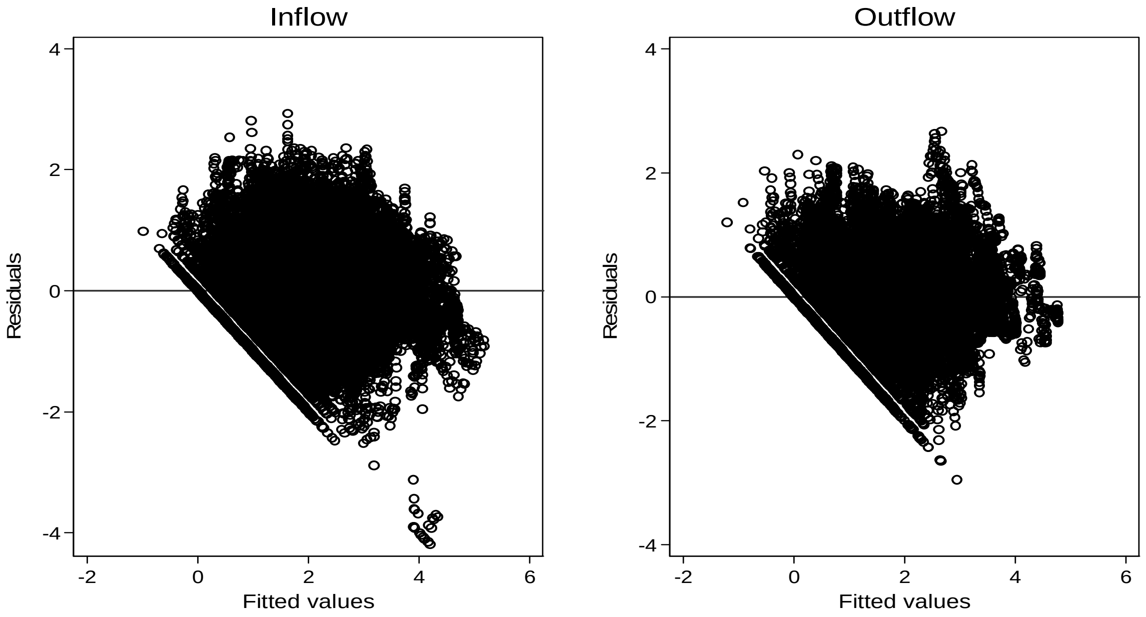


Figure S3. Regression diagnostics for inflow and outflow models



Notes: Predicted values and residuals are based on M1 in Tables 2 and 3. Because coefficients from GEE models with independent error structure are the same as the OLS coefficients, predicted values and residuals are the same as well. The straight lines along the bottom left edge of the data points in the above plots are cases where the number of migrants was 1. In such cases, log migrants = 0 and the residual therefore equals the negative of the fitted value.

APPENDIX REFERENCES

Baum, C. F.

2006 *An Introduction to Modern Econometrics using Stata*. College Station, TX: Stata Press.

Bongaarts, John and Bulatao, Rodolfo A., eds.

2000 *Beyond Six Billion: Forecasting the World's Population.* Panel on Population Projections, Committee on Population, Commission on Behavioral and Social Sciences and Education, National Research Council. National Academy Press, Washington DC.

Cohen, J. E., Roig, M., Rueman, D. C., and GoGwilt, C.

2008 "International Migration Beyond Gravity: A Statistical Model for Use in Population Projections." *Proceedings of National Academy of Science* 105(40):15269-15274.

Congressional Budget Office, United States Congress

2005 CBO's Economic Forecasting Record: An evaluation of the economic forecasts CBO made from January 1976 through January 2003. Congress of the United States.

Cui, J.

2007 "QIC Program and Model Selection in GEE." *The Stata Journal* 7(2):209-220.

Frees, E. W.

2004 *Longitudinal and panel data: analysis and applications for the social sciences*. Cambridge: Cambridge University Press.

Groemling, M.

2002 Evaluation and Accuracy of Economic Forecasts. *Historical Social Research* 27(4):242-255.

Hardin, J. W., and Hilbe, J. M.

2003 *Generalized Estimating Equations*. Boca Raton, FL: Chapman & Hall/CRC.

Horton, N. J., and Lipsitz, S. R.

1999 “Review of Software to Fit Generalized Estimating Equation Regression Models.” *The American Statistician* 53:160-169.

Liang, K. Y., and Zeger, S. L.

1986 “Longitudinal Data Analysis Using Generalized Linear Models.” *Biometrika* 73(1):13-22.

Neumayer, E.

2005 “Bogus Refugees? The Determinants of Asylum Migration to Western Europe.” *International Studies Quarterly* 49: 389-409.

Pan, W.

2001 “Akaike's Information Criterion in Generalized Estimating Equations.” *Biometrics* 57(1):120-125.

Pedersen, P. J., Pytlikova, M., and Smith, N.

2008 “Selection and Network Effects - Migration Flows into OECD Countries 1990-2000.” *European Economic Review* 52:1160-1186.

Raftery, A. E.

1995 “Bayesian Model Selection in Social Research.” *Sociological Methodology* 25:111-163.

Wooldridge, J. M.

2006 *Introductory Econometrics: A Modern Approach*, *Third Edition*. Cincinnati, OH: South Western College Publishing.