International migration beyond gravity: A statistical model for use in population projections

Joel E. Cohen*†, Marta Roig‡, Daniel C. Reuman*§, and Cai GoGwilt¶

*Laboratory of Populations, Rockefeller and Columbia Universities, 1230 York Avenue, Box 20, New York, NY 10065-6399; ‡Population Division, Department of Economic and Social Affairs, 2, United Nations Plaza DC2-1984, United Nations, New York, NY 10017; §Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot SL5 7PY Berkshire United Kingdom; and ¶321 East 54th Street Apartment 4B, New York, NY 10022

Contributed by Joel E. Cohen, August 18, 2008 (sent for review June 18, 2008)

International migration will play an increasing role in the demographic future of most nations if fertility continues to decline globally. We developed an algorithm to project future numbers of international migrants from any country or region to any other. The proposed generalized linear model (GLM) used geographic and demographic independent variables only (the population and area of origins and destinations of migrants, the distance between origin and destination, the calendar year, and indicator variables to quantify nonrandom characteristics of individual countries). The dependent variable, yearly numbers of migrants, was quantified by 43653 reports from 11 countries of migration from 228 origins and to 195 destinations during 1960–2004. The final GLM based on all data was selected by the Bayesian information criterion. The number of migrants per year from origin to destination was proportional to (population of origin)$^{0.86}$(area of origin)$^{-0.21}$(population of destination)$^{0.36}$(distance)$^{-0.97}$, multiplied by functions of year and country-specific indicator variables. The number of emigrants from an origin depended on both its population and its population density. For a variable initial year and a fixed terminal year 2004, the parameter estimates appeared stable. Multiple R^2, the fraction of variation in log numbers of migrants accounted for by the starting model, improved gradually with recentness of the data: $R^2 = 0.57$ for data from 1960 to 2004, $R^2 = 0.59$ for 1985–2004, $R^2 = 0.61$ for 1995–2004, and $R^2 = 0.64$ for 2000–2004. The migration estimates generated by the model may be embedded in deterministic or stochastic population projections.

generalized linear model | geography | population density | spatial interaction model | stochastic population projection

International migration will play an increasing role in the demographic future of nations if fertility continues to decline in most countries. In projecting international migration, the United Nations Population Division (ref. 1, paragraphs 57–59) identified the need for a demographically plausible, programmable algorithm that automatically projects a zero world balance of net migration and prevents projected net emigration from completely depleting the population of any sending country. To meet this need, we propose an algorithm (based only on demographic and constant geographic variables) for projecting future numbers of international migrants from any country or region to any other. It is comparable in transparency and generality to standard cohort-component methods of projecting births and deaths. The approach presented here is different from methods of projecting migrant flows currently practiced in international demographic institutions, the United States, European countries, and other developed countries (2–6).

Most theories of international migration draw on social, economic and/or political factors to explain migration (4, 7–9), such as differences among countries in gross domestic product, labor markets, migration policies, social networks of prior migrants, and cognitive and behavioral attributes of individuals (3, 10–11). For multidecadal demographic projections, it seems more difficult to project such nondemographic variables than it is to project demographic variables such as fertility and mortality. The proposed model assumes the availability only of constant geographic variables and of population sizes that can be projected incrementally in time by accepted demographic procedures. The model makes possible both deterministic and stochastic projections of migration and hence of population.

The intellectual antecedents of the proposed model include Zipf's (12, 13) model of intercity migration, which is one of several "gravity" models in the social sciences (6). Zipf (12) aimed to "show with supporting data that the number of persons that move between any two communities in the United States whose respective populations are P_1 and P_2 and which are separated by the shortest transportation distance, D, will be proportionate to the ratio, $P_1 \cdot P_2/D$, subject to the effect of modifying factors." Unlike Zipf, we distinguish the number of people who move from community 1 to community 2 from the number of people who move from 2 to 1. Taking logarithms of $P_1 \cdot P_2/D$ and adding an error term yields a linear model in log-transformed variables, log(migrants) = $a_0 + a_1$ log(ppnorig) + a_2 log(ppndest) + a_3 log(distance) + error, where ppnorig is the population of the origin, ppndest is the population of the destination, and the error term characterizes random deviations. Here and throughout, log refers to $\log_{10}$, and ln refers to the natural logarithm $\log_e$. Zipf posits that $a_1 = 1$, $a_2 = 1$, and $a_3 = -1$. We estimate all coefficients from data using a generalized linear model (GLM) (14).

Zipf (12) treated cities as points of negligible spatial extent. Our communities are countries or regions and our subject is international migration. To let the data reveal whether the area of a country influences its numbers of migrants, we add two terms to the above equation: log(areaorig), the log area of the origin, and log(areadest), the log area of the destination. By definition, the population density of the origin is ppnorig/areaorig, so log(density) = log(ppnorig) − log(areaorig). If the number of migrants from origin to destination depends on ppnorig and ppndest and not on their areas, then the estimated coefficients of log(areaorig) and log(areadest) should be close to zero. However, if the number of migrants from origin to destination depends on the population density of the origin and the population density of the destination, but not on their respective population numbers *per se*, then the estimated coefficients of log(areaorig) and log(areadest) should be nearly the negative of the respective estimated coefficients of log(ppnorig) and log(ppndest). The estimated coefficients of the terms for log population and log area of origin and destination reveal the relative importance of population *per se* and population density.

To allow for differences in migratory intensities among origins or destinations, we let the data reveal whether different origins and destinations had numbers of migrants different from the numbers of migrants expected on average from their respective populations, areas, and distances. We introduced four indicator variables for

Author contributions: J.E.C. designed research; J.E.C. performed research; J.E.C., D.C.R., and C.G. contributed new reagents/analytic tools; J.E.C. and M.R. analyzed data; and J.E.C., M.R., and D.C.R. wrote the paper.

The authors declare no conflict of interest.

†To whom correspondence should be addressed. E-mail: cohen@rockefeller.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0808185105/DCSupplemental.

each country that provided data. For example, Australia was a source of data on numbers of emigrants and immigrants by year. One indicator variable for Australia, orig.indicator$Australia, equaled 1 whenever Australia was the origin and equaled 0 if any other country or area was the origin. A second indicator variable, dest.indicator$Australia, similarly equaled 1 whenever Australia was the destination and equaled 0 if any other country or area was the destination. A third indicator variable, orig.is.datasource$Australia, equaled 1 if Australia was the origin and was the source of the data and otherwise equaled 0. A fourth indicator variable, dest.is.datasource$Australia, equaled 1 if Australia was the destination and was the source of the data and otherwise equaled 0. Similar indicator variables were defined for each country that provided data on its flows of immigrants and/or emigrants.

To allow for the possibility that migration rates changed over time and for the simultaneous effects of the other independent variables (populations, areas, distance, and indicator variables), we introduced year as an independent variable. We used the centered variable year minus 1985 to assure the stability of the model's estimated intercept.

To summarize in notation similar to that of Zipf's gravity model, let P be population, A be area, and D be distance. Detailed definitions of these variables are given in *Methods*. Our "starting" model was the logarithmic transformation of model Eq. **1**. This logarithmic transformation guaranteed that the number of migrants estimated by the model was positive.

$$\begin{aligned}&\text{migrants from origin to destination} \qquad \textbf{[1]}\\ &= \text{constant} \times P^{a}_{\text{orig}} P^{b}_{\text{dest}} A^{c}_{\text{orig}} A^{d}_{\text{dest}} D^{f}_{\text{orig,dest}} \times\\ &[10 \text{ if origin is Australia; 1 otherwise}]^{g} \times\\ &[10 \text{ if destination is Australia; 1 otherwise}]^{h} \times\\ &[\text{similar indicator variables for seven other origins and ten other destinations}]^{\text{coef}} \times\\ &[10 \text{ if origin is Australia and Australia is the source of the data; 1 otherwise}]^{i} \times\\ &[10 \text{ if destination is Australia and Australia is the source of the data; 1 otherwise}]^{j} \times\\ &[\text{similar indicator variables for seven other origins and ten other destinations}]^{\text{coef}} \times\\ &10^{k(\text{year}-1985)} \times\\ &[\text{lognormal error term } \exp(\varepsilon)],\end{aligned}$$

where $\varepsilon \sim N(0,\sigma^2)$ are independent.

This model contains factors of the form [10 if origin is Australia; 1 otherwise]^g. The common logarithm of this expression is $g \times$ [1 if origin is Australia; 0 otherwise], and the expression in square brackets is an indicator (or dummy) variable. The coefficient g summarizes the nonrandom effects on the expected number of migrants from Australia (in this case) apart from those due to calendar year, the populations and areas of origin and destination, and distance. The indicator variables quantified how numbers of migrants were affected by nonrandom characteristics of individual countries: migratory history, policy, and statistical completeness; economic, geographic, and cultural affinities or disparities; and effects of geographical adjacency not captured by the chosen measure of distance.

We estimated the intercept (log constant) and exponents a, b, c, d, …, which were linear coefficients in the GLM. We fitted this and other models to data from 11 countries (Australia, Belgium, Canada, Denmark, Germany, Italy, the Netherlands, Spain, Sweden, the U.K., and the U.S.A.). These countries were selected on the basis of the quality of their data on international migrants by year and places of origin and destination from 1960 to 2004. The data included 228 origins and 195 destinations of migrants. Oceania was the only destination not also an origin, so 229 countries or regions in total were named in these data.

Because of the limited quantity and quality of migration data, this article demonstrates a method and illustrates a modeling approach, rather than specifying numerical parameters definitively. Parameters and models will evolve as more and better data become available. The analysis showed that statistically simple and demographically interpretable modeling accounted for more than half the variation in the migration data. How much more than half could be accounted for by this approach with better data remains to be determined.

Results

Descriptive Bivariate Relationships. On average, but with enormous variability, the log number of migrants increased with increasing log population of origin ($r = 0.43$), increasing log area of origin ($r = 0.18$), increasing log population of destination ($r = 0.27$), and increasing log area of destination ($r = 0.10$) (Fig. 1 *A–D*). The log number of migrants decreased, on average but with enormous variability, with increasing log distance from origin to destination ($r = -0.24$) and increased weakly with year ($r = 0.01$) (Fig. 1 *E* and *F*). Log population and log area were highly correlated (Fig. 1 *G* and *H*) for origins ($r = 0.74$) and destinations ($r = 0.64$). Because these were the two highest-magnitude correlations, collinearity was not a problem in fitting the GLM. Correlations between year and log migrants, and between log(ppnorig) and log(distance), were insubstantial.

Log(ppnorig) and log(ppndest) were negatively correlated ($r = -0.19$) (Fig. 1*I*). This negative correlation could reflect the absence of data sources among countries with four million or fewer people, which may account for the absence of data points in the lower left quadrant of Fig. 1*I*. The long horizontal and vertical streaks in Fig. 1*I* represent the populations of countries that were data sources as origins and destinations, respectively, whereas the short diagonal streaks largely reflect population growth of a given (origin, destination) pair.

Model Selection. The starting linear model was fitted [supporting information (SI) Table S1] with the dependent variable log(migrants) and with the six independent variables that we call "basic" [year minus 1985, log(ppnorig), log(areaorig), log(ppndest), log(areadest), log(distance)] and all indicator variables (orig.indicator, dest.indicator, orig.is.datasource, dest.is.datasource). The Multiple R^2 was 0.5693 and the Adjusted R^2 was 0.5689 (see *Methods*).

When the stepwise algorithm with Bayesian information criterion (15) was applied to this starting model, log(areadest) was eliminated and all other independent variables were retained in the resulting "final" model (Table 1). To the four significant digits shown, the Multiple R^2 and the Adjusted R^2 were unchanged between the starting and the final models.

When the values of log(migrants) were independently and randomly permuted in each of 100 simulations, the maximum of the 100 simulated multiple R^2 values was 0.00147, far smaller than the multiple R^2 value of 0.5693 for the data. Hence, the latter multiple R^2 value could not have been an artifact of the fitting procedure alone.

When all indicator variables were suppressed and only the six "basic" independent variables were retained, the multiple R^2 value dropped to 0.4345 (Table S3). The only notable change in the coefficients of the six "basic" independent variables was the increase in the coefficient of log(areadest) from 0.0239 to 0.1604. A

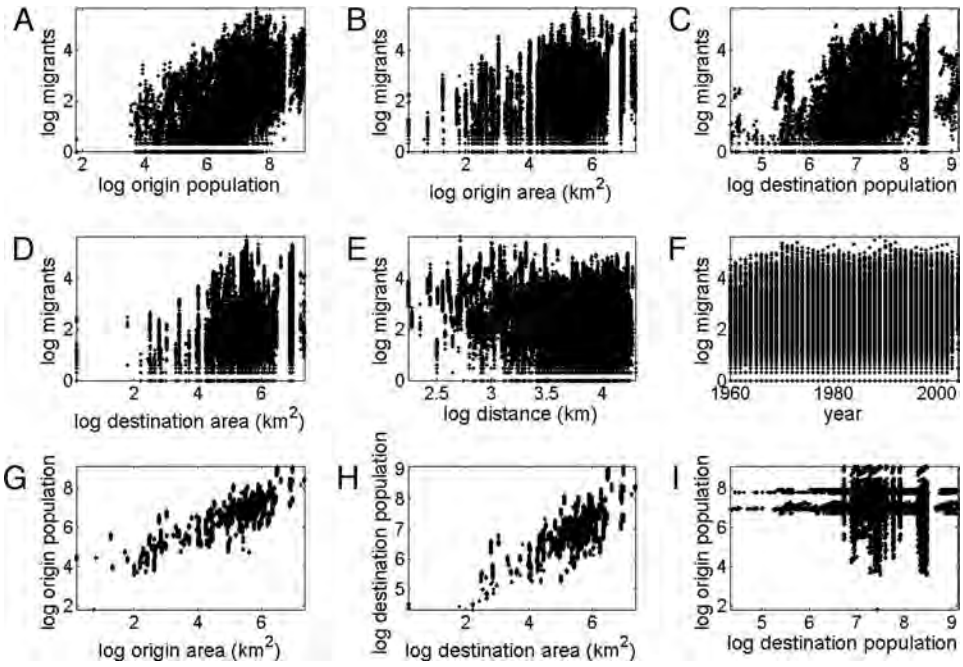

Fig. 1. Bivariate relationships in international migration data, 1960–2004. (*A*) Log number of migrants versus log population of origin. (*B*) Log number of migrants versus log area of origin. (*C*) Log number of migrants versus log population of destination. (*D*) Log number of migrants versus log area of destination. (*E*) Log number of migrants versus log distance from origin to destination. (*F*) Log number of migrants versus year. (*G*) Log population of origin versus log area of origin. (*H*) Log population of destination versus log area of destination. (*I*) Log population of origin versus log population of destination. Each plot has 43,653 points.

model that omitted the indicator variables would have misleadingly suggested that log(areadest) was fairly influential. The stepwise algorithm, however, retained the indicator variables and dropped log(areadest).

Residuals. In the final model, the interquartile range of the residuals was −0.4352 to 0.4414 log(migrants), meaning that half the time, the observed numbers of migrants fell in the interval from 36.7% to 2.763 times the predicted number of migrants (the predicted number was $10^{\text{expected log(migrants)}}$). The smallest and largest residuals were −3.2449 and 3.2918, corresponding to cases where the observed number of migrants was >1,000 times smaller or larger than the predicted number.

The largest residuals occurred at intermediate fitted values from 0.5 to 3.5, corresponding to predicted numbers of migrants from ≈3 to 3,000 (Fig. S1*A*). The scatter of the residuals was clearly not constant over the range of fitted values. This lack of homoscedasticity justified the use of the Bayesian information criterion for model selection instead of a probabilistic interpretation of *F* tests for omitted variables. The validity of the latter approach assumes homoscedasticity of residuals and independence of observations.

When the fitted values were >4 on the $\log_{10}$ scale (corresponding to 10,000 migrants per year or more), residuals were systematically negative, indicating fewer reported migrants than predicted. This pattern could result, among other possible causes, from underreporting of large migrant flows or from systematic policies intended to diminish the largest predicted flows.

Model Coefficients. In the final model (Table 1), the predicted number of migrants increased by $0.38\% = 10^{0.00163}$ per year, in addition to the changes in numbers of migrants resulting from changes over time in log(ppnorig) and log(ppndest). The predicted number of migrants was proportional to the population of origin raised to 0.86 and to the population of destination raised to 0.36. Increases in log(ppnorig) increased log(migrants) more than twice as much as increases in the log(ppndest). In light of the small standard errors of these estimates (Table 1), these exponents very probably differed from the exponents of 1 in Zipf's (12) gravity model, even if the distributional assumptions of the linear model were not precisely met. The predicted number of migrants was proportional to the distance raised to −0.97, a bit more than three standard errors from Zipf's posited exponent of −1. The predicted number of migrants was proportional to the area of origin to the power −0.21. Because $P^{0.86}A^{-0.21} = P^{0.65}(P/A)^{0.21}$, the number of migrants increased with both the population of origin (to the power of 0.65) and the population density of origin (to the power 0.21), and the population of origin contributed more to the number of migrants than did the population density of origin.

The indicator variables revealed substantial heterogeneity among countries in their propensity to send or receive migrants and in their reporting practices, given the other independent variables. According to its coefficient for orig.indicator, Australia had $13.47 = 10^{1.1295}$ times as many emigrants as expected. At the opposite extreme, Belgium had 56% as many emigrants as expected from its other characteristics. Denmark, Germany, and the Netherlands had approximately as many emigrants as expected. According to its coefficient for dest.indicator, Australia had 27.31 times as many immigrants as expected from its geographic and demographic characteristics, followed by the U.S.A. and Canada with 13.95 and 7.17 times as many immigrants as expected, respectively. At the opposite extreme, Belgium had 1.35 times as many immigrants as expected, the smallest multiple among the countries that provided data in this study. According to dest.indicator, all reporting countries had more immigrants than expected on average. The countries that provided data for this study are countries with the resources to support effective statistical systems, and such countries are likely to be attractive destinations of migration.

According to its coefficient for orig.is.datasource, when Australia reported the number of emigrants from Australia, it reported on average $50\% = 10^{-0.3038}$ of the number of emigrants from Australia that the average country that reported emigration data in this study would have reported, given Australia's population and area, the destination and year, and the propensity to emigrate from Australia estimated without regard to the source of the data. For example, Australia reported 3,971 migrants from Australia to the U.K. in 1998, whereas the U.K. reported 41,800 migrants from Australia to the U.K. in 1998 (see *Methods*). Australia reported emigrants leaving permanently. The U.K. reported immigrants staying for 1 year or longer.

At the opposite extreme from low reporters of emigration like Australia (50%) and Italy (33%), the U.K. reported 22.54 times as many emigrants as would be expected otherwise, whereas Germany reported 3.07 times as many. The U.K. data are estimates derived

Table 1. The "final" model of log migrants as a function of year, log population of origin, log area of origin, log population of destination, log distance, and indicator variables; specification of the model in R and resulting coefficients and statistics

Call:
lm (formula = logmigrants ~ I (year−1985) + logppnorig + logareaorig + logppndest + logdistance + orig.indicator + dest.indicator + orig.is.datasource + dest.is.datasource)

Coefficients	Estimate	SE	*t* value
(Intercept)	−2.5135	0.0886	−28.3730
I (year–1985)	0.0016	0.0003	5.1670
Logppnorig	0.8631	0.0083	103.6400
Logareaorig	−0.2103	0.0066	−31.9050
Logppndest	0.3604	0.0089	40.7010
Logdistance	−0.9685	0.0102	−94.5470
orig.indicatorAustralia	1.1295	0.0436	25.8900
orig.indicatorBelgium	−0.2557	0.0404	−6.3300
orig.indicatorDenmark	−0.0441	0.0409	−1.0760
orig.indicatorGermany	0.0699	0.0409	1.7080
orig.indicatorItaly	0.1844	0.0401	4.5960
orig.indicatorNetherlands	0.0250	0.0408	0.6120
orig.indicatorSweden	0.1602	0.0473	3.3840
orig.indicatorUnited Kingdom	0.2486	0.0397	6.2580
dest.indicatorAustralia	1.4362	0.0558	25.7360
dest.indicatorBelgium	0.1314	0.0524	2.5090
dest.indicatorCanada	0.8557	0.0457	18.7160
dest.indicatorDenmark	0.2560	0.0523	4.8980
dest.indicatorGermany	0.5875	0.0502	11.7020
dest.indicatorItaly	0.7551	0.0493	15.3090
dest.indicatorNetherlands	0.4805	0.0509	9.4480
dest.indicatorSpain	0.6400	0.0470	13.6210
dest.indicatorSweden	0.2528	0.0696	3.6340
dest.indicatorUnited Kingdom	0.6284	0.0491	12.7910
dest.indicatorUnited States of America	1.1444	0.0457	25.0180
orig.is.datasourceAustralia	−0.3038	0.0633	−4.8010
orig.is.datasourceBelgium	0.4595	0.0626	7.3410
orig.is.datasourceDenmark	0.2340	0.0642	3.6450
orig.is.datasourceGermany	0.4872	0.0613	7.9500
orig.is.datasourceItaly	−0.4761	0.0630	−7.5620
orig.is.datasourceNetherlands	0.2112	0.0645	3.2750
orig.is.datasourceSweden	−0.0733	0.0658	−1.1140
orig.is.datasourceUnited Kingdom	1.3529	0.0682	19.8440
dest.is.datasourceAustralia	−0.0317	0.0719	−0.4420
dest.is.datasourceBelgium	0.5479	0.0716	7.6510
dest.is.datasourceCanada	0.1462	0.0638	2.2930
dest.is.datasourceDenmark	0.2687	0.0721	3.7270
dest.is.datasourceGermany	0.5646	0.0674	8.3730
dest.is.datasourceItaly	−0.2356	0.0687	−3.4280
dest.is.datasourceNetherlands	0.4550	0.0718	6.3360
dest.is.datasourceSpain	−0.2294	0.0678	−3.3850
dest.is.datasourceSweden	0.1258	0.0831	1.5140
dest.is.datasourceUnited Kingdom	1.4979	0.0762	19.6680
dest.is.datasourceUnited States of America	NA	NA	NA

Residual standard error: 0.6957 on 43610 degrees of freedom
Multiple R^2: 0.5693, Adjusted R^2: 0.5689
F statistic: 1,372 on 42 and 43,610 DF, nominal *P* value: < 2.2e-16

Because the assumption of independence among observations was implausible, conventional measures of statistical significance were probably inapplicable. The destination.is.data.source parameter for the U.S. is labeled "NA" because one of the destination.is.data.source variables must be eliminated to prevent a singularity in estimating the coefficients. See *SI Text* for details.

mainly from a survey of arriving and departing international passengers; U.K.'s recording of emigration and immigration flows is equally complete. Australia, by contrast, focuses on entries. These differences between countries are detected by the statistical analysis and reflected in the estimated coefficients of the model.

According to its coefficient for dest.is.datasource, when the U.K. reported the number of immigrants to the U.K., its reports were on average 31.47 times the number of immigrants that would have been reported by the average reporting system over all countries in the study, given the U.K.'s population and area, the destination and year, and the propensity to immigrate to the U.K. estimated without regard to the source of the data. At the opposite extreme, Spain and Italy reported 59% and 58%, respectively, of the numbers of immigrants expected from other factors.

In general, immigration is better recorded at destinations than emigration is recorded at origins, in part because migrants often have more formal incentives to register at their destination than to deregister at their origin. The estimated coefficients of the indicator variables (Table 1) are consistent with this belief, although other interpretations are possible. For example, dest.indicator for Australia had coefficient 1.4362, greater than the coefficient 1.1295 of orig.indicator for Australia. The same inequality, dest.indicator coefficient > orig.indicator coefficient, held for all eight countries that supplied both immigration and emigration data. Similarly, the comparable inequality, dest.is.datasource coefficient > orig.is.datasource coefficient, held for all eight countries that supplied both immigration and emigration data. These inequalities indicated greater detection (or greater intensity, an alternative interpretation) of immigration than emigration in every country for which the comparison could be made.

Parameter Stability: How Much of the Past Is Relevant to the Future?

Estimated coefficients varied systematically as a function of the time interval from which data were drawn and as a function of the subset of variables selected from the starting model Eq. **1** (Table S3). The starting model rather than the final model was used for this analysis to allow for the possibility that the log area of destination might become dramatically important for some subset of 1960–2004. As it turned out, this possibility did not occur.

For each set of independent variables, the most recent five years (2000–2004) gave the highest multiple R^2. Using all variables, all data from 1960 to 2004 gave a multiple R^2 value of 0.57, whereas the 2000–2004 data gave a multiple R^2 value of 0.64 (Table S3). This higher value of the multiple R^2 may be partially due to improved quality of data in more recent years but may also be due to fewer external perturbations to migratory flows during 2000–2004 than during 1960–2004. For example, during the 45-year period of 1960–2004, the Berlin wall and the Soviet Union fell, and Germany was reunified, whereas no such events marked the 5-year period of 2000–2004, an interval only one-ninth as long. For each set of independent variables, as the initial year moved forward while the terminal year was 2004, the multiple R^2 value increased. Each set of independent variables accounted for more variation in log(migrants) when using the data from 1985 to 2004 than by adding additional data from earlier years.

All estimated coefficients were stabler when using a moving initial year than when using a moving terminal year. They were least stable when using moving tranches of 5 or 10 years' duration (Table S3). For the starting model with all variables, the standard deviation 0.3877 of the estimated intercept for the five time intervals with moving terminal year 1960–1984, 1960–1989, 1960–1994, 1960–1999, and 1960–2004 was larger than the standard deviation 0.2489 of the estimated intercept for the five time intervals with moving initial year 1965–2004, 1970–2004, 1975–2004, 1980–2004, and 1985–2004. The same inequality held for the standard deviations of the estimates of the coefficients of all six basic variables when the five time intervals with moving terminal year were compared with the five time intervals with moving initial year.

Discussion

We assembled annual data on immigrants and emigrants from 11 countries' sources and combined them with data on the populations and areas of 228 origins and 195 destinations and the distances between origins and destinations. These 43,653 reports did not suffice to cover the $228 \times 195 = 44{,}460$ possible origin-destination pairs in a single year and offered very sparse coverage over 45 years. A simple GLM was able to account for more than half the variation in log(migrants). Despite the present limitations of data, this approach may improve on existing demographic procedures for projecting international migration and may motivate the collection of better data.

The bivariate relations among variables demonstrated the need for a multivariate model. For example, the number of migrants increased with the population of origin and the population of destination, but the population of origin and the population of destination were negatively correlated. Only a multivariate model could reveal how the number of migrants depended on the populations of origin and of destination jointly.

In the final GLM, the coefficients of log(ppnorig) and log(ppndest) were both positive and <1. If either coefficient had been negative, then the estimated number of migrants could have diverged to infinity as the population of origin or destination became smaller. Because the coefficients were <1, the numbers of migrants did not increase in proportion to ppnorig or ppndest.

The GLM served two distinct purposes: understanding and prediction. For scientific understanding, we eliminated superfluous independent variables to obtain the most economical model, then interpreted the values of the coefficients in the model. For prediction, we sought as much predictive power as possible by adding variables that gave the highest coefficient of determination, provided that the parameter estimates did not become unstable (sensitive to the inclusion or exclusion of a small number of data points and/or other predictor variables). The starting and final models considered here balanced interpretability and predictive ability. Other models are discussed in *SI Text* (Table S2).

The principal problems with this method were the lack of data and the lack of comparability (discrepancies between countries in definitions and measurements) where the data existed (*Methods*). Most countries lack a system to record migration flows. Many do not publish their information on migration. The data did not consistently distinguish moves from movers (6); individuals who crossed borders multiple times may have been interpreted as multiple migrants.

Comparisons with Some Related Studies. The 2003 Technical Panel on Assumptions and Methods of the Social Security Administration (16) suggested assuming that the number of net migrants to the U.S. will grow, at least in the long run, in direct proportion to the size of the U.S.A. population. The coefficients in the final model (Table 1) suggested, by contrast, that immigration to the United States is expected to grow in proportion to the population of the United States raised to the power 0.36, times independent multiplicative effects of the calendar year. The final model also anticipates changing log(migrants) as a result of population growth in countries of origin. Likewise, according to the final model, emigration from the United States should be expected to grow in proportion to the population of the United States raised to the power 0.86, times multiplicative effects that depend on year and country-of-destination populations.

If countries' populations changed by a factor of $\lambda > 0$, holding constant all other variables and GLM coefficients, the number of migrants would be multiplied by a factor of λ^{a+b} because $(\lambda\cdot\text{ppnorig})^a(\lambda\cdot\text{ppndest})^b = \lambda^{a+b}\cdot\text{ppnorig}^a\cdot\text{ppndest}^b$, where $a + b$ is the sum of the coefficients of log(ppnorig) and log(ppndest). In the final model (Table 1), $a + b = 1.22$; so if ppnorig and ppndest both doubled, the predicted number of migrants from origin to destination would increase by a factor of $2^{1.22} = 2.34$. For moving time intervals in the starting model with all variables (Table S3), $a + b$ tended to increase with the moving initial year or moving final year or tranche. For example, for data in the time intervals 1980–2004 and 1985–2004, $a + b = 1.26$ and 1.30, respectively.

Bijak *et al.* (17) projected migratory flows among 27 European countries by multiplying the initial emigration rates by an overall trend (mobility increasing by 0.5% yearly) and temporal effects of labor market policies. For comparison, our final model estimated that the global number of migrants rose by $10^{0.00163} = 1.0038 = 0.4\%$ per year in the data of 1960–2004, apart from the multiplicative effects of population growth or decline and the indicator variables of the countries or regions of origin and destination. This agreement in estimates is remarkable considering the difference in methods, data, and context (Europe versus the globe). Raymer (18) reconstructed the migratory flows for the European Union using a different log-linear model with multiplicative components.

Some migration models are based on transition probability matrices (6). In such models, the number of immigrants is independent of the destination's population and proportional to a weighted sum of the populations of origins or of the fractions of global population in different origins. The GLMs estimated here suggest that each destination's number of immigrants is a nonlinear function of the populations of both origin and destination and of other variables.

Use in Population Projection. The projected number of migrants can be embedded in a population projection algorithm, initially ignoring age structure and then incorporating it. The initial goal is, given a vector of country population sizes $P(t)$ with elements $P(i,t)$ for country i at time t, to compute the population vector $P(t+1)$ at the next time step. The GLM can estimate the number of migrants $M(i,j,t)$ from country i to country j between t and $t+1$. Let $M(i,i,t) = 0$ for all i (despite some countries' reporting positive numbers of migrants from the country to itself). The matrix $M(t)$ with elements $M(i,j,t)$ is called the migration matrix at time t. It will be assumed that $M(i,j,t)$ obtained from the GLM approximates the number of people who were in country i at time t and in country j ($\neq i$) at time $t + 1$.

The number of emigrants from country i between t and $t+1$, denoted $E(i,t)$, is then $E(i,t) = \Sigma_j M(i,j,t)$. The vector $E(t)$ with elements $E(i,t)$ is called the emigration vector at time t. The number of immigrants to country i between t and $t + 1$, denoted $I(i,t)$, is $I(i,t) = \Sigma_h M(h,i,t)$. The vector $I(t)$ with elements $I(i,t)$ is called the immigration vector at time t. The number of net migrants to country i between t and $t + 1$, denoted $N(i,t)$, is $N(i,t) = I(i,t) - E(i,t)$. The vector $N(t)$ with elements $N(i,t)$ will be called the net migration vector at time t. Then populations of countries or regions can be projected as

$$P(i,t+1) = P(i,t) + N(i,t) + B(i,t) - D(i,t), \qquad [2]$$

where $B(i,t)$ is the number of births and $D(i,t)$ is the number of deaths projected for country i from separate models of fertility and mortality.

It has been assumed thus far that the elements of the migration matrix would be the predicted values $10^{\text{expected log(migrants)}}$ produced by the GLM, yielding a deterministic population projection. A stochastic (Monte Carlo) method would be to sample numerically from the distribution of residuals for given values of the independent variables. Then the number of migrants from an origin to a destination would become a random variable, and a population projection that incorporated the migration matrix would become a probabilistic ensemble of projections.

The migration matrix at time t depends on the indicator variables, which are estimated from data up to and including t. Given $P(t+1)$ from Eq. **2**, the simplest approach to computing the migration matrix at time $t + 1$ would be to keep the coefficients of the indicator variables constant and to update only the population vector $P(t + 1)$. A more sophisticated approach would be to

examine trends over time in the coefficients of the indicator variables and to extrapolate those trends forward to obtain updated coefficients for the indicator variables of future migration matrices.

To obtain an age-specific net migration vector, one could apply age-specific models of migration patterns for different countries and regions to $E(i,t)$ and to $I(i,t)$ (19–21) and combine that with the projected age-specific fertility and mortality vectors as in Eq. **2**.

This method of projecting international migration in combination with cohort-component methods of projecting fertility and mortality solves the two problems identified by the United Nations Population Division (1). First, the global sum of net migrants is guaranteed to be 0 when the numbers of emigrants, immigrants, and net migrants are derived from a migration matrix (22). Second, net emigration should not completely deplete the population of any sending country for realistic model parameters. In all models considered here, the exponents of the populations of origin and destination are positive and the predicted number of emigrants is a small fraction of the population of origin. Consequently, the projected number of emigrants from an origin declines to 0 as its population declines to 0. Likewise, if the destination's population declines, the models predict that fewer people will migrate there. This feature of the model depends on realistic parameter values and may not hold in all mathematically possible cases. If the coefficients of the model were unrealistic and gave an unrealistically large number of emigrants, then it is possible that the origin population could be depleted.

Open Research Questions. Many open questions remain. Why did the area of the destination influence the numbers of migrants much less than the area of origin? How well would the proposed models work for migration within a country, taking account of international migration? For a given origin-destination pair, were the residuals correlated over time or independent as the error term in the model assumed?

International migrants are mainly younger individuals of working age and their families. Migratory flows of elderly individuals may also be important. Would the fit of the models be improved by replacing total population size with a weighted average that emphasized age groups most prone to migrate, or by a simple index such as the proportion of the population age 20–34 years? Age-structure seems likely to matter to migration increasingly as all countries undergo population aging (23).

What is the long-run behavior of the projection model Eq. **2** assuming constant birth rates and death rates and constant coefficients in Eq. **1**? For example, when, if ever, does there exist a fixed vector P of population size by country and a stable growth rate ρ such that $\lim_{t\to\infty} P(t)/\rho^t = P$? If this case arises, how does ρ depend on the parameters of the basic model Eq. **2**? What irreducibility conditions on the migration matrix assure the uniqueness of P? When the migration matrix is reducible, could different "ergodic sets" of countries (sets of countries linked through migration flows) have different fixed vectors P of population size by country and different stable growth rates ρ? Under what conditions on the coefficients of log(ppnorig) and log(ppndest) can country populations snowball to infinity (in finite time or with infinite time) or vanish? How sensitive to the initial conditions $P(i,0)$ are each country's proportion of world population $P(i,t)/\Sigma_h P(h,t)$ for large t, where the summation runs over all countries? How sensitive are country-specific population growth rates $P(i,t+1)/P(i,t)$, for large t, to initial conditions? In short, what ergodic theorems hold (24)?

Methods

Data. All 43,653 data records are provided in Dataset S1. Each record contains 12 variables: a unique serial number, the year in the Western calendar (1960–2004), the name of the country or region of origin of migrants, the log population of the origin in that year, the name of the country or region of destination of migrants, the log population of the destination in that year, the log number of migrants from origin to destination in that year, the log area (square kilometers) of the origin, the log area (square kilometers) of the destination, the log great circle distance (kilometers) from the capital of the origin to the capital of the destination, the source of the migration data, and "neighbor" (see *SI Text*). Records for which the value of any variable was missing were excluded.

Each country's definitions of what constituted a migrant, of the origin or destination of a migrant, and of the accounting year were used (Table S4). Differences among definitions and in the effectiveness of collecting migration data led to hundreds of discrepancies when both the origin and the destination reported migration data in the same year. In *SI Text*, sources are listed and methods of collecting migration data are discussed.

Data Analysis. A GLM was fitted to a starting model with dependent variable log(migrants) and with all six basic independent variables [year minus 1985, log(ppnorig), log(areaorig), log(ppndest), log(areadest), and log(distance)] and all indicator variables (orig.indicator, dest.indicator, orig.is.datasource, dest.is-.datasource). The stepwise regression algorithm stepAIC was applied to this linear model to obtain a final model. Details of data management and statistical software are in *SI Text*.

ACKNOWLEDGMENTS. Earlier versions of this article were presented to the Technical Working Group on Long-Range Population Projections, Population Division, United Nations Headquarters, on June 30, 2003; the United Nations Population Division Seminar on October 21, 2004; the Stanford Workshop on Formal and Quantitative Demography on August 15, 2005; the Policy Research Division of the Population Council on January 23, 2006; and the Estimates and Projection Section of the Population Division of the United Nations on June 6, 2008. We thank for helpful comments the many participants in those presentations and Adam E. Cohen, Jakub Bijak, Joshua R. Goldstein, Ryuichi Kaneko, Ronald D. Lee, Jim Oeppen, and Hania Zlotnik. We thank Mr. and Mrs. William T. Golden and family for their hospitality during this work. Priscilla K. Rogerson assisted in preparing data and the manuscript. J. Bijak and R. Kaneko helpfully reviewed two versions of the article. This work was supported by U.S. National Science Foundation Grants DEB-9981552 and DMS-0443803.

1. United Nations Population Division (2003) *United Nations Population Projections to 2300. ESA/P/WP.184, 30 June 2003*. (UNPD, New York).
2. Fertig M, Schmidt CM (2001) *Aggregate-Level Migration Studies as a Tool for Forecasting Future Migration Streams*. In *International Migration*, ed Djajic S (Routledge, London), pp 110–136.
3. Howe N, Jackson R (2005) *Projecting Immigration: A Survey of the Current State of Practice and Theory. A Report of the CSIS Global Aging Initiative, with Contributions by Rebecca Strauss and Keisuke Nakashima* (Center for Strategic and International Studies, Washington, DC).
4. Bijak J (2008) PhD Dissertation. (Central European Forum for Migration and Population Research, Warsaw).
5. Raymer, J, Willekens F, eds (2008) *International Migration in Europe: Data, Models and Estimates* (Wiley, Chichester).
6. Rogers A (2008) Demographic modeling of the geography of migration and population: A multiregional perspective. *Geog Analysis* 40:276–296.
7. Massey DS, *et al.* (1993) Theories of international migration: A review and appraisal. *Popul Dev Rev* 19:431–466.
8. Massey DS, *et al.* (1998) *Worlds in Motion: Understanding International Migration at the End of the Millennium (IUSSP International Studies in Demography)* (Clarendon, Oxford).
9. Faist T (2000) *The Volume and Dynamics of International Migration and Transnational Social Spaces* (Oxford Univ Press, Oxford).
10. Ritchey PN (1976) Explanations of migration. *Annu Rev Sociol* 2:363–404.
11. Dorigo G, Tobler W (1983) Push-pull migration laws. *Ann Assoc Am Geogr* 73:1–17.
12. Zipf GK (1946) The P_1P_2/D hypothesis: On the inter-city movement of persons. *Am Sociol Rev* 11:677–686.
13. Zipf GK (1949) *Human Behavior and the Principle of Least Effort* (Addison Wesley, Cambridge MA).
14. McCullagh P, Nelder JA (1989) *Generalized Linear Models*. 2d ed (Chapman and Hall, London).
15. Schwarz GE (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464.
16. Technical Panel on Assumptions and Methods (2003) *Report to the Social Security Advisory Board* (Social Security Administration, Washington, DC).
17. Bijak J, Kupiszewska D, Kupiszewski M (2008) Replacement migration revisited: Simulations of the effects of selected population and labour market strategies for the ageing Europe, 2002–2052. *Popul Res Policy Rev* 27:321–342.
18. Raymer J (2008) in *International Migration in Europe: Data, Models and Estimates*, eds Raymer J, Willekens F (Wiley, Chichester), pp 209–234.
19. Rogers A, Castro LJ (1981) *Model migration schedules. Research Report-81–30* (International Institute for Applied Systems Analysis, Laxenburg, Austria).
20. Rogers A, Willekens F, Raymer J (2003) Imposing age and spatial structures on inadequate migration-flow datasets. *Prof Geogr* 55:56–69.
21. Rogers A, Castro LJ, Lea M (2004) Model migration schedules: Three alternative linear parameter estimation methods. *Mathematical Population Studies* 12:17–38.
22. Cohen JE (2008) Constant global population with demographic heterogeneity. *Demogr Res* 18:409–436.
23. Raymer J, Rogers A (2008) in *International Migration in Europe: Data, Models and Estimates*, eds Raymer J, Willekens F (Wiley, Chichester), pp 175–192.
24. Cohen JE (1979) Ergodic theorems of demography. *Bull Amer Math Soc NS* 1:275–295.

Supporting Information

Cohen *et al.* 10.1073/pnas.0808185105

SI Text

SI Results

Starting Model. The starting linear model was fitted (Table S1) with dependent variable logmigrants and with the six "basic" independent variables and all indicator variables (orig.indicator, dest.indicator, orig.is.datasource, dest.is.datasource). In this model, under the implausible assumption of independent observations and the false assumption of homoscedasticity, log area of destination had a coefficient that differed from 0 with $0.01 < P < 0.05$. All other variables (treating the indicator variables as four matrix blocks, not as vectors for individual countries) had coefficients that differed from zero with $P < 0.001$. Because the assumptions on which they are based are unjustified or incorrect, all p values are regarded as nominal rather than credible. Software calls were written in R. The tabulations of results are a "summary" of the output of the functions "lm" (R stats package) or "stepAIC" (R MASS package).

Models with More Independent Variables Than the Starting Model. A variable called "neighbor" was constructed to see whether geographical adjacency influenced the number of migrants. Two countries or other geographical units were defined to be geographically adjacent if it was possible (in principle, disregarding political or military barriers, and disregarding rivers but not oceans) to walk across a border from one to another. An adjacency matrix of 228 rows (labeled by countries of origin) and 195 columns (labeled by countries of destination) was filled with the element 1 if the corresponding row country and column country were geographically adjacent and with the element 0 otherwise. For each line of data giving the number of migrants from an origin to a destination, the value of the variable "neighbor" for that line was looked up in the adjacency matrix: neighbor(origin, destination) = 1 if origin and destination were geographically adjacent, = 0 otherwise. The addition of "neighbor" to the starting model increased multiple R^2 very slightly from 0.5693 to 0.5709. The stepwise elimination algorithm stepAIC ranked the variables of this enlarged model (based on the increment to BIC resulting from eliminating each variable in succession) in increasing order of importance as log(areadest) (least important), year, neighbor, log(ppndest), dest.is.datasource, orig.indicator, orig.is.datasource, log(areaorig), dest.indicator, log(distance) and log(ppnorig) (most important). Thus, "neighbor" ranked among the less important variables. Its coefficient indicated that being geographically adjacent increased the predicted number of migrants by a factor of $10^{0.2660910} = 1.8454$ when the influence of all other variables was taken into account. Thus, geographical adjacency less than doubled the predicted number of migrants.

The starting model and the final model allowed for multiplicative interactions of the basic variables on the original scale of measurement because, for example, log(ppnorig·ppndest) = log(ppnorig) + log(ppndest). Such products are captured by terms linear on the logarithmic scale. When we added to the starting model an indicator variable for all 228 origins (not only for the 8 origins from which we obtained data), we obtained a very substantially improved multiple R^2 but the estimated coefficients of the basic variables and indicator variables were large, apparently erratic, and uninterpretable. The results were similar when we added an indicator for all 195 destinations (not only for the 11 destinations from which we obtained data). The estimated coefficients from such apparently over-fitted models seemed not to provide a reliable basis for projecting numbers of migrants. The number of estimated parameters for the model that included indicators for all 228 origins was 265 (1 intercept, 6 area and population predictors plus year, 22 destination and destination-.is.data.source indicators, 8 origin.is.data.source indicators, and 228 origin indicators), and the number of estimated parameters for the model that included indicators for all 195 destinations was 229 (1 intercept, 6 area and population predictors plus year, 16 origin and origin.is.data.source indicators, 11 destination.is.data.source indicators, and 195 destination indicators). Both values were above the rule-of-thumb cutoff of the square root of the number of data points ($43653^{1/2} = 208.9$) for the recommended maximum number of independent variables in a linear model, indicating that the larger models are over-fitted.

Other models not reported in detail here had interactions between some or all of the "basic" variables, for example, between log(ppnorig) and log(ppndest). We were not able to interpret interaction terms such as log(ppnorig)·log(ppndest) and did not pursue such models.

We considered three models in greater detail. In the first such model, in addition to the independent variables in the starting model, log(ppnorig) interacted with both indicator variables for destinations, namely, dest.indicator and dest.is.datasource, and log(ppndest) interacted with both indicator variables for origins, namely, orig.indicator and orig.is.datasource. This model allowed the exponent of the population of origin to differ for each destination *per se* and each destination as a data source. It allowed the exponent of the population of destination to differ for each origin *per se* and for each origin as a data source.

The addition of these 38 independent variables raised the multiple R^2 to 0.5861 compared with the starting model's multiple R^2 of 0.5693, an increase of <0.02 (Table S2). The coefficient of log(ppnorig) (that is, the exponent of the population of origin) rose to nearly 1.24 while the coefficient of log(ppndest) (the exponent of the population of destination) fell from positive to −0.64. As pointed out in the main Discussion, these values outside the interval from 0 to 1 could lead to undesirable behavior of the model. The coefficients for the destination indicators for Denmark and Germany rose to >6 and declined to below −6, respectively, corresponding to factors of one million and one millionth. Many of the estimated coefficients for orig.is.datasource and dest.is.datasource became even more extreme.

To the first model just considered, in the second model we also added the interactions between year minus 1985 and each of the indicator variables, orig.indicator, dest.indicator, orig.is.datasource and dest.is.datasource. These additional terms represented the possibility that each origin or destination (*per se* or as a data source) changed in time at a rate distinct from the time-associated global average rate of change. While the multiple R^2 increased slightly to 0.5975 (Table S2), some coefficients estimated for the "basic" variables became highly unstable. For example, the coefficient of log(ppnorig) rose to 8.14. All of the coefficients of orig.is.datasource fell below −29.

We also considered a third model that contained all of the independent variables of the starting model and in addition the interactions between year minus 1985 and all of the indicator variables. The addition of these 38 independent variables raised the multiple R^2 to 0.5817 compared with the starting model's multiple R^2 of 0.5693 (Table S2). None of the parameter estimates seemed unstable or unreasonable but the increase in

descriptive power of the GLM seemed small compared with the increase in the number of independent variables.

The three extensions of the starting model considered above slightly increased descriptive power (Table S2) at the price of large numbers of additional independent variables and, in some cases, of instability in the estimated coefficients.

Models with Fewer Independent Variables than the Starting Model. To see how much demographic and geographic variables mattered in accounting for the number of migrants, we fitted a model with none of the "basic" independent variables except year minus 1985. The independent variables in this model were year minus 1985, the four indicator variables, and the interactions between year minus 1985 and the indicator variables (for a total of 78 estimated parameters, including the intercept). For this model, multiple R^2 was 0.3371 and the estimated coefficients were not apparently unstable.

When we fitted a GLM that did not include year minus 1985 or the four indicator variables, but did include the five remaining "basic" independent variables, multiple R^2 was 0.4345 (Table S3). The five demographic and geographic variables (populations of origin and destination, areas of origin and destination, distance from origin to destination) better described variation in logmigrants than did the independent variables year minus 1985 together with the four indicator variables and the interactions between year minus 1985 and the four indicator variables (78 parameters including intercept).

The 2 models considered in the 2 previous paragraphs have disjoint sets of independent variables and the same dependent variable log(migrants). The union of these 2 disjoint sets of independent variables was considered in the third model described above. When the interactions between year minus 1985 and the four indicator variables were added to the starting model, multiple R^2 was 0.5817 (Table S2), which is considerably less than $0.3371 + 0.4345 = 0.7716$. The demographic and geographic "basic" variables were not orthogonal to year and the indicator variables. Both kinds of independent variables contributed substantially to the fits achieved by the starting and final models.

For each of the 29 time intervals considered in Table S3, the multiple R^2 ranked as follows according to the independent variables included: all variables in the starting model $>$ only "year minus 1985" omitted $>$ only indicators omitted $>$ "year minus 1985" and indicators omitted. The first and last inequalities are automatic. The middle inequality is unsurprising because there were many indicator variables and only one variable for year.

SI Discussion

These models assume that population sizes vary continuously and that time changes discretely. Both assumptions differ from reality. Real population sizes change by at least one individual and real time changes continuously. These differences in discretization between the model and reality are negligible when populations are large enough and numbers of migrants are small relative to populations.

SI Methods

Data. Eleven countries (Australia, Belgium, Canada, Denmark, Germany, Italy, the Netherlands, Spain, Sweden, the United Kingdom and the United States of America) reported 29735 records of migration in which the reporting country was the destination of the migrants, and eight countries (the above 11 excluding Canada, Spain and the United States of America) reported 13918 records of migration in which the reporting country was the origin of the migrants. Reported numbers of migrants from a country or region to itself were excluded. Records of 0 migrants were also excluded.

Population data were from the United Nations (1). The main source of migration data was ref. 2, but additional migration data came from refs. 3–5.

For most countries, land area was based on estimates from the Food and Agriculture Organization (FAO) compiled by the United Nations Statistics Division (http://unstats.un.org/unsd/cdb/cdb_advanced_data_extract.asp; accessed May 2008). For several countries where land area was not available but total area (including water bodies) was provided by the UN Statistical Division, total area was used instead of land area. Estimates of land area for Czechoslovakia, Yugoslavia and the USSR, which no longer exist as national entities, were taken from the *United Nations Demographic Yearbook 1990*, when all three existed as countries. The total land area of Central America was calculated by the United Nations Population Division. The total land area of the European Union was taken from the on-line *CIA World Factbook 2006* at www.cia.gov/library/publications/the-world-factbook (accessed August 20, 2006). For composites of multiple countries (including African Commonwealth; Bangladesh, India and Sri Lanka; Caribbean Commonwealth; and United Kingdom and Ireland), an area was computed as the sum of the land areas of the component countries.

Estimating the distance entailed certain assumptions. For Bolivia, which has two capital cities, La Paz and Sucre, Sucre was arbitrarily chosen. For Yemen, which moved its capital city to Sanaa after reunification of the country in 1990, the later city was arbitrarily chosen. For regions that included multiple countries, a capital of one of the countries was chosen to represent the region (for Bangladesh, India and Sri Lanka, New Delhi was chosen; for Oceania, the capital of Samoa was chosen; for Great Britain and Ireland, London was chosen). The capital was chosen to approximate both geographic and demographic centrality, but other choices could have been made. For each chosen city, a longitude and latitude were determined from public sources. Public sources frequently disagreed on the longitude and latitude (to a precision of degrees and minutes) of the selected cities. Where multiple sources were available, the most commonly used values were accepted for latitude and longitude. The longitude and latitude values were converted to radians (lon1, lat1) for city 1 and (lon2, lat2) for city 2 with south as negative and west as negative relative to Greenwich and entered into the following formula for the great-circle distance on a sphere:

$$\text{Distance (km)} = 6372.795*\arccos(\sin(\text{lat1})*\sin(\text{lat2}) + \cos(\text{lat1})*\cos(\text{lat2})*\cos(\text{lon2}-\text{lon1})).$$

The formula is exact for spherical geometry. The Earth is an oblate spheroid, with polar radius 6356.912 km and equatorial radius 6378.388 km. The ratio of the equatorial to polar radius is 1.0034. The formula used to calculate great-circle distance uses the average great-circle radius of the Earth. The error introduced by this approximation is likely to be $<0.34\%$. This error is smaller than that introduced by several other assumptions. In particular the error is probably smaller than the assumption that the great-circle distance between capital cities is the distance relevant for international migrants, particularly when countries adjoin like the USA and Mexico.

For a great majority of countries or regions, the latitude and longitude in radians were checked against a worksheet prepared independently by Uwe Deichmann at the World Bank and kindly sent to JEC November 3, 2005. In general, there was excellent agreement, to within the error of locating the center of the cities. After distances were calculated, they were compared with a database of distances at http://dss.ucsd.edu/~kgledits/capdist.html, accessed November 24, 2005, "Distance Between Capital Cities." Again, for the pairs of countries selected, the agreement between the online database and the distances cal-

culated here was good compared with the imprecision in the location of cities and the radius of the Earth.

Countries use varied systems to collect data on migration flows, e.g., residence permits (Canada, the United States), border collection (Australia, the United Kingdom) and national population registers (several European countries). These sources were built not to gather reliable statistics but for administrative reasons closely related to the control of international migration. Statistics derived from the issuance of residence permits, for instance, reflect administrative procedures and documents rather than actual entries. They provide information on legally resident foreigners but do not capture inflows or outflows of citizens, outflows of foreigners or the movement of undocumented migrants. Border statistics reflect actual moves but gathering information from large volumes of people subject to different degrees of control (depending on citizenship, port of entry, etc.) poses numerous challenges; for example, the status of persons arriving and departing is based on documents (passports, visas) which often do not reflect their actual stay. Population registers record arrivals and departures of both nationals and foreigners. In most countries, foreigners must have a valid residence permit to register; thus, in principle, undocumented migrants are not included in statistics based on registers. However, this regulation is not strictly applied in many countries. Those in charge of registration may not be fully apprised of the legal requirements to be met for foreigners to register. Whether foreigners are recorded or not often depends on the type of accommodation they occupy, rather than on their legal status: those settling in normal housing usually register, while those staying in government hostels or other group residence may not. In fact, population registers have been used in various European countries to estimate the magnitude of undocumented migration.[†] Therefore, population registers are the most comprehensive sources of information on international migration flows. Their main drawback is that the rules for registration and deregistration vary considerably among countries (Table S4).

Not all of the information from registers and other administrative sources is published. The publications and secondary data sources available often provide information on the entries and exits of foreigners only. Among the countries included in this study, only Germany, Sweden and the United Kingdom publish information on the movement of nationals. In the German case, included among nationals are individuals of German origin (*Aussiedler*) "repatriating" to Germany.

Countries differ in the criteria they use to classify migrants. Some countries (the Netherlands, Denmark) classify migrants by country of citizenship. Others (Australia, Canada, United States of America) classify migrants by country of birth, not country or region of origin or destination. However, more and more countries are publishing data by origin and destination, so comparability should improve in the future.

Most countries lack a system to register migratory flows continuously or do not publish the information that emanates from it. The countries that generated the data are all in the developed world, and most are members of the European Union. These are currently among the few countries in the world that record flows of people entering and leaving the country. On 11 July 2007, the European Parliament adopted a regulation intended to improve and harmonize its migration registration systems (http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:199:0023:0029:EN:PDF). This regulation postdates the data analyzed here.

Efforts are under way e.g., in Latin America, Eastern Asia and Eastern Europe to improve the availability of data on international migration flows. Information for several Central American and various Asian countries is available on the web (for instance, *Sistema de Información Estadística sobre las Migraciones en Mesoamérica*–SIEMMES at http://163.178.140.43, accessed June 14, 2008). However, the quality and completeness of the data in most of these countries are still unsatisfactory.

Origins were not necessarily mutually exclusive. For example, the European Union was identified as an origin along with countries that are members of the European Union. The United Kingdom was named as an origin along with the United Kingdom and Ireland as an origin. Similar overlaps occurred among the destinations. Moreover, not all origins or destinations existed as countries throughout 1960–2004, such as Yugoslavia and Bosnia-Herzegovina.

Data Analysis. Data were arranged using Microsoft Excel 2002 SP3 and were analyzed statistically using R, Version 2.6.1, a free open-source statistical analysis system. The function stepAIC selects a linear model generated by the function lm from a specified hierarchy of linear models using a penalty function that rewards goodness of fit and penalizes the number of parameters fitted to obtain that fit. Because of the large number of data points, we used the Bayesian Information Criterion (15), which sets the multiple of the number of degrees of freedom used for the penalty to $k = \ln(43653) = 10.684$, rather than the original Akaike Information Criterion, which sets the multiple of the number of degrees of freedom used for the penalty to $k = 2$.

Four indicator variables were matrices of 43653 rows. The matrix orig.indicator had 8 columns, one for each country that reported numbers of emigrants. For example, orig.indicator$Australia had 1 in data records where Australia was the origin, even when that record's migration data were reported by another country, e.g., U.K. The 11-column matrix dest.indicator similarly specified migrants' destinations. The 8-column matrix orig.is.datasource specified if a country reported itself as the origin. For example, in orig.is.datasource$Australia, an element was 1 if Australia was the origin and Australia reported the migration data in this data record; if either of these conditions failed, orig.is.datasource$Australia was 0. The 11-column matrix dest.is.datasource specified which country reported itself as the destination.

With one exception, the multiple R^2 is used throughout the article. For comparing models with varying numbers of variables, the adjusted R^2 could be used, where $R^2_{\text{adj}} = 1 - (1 - R^2)(n - 1)/(n - k - 1)$, n being sample size and k being the number of variables (without the constant). Here, $n = 43{,}653$ and for the starting model $k = 43$, so the maximum of $(n - 1)/(n - k - 1)$ among the models considered in the main article is 1.000986, which is trivially different from 1 considering the range of variation of R^2. Consequently, we used the multiple R^2.

Table 1 omitted the estimates for dest.is.datasource for the United States of America because the sum of all of the dest.is.datasource vectors for individual reporting countries was necessarily equal to the constant vector used to estimate the intercept. One of the country vectors had to be dropped to avoid a singularity. However, the information in the vector for the United States of America entered the overall averages for this indicator and was therefore reflected in the remaining estimates.

A plot of Cook's distance versus leverage revealed no outlying data points that unduly influenced the fit of the model (Fig. S1(b)).

Do the Data or the Methods Produce the Fit? Does the final model's multiple R^2 reflect over-fitting of too many independent variables? The data could be fitted perfectly if the model had as many independent variables as data points. A rule of thumb that a linear model should not have more independent variables than the square root of the number of data points is reassuring

[†]Recaño J, Domingo A, XXV Population Conference of the International Union for the Scientific Study of Population (IUSSP), July 18–23, 2005, Tours, France.

because $(43653)^{1/2} = 208.9$ whereas the final model has 44 independent variables.

For a more definitive answer, in each of 100 simulations, the values of the dependent variable log(migrants) were independently and randomly permuted. This randomized version of log(migrants) was then fitted to the final model using the unmodified data for the independent variables. From each such fit, the multiple R^2 was recorded. (The adjusted R^2 was always smaller by definition.)

Parameter Stability: How Much of the Past Is Relevant to the Future? To examine how coefficients varied as a function of the time interval from which data were drawn and as a function of the variables included in the model, the starting model and three subsets of its variables were fitted to temporal subsets of the data selected in four different ways. The starting model differs from the final model only in including the independent variable log(areadest).

For each of four subsets of variables [namely, (*i*) all variables; (*ii*) "year minus 1985" omitted; (*iii*) indicator variables omitted, and (*iv*) "year minus 1985" and indicator variables omitted], four sets of time intervals were considered. In total, there were 29 time intervals: (*i*) fixed initial year 1960 and moving terminal year from 1984 to 2004 in 5-year steps; (*ii*) five-year nonoverlapping tranches 1960–1964, 1965–1969, …, 2000–2004; (*iii*) overlapping 10-year tranches 1955–1964 (no data were available 1955–1959 so this first tranche covered five years only), 1960–1969, 1965–1974, …, 1995–2004; and (*iv*) intervals with initial year ranging from 1960 to 1985 in five-year steps and fixed terminal year 2004.

For each subset of variables and for each time interval, seven numbers were recorded in Table S3: the intercept, the coefficients of log(ppnorig), log(areaorig), log(ppndest), log(areadest), and log(distance), and the multiple R^2. Where "year minus 1985" was not excluded, its coefficient was also recorded.

1. United Nations (2005) *World Population Prospects: The 2004 Revision* (United Nations, New York).
2. United Nations (2006) *International Migration to and from Selected Countries* (POP/DB/MIG/FL/Rev.2005).
3. Eurostat (2000) *European Social Statistics. Migration. 2000 Edition* (Eurostat, Luxembourg).
4. Migration Policy Institute (2004) *Migration Information Source.* Global Data Center. Available at www.migrationinformation.org. Accessed December 2004.
5. United Nations Statistics Division (2004) Demographic Yearbook Database. Available at unstats.un.org/unsd/demographic/products/dyb/dyb2.htm. Accessed December 2004.

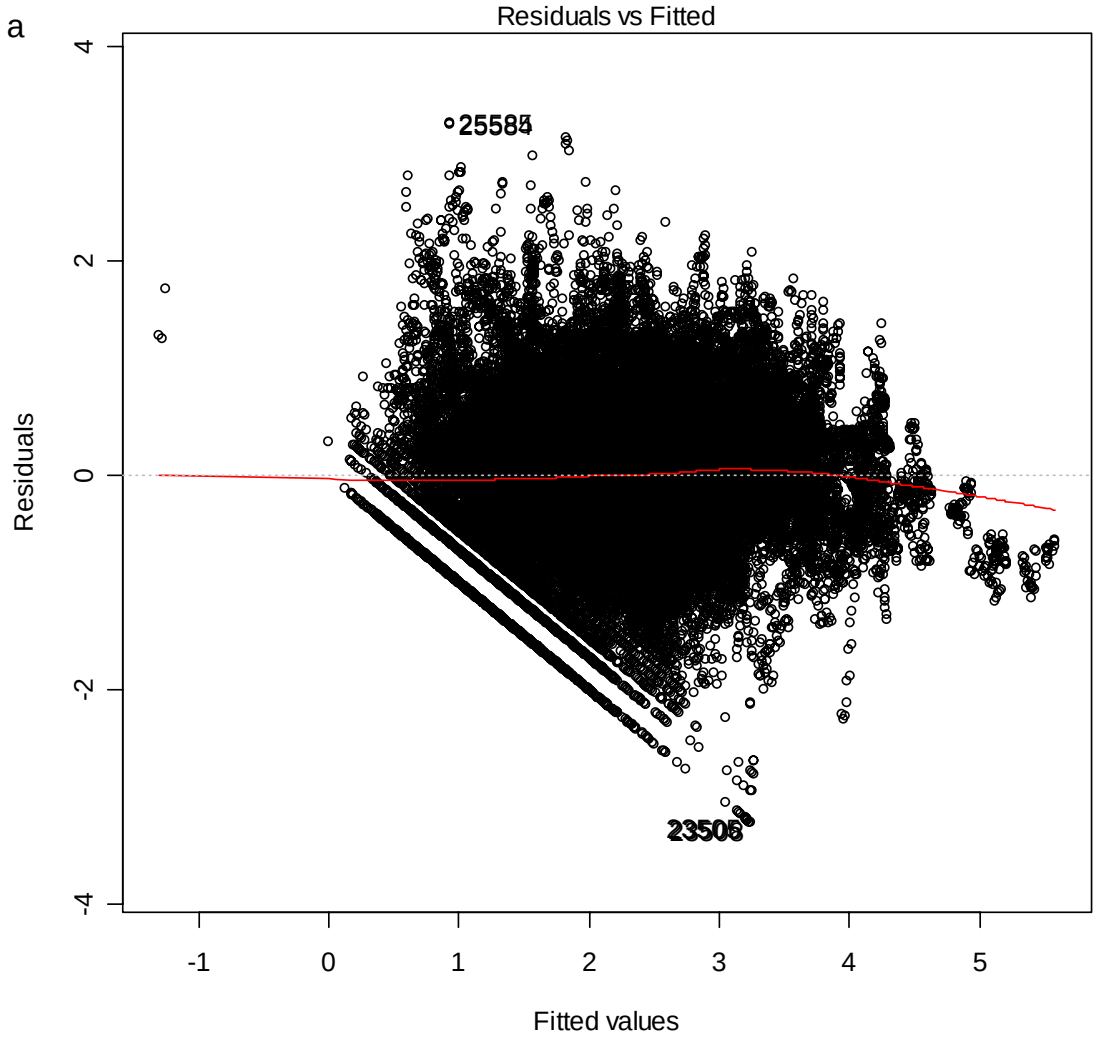

Fig. S1. Regression diagnostics for the ''final'' model (Table 1). (*a*) Residuals as a function of the fitted value of log number of migrants. (*b*) Cook's distance versus leverage: all points fell below the line labeled ''1'' so none was identified as an outlier.

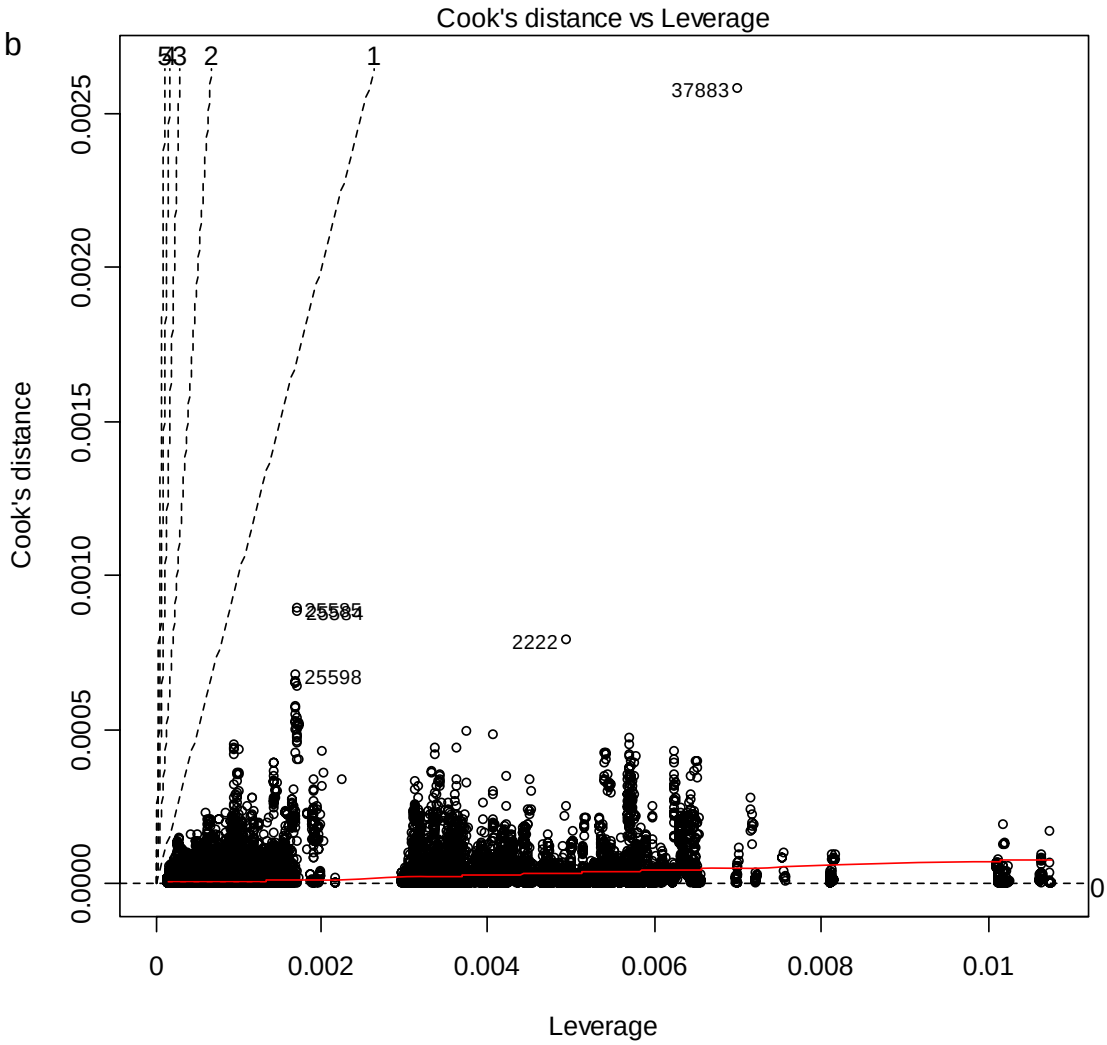

Figure S1. (continued)

Table S1. Starting model

```
Call:
lm(formula = logmigrants ~ I(year - 1985) + logppnorig + logareaorig +
    logppndest + logareadest + logdistance + orig.indicator +
    dest.indicator + orig.is.datasource + dest.is.datasource)

Residuals:
      Min        1Q    Median        3Q       Max
-3.245622 -0.435633  0.004541  0.441538  3.293094

Coefficients: (1 not defined because of singularities)
                                          Estimate     SE     t value  Pr(>|t|)
(Intercept)                             -2.4756833  0.0904315 -27.376  < 2e-16 ***
I(year - 1985)                           0.0017356  0.0003197   5.429 5.69e-08 ***
logppnorig                               0.8631499  0.0083278 103.647  < 2e-16 ***
logareaorig                             -0.2102357  0.0065929 -31.888  < 2e-16 ***
logppndest                               0.3377718  0.0140278  24.079  < 2e-16 ***
logareadest                              0.0239225  0.0115069   2.079 0.037626 *
logdistance                             -0.9702149  0.0102759 -94.416  < 2e-16 ***
orig.indicatorAustralia                  1.1302088  0.0436250  25.907  < 2e-16 ***
orig.indicatorBelgium                   -0.2562171  0.0403891  -6.344 2.26e-10 ***
orig.indicatorDenmark                   -0.0445711  0.0409475  -1.088 0.276383
orig.indicatorGermany                    0.0693162  0.0408962   1.695 0.090096 .
orig.indicatorItaly                      0.1841866  0.0401293   4.590 4.45e-06 ***
orig.indicatorNetherlands                0.0244522  0.0408387   0.599 0.549342
orig.indicatorSweden                     0.1597280  0.0473343   3.374 0.000740 ***
orig.indicatorUnited Kingdom             0.2479750  0.0397223   6.243 4.34e-10 ***
dest.indicatorAustralia                  1.4041046  0.0579072  24.248  < 2e-16 ***
dest.indicatorBelgium                    0.1489444  0.0530437   2.808 0.004988 **
dest.indicatorCanada                     0.8247913  0.0480852  17.153  < 2e-16 ***
dest.indicatorDenmark                    0.2636449  0.0523936   5.032 4.87e-07 ***
dest.indicatorGermany                    0.5996103  0.0505373  11.865  < 2e-16 ***
dest.indicatorItaly                      0.7664657  0.0496232  15.446  < 2e-16 ***
dest.indicatorNetherlands                0.5003483  0.0517456   9.669  < 2e-16 ***
dest.indicatorSpain                      0.6420857  0.0469944  13.663  < 2e-16 ***
dest.indicatorSweden                     0.2413032  0.0698006   3.457 0.000547 ***
dest.indicatorUnited Kingdom             0.6416269  0.0495353  12.953  < 2e-16 ***
dest.indicatorUnited States of America   1.1356594  0.0459375  24.722  < 2e-16 ***
orig.is.datasourceAustralia             -0.3000476  0.0633136  -4.739 2.15e-06 ***
orig.is.datasourceBelgium                0.4600040  0.0625941   7.349 2.03e-13 ***
orig.is.datasourceDenmark                0.2354601  0.0642053   3.667 0.000245 ***
orig.is.datasourceGermany                0.4853263  0.0612965   7.918 2.48e-15 ***
orig.is.datasourceItaly                 -0.4767394  0.0629507  -7.573 3.71e-14 ***
orig.is.datasourceNetherlands            0.2118823  0.0644958   3.285 0.001020 **
orig.is.datasourceSweden                -0.0734164  0.0657672  -1.116 0.264297
orig.is.datasourceUnited Kingdom         1.3506823  0.0681858  19.809  < 2e-16 ***
dest.is.datasourceAustralia             -0.0333185  0.0718665  -0.464 0.642925
dest.is.datasourceBelgium                0.5482645  0.0716074   7.657 1.95e-14 ***
dest.is.datasourceCanada                 0.1462815  0.0637547   2.294 0.021770 *
dest.is.datasourceDenmark                0.2685861  0.0720724   3.727 0.000194 ***
dest.is.datasourceGermany                0.5659672  0.0674291   8.394  < 2e-16 ***
dest.is.datasourceItaly                 -0.2357825  0.0687124  -3.431 0.000601 ***
dest.is.datasourceNetherlands            0.4557637  0.0718021   6.347 2.21e-10 ***
dest.is.datasourceSpain                 -0.2291943  0.0677617  -3.382 0.000719 ***
dest.is.datasourceSweden                 0.1273124  0.0830541   1.533 0.125311
dest.is.datasourceUnited Kingdom         1.4992516  0.0761564  19.686  < 2e-16 ***
dest.is.datasourceUnited States of America         NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6957 on 43609 degrees of freedom
Multiple R^2: 0.5693,     Adjusted R^2: 0.5689
F statistic:  1341 on 43 and 43609 DF,  P value: < 2.2e-16
```

The dependent variable is logmigrants. The independent variables are year minus 1985, logppnorig, logareaorig, logppndest, logareadest, logdistance, orig.indicator, dest.indicator, orig.is.datasource, and dest.is.datasource. Residuals are observed logmigrants minus expected logmigrants based on the fitted model.

Table S2. Multiple R^2 of the starting model and the three extensions of it

	No interactions of time with indicator variables	Interactions of time with indicator variables
No interactions of origin population with destination indicator variables or of destination population with origin indicator variables	0.5693 [starting model (Table S1)]	0.5817 (third additional model)
Interactions of origin population with destination indicator variables and of destination population with origin indicator variables	0.5861 (first additional model)	0.5975 (second additional model)

Table S3. Parameters in the starting model for varying subsets of independent variables and for data from varying intervals of time. Subsets of variables: (a) All variables. (b) Only ‘year minus 1985’ omitted. (c) Only indicators omitted. (d) ‘year minus 1985’ and indicators omitted. Time intervals: (1) Fixed initial year 1960, final year moving from 1984 to 2004. (2) Five-year tranches, moving from 1960-1964 to 2000-2004. (3) Ten- year tranches, moving from 1955-1964 to 1995-2004. (4) Moving initial year from 1960 to 1985, fixed terminal year 2004. The number of data lines in each time interval is given for (a) All variables, and is not repeated for the other subsets of variables.

(a) All variables.

(1) Fixed initial year 1960, final year moving from 1984 to 2004.

Moving.terminal.year	1984	1989	1994	1999	2004
Number of data lines	18389	24262	30930	39522	43653
(Intercept)	-1.5692	-1.7743	-1.9703	-2.3783	-2.4757
year-1985	-0.0040	-0.0024	0.0002	0.0002	0.0017
logppnorig	0.8291	0.8482	0.8551	0.8582	0.8631
logareaorig	-0.2364	-0.2384	-0.2309	-0.2169	-0.2102
logppndest	0.2314	0.2419	0.2683	0.3262	0.3378
logareadest	0.0547	0.0563	0.0424	0.0277	0.0239
logdistance	-0.9454	-0.9458	-0.9443	-0.9628	-0.9702
Moving.terminal.year.R^2	0.5635	0.5709	0.5730	0.5733	0.5693

(2) Five-year tranches, moving from 1960-1964 to 2000-2004.

5.year.tranche	1964	1969	1974	1979	1984	1989	1994	1999	2004
Number of data lines	1788	2422	3977	4740	5462	5873	6668	8592	4131
(Intercept)	0.6396	0.0603	-2.0670	-1.4210	-1.6516	-2.4278	-2.6442	-3.4927	-3.6341
year-1985	0.0245	0.0295	-0.0119	-0.0067	-0.0177	0.0234	-0.0004	-0.0025	0.0308
logppnorig	0.4060	0.5945	0.8875	0.8628	0.8470	0.8991	0.8659	0.8544	0.8215
logareaorig	-0.0434	-0.1480	-0.2727	-0.2736	-0.2348	-0.2435	-0.2054	-0.1658	-0.1384
logppndest	0.0785	0.2117	0.3389	0.2398	0.2027	0.2820	0.3682	0.5151	0.4856
logareadest	0.1055	0.0559	0.0119	0.0528	0.0599	0.0470	-0.0148	-0.0597	-0.0701
logdistance	-0.5583	-0.7514	-1.0990	-1.0465	-0.9198	-0.9487	-0.9292	-0.9699	-0.9450
5.year.tranche.R^2	0.5589	0.6859	0.5743	0.5696	0.5600	0.6082	0.6046	0.6072	0.6373

(3) Ten- year tranches, moving from 1955-1964 to 1995-2004.

10.year.tranche	1964	1969	1974	1979	1984	1989	1994	1999	2004
Number of data lines	1788	4210	6399	8717	10202	11335	12541	15260	12723
(Intercept)	0.6396	0.0848	-1.7217	-1.7296	-1.5223	-2.0310	-2.6079	-3.1522	-3.7839
year-1985	0.0245	0.0203	-0.0084	-0.0081	-0.0017	0.0023	0.0098	-0.0070	0.0254
logppnorig	0.4060	0.5286	0.8270	0.8788	0.8541	0.8739	0.8837	0.8650	0.8493
logareaorig	-0.0434	-0.1078	-0.2393	-0.2744	-0.2537	-0.2388	-0.2222	-0.1807	-0.1573
logppndest	0.0785	0.1743	0.2930	0.2847	0.2233	0.2427	0.3305	0.4551	0.5131
logareadest	0.1055	0.0707	0.0287	0.0348	0.0553	0.0603	0.0137	-0.0309	-0.0644
logdistance	-0.5583	-0.6998	-0.9921	-1.0697	-0.9777	-0.9358	-0.9392	-0.9713	-0.9757
10.year.tranche.R^2	0.5589	0.6325	0.6025	0.5681	0.5625	0.5827	0.6009	0.5983	0.6059

(4) Moving initial year from 1960 to 1985, fixed terminal year 2004.

Moving.initial.year	1960	1965	1970	1975	1980	1985
Number of data lines	43653	41865	39443	35466	30726	25264
(Intercept)	-2.4757	-2.5530	-2.6551	-2.7354	-2.9372	-3.1787
year-1985	0.0017	0.0022	0.0038	0.0052	0.0060	0.0080
logppnorig	0.8631	0.8755	0.8796	0.8759	0.8760	0.8776
logareaorig	-0.2102	-0.2155	-0.2168	-0.2100	-0.1985	-0.1892
logppndest	0.3378	0.3492	0.3631	0.3671	0.3861	0.4244
logareadest	0.0239	0.0196	0.0153	0.0148	0.0072	-0.0119
logdistance	-0.9702	-0.9861	-0.9968	-0.9852	-0.9753	-0.9829
Moving.initial.year.R^2	0.5693	0.5737	0.5715	0.5748	0.5783	0.5877

(b) Only ‘year minus 1985’ omitted.

(1) Fixed initial year 1960, final year moving from 1984 to 2004.

Moving.terminal.year.noyear	1984	1989	1994	1999	2004
(Intercept)	-1.4725	-1.7102	-1.9767	-2.3838	-2.5360
logppnorig	0.8289	0.8471	0.8554	0.8585	0.8670
logareaorig	-0.2340	-0.2361	-0.2312	-0.2172	-0.2138
logppndest	0.2191	0.2316	0.2695	0.3274	0.3513
logareadest	0.0626	0.0634	0.0416	0.0269	0.0141
logdistance	-0.9514	-0.9499	-0.9439	-0.9624	-0.9671
Moving.terminal.year.noyear.R^2	0.5630	0.5706	0.5730	0.5733	0.5691

(2) Five-year tranches, moving from 1960-1964 to 2000-2004.

5.year.tranche.noyear	1964	1969	1974	1979	1984	1989	1994	1999	2004
(Intercept)	0.0388	-0.4972	-1.9124	-1.3619	-1.5829	-2.3981	-2.6471	-3.5210	-3.1545
logppnorig	0.4080	0.5992	0.8891	0.8624	0.8457	0.8999	0.8659	0.8542	0.8249
logareaorig	-0.0444	-0.1514	-0.2729	-0.2733	-0.2342	-0.2441	-0.2054	-0.1657	-0.1395
logppndest	0.0822	0.2165	0.3367	0.2388	0.1996	0.2858	0.3682	0.5148	0.4893
logareadest	0.1035	0.0520	0.0129	0.0533	0.0621	0.0449	-0.0147	-0.0595	-0.0726
logdistance	-0.5540	-0.7517	-1.0998	-1.0470	-0.9197	-0.9489	-0.9292	-0.9699	-0.9461
5.year.tranche.noyear.R^2	0.5576	0.6846	0.5741	0.5695	0.5594	0.6073	0.6046	0.6072	0.6364

(3) Ten- year tranches, moving from 1955-1964 to 1995-2004.

10.year.tranche.noyear	1964	1969	1974	1979	1984	1989	1994	1999	2004
(Intercept)	0.0388	-0.4128	-1.5908	-1.6259	-1.5065	-2.0379	-2.5806	-3.2146	-3.4969
logppnorig	0.4080	0.5346	0.8291	0.8789	0.8537	0.8744	0.8863	0.8637	0.8565
logareaorig	-0.0444	-0.1129	-0.2386	-0.2736	-0.2534	-0.2392	-0.2239	-0.1796	-0.1605
logppndest	0.0822	0.1849	0.2877	0.2800	0.2222	0.2442	0.3351	0.4528	0.5254
logareadest	0.1035	0.0619	0.0321	0.0374	0.0560	0.0591	0.0104	-0.0281	-0.0712
logdistance	-0.5540	-0.6868	-0.9944	-1.0719	-0.9778	-0.9357	-0.9399	-0.9723	-0.9798
10.year.tranche.noyear.R^2	0.5576	0.6298	0.6021	0.5677	0.5625	0.5827	0.6002	0.5979	0.6028

(4) Moving initial year from 1960 to 1985, fixed terminal year 2004.

Moving.initial.year.noyear	1960	1965	1970	1975	1980	1985
(Intercept)	-2.5360	-2.6142	-2.7371	-2.8045	-2.9618	-3.1598
logppnorig	0.8670	0.8802	0.8876	0.8852	0.8836	0.8845
logareaorig	-0.2138	-0.2194	-0.2226	-0.2163	-0.2037	-0.1934
logppndest	0.3513	0.3635	0.3833	0.3878	0.4019	0.4369
logareadest	0.0141	0.0092	0.0003	-0.0010	-0.0052	-0.0214
logdistance	-0.9671	-0.9836	-0.9945	-0.9837	-0.9745	-0.9830
Moving.initial.year.noyear.R^2	0.5691	0.5733	0.5705	0.5734	0.5770	0.5863

(c) Only indicators omitted.

(1) Fixed initial year 1960, final year moving from 1984 to 2004.

Moving.terminal.year.noindicator	1984	1989	1994	1999	2004
(Intercept)	-4.6175	-4.6936	-4.7396	-4.7681	-4.7955
year-1985	-0.0079	-0.0068	-0.0044	-0.0051	-0.0031
logppnorig	0.8756	0.8895	0.8886	0.8626	0.8609
logareaorig	-0.2332	-0.2306	-0.2152	-0.1849	-0.1767
logppndest	0.5415	0.5408	0.5447	0.5594	0.5582
logareadest	0.1410	0.1451	0.1431	0.1484	0.1604
logdistance	-0.7821	-0.7929	-0.7981	-0.8244	-0.8390
Moving.terminal.year.noindicator.R^2	0.4192	0.4317	0.4334	0.4328	0.4355

(2) Five-year tranches, moving from 1960-1964 to 2000-2004.

5.year.tranche.noindicator	1964	1969	1974	1979	1984	1989	1994	1999	2004
(Intercept)	-3.1777	-5.8939	-4.2080	-3.9826	-4.4213	-4.9961	-4.8138	-4.7483	-6.6734
year-1985	0.0319	0.0238	-0.0110	-0.0101	-0.0165	0.0205	-0.0305	-0.0129	0.1056
logppnorig	0.7191	0.9297	0.8331	0.8671	0.8846	0.9326	0.8788	0.7567	0.8670
logareaorig	-0.2745	-0.2914	-0.1709	-0.2397	-0.2301	-0.2283	-0.1563	-0.0603	-0.1343
logppndest	0.4562	0.6577	0.5131	0.5614	0.5306	0.5376	0.5697	0.6180	0.6231
logareadest	0.1754	0.1810	0.1753	0.0958	0.1074	0.1571	0.1264	0.1498	0.1553
logdistance	-0.4620	-0.5838	-0.9102	-0.9022	-0.7864	-0.8215	-0.8060	-0.9017	-0.9726
5.year.tranche.noindicator.R^2	0.3628	0.4393	0.4033	0.4491	0.4390	0.4734	0.4465	0.4392	0.4899

(3) Ten- year tranches, moving from 1955-1964 to 1995-2004.

10.year.tranche.noindicator	1964	1969	1974	1979	1984	1989	1994	1999	2004
(Intercept)	-3.1777	-5.0997	-5.1361	-4.0959	-4.1998	-4.6928	-5.0153	-4.7847	-5.3505
year-1985	0.0319	0.0107	-0.0097	-0.0106	-0.0074	-0.0023	0.0041	-0.0186	0.0282
logppnorig	0.7191	0.8398	0.8811	0.8527	0.8763	0.9097	0.9047	0.8101	0.7891
logareaorig	-0.2745	-0.2809	-0.2139	-0.2098	-0.2349	-0.2283	-0.1909	-0.1026	-0.0807
logppndest	0.4562	0.5711	0.5562	0.5405	0.5458	0.5360	0.5537	0.5983	0.6128
logareadest	0.1754	0.1838	0.1926	0.1308	0.1015	0.1340	0.1421	0.1407	0.1667
logdistance	-0.4620	-0.5383	-0.7961	-0.9034	-0.8385	-0.8065	-0.8134	-0.8624	-0.9205
10.year.tranche.noindicator.R^2	0.3628	0.4050	0.4108	0.4279	0.4433	0.4563	0.4576	0.4413	0.4529

(4) Moving initial year from 1960 to 1985, fixed terminal year 2004.

Moving.initial.year.noindicator	1960	1965	1970	1975	1980	1985
(Intercept)	-4.7955	-4.8089	-4.7075	-4.7743	-4.9000	-5.0065
year-1985	-0.0031	-0.0029	-0.0018	-0.0006	0.0006	0.0021
logppnorig	0.8609	0.8637	0.8564	0.8575	0.8540	0.8447
logareaorig	-0.1767	-0.1731	-0.1665	-0.1656	-0.1522	-0.1342
logppndest	0.5582	0.5647	0.5635	0.5688	0.5699	0.5784
logareadest	0.1604	0.1583	0.1528	0.1507	0.1571	0.1649
logdistance	-0.8390	-0.8558	-0.8697	-0.8651	-0.8589	-0.8716
Moving.initial.year.noindicator.R^2	0.4355	0.4402	0.4423	0.4473	0.4481	0.4514

(d) ‘year minus 1985’ and indicators omitted.

(1) Fixed initial year 1960, final year moving from 1984 to 2004.

Moving.terminal.year.noindicatororyeai	1984	1989	1994	1999	2004
(Intercept)	-4.4567	-4.5718	-4.6697	-4.6913	-4.7476
logppnorig	0.8677	0.8810	0.8814	0.8502	0.8531
logareaorig	-0.2227	-0.2197	-0.2065	-0.1714	-0.1683
logppndest	0.5282	0.5279	0.5351	0.5454	0.5503
logareadest	0.1502	0.1549	0.1510	0.1617	0.1669
logdistance	-0.7918	-0.8019	-0.8043	-0.8333	-0.8443
Moving.terminal.year.noindicatororyeai	0.4169	0.4293	0.4320	0.4303	0.4345

(2) Five-year tranches, moving from 1960-1964 to 2000-2004.

5.year.tranche.noindicatororyear	1964	1969	1974	1979	1984	1989	1994	1999	2004
(Intercept)	-3.9305	-6.3303	-4.0678	-3.8993	-4.3575	-4.9600	-5.0374	-4.8988	-5.1105
logppnorig	0.7239	0.9317	0.8343	0.8667	0.8838	0.9331	0.8783	0.7557	0.8683
logareaorig	-0.2772	-0.2929	-0.1708	-0.2393	-0.2299	-0.2288	-0.1552	-0.0595	-0.1309
logppndest	0.4562	0.6582	0.5133	0.5610	0.5298	0.5386	0.5682	0.6169	0.6015
logareadest	0.1748	0.1800	0.1741	0.0963	0.1073	0.1566	0.1290	0.1508	0.2015
logdistance	-0.4600	-0.5828	-0.9105	-0.9027	-0.7870	-0.8216	-0.8059	-0.9020	-0.9708
5.year.tranche.noindicatororyear.R^2	0.3606	0.4384	0.4031	0.4489	0.4385	0.4727	0.4448	0.4389	0.4771

(3) Ten- year tranches, moving from 1955-1964 to 1995-2004.

10.year.tranche.noindicatororyear	1964	1969	1974	1979	1984	1989	1994	1999	2004
(Intercept)	-3.9305	-5.3236	-4.9437	-3.9898	-4.1491	-4.6910	-4.9975	-4.9424	-4.9999
logppnorig	0.7239	0.8441	0.8780	0.8536	0.8746	0.9095	0.9052	0.8041	0.7875
logareaorig	-0.2772	-0.2845	-0.2096	-0.2089	-0.2339	-0.2280	-0.1914	-0.0971	-0.0796
logppndest	0.4562	0.5717	0.5482	0.5413	0.5439	0.5358	0.5543	0.5924	0.6052
logareadest	0.1748	0.1800	0.1967	0.1292	0.1029	0.1342	0.1414	0.1481	0.1822
logdistance	-0.4600	-0.5339	-0.8007	-0.9048	-0.8384	-0.8068	-0.8133	-0.8656	-0.9206
10.year.tranche.noindicatororyear.R^2	0.3606	0.4043	0.4103	0.4271	0.4429	0.4563	0.4575	0.4388	0.4488

(4) Moving initial year from 1960 to 1985, fixed terminal year 2004.

Moving.initial.year.noindicatororyear	1960	1965	1970	1975	1980	1985
(Intercept)	-4.7476	-4.7777	-4.6998	-4.7733	-4.8987	-4.9955
logppnorig	0.8531	0.8578	0.8536	0.8567	0.8545	0.8459
logareaorig	-0.1683	-0.1671	-0.1637	-0.1650	-0.1526	-0.1352
logppndest	0.5503	0.5593	0.5616	0.5683	0.5702	0.5789
logareadest	0.1669	0.1625	0.1546	0.1512	0.1569	0.1648
logdistance	-0.8443	-0.8591	-0.8711	-0.8654	-0.8587	-0.8711
Moving.initial.year.noindicatororyear.R^	0.4345	0.4394	0.4420	0.4473	0.4481	0.4513

Table S4. Data sources and definitions

Country	Type of source	Classification by country of	In-migrants duration of stay	Out-migrants duration of stay	Citizenship of migrants
Australia	Border collection	Birth	Permanent residence*	Permanent departures†	All
Belgium	Population register	Previous/intended residence	3 months or longer	One year or longer	Foreigners
Canada	Residence permits	Birth	Permanent residence*		Foreigners
Denmark	Population register	Citizenship	3 months or longer	Permanent departures	Foreigners
Germany‡	Population register	Previous/intended residence	3 months or longer	3 months or longer	All
Italy	Population register	Previous/intended residence	3 months or longer§	Permanent departures	All
Netherlands	Population register	Citizenship	4 months or longer¶	8 months or longer¶	Foreigners
Spain‡	Population register	Previous residence	3 months or longer§		All
Sweden	Population register	Previous/intended residence	1 year or longer	1 year or longer	All
U.K.	Border collection and survey	Previous/intended residence	1 year or longer	1 year or longer	All
U.S.	Residence permits	birth	Permanent residence*		Foreigners

*Includes persons who obtain permanent residence permits, regardless of their actual entry date and of their intended period of stay.
†Until 1984, data refer to former settlers departing. Since 1985, data refer to permanent departures.
‡German criteria for the duration of stay vary, depending on the regulations of the federal states (*Länder*). Migrants are required to notify the authorities each time they cross national boundaries. Thus the statistics report migrations rather than migrants.
§Foreigners intending to stay in the country for at least three months as well as citizens returning after having resided abroad.
¶Up to September 1994, included persons intending to stay for 6 months or longer and to leave for one year or longer.

Other Supporting Information Files

Table S3
Dataset S1