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 calculated 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 non-overlapping 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.

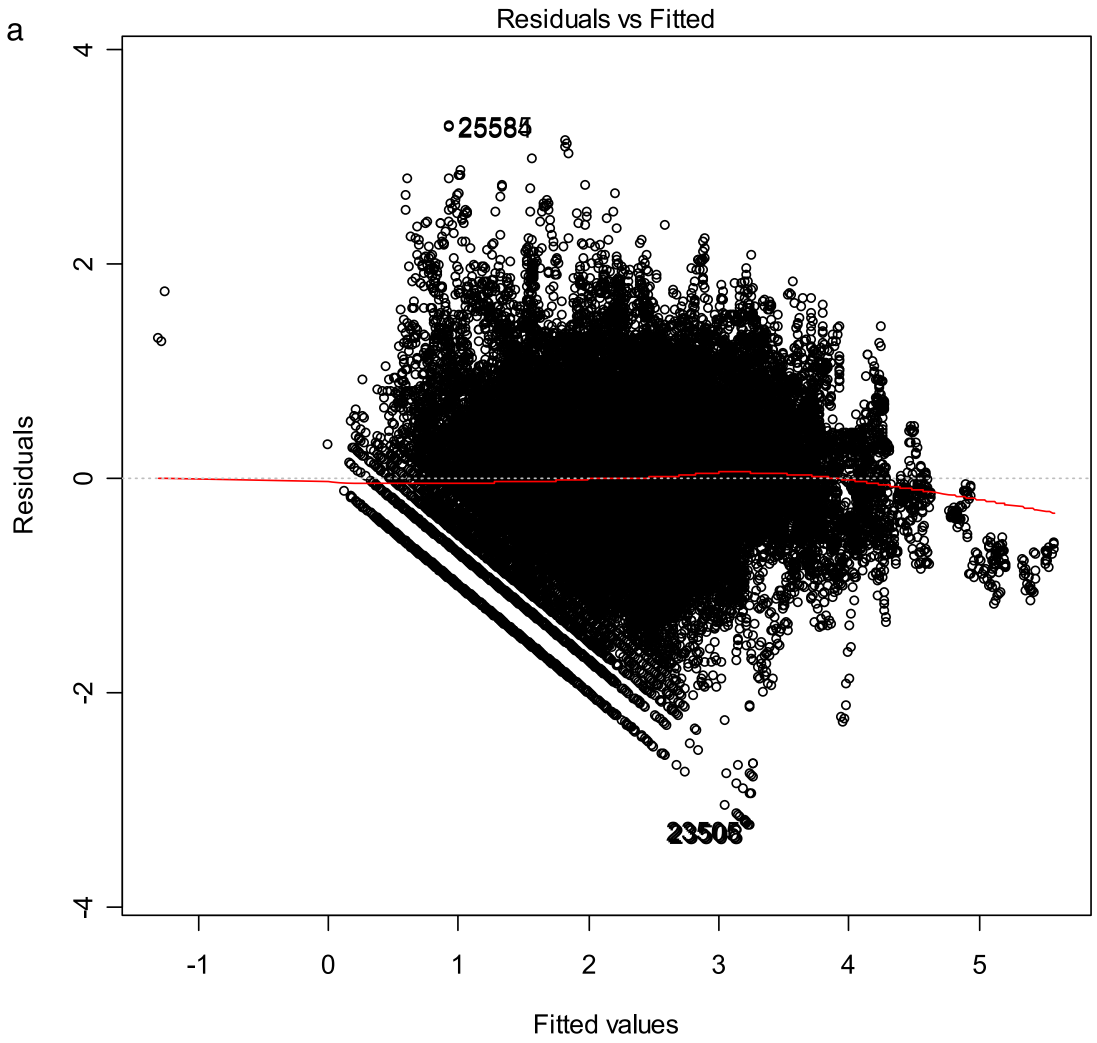

Fig. S1. Regression diagnostics for the ''final'' model (Table 1). (*a*) Residuals as a function of the fitted value of log number of migrants. (*b*) Cook's distance versus leverage: all points fell below the line labeled ''1'' so none was identified as an outlier.

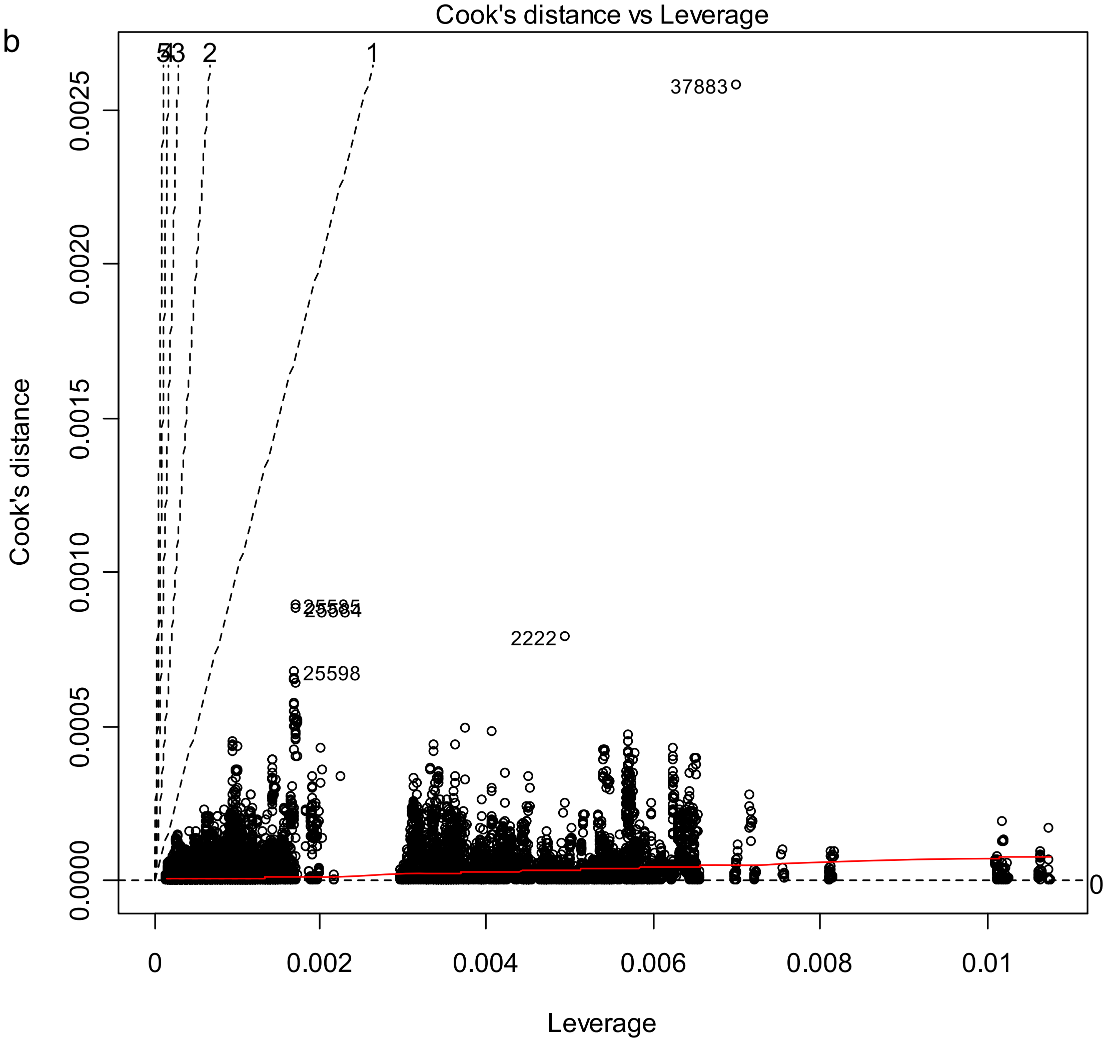

Figure S1. (continued)

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 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