Global Estimated Net Migration Grids by Decade: 1970-2000

May 2015

Socioeconomic Data and Applications Center (SEDAC)
Center for International Earth Science Information Network (CIESIN)
Columbia University
61 Route 9W
P.O. Box 1000
Palisades, NY 10964
Phone: 1 (845) 365-8920
FAX: 1 (845) 365-8922

Please address comments to SEDAC User Services:
http://sedac.ciesin.columbia.edu/help

This document outlines the basic methodology used to construct the Global Estimated Net Migration Grids by Decade: 1970-2000 data set. Please see the disclaimer and use restrictions at the end of the document, as well as the recommended citation below. We appreciate feedback regarding these data sets, such as suggestions, discovery of errors, difficulties in using the data, and format preferences.

Important note: Owing to data limitations, these net migration estimates contain many uncertainties. Uncertainties are highest at the individual grid cell level and for small areas, such as second administrative level units (e.g., counties, municipios). Therefore, these data are not recommended for use in analyses of small areas. They are best used, as they were in the original Foresight study, for analyses over broader areas such as ecosystems or climate zones. Users are encouraged to read the full documentation in order to understand the limitations.

Contents

I. Introduction

The Global Estimated Net Migration by Decade: 1970-2000 data set provides estimates of net migration over the three decades from 1970 to 2000. Net migration represents the number of people migrating into an area *minus* the number of people migrating out. The authors do not distinguish between internal and international migration. Because of the lack of globally consistent data on migration, indirect estimation methods were used. The authors relied on a combination of data on spatial population distribution for four time slices (1970, 1980, 1990, and 2000) and subnational rates of natural increase in order to derive estimates of net migration on a grid cell basis. Net migration was estimated by subtracting the population in time period 2 from the population in time period 1, and then subtracting the natural increase (births minus deaths).

The data were produced for the United Kingdom Foresight project on Migration and Global Environmental Change. The full report, *Estimating net migration by ecosystem and by decade: 1970-2010*, is available at http://www.ciesin.columbia.edu/documents/foresight-2011-oct.pdf. A peer reviewed open access article illustrating the application of the net migration grids to a research problem work can be found at:

> de Sherbinin, A., M. Levy, S.B. Adamo, K. MacManus, G. Yetman, V. Mara, L. Razafindrazay, B. Goodrich, T. Srebotnjak, C. Aichele, and L. Pistolesi. 2012. Migration and Risk: Net Migration in Marginal Ecosystems and Hazardous Areas. *Environmental Research Letters*. 7 045602. http://dx.doi.org/10.1088/1748-9326/7/4/045602.

Downloads

The data are available from the Global Estimates for Net Migration web page http://sedac.ciesin.columbia.edu/data/set/popdynamics-global-est-net-migration-grids-1970-2000 in GeoTIFF format for the decades 1970-1980, 1980-1990 and 1990-2000 in half-degree and 30 arc-second grid cell resolution. The data are stored in geographic coordinates of decimal degrees based on the World Geodetic System spheroid of 1984 (WGS84).

Downloaded files need to be uncompressed using either WinZip (Windows file compression utility) or a similar application before they can be opened by GIS software. Users should expect an increase in the size of downloaded data after decompression.

II. Data Processing and Methodology

The lack of subnational migration data for the forty-year time span considered by this project means that the authors needed to use indirect estimation methods to derive spatially explicit estimates of migration. Their basic methods can be summarized as follows, with details presented in the remainder of the section.

1. The authors utilized the History Database of the Global Environment, Version 3.1 (HYDE)[1] population grids for the years 1970, 1980, 1990, and 2000 to create one degree grids representing the rates of change in population for each decade. This makes optimal use of the HYDE data set, which is produced to provide a consistent decadal time series of population distribution over several centuries.
2. They applied those rates to the Global Rural-Urban Mapping Project, Version 1 (GRUMP)[2] population grids for 2000, producing "backcast" grids[3] to 1970, 1975, 1980, 1985, 1990, and 1995, and forecast grids to 2005 and 2010. This ensured that the global population data set with the greatest number of census inputs was utilized to spatially allocate population in one time slice, and also enabled the analysis to be conducted at the higher resolution of the GRUMP product (30 arc-second resolution for GRUMP vs. 5 arc-minute resolution for HYDE).
3. They adjusted the global grids to match country totals from the UN population estimates for the given year. This was done proportionally by calculating the ratio of the backcast and forecast grids summed by country for each time slice to the UN estimate for each country for that time slice and then applying that ratio to the population count grids for each year.
4. In order to estimate that portion of population growth that is due to natural increase (births *minus* deaths) for each grid cell in each decadal period, They applied subnational observed and imputed rates of natural increase (crude birth rates *minus* crude death rates) to the population grid at the beginning of each time to come up with decadal estimated natural increase. Similar to step 3 above, they adjusted the natural increase grids to match the UN estimates of natural increase at the country level.
5. Next, for each decade, they subtracted the population in time 1 (e.g., 1970) from the population in time 2 (e.g., 1980) in order to come up with the change in population in that grid cell, and then subtracted the natural increase in that grid cell (from step 4) in order to come up with an estimate of net migration for that grid cell in that decade. This is based on the population balancing equation:

   Population growth = (births - deaths) + (net migration)

   Which, when net migration is unknown, can be solved as follows:

   Net migration = population growth - (births - deaths)

---

[1] For further information on HYDE refer to Appendix B of the Foresight report *Estimating net migration by ecosystem and by decade: 1970-2010.*

[2] The GRUMP collection: http://sedac.ciesin.columbia.edu/data/collection/grump-v1

[3] The "backcast" grids "Global Population Count Grid Time Series Estimates" and "Global Population Density Grid Time Series Estimates".

Detailed Data and Methods

To conduct this modeling exercise the authors chose to use the GRUMP Population Count Grid, which represents an urban reallocation of the Gridded Population of the World, Version 3 (GPW) using night-time lights and other urban spatial extents and an algorithm that "pulls" population from larger administrative units out of rural areas and into urban areas. The alternative high resolution gridded population data product is Oak Ridge National Laboratory's Landscan 2008 (earlier versions are not available), which represents a modeled population surface at a 30 arc-second resolution. Although Landscan uses 8,205,582 census inputs for the United States, it only uses census data from only 79,590 administrative units outside the United States and then applies a multi-layered, dasymetric, spatial modeling approach to reallocate populations based on layers representing land use/land cover, high resolution satellite imagery, transportation networks, elevation, and slope, among others. The precise reallocation algorithm is not documented.

In contrast, GRUMP is based on population data from GPW, which uses 338,863 census units outside of the US[4], and is only lightly modeled using documented methods. It is worth noting, however, that the average population reporting unit size varies considerably by region, from 9,433 and 7,042 sq. km in Africa and Asia, respectively, to 5,744 sq. km in South America, 2,516 sq. km in Europe, and 1,094 sq. km in the rest of the Americas. This variability in the size of census unit is somewhat mitigated by the algorithm that pulls populations into urban areas, but nevertheless, in developing regions, and regions with large areas in sparsely populated drylands, there is generally less certainty regarding the spatial location of populations, and this will affect estimates of net migration.

To ensure that they had consistent rates of population change over the four decadal periods, they applied a grid representing the rate of population change per decade derived from the HYDE population grids for the years 1970, 1980, 1990, and 2000. The HYDE grids are adjusted at the country level to match the country totals from the UN Population Division's *World Population Prospects, 2008 Revision*. Although HYDE is distributed on a 5 arc-minute resolution, the rates were calculated on a one-degree resolution in order to average over a wider area and reduce the impact of decade-on-decade population variability inherent in higher resolution grid cells. A moving window was also applied in order fill in gaps in the HYDE-derived rates for areas that had no population in HYDE but observed population values in GRUMP.

One drawback of HYDE is that many small island states are not included in the data set[5], meaning that their coastal and island ecosystem estimates are not taking into account these countries. They have tallied net migration data from alternate sources for these islands[6].

The GRUMP Population Count Grid for the year 2000 was "backcast" to 1970, 1980, and 1990, and was projected to the year 2010 by multiplying the HYDE rates times the population grids. For the most part

[4] Summary information for GPWv3 input units can be found in Table A3 of the Foresight report *Estimating net migration by ecosystem and by decade: 1970-2010*. http://www.ciesin.columbia.edu/documents/foresight-2011-oct.pdf

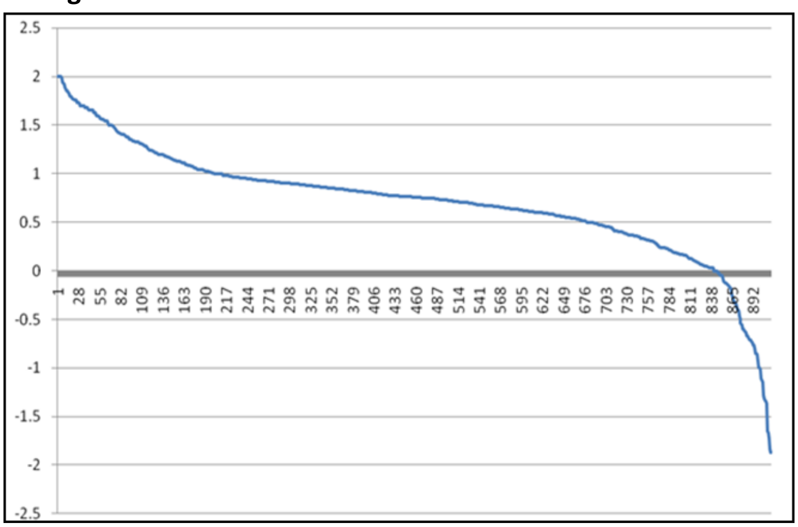
[5] A list of missing states is included at the end of Appendix B in the above report.

[6] See table A13 from the above report.

negative rates were used for backcasting and positive rates for forecasting, but in selected areas of depopulation over the course of each decade the sign for the rates was reversed. In each case they adjusted the gridded country totals so that they equal the UN *World Population Prospects, 2008 Revision* country population totals for each time period. In this way all population data were consistent with the UN *World Population Prospects, 2008 Revision,* which represents a harmonized time series of country-level demographic data.[7] A population change grid for each decade was derived by subtracting the population at the beginning of the time period (e.g., 1970) from the population at the end of the time period (e.g., 1980).

In a pilot effort, they applied national level rates of natural increase (crude birth rates minus crude death rates) from the *World Population Prospects* to population grids to derive decadal estimates of natural increase. However, this approach ignored the fact that there is substantial subnational variation in rates of natural increase (RNIs). Culling data on urban and rural crude birth and death rates (CBRs and CDRs) from the United Nations *Demographic Yearbooks*[8] published from 1970 to 2008, and deriving urban and RNIs (CBRs minus CDRs), they found a high degree of variation within countries. Figure 1 shows the ratio of urban to rural RNIs within the range of +2 to -2, which represents 85% of the country-year combinations for which they had observed data (900 out of 1,070 cases). There is no clustering around 1 RNI, which is what one would expect if there were no difference in urban and rural rates.

Figure 1. Ratio of Urban to Rural Rates of Natural Increase



Source: UN Demographic Yearbook data.

[7] The authors utilized year 2000 boundaries and country definitions for all processing steps. Countries that were separated in 1970 such as East and West Germany were treated as one entity; countries that were part of larger countries in the 1970s such as the republics of the former USSR and many Eastern European countries were treated as though they were separate entities throughout all four decades.

[8] The statistics presented in the Demographic Yearbook are national data provided by official statistical authorities unless otherwise indicated. The primary source of data for the Yearbook is a set of questionnaires sent annually by the United Nations Statistics Division to over 230 national statistical services and other appropriate government offices. Data reported on these questionnaires are supplemented, to the extent possible, with data taken from official national publications, official websites and through correspondence with national statistical services. In the interest of comparability, rates, ratios and percentages have been calculated by the Statistics Division of the United Nations, except for crude birth rate and crude death rate for some countries or areas as noted.

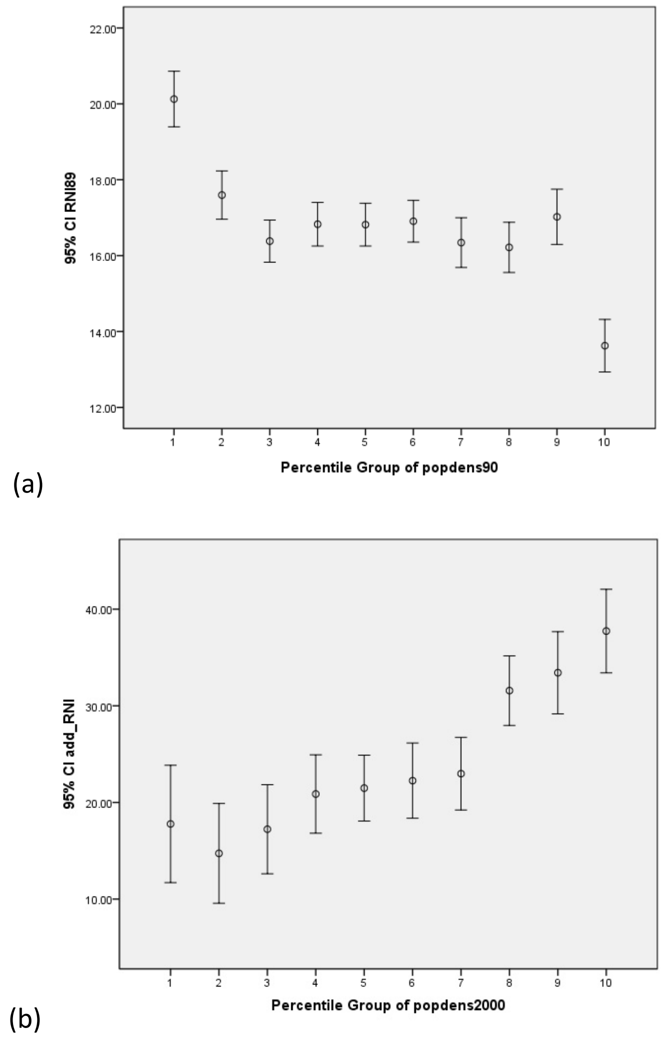
The authors hypothesized that RNIs can be predicted based on where a particular grid cell lies on an urban to rural gradient as measured by population density. They tested this hypothesis for subnational data on RNIs for two countries: China and the United States. For China data for 2,315 districts for 1989-90 from the CIESIN China Dimensions data collection[9] yielded a fairly clear gradient from higher RNIs in low density rural areas to lower RNIs in high density urban areas (Figure 2a). For the China data set, the mean RNI was 17 per 1,000 population, with a standard deviation of 5.2. For the US they used data for 3,194 counties and county equivalents from the US Census Bureau for the year 2000 and found that, contrary to China, RNIs tend to increase over the density gradient, rising from around 20 per 1,000 to more than 30 per 1,000 for the top three deciles in terms of population density (Figure 2b). For the US data set the mean RNI was 25 per 1,000 population with a standard deviation of 4.7.

The empirical data confirmed the authors' hunch that there is a systematic relationship between RNIs and population density, though that relationship varies by development level (Figure 3). They therefore felt that it was preferable to assume some level of subnational variation, even if population density is an imperfect predictor, rather than assume that RNIs are constant throughout a country. This presented a further challenge, however, because of the lack of a globally consistent database of urban and rural RNIs by country that covers the 40 year time period from 1970-2010. As a significant subcomponent of this project, they created a database of urban and rural CBRs and CDRs based on available data and imputation methods. They did this by compiling data on urban and rural CBRs and CDRs from the UN *Demographic Yearbooks* and the Demographic and Health Surveys (CBRs only), and then imputing the missing values. To impute missing values (more than 32,000 country-year urban and rural CBRs and CDRs), they combined 5,016 observed values with as many auxiliary variables as they could obtain that might help explain patterns of urban and rural birth and death rates. Two models were used, Multiple Imputation (mi) and Amelia. Although the results are subject to uncertainties and though the results were generally better for the CBRs than the CDRs (which proved more difficult to predict from available data), they feel this approach is better than ignoring subnational variation in RNIs. A total of 13 imputation runs were produced – eight runs from the multiple imputation package associated with the R statistical language and environment and five runs from the Amelia cross-sectional time series imputation package, which is also available for R.

[9] China Dimensions collection: http://sedac.ciesin.columbia.edu/data/collection/cddc

Figure 2. RNIs (y-axis) Across the Rural-to-Urban Population Density Gradient:

(a) China (1989-90) and (b) United States (2000)

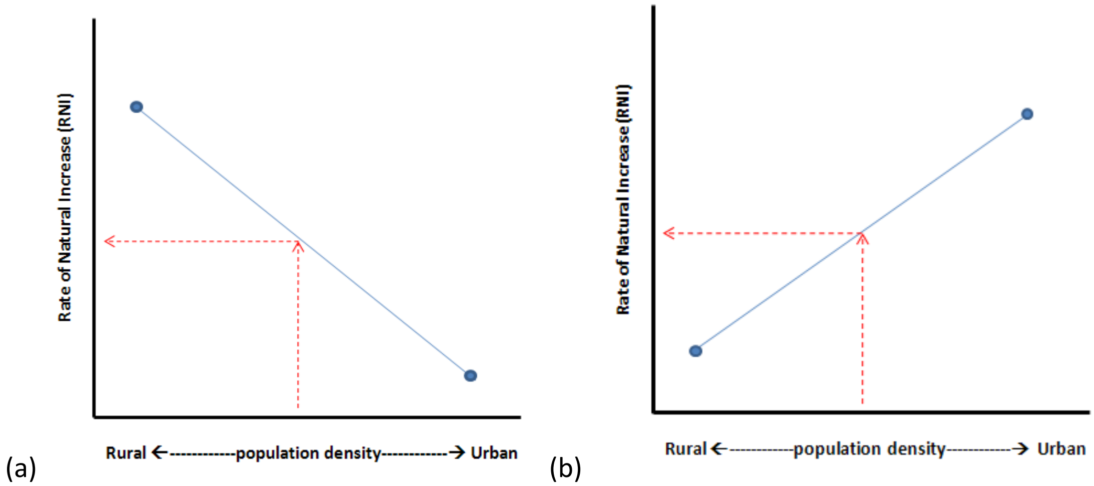


Rural ←-----------population density-----------→ Urban

The United States represented a special case of a demographically significant country with no observed urban/rural rates. Although the US does not report these data to the *Demographic Yearbook*, county level data on births and death rates by decade are available from the US Census Bureau. Given this special case, the authors replaced the imputed data for the US with estimated decadal rates of natural increase from the US Census Bureau. These are not truly "observed" data, in the sense of being based on a population registry, but they come very close. They averaged the rates across urban and rural US counties based on population density; the top three deciles in county-level population density were classified as urban based on natural breaks in the RNIs (Figure 2b).

Once they had completed the imputations, they began a series of processing steps to develop spatial estimates of net migration on a 30 arc-second grid. The remainder of this section details the processing steps they utilized in order to develop the subnational estimates of net migration.

Figure 3. Generalized Relationship Between Population Density and RNIs: (a) for most developing countries, and (b) for most developed countries



Note: The authors did not *a priori* assign slopes to countries; the slopes were based on the RNI data (observed and imputed) and observed population densities.

Step 1. The authors' approach assumed that there is a consistent relationship between RNI and population density, such that with population density information for any given grid cell one could derive the RNI based on the slope and intercept (Figure 3). In order to establish the slope and the intercept for this relationship between RNIs and population density, they already had average urban and rural RNIs, so they then calculated average urban and rural population densities by country in order to establish the points pinning both ends of the line in Figure 3. The GRUMP data set derives its urban extents from circa 1995 night-time lights satellite imagery for larger settlements, and for smaller settlements without night-time lights signatures it uses buffered points. For each country, using ArcMap 10, they calculated the average urban and rural population densities for 1995 based on the GRUMP delineation of urban extents. Using SPSS, they then used these densities and the average decadal urban and rural "RNIs" (converted to proportions that are multiplied by the decade start population to arrive at actual change in population over a decade) to obtain the slopes and intercepts for the relationship between population density and RNIs for the 1990s:

slope_1990s = (urban_rni_1990s-rural_rni_1990s)/ (urban_density1995-rural_density1995)

intercept_1990s = urban_rni_1990s - (slope_1990s*urban_density)

Step 2. Because urban extents for the other decades was not available, the slope was adjusted by decade using a "pseudo-slope" formula, as demonstrated for the decade of the 1970s:

pseudo_slope_1970s = (urban_RNI_1970s-rural_RNI_1970s)/rural_RNI_1970s

The pseudo-slope has as many properties as possible as the slope, in the absence of knowledge of the precise location on the x-axis of population density: a) it varies proportionally with the slope, and b) it has the same sign as the slope. In order to calculate the slope for the 1970s, 1980s, and the 2000s, they used the following formula, which in this instance calculates the slope for the 1970s:

slope_1970s = slope_1990s *(pseudo_slope_1970s/pseudo_slope_1990s)

This adjustment factor has the following desirable characteristics: a) if the slope needs to reverse sign (because the urban/rural relationship reverses) then the slope reverses sign; and b) the slope changes in the right direction (if it needs to steepen, it steepens; if it needs to flatten, it flattens). They did not change the intercept, but instead relied upon the intercepts for each imputation run from the 1990s.

Step 3. The next step was to create an RNI grid. For each grid cell, the RNI is derived from the population density in that grid cell. The generic formula was as follows:

rni _decadal_period = intercept + (slope_decadal_period * density_start_of_decade)

Or, as examples from the imputation runs:

rni_run1_1970s = intercept1_1990s + (slope1_1970s * density_1970s).

rni_run12_1990s = intercept12_1990s + (slope12_1990s * density_1990s)

Step 4. In this step, they multiplied the population counts grid and the RNI grid. At the pixel level they calculated the "implied" natural increase – that is the natural increase that a particular model run implies for that grid cell:

ni_pixel_implied_decade = rni_decadal_period * population_gridcell

Or, as examples from the imputation runs:

ni_pixel_implied_run3_1990s = rni_run3_1990s * population_gridcell_1990

Step 5. In this step they summed the natural increase in all grid cells to come up with a country total of natural increase, as follows:

country_ni_implied_decade = Σ (rni_decadal_period * population_start_of_decade)

Or as an example for imputation run 3 in the 1990s:

country_ni_implied3_1990s = Σ(rni_run3_1990s * population_1990)

Step 6. In this step they adjusted the pixel level natural increase (ni) estimates so that they total to the UN ni at the country level. First, the country level summed ni was compared to the ni reported for that country by the *World Population Prospects 2008* and the difference was calculated. Next, the absolute value of all pixels was summed at the country level, and a weight matrix was developed by dividing the absolute value of each pixel by the sum of the absolute value of all pixels in the country. The weights were then multiplied by the difference between the implied (or calculated) ni and the ni from the *World Population Prospects 2008* in order to produce a matrix of pixel-level adjustment factors. The adjustment factors were summed with the initial ni estimates to produce a matrix of UN adjusted natural increase. The generic formulas for this were as follows:

ni_diff_decade = UN_country_ni_decade - country_ni_implied_decade

country_sum_abs_ni_decade = Σ(abs(ni_pixel_implied_decade))

ni_pixel_weight_decade = abs(ni_pixel_implied_decade) / country_sum_abs_ni_decade

ni_pixel_adjustment_factor_decade = ni_pixel_weight_decade * ni_diff_decade

ni_pixel_adjusted_decade = ni_pixel_implied_decade + ni_pixel_adjustment_factor_decade

Or as an example, for imputation run 3 in the 1990s:

ni_diff_decade = country_ni_1990s - country_ni_implied_run3_1990s

country_sum_abs_ni_run3_1990s = Σ(abs(ni_pixel_implied_run3_1990s))

ni_weights_run3_1990s = abs(ni_pixel_implied_run3_1990s) / country_sum_abs_ni_run3_1990s

ni_pixel_adjustment_factor_run3_1990s = ni_pixel_weight_run3_1990s * ni_diff_run3_1990s

ni_pixel_adjusted_decade = ni_pixel_implied_run3_1990s + ni_pixel_adjustment_factor_run3_1990s

Step 7. The final step involved subtracting the decadal natural increase grids (based on the 13 imputation runs) from the decadal population change grid to arrive at a residual, and it is this residual that the authors are terming "net migration" at the pixel level, as follows:

nm_pixel_decade = pop_change_pixel_decade - ni_pixel_adjusted_decade

or as an example, for imputation run 3, 1990s:

nm_pixel_run3_1990s = pop_change_pixel_1990s - ni_pixel_final_run3_1990s

Through these methods they were able to estimate net migration for each decade for each grid cell based on 13 imputation runs. They further processed these runs in order to remove rounding errors by ecosystem, so that the global net migration totals for each decade summed to less than +/- 1 persons. With 13 runs, they were able to develop an average and a standard deviation of the model runs for net migration for each grid cell, which represents a "pseudo" error bar for their estimates. But they caution that the actual numbers represent net migration plus or minus some unknown error term per grid cell. Nevertheless, because of the methodology they followed for this work, the sum of net migration of all grid cells in any given country equals the total net migration per country according to the *World Population Prospects 2008*. They have validated that the sum of net migration on a country level is consistent with the UN estimates; so the only difference in spatial distribution in net migration at the subnational level is due to the differences in slopes and intercepts generated by the urban and rural RNIs from the imputation runs.

Although unable to precisely quantify the amount of error in their estimates, the authors were able to characterize the precision and accuracy of the data inputs. This assessment is found in Appendix E. Note that they could not fully assess the accuracy of the United Nations *World Population Prospects 2008* data set, and therefore any issues with those data (for example, errors in national decadal natural increase or net migration levels) will affect their results. The UN Population Division provides extensive documentation but given resource constraints they were not able to fully characterize the uncertainties for any country-decade combination, though they do assess frequency of censuses[10], which is an important underpinning of both their work and the UN data.

III. Acknowledgements

This work reflects the contributions of a large number of individuals. Alex de Sherbinin served as PI and lead author of the report, and Marc Levy served as co-PI and also led the development of the methodology. Susana Adamo led data collection for the imputation model inputs, with contributions from Cody Aichele and Linda Pistolesi, and also led the literature review and the results write up. Ben Goodrich (Post-Doctoral Researcher at the Applied Statistics Center within the Institute Social and Economic Research and Policy, Columbia University) and Tanja Srebotnjak (independent consultant) ran the imputation models for generating urban and rural crude birth rates and death rates and provided statistical advice. Greg Yetman led the geospatial processing with assistance from Kytt MacManus, Liana Razafindrazay, and Cody Aichele.

This project was funded by the United Kingdom Government Office for Science as part of the Foresight Project on Global Environmental Migration. Additional support for documentation and dissemination was provided by NASA under Contract NNG13HQ04C for the continued operation of the Socioeconomic Data and Applications Center (SEDAC).

IV. Disclaimer

[10] See Appendix E "Characterizing Precision and Accuracy of Data Inputs" of the Foresight report Estimating net migration by ecosystem and by decade: 1970-2010. http://www.ciesin.columbia.edu/documents/foresight-2011-oct.pdf

V. Use Constraints

Users are free to use, copy, distribute, transmit, and adapt the work for commercial and non-commercial purposes, without restriction, as long as clear attribution of the source is provided. Users of the data are encouraged to acknowledge CIESIN as the source used in the creation of any reports, publications, new data sets, derived products, or services resulting from their use. CIESIN also requests reprints of any publications acknowledging CIESIN as the source and requests notification of any redistribution efforts.

VI. Recommended Citations

Data Set:

de Sherbinin, A., M. Levy, S. Adamo, K. MacManus, G. Yetman, V. Mara, L.Razafindrazay, B. Goodrich, T. Srebotnjak, C. Aichele, and L. Pistolesi. 2015. Global Estimated Net Migration Grids by Decade: 1970-2000. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). http://dx.doi.org/10.7927/H4319SVC. Accessed DAY MONTH YEAR.

Scientific Publication:

de Sherbinin, A., M. Levy, S. Adamo, K. MacManus, G. Yetman, V. Mara, L.Razafindrazay, B. Goodrich, T. Srebotnjak, C. Aichele, and L. Pistolesi. 2012. Migration and Risk: Net Migration in Marginal Ecosystems and Hazardous Areas. *Environmental Research Letters 7(4): 045602*. http://iopscience.iop.org/1748-9326/7/4/045602.