Estimating Net Migration By Ecosystem and By Decade: 1970-2010

Final Report – September 2011

Provided to the UK Government's Foresight Environmental Migration Project

By the Center for International Earth Science Information Network (CIESIN),
The Earth Institute at Columbia University

Contents

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 Achille and Linda Pistolesi, and also led the literature review, the results write up, and Appendix D. 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, provided statistical advice, and co-authored Appendix C. Greg Yetman led the geospatial processing with assistance from Kytt MacManus, Liana Razafindrazay, and Cody Achille. Valentina Mara compiled the final tables and graphs in Appendices A and E. The authors are grateful for the comments of two external reviewers.

Executive Summary

This project sought to generate estimates of net migration, both domestic and international, by ecosystem over the four decades from 1970 to 2010. Because of the lack of globally consistent data on migration, indirect estimation methods were used. We relied on a combination of data on spatial population distribution for five time slices (1970, 1980, 1990, 2000, and 2010) and subnational rates of natural increase in order to derive estimates of net migration which were then summed by ecosystem. We ran 13 geospatial net migration estimation models based on outputs from the same number of imputation runs for urban and rural rates of natural increase. We took the average and standard deviation of the runs to produce the results described in Section IV and Annex A (see Maps A1-A4 for a global map depiction of the results by decade).

In summary, and noting that ecosystems are not mutually exclusive (the same grid cell can be counted for example in cultivated, island, and coastal ecosystems), we found that:

- Most out-migration occurs over large areas, reflecting its largely rural character, whereas areas of net in-migration are typically smaller which reflects its largely urban character.
- Coastal ecosystems (as defined circa 2000) have experienced the highest levels of net in-migration, with levels ranging from ~30m in the 1970s to 1980s to +82m in the 2000s.
- Inland Water ecosystems have experienced the second highest levels of net in-migration, with levels ranging from +23m in the 1980s to +53m in the 2000s
- Mountain, Forest, Cultivated, and Dryland ecosystems all show high levels of net out-migration, ranging from -12m to -43m across all decades. Mountain ecosystems have the highest net out-migration over the four decades, totaling -126m. The patterns across these ecosystems is consistent with global trends in rural-to-urban migration over the past 40 years.
- Considering their generally small populations, Island ecosystems have high levels of out-migration, ranging from -3m to -4.5m.
- The largest population countries such as China and India tend to drive global results for all the ecosystems found in those countries.
- There are large standard deviations for the Asia model runs, especially in the decade from 2000 to 2010. This is due to small variations in rates of natural increase generated by the model runs, which when multiplied by large populations results in large standard deviations.

There are a number of uncertainties and potential sources of error in these estimates. The uncertainties include measurement errors in the spatial and tabular data sets used, potential biases in the results of the imputed time series of urban and rural rates of natural increase, and issues arising from the simplifying assumptions we applied in our processing steps. These uncertainties are addressed in greater detail in Section V, along with efforts to evaluate our results. However, we note here that the lack of observed population distribution data from 2010 round censuses means that the results for the 2000s are subject to greatest uncertainty.

I. Introduction

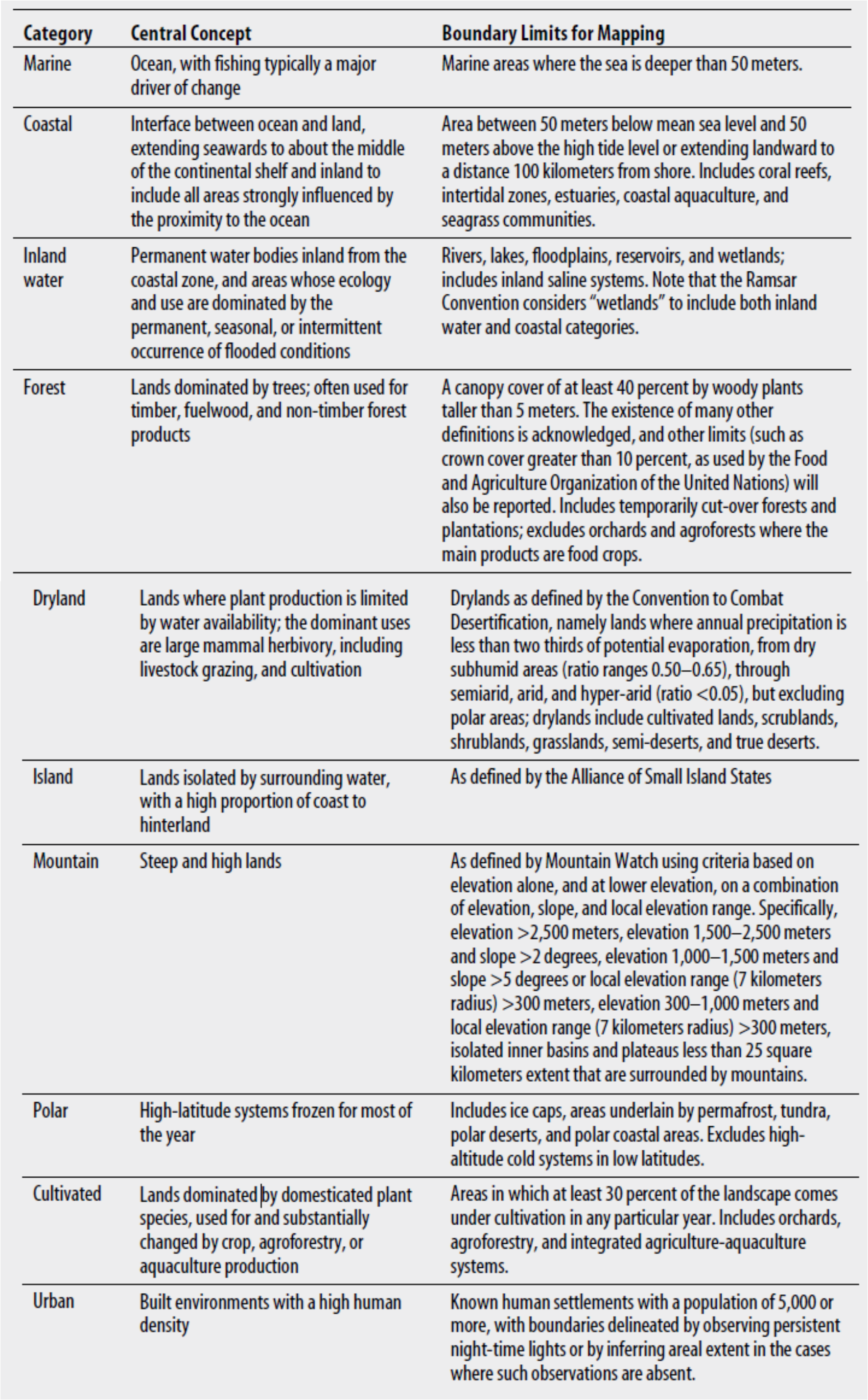
Under this project we generated estimates of decadal net migration flows by ecosystem type for the period 1970-2010. This report provides a brief review of the literature on migration by ecosystem type (Section II), and then describes the data and methods (Section III) and results (Section IV) of this modeling exercise. Evaluation of the results and a discussion of uncertainties and the steps that would need to be taken to reduce them are addressed in Section V.

The absence of globally accurate data on migration flows means that indirect estimation methods are necessary in order to estimate the number of net migrants for any given ecosystem over any given time period (Appendix D describes the problems with currently available migration data). Given the lack of direct measurements, our task was to develop estimates of net migration using the data that are currently available – time series population distribution grids combined with UN and other data on birth and death rates. We began with a high spatial resolution gridded population data set for the year 2000 and backcast and projected this grid using consistent rates so as to obtain population grids for the years 1970, 1980, 1990, 2000, and 2010. We then calculated the change in population per decade per grid cell by subtracting the decadal grid at the beginning of each decade from the decadal grid at the end of each decade. The result was a decadal population change grid. We then applied decadal rates of natural increase (birth rates *minus* death rates) to the population grids at the beginning of each decade to create estimates of natural increase per grid cell over the decade. Finally, we subtracted the natural increase grid from the population change grid for each decade, which yields a residual which we termed "net migration".[1] We recognize that the residual is in fact net migration plus some unknown error term. We sought to reduce this error term as much as possible by applying differential rates of natural increase across an urban-to-rural population density gradient based a combination of observed and imputed rates.

The ecosystem categories used in this analysis were drawn from the Millennium Ecosystem Assessment (MA), which developed a global map of ecosystems based on categories such as Drylands, Mountains, and Coastal, each with different numbers of subcategories (see Box 1 and Table A1). The categories were not mutually exclusive (e.g. Drylands and Coastal can and do overlap), so the net migration numbers by ecosystem presented in Section III and Appendix A do not sum to the global total net migration in any given decade. Although some ecosystem boundaries (namely forest and cultivated systems) have undoubtedly changed over four decades, there is no ecosystem boundary data set for earlier decades, and the categorization of ecosystems was sufficiently generalized that this may mitigate this shortcoming.

[1] Note that our methods cannot distinguish between international and domestic (or internal) migration, though we do constrain country-level migration to equal UN estimates.

Box 1. Millennium Ecosystem Assessment Reporting Categories



Category	Central Concept	Boundary Limits for Mapping
Marine	Ocean, with fishing typically a major driver of change	Marine areas where the sea is deeper than 50 meters.
Coastal	Interface between ocean and land, extending seawards to about the middle of the continental shelf and inland to include all areas strongly influenced by the proximity to the ocean	Area between 50 meters below mean sea level and 50 meters above the high tide level or extending landward to a distance 100 kilometers from shore. Includes coral reefs, intertidal zones, estuaries, coastal aquaculture, and seagrass communities.
Inland water	Permanent water bodies inland from the coastal zone, and areas whose ecology and use are dominated by the permanent, seasonal, or intermittent occurrence of flooded conditions	Rivers, lakes, floodplains, reservoirs, and wetlands; includes inland saline systems. Note that the Ramsar Convention considers "wetlands" to include both inland water and coastal categories.
Forest	Lands dominated by trees; often used for timber, fuelwood, and non-timber forest products	A canopy cover of at least 40 percent by woody plants taller than 5 meters. The existence of many other definitions is acknowledged, and other limits (such as crown cover greater than 10 percent, as used by the Food and Agriculture Organization of the United Nations) will also be reported. Includes temporarily cut-over forests and plantations; excludes orchards and agroforests where the main products are food crops.
Dryland	Lands where plant production is limited by water availability; the dominant uses are large mammal herbivory, including livestock grazing, and cultivation	Drylands as defined by the Convention to Combat Desertification, namely lands where annual precipitation is less than two thirds of potential evaporation, from dry subhumid areas (ratio ranges 0.50–0.65), through semiarid, arid, and hyper-arid (ratio <0.05), but excluding polar areas; drylands include cultivated lands, scrublands, shrublands, grasslands, semi-deserts, and true deserts.
Island	Lands isolated by surrounding water, with a high proportion of coast to hinterland	As defined by the Alliance of Small Island States
Mountain	Steep and high lands	As defined by Mountain Watch using criteria based on elevation alone, and at lower elevation, on a combination of elevation, slope, and local elevation range. Specifically, elevation >2,500 meters, elevation 1,500–2,500 meters and slope >2 degrees, elevation 1,000–1,500 meters and slope >5 degrees or local elevation range (7 kilometers radius) >300 meters, elevation 300–1,000 meters and local elevation range (7 kilometers radius) >300 meters, isolated inner basins and plateaus less than 25 square kilometers extent that are surrounded by mountains.
Polar	High-latitude systems frozen for most of the year	Includes ice caps, areas underlain by permafrost, tundra, polar deserts, and polar coastal areas. Excludes high-altitude cold systems in low latitudes.
Cultivated	Lands dominated by domesticated plant species, used for and substantially changed by crop, agroforestry, or aquaculture production	Areas in which at least 30 percent of the landscape comes under cultivation in any particular year. Includes orchards, agroforestry, and integrated agriculture-aquaculture systems.
Urban	Built environments with a high human density	Known human settlements with a population of 5,000 or more, with boundaries delineated by observing persistent night-time lights or by inferring areal extent in the cases where such observations are absent.

Source: Hassan et al. 2005; A technical description of the methods used to derive these layers is found in Appendix Table 2.2 of Defries and Pagiola 2005.

II. Literature Review

Ecosystems are expected to be affected by climate change processes such as warmer temperatures, rainfall variability, extreme events, and sea level rise. This will have major effects on human populations as ecosystem services are key providers of life's basic needs. Any change in their characteristics has the potential of affecting livelihoods, income, and migration trends (Corvalan et al. 2005:2, Warner et al. 2009, Adamo and de Sherbinin forthcoming), and may also lead to civil or interstate conflict, which itself is a precursor to population displacements (WBGU 2007, Campbell et al. 2007).

In 1990, the IPCC’s First Assessment Report already suggested that the greatest effect of climate change on society could be human migration, meaning involuntary forms of displacement and relocation (OSCE 2005). In 2007, the IPCC’s Fourth Assessment Report highlighted the significance of already established migrant networks and patterns as part of the inventory of adaptation practices, options and capacities available to face climate change impacts (Adger *et al.* 2007:736).

Climate change and ecosystem impacts will create different kinds of migration responses. Studies have shown that environmental displacements take place mostly within national boundaries (Adamo and de Sherbinin *forthcoming*, EACH-FOR 2009). Nevertheless, climate change will likely cause an up-tick in international migration not only for those countries most often cited (e.g., Small Island States) but also for those that will experience increasing frequency in climate hazards such as drought and floods (Hugo 1996, Brown 2007, ADB 2011).

In this section we review the literature on migration and ecosystems, starting with a brief overview, then turning to a presentation of the few studies that have considered migration by different ecosystem type. For the most part this literature focuses on migration associated with processes such as land cover change or loss of ecosystem function, which in turn are driven by processes of agricultural expansion, economic development, and globalization. Suitability for cultivation or development are the primary factors associated with high in-migration, whereas lack of suitability or isolation from markets tends to fuel out-migration. In their meta-analysis of 108 cases of agricultural intensification, Keys and McConnell (2005) described several cases of large-scale migrations or resettlements associated with the establishment of plantations (Schelhas 1996), the construction of roads (Conelly 1992), or political events (Kasfir 1993). There is comparatively less written on the inherent characteristics of ecosystems that make them attractive to migrants – though certainly agricultural and development potential are part of the characteristics that make certain ecosystems more or less hospitable for new comers.

Migration patterns and ecosystems

From an “ecosystem” point of view, human migration is a driver of ecosystem, biodiversity and land use changes (Meyerson et al. 2007). These changes range from deforestation due to clearing for pastures and crops to urban and suburban sprawl and the abandonment of rural areas (e.g. Geist and Lambin 2004, Magdalena 1996, Aide and Grau 2004). In turn, changes in ecosystems (particularly in the quantity and quality of services) are among many drivers of migration (e.g. Adamo and de Sherbinin forthcoming, de Sherbinin et al. 2007, Henry 2004). Examples include declining land productivity and changing rainfall patterns.

Heterogeneity, however, is the rule when dealing with migration, ecosystems and climate change effects (ADB 2011, Warner et al. 2009). Different ecosystems present different opportunities and challenges to human settlement. Regional diversity is clear in the mechanisms that link environment and migration dynamics, and it is also evident in terms of data availability and accuracy of estimates of environmentally induced displacements.

This uneven distribution requires a mixed approach combining ecosystems with national and subnational boundaries, in order to account for local dynamics and policy developments in assessing environmental displacement. For example, cultivated systems and coastal ecosystems tend to show higher human well-being, while drylands display lower human well-being (Levy et al. 2005), resulting in different 'pull' and 'push' factors for migration flows.

Urbanization (the proportion of people living in urban settlements) seems to be the exception to the heterogeneity rule. The redistribution of population toward urban areas is evident in most ecosystems (Aide and Grau 2004, Grau and Aide 2008). This trend is most visible in coastal areas (MacGranahan et al. 2007) but is also evident in drylands (Balk et al. 2009, Barbieri et al. 2010), forests (Uriarte et al. 2010) and mountains (Riebsame et al. 1996). Rural-urban migration often fuels urban growth, but as cities become larger the component of growth due to natural increase is often greater than that due to migration (Montgomery 2008).

Coastal areas

A growing proportion of the world population (about 40% in 1995) lives in coastal areas. Settlements are increasingly urban (Curran 2002, Balk et al. 2009) although in some countries (Vietnam, Bangladesh, Egypt, Mauritania, Cambodia) a large proportion of the rural population also lives in coastal areas (McGranahan et al. 2007).

A large part of this accelerated population growth in coastal areas is attributed to in-migration (Curran 2002, Agardy and Alder 2005). Population mobility in coastal areas includes permanent migration, seasonal labor migration, and tourists. This attraction or 'pull' effect of coastal areas derives from their endowment of natural resources (for example natural amenities, exploitation of mangroves, fishing), communication and transportation facilities, and diversity of work opportunities. On the other hand, rural communities in coastal areas, particularly those heavily dependent on natural resources, have also witnessed out-migration due to changes in the original conditions (e.g. Hamilton and Butler 2002) such as depletion of fisheries.

Low elevation coastal zones are particularly vulnerable to storms, storm surges, and sea level rise (MacGranahan et al. 2007), and population growth in coastal areas places more people potentially in harm's way, which could mean that migration out of the near coastal areas will increase in the future (Balk et al. 2009, Wheeler 2011).

Drylands

Drylands (arid, semiarid, and dry sub-humid areas) cover about 40% of the Earth's land surface and house more than 2 billion people, 90% of them in developing countries (IIED 2008). Overall, population growth is higher and human well-being is lower among drylands populations (Levy et al. 2005, IIED 2008, Safriel and Adeel 2005). Drylands tend to be less urbanized than coastal ecosystems, with approximately 45% of the population living in urban areas (Balk et al. 2009).

Climate change threats to drylands include increasing water shortages (especially in semiarid and dry sub-humid areas) and frequency of droughts, and declining flows in rivers depending on glacier melt (IIED 2008, Adamo and de Sherbinin forthcoming). Other areas may witness an increase in rainfall , although concomitant increases in temperature may offset the benefits (Safriel et al. 2005) .

Population mobility in drylands is a very common household livelihood strategy, composed of different types of movements (permanent, temporary and seasonal) into, outside and within arid lands (Rain 1999). The Millennium Ecosystem Assessment concluded (with medium certainty) that droughts and land degradation (particularly losses in productivity) were key factors behind migration from drylands (Safriel and Adeel 2005). Inter- and intra-annual changes in water availability due to climate change are expected to have an effect on migration patterns (e.g. Barbieri et al. 2010, Feng et al. 2010).

Mountains

In developed countries some mountain zones have attracted migrants who are seeking the amenities associated with mountain areas. For example, migration to the Rocky Mountains in the United States increased significantly in the last two decades (e.g. Riebsame et al. 1996). According to Shumway and Otterstrom (2004), "In the Mountain West, a number of counties with service-based economies are located in areas with high levels of environmental or natural amenities, creating what has been termed the 'New West.' Migration to the rural parts of the Mountain West, and the income transfers associated with migration, are increasingly concentrated within these New West counties."

Similar patterns have been seen in the Alps, with migration being spurred through the development of resorts and retirement communities, and also in advanced developing countries, for example in Chile (Hidalgo et al. 2009) and in Argentina (Gonzalez et al. 2009). In other mountain zones of the developed world, such as certain regions of the Massif Central in France and of Appalachia in the United States, there has been out migration and depopulation (André 1998).

In much of the developing world, mountain areas have been areas of net out-migration and population loss as people flee so-called "spatial poverty traps" (Scott 2006) – areas with low market access and poor infrastructure – for regions with greater market penetration and infrastructure (Xu 2008, Körner and Ohsawa 2005, Valdivia et al. 2010).

Forests and Cultivated Ecosystems

Migration to the agricultural frontier has been one of several contributors to deforestation in the tropics and dry forest areas, acting in combination with agriculture and pasture expansion and commercial logging (Carr 2009, Geist and Lambin 2002). In the world's iconic forest frontier, the Amazon, migration processes in recent years have reversed due to factors linked to urbanization and modernization of agriculture, and this has led in some cases to forest recovery (Aide and Grau 2004, Grau and Aide 2008, Barbieri et al. 2009)

Depending on land tenure, type of agriculture (commercial or subsistence), the degree of modernization, the relationship with markets, etc., cultivated ecosystems could either attract or expulse population. For example, the expansion of soybeans and biofuels – which usually require minimal labor inputs – has been associated with out migration of farm laborers and smallholders households, while the adoption of labor intensive farm systems are associated with population retention and seasonal migration (Craviotti and Soverna 1999, Grau and Aide 2008).

III. Methodology

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

1. We utilized the HYDE (History Database of the Global Environment) 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. We applied those rates to the Global Rural-Urban Mapping Project (GRUMP) (CIESIN 2011) population grids for 2000, producing "backcast" grids 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). [2]

3. We 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, we 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

[2] Table A2 presents a conversion of grid cell resolutions in East-West Arcs to distances.

to step 3 above, we adjusted the natural increase grids to match the UN estimates of natural increase at the country level.

5. Next, for each decade, we 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)

6. Using zonal statistics in ArcGIS, we produced aggregations by Millennium Ecosystem Assessment (MA) ecosystems (six core ecosystems and 34 sub-systems) to come up with estimates of net migration per decade per ecosystem (see Table A1). We retained country identifiers so that analyses can be performed for any country-ecosystem combination (e.g. drylands of Africa).

Detailed Data and Methods

To conduct this modeling exercise we chose to use the Global Rural-Urban Mapping Project (GRUMP) version 1 (CIESIN et al. 2011) population grid, which represents an urban reallocation of the Gridded Population of the World v.3 (GPWv3) 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 (Balk et al. 2004 and 2010). 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, outside the United States it only uses census data from only 79,590 administrative units 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 (Bright *personal communication*). The precise reallocation algorithm is not documented.

In contrast, GRUMP is based on population data from GPWv3, which uses 338,863 census units outside of the US (Table A3), 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 of drylands, there is generally less certainty regarding the spatial location of populations, and this will affect estimates of net migration (see Appendix E).

To ensure that we had consistent rates of population change over the four decadal periods, we applied a grid representing the rate of population change per decade derived from the History Database of the Global Environment version 3.1 (HYDEv3.1) population grids for the years 1970, 1980, 1990, and 2000. The HYDEv3.1 grids are adjusted at the country level to match the country totals from the UN Population Division's *World Population Prospects, 2008 Revision* (UN 2009). A detailed description of the HYDE data set and its evolution is provided in Appendix B. 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, meaning that our coastal and island ecosystem estimates are not taking into account these countries. A list of missing states is included at the end of Appendix B. We have tallied net migration data from alternate sources for these islands (UN 2009, Census Bureau's International database), and have provided separate tables of these results (Table A13).

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 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 we adjusted the gridded country totals so that they equal the UN *World Population Prospects, 2008 Revision* (UN 2009) 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.[3] 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).

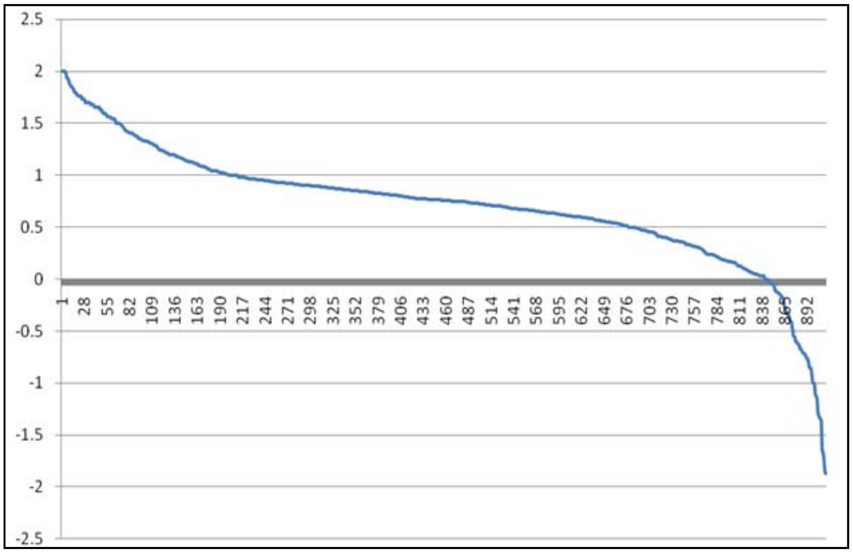
In a pilot effort, we applied national level rates of natural increase (crude birth rates minus crude death rates) from the *World Population Prospects* (UN 2009) to population grids to derive decadal estimates of natural increase. However, this approach ignored the fact that there is substantial subnational 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*[4] published from 1970 to 2008, and deriving

[3] We 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.

[4] 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

urban and RNIs (CBRs minus CDRs), we 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 we had observed data (900 out of 1,070 cases). There is no clustering around 1, 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.

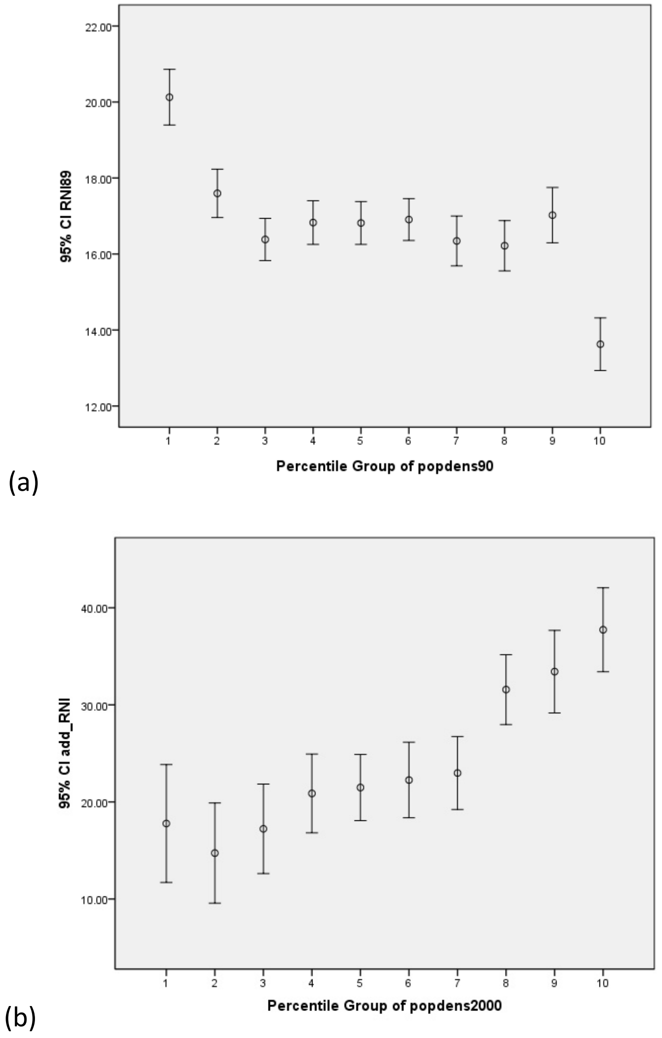
We hypothesized that RNIs can be predicted based on where a particular grid cell lies on an urban to rural gradient as measured by population density. We tested this hypothesis for subnational data on RNIs for two countries: China and the United States. For China, we used data for 2,315 districts for 1989-90 from the CIESIN China Dimensions data collection (CITAS et al. 1997) and found 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 we 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 our hunch that there is a systematic relationship between RNIs and population density, though that relationship varies by development level (Figure 3). We 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, we created a database of urban and rural CBRs and CDRs based on available data and

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.

imputation methods. We 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), we combined 5,016 observed values with as many auxiliary variables as we could obtain that might help explain patterns of urban and rural birth and death rates (see Table C1). Two models were used, Multiple Imputation (mi) and Amelia, and the methods and results are described in Appendix C. Although the results are subject to uncertainties (see Appendix C and Section V on Next Steps), and though the results were generally better for the CBRs than the CDRs (which proved more difficult to predict from available data), we feel this approach is better than ignoring subnational variation in RNIs. We obtained a total of 13 imputation runs – 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.

Figure 2. RNIs (y-axis) Across the Rural-to-Urban Population Density Gradient: (a) China (1989-90) and (b) United States (2000)



(a)

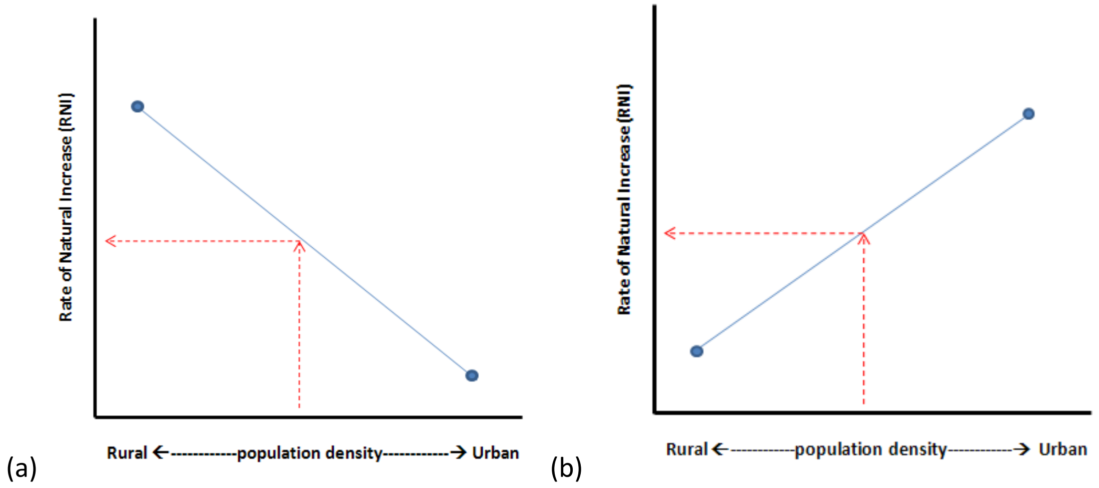
(b)

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, we 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. We 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 we had completed the imputations, we began a series of processing steps to develop spatial estimates of NM on a 30 arc-second grid. The remainder of this section details the processing steps we 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: We 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. Our 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, we already had average urban and rural RNIs, so we 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, we calculated the average urban and rural population densities for 1995 based on the GRUMP delineation of urban extents. Using SPSS, we 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 we did not have urban extents for the other decades, we needed to adjust the slope 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, we 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). We did not change the intercept, but instead relied upon the intercepts for each imputation run from the 1990s.

Step 3. Our 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, we multiplied the population counts grid and the RNI grid. At the pixel level we 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 we 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 we 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

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 we 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 we were able to estimate net migration for each decade for each grid cell based on 13 imputation runs. We further processed these runs in order to remove rounding errors by ecosystem, so that the global NM totals for each decade summed to less than +/- 1 persons. With 13 runs, we were able to develop an average and a standard deviation of the model runs for net migration (NM) for each grid cell, which represents a "pseudo" error bar for our estimates. But we must caution that the actual numbers represent net migration plus or minus some unknown error term per grid cell. Nevertheless, because of the methodology we followed for this work, the sum of net migration of all grid cells in any given country is very close to the total net migration per country according to the *World Population Prospects 2008*. We 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 we were unable to precisely quantify the amount of error in our estimates, we were able to characterize the precision and accuracy of the data inputs. This assessment is found in Appendix E. Note that we 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 our results. The UN Population Division provides extensive documentation but given resource constraints we were not able to fully characterize the uncertainties for any country-decade combination, though we do assess frequency of censuses in Appendix E, which is an important underpinning of both our work and the UN data.

IV. Results

Before describing the results, it should be emphasized again that the methodology employed here was experimental and the results are therefore to be treated as broad estimates of likely net migration flows by decade. Furthermore, as mentioned above, results for the decade of the 2000s are even more speculative because we lacked observed population distributions for 2010 with which to generate an accurate measure of decadal population growth at the pixel level. We cannot put precise quantitative estimates on the levels of error (the pseudo-error bars produced by the imputation runs are not genuine error bars), but we do discuss the uncertainties and the measures that would be required to reduce them in Section V.

Maps A1-A4 provide net migration results by decade of our global modeling without ecosystem masks; Map Sets A1-A3 provide results per decade for coastal, mountain, and dryland ecosystems, respectively. Tables A5-A10 provide results by decade for all ecosystems and then each ecosystem in turn, broken out by major UN regions, China, and the United States. Figures A1-A7 provide line graphs depicting the net migration figures by ecosystem and region with pseudo-error bars (except for Figure A1) representing the standard deviation for each model. Note that the standard deviations are in proportion to population size, so the high standard deviations in model outputs for Asia reflects the larger populations in that region.

At the request of the Foresight Project, we have also produced maps (Map Set A4) and figures (Figures A8-A13 for the Mediterranean region, though we do not specifically discuss the results here.

All Ecosystems

Results are found in Table A5. As expected, patterns of net migration by decade show regional variations following the classical developed/developing divide. Europe, Oceania and North America show a positive balance over the time period (1970-2010), while Africa, Asia and Latin America and the Caribbean have a negative balance. There are internal variations within regions: Southern Africa and Western Asia have positive balances; results are mixed for Eastern Europe; and Melanesia's NM balance is consistently negative.

Coastal Ecosystems

Results are found in Table A6, Figure A2, and Map Set A1. Globally, positive NM in coastal ecosystems is in the range of 30m in the 1970s and 1980s to 82m in the 2000s. The overall trend is upwards, with a more than doubling of the levels over four decades. This overall trend holds at the regional level, with magnitudes largely driven by the size of net migration in coastal ecosystems in Asia. However, there are some sub-regional exceptions: NM is negative over the whole period for coastal ecosystems located in northern Africa, Central America, the Caribbean, and Melanesia (except for the 2000s).

We also found very high levels of coastal out-migration in Canada across the first three decades, ranging from -1.3m in the 1970s to -2.1m in the 1990s. While this corresponds to a period of economic downturn in the Maritime provinces owing to the collapse of cod stocks and closure of fisheries, the 1990s net outmigration would represent roughly 10% of the Canadian population at that time. In the absence of corroborating evidence, we are not overly confident in these results.[5] The trend reversed in the 2000s (+0.7m NM), which may be owing to strong international migration to area around Vancouver. The reversal of NM in the coastal US during the 2000s also seems puzzling (-1m during the decade). Hurricane Katrina displaced several hundred thousand people, but our data inputs would not have easily picked up on this change.

Looking at the map insets in Map Set A1, the high levels of net in-migration in coastal China stands out throughout the four decades, though there are rural coastal areas in the 1980s and 200s that show significant out-migration. Northern Germany and the Netherlands also show significant areas of net in-migration.

[5] Canada was one of the countries with no data on urban/rural rates of natural increase, which is probably increasing the uncertainty of the results for this country.

Mountain Ecosystems (higher altitude systems)

Results are found in Table A7, Figure A3, and Map Set A2. Global net migration in the upper mountain ecosystems (we excluded lower montane ecosystems from our analysis) is consistently negative over the four decades, and this trend is quite consistent at the regional level. Negative balances range from -22m in the 1970s and 1980s to -43m in the 2000s, with Asian upper mountain ecosystems leading the way.

Europe has had largely negative net-migration, owing largely to strong outmigration in Southern Europe. By contrast, Eastern Europe shows net migration into mountain systems. For North America, net migration is solidly positive for all four decades, probably reflecting the amenity migration to mountain areas described in the literature review.

Cultivated ecosystems

Results are found in Table A8 and Figure A4. Globally, cultivated ecosystems show negative net migration over the whole period, with a negative balance of ~12m in both the 1970s and 1980s, peaking at -34m in the 1990s, then coming down slightly to 23m in the 2000s. This system, together with mountain systems, may be the source fueling much of the coastal net in-migration. Africa, Asia and Latin America show negative trends in net out-migration over the four decades, starting negative and becoming increasingly negative. South Asia and Eastern Asia (driven mostly by China) have extraordinarily high rates of negative NM during the 1990s (approximately -16.5m each). Maps A3 and A4 depict larger areas of blue (signaling strong out-migration) during this decade in India and China. This was a decade of extraordinary rural-urban migration throughout the developing world, which may explain these trends.

There are strong regional differentials, however, with developed regions generally showing positive NM in cultivated systems. In Europe, trends show generally positive and increasing net migration driven entirely by Western Europe. In North America, net migration is positive and increasing throughout the four decades.

Forest ecosystems

Results are found in Table A9 and Figure A5. Globally, forest ecosystems present negative NM for the four decades ranging from -19m in the 1970s to -39m in the 2000s. This pattern is pretty consistent at the regional and even sub-regional levels; in fact only Eastern Europe, North America, and Australia and New Zealand show net in-migration to forest ecosystems across all decades.

Asia shows strongly negative net-migration throughout the four decades, with levels approaching -30m in the 2000s. The evaluation of China results (see Section V and Table A14) suggests that our model is probably overestimating forest outmigration by a factor of three during the 1990s.

Inland Waters

Results are found in Table A9 and Figure A6. By contrast with cultivated and forest systems, inland waters shows positive net migration globally with a generally positive trend, peaking at positive 53m net migrants in the 2000s. As with other systems, this is closely linked to Asia's magnitudes. Major growth in cities situated near inland waters in China may be driving the Asia trends. This tendency is consistent over time, and across regions and sub-regions. The exception is Latin America and the Caribbean, but numbers there are comparatively small with very large standard deviations.

Dryland ecosystems (excluding hyperarid)

Results are found in Table A10 , Figure A7, and Map Set A3. Global net migration in dryland ecosystems is negative over the whole period, with an abrupt increase in magnitude in the 1990s and 2000s: from a negative NM of around -10.5m in the 1970s and 1980s to a negative NM of about -24m in the 1990s and -38m in the 2000s. Given the much reported decline in dryland ecosystem services over the past 40 years, which was partly responsible for the creation of the Convention to Combat Desertification, this trend is not overly surprising.

At the regional level, net migration in developing regions is negative and trending downwards throughout the four decades. However there is an unexplained positive net migration in Asia in the 1980s, largely driven by China. This may have had something to do with government-led efforts to settle people and establish irrigated agriculture in the dry western portions of the country. Whatever led to the increase, this was quickly reversed in the 1990s and 2000s.

By contrast with the rest of the world, North America shows strong positive net migration during the entire period, driven largely by migration to the "sunbelt" of the US Southwest. This is consistent with observed data.

Polar ecosystems

Table A11 provides the results for Polar ecosystems. The numbers are quite a bit smaller in this system, with negative levels in the 1970s (driven mostly by Canada), followed by positive net migration for the remaining decades. This trend is almost wholly explained by NM in Europe and North America. The reversal of trends in the last three decades and may be partially explained by the development of the oil industry.

Island ecosystems

The results here presented in Table A12 (and associated Table A13 for islands smaller 300,000 population). Net migration in island ecosystems was negative across all decades, ranging from -1.9m in the 2000s to a maximum of -4.5m in the 1980s. The largest levels of negative net migration were in the Caribbean, South-Eastern Asia, and North America. The Caribbean net out-migration can be explained by labor market demands in North America and comparatively stagnant economies in the Caribbean. It is more difficult to explain the negative NM in North America, which was driven by spikes of island outmigration in Canada in the 1970s and the 1990s.

Evolution of NM in East Asia is irregular, flipping from positive in the 1970s to negative in the 1980s, and back to positive again in the last two decades. These trends are difficult to interpret without more local knowledge.

Overall Assessment

Overall, the patterns identified in the maps and tables need to be examined in light of auxiliary information. Some of the patterns conform to what would be expected, but given the methodological challenges associated with indirect estimation methods and the uncertainties in the data, we cannot always explain the patterns of estimated net migration. We turn next to an evaluation of the results using observed data for rates of natural increase in China and for net migration in the US.

V. Evaluation of Results and Next Steps

In this section we first review some steps we took to evaluate the results using alternative data for China and the United States. Then we identify the next steps that would be required in order to reduce the uncertainty in the results.

Evaluation of Results

To evaluate our results[6], we utilized alternative model runs for China based on observed rates of natural increase at the district level for 1990, and we utilized county level estimates of net migration for the United States from 2000-2009 produced by the US Census Bureau.

China evaluation

For China, we used a data set for 2,315 districts for 1989-90 from the CIESIN China Dimensions data collection (CITAS et al. 1997) which includes rates of natural increase by district. We assumed these rates remain constant over the 1990s in order to obtain a percentage change owing to natural increase for high density urban areas and lower density rural areas during the decade. Those transformed rates of natural increase were utilized to calculate the slope and intercept per Step 1 of the processing steps. All the remaining processing steps were the same.

A comparison of results obtained using the observed RNIs and the imputed RNIs is presented in Map A5. The map shows very few differences, and the Pearson's r coefficient for the relationship between the mean net migration results by ecosystem presented in Table A14 is 0.896 (p<.01). Figure A16 shows a scatter plot that demonstrates a very high correlation between the NM results obtained from observed and imputed RNIs, though the slope is steeper than a 1:1 relationship, suggesting that for some ecosystems the NM derived from imputed RNIs is more strongly negative (especially cultivated, forest and mountain ecosystems), and for the coastal ecosystem it is much more strongly positive (26m vs. 15m for the observed RNIs). The primary outlier is mountain ecosystems, which show close to zero net migration for the NM derived from observed RNIs but fairly high negative NM of -19m for the NM derived from imputed RNIs. However, overall the results suggest that for China the spatial distribution of net migration is adequately captured by the scenarios based on the imputed RNIs.

[6] Note that we do not use the term "validation" since validation implies that we actually have some true measure against which to measure errors. What we are doing here is using measures derived from alternative data sources to try to get a better sense of the uncertainties in the results.

United States evaluation

We obtained data on county-level net migration from 1990-1999 and 2000-2009 from the US Census Bureau.[7] Note that there is in fact no measurement program for migration flows in the US, and the only observations made by the Census Bureau is for migrant stocks at the time of the decennial censuses.[8] So net migration levels are inferred from decennial census results and inter-census surveys.

For the 1990s, the Census Bureau results suggested a total NM of 7.48m, whereas the model results showed a global NM for the US of 16.16m. The UN *World Population Prospects 2008* estimates 1990s NM for the US at 14.54m So, for some reason the US Census Bureau estimates a NM of roughly half the level of the UN and our UN constrained modeling effort. Although the magnitude of NM by ecosystem is off for this reason (see Table A15), the actual correlation in the relative magnitude of NM by ecosystem between our modeled results and the Census Bureau data is quite high, with an r-square of 0.81 (Pearson's r = .899, p<.01) (see Figure A18). Map A6 also shows a strong correspondence in overall patterns, recognizing that our modeling effort produced results on a 30 arc-second (~1km) pixel basis, whereas the Census Bureau NM is spread out over entire counties, which in the western US can be quite large. So it may be that our model results actually do a better job of allocating NM to the urban locales where much in-migration is occurring, while still depicting the relatively spread out and rural character of out-migration.

For the 2000s, it is important to note that the Census Bureau data were developed *prior to* the 2010 census and therefore the data represent estimates that have a considerable levels of uncertainty. Total NM for the US for the decade was estimated by the census to be at 8.94m, whereas the UN *World Population Prospects* estimates a US NM of 10.73m and the modeling effort found a total NM for the decade of 10.44m. Map A7 shows the NM estimates based on the Census Bureau data and those produced by this project. As with the 1990s, the patterns look quite similar when accounting for the county-level reporting of the US estimates. Table A16 and Figure A18 show some divergence at the ecosystem level between the estimates developed by this modeling effort, with for example the modeled results showing -1m NM in the coastal zone during the decade whereas the Census Bureau shows a more likely +1.8m. Nevertheless, in aggregate the correlation in results at the ecosystem level is fairly high (Pearson's r = .713, p<.05).

Overall, the results for China suggests that the model can produce reasonably robust results without adequate observational data on rural and urban rates of natural increase. Although the results of our models for NM in the US were reasonably good (given the constraints imposed by using UN data for all countries), it should be noted that an earlier model run that did not have the benefit of observed county-level data on RNIs, and which therefore relied entirely on imputed RNIs, did not produce very good results at the ecosystem level. Comparing the earlier modeled results with the Census Bureau estimates, the Pearson's r was only .21 (not significant) for the 1990s and .17 (not significant) for the 2000s. This underscores the importance of observed RNIs to our results. We turn next to the

[7] The data for the 1990s were obtained from http://www.census.gov/popest/archives/1990s/CO-99-04.html
And the data for the 2000s were obtained from http://www.census.gov/popest/counties/.
[8] For more on migration data types, see Appendix D.

uncertainties in the NM results, which stem largely from the imputation process that was used to estimate RNIs, and the steps that would need to be taken to reduce uncertainties.

Next Steps

There are a number of uncertainties in our results that would take time and effort to resolve. We present here the biggest uncertainties, which relate to the imputation results, the relationship between population density and natural increase, and the census inputs. In each sub-section we also examine the steps that would be required to improve results.

Imputation Results

Probably the biggest uncertainties relate to the imputation methods used to impute rates of natural increase, which were part of our indirect estimation methodology. To recap, in order to obtain subnational variation in the estimates of net migration, we subtracted subnational natural increase from the subnational change in population in each decade. The subnational natural increase, in turn, was obtained by multiplying rates of natural increase times the population at the beginning of the decade, with the rates varying subnationally as a function of population density. Ultimately, this method depended heavily on imputed urban and rural crude birth rates and crude death rates, which were used to derive urban and rural rates of natural increase (RNIs). Although we had national level birth and death rates from the *World Population Prospects 2008* at five year increments, the imputation procedures introduce uncertainty because for many countries we had either sparse or no observed urban/rural birth or death rates and thus had to impute much or all of the forty year annual time series.[9]

A good deal of the variance in the NM estimates across our model runs could be traced to variance in the imputed urban and rural RNIs. This generated high variance in slopes across the model runs, and the slopes were used to convert the urban densities to a grid of rates of natural increase. The high variance in rates of natural increase multiplied by population per grid cell tended to generate higher variance in regions with larger populations such as Asia, and lower variance in regions with sparse population such as Oceania. We considered applying a Loess smoothing algorithm to the mi imputation results, much as it was applied to the Amelia results (see Appendix C). This would have substantially reduced the variance in the NM estimates around the mean. However, it would not necessarily have resulted in more accurate estimates of RNIs and ultimately net migration.

The issue of missing data has been a topic of substantial research interest. Missing values pose more than a nuisance to analysts. They increase scientific uncertainty, create additional challenges in the application of statistical analysis software and can call the representativeness of the results into doubt. The field of statistical imputation methods has therefore received continuous interest and has grown exponentially with the advent of cheap computing power and Bayesian methods. Seminal work was done in particular by Rubin, Little, and Schafer but many others also contributed theory and software.

[9] In Appendix E we review the number of urban/rural birth and death rate inputs for the imputation process. The great number of inputs can be found in the former Soviet Union, whereas many other regions – including North America – had zero inputs. As discussed above, we ultimately used US Census Bureau data instead of imputed RNIs for the United States.

Multiple imputation has evolved into a more widely used technique to not only fill data gaps but to do so in a way that captures both the uncertainty in the data and in the imputation itself.

Multiple imputation is also the approach used in this project, albeit it was implemented in two different ways as is explained elsewhere in the report. The basic approach of multiple imputation is to generate more than *m* (m>1 but generally fewer than 10) completed data sets, analyze them separately and then combine the results using functions for the mean of the m individual estimates and the associated variance derived by Rubin (1987).

An important assumption has to be made regarding the so-called missing-data generating mechanism, which is specified separately from the data generating process in a likelihood-based approach. Three different scenarios are possible. In the simplest but also most restrictive sense, the missing data are produced in a process that is completely independent from the data process. That is, one could essentially toss a coin to decide which data points to knock out. It is called missing completely at random (MCAR) and allows the analyst to ignore the missing data pattern. A valid but sometimes still inefficient way to deal with MCAR data is to remove incomplete observations through list-wise deletion.

More often, the missing data mechanism is in some way related to other, observed data, and is referred to as missing at random (MAR). A hypothetical example for MAR data is that data are missing because they have exceeded a threshold in another control variable, for example, study participants are excluded from a blood pressure medication because they have excessive weight. It turns out that the MAR situation is quite widely applicable and can be dealt with easily in the likelihood analysis context. Specifically, under MAR the probability that an observation is missing may depend on the observed values but not the missing values. In addition, it is assumed that the parameters of the data model and the parameters of the missing data indicators are distinct, which means the missing data mechanism is said to be ignorable and inference can rely on the observed data alone. Lastly, in the adverse event that the data are missing because of their unobserved value, the analyst is faced with a missing not at random (MNAR) problem and missing and observed data models do not separate out nicely and an explicit missing data model has to be specified and included in the likelihood function for the observed data model. Since this often requires tailor-made solutions, the MNAR case is not as widespread as the MAR assumption.

The two multiple imputation approaches used in this study rely on the MAR assumption but differ in their implementation of the data model as is described in Appendix C. The decision on what procedure or model to use (different software packages such as R mi and SAS PROC MI come with different solutions) and should be evaluated with respect to their performance for the given data. Diagnostic tools are available these days and validation studies can also be performed by, for example, removing observed values, imputing them and comparing the imputed data with the observed value that was removed. It is noted though that multiple imputation is often not concerned with imputing the most accurate value but with maintaining the distributional characteristics of the data, e.g., the mean and variance-covariance structure.

The objective in this project was to fill the considerable gaps in crude birth and death rates for urban and rural areas (uCBR, rCBR and uCDR, rCDR). Missingness was approximately 96.8 percent in this panel data set. A multiple imputation approach tailored to this problem must therefore take the following into account:

- High fraction of missingness (but not necessarily missing information);
- Some countries without any observed values;
- Substantial autocorrelation due to the time series nature of the data and the fact that change in CBR and CDR manifests relatively slowly;
- Likely spatial correlation since neighboring countries and regions often share similarities;
- Lack of complete covariates to draw from; and
- Correlation between the panel series that are to be imputed, i.e., between the urban and rural CBRs and CDRs.

The two approaches applied have different strengths regarding these challenges. The mi procedure harnesses the power of all available covariates, spatial relatedness and the existing relationships between the urban and rural CBRs and CDRs within a country. The Amelia procedure on the other hand exploits the time series nature of the data, cross-country association, as well as the explanatory power of covariates (albeit in a different way). Amelia does not take the relatedness of the CBRs and CDRs into account.

Future work would seek to address deficiencies in the respective imputation models through a variety of measures, some of which are generic to both models, and some of which are model specific, and which fall into two categories: improvements to the model inputs and to the model specification. Assuming as we did an MAR missing data generating mechanism, one possibility is to work on assembling better or more complete covariates, which have known and empirically demonstrated associations with the imputation variables. They should if possible be available at the same geographical-political resolution, i.e., the urban and rural dichotomy of the CBRs and CDRs, but at least the mi procedure can also extract some value from country-level data.

The Amelia imputations were mainly limited by computational power and therefore could not use all available covariate information. In addition, incomplete covariates are also imputed, which necessitates finding a good balance between explanatory variables and the need to also fill their gaps using the imputation variables and other covariates (which essentially creates a circular problem). Running Amelia on powerful computers in parallel processes could save time and allow expansion of the model.

Countries without any information will remain a particular challenge. However, if only the imputation variables are missing but some data on covariates are available the models will treat them more like other countries with similar covariate information. This could be examined further to see if these similarities are indeed meaningful, especially for “unique” countries such as China and India but also small island states and others.

Model specification is another aspect that can be tested more systematically. The currently used models were built primarily on the basis of conversations with experts such as demographers, using simple linear, bivariate correlation analysis and exploratory graphs and scatterplots. Functional analysis of the relationships between variables to be imputed and covariates, as well as their interactions, could be further investigated and used to improve model specification.

A model specification issue specific to the mi package used in this analysis is that it does not currently take into account the temporal structure of the data. In other words, the mi package models and imputes a missing value for country i and time t as if it were conditionally independent of time (t - 1), time (t + 1), and all the other time periods. As a result, the estimated error variance is likely to be too small and the imputed values are unlikely to be consistent with the dominant trends in the observed data. For example, in most countries birth rates and death rates tend to decline over time.

When modeling (complete) time-series cross-section data, a popular specification is to model the response as a function of a country-specific intercept, the lag of the dependent variable, the current values of the explanatory variables, and the lag of the explanatory variables. Additional lags can be included if necessary but often are not necessary when the data are measured annually. It would be sensible and not terribly difficult for the mi package to use this specification when iteratively modeling and imputing time-series cross-section data with missing values.

There are two substantive hurdles that have not been fully overcome yet, mostly because time-series cross-section data is not a top priority for mi development. The first, as mentioned before, is that for many countries, there is no observed data on birth and/or death rates at the urban and rural levels. In that situation, a country-specific intercept is at best weakly identified by the data, and we would be forced to use the cross-sectional variation in the data to impute the missing values. The approach used in this analysis is to treat the country-intercepts as random draws from a normal distribution and to estimate the unknown variance.

The second hurdle is that if the specification were to include the lag of the dependent variable and/or explanatory variables, then it becomes necessary to impute the value(s) one year *before* the first observed value. For example, since our dataset starts in 1970, in order to model an outcome as a function of variables in 1969, we need to impute the relevant values for 1969. The usual procedure in the mi package of modeling and imputing the 1969 data does not exactly work, because in order to model the 1969 data, we would need the 1968 data and so on.

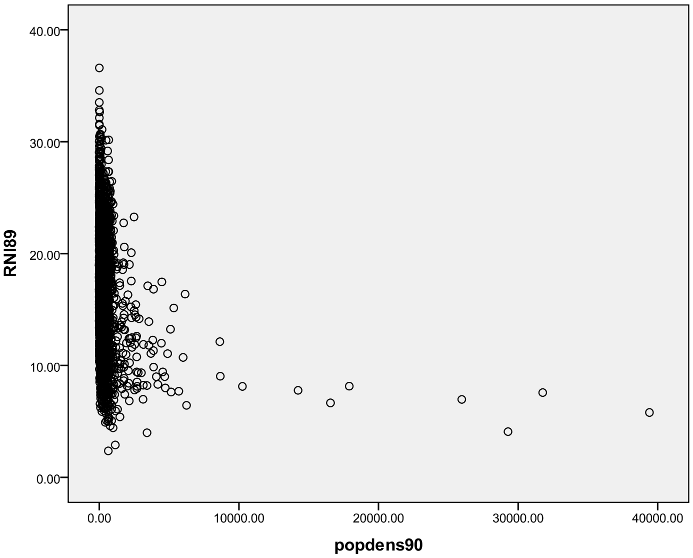
If these two hurdles could be overcome and sufficient time were available, then it would be possible to impute missing values with the mi package in a time-series cross-section context in a way that took the dynamic nature of the data into account and allowed for cross-country heterogeneity. The Amelia package already includes some options that are geared toward time-series cross-section data, but in this case (and others) Amelia did not produce very smooth series, which forced us to smooth them further with the Loess procedure, discussed in Appendix C.

It can be concluded that there are still possibilities and opportunities to learn more about the nature of the missing data and to identify ways to use the available information more effectively through improvements in model specification.

Relationship Between Density and Rate of Natural Increase

Although we hypothesize that there is a systematic relationship between the rates of natural increase and population density, the empirical evidence suggests a much more varied relationship. For example, without binning of the density levels, as was done for Figure 2, a simple scatter plot between district level population density and RNIs in China suggests a more varied relationship (Figure 4). While indeed most densely settled urban areas have consistently low RNIs, the rural areas represent a much more varied picture, which may be due to the effects of the one-child policy.

Figure 4. Scatter Plot of RNIs (1989) against Population Density (1990) for China



Thus, for developing countries, while it may be true that on average the densest urban areas have lower RNIs than other areas, the relationship between density and RNI for other areas is more complex and undoubtedly has to do with proximity to areas of economic activity and higher development levels. The relationship in developed countries is even more complex. Nevertheless, we feel justified in applying this simplifying assumption because the alternative, assuming uniform natural increase in all areas of a country, seems worse. Yet further research into the nature of the relationship between density and RNIs would improve the specification of the spatial modeling portion of this work.

The Census Inputs

There are several ways in which the census data inputs utilized in this project vary with respect to precision and accuracy, and some of these can be specified quantitatively. In Appendix E we review the sources of variation and present some high-level indicators related to the size of the census input units used in the gridded population products and frequency of censuses. The mean size of the census input units that underlie the population grids vary significantly by region and by ecosystem. Smaller sized units are generally most desirable, yet the size in square kilometers ranges from a ~100 sq km in Oceania to 60,000 sq km in Northern Africa and Australia, and for ecosystems the census inputs vary from a few 1,000 sq km in cultivated ecosystems to more than 60,000 sq km in high mountain and inland water ecosystems. There is little that we are able to do to redress these deficiencies other than to note that the number census inputs per country has tended to increase over time, which results in more accurate spatial allocation of populations .

A significant concern is the lack of observed data for population distribution in 2010. Because of the lag between completion of censuses and publication of results, and when one adds the time required to compile and grid census results, we were unable to use census data to map the 2010 population distribution. In other words, the 2010 population distribution and the 2000-2010 population growth was largely an extrapolation of 1990-2000 sub-national trends, but adjusted at the country level to the *World Population Prospects 2008* country level estimates. This implies that the 2000-2010 NM estimates have a good deal more uncertainty than the other estimates. The only real solution to this problem is to wait for 2010 round census results, which are beginning to be published.

Appendix A. Tables and Figures

Table A1. Millennium Ecosystem Assessment Systems and Subsystems

CULTIVATION		
VALUE	DESCRIP	AG_SHARE
1	Cropland	
2	Pasture	
3	Cropland / Pasture	
4	Agriculture with forest	60-80% ag
5	Agriculture with other vegetation	60-80% ag
6	Agriculture / Forest mosaic	approx. 50% ag
7	Agriculture / Other mosaic	approx. 50% ag
8	Forest with agriculture	20-40% ag
9	Other vegetation with agriculture	20-40% ag
10	Agriculture / 2 other land cover types	approx. 30% ag

DRY		
VALUE	DESCRIP	
2	Dry subhumid	
3	Semiarid	
4	Arid	
5	Hyperarid	

FORESTED		
VALUE	DESCRIP	
1	Tree Cover, broadleaved, evergreen	
2	Tree Cover, broadleaved, deciduous, closed	
3	Tree Cover, broadleaved, deciduous, open	
4	Tree Cover, needle-leaved, evergreen	
5	Tree Cover, needle-leaved, deciduous	
6	Tree Cover, mixed leaf type	
7	Tree Cover, regularly flooded, fresh	
8	Tree Cover, regularly flooded, saline, (daily var	
9	Mosaic: Tree cover / Other natural vegetation	
10	Tree Cover, burnt	

INLAND WATER		
VALUE	DESCRIP	
1	Lake	
2	Reservoir	
3	River	
4	Freshwater marsh, floodplain	
5	Swamp forest, flooded forest	
6	Pan, brackish/saline wetland	
7	Bog, fen, mire	
8	Intermittent wetland/lake	
9	50-100% wetland	

Table A1. Millennium Ecosystem Assessment Systems and Subsystems (continued)

ISLAND		
VALUE	DESCRIP	
1	Continental State Island / Inhabited / <= 2km	
2	Continental State Island / Inhabited / > 2km	
3	Continental State Island / Uninhabited / <= 2km	
4	Continental State Island / Uninhabited / > 2km	
5	Island State / Inhabited / <= 2km	
6	Island State / Inhabited / > 2km	
7	Island State / Uninhabited / > 2km	

MOUNTAIN		
VALUE	DESCRIP	
1	Humid tropical hill	
2	Humid tropical lower montane	
3	Humid tropical upper montane	
4	Humid temperate hill and lower montane	
5	Humid temperate lower/mid montane	
6	Humid temperate upper montane and pan-mixed	
7	Humid temperate alpine/nival	
8	Humid tropical alpine/nival	
9	Dry tropical hill	
10	Dry subtropical hill	
11	Dry warm temperate lower montane	
12	Dry cool temperate montane	
13	Dry boreal/subalpine	
14	Dry subpolar/alpine	
15	Polar/nival	

POLAR		
VALUE	DESCRIP	SHORT_DESC
1	Ice	ice
2	barrens and prostrate dwarf shrub tundra (includes rock/lichens, and prostrate tundra)	barrens and prostrate dwarf shrub tundra
3	graminoid, dwarf-shrub, and moss tundras	graminoid, dwarf-shrub, and moss tundras
4	forest tundra (inclu. low shrub tundra)	forest tundra
5	ice (Antarctica)	ice (Antarctica)

COASTAL		
VALUE	DESCRIP	
1	COASTAL	

Source: Millennium Ecosystem Assessment

Table A2. Grid Cell Resolutions In Relation to Distance on a Side and Area

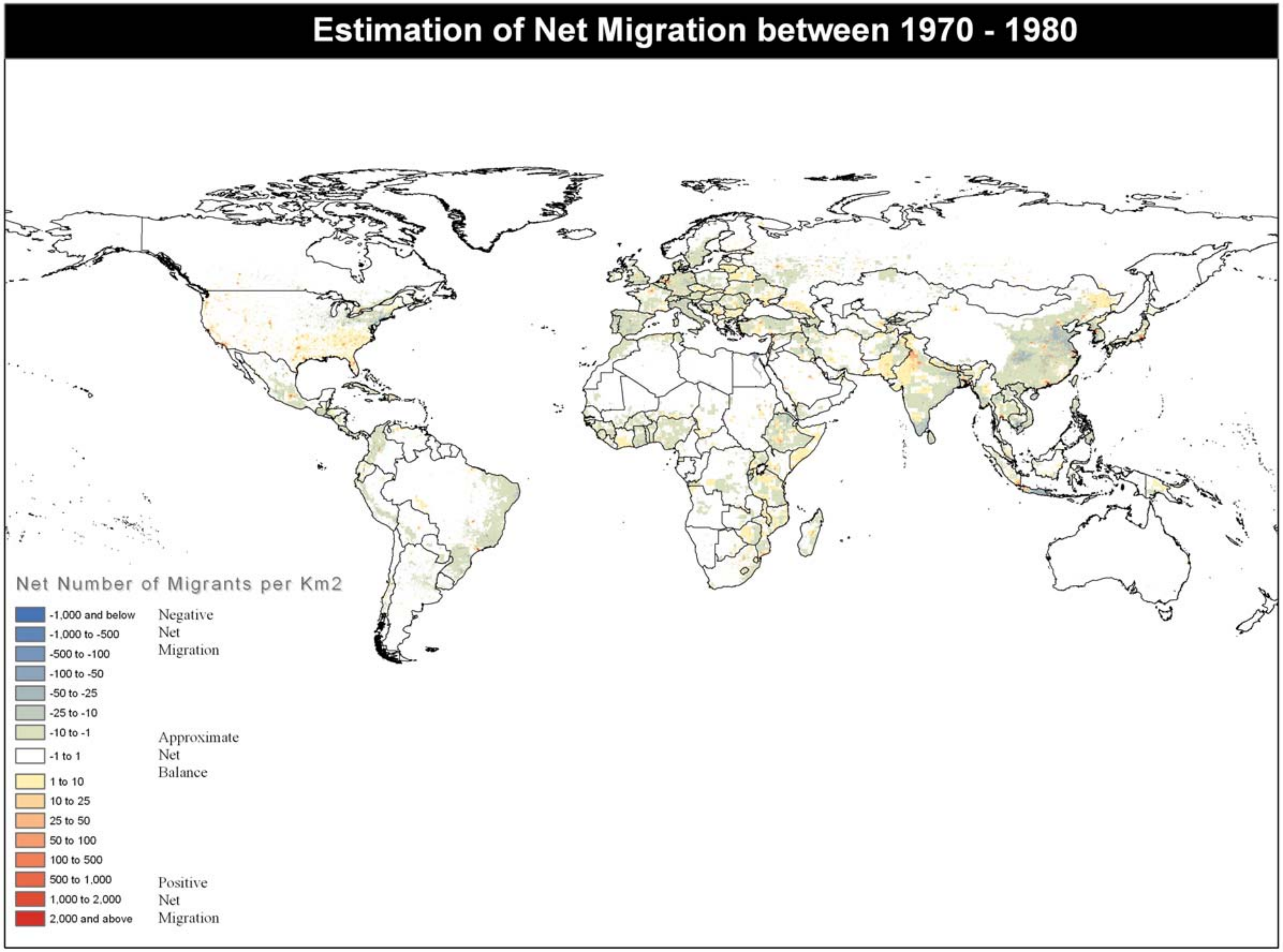
East-West Arcs	Distance on a side at equator	Area at equator
1 degree	111.32 km	12,392.14 km^2
5 minute (.083°)	9.3 km	86.49 km^2
2.5 minute (.042°)	4.65 km	21.62 km^2
30 arc-sec (.0083°)	0.93 km (~1km)	0.87 km^2
East-West Arcs	**Distance at 45°**	**Area at 45°**
1 degree	78.85 km	6,217.32 km^2
2.5 minute	3.275 km	10.73 km^2

Table A3. Summary Information on Input Units for Gridded Population of the World v3, by Continent

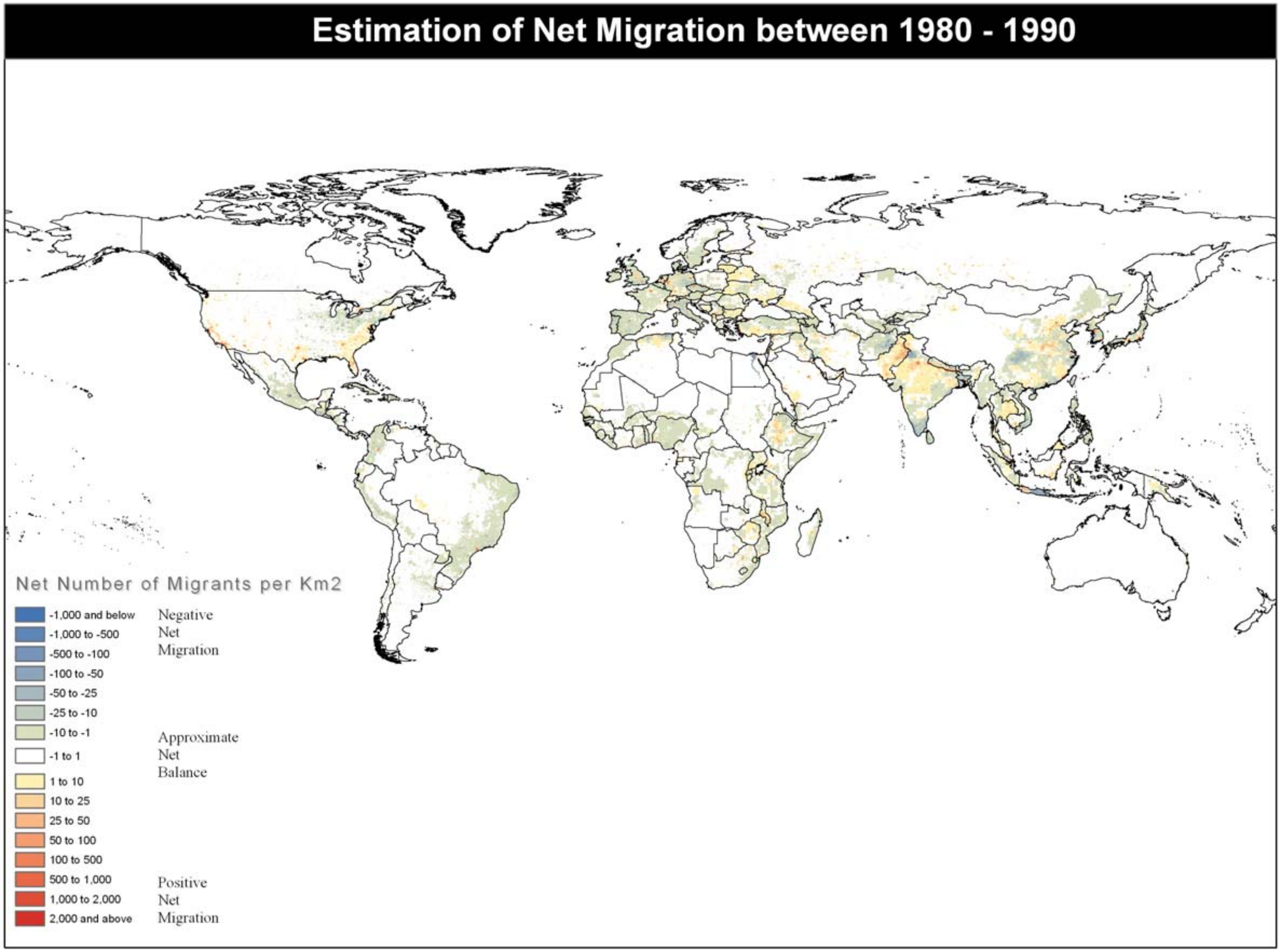
Continent				
	Modal Level*	Total Number of Units	Average Resolution	Average Persons per Unit
Africa	2	109,138	73	166
Asia	2	88,782	53	276
Europe	2	91,086	25	112
North America	2	74,421	29	83
Oceania	1	2,153	25	27
South America	2	10,919	68	49
Global	2	376,499	46	144

Source: Balk et al. 2010

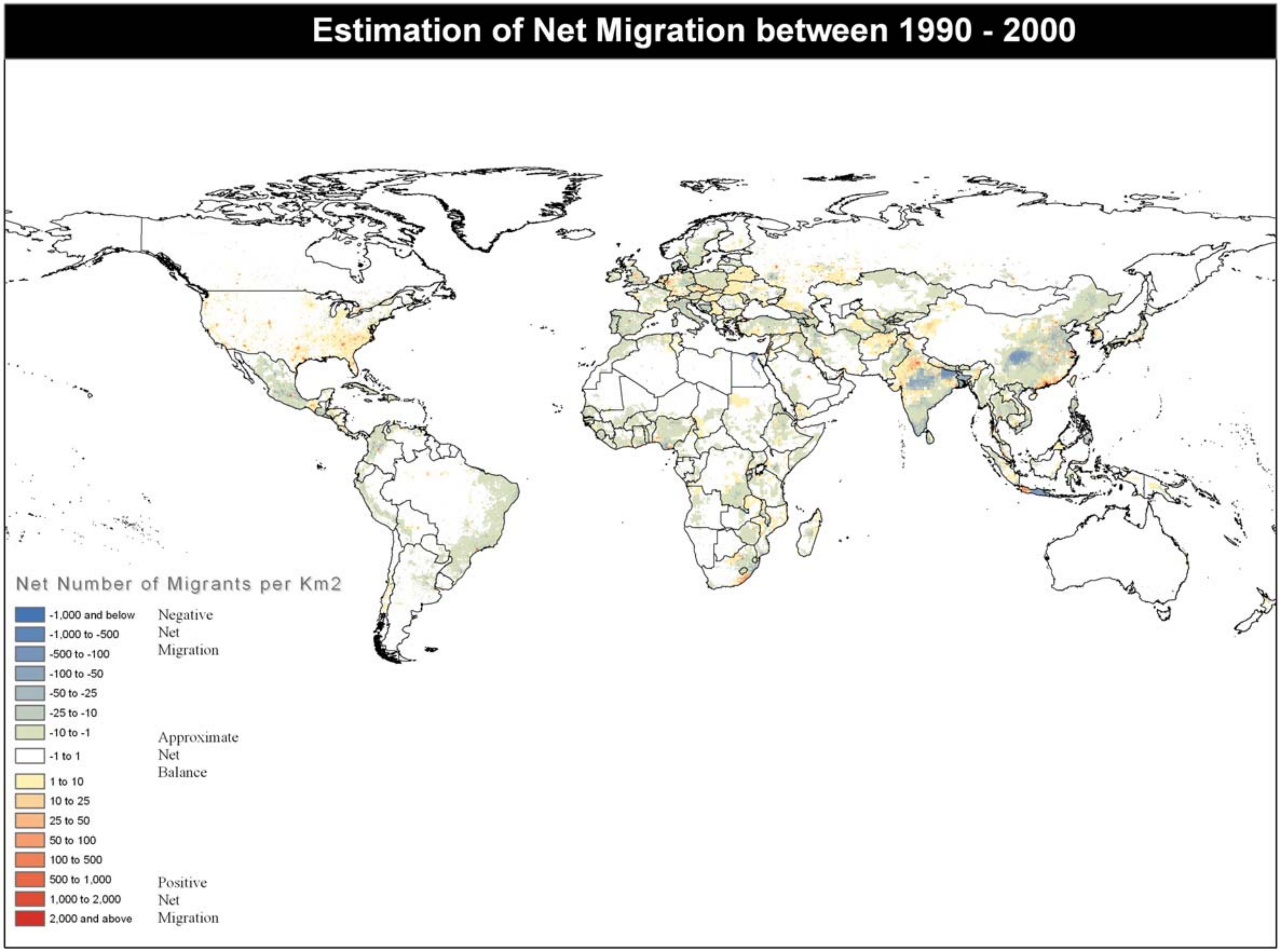
Map A1. Estimated Net Migration 1970s



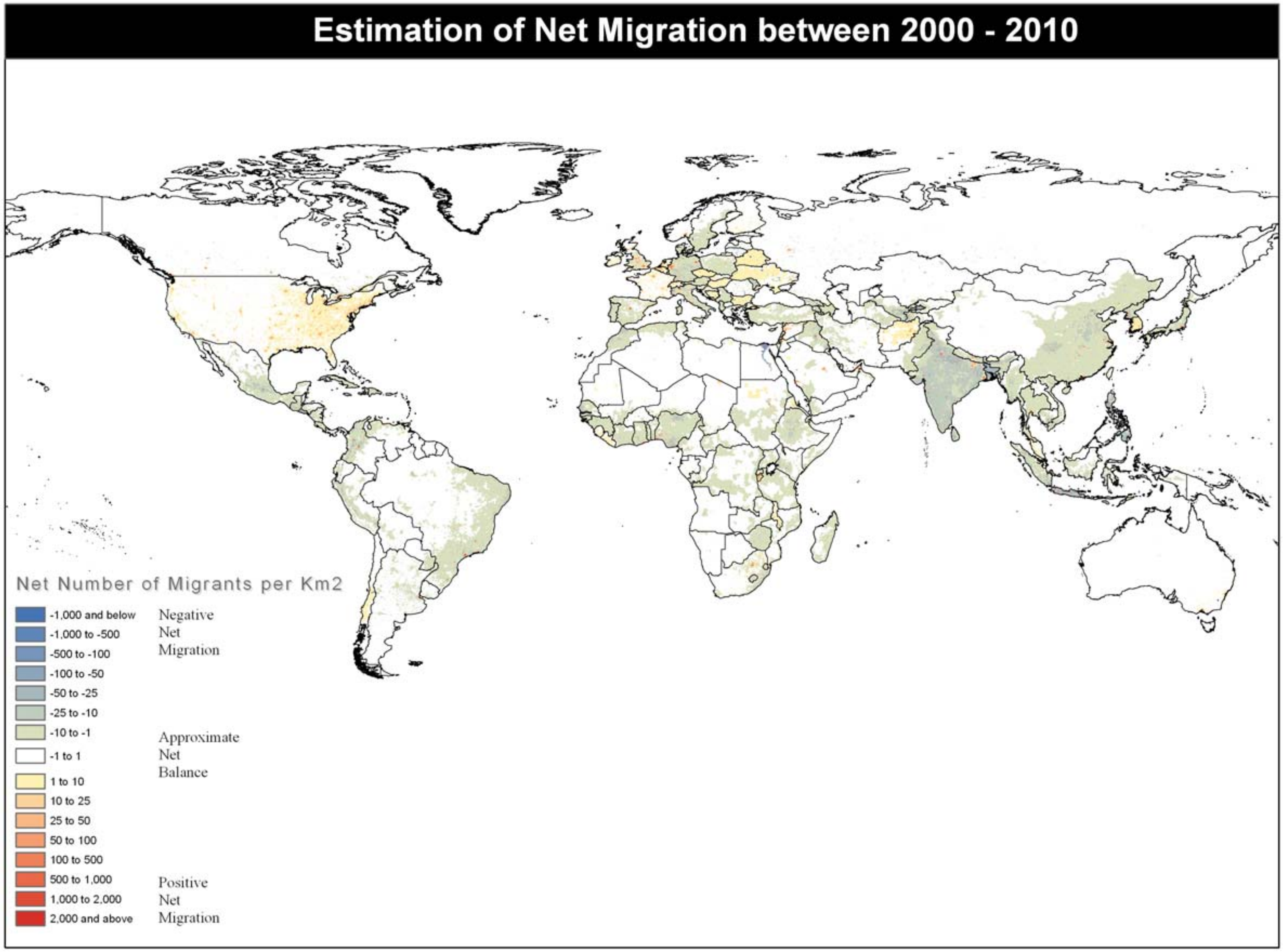
Map A2. Estimated Net Migration 1980s



Map A3. Estimated Net Migration 1990s



Map A4. Estimated Net Migration 2000s*



* Note: These estimates must be treated with particular caution because they are not based on observed population distributions in 2010. See the text for details.

Table A5. Net Migration for All Ecosystems

Ecosystem/UN_region	All ecosystems							
	1970		1980		1990		2000	
	Average	*StDev*	Average	*StDev*	Average	*StDev*	Average	*StDev*
Global	**0**	*0*	**0**	*0*	**0**	*0*	**0**	*0*
Africa	**-2,986,286**	*2,198*	**-2,766,992**	*1,195*	**-4,014,701**	*7,539*	**-5,420,791**	*5,679*
Northern Africa	-1,638,324	*503*	-1,161,216	*503*	-2,596,270	*2,511*	-2,433,764	*1,500*
Middle Africa	-126,230	*547*	-133,215	*247*	-90,002	*1,906*	-33,992	*1,228*
Western Africa	-186,527	*465*	-1,169,938	*469*	-922,184	*3,087*	-1,387,243	*1,317*
Eastern Africa	-1,245,616	*1,183*	-315,295	*509*	-1,361,623	*1,682*	-2,872,478	*2,679*
Southern Africa	210,411	*190*	12,672	*275*	955,378	*1,221*	1,306,685	*1,005*
Europe	**2,963,155**	*83,719*	**4,495,732**	*91,701*	**9,970,784**	*80,473*	**15,319,340**	*17,182*
Northern Europe	-69,297	*141,722*	157,709	*137,596*	354,638	*162,509*	2,908,449	*642*
Western Europe	2,616,636	*495*	3,273,176	*1,019*	5,206,092	*2,033*	4,065,874	*1,323*
Eastern Europe	-416,312	*139,097*	859,416	*107,644*	3,150,054	*177,101*	79,458	*1,173*
Southern Europe	832,129	*1,562*	205,431	*1,569*	1,259,999	*4,348*	8,265,558	*17,404*
North America	**7,314,970**	*82,312*	**9,842,519**	*92,298*	**15,170,834**	*82,094*	**12,859,641**	*917*
Northern America	7,314,970	*82,312*	9,842,519	*92,298*	15,170,834	*82,094*	12,859,641	*917*
Latin America and the Caribbean	**-4,175,727**	*2,263*	**-7,845,732**	*1,573*	**-7,237,869**	*13,173*	**-10,989,066**	*4,571*
Central America	-2,583,320	*1,806*	-5,324,188	*919*	-4,270,031	*7,269*	-6,782,638	*2,499*
Caribbean†	-1,133,976	*148*	-1,361,950	*88*	-1,183,922	*270*	-1,199,597	*1,045*
South America	-458,431	*1,155*	-1,159,594	*1,091*	-1,783,915	*8,945*	-3,006,831	*4,065*
Asia	**-3,658,356**	*5,786*	**-4,696,559**	*2,174*	**-14,768,469**	*19,718*	**-12,942,368**	*13,979*
Western Asia	1,273,059	*519*	1,175,888	*307*	-49,249	*2,987*	2,285,370	*1,263*
South-Central Asia	-2,219,135	*1,324*	-4,526,475	*1,389*	-9,747,041	*13,239*	-8,937,475	*6,275*
Eastern Asia	-934,269	*6,244*	120,729	*2,516*	-1,283,993	*23,997*	-3,344,508	*13,606*
South-Eastern Asia	-1,778,010	*1,760*	-1,466,701	*940*	-3,688,187	*7,754*	-2,945,755	*4,914*
Oceania	**542,243**	*49*	**971,031**	*65*	**879,422**	*532*	**1,173,245**	*69*
Australia and New Zealand	657,523	*39*	1,128,733	*57*	1,018,520	*446*	1,294,376	*61*
Melanesia†	-49,840	*31*	-89,918	*17*	-90,282	*87*	-66,431	*37*
Micronesia†	0	*0*	0	*0*	0	*0*	0	*0*
Polynesia†	-65,440	*2*	-67,784	*1*	-48,817	*6*	-54,701	*4*
China	**-1,359,034**	*488,636*	**-542,218**	*235,839*	**-1,272,109**	*327,752*	**-3,658,773**	*446,928*
USA	**7,934,280**	*72,736*	**9,584,103**	*67,759*	**16,154,882**	*56,056*	**10,442,903**	*912*

†Excludes small islands with populations under 300,000.

Table A6. Net Migration for the Coastal Ecosystem*

Ecosystem/UN_region	Coastal							
	1970		1980		1990		2000	
	Average	*StDev*	Average	*StDev*	Average	*StDev*	Average	*StDev*
Global	**30,504,221**	*15,199,543*	**27,143,283**	*14,652,931*	**72,984,160**	*23,338,767*	**82,094,081**	*33,472,770*
Africa	**2,144,952**	*1,188,626*	**1,662,674**	*1,626,173*	**4,230,605**	*1,584,188*	**5,447,186**	*1,713,677*
Northern Africa	-1,298,326	*216,063*	-1,437,626	*913,368*	-1,979,162	*502,953*	-1,772,625	*632,095*
Middle Africa	161,623	*190,072*	488,602	*279,402*	1,079,794	*180,475*	626,745	*280,445*
Western Africa	2,124,722	*532,937*	2,590,294	*765,101*	3,925,136	*995,296*	4,780,628	*823,567*
Eastern Africa	1,121,316	*391,820*	-75,880	*485,203*	1,017,358	*449,679*	1,265,445	*763,199*
Southern Africa	35,616	*84,008*	97,284	*133,786*	187,478	*191,082*	546,993	*287,308*
Europe	**2,877,503**	*765,567*	**3,435,037**	*787,846*	**3,630,706**	*1,000,721*	**6,289,301**	*1,516,045*
Northern Europe	328,269	*159,706*	774,972	*278,258*	609,382	*271,281*	2,247,105	*621,693*
Western Europe	333,965	*127,590*	191,798	*239,840*	707,948	*246,654*	670,919	*270,074*
Eastern Europe	297,067	*234,109*	374,463	*101,587*	-117,810	*181,185*	-220,443	*126,831*
Southern Europe	1,918,203	*746,173*	2,093,805	*900,096*	2,431,186	*819,761*	3,591,719	*1,022,454*
North America	**-149,265**	*111,789*	**2,061,153**	*100,181*	**404,874**	*131,374*	**-179,036**	*338,032*
Northern America	-149,265	*111,789*	2,061,153	*100,181*	404,874	*131,374*	-179,036	*338,032*
Latin America and the Caribbean	**3,196,119**	*1,124,868*	**4,855,932**	*1,417,857*	**4,295,320**	*2,138,264*	**4,968,650**	*2,314,933*
Central America	-240,794	*923,973*	-733,134	*1,054,742*	-1,265,953	*985,642*	-1,440,650	*1,493,172*
Caribbean†	-513,645	*124,569*	-87,884	*130,609*	-240,061	*87,942*	-225,013	*247,450*
South America	3,950,558	*1,158,359*	5,676,949	*1,660,623*	5,801,333	*2,344,572*	6,634,314	*1,796,662*
Asia	**22,003,638**	*13,718,904*	**14,276,143**	*12,151,137*	**59,571,112**	*21,452,381*	**64,825,722**	*31,007,417*
Western Asia	1,203,989	*193,816*	904,228	*190,475*	1,474,870	*362,057*	2,211,387	*677,769*
South-Central Asia	99,654	*3,478,467*	-1,779,532	*3,830,310*	16,932,248	*8,069,249*	30,165,585	*13,226,772*
Eastern Asia	12,924,933	*10,923,218*	7,090,779	*6,704,996*	28,633,201	*13,883,901*	17,662,878	*17,023,951*
South-Eastern Asia	7,775,062	*3,398,815*	8,060,668	*6,396,541*	12,530,794	*6,361,181*	14,785,872	*7,821,281*
Oceania	**431,274**	*169,360*	**852,344**	*240,790*	**851,542**	*285,533*	**742,257**	*217,087*
Australia and New Zealand	463,151	*172,085*	893,831	*239,336*	864,867	*274,584*	721,493	*205,745*
Melanesia†	-20,523	*6,426*	-29,292	*9,442*	-5,143	*12,587*	27,851	*18,814*
Micronesia†	0	*0*	0	*0*	0	*0*	0	*0*
Polynesia†	-11,353	*1,093*	-12,195	*1,145*	-8,182	*1,270*	-7,087	*1,808*
China	**11,834,456**	*9,800,265*	**4,625,581**	*6,514,140*	**26,351,021**	*13,637,352*	**15,219,039**	*16,505,812*
USA	**1,443,290**	*1,001*	**3,115,996**	*1,354*	**2,941,103**	*1,788*	**-1,052,101**	*326*

* See Table A13 for net migration for small island states by decade.

†Excludes small islands with populations under 300,000.

Table A7. Net Migration for Mountain Ecosystems *

Ecosystem/UN_region	Mountain							
	1970		1980		1990		2000	
	Average	StDev	Average	StDev	Average	StDev	Average	StDev
Global	**-23,384,687**	12,410,812	**-22,743,723**	10,260,815	**-36,900,825**	13,599,185	**-43,247,703**	18,179,487
Africa	**-2,795,249**	954,332	**-2,001,789**	1,096,371	**-4,293,636**	1,426,037	**-6,540,535**	2,027,037
Northern Africa	-471,098	344,456	-1,033,836	377,266	-924,313	283,464	-1,419,950	469,966
Middle Africa	-10,549	265,368	201,365	222,547	-667,189	221,695	-324,319	525,892
Western Africa	-267,874	54,230	-257,945	72,841	-428,033	99,637	-597,482	94,981
Eastern Africa	-1,875,783	321,646	-428,575	451,500	-2,657,177	703,033	-3,303,646	718,348
Southern Africa	-169,945	267,588	-482,798	418,663	383,076	609,864	-895,138	817,175
Europe	**-2,794,237**	1,259,666	**-3,185,719**	1,941,847	**-2,991,756**	1,365,962	**-1,813,462**	1,621,850
Northern Europe	144,371	101,791	252,308	59,398	141,569	90,206	-24,898	84,524
Western Europe	-648,100	246,390	-472,825	669,154	-201,737	422,398	-305,602	538,654
Eastern Europe	311,112	100,255	279,404	86,567	740,879	103,049	174,034	129,699
Southern Europe	-2,601,619	1,124,330	-3,244,605	1,587,336	-3,672,468	1,213,977	-1,656,996	1,311,285
North America	**4,133,098**	91,869	**4,852,499**	80,113	**7,566,917**	107,113	**962,697**	108,488
Northern America	4,133,098	91,869	4,852,499	80,113	7,566,917	107,113	962,697	108,488
Latin America and the Caribbean	**-2,896,616**	1,779,519	**-5,984,636**	1,829,444	**-7,291,786**	1,565,120	**-7,064,234**	2,554,582
Central America	-1,029,205	1,692,381	-2,782,932	1,722,623	-3,670,955	1,440,054	-3,573,061	2,197,039
Caribbean†	-131,459	84,524	-388,753	82,425	-252,132	48,891	-318,431	161,720
South America	-1,735,952	385,006	-2,812,951	526,356	-3,368,700	444,479	-3,172,742	834,735
Asia	**-19,023,269**	10,559,457	**-16,349,197**	7,125,313	**-29,788,068**	11,370,905	**-28,714,062**	13,790,577
Western Asia	-1,714,296	583,871	-2,217,557	357,824	-2,907,732	888,166	-1,623,636	1,307,681
South-Central Asia	-1,089,525	1,395,605	-4,074,997	2,339,368	-1,792,684	1,831,478	-7,762,292	2,719,818
Eastern Asia	-12,670,688	9,195,309	-6,007,673	5,230,625	-20,187,170	9,809,740	-12,538,653	10,739,607
South-Eastern Asia	-3,548,759	1,220,764	-4,048,970	2,041,638	-4,900,483	2,375,998	-6,789,481	2,634,012
Oceania	**-8,414**	33,294	**-74,881**	49,886	**-102,495**	69,033	**-78,107**	70,479
Australia and New Zealand	-18,327	33,773	-45,580	44,041	-37,357	48,767	26,730	34,530
Melanesia†	9,913	7,335	-29,301	16,317	-65,138	23,675	-104,837	41,523
Micronesia†	0	0	0	0	0	0	0	0
Polynesia†								

China	**-11,973,460**	8,670,885	**-4,551,491**	5,001,721	**-19,052,948**	9,717,077	**-11,199,708**	10,558,251
USA	**3,885,519**	59,978	**4,536,311**	59,276	**7,237,223**	51,291	**1,014,054**	48

* Includes only Humid tropical upper montane, Humid temperate upper montane and pan-mixed, Humid temperate alpine/nival, Humid tropical alpine/nival, Dry cool temperate montane, Dry boreal/subalpine, Dry subpolar/alpine, and Polar/nival. Lower and lower/mid montane and hill ecosystems were removed.

†Excludes small islands with populations under 300,000.

Table A8. Net Migration for Cultivated Systems*

Ecosystem/UN_region	Cultivated							
	1970		1980		1990		2000	
	Average	*StDev*	Average	*StDev*	Average	*StDev*	Average	*StDev*
Global	**-12,751,417**	*6,176,736*	**-11,923,479**	*5,906,406*	**-34,396,406**	*9,772,862*	**-22,958,756**	*15,823,015*
Africa	**-1,495,729**	*544,253*	**-1,364,092**	*1,583,953*	**-3,912,595**	*801,574*	**-4,738,416**	*1,260,911*
Northern Africa	-1,487,364	*370,114*	-2,786,520	*1,444,876*	-2,840,739	*747,789*	-2,605,911	*1,185,644*
Middle Africa	530,628	*447,912*	783,028	*318,763*	240,420	*282,248*	559,533	*452,681*
Western Africa	-487,110	*83,498*	-646,698	*94,227*	-560,978	*151,634*	-1,338,239	*172,035*
Eastern Africa	-283,754	*165,439*	1,214,065	*151,227*	-840,802	*203,642*	-1,899,180	*329,058*
Southern Africa	231,869	*37,233*	72,033	*60,559*	89,505	*88,918*	545,380	*148,668*
Europe	**603,220**	*883,708*	**835,855**	*1,081,603*	**7,183,384**	*579,928*	**10,867,666**	*825,366*
Northern Europe	-280,159	*160,292*	-540,222	*179,851*	178,532	*191,455*	2,006,647	*82,674*
Western Europe	2,145,282	*474,923*	2,480,518	*604,593*	4,418,001	*402,250*	3,762,325	*350,940*
Eastern Europe	-1,036,850	*389,510*	-121,131	*313,548*	3,162,103	*415,027*	358,010	*272,734*
Southern Europe	-225,053	*550,899*	-983,311	*559,626*	-575,252	*546,071*	4,740,684	*756,552*
North America	**4,202,477**	*157,225*	**3,859,522**	*135,915*	**10,410,234**	*187,103*	**11,430,717**	*209,341*
Northern America	4,202,477	*157,225*	3,859,522	*135,915*	10,410,234	*187,103*	11,430,717	*209,341*
Latin America and the Caribbean	**-6,292,604**	*2,312,168*	**-9,404,610**	*3,625,004*	**-11,004,362**	*3,687,631*	**-11,362,221**	*2,890,131*
Central America	-1,066,197	*1,593,729*	-2,428,767	*1,637,693*	-3,111,637	*1,645,960*	-2,953,331	*2,303,728*
Caribbean†	-341,871	*82,674*	-430,255	*80,891*	-472,265	*78,410*	-455,784	*97,539*
South America	-4,884,536	*906,502*	-6,545,589	*2,192,734*	-7,420,460	*2,615,799*	-7,953,106	*2,018,541*
Asia	**-9,791,249**	*5,762,592*	**-5,897,132**	*3,966,188*	**-37,171,530**	*9,536,413*	**-29,334,870**	*15,215,344*
Western Asia	-2,898	*235,231*	-1,056,750	*175,988*	281,695	*389,717*	539,628	*562,301*
South-Central Asia	-1,585,972	*1,826,882*	-636,154	*2,051,250*	-16,570,948	*3,356,938*	-15,805,450	*6,615,303*
Eastern Asia	-7,266,576	*5,097,331*	-4,330,315	*4,139,991*	-16,716,271	*8,650,442*	-12,723,684	*11,782,858*
South-Eastern Asia	-935,803	*936,691*	126,087	*850,705*	-4,166,006	*1,449,555*	-1,345,363	*2,027,190*
Oceania	**22,468**	*78,139*	**46,979**	*95,792*	**98,463**	*98,028*	**178,369**	*90,430*
Australia and New Zealand	-3,596	*78,092*	41,851	*95,686*	105,524	*96,965*	181,560	*87,541*
Melanesia†	26,064	*1,233*	5,128	*592*	-7,061	*1,683*	-3,191	*4,090*
Micronesia†	0	*0*	0	*0*	0	*0*	0	*0*
Polynesia†								

China	**-7,779,642**	*4,813,717*	**-5,164,043**	*4,108,277*	**-17,230,841**	*8,604,270*	**-13,104,767**	*11,694,991*
USA	**3,784,349**	*458*	**3,331,267**	*254*	**9,865,971**	*1,682*	**10,488,022**	*522*

* Includes all but Pasture.

†Excludes small islands with populations under 300,000.

Table A9. Net Migration for Forest Ecosystems

Ecosystem/UN_region	Forest							
	1970		1980		1990		2000	
	Average	*StDev*	Average	*StDev*	Average	*StDev*	Average	*StDev*
Global	**-19,276,493**	*10,048,550*	**-21,256,821**	*9,344,751*	**-25,446,233**	*11,737,066*	**-39,646,045**	*15,621,282*
Africa	**-3,547,684**	*1,688,931*	**-6,265,233**	*1,884,644*	**-6,725,177**	*2,052,899*	**-9,165,396**	*3,741,378*
Northern Africa	-53,879	*61,687*	-455,273	*65,416*	-208,071	*69,621*	-309,044	*135,640*
Middle Africa	-1,313,662	*1,142,480*	-2,533,891	*1,120,757*	-3,662,361	*1,111,051*	-3,603,083	*2,377,162*
Western Africa	-690,189	*112,900*	-845,474	*152,610*	-1,429,178	*255,639*	-1,626,148	*245,208*
Eastern Africa	-1,180,794	*413,758*	-1,851,349	*632,397*	-1,450,680	*602,758*	-2,734,268	*723,459*
Southern Africa	-309,159	*234,201*	-579,246	*369,409*	25,114	*515,083*	-892,853	*708,743*
Europe	**-1,973,588**	*993,135*	**-1,530,933**	*1,798,733*	**-902,946**	*1,213,569*	**-457,463**	*1,689,265*
Northern Europe	-184,746	*182,517*	132,125	*153,656*	-402,583	*170,222*	-373,865	*48,442*
Western Europe	-532,857	*305,231*	-334,789	*833,652*	231,507	*517,389*	-197,569	*667,740*
Eastern Europe	387,805	*402,378*	1,005,523	*402,655*	1,614,976	*295,200*	498,964	*291,295*
Southern Europe	-1,643,790	*726,932*	-2,333,792	*1,029,077*	-2,346,845	*790,992*	-384,994	*805,944*
North America	**2,763,800**	*280,052*	**3,889,078**	*323,670*	**7,333,533**	*354,307*	**5,190,364**	*531,322*
Northern America	2,763,800	*280,052*	3,889,078	*323,670*	7,333,533	*354,307*	5,190,364	*531,322*
Latin America and the Caribbean	**-1,210,162**	*305,703*	**-3,747,611**	*480,562*	**-2,571,922**	*670,580*	**-6,643,599**	*870,841*
Central America	-529,533	*439,039*	-2,062,077	*491,096*	-710,688	*501,597*	-2,740,806	*800,842*
Caribbean†	-233,032	*48,442*	-402,286	*56,467*	-303,993	*39,446*	-291,875	*91,237*
South America	-447,597	*200,157*	-1,283,247	*593,959*	-1,557,241	*810,092*	-3,610,918	*683,673*
Asia	**-15,246,655**	*8,035,745*	**-13,527,726**	*5,848,323*	**-22,445,192**	*9,705,012*	**-28,666,429**	*11,514,559*
Western Asia	-154,390	*151,983*	-219,024	*84,108*	-325,461	*111,248*	-230,117	*124,311*
South-Central Asia	-3,139,322	*598,593*	-3,209,997	*512,219*	-6,868,506	*1,787,300*	-10,229,244	*2,077,118*
Eastern Asia	-8,303,324	*6,994,241*	-6,356,738	*4,102,591*	-10,816,066	*7,291,319*	-10,261,098	*8,098,845*
South-Eastern Asia	-3,649,619	*1,337,596*	-3,741,968	*2,328,669*	-4,435,160	*2,652,763*	-7,945,969	*3,198,308*
Oceania	**-62,203**	*146,822*	**-74,396**	*210,214*	**-134,529**	*249,944*	**96,478**	*217,542*
Australia and New Zealand	76,010	*148,369*	90,636	*208,521*	13,476	*229,979*	247,589	*187,479*
Melanesia†	-138,213	*9,001*	-165,031	*19,412*	-148,006	*25,604*	-151,111	*43,048*
Micronesia†	0	*0*	0	*0*	0	*0*	0	*0*
Polynesia†								

China	**-7,283,813**	*5,903,549*	**-4,003,819**	*3,526,180*	**-9,094,914**	*7,001,862*	**-8,177,860**	*7,578,934*
USA	**4,501,263**	*39,795*	**5,092,257**	*32,599*	**10,030,185**	*24,887*	**5,331,891**	*274*

†Excludes small islands with populations under 300,000.

Table A9. Net Migration for Inland Water Ecosystems

Ecosystem/UN_region	Inland waters							
	1970		1980		1990		2000	
	Average	*StDev*	Average	*StDev*	Average	*StDev*	Average	*StDev*
Global	**26,239,755**	*11,986,824*	**22,964,579**	*11,901,504*	**48,317,797**	*16,879,852*	**53,198,650**	*22,631,857*
Africa	**3,776,768**	*1,300,227*	**4,862,038**	*2,092,590*	**6,955,777**	*1,996,716*	**8,953,533**	*2,998,404*
Northern Africa	726,704	*270,255*	1,654,480	*1,072,488*	1,174,516	*709,598*	2,189,453	*1,248,683*
Middle Africa	1,096,921	*779,031*	1,495,776	*599,422*	2,838,061	*795,438*	2,576,963	*1,784,210*
Western Africa	2,033,948	*511,900*	1,940,994	*781,998*	3,091,941	*886,848*	4,257,670	*799,519*
Eastern Africa	-94,052	*157,479*	-201,280	*174,422*	-304,840	*140,680*	-203,265	*293,994*
Southern Africa	13,248	*32,212*	-27,933	*50,175*	156,099	*55,742*	132,712	*28,917*
Europe	**1,057,203**	*827,424*	**2,198,234**	*1,684,662*	**3,773,348**	*1,475,357*	**3,157,297**	*1,589,179*
Northern Europe	319,531	*220,483*	616,477	*243,462*	536,062	*274,516*	1,554,861	*442,343*
Western Europe	1,944,772	*849,671*	2,749,223	*1,598,918*	2,592,941	*1,086,929*	1,896,116	*1,203,084*
Eastern Europe	-605,643	*488,255*	-284,984	*350,172*	1,429,208	*406,176*	-116,653	*291,702*
Southern Europe	-601,457	*354,738*	-882,483	*353,654*	-784,863	*247,088*	-177,027	*284,524*
North America	**4,006,605**	*131,072*	**5,205,753**	*91,505*	**6,791,637**	*153,567*	**3,252,150**	*211,235*
Northern America	4,006,605	*131,072*	5,205,753	*91,505*	6,791,637	*153,567*	3,252,150	*211,235*
Latin America and the Caribbean	**-278,593**	*1,796,910*	**-166,915**	*1,881,425*	**2,344,029**	*1,820,756*	**-258,084**	*3,060,003*
Central America	-231,274	*1,794,233*	-382,117	*1,934,581*	1,746,485	*1,749,343*	39,758	*2,821,734*
Caribbean†	-148,496	*44,503*	-217,717	*45,295*	-159,187	*33,931*	-125,803	*37,764*
South America	101,176	*221,385*	432,920	*342,932*	756,730	*299,920*	-172,039	*516,937*
Asia	**17,373,072**	*10,244,739*	**10,447,294**	*8,535,855*	**28,050,171**	*15,391,386*	**37,739,024**	*18,713,311*
Western Asia	962,871	*127,105*	988,978	*113,848*	1,068,785	*287,103*	900,229	*404,637*
South-Central Asia	-299,285	*1,423,366*	-508,087	*940,517*	1,551,761	*3,960,461*	16,414,279	*6,073,880*
Eastern Asia	10,722,804	*7,947,469*	4,955,685	*4,902,057*	16,812,375	*9,880,114*	9,812,356	*10,974,587*
South-Eastern Asia	5,986,682	*2,794,146*	5,010,717	*5,374,944*	8,617,250	*4,886,621*	10,612,161	*6,323,496*
Oceania	**304,700**	*32,643*	**418,175**	*31,271*	**402,835**	*39,157*	**354,729**	*42,660*
Australia and New Zealand	267,470	*29,242*	379,579	*24,678*	382,024	*31,260*	298,864	*19,950*
Melanesia†	41,527	*4,074*	44,213	*9,422*	24,837	*16,745*	61,792	*26,308*
Micronesia†	0	*0*	0	*0*	0	*0*	0	*0*
Polynesia†	-4,296	*1,236*	-5,617	*1,130*	-4,027	*1,733*	-5,926	*3,177*

China	**10,233,395**	*7,698,982*	**3,943,029**	*4,839,183*	**16,286,948**	*9,835,224*	**9,510,199**	*10,858,893*
USA	**5,434,782**	*72,574*	**5,958,090**	*65,495*	**8,857,790**	*53,429*	**2,030,171**	*365*

†Excludes small islands with populations under 300,000.

Table A10. Net Migration for Dryland Ecosystems*

Ecosystem/UN_region	Drylands							
	1970		1980		1990		2000	
	Average	*StDev*	Average	*StDev*	Average	*StDev*	Average	*StDev*
Global	**-10,516,579**	*4,951,097*	**-10,374,055**	*5,469,493*	**-24,167,885**	*6,996,494*	**-38,382,564**	*11,661,890*
Africa	**-5,052,978**	*641,987*	**-8,703,231**	*1,154,282*	**-8,139,583**	*1,427,998*	**-10,567,916**	*1,286,780*
Northern Africa	-1,460,461	*255,234*	-1,863,148	*374,156*	-3,289,764	*277,441*	-3,190,984	*275,067*
Middle Africa	221,673	*93,340*	129,277	*138,037*	792,765	*100,457*	284,039	*195,939*
Western Africa	-2,243,963	*524,035*	-3,250,066	*700,028*	-4,567,458	*969,115*	-5,081,854	*1,088,235*
Eastern Africa	-1,522,563	*166,287*	-3,496,587	*282,933*	-2,280,299	*442,718*	-2,412,802	*456,728*
Southern Africa	-47,665	*169,115*	-222,706	*271,876*	1,205,174	*404,397*	-166,315	*367,875*
Europe	**-2,345,018**	*696,242*	**-2,209,334**	*872,554*	**727,774**	*672,629*	**1,041,311**	*1,016,088*
Northern Europe	-64,393	*1,404*	-40,225	*5,464*	-96,897	*3,995*	-79	*272*
Western Europe	49,657	*62,258*	-81,528	*40,546*	-95,064	*29,126*	-3,561	*36,754*
Eastern Europe	-949,202	*175,801*	-143,563	*105,294*	1,647,874	*93,812*	8,976	*232,572*
Southern Europe	-1,381,080	*755,261*	-1,944,018	*928,279*	-728,138	*610,440*	1,035,976	*885,906*
North America	**4,181,696**	*77,726*	**3,891,953**	*76,619*	**6,907,934**	*108,591*	**2,678,959**	*131,611*
Northern America	4,181,696	*77,726*	3,891,953	*76,619*	6,907,934	*108,591*	2,678,959	*131,611*
Latin America and the Caribbean	**-3,901,348**	*2,825,701*	**-4,976,008**	*3,561,389*	**-9,041,412**	*3,608,942*	**-8,520,219**	*4,047,024*
Central America	-1,456,176	*2,698,110*	-1,734,821	*2,972,629*	-5,356,753	*2,648,881*	-3,587,938	*3,783,031*
Caribbean†	-161,540	*116,181*	8,184	*185,667*	165,775	*123,340*	36,442	*88,208*
South America	-2,283,632	*725,462*	-3,249,371	*965,465*	-3,850,434	*1,432,168*	-4,968,724	*1,378,847*
Asia	**-3,396,570**	*3,480,071*	**1,598,429**	*1,586,882*	**-14,598,690**	*4,715,871*	**-23,059,970**	*9,133,360*
Western Asia	-419,975	*220,028*	-1,253,807	*221,486*	-1,582,500	*625,247*	545,691	*1,140,480*
South-Central Asia	785,373	*1,447,820*	356,273	*1,225,423*	-11,318,740	*3,524,224*	-19,199,024	*7,791,012*
Eastern Asia	-4,671,966	*2,446,560*	1,827,890	*1,540,289*	-2,192,152	*2,840,159*	-4,585,595	*3,632,131*
South-Eastern Asia	909,998	*343,922*	668,073	*1,075,895*	494,701	*741,119*	178,958	*1,783,549*
Oceania	**-2,360**	*80,656*	**24,137**	*97,085*	**-23,908**	*100,948*	**45,271**	*87,228*
Australia and New Zealand	-6,042	*80,624*	19,686	*97,104*	-29,536	*102,309*	36,204	*89,834*
Melanesia†	3,682	*70*	4,451	*916*	5,628	*1,918*	9,067	*3,960*
Micronesia†								
Polynesia†								

China	**-4,650,766**	*2,447,093*	**1,839,184**	*1,539,493*	**-2,083,405**	*2,841,297*	**-4,585,784**	*3,632,297*
USA	**3,841,379**	*217*	**3,768,877**	*171*	**6,722,214**	*1,026*	**2,429,264**	*154*

* Includes Dry Subhumid, Semiarid, and Arid subsystems, but not Hyperarid sybsystems.

† Excludes small islands with populations under 300,000.

Table A11. Net Migration for Polar Ecosystems

Ecosystem/UN_region	Polar							
	1970		1980		1990		2000	
	Average	*StDev*	Average	*StDev*	Average	*StDev*	Average	*StDev*
Global	**-992,046**	*156,914*	**141,566**	*130,505*	**1,076,110**	*147,530*	**21,792**	*31,023*
Africa								
Northern Africa								
Middle Africa								
Western Africa								
Eastern Africa								
Southern Africa								
Europe	**-150,491**	*94,914*	**327,944**	*54,233*	**897,189**	*86,657*	**23,015**	*25,796*
Northern Europe	*83,475*	*105,423*	272,535	*73,020*	299,380	*100,587*	7,308	*20,581*
Western Europe								
Eastern Europe	*-233,966*	*82,300*	55,409	*104,953*	597,808	*110,270*	15,707	*14,700*
Southern Europe								
North America	**-841,317**	*88,643*	**-186,197**	*96,236*	**179,125**	*89,892*	**-1,102**	*9,557*
Northern America	-841,317	*88,643*	-186,197	*96,236*	179,125	*89,892*	-1,102	*9,557*
Latin America and the Caribbean								
Central America								
Caribbean								
South America								
Asia	**-239**	*34*	**-181**	*27*	**-204**	*0*	**-121**	*6*
Western Asia								
South-Central Asia	-239	*34*	-181	*27*	-204	*0*	-121	*6*
Eastern Asia								
South-Eastern Asia								
Oceania								
Australia and New Zealand								
Melanesia								
Micronesia								
Polynesia								

China								
USA	**756,780**	*63,343*	**817,267**	*56,074*	**2,662,604**	*40,245*	**7,033**	*0*

Table A12. Net Migration for Island Ecosystems*

Ecosystem/UN_region	Island							
	1970		1980		1990		2000	
	Average	*StDev*	Average	*StDev*	Average	*StDev*	Average	*StDev*
Global	**-2,974,210**	*562,559*	**-4,548,724**	*434,454*	**-3,575,448**	*483,853*	**-1,889,641**	*742,835*
Africa	**-96,520**	*20,407*	**-51,192**	*21,384*	**94,571**	*22,502*	**29,604**	*41,545*
Northern Africa	-10,921	*1,084*	-3,330	*1,967*	2,992	*2,394*	-9,886	*2,894*
Middle Africa	-38,308	*3,568*	12,283	*980*	20,164	*1,158*	-953	*3,094*
Western Africa	-50,239	*2,060*	-39,248	*3,672*	18,780	*3,964*	-32,781	*2,524*
Eastern Africa	2,951	*19,177*	-20,893	*24,563*	52,641	*26,060*	73,230	*41,376*
Southern Africa	-2	*1*	-3	*2*	-6	*2*	-5	*3*
Europe	**86,162**	*269,447*	**-476,781**	*274,244*	**1,068,005**	*239,220*	**2,728,675**	*277,818*
Northern Europe	135,048	*204,098*	63,807	*200,461*	1,280,721	*181,111*	2,628,328	*203,810*
Western Europe	-8,219	*9,907*	-27,601	*22,359*	-19,346	*14,286*	-10,830	*16,154*
Eastern Europe	30,340	*7,660*	47,347	*6,853*	9,943	*3,525*	5,615	*9,742*
Southern Europe	-71,008	*244,488*	-560,335	*310,977*	-203,313	*248,532*	105,561	*258,683*
North America	**-1,738,358**	*105,655*	**-394,731**	*99,196*	**-2,799,245**	*112,410*	**-1,250,022**	*92,195*
Northern America	-1,738,358	*105,655*	-394,731	*99,196*	-2,799,245	*112,410*	-1,250,022	*92,195*
Latin America and the Caribbean	**-994,083**	*50,475*	**-1,210,829**	*41,609*	**-1,091,994**	*20,990*	**-1,130,246**	*61,204*
Central America	5,105	*3,924*	7,795	*5,003*	6,800	*7,580*	-12,597	*12,477*
Caribbean†	-1,116,374	*40,323*	-1,380,492	*45,095*	-1,252,019	*28,032*	-1,236,072	*52,781*
South America	117,186	*21,149*	161,868	*12,392*	153,225	*22,240*	118,423	*31,025*
Asia	**-121,688**	*393,600*	**-2,223,978**	*287,066*	**-851,313**	*469,766*	**-2,310,161**	*655,626*
Western Asia	996	*21,949*	28,932	*32,341*	23,081	*17,492*	112,588	*37,573*
South-Central Asia	-402,230	*29,597*	-594,291	*49,877*	-697,333	*73,875*	-802,791	*77,833*
Eastern Asia	886,034	*403,611*	-1,311,874	*260,648*	2,480,595	*523,847*	1,066,651	*596,279*
South-Eastern Asia	-606,487	*87,749*	-346,744	*114,725*	-2,657,656	*134,491*	-2,686,609	*159,798*
Oceania	**-109,724**	*20,510*	**-191,213**	*27,489*	**4,528**	*32,496*	**42,510**	*28,162*
Australia and New Zealand	-17,564	*18,757*	-57,909	*27,031*	124,503	*30,867*	147,621	*20,744*
Melanesia†	-41,406	*4,605*	-79,601	*5,627*	-81,101	*3,551*	-61,568	*11,031*
Micronesia†	0	*0*	0	*0*	0	*0*	0	*0*
Polynesia†	-50,753	*2,016*	-53,703	*1,902*	-38,874	*2,635*	-43,543	*2,913*

China	**323,684**	*383,410*	**-1,046,509**	*261,939*	**1,968,561**	*521,790*	**734,250**	*593,107*
USA	**-121,480**	*54*	**756,692**	*94*	**-59,764**	*317*	**-1,234,387**	*75*

† Excludes small islands with populations under 300,000.

* See also Table A13 for islands smaller than 300,000 persons.

Table A13. Net Migration for Most Small Island States Excluded in the HYDE Rates

Country	NM1970s	NM1980s	NM1990s	NM2000s
Bermuda	66	0	1,464	1,605
Cook Islands	-7,402	-4,199	-5,909	-6,701
Fed. State of Micronesia	-9,000	-2,000	-16,000	-19,000
Guam	-3,000	1,000	-8,000	1,000
Maldives	-11	47	32	109
Marshall Islands	-5,758	363	-7,061	-4,086
Northern Mariana Islands	-7,810	14,807	13,358	-30,622
Nauru	-1,187	-419	-1,863	-2,758
Palau	-756	-173	2,523	329
French Polynesia	6,000	2,000	1,000	2,000
Saint Helena	343	-242	0	0
Seychelles	-2,955	-5,176	-1,960	920
Tuvalu	196	-245	-796	-918
Wallis and Futuna	-284	292	-1,198	-1,002
Greenland	37	0	-5,484	-3,448
Total	-31,521	6,056	-29,894	-62,572

Sources: United Nations *World Population Prospects 2008* and US Census Bureau International Database

Figure A1. Estimated Net Migration for All Ecosystems

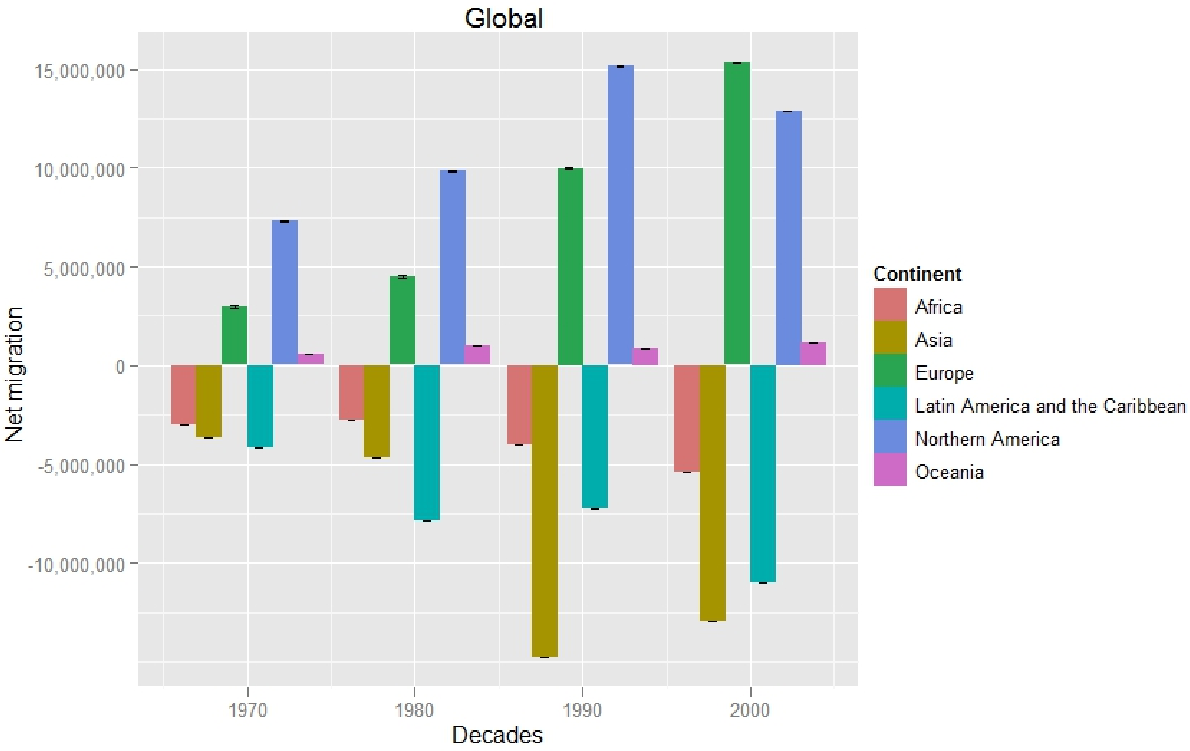


Figure A2. Estimated Net Migration for the Coastal Ecosystem (error bars represent the 95% confidence intervals of model run outputs)

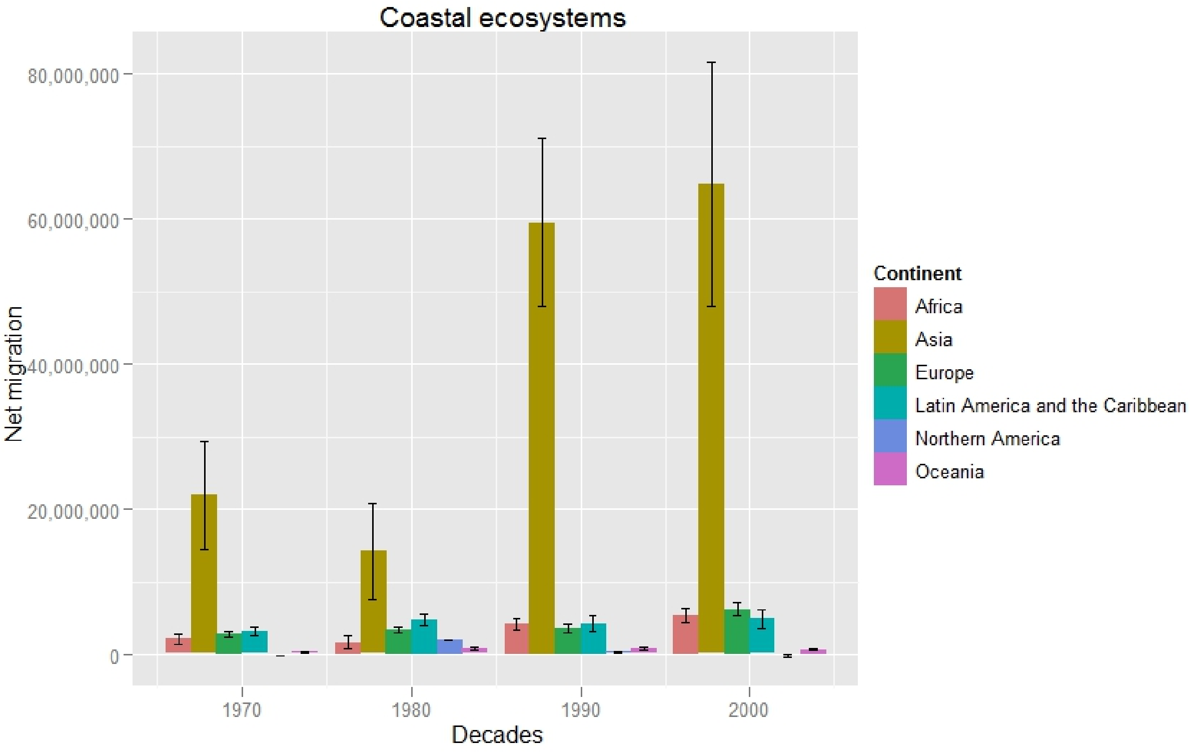
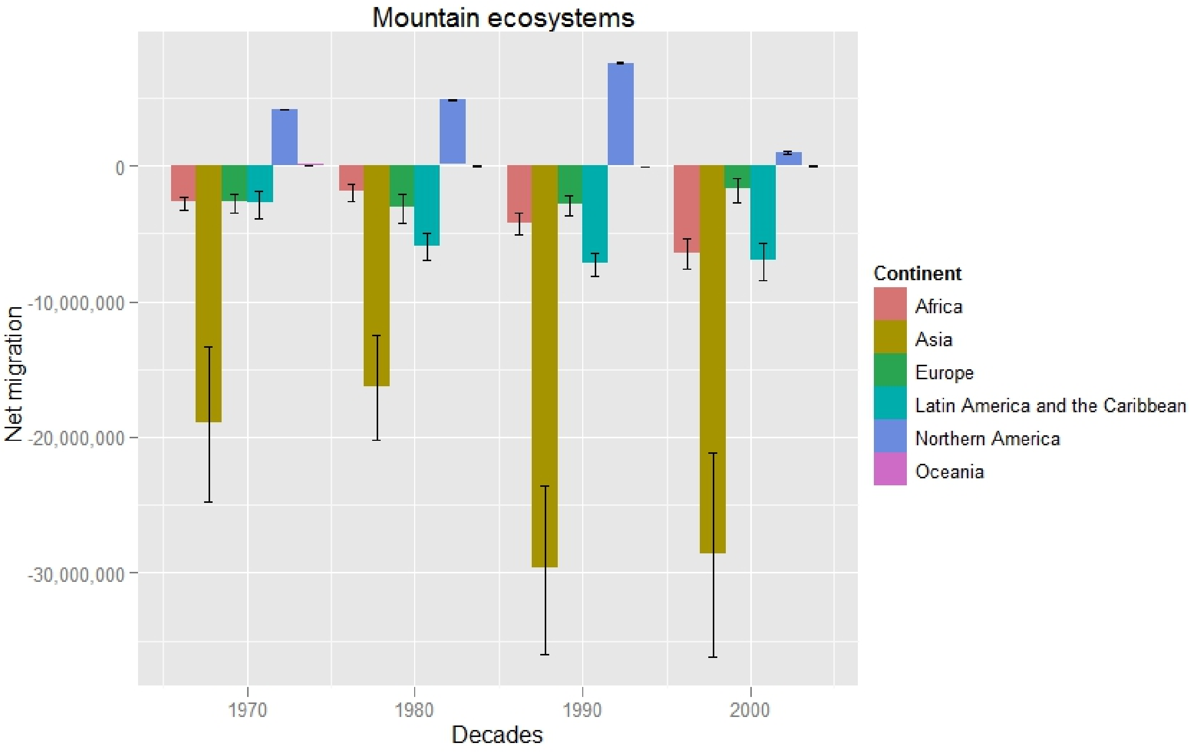
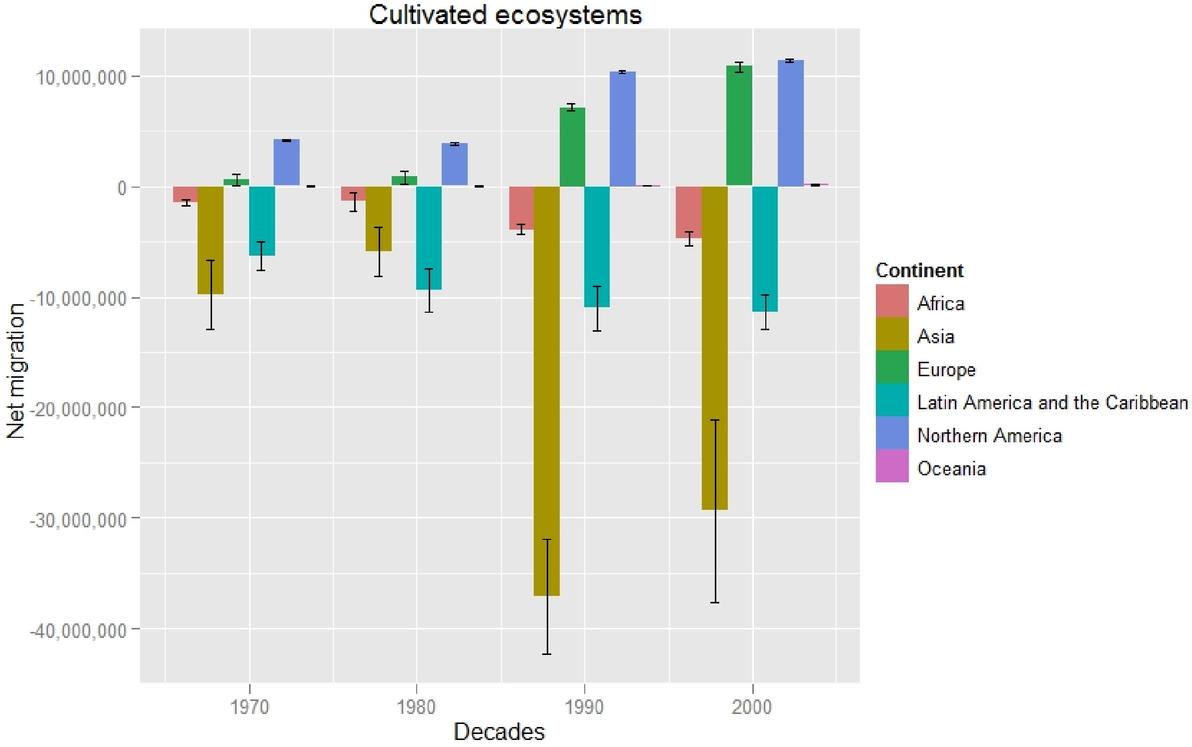


Figure A3. Estimated Net Migration for the Mountain Ecosystems* (error bars represent the 95% confidence intervals of model run outputs)



* Includes only Humid tropical upper montane, Humid temperate upper montane and pan-mixed, Humid temperate alpine/nival, Humid tropical alpine/nival, Dry cool temperate montane, Dry boreal/subalpine, Dry subpolar/alpine, and Polar/nival. Lower and lower/mid montane and hill ecosystems were removed.

Figure A4. Estimated Net Migration for the Cultivated Ecosystems* (error bars represent the 95% confidence intervals of model run outputs)



* Includes all but Pasture lands.

Figure A5. Estimated Net Migration for the Forest Ecosystems (error bars represent the 95% confidence intervals of model run outputs)

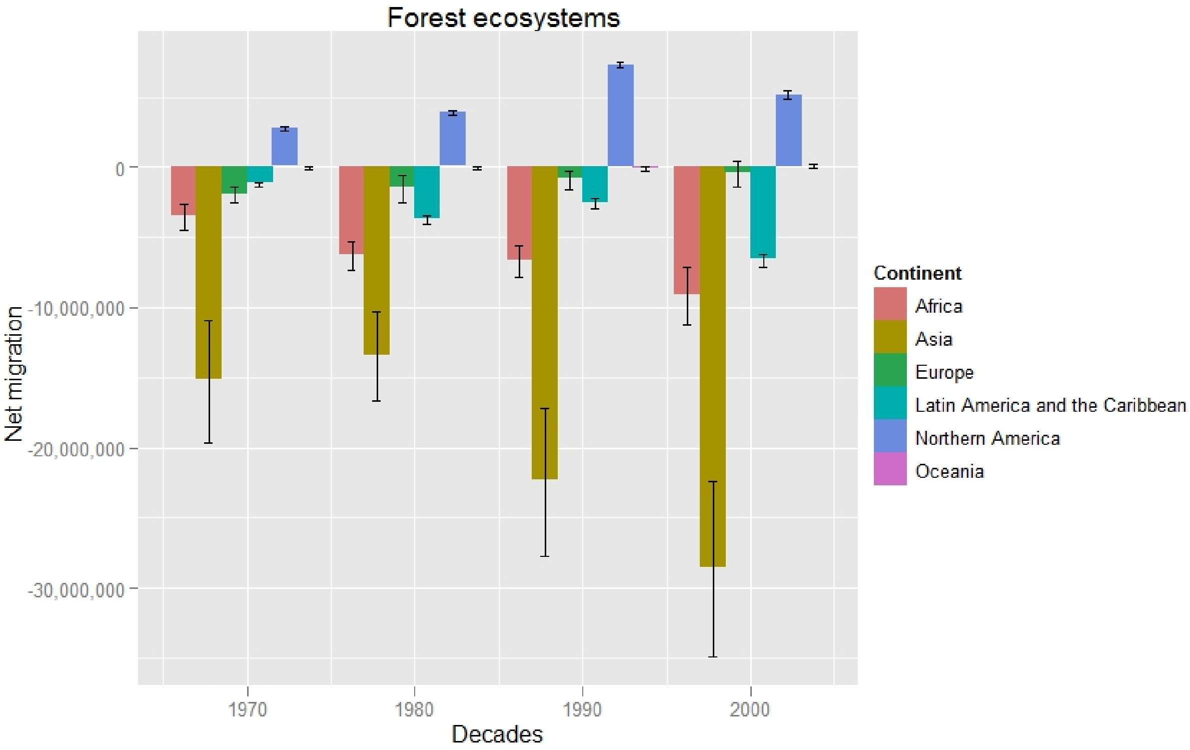


Figure A6. Estimated Net Migration for the Inland Water Ecosystems (error bars represent the 95% confidence intervals of model run outputs)

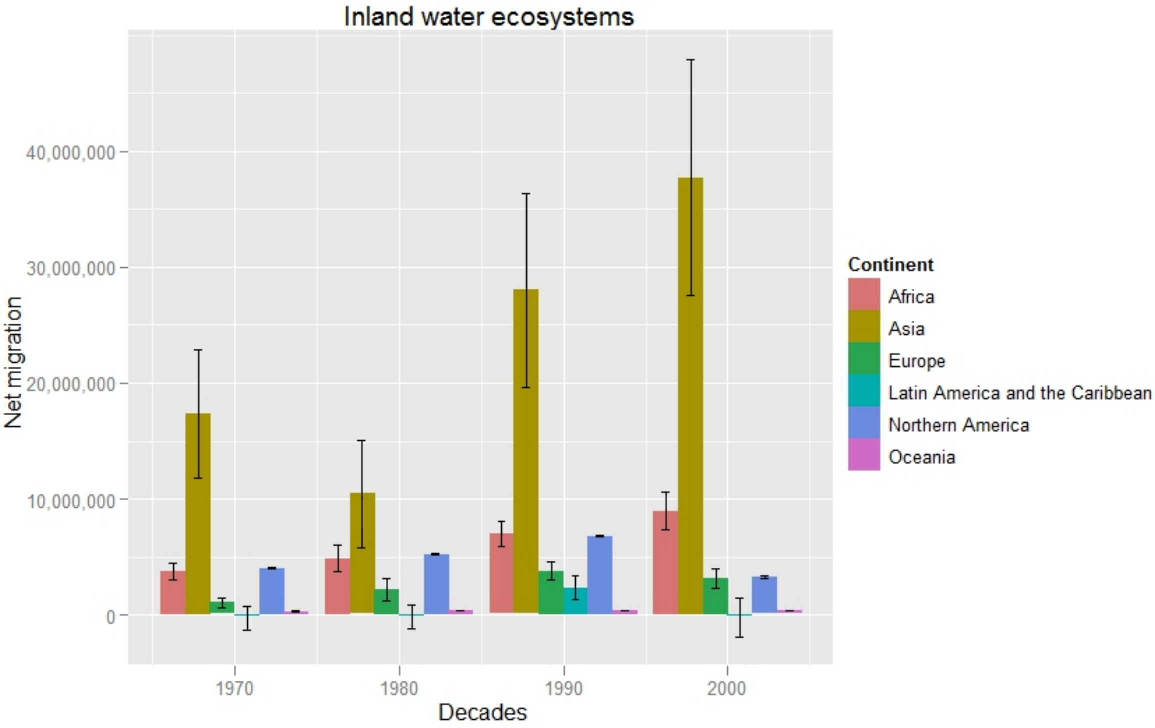
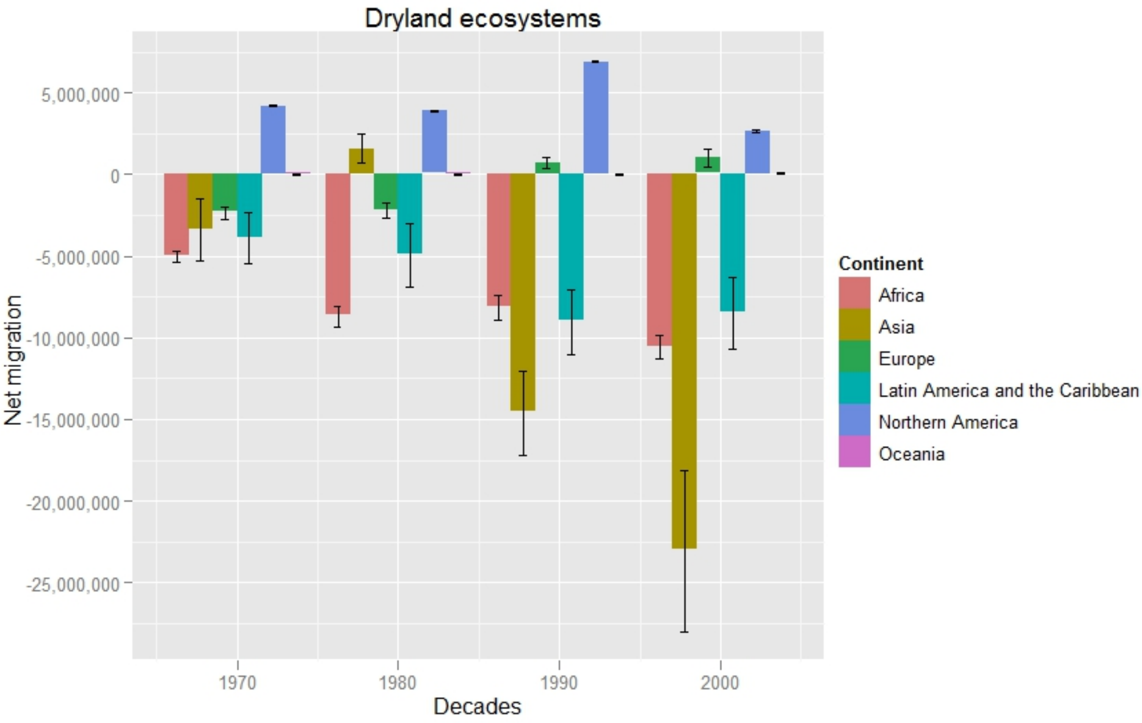
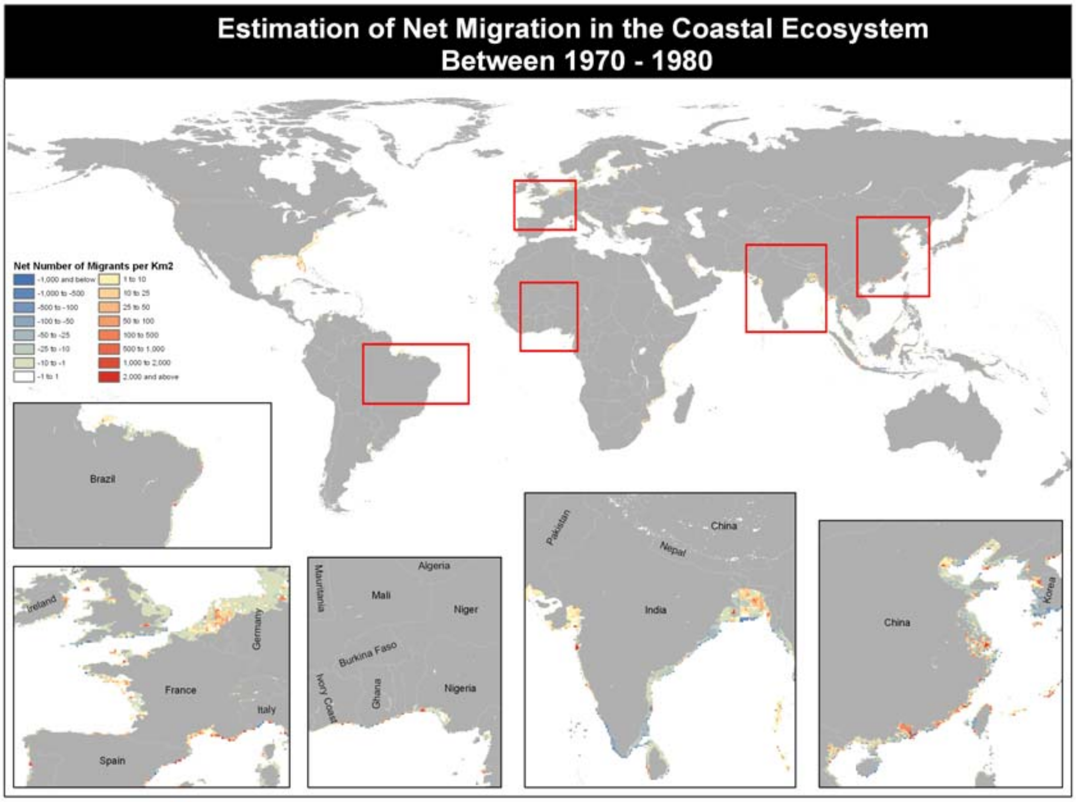


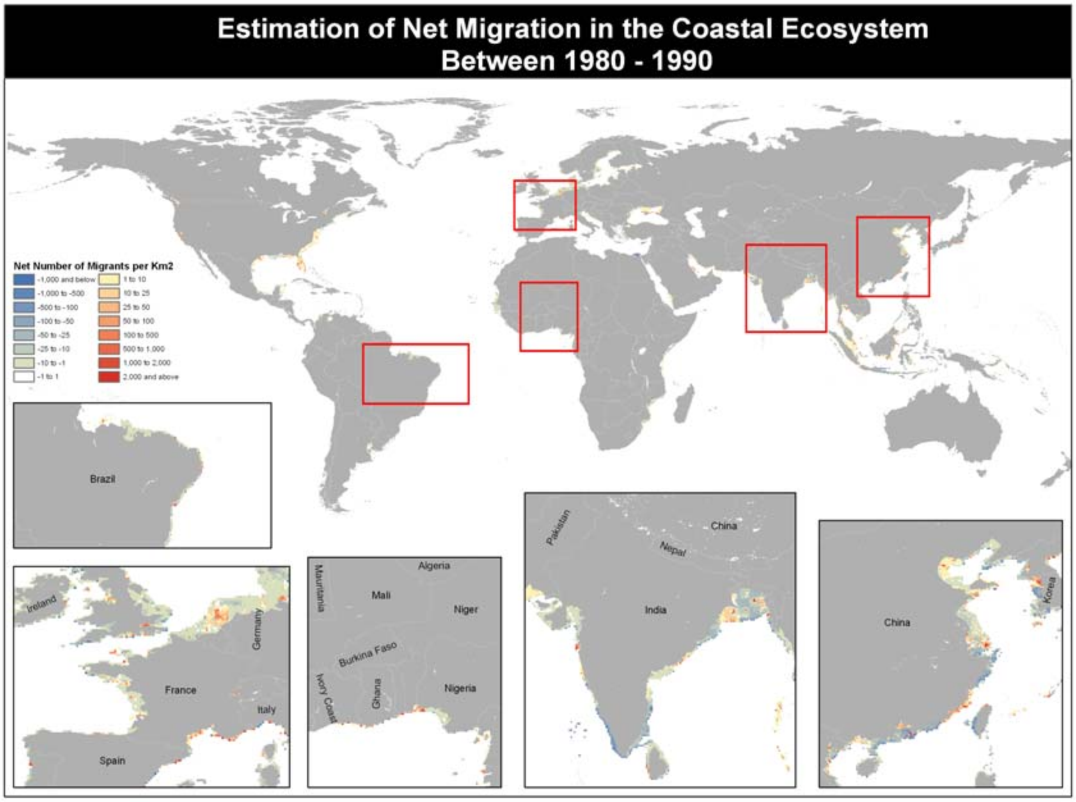
Figure A7. Estimated Net Migration for the Dryland Ecosystems* (error bars represent the 95% confidence intervals of model run outputs)

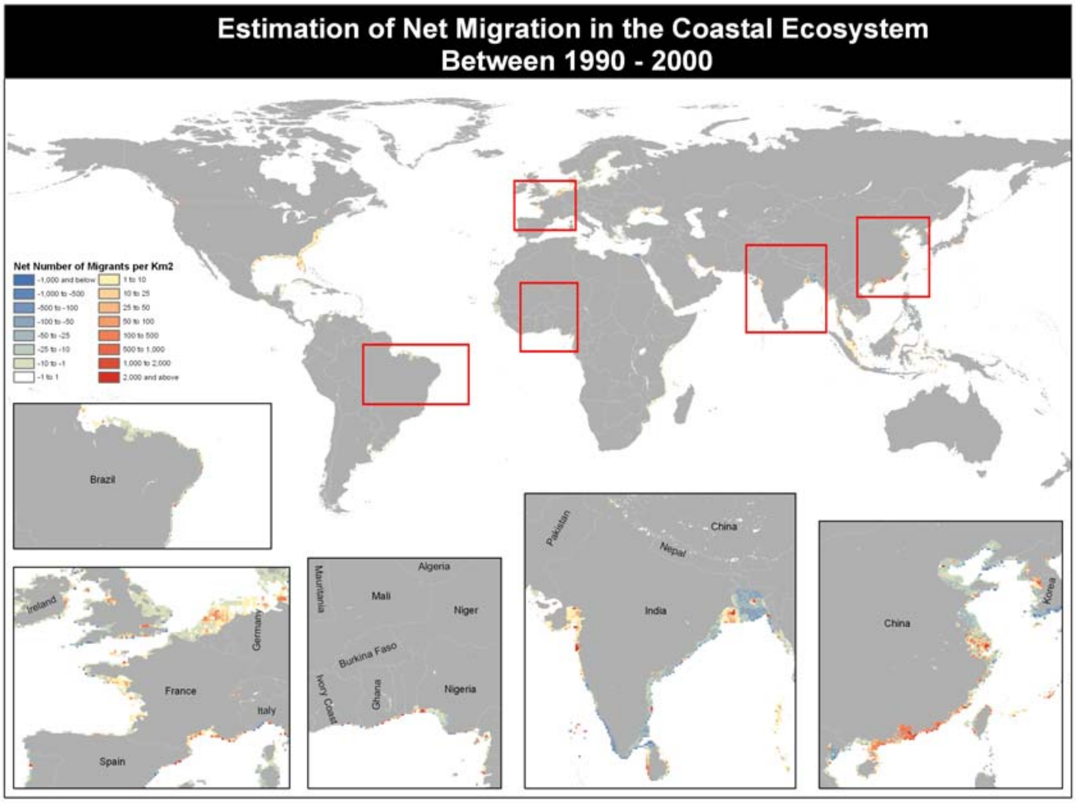


* Does not include Hyperarid subsystems.

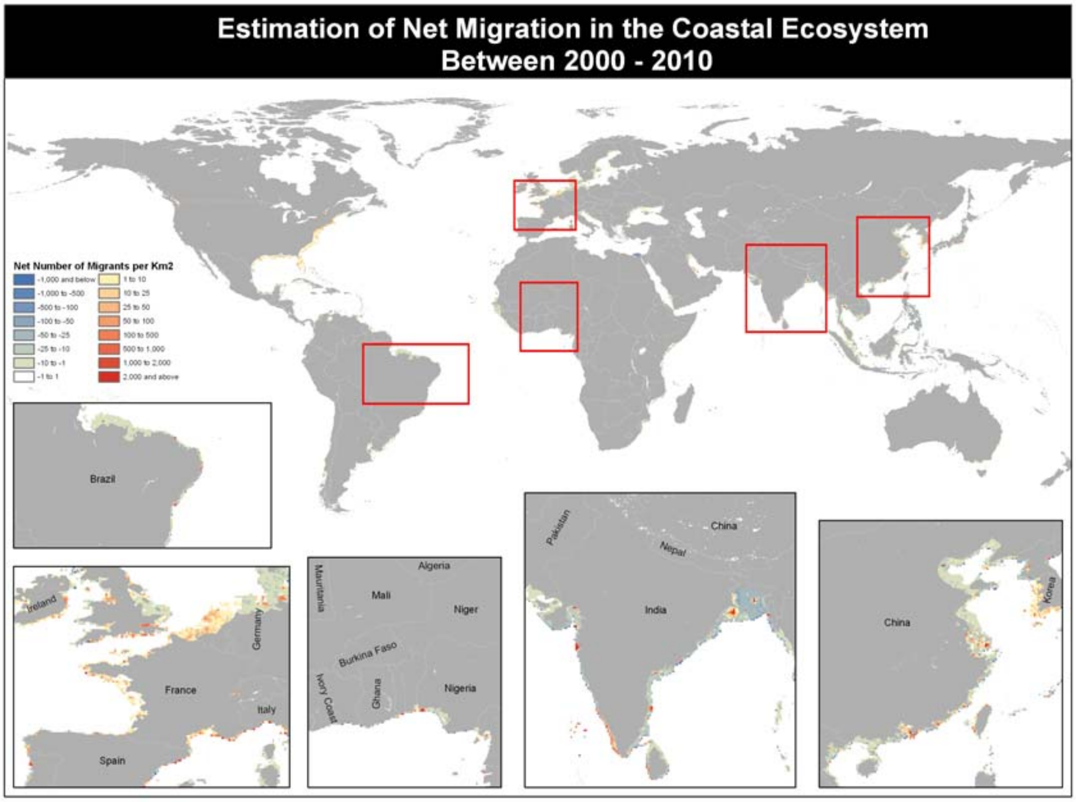
Map Set A1. Maps of Estimated Net Migration in Coastal Ecosystems by Decade







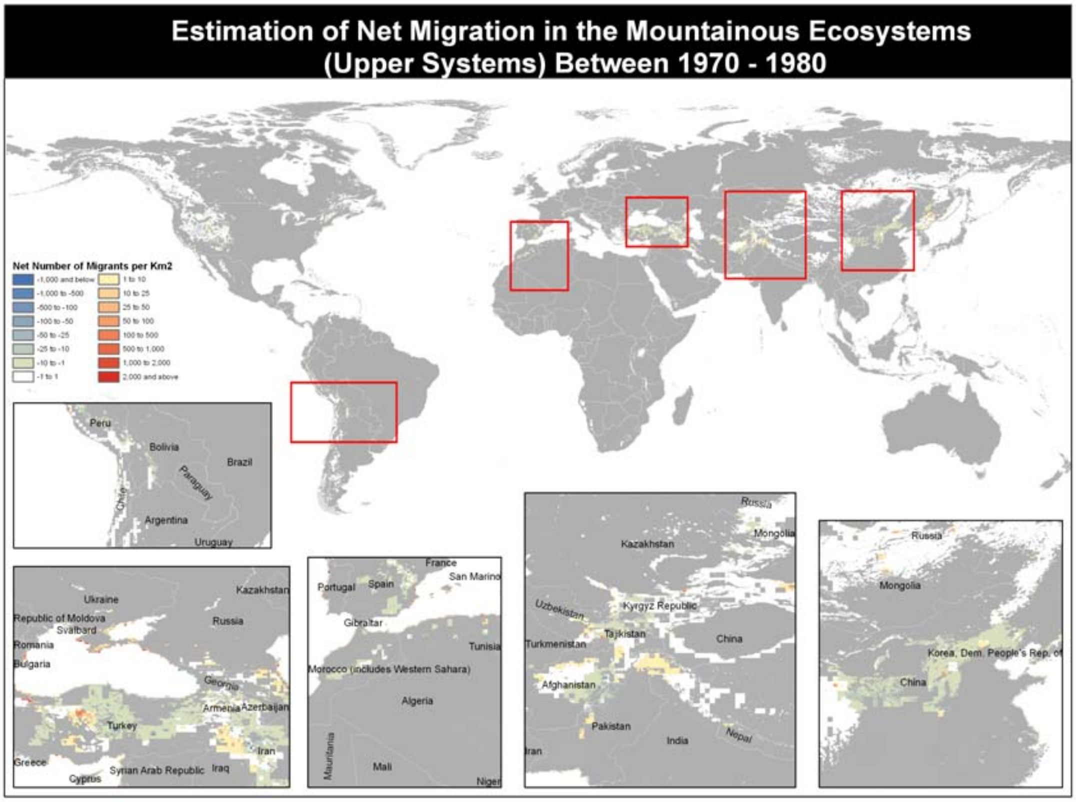
Estimation of Net Migration in the Coastal Ecosystem Between 1990 - 2000

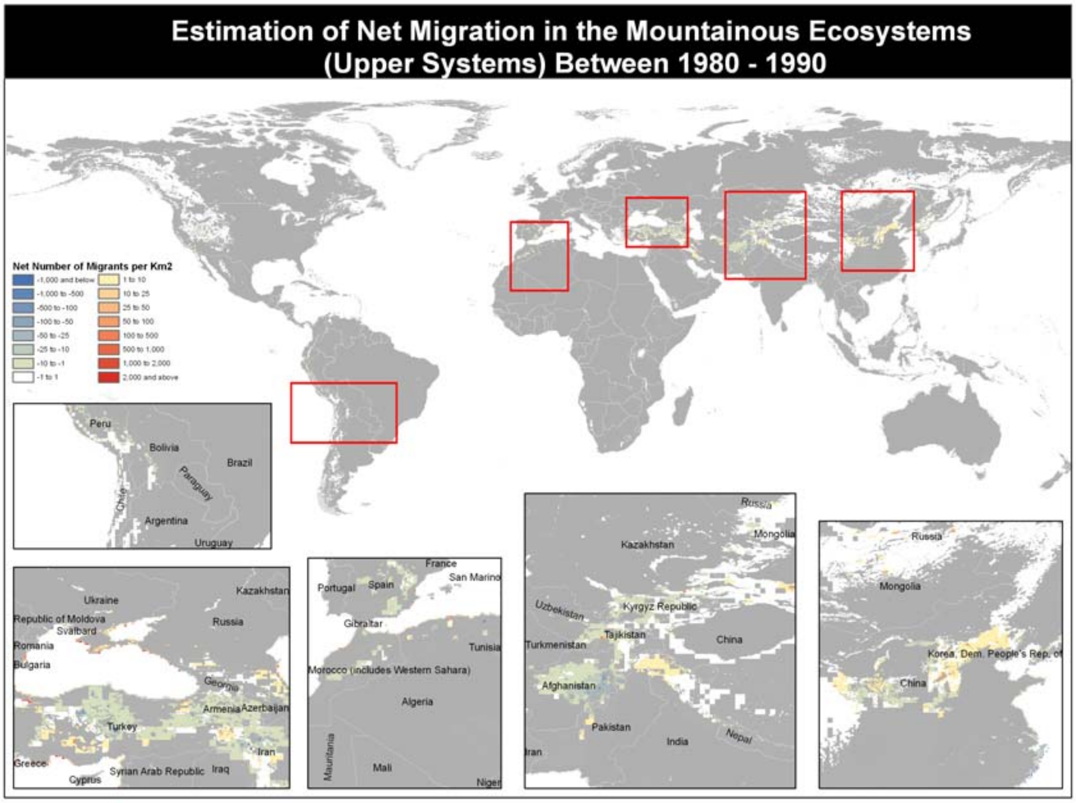


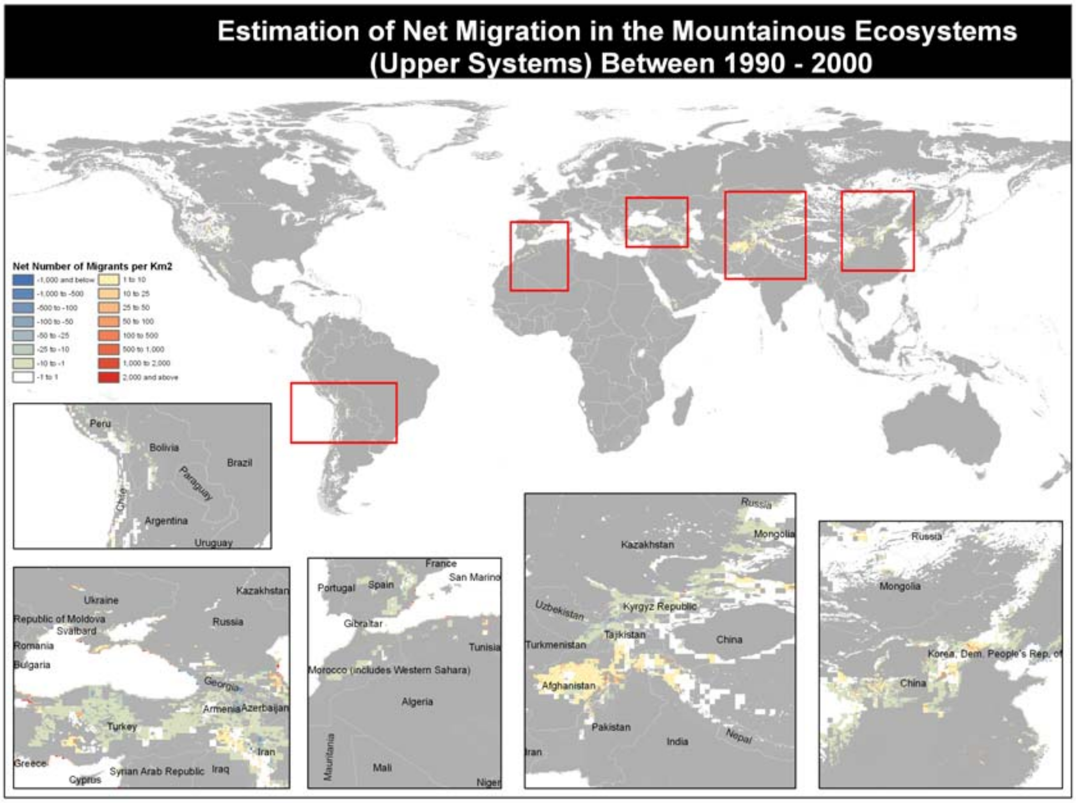
Estimation of Net Migration in the Coastal Ecosystem Between 2000 - 2010

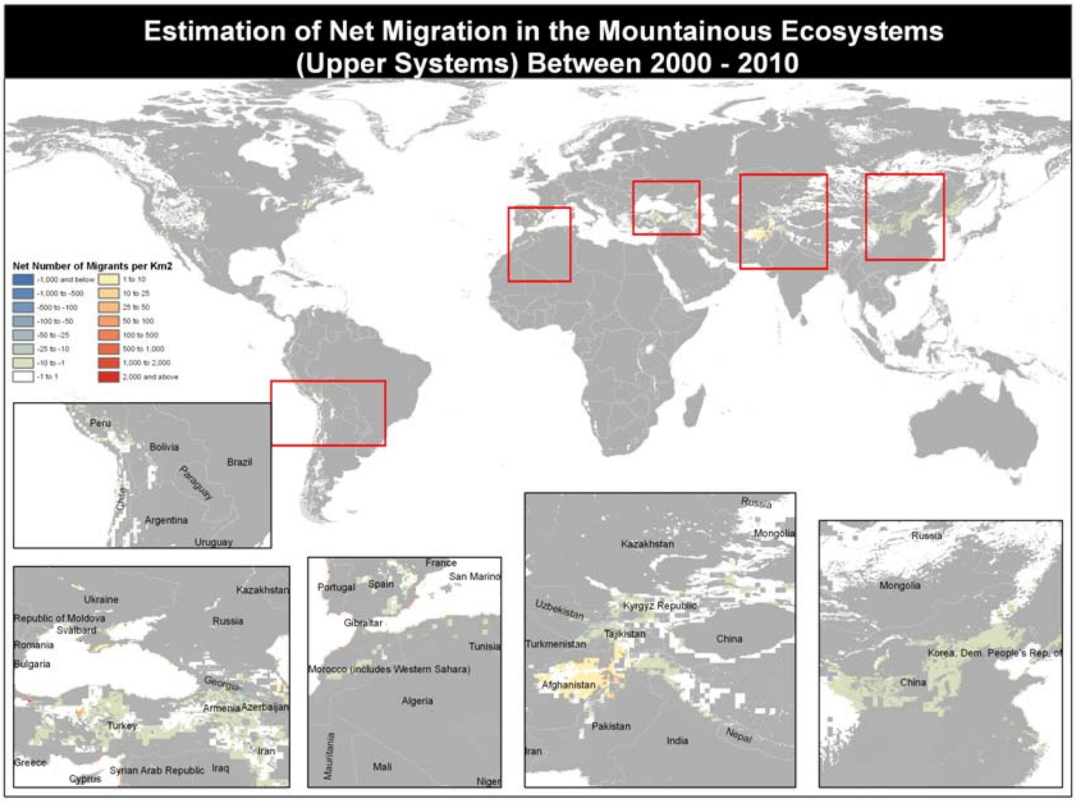
Note: 2000-2010 NM estimates are not based on observed population distributions in 2010.

Map Set A2. Maps of Estimated Net Migration in Mountain Ecosystems by Decade



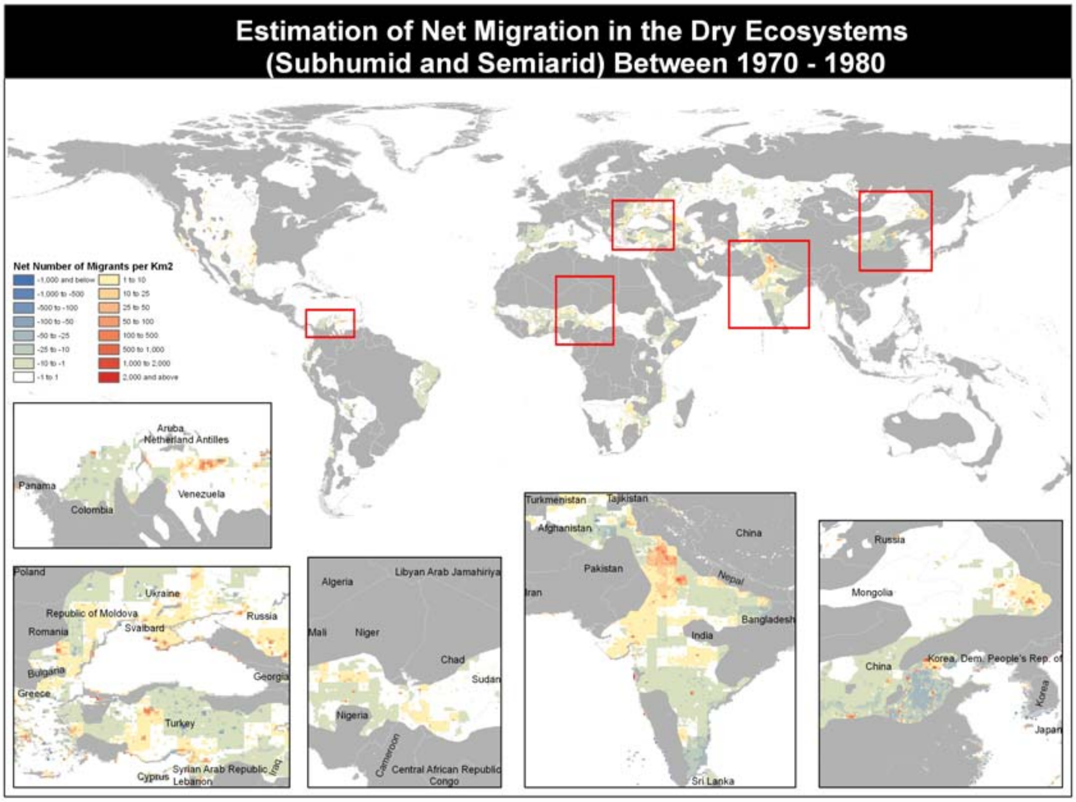


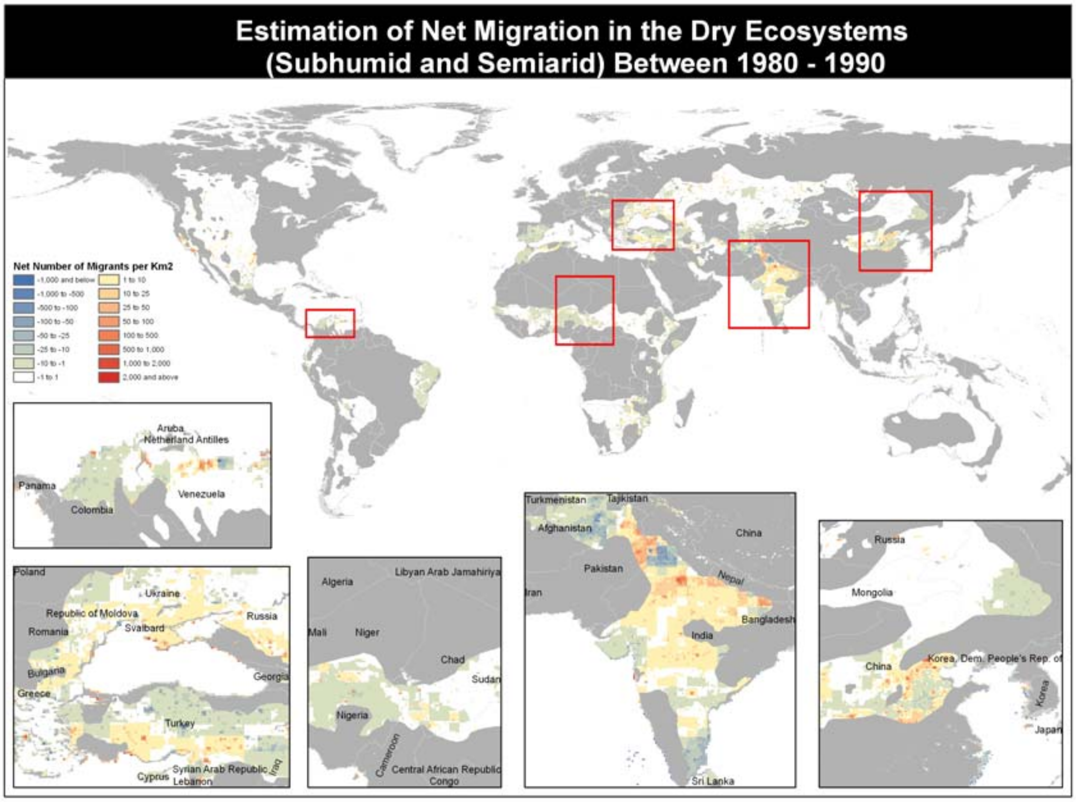


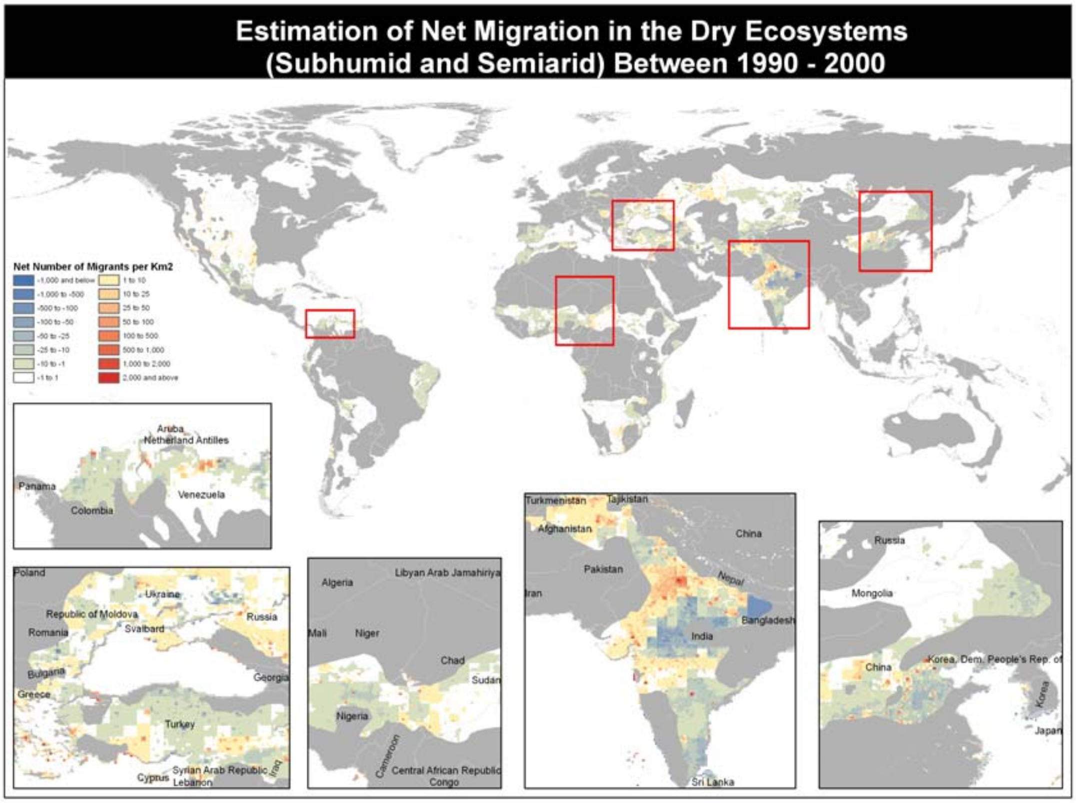


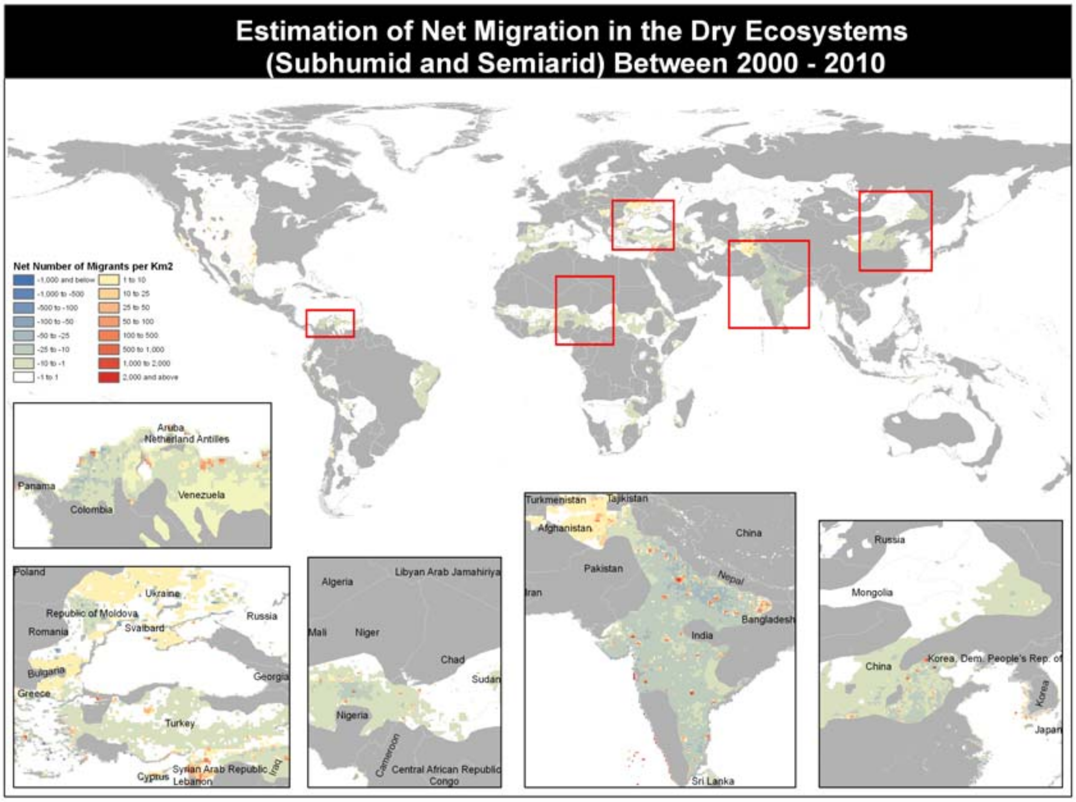
Note: 2000-2010 NM estimates are not based on observed population distributions in 2010.

Map Set A3. Maps of Estimated Net Migration in Dryland Ecosystems by Decade



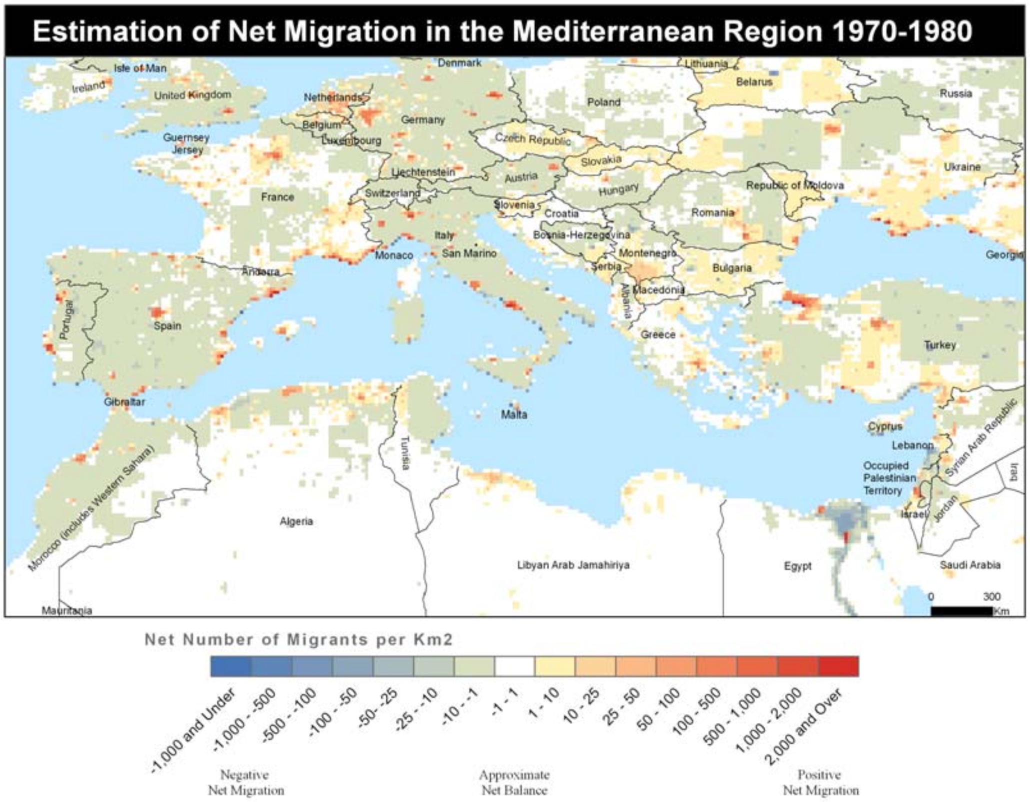


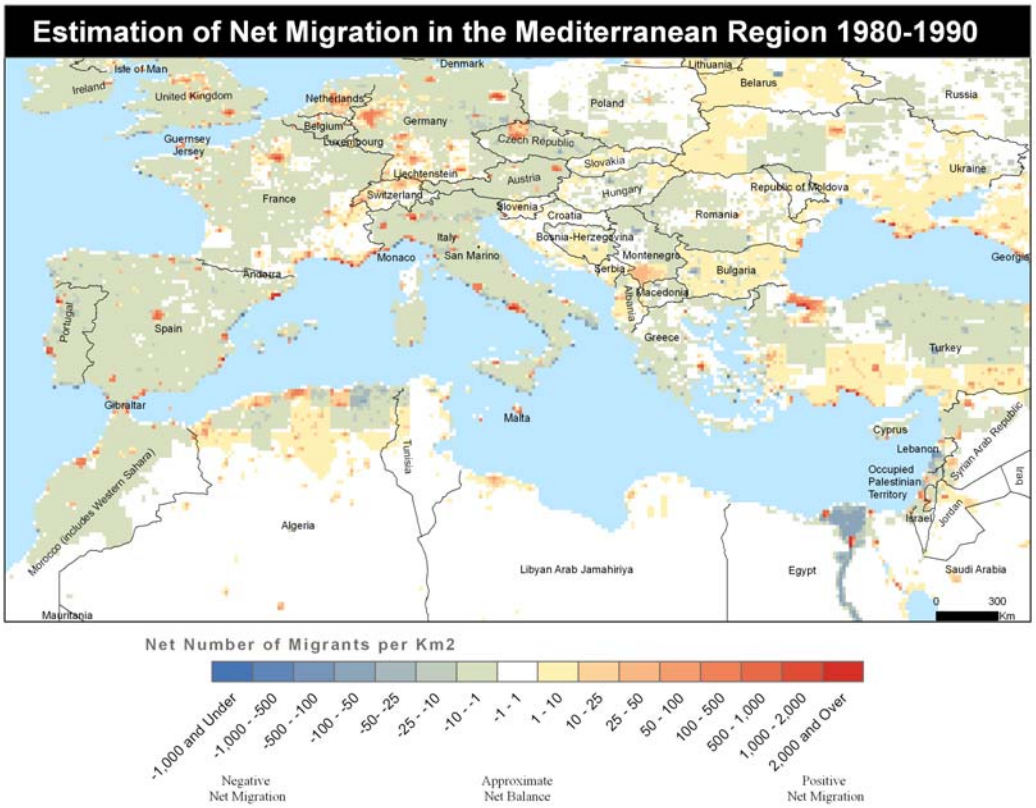


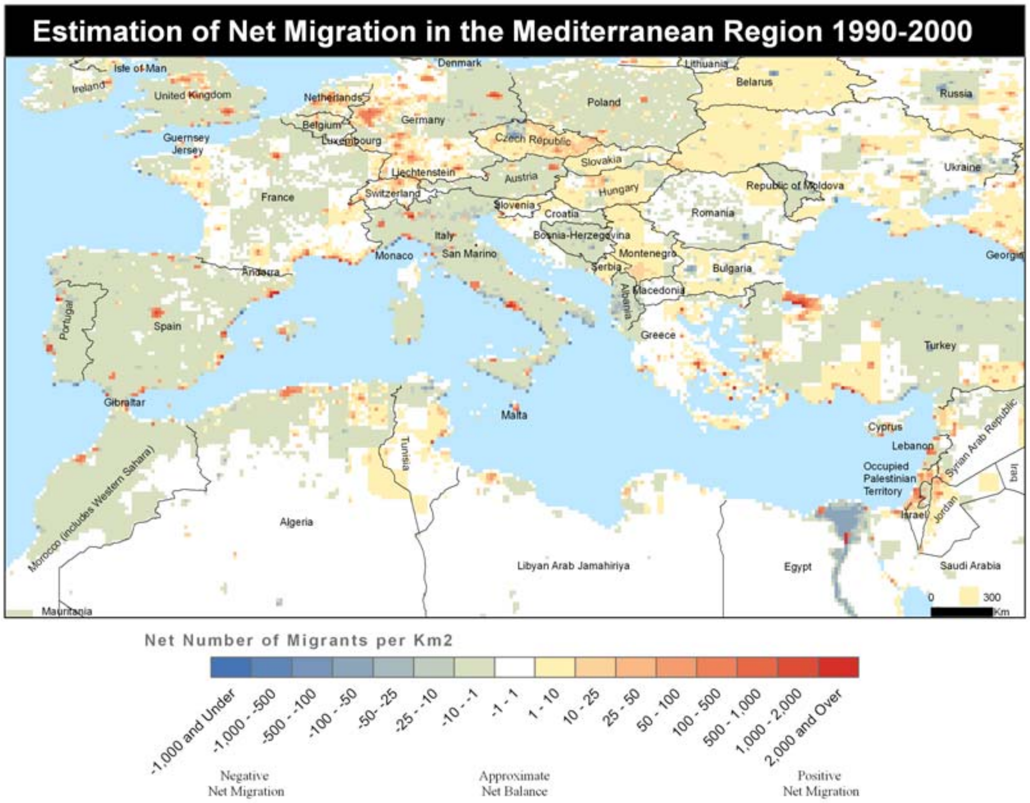


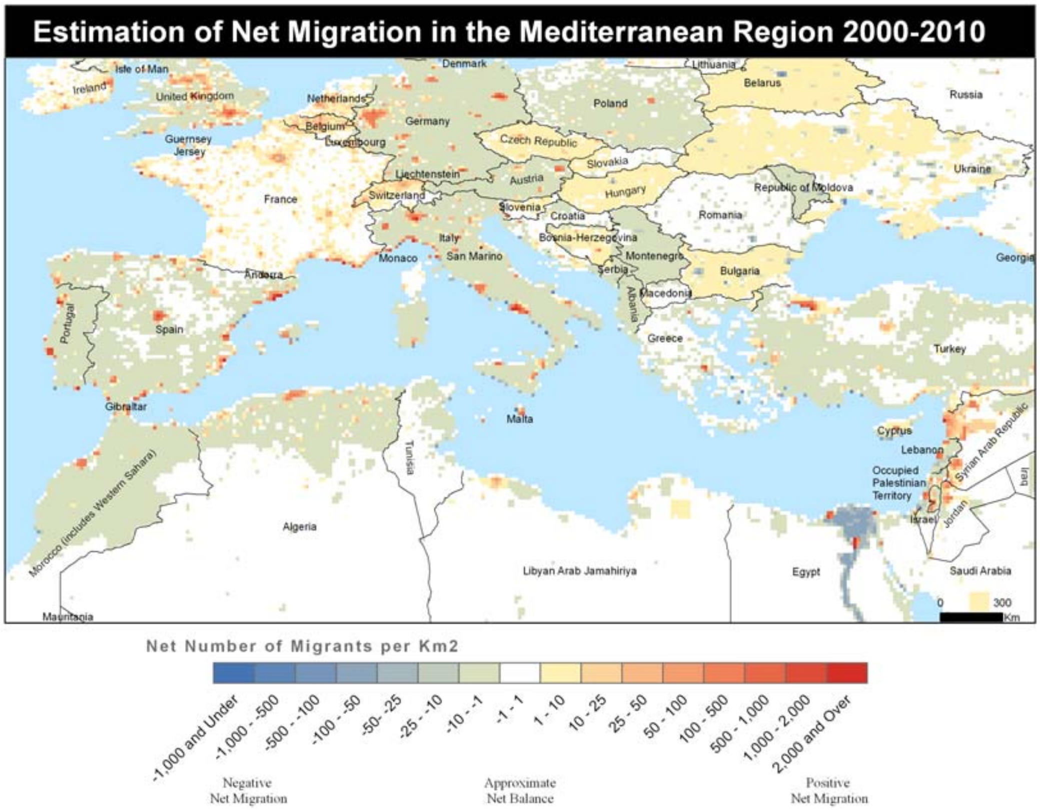
Note: 2000-2010 NM estimates are not based on observed population distributions in 2010.

Map Set A4. Mediterranean Estimated Net Migration by Decade









Note: 2000-2010 NM estimates are not based on observed population distributions in 2010.

Figure A8. Mediterranean Estimated Net Migration for the Coastal Ecosystem (error bars represent the 95% confidence intervals of model run outputs)

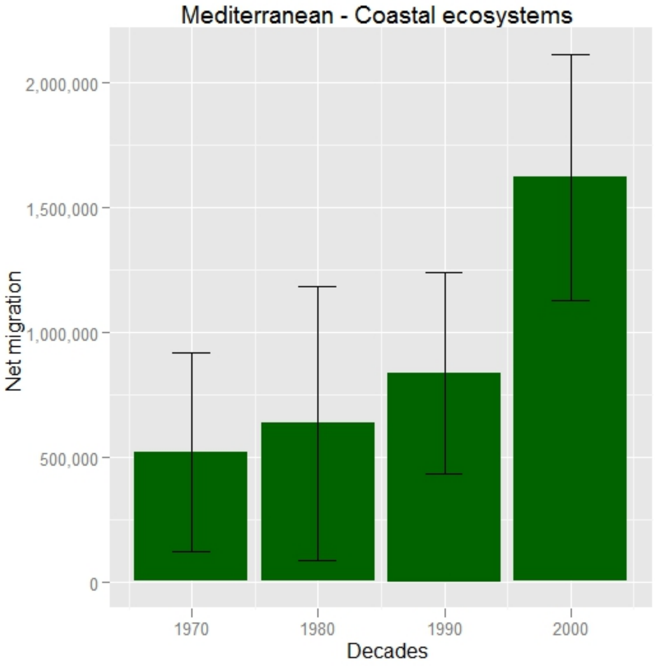
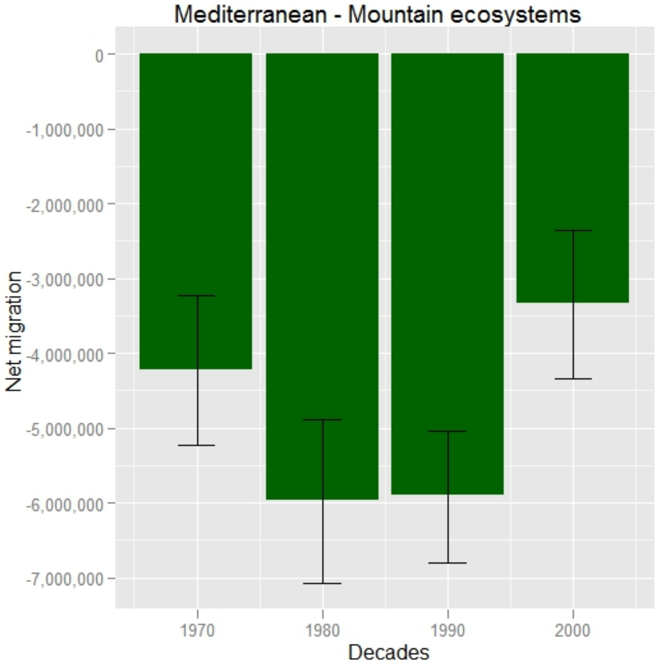
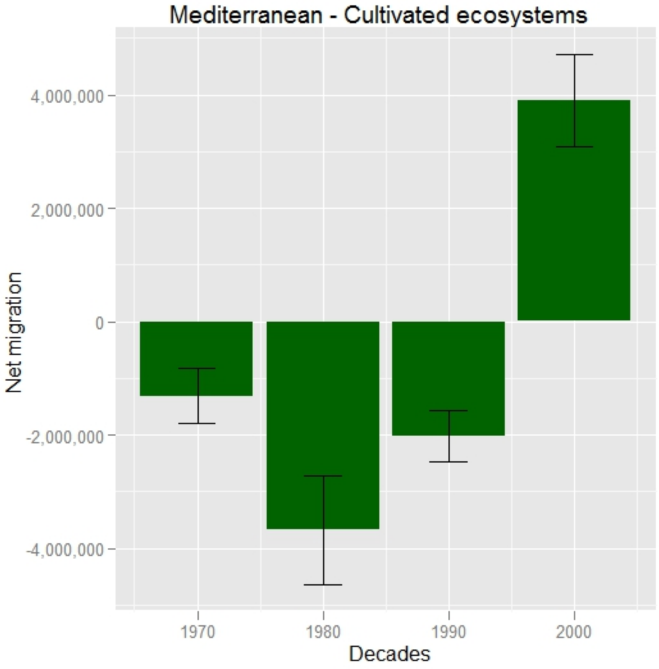


Figure A9. Mediterranean Estimated Net Migration for the Mountain Ecosystem* (error bars represent the 95% confidence intervals of model run outputs)



* Includes only Humid tropical upper montane, Humid temperate upper montane and pan-mixed, Humid temperate alpine/nival, Humid tropical alpine/nival, Dry cool temperate montane, Dry boreal/subalpine, Dry subpolar/alpine, and Polar/nival. Lower and lower/mid montane and hill ecosystems were removed.

Figure A10. Mediterranean Estimated Net Migration for the Cultivated Ecosystem* (error bars represent the 95% confidence intervals of model run outputs)



* All but pasture.

Figure A13. Mediterranean Estimated Net Migration for the Forest Ecosystem (error bars represent the 95% confidence intervals of model run outputs)

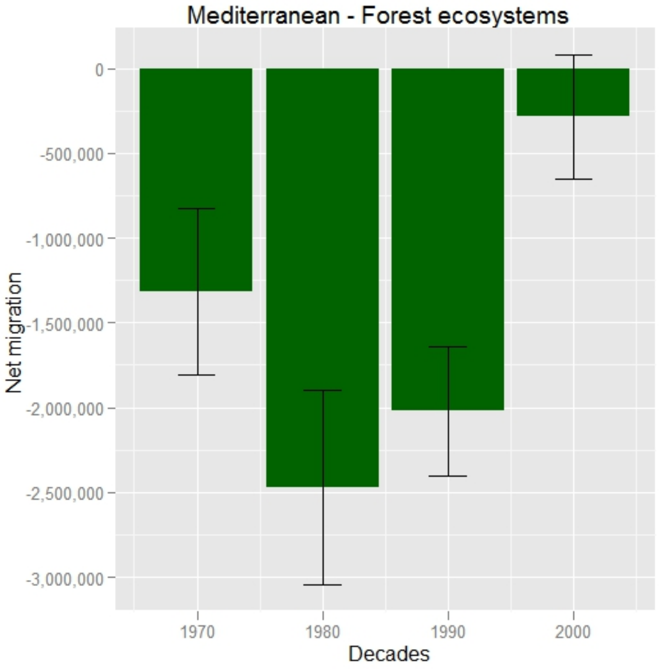


Figure A14. Mediterranean Estimated Net Migration for the Inland Water Ecosystem (error bars represent the 95% confidence intervals of model run outputs)

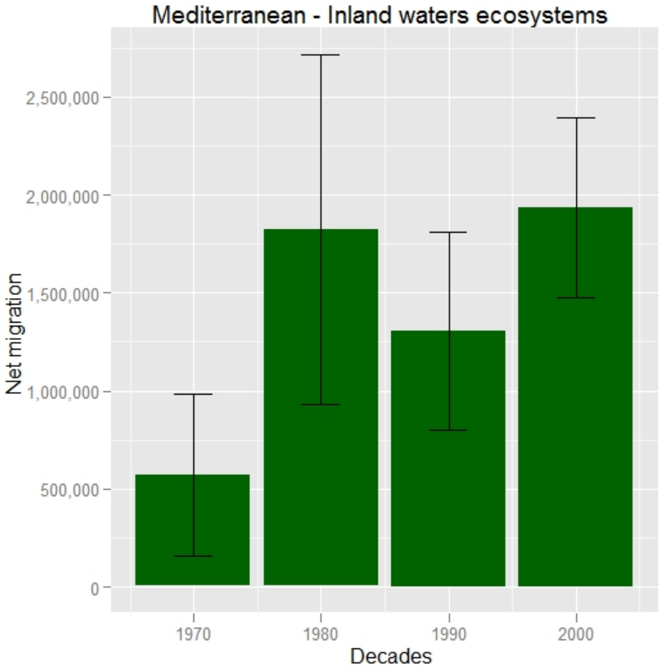
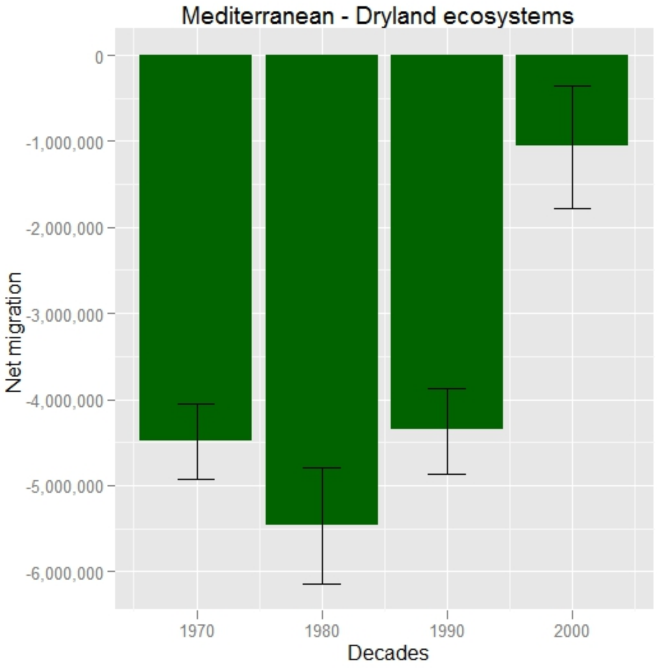


Figure A15. Mediterranean Estimated Net Migration for the Dryland Ecosystem* (error bars represent the 95% confidence intervals of model run outputs)



* Does not include the hyperarid ecosystem

Map A5. Comparison of Net Migration Estimates Using Imputed RNIs vs. Observed RNIs, China, 1990s

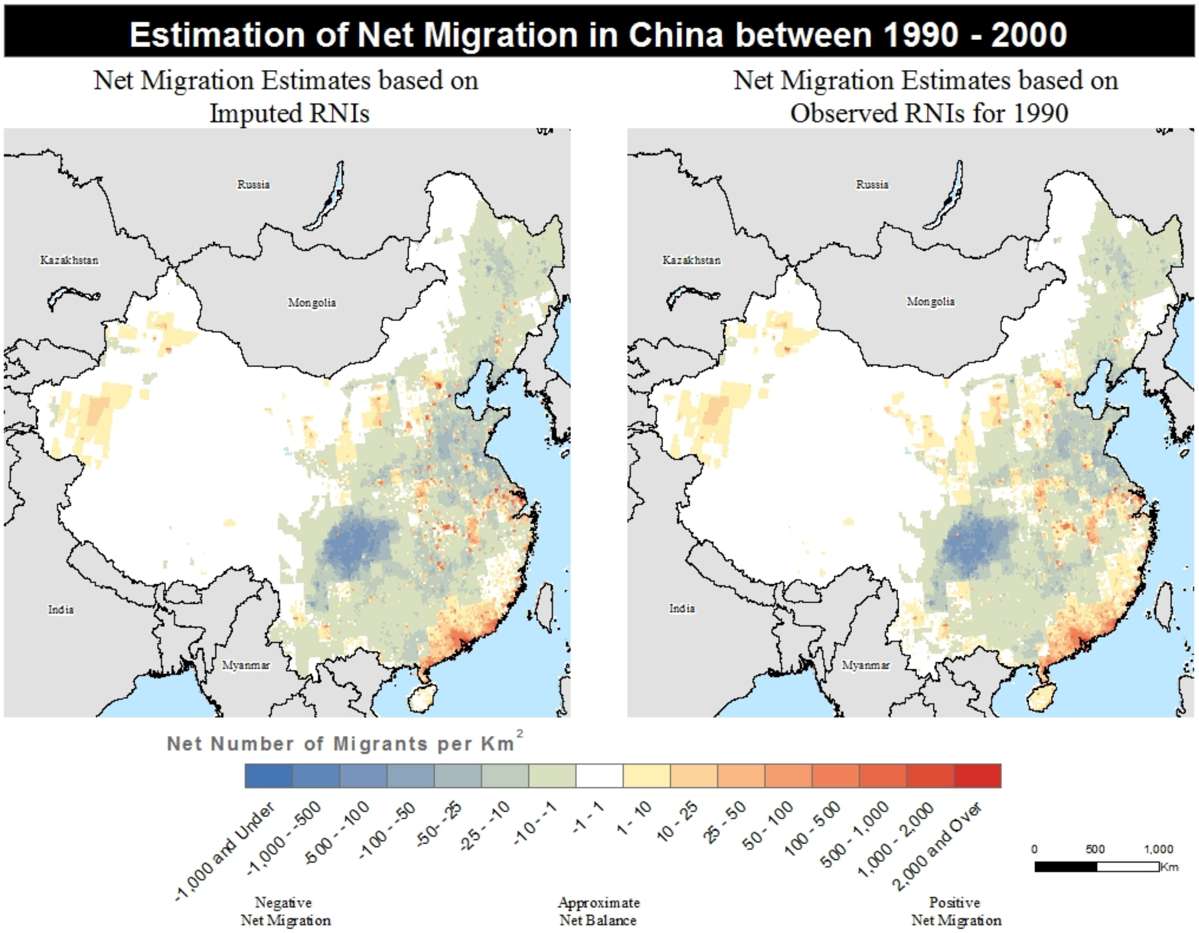
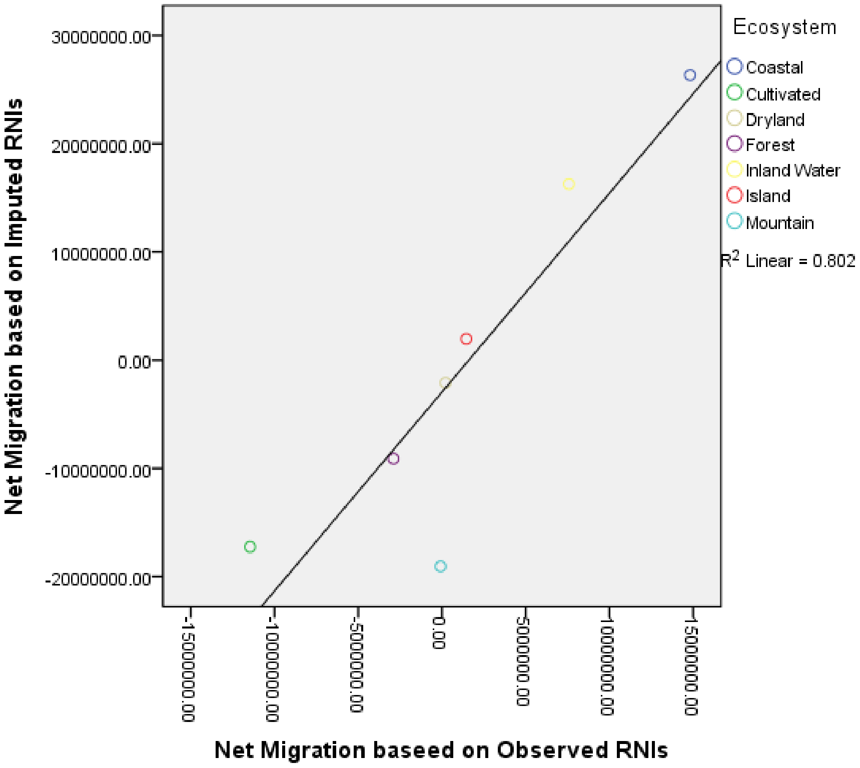


Table A14. China Net Migration Estimates based on Observed RNIs and Imputed RNIs, for the Decade of the 1990s

Ecosystem	NM Based on Observed RNIs (mean)	NM Based on Observed RNIs (SD)	NM Based on Imputed RNIs (mean)	NM Based on Imputed RNIs (SD)	Observed Within SD?
Coastal	14,819,400	13,054	26,344,387	13,634,899	yes
Mountain	-85,562	1,633	-19,057,800	9,718,136	no
Cultivated	-11,463,876	38,303	-17,248,133	8,608,419	yes
Forest	-2,892,429	9,568	-9,099,173	7,002,821	yes
Inland Water	7,584,715	14,769	16,279,880	9,833,004	yes
Dryland	207,745	8,967	-2,088,252	2,842,969	yes
Island	1,454,086	1,193	1,967,995	521,636	yes

Figure A16. Scatter Plot of China NM Estimates based on Observed RNIs and Imputed RNIs, for the Decade of the 1990s



Map A6. Comparison of Net Migration Estimates, US Census Bureau and Modeled Data, 1990s

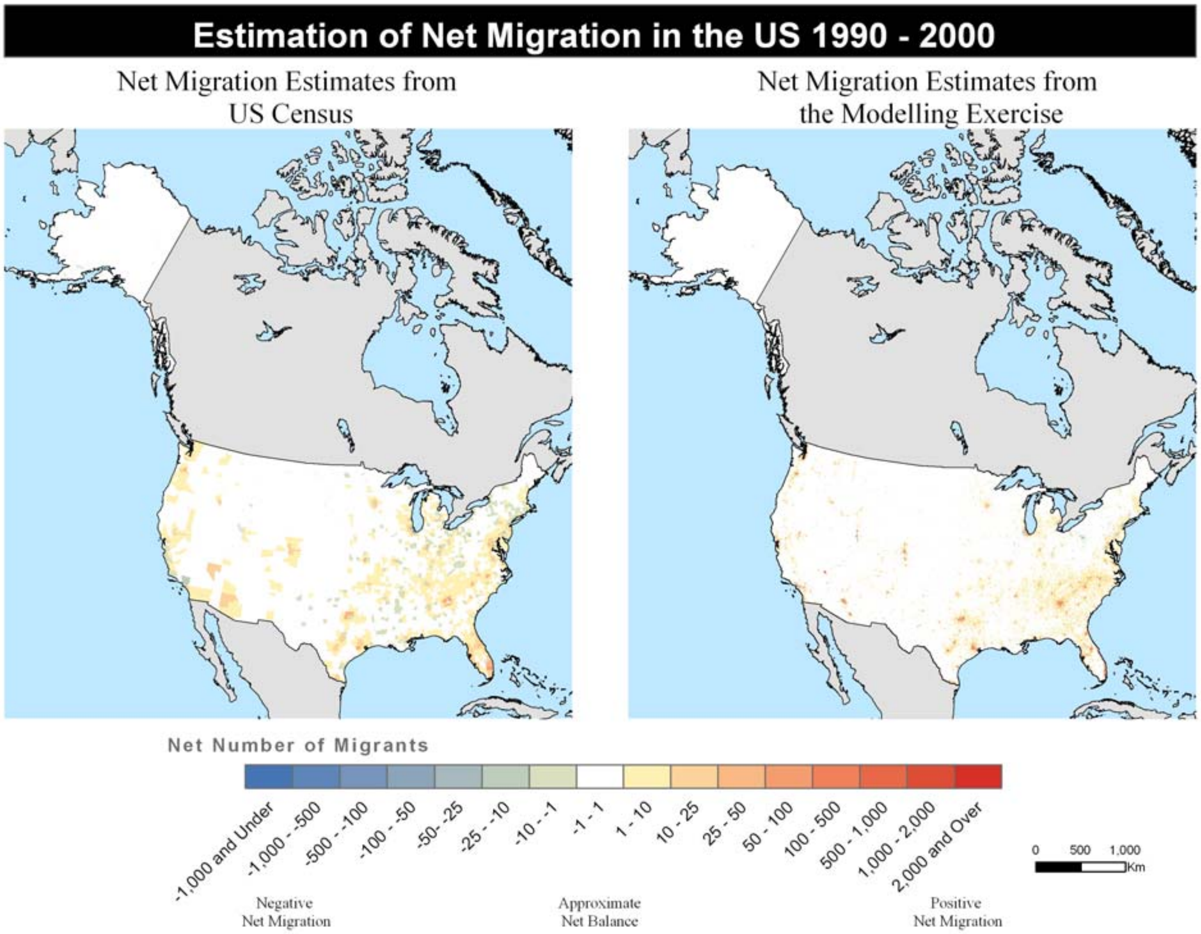
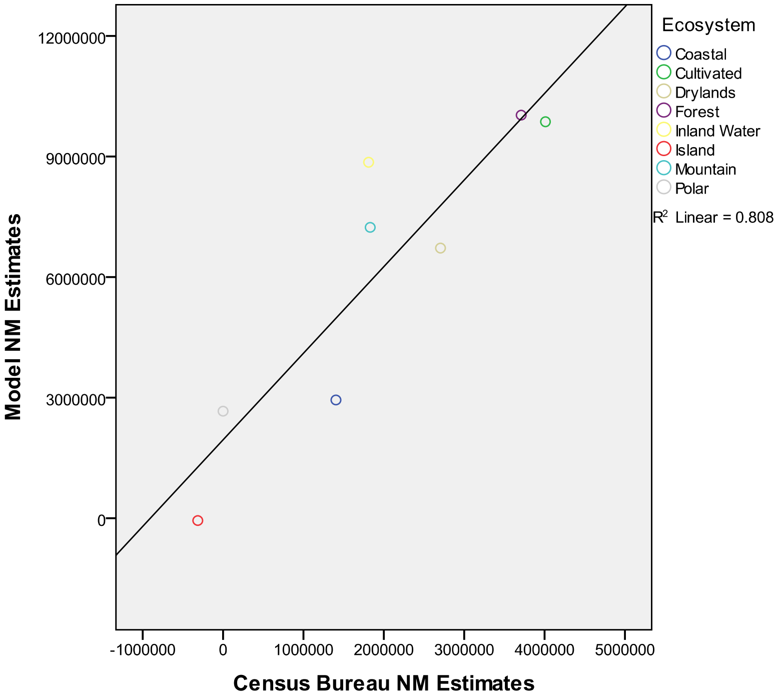


Table A15. US Net Migration by Ecosystem, US Census Compared Estimates to Modeled Estimates, decade of the 1990s

Ecosystem	US Census NM	Model NM (mean)
Coastal	1,404,241	2,941,103
Mountain	1,830,652	7,237,223
Cultivated	4,011,386	9,865,971
Forest	3,709,322	10,030,185
Inland Water	1,811,048	8,857,790
Dryland	2,705,433	6,722,214
Polar	-329	2,662,604
Island	-314,599	-59,764

Figure A17. Scatter Plot of NM Estimates from the US Census Bureau and Model Ouputs for Ecosystems, Decade of the 2000s



Map A7. Comparison of Net Migration Estimates, US Census Bureau and Modeled Data, 2000s

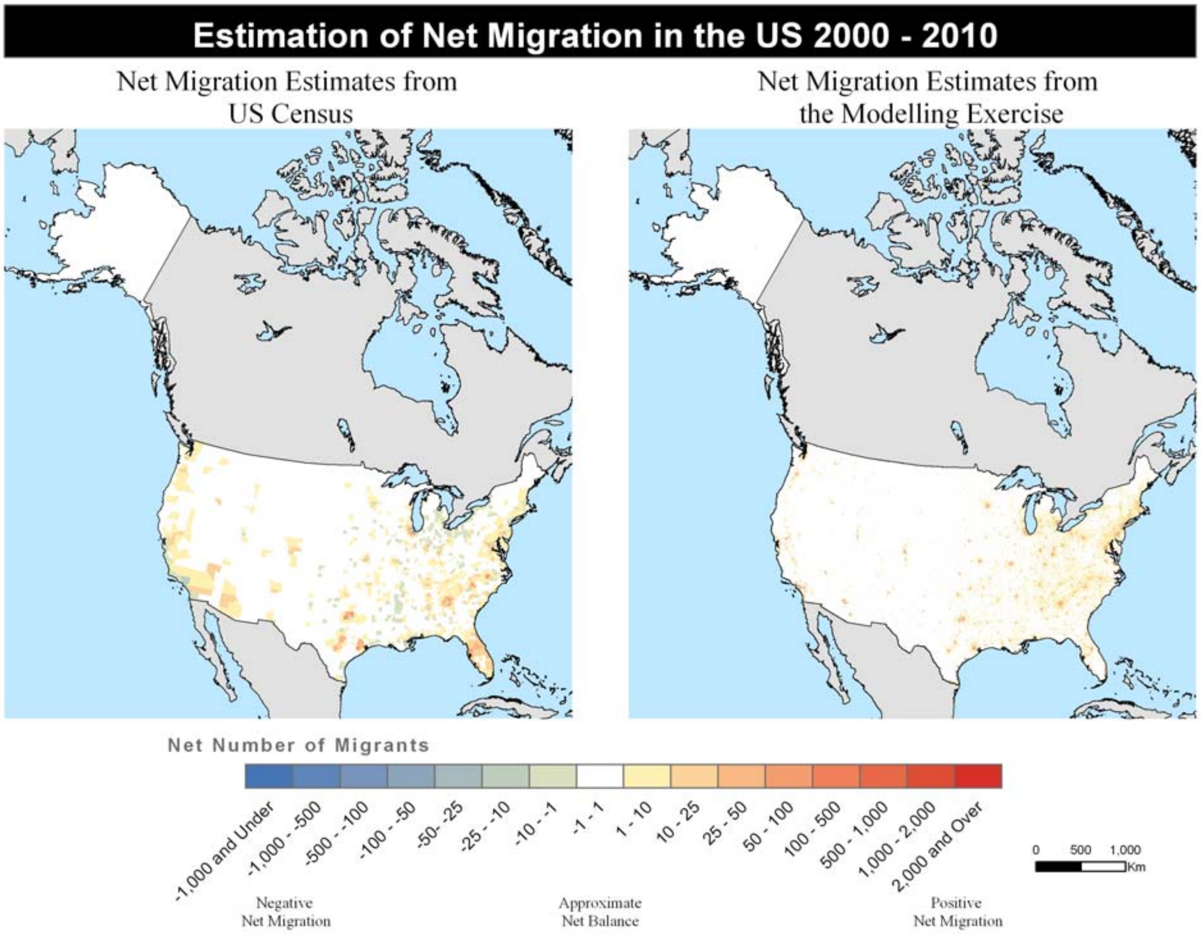
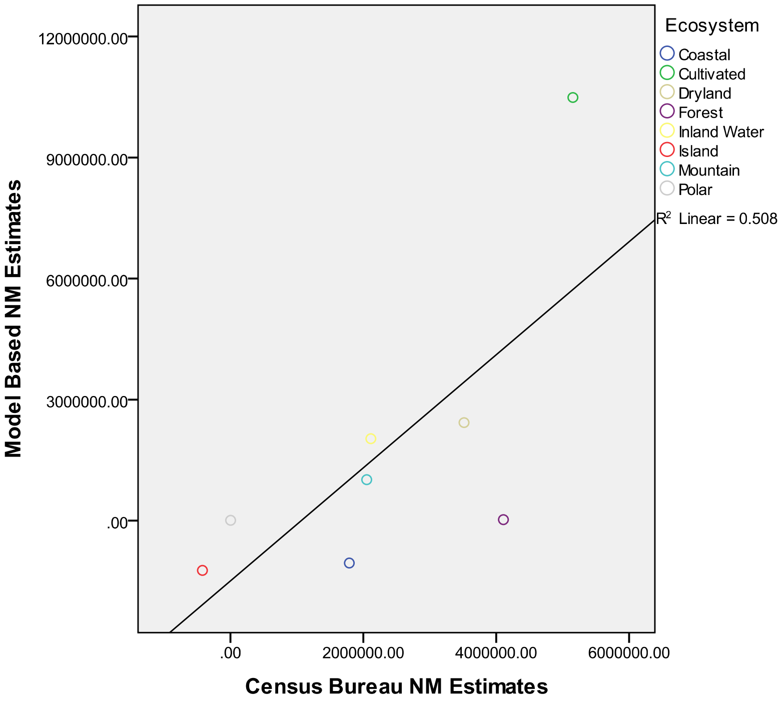


Table A16. US Net Migration by Ecosystem, US Census Compared Estimates to Modeled Estimates, decade of the 2000s

Ecosystem	US Census NM	Model NM (mean)
Coastal	1,789,568	-1,052,101
Mountain	2,051,641	1,014,054
Cultivated	5,153,437	10,488,022
Forest	4,107,755	24,887
Inland Water	2,113,206	2,030,171
Dryland	3,516,311	2,429,264
Polar	3,305	7,033
Island	-420,219	-1,234,387

Figure A18. Scatter Plot of NM Estimates from the US Census Bureau and Model Ouputs for Ecosystems, Decade of the 2000s



Appendix B. HYDE – History Database of the Global Environment

The History Database of the Global Environment (HYDE) was originally designed for testing and validation of the Integrated Model to Assess the Global Environment (the IMAGE Model , see Solomon, 1994 and Goldewijk & van Drecht, 2006). As one of the basic driving factors influencing environmental change, a global 100-year historical population map was one of the integral components of the HYDE database.

The National Center for Geographic Information and Analysis (NCGIA) 0.5 x 0.5 degree longitude/latitude population density database (Tobler et al. 1995) was used as the starting point for the first version of the HYDE global historical population map covering the period 1890-1990 (Goldewijk & Battjes 1997). IMAGE 2.1 country borders were overlaid on the NCGIA database. NCGIA database grid cells belonging to IMAGE 2.1 countries were aggregated to country totals. Country totals were adjusted to match country totals from the UN population database for the year 1994. Countries large enough to cover at least one grid cell were assigned to one of 13 IMAGE 2.1 geographic regions; the remaining countries were omitted. Historical country population data points from Mitchell (1975, 1982, 1983) were linked to the 1950 base year of the UN database using a country specific logistic curve determined through the earliest available data point and the 1950 UN data point. The growth rate, x, of the curve is calculated by:

$$x=(POP_{rec}/POP_{his})^{1/(Trec-This)} -1$$

where POP_{rec} and POP_{his} are the most recent and most historic population estimates, respectively, and T_{rec} and T_{his} are the years of the population estimates. The curve was used to calculate earlier historic values which were then checked against other available sources.

In cases where the T_{rec}-T_{his} was too small, thereby producing a skewed 1890 estimate, the regional growth rates of Durand (1977) were used instead.

Finally, adopting the assumption that high population density areas remain in the same place over time, the population distribution represented by the NCGIA database was applied to the HYDE country totals and population densities were scaled to a 0.5 deg x 0.5 deg lon/lat grid.

Goldewijk, & Battjes (2001) presented HYDE 2, updated and extended using the UN country population database for the period 1950-1995 (United Nations , 1997), and historical country population estimates (for the period 1750-1993) from Mitchell (1993;1998a,b) and (for the period 1820-1992) from Maddison (1994). These estimates were scaled to match UN country population data for the year 1950 in order to create a ~300-year (1700-1990) dataset consistent with the UN population database. Data gaps were filled in using a logistic curve (presumably as described in Goldewijk, C.G.M. and Battjes, J.J., 1997), and checked against other available estimates. The same methodology described above (Goldewijk, C.G.M. and Battjes, J.J., 1997) was used to generate updated gridded population density maps. To allow for at least some internal population changes in very large countries like the United State, Canada, India, China, etc…, sub-national data derived from Mitchell (1993; 1998a,b) was used.

HYDE version 3 (Goldewijk 2005), incorporated sub-national level historical population data acquired from Populstat database (Lahmeyer 2004) and the Gazetteer (2004) (crosschecked with data from Mitchell (1993, 1998a, b), Madison (1995) and many local country studies), and applied population growth rates from Grigg (1987) to all of the administrative units for which no data were available for historical time periods. The new database covered a 300-year period (from 1700-2000), and was built upon a new global sub-national level administrative boundary map developed by Goldewijk, de Man, Meijer, & Wonink (2004) of National Institute for Public Health and the Environment (RIVM). The new map consisted of 222 countries divided into 3441 administrative units, and provided the framework for data collection. ISO3166-2 level coding was used, if available, for all countries in the world (many of the historical population data sources used in this study are provided at this level). If data were not available at the ISO3166-2 level, they were converted to match that level.

HYDE 3.1, which was applied to this Foresight Project net migration modeling study, population totals for each country match the exact United Nations World Population Prospect (2008 revision) population numbers after 1950, except for Taiwan, in which the authors used data from Taiwan National Statistics instead (Godewijk *personal communication*).

Historical time series were constructed at the subnational level where the data were robust, and resulting country totals were checked against other sources where possible. Data gaps were filled through interpolation, and where no data were available, regional growth rates given by Grigg (1987) were used to hindcast to the base year 1700.

For cases in which historical data consisted of country totals while recent data consisted of sub-national totals, the ratio of all sub-national units for a given year compared to the country total for that year was applied to the historical country total to obtain historical sub-national population numbers. In other words, the spatial differentiation within a country was assumed to have remained constant over time.

Given that the purpose of the study is to present a broad demographic overview of the past 300 years, useful for climate change modelers, the author feels it is defensible to use growth rates as published in literature in combination with other available sources, and asserts growth rates from Grigg (1987) for 10 world regions are generally in agreement with rates found in other studies (UN, 2003; Durand, 1974). Grigg's regional growth rates were applied to all countries within a region, and, when no sub-national growth rates were available, to all administrative units in a particular country of that region.

To address the uncertainties inherent in applying a single regional growth rate to all country/province levels in that region, plus and minus 5, 10, and 20 percent uncertainty intervals were computed on growth rates, yielding a bandwidth in the total population numbers for each country, and accumulating into regional bandwidths and a global one (Goldewijk & van Drecht 2006: 100-101).

Historical population numbers were downscaled on a sub-national basis, using statistics and the literature, to the 5 minute Landscan population counts map (Landscan, 2006) to produce HYDE 3 population density maps on a 5 by 5 minute resolution for 10 year time steps for the period 1700-1970 (Goldewijk 2005).

Goldewijk et al.(2010) and Goldewijk et al. (2011) present, HYDE3.1, a revision and extension HYDE 3.0, including updated and internally consistent historical population estimates for the extended period of 10,000 BC to AF 2,000, i.e. the whole Holocene. National historical population estimates are based on historical population numbers of McEvedy and Jones (1978), Livi-Bacci (2007), Maddison (2001). Supplemented with subnational population numbers from Populstat (Lahmeyer 2004) and other sources, time series were constructed for each sub-national administrative unit of every country of the world. Current administrative units and their boundaries were kept constant over time, and historical sources were adjusted to match the current boundaries of HYDE 3.1 (i.e. "by taking fractions of former larger administrative units" (p 566)). Country and regional totals and pop density were estimated and the resulting pop growth rates in percent per year per time period were computed.

Spatial distribution for recent time periods was depicted by using weighing maps based on the 30" x 30" latitude/longitude Landscan (2006) population density map. Hindcasting required gradually replacing these with weighing maps based on proxies such as distance to water and soil suitability. (Goldewijk et al., 2011)

Uncertainties

The author's acknowledge that although the HYDE sources of historical population have been reviewed extensively, the further back one goes in time (pre-1950), the actual data on population distribution is sparse, and so the team developing HYDE had to rely heavilty on 'educated guesses'. Therefore, prior to the 1950s the numbers must be treated with care, and especially so for the pre-1700 period. Despite this caution, the authors believe the hindcast estimates fall well within the range of those found literature, and both the estimates and resulting growth rates seem a "reasonable reconstruction of historical population trends" (Goldewijk et al., 2010).

Authors attempted to quantify uncertainty in total pop estimates by introducing 'lower' and 'upper' range beside the HYDE 3.1 estimate, based on the high end of the literature estimates. These estimates yield an increasing uncertainty range going back in time, e.g. +-1% in AD 2000, …,+-25% in AD 1700, …, +-100% in 10,000 BC. Considering the min and max results as extremes (since the high end of the lit estimates were used) HYDE 3.1 can be considered a reasonable scenario for historical population growth (Goldewijk et al., 2010).

Table B1. Islands Missing from HYDE v.3.1

Island	population (year 2000)
Bermuda	62,960
Cook Islands	19,601
Federated State of Micronesia	122,692
Guam	155,080
Maldives	290,923
Marshall Islands	51,127
Northern Mariana Islands	72,736
Niue	--
Nauru	12,218
Pitcairn	--
Palau	19,175
French Polynesia	233,167
Saint Helena	77,187
Svalbard	--
Seychelles	77187
Tokelau	--
Tuvalu	10156
Wallis and Futuna	14454
Greenland	55974

Appendix C: Methods for Imputing Rates of Natural Increase

This Appendix includes sections that describe the imputation method for developing urban and rural rates of natural increase using two different statistical packages, and a third section that compares the two approaches. The purpose of the modeling was to impute missing values for urban and rural crude birth rates (CBRs) and crude death rates (CDRs) across all countries and for every year in the 41 year time span (1970-2010) in order to obtain urban and rural rates of natural increase. We had 5,016 observed urban/rural CBRs and CDRs across 231 countries and four decades: 766 for the 1970s, 1,198 for the 1980s, 1,458 for the 1990s, and 1,594 for the 2000s. Sources included the UN Demographic Yearbook (DY) for CBRs and CDRs and the Demographic and Health Surveys (DHS) for CBRs only. In all, a total of 32,868 data points needed to be imputed. We had an extensive set of mostly national-level time series ancillary data with which to carry out the imputations (Table C1). Note that we subsequently replaced the imputed data for the US with observed decadal rates of natural increase from the US Census Bureau, averaging 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 RNI data). Section C1 describes the imputation methods using the *mi* package for R, Section C2 describes the imputation methods using the Amelia package for R, and Section C3 compares the *mi* and Amelia packages.

Section C1. Description of Imputation Methods using Multiple Imputation

1. Working data set

The working data set included the following variables (see Table C1 for the codebook): isocode, year, countryname, uncode, totpop, urbpop, rurpop, gdppc, watsup, cntrycbr, cntrycdr, cbr_cbidb, cdr_cbidb, dyburbancbr, dyburbancdr, dybruralcbr, dybruralcdr, dhsurbancbr, dhsruralcbr, oecd, hiopec, rurpov, urbpov, pctubanjmp, pctruraljmp, urbwatsup, rurwatsup, un_region, un_majorarea, un_development_group, agri_kd, femurb15_49, urban_cwr, femrur15_49, rural_cwr, urban1q0, urban4q1, urban5q0, rural1q0, rural4q1, rural5q0, total1q0, total4q1, total5q0, literacyurban, literacyrural, literacytotal, tot_prop60, urb_prop60, rur_prop60, acsat, urbacsat, ruracsat, rural_birth_doc, urban_birth_doc, rural_measles, urban_measles, rural_mortality, urban_mortality, umdg7_water, rmdg7_water, totmdg7_water, umdg7_sant, rmdg7_sant, totmdg7_sant.

It also included the following 199 countries/entities:

Afghanistan, Albania, Algeria, Andorra, Angola, Antigua and Barbuda, Argentina, Armenia, Australia, Austria, Azerbaijan, Bahamas, Bahrain, Bangladesh, Barbados, Belarus, Belgium, Belize, Benin, Bhutan, Bolivia, Bosnia and Herzegovina, Botswana, Brazil, Brunei Darussalam, Bulgaria, Burkina Faso, Burundi, Cambodia, Cameroon, Canada, Cape Verde, Central African Republic, Chad, Chile, China, China, Hong Kong SAR, China, Macao SAR, Colombia, Comoros, Congo, Costa Rica, Croatia, Cote d'Ivoire, Cuba, Cyprus, Czech Republic, Democratic Republic of the Congo, Dem. People's Republic of Korea, Denmark, Djibouti, Dominica, Dominican Republic, Ecuador, Egypt, El Salvador, Equatorial Guinea, Eritrea, Estonia, Ethiopia, Fiji, Finland, France, French Guiana, Gabon, Gambia, Georgia, Germany, Ghana, Greece, Grenada, Guatemala, Guinea, Guinea-Bissau, Guyana, Haiti, Holy See, Honduras, Hungary, Iceland, India, Indonesia, Iran (Islamic Republic of), Iraq, Ireland, Israel, Italy, Jamaica, Japan, Jordan, Kazakhstan, Kenya, Kiribati, Kuwait, Kyrgyzstan, Lao People's Democratic Republic, Latvia, Lebanon, Lesotho, Liberia, Libyan Arab Jamahiriya, Liechtenstein, Lithuania, Luxembourg, Madagascar, Malawi, Malaysia, Maldives, Mali, Malta, Marshall Islands, Mauritania, Mauritius, Mexico, Micronesia (Fed. States of), Monaco, Mongolia, Montenegro, Morocco, Mozambique, Myanmar, Namibia, Nauru, Nepal, Netherlands, New Zealand, Nicaragua, Niger, Nigeria, Norway, Occupied Palestinian Territory, Oman, Pakistan, Palau, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Poland, Portugal, Puerto Rico, Qatar, Republic of Korea, Republic of Moldova, Romania, Russian Federation, Rwanda, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Samoa, San Marino, Sao Tome and Principe, Saudi Arabia, Senegal, Serbia, Seychelles, Sierra Leone, Singapore, Slovakia, Slovenia, Solomon Islands, Somalia, South Africa, Spain, Sri Lanka, Sudan, Suriname, Swaziland, Sweden, Switzerland, Syrian Arab Republic, Tajikistan, TFYR Macedonia, Thailand, Timor-Leste, Togo, Tonga, Trinidad and Tobago, Tunisia, Turkey, Turkmenistan, Tuvalu, Uganda, Ukraine, United Arab Emirates, United Kingdom, United Republic of Tanzania, United States of America, Uruguay, Uzbekistan, Vanuatu, Venezuela (Bolivarian Republic of), Viet Nam, Western Sahara, Yemen, Zambia, Zimbabwe

2. Specification of the imputation model

We used a customized version of the mi package for R to perform multiple imputation. In the mi package, each variable with missingness is modeled as a function of all other variables in an iterative process. The main differences between the mi runs and runs produced using the Amelia model (Section C1) are as follows:

- Amelia used a gap filled annual time series of crude birth rates and death rates (allcbrcountry and allcdrcountry) based on interpolations between the five year rates available from the United Nations *World Population Prospects 2008*.

- In Amelia, four separate imputation models were specified — one each for urban CBR, urban CDR, rural CBR, and rural CDR — using for each run a different 8 to 10 variable subset of the original 65 variables. This is much faster but utilizes less multivariate information when imputing the missing values, and in particular does not utilize the information that each of these four variables contains about the other three.

Each variable with missingness is modeled using a "random intercept, random coefficient" linear model. In other words, the intercepts over all country-years form a normal distribution while the coefficients for gross domestic product per capita (gdppc) and agriculture value added per worker (agri_kd) over all countries form normal distributions (unless gdppc or agri_kd is the dependent variable being modeled at that stage of the loop over all variables). The coefficients for the other variables are considered "fixed" and estimated. This specification allows a considerable amount of heterogeneity across countries in how CBRs and CDRs (and other variables) relate to development.

I ran 20 iterations for each of eight independent chains. One "iteration" is essentially one complete loop so that each variable with missingness is modeled and imputed once and then the process repeats on the next iteration. The completed dataset reflected the state of the process at the end of the 20th iteration, for each of the eight chains. On iteration zero, no variable is modeled. Instead, the observations that are missing are filled in with random draws from a uniform distribution whose minimum and maximum are the observed minimum and maximum for that variable and country. When a country had no observed data on a variable, the minimum and maximum were taken from the observed minimum and maximum for that variable across all countries.

Thus, on subsequent iterations, the dataset was provisionally filled in and each variable can be modeled. The following variables always entered the models — regardless of which side of the equation they happen to be on at a given time — in (natural) logarithm form: total population (totpop), urban population (urbpop), rural population (rurpop), GDP per capita (gdppc), and agriculture value added per worker (agri_kd). For six "city-states" (e.g. Hong Kong, Singapore), there is no rural population in some years, in which case I artificially changed the observed values to 1 (thousand) in order to take the logarithm. When they are on the left-hand side (to impute their missing values) I used the following steps which approximate the idea of "posterior predictive distribution" imputation:

1. Estimate the model (using the blmer function in the forthcoming blmer R package by Vincent Dorie at Columbia University);
2. Given the estimated parameters, draw new parameters from the multivariate distribution implied by the estimates;
3. Using the new parameters, construct a linear predictor for each observation that is missing in the original dataset; and
4. For each observation that is missing in the original dataset, draw an imputed value from a normal distribution with expectation equal to its linear predictor and variance equal to the previously drawn error variance. This overwrites whatever was the previous imputation for that observation.

When the imputation is finished, these variables were transformed back into their original scale using the exp() function. Other than that, there was no "post-processing".

The other variables, which are typically ratios of some sort, were not transformed but cannot be negative. The following steps were taken:

5. Estimate the model (again using blmer);
6. Given the estimated parameters, draw new parameters from the multivariate distribution implied by the estimates;
7. Using the new parameters, construct a linear predictor for each observation that is missing in the original dataset; and

8. For each observation that is missing in the original dataset, draw an imputed value from a normal distribution **truncated at zero** with expectation equal to its linear predictor and variance equal to the previously drawn error variance. This overwrites whatever was the previous imputation for that observation.

Thus, these variables needed no additional transformation or post-processing at the end.

3. Convergence

In principle, one should verify that the process converges for all unknowns. However, there are almost 400,000 missing values in the original dataset (including the auxiliary variables used to estimate urban and rural CBRs and CDRs) plus thousands of parameters to estimate. Thus, it was infeasible to do a rigorous convergence analysis in the time available (which also would have necessitated many more iterations). Instead, we tend to judge the convergence in mi by whether the variance in variable means (across the eight chains) is small relative to the mean (across the eight chains) of the variable variances. In other words, is the additional variance induced by the fact that the eight chains are not exactly in the same spot small relative to the total variance in a variable.

In this case, that seemed to be true despite only running 20 iterations. However, the total variance in each variable tends to be fairly large due to vast cross-country differences between the developed and developing worlds. In principle, it would probably have been better to judge the convergence on a country-by-country basis rather than for the world, although some countries have very little variance over time in some variables.

4. Analysis

The main limitation of this way of modeling the variables and then imputing the values is that it does not rigorously take into account the time-series nature of the data. Thus, there is almost assuredly autocorrelation and heteroskedasticity, which induce no bias but do compromise the estimates of uncertainty. Typically, the estimated uncertainty is biased downward. The main advantage is that all the available information is used so that the joint distribution of the completed variables is approximately correct under the assumption that the models are approximately correct. In the case of the Amelia runs, where the four important variables are imputed separately, each with a different subset of included variables, nothing ensures that the completed *joint* distribution of urban CBRs, urban CDRs, rural CBRs, and rural CDRs is internally coherent or coherent with national CBRs and CDRs. Amelia is only ensuring that each of these four variables is coherent with the subset of variables that were imputed along with it.

The main tool to judge the quality of the imputations was a set of four plots by country for the years between 1970 and 2010, averaging over the eight completed datasets. The four plots were

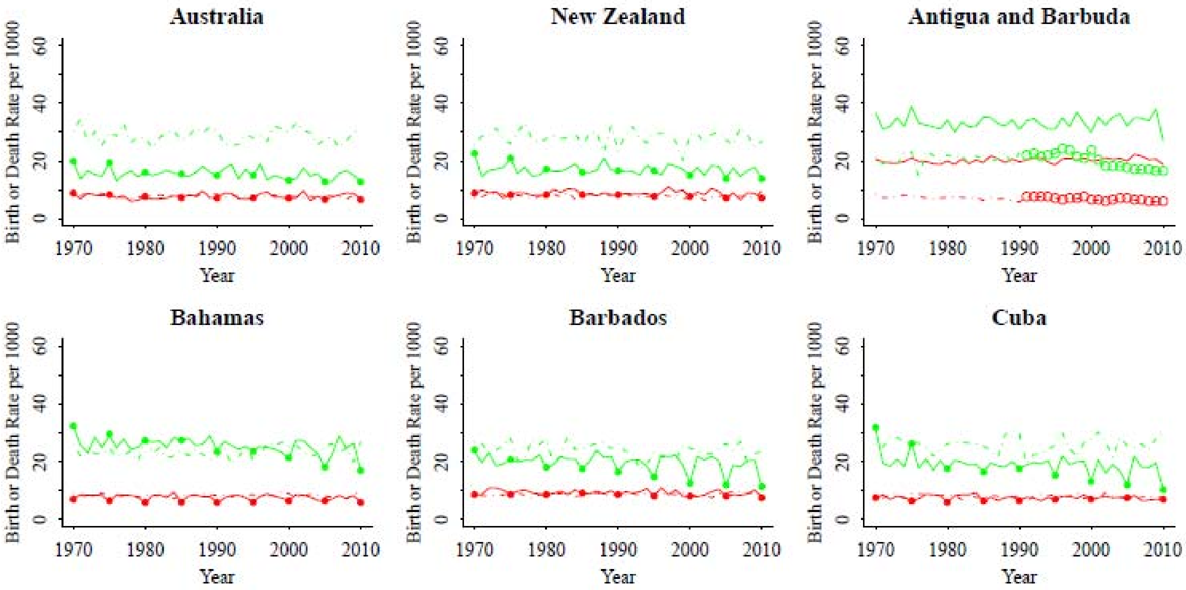
- Country-level CBRs and CDRs from the UN *World Population Prospects* (cntrycbr and cntrycdr) with estimates for every five years compared to Country level CBRs and CDRs from the US Census Bureau's International Database (cbr_cbidb and cdr_cbidb) (Figure C1).

- Urban CBRs and CDRs from the UN *Demographic Yearbook* (dyburbancbr and dyburbancdr) compared to rural CBRs and CDRs from the same source (dybruralcbr and dybruralcdr) (Figure C2)

- Urban CBRs and CDRs from the UN *Demographic Yearbook* (dyburbancbr and dyburbancdr) compared to imputations for Urban CBRs and CDRs for the DHS time series, which only had observations for urban and rural CBRs (dhsurbancbr and dhsurbancdr).

- Country-level CBRs from the UN *World Population Prospects* (cntrycbr) compared to weighted averages of the imputed values for the Demographic Yearbook series (dyburbancbr and dybruralcbr) based urban and rural populations in the country, and analagously for death rates (Figure C3).

Figure C1. CBRs (green) and CDRs (red) imputed for the United Nations World Population Prospects series (solid line) and Census Bureau (dashed line). Filled circles represent the UN published estimates, and open circles represent the Census Bureau's observed data.



I was primarily looking for rough coherence across different *sources* of data on the same conceptual variables (i.e., comparing imputed CBR series to one another). In the second case, there is no reason to think that urban vital statistics would be the same as rural vital statistics, but we would expect that urban birth and death rates to be consistently lower than the rural counterparts within a country. The plots for some countries were more coherent than for other countries, but overall I felt that coherence was lacking.

Although the plots indicated fully observed data points by superimposing circles on the lines, I was less concerned with whether the imputed values were coherent with the observed values for a couple of reasons. First, in many cases, there were no observed data on vital statistics at the subnational level. Second, in many cases, the observed data are presumably measured with considerable error, so it could be the case that the model is implying imputations that are consistent with the true concept, but the observed measurement of that concept is off. That said, in many countries, the imputed values were consistently higher or lower than the observed values.

Figure C2. CBRs (green) and CDRs (red) imputed for the urban areas (solid line) and rural areas (dashed line). Observed data from the UN Demographic Yearbook represented by circles.

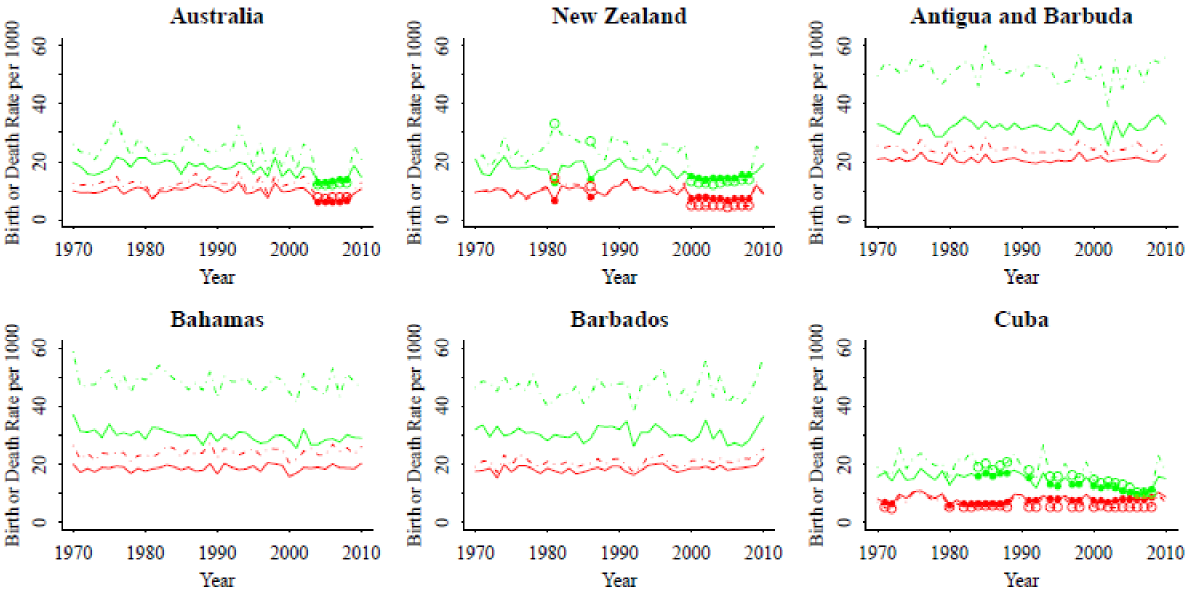
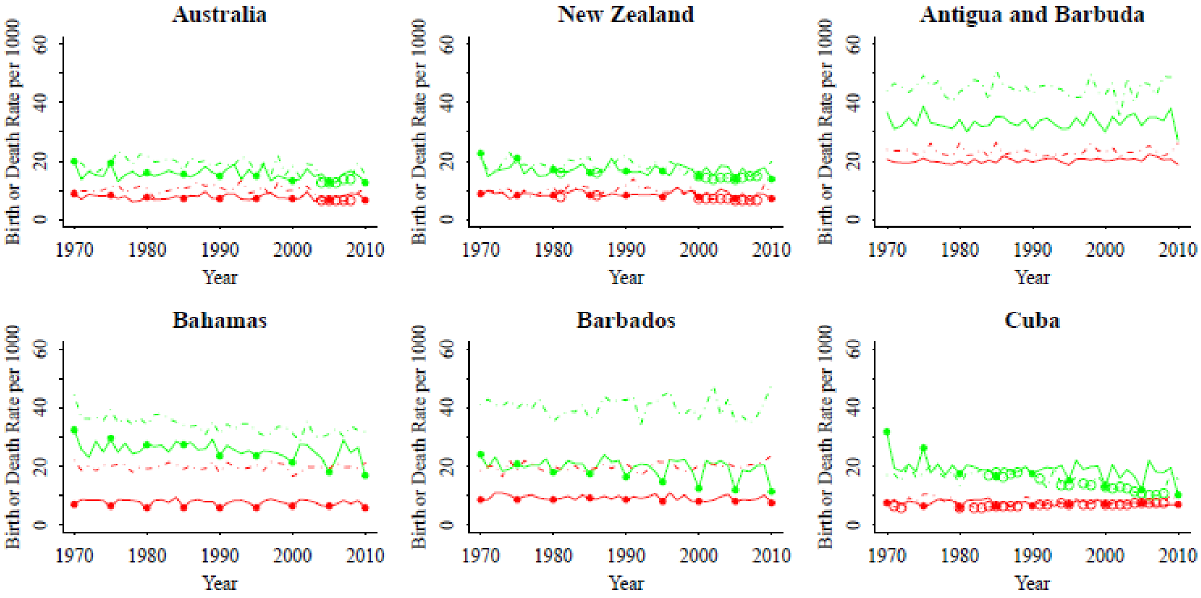


Figure C3. Country level CBRs (green) and CDRs (red) imputed for the United Nations World Population Prospects series (solid line) compared to a Weighted Average of the UN Demographic Yearbook urban/rural CBRs and CDRs (dashed line). Filled circles represent the UN published estimates, and open circles represent the Demographic Yearbook's observed data.



For each country annual rates of natural increase (RNIs) were calculated by subtracting the CDR from the CBR. The decadal urban rates of natural increase (percent change due to natural increase over the decade) were calculated by multiplying the annual rates of natural increase for urban areas times the urban population and summing the population growth due to natural increase and dividing by the urban population at the start of the period. The same steps were used to calculate the rural rates.

Note that it is possible that the imputations are largely correct on average within a decade but poor in any particular year, in which case the decadal statistics are presumably viable for their intended purpose. However, it is also possible that the imputations are, for example, systematically too low in the 1970s and systematically too high in the 1980s, in which case the migration estimates for that country between 1970 and 1990 would be biased.

Section C2. Imputation Using Amelia and Post-Imputation Processing Steps

1. Working data set

This imputation was carried out using the same data set as described in Section C1. I re-created the unique identifier (uniID) by concatenating ISO3 code and year.

The urban population variable (urbpop) was calculated as the difference between total population (totpop) and rural population (rurpop) because the data set did not have decimal values for urban population anymore. This step was not used in the imputations itself but was used in the post-processing of the imputed values, e.g., the adjustment to total country CBR and CDR and calculation of NRIs.

2. Identifying the variables for inclusion in the imputation model

I calculated the bivariate correlation matrix of all continuous and ordinal variables in the working data set using Kendall's correlation coefficient on all pairwise complete data points for each variable pair. From this matrix, I identified the variables with (a) the highest correlation with the outcome or target variables (i.e., with urban and rural CBR and CDR from the Demographic Yearbook) and (b) reasonably high tempo-spatial coverage to contribute meaningfully to informing the imputation model and produce sensible imputations.

3. Running the imputations in Amelia

I ran the following imputation models:

Urban CBR:

To impute the urban CBRs, I used five auxiliary variables: GDP per capita (gdppc), the interpolated UN World Population Prospects CBRs and RNIs (allcbrcountry and allnircountry), and dummy variables for more developed and less developed countries (mdr and ldr). The complete inputs were: "ISOcode", "year", "uniID", "gdppc", "allcbrcountry", "allnircountry", "dyburbancbr", "dybruralcbr", "mdr", and "ldr".

The imputation model was specified as follows:

```
a.out <- amelia(dats[,vars.imp], m = 5, p2s = 1,frontend = FALSE, idvars = "uniID", ts = "year", cs =
"ISOcode", polytime = 1, splinetime = 3, intercs = TRUE, lags = NULL, leads = NULL, startvals = 1,
tolerance = 0.0001, logs = "gdppc", sqrts = NULL, lgstc = NULL, noms = c("mdr","ldr"), ords = NULL,
incheck = TRUE, collect = FALSE, arglist = NULL,  empri = 500, priors = NULL, autopri = 0.5, emburn =
c(0,0), bounds = NULL, max.resample = 100)
```

Urban CDR:

To impute the urban CDRs, I used four auxiliary variables: The interpolated UN World Population Prospects CDRs (allcdrcountry), the country literacy rate (iteracytotal), and dummy variables for more developed and less developed countries (mdr and ldr). The complete inputs included: ”ISOcode", "year", "uniID", "allcdrcountry", "dyburbancdr", "mdr", "ldr", and "literacytotal".

The imputation model was specified to produce m=5 separate, completed data sets with a linear effect of time (year) that can vary across countries and which is smoothed by third degree polynomials. Per capita GDP is log transformed, while degree of development expressed in classifications such as least developed were included as nominal variables. The model also uses ridge priors, which shrinks the covariances toward zero while retaining the empirical variances and mean structure. Ridge priors are useful if there is multicollinearity in the data. Starting point for each imputation chain is an identity matrix:

Urban CDR:

To impute the urban CDRs, I used four auxiliary variables: The interpolated UN World Population Prospects CDRs (allcdrcountry), the country literacy rate (iteracytotal), and dummy variables for more developed and less developed countries (mdr and ldr). The complete inputs included: ”ISOcode", "year", "uniID", "allcdrcountry", "dyburbancdr", "mdr", "ldr", and "literacytotal" and model specifications were otherwise the same as for urban CBR.

Rural CBR:

To impute the rural CBRs, I used three auxiliary variables: GDP per capita (gdppc), the interpolated UN World Population Prospects CBRs (allcbrcountry), and the rural literacy rate (literacyrural). The complete inputs included: "ISOcode", "year", "uniID", "gdppc", "allcbrcountry", "dybruralcbr", and "literacyrural". Model specification followed the principles of urban CBR.

Rural CDR:

To impute the rural CDRs, I used four auxiliary variables: The interpolated UN World Population Prospects CDRs (allcdrcountry) and dummy variables for more developed, less developed, and less developed without least developed countries (mdr, ldr, ldr_noleast). The complete inputs included: "ISOcode", "year", "uniID", "allcdrcountry", "dybruralcdr", "mdr", "ldr", and "ldr_noleast" and model specification followed the same principles as for urban CBR.

4. Postprocessing

Following the generation of 5 sets of Amelia imputations for urban and rural CBR and CDR, respectively (for 20 sets of imputed values in total), I ran a LOESS local regression smoother over the imputed data set (over imputed and observed data together). The objective for doing this was to smooth out the otherwise measurably volatile imputations (a result of the weak imputation models).

Following the LOESS smoothing, I replaced the smoothed observed values again with their raw observed values because we wanted to keep the observed values and did not want to replace them with smoothed values.

5. Calculation of NIs and decadal RNIs

To calculate the natural (crude) increase, the five model runs for urban CBR and urban CDR were combined to create five runs for urban RNIs, and the same was done for rural CBRs and CDRs. In the six entities where the rural population was 0, this was changed to 1 (thousand) in order to be able to calculate rural decadal RNIs.

The decadal urban rates of natural increase (percent change due to natural increase over the decade) were calculated by multiplying the annual rates of natural increase for urban areas times the urban population and summing the population growth due to natural increase and dividing by the urban population at the start of the period. The same steps were used to calculate the rural rates.

Table C1. Variable Code Book

variable name	type	format	label	Source
countryname	str52	%52s	Country Name	
uncode	int	%8.0g	United Nations country code	
ISOcode	str4	%9s	ISO 3-character country code	
uniID	str7	%9s	Unique ID - ISOcode + year	
year	int	%9.0g	Calendar year	
totpop	float	%9.0g	Total Population (thousands) (UN)	WUP 2009
urbpop	str7	%9s	urban population string (UN)	WUP 2009
_urbpop	long	%10.0g	Urban Population (thousands) (un)	WUP 2009
rurpop	float	%9.0g	rural population (thousands) (UN)	WUP 2009
pcturb	str6	%9s	percent urban string (UN)	WUP 2009
_pcturb	double	%10.0g	percent urban numeric (UN)	WUP 2009
cntrycbr	float	%9.0g	original country CBR (per thousand) (UN)	WPP 2008
allcbrcountry	double	%10.0g	interpolated country CBR (per thousand)	Calculated
cntrycdr	float	%9.0g	original country CDR (per thousand) (UN)	WPP 2008
allcdrcountry	double	%10.0g	interpolated country CDR (per thousand)	Calculated using STATA linear interpolation
cntrynir	float	%9.0g	original country NIR (per thousand) (UN)	WPP2008
allnircountry	double	%10.0g	interpolated country NIR (per thousand)	calculated
cbr_cbidb	float	%9.0g	CBIDB Births per 1,000 population	US Census Bureau International Database
cdr_cbidb	float	%9.0g	CBIDB Deaths per 1,000 population	US Census Bureau International Database
nir_cbidb	float	%9.0g	CBIDB Natural increase rate (per thousand)	US Census Bureau International Database
gdppc	float	%9.0g	GDP per capita 2000 constant US dollars	Several
gdp_source	str8	%9s	Source of GDP data	See Categories
dyburbancbr	float	%9.0g	DYB Urban Crude Birth Rate	UN SD Demographic Year Book
dyburbancdr	float	%9.0g	DYB Urban Crude Death Rate	UN SD Demographic Year Book
dybruralcbr	float	%9.0g	DYB Rural Crude Birth Rate	UN SD Demographic Year Book
dybruralcdr	float	%9.0g	DYB Rural Crude Death Rate	UN SD Demographic Year Book
dhsurbancbr	float	%9.0g	DHS Urban Crude Birth Rate	DHS Surveys through Statcompiler

variable name	type	format	label	Source
dhsruralcbr	float	%9.0g	DHS Rural Crude Birth Rate	DHS Surveys through Statcompiler
dhs_region	str29	%29s	DHS Regional Grouping	DHS Surveys through Statcompiler
country_name	str50	%50s	Country Name (from merges)	
mdr	byte	%8.0g	More Developed Region	UN POP DIV regional grouping WPP2008
ldr	byte	%8.0g	Less Developed Region	UN PD regional grouping WPP2008
ltdr	byte	%8.0g	Least Developed Region	UN PD regional grouping WPP2008
ldr_noleast	byte	%8.0g	Less Developed Region without Least Developed	UN PD regional grouping WPP2008
oecd	byte	%8.0g	OECD countries	
ssa	byte	%8.0g	Sub-Saharan Africa	
hiopec	byte	%8.0g	High Income OPEC	
asia	byte	%8.0g	Asia countries	
fsu	byte	%8.0g	Former Soviet Union	
mena	byte	%8.0g	Middle East and North Africa	
whsrg	byte	%8.0g	Western Hemisferio South of Rio Grande	
rurpov	float	%8.0g	Poverty headcount ratio at rural poverty line (% of rural population) (WDI)	WDI
urbpov	float	%8.0g	Poverty headcount ratio at urban poverty line (% of urban population) (WDI)	WDI
_mergeA	byte	%8.0g		System variable
_mergeB	byte	%8.0g		System variable
pctubanjmp	byte	%8.0g	% pop urban (JMP)	
pctruraljmp	byte	%8.0g	% pop rural (JMP)	
urbwatsup	int	%8.0g	% urban pop with water supply	
rurwatsup	int	%8.0g	% rural pop with water supply	
watsup	int	%8.0g	% total pop with water supply	
urbacsat	byte	%8.0g	% urban pop with access to improved sanitation	
ruracsat	byte	%8.0g	% rural pop with access to improved sanitation	
acsat	int	%8.0g	% total pop with access to improved sanitation	
_merge	byte	%8.0g		System variable

variable name	type	format	label	Source
dhsregion	str29	%29s	DHS region	Demographic and health survey
urban1q0	str5	%9s	Urban1q0 = infant mortality (less than 1 year 0ld) urban areas	Demographic and health survey
urban4q1	str5	%9s	Urban4q1 = child mortality (between ages 1 and 4)	Demographic and health survey
urban5q0	str5	%9s	Urban5q0 = under-5 mortality	Demographic and health survey
rural1q0	str5	%9s	Rural1q0	Demographic and health survey
rural4q1	str5	%9s	Rural4q1	Demographic and health survey
rural5q0	str5	%9s	Rural5q0	Demographic and health survey
total1q0	str5	%9s	Total1q0	Demographic and health survey
total4q1	str5	%9s	Total4q1	Demographic and health survey
total5q0	str5	%9s	Total5q0	Demographic and health survey
literacyurban	str5	%9s	LiteracyUrban : Percent distribution of women by level of schooling attended and by level of literacy, and percent literate, according to background characteristics	Demographic and health survey
literacyrural	str5	%9s	LiteracyRural	Demographic and health survey
literacytotal	str5	%9s	LiteracyTotal	Demographic and health survey
development_reg	str11	%11s	UN Pop Division Development regions	
agri_kd	float	%8.0g	Agriculture value added per worker (constant 2000 US$)	WDI
UMDG7_water%			MDG 7 Population using improved drinking-water sources (%) urban	WHO ????
RMDG7_water%			MDG 7 Population using improved drinking-water sources (%) rural	WHO ????
TotMDG7_water%			MDG 7 Population using improved drinking-water sources (%) total	WHO ????
UMDG7_sant%			MDG 7 Population using improved sanitation (%) urban	WHO ????
RMDG7_sant%			MDG 7 Population using improved sanitation (%) rural	WHO ????
TotMDG7_sant%			MDG 7 Population using improved sanitation (%) total	WHO ????
whourbancbr			WHO urban CBR	World Health Organization (WHO)
whourbancdr			WHO urban CDR	WHO
whoruralcbr			WHO rural CBR	WHO
whoruralcdr			WHO rural CDR	WHO

variable name	type	format	label	Source
V1			Rural births attended by skilled health personnel %, WHS2010-RurBirth-SHP	WHS2010-Health-Inequalities.csv
V2			Urban births by skilled health personnel %, WHS2010-UrbBirth-SHP	WHS2010-Health-Inequalities.csv
V5			Rural measles immunization coverage among 1 year olds WHS2010-RurMeasImm-1yo	WHS2010-Health-Inequalities.csv
V6			Urban measles immunization coverage among 1 year olds, WHS2010-UrbMeasImm-1yo	WHS2010-Health-Inequalities.csv
V9			Rural under five mortality rate (probability of dying by age 5 per 1000 live births), WHS2010-RurUnder5Mort	WHS2010-Health-Inequalities.csv
V10			Urban under five mortality rate (probability of dying by age 5 per 1000 live births), WHS2010-UrbUnder5Mort	WHS2010-Health-Inequalities.csv
total_pop	float	%9.0g	Total Population (thousands)	United Nations - Population division - Urban and Rural Population by Age and Sex, 1950-2005
tot60plus	float	%9.0g	Total population age 60 and older (thousand)	United Nations - Population division - Urban and Rural Population by Age and Sex, 1950-2005
tot_prop60	float	%9.0g	Proportion of total population age 60 and older	United Nations - Population division - Urban and Rural Population by Age and Sex, 1950-2005
urbpop	float	%9.0g	Total urban population (thousand)	United Nations - Population division - Urban and Rural Population by Age and Sex, 1950-2005
urb60plus	float	%9.0g	urban population age 60 and older (thousand)	United Nations - Population division - Urban and Rural Population by Age and Sex, 1950-2005
urb_prop60	float	%9.0g	Proportion of urban population age 60 and older	United Nations - Population division - Urban and Rural Population by Age and Sex, 1950-2005
rurpop	float	%9.0g	Total rural population (thousand)	United Nations - Population division - Urban and Rural Population by Age and Sex, 1950-2005
rur60plus	float	%9.0g	Rural population age 60 and older (thousand)	United Nations - Population division - Urban and Rural Population by Age and Sex, 1950-2005
rur_prop60	double	%10.0g	Proportion of rural population age 60 and older	United Nations - Population division - Urban and Rural Population by Age and Sex, 1950-2005
child0_4urb	float	%9.0g	Urban Children ages 0-4	United Nations - Population division - Urban and Rural Population by Age and Sex, 1950-2005
femurb15_49	float	%9.0g	urban women ages 15-49	United Nations - Population division - Urban and Rural Population by Age and Sex, 1950-2005

variable name	type	format	label	Source
urban_cwr	float	%9.0g	urban child-woman ratio	United Nations - Population division - Urban and Rural Population by Age and Sex, 1950-2005
child0_4rur	float	%9.0g	Rural Children ages 0-4	United Nations - Population division - Urban and Rural Population by Age and Sex, 1950-2005
femrur15_49	float	%9.0g	Rural women ages 15-49	United Nations - Population division - Urban and Rural Population by Age and Sex, 1950-2005
rural_cwr	str11	%11s	rural_CWRrural child woman ratio	United Nations - Population division - Urban and Rural Population by Age and Sex, 1950-2005
_rural_cwr	double	%10.0g	numerical rural child woman ratio	United Nations - Population division - Urban and Rural Population by Age and Sex, 1950-2005
_merge	byte	%8.0g		system variable

Section C3. Comparing the mi and Amelia Packages for R

1. The R package mi for multiple imputations using chained equations

The descriptions given in the following paragraphs are based on the article “Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box” by Yu-Sung Su, Andrew Gelman, Jennifer Hill and Masanao Yajima in the Journal of Statistical Software[10], the most recent version available for download (version 0.09-14 from 2011-4-25) and from practical examples run in R v. 2.13.1.

The mi package is based on chained equation regression imputation, which requires the specification of conditional models for each imputation variable conditional on predictor variables. The imputation algorithm sequentially iterates through the variables to impute the missing values using the specified model until the convergence criterion is satisfied.

The analyst works through the following steps prior to running the mi() function and subsequent imputation analytics. They are aimed at giving more control over the imputation process and hence taking the “black box” image out of multiple imputation.

1. Setup
   a. Display of missing data patterns.
   b. Identifying structural problems in the data and preprocessing.
   c. Specifying the conditional models.
2. Imputation
   a. Iterative imputation based on the conditional model.
   b. Checking the fit of conditional models.
   c. Checking the convergence of the procedure.
   d. Checking to see if the imputed values are reasonable.
3. Analysis
   a. Obtaining completed data.
   b. Pooling the complete case analysis on multiply imputed datasets.
4. Validation
   a. Sensitivity analysis.
   b. Cross validation.

Important input information includes (a) the order in which the incomplete variables are to be imputed. This is not a trivial question since the different orderings will yield different results. One question that arises in this context if variables with small amounts of missingness should be imputed prior to variables with higher fractions of missing values. Variable types also need to be specified correctly in order to

[10] http://www.stat.ucla.edu/~yajima/Publication/mipaper.rev04.pdf

ensure that the imputed values are sensible, e.g., fall within the permissible range of values for a given variable. The mi package offers preprocessing capabilities to identify up to 11 variable types and transform them appropriately (e.g., for positive continuous data or fractions). As is the case with all existing imputation programs, variables with 100% missingness are not imputed. The same holds for a variable that is completely collinear to another variable in the data set (the mi package checks for that and imputes the excluded variable using the linear relationship between the two collinear variables). Additive constraints, on the other hand are harder to detect and is dealt with by adding a user-controlled level of noise produced from an artificial set of prior distributions and adding the noise to the observed data (and hence preserving many of the variable's distributional characteristics). A very useful capability of the mi package is the use of Bayesian model fitting algorithms. Currently implemented are bayesglm() with Gaussian functions, binomial family with logit link function and quasi-poisson families for overdispersion models as well as bayespolr() for ordered logistic or probit modeling with independent normal, t, or Cauchy prior distribution for the coefficients.

Convergence of imputation chains is a notoriously tricky issue to ascertain and the mi package offers several parametric, statistical and graphical options for assessing it. To begin, mi() monitors the mixing of each variable by the variance of its mean and standard deviation within and between different chains of the imputation. Convergence is assumed if the R.hat statistic, i.e., the difference of the within and between variance is trivial, is smaller than 1:1 (Gelman et al. 2004). Additionally, by specifying mi(data, check.coef.convergence = TRUE, ...), users can check the convergence of the parameters of the conditional models.

Imputation of incomplete data sets is only a means towards a greater end, i.e., the actual analysis of the data. This means that the imputation model needs to be chosen wisely because it can be assumed that it is often not the model used to analyze the final data set(s). Therefore, model assessment should be an integral part of multiple imputations and the mi package contains several features for this purpose (the following excerpt is from the paper in Journal of Statistical Software):

Our mi addresses this problem with three strategies.

1. Imputations are typically generated using models, such as regressions or multivariate distributions, which are fit to observed data. Thus the fit of these models can be checked (Gelman et al. 2005).

2. Imputations can be checked using a standard of reasonability: the differences between observed and missing values, and the distribution of the completed data as a whole, can be checked to see whether they make sense in the context of the problem being studied (Abayomi et al. 2008).

3. We can use cross-validation to perform sensitivity analysis to violations of our assumptions. For instance, if we want to test the assumption of missing at random, after obtaining the completed dataset (original data plus imputed data) using mi, we can randomly create missing values on these imputed datasets and re-impute the missing data (Gelman et al. 1998). By comparing the

imputed dataset before and after this test, we can assess the validity of the missing at random assumption.

1.1 Advantages and their costs in the mi package

The mi package is a new and powerful addition to the toolbox of statisticians, researchers and analysts who must deal with incomplete or missing data. It has a breadth of features that expand its applicability and give the users a higher level of control than is the case with many other software tools, although SAS MI and the R package Amelia are also heavily user-driven. However, additional features and control parameters in multiple imputation usually come at a cost. The mi user has to make a rather large number of decisions on what variables to impute, what model to use for it, and how to check if the results are meaningful. This control is useful for experienced users but may be a challenging task for the occasional practitioner. Understanding basic statistical theory and regression modeling concepts is therefore recommended. In addition, the use of a chained equation regression approach (in a Bayesian framework) puts considerable demands on the computational capabilities of the user. While small data sets generally impute in a few minutes, larger and large data sets may take hours and days to complete. The mi() function is powerful in that it allows the individual specification of the imputation model for each variable. It may, however, also exponentially grow the risk of model misspecification, especially if the underlying data generating process is assumed to be more complicated than it is in reality. Parsimony may be sacrificed for a misguided need for complexity. On the other hand, including rather more than fewer predictors in the regression model can assist with the Missing At Random assumption.[11] And while the mi package can already handle a fairly wide range of data types, time series data (and panel data) with their inherent autocorrelation (spatial-temporal correlation) structure are not yet included.

Convergence diagnostics remain an active field of statistical research and while methods and visual displays have become more sophisticated, there is still debate on what statistics to use. In the mi package the risk is to abort the iteration process too quickly (the default is 30 iterations), especially if iterations take a long time. Experiences with other glm and glmm modeling indicate that convergence can sometimes take several hundred iterations.

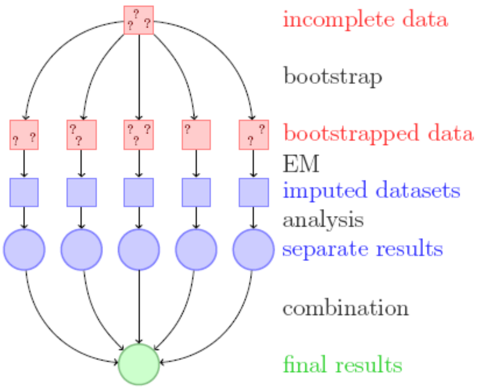
[11] Missing At Random (MAR) means that the probability of a single value being missing depends only on other observed values but not the missing value itself. It is the most widely used assumption in imputation software. Missing Completely At Random (MCAR) is the strictest assumption and means that the probability of a value being missing is independent of both the missing value and other observed values. Not Missing At Random (NCAR) means that the probability of a value being missing may also depend on the missing value itself. This form of missing data generating mechanism is known to exist in certain survey and data collection areas, e.g., when asking people about the income: high and low income earners are less likely to state their true income.

2. The R package Amelia for multiple imputations of cross-sectional time series data

The following information is sourced from the Amelia documentation[12], the current version of Amelia (1.5-0 of 23 Nov 2010) and practical examples included in the documentation. Amelia is a program for multiple imputation of cross-sectional time series data based on the MAR assumption. It uses the EMB algorithm, which combines the classical Expectation-Maximization algorithm of Dempster, Laird and Rubin (1977) with a bootstrapping component. The EMB algorithm in particular, saves time by first generating *m* bootstrapped sets of the data and then imputing each of these to generate *m* completed sets using the posterior distribution of the complete-data parameter distribution. The general approach is schematized in Figure 1.

The imputation model is likely to differ from the analysis model, however, when considering using Amelia to multiply impute missing data, the first step should be to identify the variables to include in the imputation model. Any variable that will or is likely to be in the analysis model should if meaningful and feasible also be in the imputation model, including any transformations or interactions of variables that will appear in the analysis model. Data permitting, inclusion of more information can be beneficial: since imputation is predictive, any variables that would increase predictive power should be included in the model, even if including them in the analysis model would produce bias in estimating a causal effect (such as for post-treatment variables) or collinearity would preclude determining which variable had a relationship with the dependent variable (such as including multiple alternate measures of GDP).

Figure 1: Overview of the steps taken in Amelia to produce *m* completed data sets (Source: Amelia documentation).



The basic imputation model of Amelia is a multivariate normal due to its useful properties for fitting and sampling from it. Transformations of the variables should be considered if they would more closely fit the multivariate normal assumption: correct but omitted transformations will shorten the number of

[12] http://r.iq.harvard.edu/docs/amelia/amelia.pdf

steps and improve the fit of the imputations. Amelia allows the specification of certain transformations (log, logistic, square root) as well of some variable types such as ordered categorical and nominal. The output is already back-transformed.

Amelia is specifically designed to deal with cross-sectional time series data, that is it requires identifiers for the time and the cross-section variables and has functionality build in to deal with autocorrelation and cross-sectional effects. The documentation states: “Many variables that are recorded over time within a cross-sectional unit are observed to vary smoothly over time. In such cases, knowing the observed values of observations close in time to any missing value may enormously aid the imputation of that value. However, the exact pattern may vary over time within any cross-section. There may be periods of growth, stability, or decline; in each of which the observed values would be used in a different fashion to impute missing values. Also, these patterns may vary enormously across different cross-sections, or may exist in some and not others. Amelia can build a general model of patterns within variables across time by creating a sequence of polynomials of the time index.” Polynomials of time of up to third degree are possible and they can be interacted with cross-sectional units to allow the patterns over time to vary between cross-sectional units. Alternatively, the temporal correlation aspect can be added to the model by including “lags” and “leads”. The latter may strike statisticians as unusual but the goal is not to build meaningful causal models but good predictive models for the missing data and there is no reason to believe that the future does not contain useful information about the past. Lastly, Amelia can incorporate prior believe about the missing data and their distribution in a number of ways: Ridge priors and observation-level priors. Constraints on the range of values an imputation variable can take on can also be controlled by specifying logical bounds and Amelia implements them via rejection sampling.

For imputation and model diagnostics Amelia also has some built-in capacity that includes plots of the densities of observed and imputed values (also available in the mi package). Overimputing allows for each of the observed values to be treated as missing and produces several hundred imputed values for observed data that can be used to judge the quality of the imputations. The mean of the imputed values should be close to the observed value. Overdispersed starting values for the imputation chains helps to avoid a known pitfall of the EM algorithm, i.e., getting stuck in a local region of the parameter space and producing non-maximum likelihood estimates. Time series plots demonstrate if the different chains using dispersed starting values converge over the course of the iterations to the same parameter value.

2.1 Advantages and their costs in the Amelia package

Amelia is a multi-purpose imputation tool aimed at giving the user considerable flexibility in specifying imputation models and control parameters. As with the mi and other software tools, it is still recommended that the user has a basic understanding of the problems occurring in missing data situations and the options to deal with them.

A general concern with Amelia is that many data commonly fail to fit to a multivariate normal distribution and using a multivariate normal model in Amelia is one of its pertinent critiques. Nonetheless, much evidence in the literature (discussed in King et al. 2001) indicates that the multivariate normal model used in Amelia usually works well for the imputation stage even when discrete or non-normal variables are included and when the analysis stage involves these limited dependent variable models.

Amelia's flexibility in model specification is fairly high – allowing for various classes of variables, transformations, polynomials in time and their interaction with the cross-section variable – but it is not infinite and may not fit all situations. Amelia currently does not allow the use of random effects models.

Just as in the case of the mi package, large data sets and/or large fractions of missing data slow down the imputation algorithm due to their impact on the EM algorithm. The algorithm may fail altogether to converge.

Imputation of variables that are logically linked such as through theoretical and/or empirical relationships, cannot easily be accommodated in Amelia. Two possibilities are the use of observation-level priors and logical bounds on the values that can be imputed (e.g., if A=B+C and A is incomplete, then a logical bound for A could be B+C+/-delta).

3. Comparison of the mi and Amelia packages in the context of estimating net migration

The analysis presented in this report used both the mi and Amelia software tools to produce custom-tailored imputations of urban and rural crude birth and death rates. Each uses certain options and must also be seen within the constraints of available time and computing power. For example, the Amelia procedure was run with largely reduced sets of predictor variables because their inclusion extended the run time by a factor of 10 on the available personal computer and Macbook Pro. The mi procedure was run in parallel chains on a computer cluster, which allowed the use of all available information in a unified modeling framework, i.e., simultaneous imputation of urban/rural birth and death data as opposed to their separate imputation in Amelia. On the other hand, time intervals between iterations were also considerable and therefore only 20 iterations were run, which may affect the convergence of the parameter estimates.

Appendix D. Issues with Currently Available Migration Data

In 2010, it is estimated that about 214 million people lived in a country different from the one they were born, or about three percent of the world population (UN Population Division 2010). As for domestic migration, some estimates suggest that around 740 million people have migrated between level 1 administrative units (states and provinces) – that is, they live in their country of origin but have moved away from their town or region of birth (Bell and Muhidin 2009). Migration has greatly accelerated with economic globalization, yet the research community is stymied in its ability to characterize international and even domestic migration because of poor quality data and divergent definitions.

Migration has been defined as a multidimensional and multifaceted phenomenon for which there is no all-embracing theory (Portes 1997; Brettel 2000). In order to understand it, it is necessary to adopt a broad conceptual approach, incorporating multiple levels of analysis within a longitudinal perspective, keeping in mind that migration behavior is embedded in social contexts and has temporal and spatial dimensions (Portes 1997; Massey 1990a, 1990b; White and Lindstrom forthcoming).

Stocks and flows

Population mobility can be measured in terms of stocks or flows (Bilsborrow et al. 1997:51). Migrant stock, a static measure, is the number of people who identify themselves or are identified as migrants at a certain point in time. This count can be obtained through census questions relating to birth location, country of origin, or location of the individual as of the last census.

Mobility can also be measured in terms of flows (inflows and outflows), which are counts of people moving into or out of an area over some period of time, generally a calendar year. This is a more problematic approach because “flows represent the dynamics of the process” and “they are considerable less tractable than stock measures” (Bilsborrow et al. 1997). For example, people entering and leaving the country several times in one calendar year could be counted just once (just one person) or more than once (several moves), depending on the time criteria when defining mobility.

Comparability problems in migration analysis

Cross country comparisons face several challenges due to differences in collection practices, for example: (a) differences in the way migration is defined and measured, and the type of data derived; (b) issues of temporal comparability (length of the interval); (c) differences in coverage of population and quality of data; and (d) the spatial units used, the division of space and the measurement of distance, which in turn is related as how migration is defined (Bell et al. 2002; Parsons et al. 2007).

Regarding (a), common sources of migration data are population censuses and registers, which report transitions (movers) and events (moves) respectively (Bell et al. 2002). Migration surveys provide richer data (e.g., places of residence, number of moves) but generally they cover small areas. In all cases, the selection of space and time frameworks affects the observation and measurement of the intensity and geographic pattern of migration flows. Temporal comparability across countries just makes these issues even more complicated.

Age structure, a key to migration selectivity, affects aggregate levels of mobility, geographic patterns of movement and timing of migration. Quality of data is critical, and varies widely. Census undercounts, for example, are not random, but selective of certain groups, among them migrants. The analysis of migration is affected by the modifiable areal unit problem. Decisions about geographies are often restricted because data are only available for administrative units, which may or may not serve the needs of the research question or the problem at hand. Another issue affecting comparability over time is changes in the number and boundaries of the administrative units.

Finally, variation in times intervals also affects comparability (Bell and Muhidin 2009). Discontinuities in the data due to country breakdown should be carefully monitored. Examples include the former Yugoslavia and Soviet Union.

Sources of Migration Data

Note that the full extent of the information collected on censuses and surveys may not be available because it has not been coded, remaining as raw data (Black and Skeldon 2009).

I. Population Censuses

A number of sources are derived from national population censuses complemented or not with other sources, as aggregates, microdata or both. Place of residence five years or one year before the census (recent migration), place of birth / citizenship (life time migration), and place of previous residence (with no defined time period) are common measures of migration in these sources.

UN Global Migration Database and the International Migration Stock: This database is the base of the International Migration Stock (UN Population Division 2010). The original version includes sex and age of immigrants and emigrants, but this has still to be incorporated in the IMS.

IPUMS International Collection: The objective of this collection from the Minnesota Population Center is to inventory, preserve, harmonize, and disseminate census microdata. It currently contains 55 countries, with several censuses for most of them. Data are coded and documented consistently across countries and over time to facilitate comparative research. The classical census migration questions (place of birth and place of residence at a fixed point before the census date) are divided in the IPUMS into two groups of variables at the individual level: nativity and birthplace variables; and migration. The main limitation is the small number of countries included.

The Bilateral Migrant Stock Database (Development Research Centre on Migration, Globalisation and Poverty (Migration DRC)): This database comprises two origin-destination matrices, at the country level, for 226 countries and territories, based on the 2000 round of censuses. One matrix records foreign born population by country of birth and the second the population by nationality (Parsons et al. 2007:3). The main objective of the database is "to include as many of the world's migrants as possible, to assign them *all* to specific countries of origin with the highest degree of accuracy and to produce as full and comparable a bilateral database of international migration stocks as is possible" (Parsons et al. 2007: 9). Last available data from the original source was preferred, census mainly but also population registers. Data on both foreign born and foreign nationals were compiled where feasible. Population registers were then drawn upon where censuses were unavailable for the 2000 round of censuses. In some cases where neither source was available, data were obtained from reliable secondary sources that cite the original. Some regions of the world provide significantly better data than others, and some simply do not exist in the public domain or even at all. While the data for Europe, the Americas and much of Oceania are of a fair standard, the data for parts of Asia and much of Africa are of more dubious quality.

UN World Population Prospects: Include net migration rate and net migration for five year periods for approximately 198 countries, between 1950 and 2010.

Others: CELADE's IMILA and MIALC databases and World Bank's International Remittances**.**

II. Registers and Administrative Records

These can be considered as continuous systems. These include border statistics for international migration, immigration registration in port of entry and departure (which may or may not be publicly available). Less commonly a register can also be used for internal migration, such as the household registration system in China (hukou system), or population registers in several European countries.

III. Surveys

Specialized migration surveys are of course the best tool for generating detailed data (Black and Skeldon 2009). Global guidelines have been developed (Bilsborrow et al. 1984) as migration surveys require careful design and are usually expensive. It could take different forms: retrospectives migration histories (trajectories), or longitudinal surveys with several rounds of data collection. Examples include the Mexican Migration Project (MMP), the Latin American Migration Project (LAMP), and the Reseau Migrations et Urbanisation en Afrique de l'Ouest (REMUAO).

Other surveys such as the Demographic and Health Surveys (DHS) and the Multiple Indicator Cluster Surveys (MICS) also provide some data on migration, although not consistently.

Periodic surveys on labor forces and living conditions, and periodic household surveys on income and family expenses could also be used for migration statistics. Migration questions are similar to the ones in the population censuses: place of birth, place of previous residence. Household surveys were used by the World Bank for the 2009 report.

Appendix E: Characterizing Precision and Accuracy of Data Inputs

There are several ways in which the data inputs utilized here vary with respect to precision and accuracy, and some of these can be specified quantitatively. Here we review the sources of variation and present some high-level indicators.

Size of census input units

The gridded census data set we utilize (CIESIN 2011) has as inputs maps provided by national census authorities or other bodies assigning population totals to spatial units. These spatial units are typically administrative areas such as counties or districts; sometimes they are even smaller but sometimes they are larger. The smaller the input unit, the greater the spatial precision and the higher the accuracy of our gridded map. So, for Figures E1-E3 larger numbers reflect bigger average size of input units and therefore greater uncertainty.

The only consistent correlate to census input size is population density: higher densities are associated with smaller input units. Levels of per-capita income are not correlated with this measure. Another correlate, which is more of an artifact of geopolitics, is that small countries tend to have smaller input units, simply because their country size sets an upper limit which is comparatively small. Therefore the small island states have greater spatial precision than average.

Figure E1. Mean Size of Original Population Data Input Units in Square Kilometers by Ecosystem

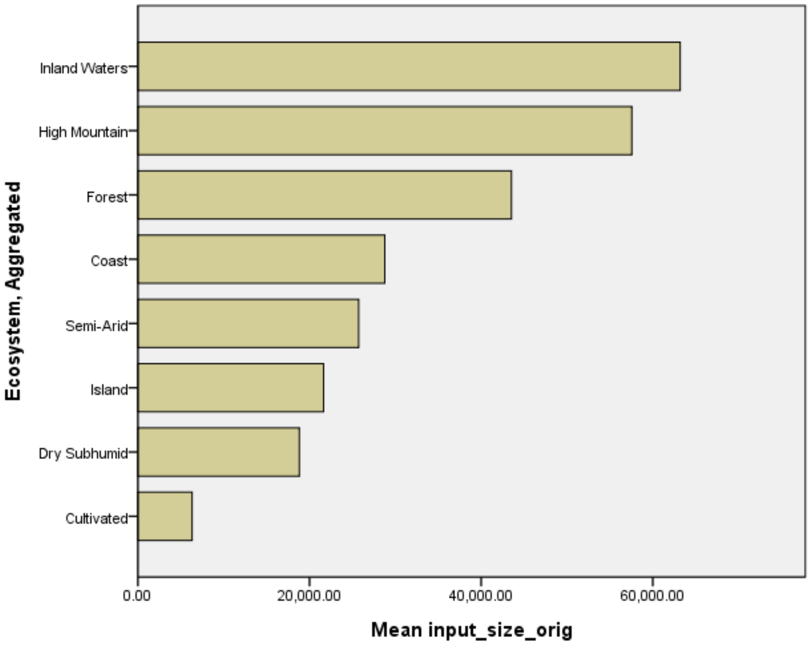


Figure E2. Mean Size of Original Population Data Input Units in Square Kilometers by UN Region

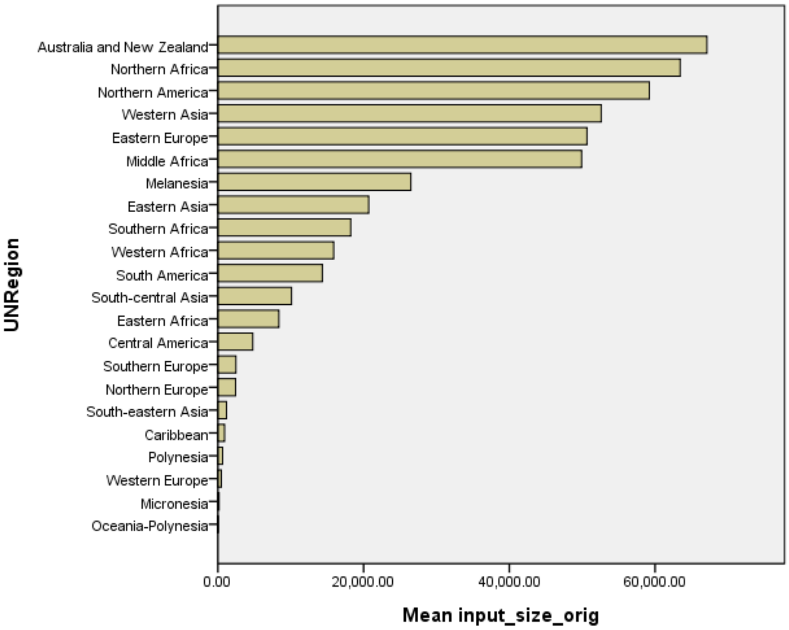
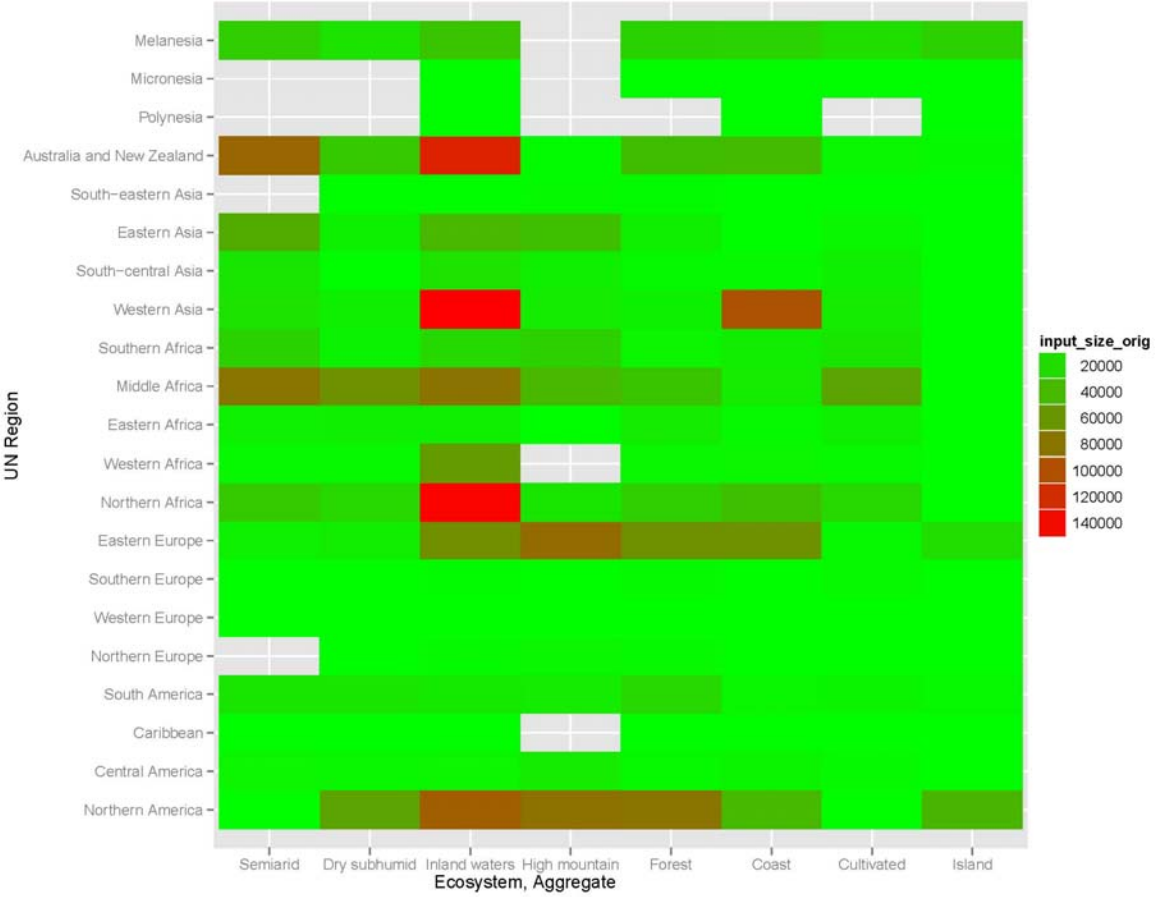


Figure E3. Mean Size of Original Population Data Input Units in Square Kilometers by Ecosystem and UN Region



Frequency of censuses

In general most countries take a census every ten years; however some countries supplement decadal censuses with additional mid-decade censuses, and some are not able to undertake a census every decade. Failure to undertake a decadal census is usually associated with war or other political disruptions. Because political disruption is a driver of movement we lack data inputs where we would most want them. Although there is not much variation in frequency of censuses by ecosystem (Figure E4), there is substantial variation by region (Figure E5).

Figure E4. Mean Number of Censuses from 1970-2010 by Ecosystem

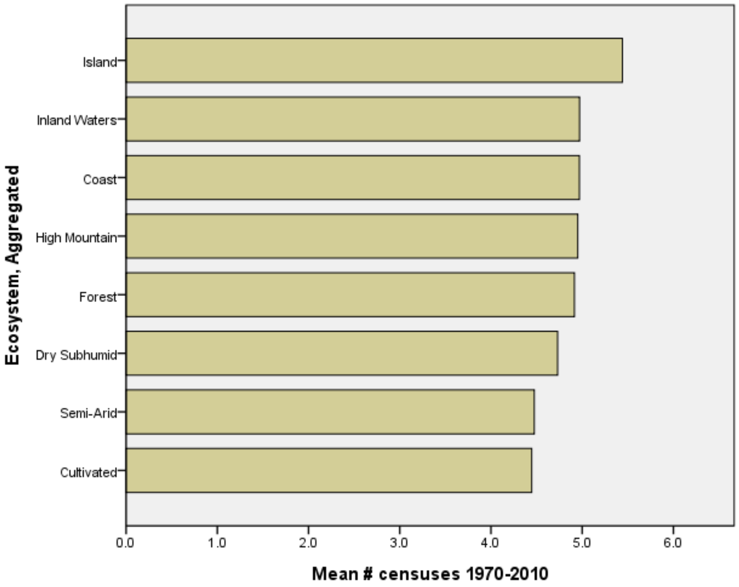


Figure E4. Mean Number of Censuses from 1970-2010 by UN Region

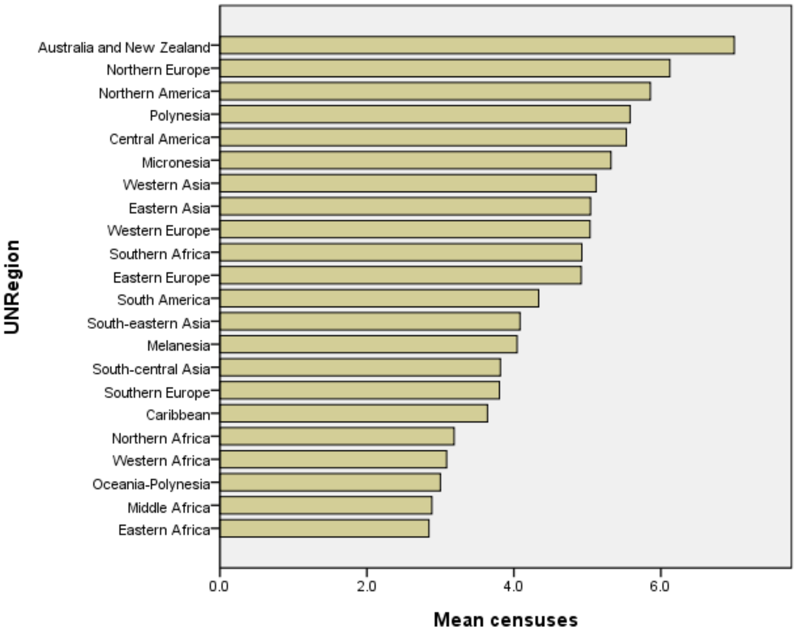
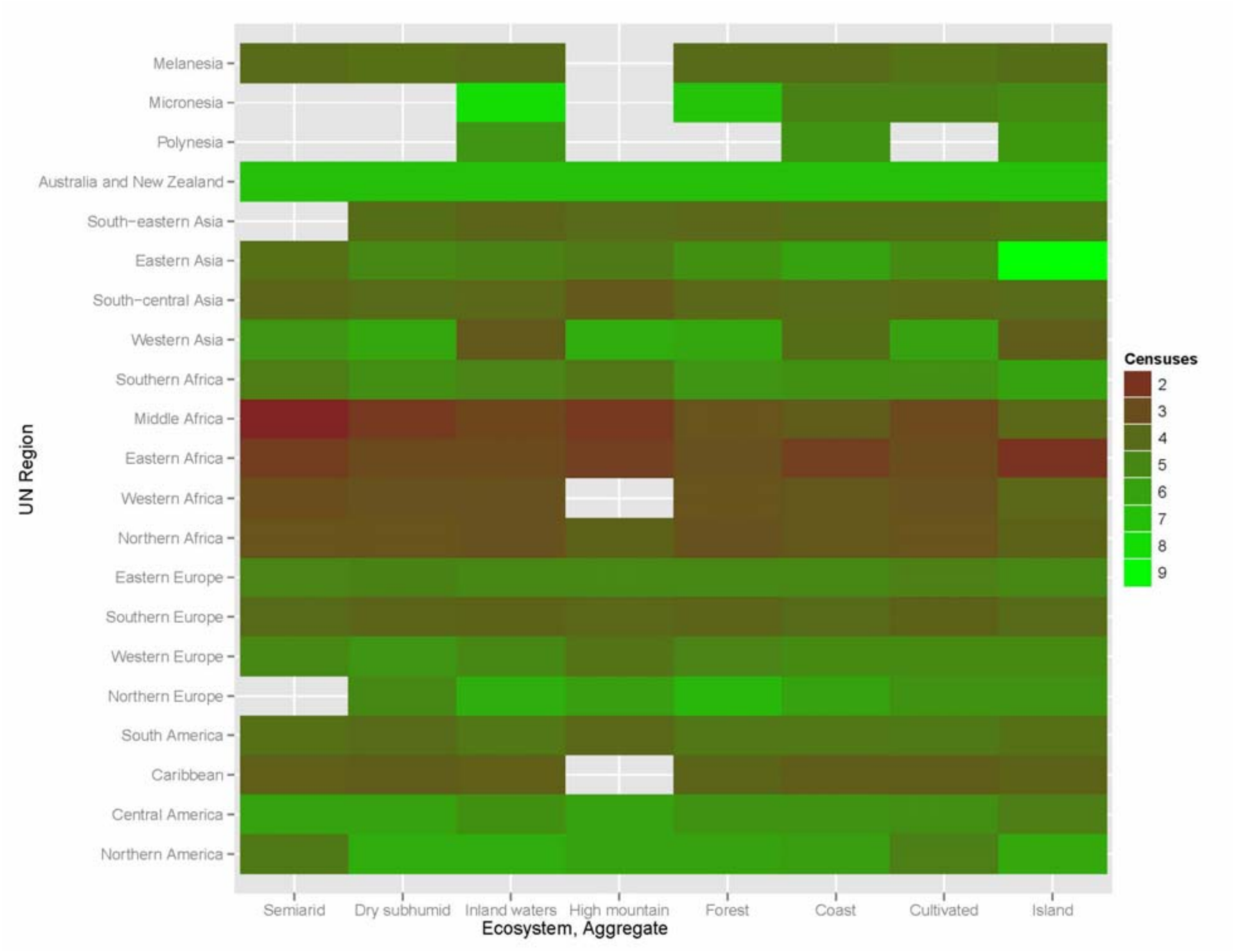


Figure E5. Mean Number of Censuses from 1970-2010 by UN Region and Ecosystem



Subnational variation in birth and death rates

Because our method infers net migration from a comparison of observed population change with the population change expected from birth and death rates, our accuracy depends on the degree to which we can estimate spatially specific values for birth rates and death rates. There are few global databases that permit such inferences, although many countries have their own sources of data that would be relevant. Because of time limitations we were limited to two global sources of information – the UN *Demographic Yearbooks* and the Demographic and Health Surveys. Both sources provide estimates for many countries of birth rates and death rates stratified by urban and rural areas. For China, which had no such information in either source, we utilized country-specific information for 1990 (CITAS et al. 1997) to make sure that such a large and ecologically diverse area had subnational inputs on birth and death rates.

Relying exclusively on urban/rural differences as the basis for inferring spatial patterns of birth and death rates is itself a source of inaccuracy. Due to our previous experience generating subnational maps of infant mortality rates we have some idea of that level of effort that would be required to generate higher quality spatial estimates of birth and death rates. Such a level was not feasible in the time frame available to us for the present effort.

We use as a quantitative indicator of accuracy the number of data points, by country, reflecting urban/rural differences in birth and death rates. As can be seen from the following graphs, the number of observed birth and death rates varies substantially by region – with many of the republics of the former Soviet Union and Eastern European countries having a large number of data points up until the late 1980s, but comparatively few observations in Oceania, Africa, and even North America. This may be a function of national reporting – the data certainly exist for many developed countries to compute urban and rural birth and death rates, but the interest in such data is limited, so these countries do not report their data to the United Nations.

Figure E6. Mean Number of Urban/Rural Birth or Death Rate Data Points from 1970-2010 by Ecosystem

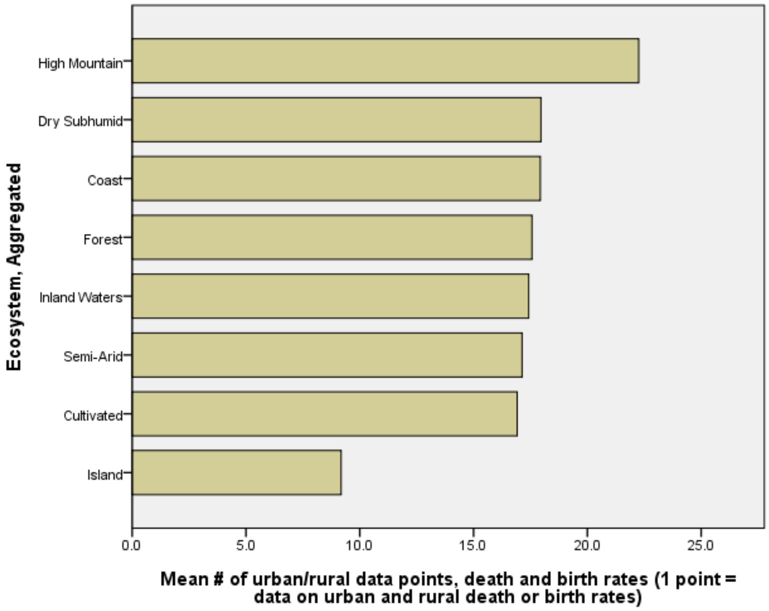


Figure E7. Mean Number of Urban/Rural Birth or Death Rate Data Points from 1970-2010 by Region

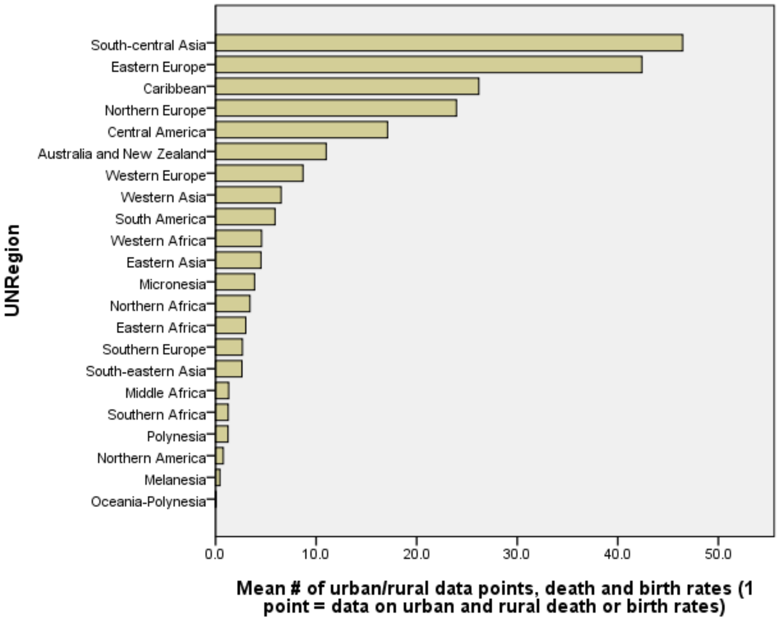
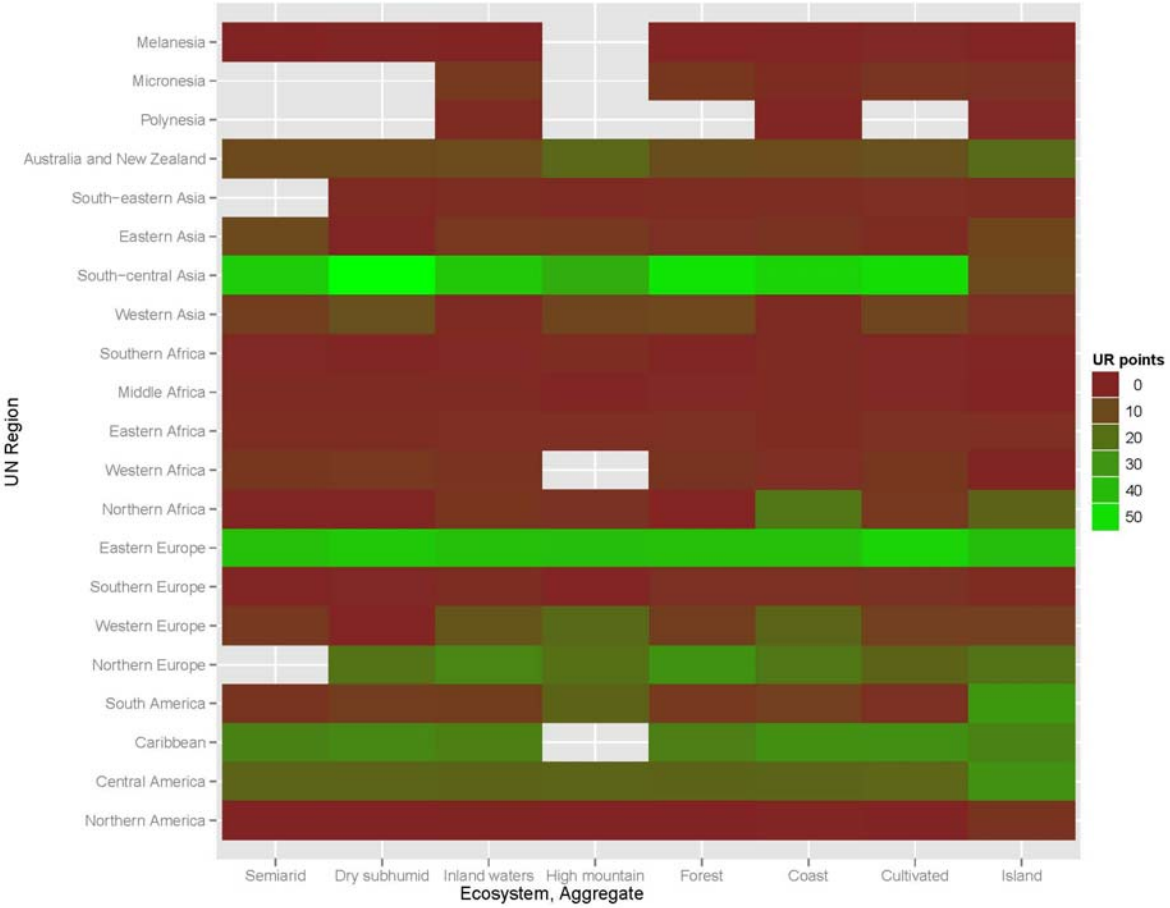


Figure E8. Mean Number of Urban/Rural Birth or Death Rate Data Points from 1970-2010 by UN Region and Ecosystem



References

Abayomi, K., A. Gelman and M. Levy (2008). Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society, Series C Applied Statistics*, 57(Part3): 273-291

Adamo S. and de Sherbinin, A. forthcoming. "The impact of climate change on the spatial distribution of populations and migration" In *Proceedings of the Expert Group Meeting on Population Distribution, Urbanization, Internal Migration and Development*. Edited by United Nations. Population Division: UNDESA.

Adger, W., S. Agrawala, and M. M. Q. Mirza. 2007. "Assessment of adaptation, practices, options, constraints and capacity." in *Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.*, edited by Martin Parry. Cambridge: IPCC / Cambridge University Press.

Aide, M. and H. R. Grau (2004). "Globalization, Migration, and Latin American Ecosystems." Science **305**(5692): 1915-16.

André, M-F. 1998. "Depopulation, Land-Use Change and Landscape Transformation in the French Massif Central." *Ambio* 27(4):351-53.

Balk, D., F. Pozzi, G. Yetman, U. Deichmann, and A. Nelson. 2004. The "Distribution of People and the Dimension of Place: Methodologies to Improve the Global Estimation of Urban Extents," Available at http://sedac.ciesin.columbia.edu/ gpw/docs/UR_paper_webdraft1.pdf

Balk, D., G. Yetman, and A. de Sherbinin. 2010. "Construction of Gridded Population and Poverty Data Sets from Different Data Sources." Proceedings of European Forum for Geostatistics Conference 5 - 7 October, 2010 Tallinn, Estonia. Available at http://www.efgs.info/geostat-project/efgs-conference-2010-e-proceedings/e-proceedings_EFGS_2010_Deliverable_WP4.pdf/view (accessed 28 February 2011).

Barbieri, A., E. Domingues, et al. (2010). "Climate change and population migration in Brazil's Northeast: scenarios for 2025–2050." Population & Environment **31**(5): 344-370.

Barbieri, A., R.L.M. Monte-Mor, and R. Bilsborrow (2009). "Towns in the Jungle: Exploring Linkages Between Rural-Urban Mobility, Urbanization and Development in the Amazon", In A. de Sherbiniin, A. Rahman, A. Barbieri, J.C. Fotso, and Y. Zhu (eds.). *Urban Population-Environment Dynamics in the Developing World: Case Studies and Lessons Learned*. Paris: Committee for International Cooperation in National Research in Demography (CICRED).

Bell, M. and S. Muhidin (2009). Cross-national comparisons of internal migration. *Human Development Research Paper*, UNDP.

Black, R. and R. Skeldon (2009). "Strengthening data and research tools on migration and development." *International Migration* **47**(5): 3-22.

Brettell, C. 2000. “Theorizing migration in anthropology. The social construction of networks, identities, communities and globalscapes”. In Brettell, C. and J. Hollifield, eds. *Migration theory: talking across disciplines*. New York, Routledge. Pp. 98-135.

Bright, E.A. *personal communication,* January 2011.

Campbell, K.M., J. Gulledge, J.R. McNeill, J. Podesta, P. Ogden, L. Fuerth, R.J. Woolsey, A.T.J. Lennon, J. Smith, R. Weitz, and D. Mix. 2007. *The Age of Consequences: The Foreign Policy and National Security Implications of Global Climate Change*. Washington, DC: Center for Strategic and International Studies and Center for New American Security.

Carr, D. (2009). "Rural migration: The driving force behind tropical deforestation on the settlement frontier." Progress in Human Geography **33**(3): 355-378.

Center for International Earth Science Information Network (CIESIN), Columbia University; International Food Policy Research Institute (IFPRI); The World Bank; and Centro Internacional de Agricultura Tropical (CIAT). 2011. Global Rural-Urban Mapping Project (GRUMP), Version 1: Population Counts. Palisades, NY: Socioeconomic Data and Applications Center (SEDAC), Columbia University. Available at http://sedac.ciesin.columbia.edu/gpw.

China in Time and Space (CITAS), University of California-Davis; China in Time and Space (CITAS), University of Washington; Center for International Earth Science Information Network (CIESIN) (1997). China Dimensions Data Collection: China County-Level Data on Population (Census) and Agriculture, Keyed to 1:1M GIS Map. Saginaw, MI: CIESIN. Available at http://sedac.ciesin.columbia.edu/china/.

Conelly, W.T., 1992. Agricultural intensification in a Philippine frontier community: impact on labor efficiency and farm diversity. *Human Ecology* 20 (2), 203–223.

Corvalan, C., S. Hales, et al. (2005). Ecosystems and human well-being : health synthesis. A report of the Millennium Ecosystem Assessment. Geneva, World Health Organization

Craviotti, C. and S. Soverna. (1999). Sistematización de casos de pobreza rural. Buenos Aires, PROINDER.

DeFries, R., S. Pagiola et al. (2005). “Analytical Approaches for Assessing Ecosystem Condition and Human Well-being.” Chapter 2 in: R. Hassan, R. Scholes, and N. Ash (eds), *Ecosystems and human well-being : current state and trends : findings of the Condition and Trends Working Group*. Washington, DC: Island Press.

de Sherbinin, A., L. VanWey, K. McSweeney, R. Aggarwal, A. Barbieri, S. Henry, L. Hunter, W. Twine, and R. Walker (2007). “Household Demographics, Livelihoods and the Environment.” *Global Environmental Change*, Vol. 18, pp. 38-53.

Environmental Change and Forced Migration Scenarios (EACH-FOR). (2009). *Project Synthesis*. Available at http://www.each-for.eu/documents/EACH-FOR_Synthesis_Report_090515.pdf.

Feng, S., A. Krueger, et al. (2010). "Linkages among climate change, crop yields and Mexico–US cross-border migration." PNAS **107**(32): 14257-14262.

Gazetteer, The (2004). The World Gazetteer, Current population figures for cities, towns and administrative divisions of all countries largest cities of the world, last visited june 2004, http://www.world-gazetteer.com/home.htm.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). Bayesian Data Analysis (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D. F., & Meulders, M. (2005). Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics*, *61*, 74-85. doi:10.1111/j.0006-341X.2005.031010.x

Gelman, A., G. King; C. Liu (1998). Not asked and not answered: Multiple imputation for multiple surveys. *Journal of the American Statistical Association*, Vol. 93, No. 443. (Sep., 1998), pp. 846-857.

Gerland, P. *personal communication*, 11 March 2011.

Goldewijk, K.K. (2005) Three centuries of global population growth: A spatial referenced population density database for 1700-2000. *Population and Environment*, 26:343-367.

Goldewijk, K.K., & Van Drecht, G. (2006) HYDE 3: current and historical population and land cover. *Integrated modeling of global environmental change. An overview of IMAGE 2.4* (ed. By A.F. Bouwman, T. Kram and K. Klein Goldewijk) pp. 93-111. Netherlands Environmental Assessment Agency, Bilthoven, The Netherlands

Goldewijk, K.K., (2001). Estimating Global Land Use Change over the Past 300 Years: The HYDE Database: *Global Biogeochemical Cycles* 15, (2), 417–433.

Goldewijk, K.K., Beusen, A. & Janssen, P. (2010) Long term dynamic modeling of global population and built-up area in a spatially explicit way: HYDE 3.1. *The Holocene*, 20, 565-573

Goldewijk, K.K., C .G.M., and J.J. Battjes, A hundred year (1890-1990) database for integrated environmental assessment (HY DE, version1 .1). *Rep.4 22514002*, National Institute of Public Health and the Environment (R IVM), Bilthoven, The Netherlands, 1997.

Goldewijk, K.K., de Man, R., Meijer, J., & Wonink, S. (2004). *The Environmental Assessment Agency. 2004 World Regions and Subregions*. National Institute for Public Health and the Environment (RIVM), KMD Memo M001/04.

Goldewijk, K.K.,, The role of historical GIS data in integrated models of global change, in *Proceedings of the Third Joint European Conference and Exhibition on Geographical Information*, pp. 317-328, IOS Press, Vienna, 1997.

Gonzalez, R., A. Otero, L. Nakayama, and S. Marioni. 2009. "Tourism mobilities and amenity migration: problems and contradictions in the development of mountain resorts" *Revista de Geografía Norte Grande* (44):75-92.

Grigg, D. (1987). The industrial revolution and land transformation. In Wolman M.G. Fournier F.G.A. (Eds.), *Land Transformation in agriculture*, Chichester, New York: SCOPE 32, John Wiley & Sons.

Hassan, R., R. Scholes, and N. Ash (eds). 2005. *Ecosystems and human well-being : current state and trends : findings of the Condition and Trends Working Group*. Washington, DC: Island Press. Available at http://www.maweb.org/documents/document.765.aspx.pdf (Accessed on 24 March 2011).

Henry, S., B. Schoumaker, and C. Beauchemin (2004). The impact of rainfall on the first out-migration: A multi-level event-history analysis in Burkina Faso. *Population and Environment* **25**, 423-460.

Hidalgo, R., A. Borsdorf, and F. Plaza (2009). "Pleasure lots near Santiago de Chile and Valparaiso. Amenity migration the Chilean way?" *Revista de Geografía Norte Grande* (44):93-112.

Kasfir, N. (1993). Agricultural transformation in the Robusta coffee/banana zone of Bushenyi, Uganda. In: Turner, II, B.L., Hyden, G., Kates, R. (Eds.), Population Growth and Agricultural Change in Africa. University Press of Florida, Gainesville, pp. 41–79.

Keys, E., and W. McConnell (2005). Global change and the intensification of agriculture in the tropics. *Global Environmental Change* 15:320–337.

Körner C, Ohsawa M. (2005). Mountain system. Ecosystems and Human Well Being, Current states and trends. Millennium Ecosystem Assessment. Washington, DC: Island Press.

Lahmeyer, J. (2004). Populstat database. Growth of the population per country in a historical perspective, including their administrative divisions and principal towns, http://www.library.uu.nl/wesp/populstat/populhome.html.

ORNL (2006). Landscan global population database, the 2004 revision. Oak Ridge National Laboratory. Available at: http://www.ornl.gov/landscan

Livi-Bacci M (2007) A Concise History of World Population. Fourth edition. Oxford: Blackwell Publishing.

Maddison, A. (1995). *Monitoring the World Economy*. Paris: OECD Development Centre pp. 1820–1992.

Massey, D. 1990a. “Social structure, household strategies, and the cumulative causation of migration”. *Population Index.* 56(1):3-26

Massey, D. 1990b. “The social and economic origins of immigration”. *Annals of the American Academy of Political and Social Science.* 510:60-72.

McEvedy C, Jones R (1978) World Atlas of Population History. Hammondsworth: Penguin Books Ltd.

Mitchell B. R. (1993). *International Historical Statistics*, The Americas: 1750–1988, p. 817. Indianapolis, IN: MacMillan.

Mitchell, B. R. (1998a). *International historical statistics*, Europe: 1750–1993, 4th Ed., In: MacMillan, Indianapolis, Ind. p. 959.

Mitchell, B. R. (1998b). *International Historical Statistics*, Africa, Asia & Oceania: 1750–1993, 1113 pp, 3rd Ed. Indianapolis, IN: MacMillan.

Montgomery, M. (2008). "The urban transformation of the Developing World." Science **319**: 761-764.

OSCE. Economic and Environmental Activities. 2005. "Background paper for Session III." in *13th Economic Forum*. Vienna.

Parsons, C., R. Skeldon, et al. (2007). Quantifying international migration: a database of bilateral migrant stocks. Policy Research Working Paper. World Bank

Portes, A. 1997. "Immigration theory for a new century: some problems and opportunities". *International Migration Review.* 31(4):799-825

Rain, D. (1999). *Eaters of the Dry Season: Circular Labor Migration in the West African Sahel* . Boulder, CO: Westview Press.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys,* New York: John Wiley & Sons, Inc.

Schelhas, J. (1996). Land use choice and change: intensification and diversification in the lowland tropics of Costa Rica. *Human Organization* 55 (3): 298–306.

Scott, L. (2006). Chronic Poverty and the Environment: A Vulnerability Perspective. Chronic Poverty Research Centre Working Paper No. 62. Overseas Development Institute, London, UK.

Solomon, A. (Ed.), Report from the IMAGE 2 advisory board meeting in Amsterdam, 20-22 June 1994. *NRP Report no. 00-13*, Nat. Res. Prog. on Global Air Pollut. and Clim. Change, Bilthoven, The Netherlands, 1994.

Tobler, W., U. Deichmann, J. Gottsegen, and K. Maloy (1995). The global demography project. *Tech. Rep. TR-95-6*, National Center for Geographic Information and Analysis (NCGIA), Santa Barbara, California.

UN Population Division (2010). *International Migration Stock: The 2009 Revision*. New York: United Nations.

United Nations Statistics Division. *Demographic Yearbook*. Several editions published from 1972 to 2008. http://unstats.un.org/unsd/demographic/products/dyb/dyb2.htm

United Nations, Department of Economic and Social Affairs, Population Division (2009). *World Population Prospects: The 2008 Revision.* New York: United Nations.

United Nations, Department of Economic and Social Affairs, Population Division (1997). *World Population Prospects: The 1996 Revision.* New York: United Nations.

Valdivia, C., A. Seth, J. L. Gilles, M. Garcia, E. Jimenez, J. Cusicanqui, F. Navia, and E. Yucra. 2010. "Adapting to Climate Change in Andean Ecosystems: Landscapes, Capitals, and Perceptions Shaping Rural Livelihood Strategies and Linking Knowledge Systems." *Annals of the Association of American Geographers* 100(4):818-34.

Warner, K., C. Enrhart, A. de Sherbinin, S. Adamo, and T. Chai-Onn. (2009). In search of shelter: mapping the effects of climate change on human migration and displacement. A policy paper prepared for the 2009 Climate Negotiations. Bonn, United Nations University, CARE, and CIESIN-Columbia University.

WBGU (German Advisory Council on Global Change). 2007. *Climate Change as Security Risk*. Berlin: WBGU.

White, M. and D. Linstrom. 2005. “Internal migration”. In Poston D. and M. Micklin. *Handbook of Population,* Springer.

Xu, J., R. Sharma, et al. (2008). "Critical linkages between land-use transition and human health in the Himalayan region" *Environment International*, **34**: 239-247.