

Transforming Population Data for Interdisciplinary Usages: From census to grid

Uwe Deichmann
The World Bank
1818 H Street
Washington, DC

Deborah Balk
CIESIN
Columbia University
PO Box 1000
Palisades, NY 10964

Greg Yetman
CIESIN
Columbia University
PO Box 1000
Palisades, NY 10964

1 October 2001

Acknowledgments

We would like to thank Chris Small, Ion Mateescu, and Chandra Giri. Funding for the writing of this paper was provided by the NASA Socioeconomic Data and Applications Center (SEDAC) to Columbia University under contract NAS5-98162.

In studies of population and the environment, it is necessary that population and environmental data are reduced to a scale or scales that mutually compatible for analysis. Typically this has been done by aggregating individuals residing in common geographic or administrative areas. The “area concept” commonly introduces the problem of “ecological fallacy whereby the population is defined by its territory rather than by its demographic or social distinctiveness...It makes more geographical sense for readily identifiable areas such as the globe, continents, and islands than for many countries and administrative units which have little environmental identity or uniformity,” (Clarke, 1995, p. 7).

Despite this, population data are routinely collected by censuses and surveys and compiled for political or administrative units. This approach, while essential for certain types of analyses, is limiting for cross-disciplinary studies particularly those of the environment. Such studies require data to be referenced to a uniform coordinate system rather than irregular administrative units. Using fairly simple techniques and good data, conversion between data organized at the administrative or political unit and that which is in a uniform system (such as latitude-longitude quadrilaterals) can be made. Due to the volume of the data and the frequency of data collection, these conversions can only be done periodically. This paper discusses the methods used to make this conversion, associated error, the strengths and shortcomings of this approach in creating the Gridded Population of the World (GPW) data base (v2), shown in Figure 1.

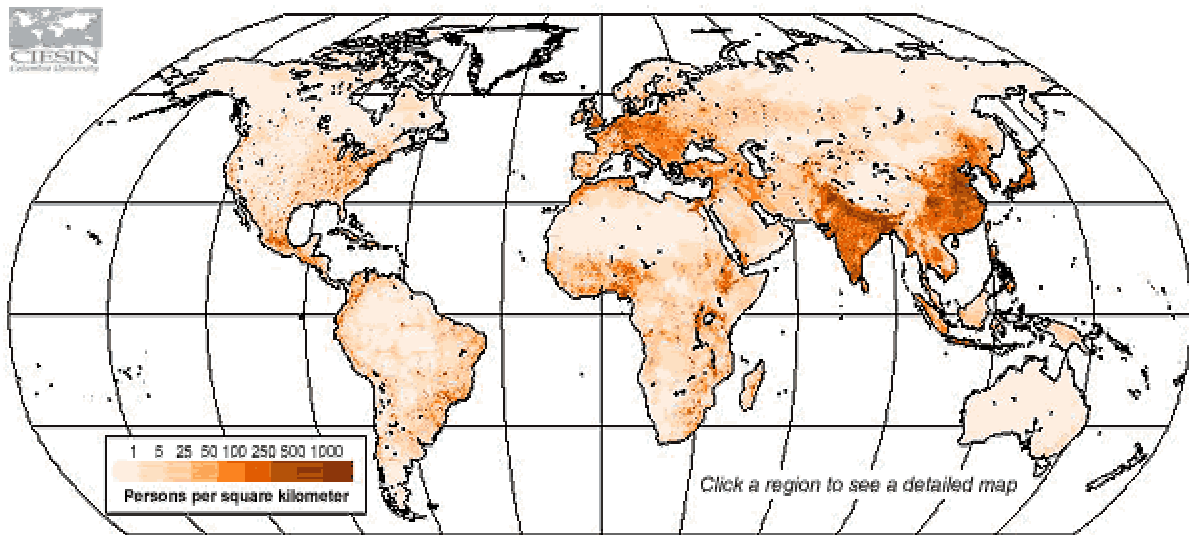


Figure 1. Global population density, 1995 (United Nations adjusted, Robinson Projection)

Resulting from a 1994 workshop on Global Demography, there was consensus that a consistent global database of population totals in raster format¹ would be invaluable for interdisciplinary study. This led to the initial version of GPW which was released in 1995 (Tobler et al., 1995). Since that first release, higher resolution data sets have been compiled for various regions of the world including [Africa](#), [Asia](#) (including [Russia](#)) and Latin America. In addition, public and commercial data producers have created high resolution administrative boundary data sets that are often linked to comprehensive census data sets. The emphasis for the version 2 update was placed on improving the resolution of the input data layers of

¹ Also known as a grid, a raster data set is a type of tessellation (mosaic) that divides a surface into uniform cells or pixels. The raster data model is common for representing phenomenon that vary continuously over a surface.

administrative boundaries and on producing better population estimates for each unit. No effort was made to “model” population distribution by distributing population totals over the grid cells that fall into each unit. Instead, the gridding approach used a simple proportional allocation of administrative unit population totals over grid cells. Thus, no ancillary data were used to predict population distribution or to revise the population estimates as in the approaches reviewed in Deichmann (1996a) and implemented in various forms, for example, by Honeycutt and Wojcik (1990), Deichmann and Eklundh (1991), Sweitzer and Langas (1995), Veldhuizen et al. (1995), Deichmann (1996b and 1998) or the LandScan Project (Dobson and Bright, 2000). The following sections describe the details of the gridding methodology used for the latest version of GPW.

1. INPUT DATA

The inputs into this project are relatively straightforward: administrative boundary data and population estimates associated with those administrative units.

1. A. Administrative unit boundaries

Geographic Information System (GIS) data sets of administrative or statistical reporting units are produced by national statistical and mapping agencies, research projects, and commercial data vendors. For improvements to the GPW v1 database, this version relied mostly on publicly available boundary data sets produced for Africa, Asia and Latin America. Additional boundary data sets—for Europe, Canada, Australia/New Zealand, India, Malaysia, and the newly independent states of the former Soviet Union—were obtained from commercial data vendors or statistical agencies that sell boundary data on license. The boundary data sources for each country are listed in the table of [country-specific documentation](#).

The use of data from multiple sources and the variable size of countries required that we treat some units that would not be treated politically as separate countries as such (e.g., Puerto Rico). The use of these units as countries in the production of GPW is not intended to represent them as recognized or unrecognized national political entities. We merely followed the UN list of countries and territories used, for example, in the World Population Prospects that is published by the Population Division (UN, 1998). (Details are contained within the table of [country-specific documentation](#).)

To ensure consistency at international borders, most national boundaries in the source data were replaced by the political boundaries from the Digital Chart of the World (DCW). While not perfect, these boundaries are the most widely used template for global or continental GIS studies. Exceptions are parts of Europe, North America and Australia, which either already had matching international boundaries or where the quality of international boundaries was much better than DCW.

In total, we assembled boundaries for 127,082 administrative units; 60,911 of these units are census tracts in the United States. Even without the very detailed information for the USA, however, the database provides significantly higher resolution than the previous version of GPW which was based on about 19,000 units (Tobler et al., 1997). This resolution also far surpasses that of other global gridded databases, such as LandScan (Dobson, 2000). Resolution (area) of the administrative units vary by country. The table in the [country-specific documentation](#) shows the average resolution, along with other summary information, for each country. The average resolution can be thought of as the “cell size” if all units in a country were square and of equal size. It is calculated as follows:

$$\text{Mean resolution in km} = \sqrt{(\text{country area})/(\text{number of units})}$$

Table 1 shows the countries with the highest and the lowest available resolution (ignoring very small countries and areas, most of which consist of only one administrative unit). Resolution is to some extent determined by the geographic size and average population density of a country. That is, smaller countries have a relatively higher resolution even before adjusting for the number of administrative units. Some of the highest resolution countries are relatively small (e.g., Luxembourg, El Salvador) and a comparatively larger number of administrative units are generally necessitated by the presence of relatively densely-distributed populations (e.g., India², Netherlands). In general, however, there is no consistency between countries with regard to the resolution of administrative units. This is, in part, due to data availability which varies by country, and also to the fact that administrative units are often based on historic rather than designed boundaries. It should be noted, however, that the designation of administrative levels is sometimes ambiguous. For instance, some countries have geographic regions, which serve no administrative purpose, but may be used for statistical data reporting.

Among the countries with the lowest resolution, some include vast, largely uninhabited areas, where administrative units tend to be very large (e.g., Greenland, Libya). For other countries in this list, higher resolution administrative units boundaries were simply not available for this project (e.g., Bosnia, Iran). Some are a combination of each problem (e.g., Chad, Egypt). The variation in mean resolution depends considerably on combination of geographic and demographic characteristics of the given country and thus are not always comparable. For example, level-three administrative units in Canada can vary from a city-district to large tracts of uninhabited land whereas the same level in the continental United States varies much less in area.

Table 1. Countries with the highest and lowest available resolution, by average resolution (km).

Highest resolution	km	Lowest resolution	km
Switzerland	3.7	Egypt	196.3
Luxembourg	4.7	Algeria	222.8
Portugal	4.7	Bosnia Herzegovina	226.3
Belgium	7.2	Iran (Islamic Republic of)	262.0
Spain	7.9	Angola	263.2
Netherlands	8.1	Libyan Arab Jamahiriya	265.3
El Salvador	8.9	Mongolia	294.9
Puerto Rico	10.8	Chad	302.8
Slovenia	11.7	Saudi Arabia	374.2
France	12.2	Greenland	380.8

Table 2 show a summary of the administrative-level used in the creation of GPW, version 2. Of the 226 political-administrative units in our data base, only one was a level-four (Australia). All of the level-zeros were city-states or islands. But only 8 percent of all countries—or 10 percent of all countries with level 1 data or higher—were based on level-three data. In some instances, we had higher level data

² For India, we purchased data for the third-level administrative unit (i.e., *tehsil*), although fourth-level data is reportedly also available for purchase.

of one kind or another. For example, although we had level-two spatial data for Guyana and Suriname, we used level-one data, because we did not have the corresponding population data.

Table 2. Summary of administrative levels.

<i>Administrative Level</i>	<i>Frequency</i>	<i>Cumulative %</i>	<i>US equiv</i>
0	47	21.17%	Nation
1	68	51.80%	State
2	88	91.44%	County
3	18	99.55%	Tract
4	1	100.00%	Block
Total	222		

1.B. Population estimates

We collected the most recent population estimates available.³ The dates of the most recent censuses range from 1967 to 1999. For each administrative unit we produced a population estimate for 1990 and 1995. For a small share of the countries (14 of the total, or roughly 6%), we had census figures or official estimates for each of these years. For most of the remaining countries, we used two recent census totals or official estimates to compute an average annual population growth rate, as follows:

$$r = \frac{\log e \frac{P_2}{P_1}}{t},$$

where r is the average rate of growth, P_1 and P_2 are the population totals for the first and second reference years, and t is the number of years between the two census enumerations. This rate was then applied to the census figures to interpolate or extrapolate population totals to 1990 and 1995. For example, the 1995 estimate is calculated:

$$P_{1995} = P_1 e^{rt}$$

Thirty-eight countries had only one population estimate. This includes newly formed states (e.g., Croatia, Palestinian National Authority) as well as countries that for either economic or political reasons have not conducted a census or released census results since 1990 (e.g., Afghanistan, Albania). For countries for which we had no subnational boundaries (as in the case of most of the Pacific and Caribbean Islands), the national level UN estimate were used (United Nations 1999). Only a few larger countries had no population estimates (e.g., Bosnia Herzegovina, Kuwait, Singapore) that were available to us. Details on the sources for the population data are listed in the [country-specific documentation](#).

The best-possible geographic unit was used to estimate population change between 1990 and 1995. For example, for the United States, we have tract-level (i.e., administrative-level 3) data population in 1990, but county-level (i.e., administrative-level 2) data for 1990 and 1995. We therefore applied the county-level change from 1990 to 1995 to all tracts in a given county.

³ Subject to our project budget constraints. Although many countries freely provide population, and boundary data, others do not. We spent roughly US\$20,000 in 1999 to acquire more recent boundary and (for some of these countries) population data for Australia, Canada, India, Malaysia, New Zealand, most of Western and Central Europe and the former Soviet Republics.

Table 3 shows the countries with the lowest and highest average population per administrative unit. The range spans vastly, from 1,500 to nearly 3.5 million per unit. As with resolution, this indicator varies considerably by country. Again this is partly a reflection of the detail of available data (e.g., Bosnia), partly due to the size of the country (Singapore), or a combination of these (e.g., Korea). Even more than with resolution, this indicator clearly identifies countries for which higher resolution boundaries and population figures should be compiled in future updates of this database.

Table 3. Countries with the lowest and highest average population per administrative unit (1995 UN estimates)

Lowest pop per unit	Population (1000s)	Highest pop per unit	Population (1000s)
Iceland	1.5	Korea, Dem. People's Rep. Of	1,853.2
Portugal	2.3	Pakistan	1,946.3
Switzerland	2.4	Former Yugoslav Rep. of Macedonia	1,963.5
Luxembourg	3.4	Egypt	2,395.4
Greenland	3.7	Iran (Islamic Republic of)	2,596.8
United States of America	4.5	Yugoslavia	2,641.7
Canada	4.6	Japan	2,669.6
Spain	4.8	Korea, Republic of	2,996.6
French Guyana	7.0	Singapore	3,320.7
New Zealand	9.9	Bosnia Herzegovina	3,415.4

In some instances, the estimated national population figures which were based on population totals reported by the country's statistical institute or another official source did not closely match the country totals reported in the United Nation's *World Population Prospects*, the most widely used source for population estimates at the national level (United Nations 1999). The UN estimates often reflect adjustments of nationally reported figures to compensate for over- or under-reporting. We used the ratio of the UN estimate of national total population to the country total of our estimates to produce a second set of 1990 and 1995 figures for each administrative unit. These adjusted administrative unit figures are thus based on a uniform inflation or deflation of each estimate.

There is a fair degree of consensus with the UN estimates. Over 50 percent of the countries have national estimates that are less than 2.5 percent different, on average between 1990 and 1995, from the UN estimates. However, as Table 4 shows, some countries have larger discrepancies with the UN estimates. For the roughly 50 island nations or city-states (e.g., Macau, Singapore, Holy See)⁴ in the data base, there was no need for subnational population data as the land area of these countries is too small

⁴ These islands, city-states, and small countries are: Réunion, Mauritius, Anguilla, Antigua and Barbuda, Aruba, Bahamas, Barbados, Bermuda, British Virgin Islands, Dominica, Grenada, Martinique, Montserrat, Saint Kitts and Nevis, Saint Lucia, St. Vincent and the Grenadines, Turks and Caicos Islands, United States Virgin Islands, Falkland Islands (Malvinas), Bahrain, Kuwait, Macau, Maldives, Qatar, Singapore, Andorra, Faeroe Islands, Gibraltar, Holy See, Liechtenstein, Monaco, San Marino, American Samoa, Cook Islands, Federated States of Micronesia, Fiji, French Polynesia, Guam, Kiribati, Marshall Islands, Nauru, New Caledonia, Niue, Northern Mariana Islands, Palau, Pitcairn, Samoa, Solomon Islands, Tonga, Tuvalu, Vanuatu, Wallis and Futuna Islands. In addition, UN estimates were used for the newly formed countries of Bosnia Herzegovina and the former Yugoslav Rep. of Macedonia because no other estimates were available. Many of these small countries completed a census within the last 10 years (see <http://www.un.org/Depts/unsd/demog/cendate/index.html>).

relative to other countries in the database. Thus, DCW national boundary data and UN national-level population estimates were used.

Table 4. Countries with the greatest discrepancy between their national estimates or projections and those of the UN, 1990 and 1995

Underestimate	1990	1995	Overestimates	1990	1995
Somalia	-11.6	-20.3	Paraguay	8.8	5.8
Gabon	-7.5	-15.0	Cape Verde	5.0	9.6
Zambia	-8.0	-13.1	Chad	5.1	9.9
Comoros	-7.6	-13.4	Mali	8.4	7.0
Djibouti	-7.5	-12.4	Guinea	8.4	8.1
Eritrea	-6.7	-10.0	Guyana	11.2	11.1
Libyan Arab Jamahiriya	-2.9	-12.0	Turkmenistan	12.9	12.9
Mauritania	-2.2	-12.7	Mozambique	12.3	15.4
Guadeloupe	-6.3	-7.4	Guatemala	14.0	14.3
Angola	-8.4	-4.6	Jordan	25.3	22.7

1.C. Continental summaries of input data

North America has, on average, the highest resolution boundaries, followed by Europe, Asia and South America. The average population of an administrative unit is also lowest in North America, followed by Oceania and Europe. Average population by unit is highest for heavily populated Asia.

Table 5. Continental summaries of mean resolution and population estimates per unit

Continent	Number of administrative units	Pop 90 (UN) '000	Pop 95 (UN) '000	Mean resolution (km)	Mean population per unit '000
Africa	5,939	614,769	696,963	72	117
Asia	13,861	3,179,952	3,435,376	48	248
Europe	27,101	721,995	727,691	29	27
Oceania	1,739	26,411	28,487	70	16
North America	71,024	427,376	456,165	19	6
South America	7,441	295,085	320,551	49	43

2. GRIDGING APPROACH

The input data on administrative unit boundaries and population totals were used to produce raster grids showing the estimated number of people residing in each grid cell. In contrast to previous efforts, we did not distribute population within each administrative unit—either on the basis of proximity to large towns, infrastructure and other factors influencing population distribution (as in the Africa and Asia data sets ; see Deichmann 1996, 1998); or based on a smoothing method that assumes that grid cells close to administrative units with higher population density tend to contain more people than those close to low density units. The second option was implemented using Waldo Tobler's *smooth pycnophylactic*

interpolation in the first GPW database (Tobler et al. 1995, 1997). The new raster grids are thus similar to the unsmoothed grids of the previous version of GPW. The cell size for the new grids is 2.5 arc minutes a side or about 5 km at the equator. Figure 3 below illustrates the cell size in relation to the administrative units for the Dominican Republic. The cell outlined in blue is used below to illustrate the gridding approach in more detail, as shown in Figure 4.



Figure 3. Grid cell size in relationship to administrative boundaries, Dominican Republic

In contrast to the unsmoothed grids for the previous version of GPW, we used a different gridding approach for this update. In the first version, a standard GIS polygon-to-grid conversion function was used. This function assigned a grid cell to a specific polygon based on a simple majority rule. This has a number of disadvantages: Grid cells that contain parts of several administrative units are assigned to only one unit, and units that are smaller than a cell may be lost. We, therefore, implemented a proportional allocation of population from administrative units to grid cells in the new version.

Proportional allocation works on the assumption that the variable being modeled—in this case population—is distributed evenly over the administrative unit. Grid cells are assigned a portion of the total population for the administrative unit they fall within dependent on the proportion of the area of administrative unit that the grid cell takes up. A simple example of proportional allocation (also known as areal weighting) would be an administrative unit with a population of 5000 that is filled exactly with 100 grid cells – each grid cell would be assigned a population of 50. In the creation of the population grids, the actual implementation of areal weighting uses the administrative unit’s population density and the area of overlap between administrative unit and grid cell to calculate each unit’s contribution to the cell population total. Figure 2 and Table 6 illustrate this for a grid cell in the Dominican Republic.

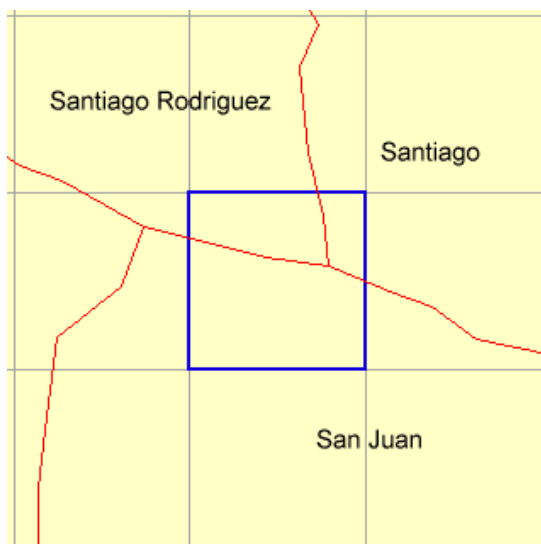


Figure 4. Detail of gridding approach for cells containing boundaries,

Table 6. Areal weighting scheme to allocation population whose boundaries cross grid cells

Administrative unit name	Administrative unit density (persons / sq km)	Area of overlap (sq km)	Population estimate for grid cell
Santiago Rodriguez	64.2	5.3	340
Santiago	246.5	2.2	542
San Juan	75.9	12.8	972
Total for cell	91.3	20.3	1854

Since larger water bodies and the presence of ice (glaciers or ice caps) can significantly distort actual population density within administrative units we used a mask (or filter) consisting of the larger lakes and ice-covered areas in the DCW. We implemented this gridding routine for each country individually and later merged the national grids to produce continental and global raster data sets of population counts (i.e., persons residing in each grid cell). Population grids, for 1990 and 1995—both unadjusted and adjusted to match the UN estimates—are available for the [global, continental, and country coverages](#) . In addition, the 2.5 arc minute grids have been aggregated to produce lower resolution grids with accurate population totals for use in applications, such as climate modeling, which require data aggregated to 0.5 or 1.0 degree grid cells.

2.A. Grid cell area

Since the grids use the latitude/longitude reference system, the actual size of a grid cell in square kilometers varies as a function of latitude, with a maximum cell size of about 21 square kilometers at the equator, a cell size of about 15 square kilometers at 45° and 5 square kilometers at 75°. We, therefore, produced a fifth grid which shows for each grid cell the total land area in that grid cell. This is actually the

grid cell's area net of water bodies (lakes, oceans, or ice-covered areas)! Dividing the grids of population counts by the area grid yields population density grids which can be used for mapping and analysis. The figure below shows population density by grid cells at 2.5 minute resolution for Haiti.

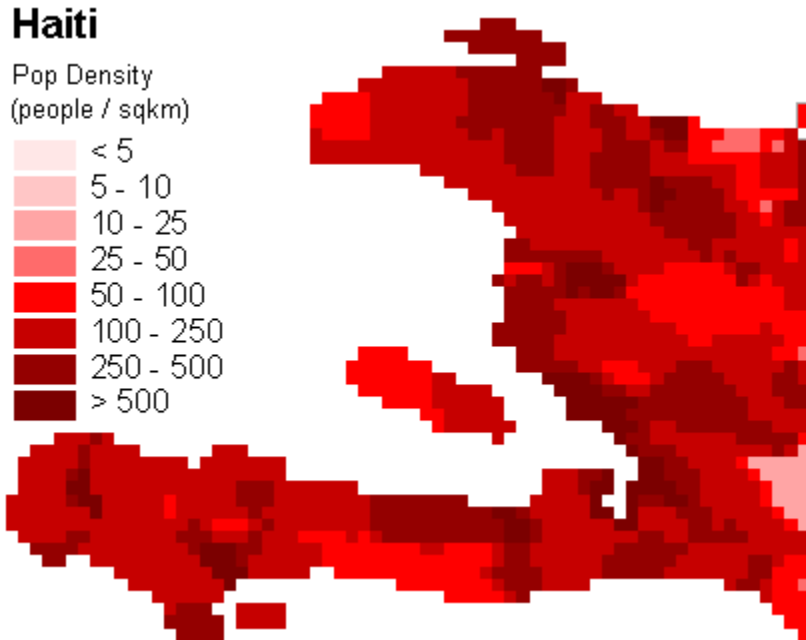


Figure 4. Gridded Population Density, 2.5 minute resolution, for part of Haiti

For grid cells in coastal areas or those bordering lakes, the cell's actual land area can be considerably smaller than that for neighboring cells that are completely on land. Cartographically, this means that grid cells of population density will be shaded completely, even if only a small portion of the cell is covered by land. Figure 5 below, for instance shows grid cells and administrative units for a small area in the north of Haiti including the *Ile de la Tortue*.

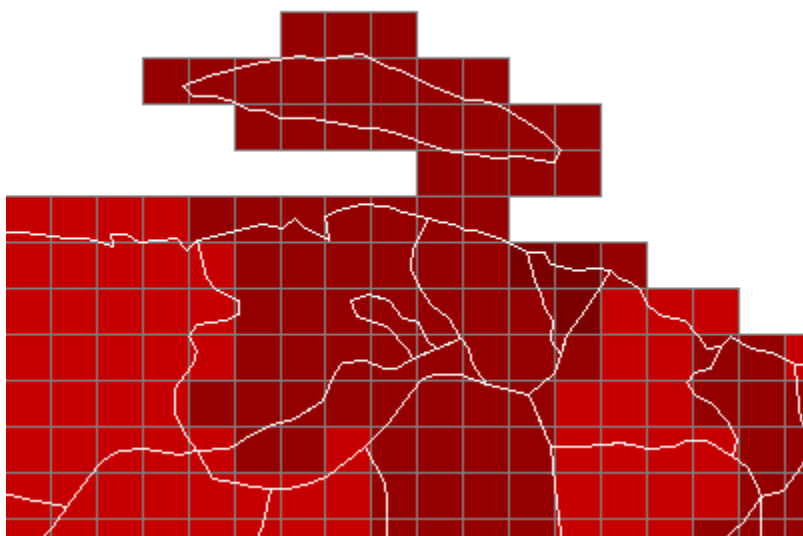


Figure 5. Example of grid cells and administrative units for a small area in northern Haiti.

As an example, the cell in the center of the top row has a land area value of only 0.97 square km. With a population density of 286.7, 278 persons are assigned to that cell. The cell immediately below with a land area of 20.14 square km and the same density contains an estimated 5774 people. This approach thus exaggerates the land area of a country in cartographic displays (however, if desired, grid cells with small land areas can be masked easily for mapping using a threshold applied to the area grid). But computations using these grids are more exact than they would be using a standard GIS provided polygon-to-grid routine in which grid cells that are located in coastal areas would be completely allocated to either land or water areas.

3. OUTPUT

3.A. Global estimates

Global population is estimated for 1990 is 5,205,349,175, and 5,653,163,865 for 1995 using our input data without gridding. The population estimates based on the gridding, for these same years, respectively, are 5,203,638,426 and 5,653,914,832 persons. This amounts to about a 0.03 percent difference in the estimates—a very close estimation by most accounts. The differences are due to rounding errors introduced by the gridding algorithm.

The impact of the improvement in the effective spatial resolution on population estimation is demonstrated in Figure 2. Small and Cohen (1999) show that the considerable increase in the number of administrative units (from 19,000 to more than 120,000) locates 75 percent of the world's population to within 50 km spatial resolution whereas the original GPW located only 37 percent to this level of detail. (The interior box in the figure shows the distribution of the effective spatial resolution for population counts in GPW v. 2 (grey bars), GPW v. 1 (red line), and the US (green line).)

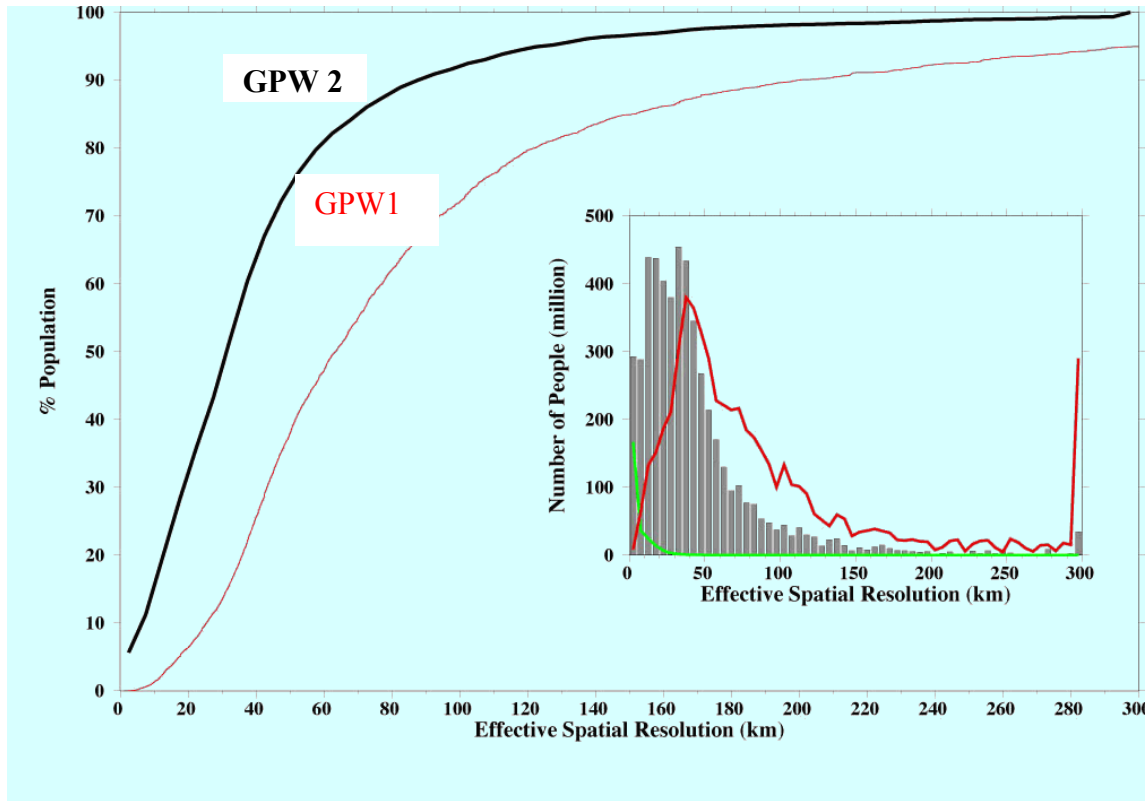


Figure 2.

Source: Small and Cohen, 1999

3.B. Products available

We have produced a number of geospatial and tabular products from this process. Spatial data available include land area, population counts and population densities by country, continent and for the world. Additional spatial data layers available only at global or continental levels include a grid of country identifiers, country boundaries and coastlines accurate to the resolution of the data set, and administrative-unit centroid locations with population and other information attached. A number of population estimates have also been calculated based on the spatial data and made available in tabular format. These include population density averages calculated by country, by proximity to coastline, by proximity to rivers, and by proximity to lakes. Population totals have been calculated by biomes, by climatic zones, by altitude ranges, by slope categories, in biodiversity 'hot spot' areas, and by ecoregions. These are all disseminated freely from the GPW [website](#).

3.C. Analytic results

Unlike conventional population data sets, GPW allows us to calculate the proximity of population to any number of the earth's physical characteristics, such as in the tabular data products described above. For example, GPW suggests that persons tend to live along low-lying coastal and riverine areas (Small and Cohen, 1999; WRI, 2000), and that these estimates are quite a bit lower than oft-cited global "estimates" which can range to as high as two-thirds of the earth's population. Nearly 150 countries (out of nearly 200) have 50% or more of their population within 200 kilometers of a coast and 50 countries (excluding island nations and city-states) have their entire population within 200 kilometers of a coast; 35 of these countries have their populations entirely situated within 100 kilometers of coasts.

Slightly more than one-third of the global population live at elevations less than 100 meters, and about 55 percent live at elevations between 100 and 1500 meters. However, Cohen and Small (1998, 1999) show that most of the world's population residing at low elevations occurs at moderate population densities rather than at high densities typical of large cities. Further, they (Cohen and Small, 1998) find that almost all the land on earth lower than 1000 meters is occupied, at a density of at least 1 person per 147 square kilometers. The average for all occupied land area is about four times as high.

As for proximity to natural hazards, about 116 million persons, in 53 countries, live within 50 km of an epicenter of a recent major earthquake (i.e., earthquakes in the year 2000 measuring 5 or higher on the Richter scale). About three times as many—368 million persons, in 63 countries—live within 100 km of an epicenter. Small and Naumann (2001) further estimate that about 9 percent (455 million people) of the world's population lived within 100 km of an historically active volcano and 12 percent within 100 km of a volcano believed to have been active during the last 10,000 years (i.e., the Holocene Epoch). They estimate that the land around the 703 volcanoes with recorded historic eruptions had a median population density of 23 people per square kilometers within 200 km; this compares to the global median density of 4.3 people for square kilometers for all occupied land area (Small and Naumann, 2001).

Global population is also distributed widely across different climatic zones. Of the six broad Köppen's climatological classifications, 40 percent of the world's population live in temperature zones and greater than 22 percent each live in dry and tropical zones. The remaining 10 and 5 percent, roughly, of global population inhabit cold and water (oceans and large lakes) zones; less than one percent live in polar zones. Among 130 predominantly non-island nations, 30 percent of nations have populations residing in only one of these broad climate zones and 58 percent have populations residing in two or three different zones.

4. SOURCES OF ERROR

Although the gridded global population estimates are extremely close to those generated from administrative units, there are potential errors nevertheless inherent in the data and the methods.

4.A. Population estimates

There are several sources of possible error in the population estimates. Loosely these fall into four categories: the accuracy of the interpolation method, the timeliness of the census estimates, the number of estimates (one or two), and the accuracy of those estimates.

The method of interpolation and extrapolation we have used assumes a constant rate of growth for the years between the intervals, an assumption which is not true especially under conditions under rapid population growth or decline. For example, in places where significant population displacement has occurred since the last enumeration (e.g., former Yugoslavia, Rwanda, Uganda), the source data used to create GPW would not reflect these movements.

In general, as Figure 6 shows, the longer the period between the last census and the reference dates, the less reliable (lighter shades) are the population estimates. It is for this reason that the recency of the population estimates is important. Our inputs range from 1967 to 1999, but most countries have at least one estimate in the 1990s, and 87 percent had a first reference in the 1980s. For the second reference year, there were also 87 entries in the 1990s. Because many countries had estimates close to one but not

both target years the left-side panel of Figure 6 is shaded darker than the right-side panel; in other words the data around 1990 are more reliable than those around 1995.⁵

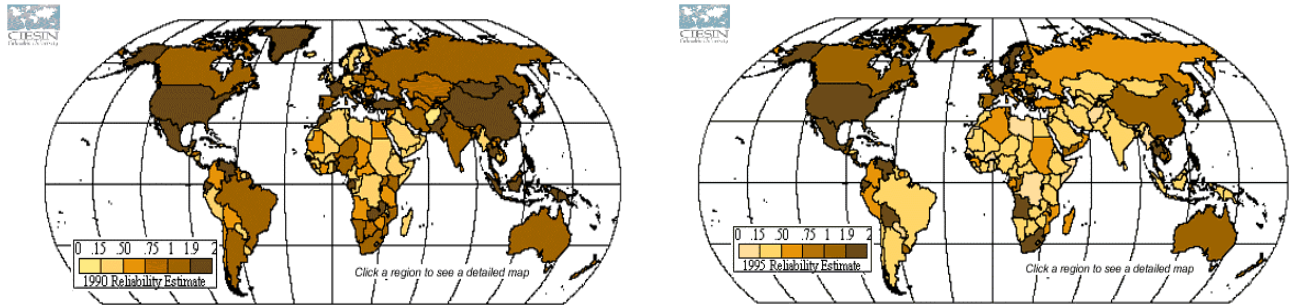


Figure 6. Temporal reliability of population data inputs, 1990 and 1995. (Robinson Projection).

As mentioned above, 38 countries—mostly islands—only had data for one year, 77 percent had two data points where at least one point was one other than our reference years, and 14 countries had data for both our reference years. For countries where data for only one reference year were available, national-level population estimates from the UN were used to calculate population change rates for the extrapolation/interpolation to the reference years. While the UN estimates are reliable at the national level, they provide no subnational information about population change. The national-level change rates derived from extrapolation were applied uniformly to subnational units, masking differences in internal population changes in the output for these countries.

Lastly, national statistical offices vary in their degree of accuracy. For example, Nigeria conducted a census in 1963 and then not again until 1991, in part because a census undertaken in 1973 was declared invalid, resulting in an estimate of 31 million fewer inhabitants than the World Bank estimated (Porter, 1992). Great efforts and expense were made to insure reliability of the 1991 census, yet much uncertainty remains concerning the accurate size of the population of Nigeria (e.g., Hollos, 1992). Among other problems, population counts indicate ethnic and geopolitical constituencies, and western

⁵ Temporal “reliability” is inversely proportional with the distances from reference years. The index scales to cover the range of values between 0 and 2. The value maximizes at 2 if one of the reference years coincide with the base year. If the reference year was far in the past (or future), the reliability of the estimate obtained through interpolation or extrapolation is low because of the assumptions built into estimating uniform change. The index of reliability is calculated, separately for each base year, 1990 and 1995, as follows:

$$R_b = \text{MIN}(2, \frac{1}{\text{MAX}(0.5, |b - Y1|)} + \frac{1}{\text{MAX}(0.5, |b - Y2|)})$$

where R_b is reliability in a given base year, b is a base year, $Y1$ is the year of the first estimate and $Y2$ is the year of the second estimate.

cultural constructs about households, thus making estimates difficult to obtain and results potentially contentious.

Although the UN estimates are generally considered reliable, the details upon which their estimates are made are not part of their public documentation (United Nations, 1999). Even though some countries have large differences between population totals reported in their national census and statistical publications and the UN estimates, it may not always be evident which one is more accurate. More often than not, however, the uncertainty involved in the interpolation or extrapolation of population data values is lower than the uncertainty in the original census figures or estimates. Users concerned about this type of uncertainty would be well served by using the UN-adjusted data, which ensures their correspondence with a widely used benchmark.

4.B. Boundaries

For a number of the countries in GPW, the accuracy of the source digital boundary data was poorly documented or absent. Information on the scale of the source maps, any generalization or thinning of the boundaries, the exact date that the boundaries represented and statements about the locational accuracy of the boundaries were often incomplete or missing. However, given that GPW is a global product with a relatively large grid cell size, greater detail in terms of the number of administrative units was deemed more important than high positional accuracy. If faced with a choice where error levels were uncertain, the more detailed data were chosen.

Similar to the population estimates, the timeliness of the boundary data can be important for accurate results. Redistricting, political change (e.g., the creation of new countries), and alterations of subnational divisions, can all change the location and number of units within the boundary data sets. Having up-to-date boundary data is important for matching population estimates correctly. As mentioned above, the poor documentation of the source data can make matching population estimates and boundaries difficult. In some cases, boundaries varied from the administrative units reported in the population data. In these cases, the population data were aggregated or distributed among the administrative units as appropriate.

The variability in the size of administrative units is evident from examination of maps produced from GPW. Variability of unit size between and within country data is present globally in the data, but most notably in North Africa, Northern Asia, South America, Australia and Canada. Often, different levels of administrative units used to create GPW result in visible differences between adjacent countries. These discrepancies can be eliminated in future updates with higher resolution data.⁶

Typically, going to the next lower administrative level increases the number of units by at least an order of magnitude, with corresponding increases in average resolution. For example, the U.S. data used in the first version of GPW, counties, were replaced by higher resolution census tracts in version 2. This increased the number of units by a factor of almost 20 (from 3,140 counties to 60,911 census tracts), while the average resolution decreased from 55 to 13 km. Variability in the size of the administrative units decreases as the administrative level increases (from a variance of 98,149,276 and standard deviation of 9,907 for counties to a variance of 2,189,485 and a standard deviation of 1,480 for census tracts; both sets of statistics are based on the area of the units in square kilometers).

⁶ In some geographic large countries whose population is very unevenly distributed (e.g., Canada, Australia) these large blocky areas are not likely to change with better boundary data in the absence of major demographic influx.

The high variability in the size of both levels of administrative unit (county and census tract) is due to how the units were defined. Land units (and the counties made of these units) in the U.S. and Canada were defined by a combination of unsystematic and systematic subdivisions (Campbell, 1991). Unsystematic partitioning such as metes and bounds was a common methodology in Eastern North America, while systematic partitioning was common in the Western portions of North America. The form and area of counties in Canada and the U.S.—generally, a mosaic pattern in the east and a rectilinear pattern in the west—reflects the influence of these partitioning systems on the formation of counties. Census tracts are an aggregation of blocks, the smallest unit distributed by the census, which are defined by the distribution of households. Census tracts are designed to be relatively stable units with homogeneous population and economic characteristics and a population of between 1,500 and 8,000 people upon creation (U.S. Census Bureau, 2000). Even though blocks (and census tracts) are forced to nest within counties and states, they are not heavily influenced by political factors in their formation. So we can see that in the U.S., the variability in size of administrative units at the level of census tract is mostly due to variation in the distribution of population.

More generally, this relationship between size and population distribution depends on how each country defines its units; it does not hold true for all units or levels, or in all countries. However, for small and medium sized countries by the second administrative level, large administrative units (relative to others within the country) typically represent low-density populations while small units represent higher densities; for larger countries (e.g., the US or Canada), the third administrative level is needed before this holds true. This is confirmed by overall trends in GPW; areas of low spatial resolution in GPW are typically areas of low population density.

4. C. Positional accuracy

Except in places of very dense administrative boundaries (e.g., urban centers like Manhattan, Brisbane, or Paris), grid cell values are only representative of a portion of the administrative units within which they fall. The count contained within any one cell does not necessarily represent actual population at that location. Instead, the count within a single cell represents the portion of population that would be present in that cell if the persons residing in an administrative unit were spread evenly over the whole administrative unit.

The variable nature of the positional accuracy of the input boundary data prohibits a universal statement about the positional accuracy of GPW. However, where information was available for the source data, the majority of the input data were of similar or higher resolution than the DCW (1:1 million) to which national boundaries were matched. GPW is appropriate for use at small scale (global and large regions). Integration of GPW with local data at medium and large scales (scales larger than 1:1 million) may be subject to considerable uncertainty.

5. NEXT STEPS

We are in process of updating GPW for estimates of population in the year 2000, although no estimated release date has been set in part because most countries have not yet released their censuses for circa 2000. Additionally, when new estimates for 1990 or 1995 become available, the 1990 and 1995 coverages are also being updated.

In addition to these updates, we are creating two new related sets of georeferenced population data. The first database is analogously similar to the current enterprise: a higher-resolution (1 kilometer) grid of population, housing, income and land cover in the United States. Additional demographic attributes (e.g., age distributions) may also be gridded. This arises out of the interest to have more

complex, and information from which causal or consequential behaviors may be inferred in relationship to gridded physical data (e.g., household characteristics and emissions). The second database, currently being compiled, consists of tabular and spatial information on the location, extent and population of human settlements. This database aims to include small settlements, which are notably overlooked by the United Nations and other agencies which tend to focus on cities with populations greater than 100,000 persons. When combined with GPW, we anticipate being able to generate a complete urban-rural surface of the world's population. Both of these databases are currently being created, with expected releases sometime in 2002.

Appendix. Short outline of the chronological methodology used to generate GPW version 2.

For each country or area, the following steps were carried out:

1. Obtain digital administrative boundaries and population data.
2. Estimate 1990 and 1995 population by administrative district (P90 and P95) and link to the digital administrative map.
3. Create alternative population estimates for 1990 and 1995 (P90A and P95A) by adjusting each figure uniformly so that the national total matches the UN World Population Prospects estimate.
4. Overlay a digital map of lakes and ice fields and set the population estimate for the lake and ice areas to zero.
5. Compute population densities as pop/km², now net of lake areas, for each administrative unit.
6. Create a regular grid in vector GIS format (“fishnet”) with a resolution of 2.5 arc minutes and overlay with the administrative units boundaries.
7. Calculate the area in km² for each resulting polygon of overlap. Multiply this area with the corresponding administrative unit’s population density to get a population estimate for each polygon of overlap.
8. For each of the four population estimates and the land area in km², aggregate all polygons of overlap that belong to a given grid cell. Then, link these grid cell totals back to the original regular grid (“fishnet”).
9. Convert this result to five raster GIS data sets: one each for population in 1990 and 1995—adjusted and unadjusted, and one for land area.
10. Create continental and global grids by adding the individual country or area grids together.

REFERENCES

- Campbell, J. (1991), Locational and Land-Partitioning Systems (p168-181). In: Campbell, J. (1991). *Map Use and Analysis*. Oxford: Wm. C. Brown Publishers.
- Clarke, J. (1995), Population and the Environment: complex interrelationships. *Population and the Environment: The Linacre Lectures 1993-94*. Bryan Cartledge, ed. Oxford: Oxford University Press.
- Cohen, J. and Small, C. (1998), Hypsographic demography: The global distribution of population with altitude, *Proceedings of the National Academy of Sciences*, November.
- Deichmann, U. and L. Eklundh (1991), Global digital datasets for land degradation studies: A GIS approach, United Nations Environment Programme, Global Resource Information Database, Case Study No. 4, Nairobi, Kenya.
- Deichmann, U. (1996a), A review of spatial population database design and modeling, paper prepared for the *UNEP/CGIAR Initiative on the Use of GIS in Agricultural Research*, National Center for Geographic Information and Analysis, Santa Barbara; www.ncgia.ucsb.edu/Publications/Tech_Reports/96/96-3.PDF .
- Deichmann, U. (1996b), Asia medium resolution population database documentation, Database documentation and digital database prepared in collaboration with *UNEP/GRID Geneva for the UNEP/CGIAR Initiative on Use of GIS in Agricultural Research*, National Center for Geographic Information and Analysis, University of California, Santa Barbara; www.grid.unep.ch and grid2.cr.usgs.gov .
- Deichmann, U. (1998), Africa medium resolution population database documentation, Database documentation and digital database prepared in collaboration with *UNEP/GRID Sioux Falls and World Resources Institute*, National Center for Geographic Information and Analysis, University of California, Santa Barbara; grid2.cr.usgs.gov .
- Digital Chart of the World. Available at: <http://www.maproom.psu.edu/dcw/>
- Dobson, J. E., E. A. Bright, P. R. Coleman, R.C. Durfee, B. A. Worley (2000), LandScan: A Global Population Database for Estimating Populations at Risk, *Photogrammetric Engineering and Remote Sensing*. 66(7): 849-857.
- Hollos, Marida (1992), Why is it difficult to take a census in Nigeria? The problem of indigenous conceptions of households. *Historical Methods*, 25 (1): 12-19.
- Honeycutt, D. and J. Wojcik (1990), Development of a population density surface for the conterminous United States, *Proceedings GIS/LIS*, Anaheim, Vol. 1, 484-496.
- Porter, G. (1992), The Nigerian Census Surprise, *Geography* 77 (4): 371-374.
- Small, C. and J.E. Cohen (1999), Continental Physiography, Climate and the Global Distribution of Human Population, Proceedings of the *International Symposium on Digital Earth*, Beijing China, 11/1999, p.965-971. www.LDEO.Columbia.edu/~small/population.html
- Small, C. and T. Naumann (2001), The Global Distribution of Human Population and Recent Volcanism, available at: <http://www.ldeo.columbia.edu/~small/PopVol.html>

- Sweitzer, J. and Langaas, S. (1995), Modelling population density in the Baltic Sea States using the Digital Chart of the World and other small scale data sets. In Gudelis, V. Povilanskas, R. and Roepstorff, A. (eds.). Coastal Conservation and Management in the Baltic Region. Proceedings of the EUCC -WWF Conference, 2-8 May 1994, Riga - Klaipeda - Kaliningrad., pages 257-267. www.grida.no/baltic/techrep/eeuc.pdf .
- Veldhuizen, J. van, R. van de Velde and J. van Woerden (1995), Population mapping to support environmental monitoring: some experiences at the European scale, *Proceedings, Sixth European Conference on Geographical Information Systems*, EGIS Foundation, Utrecht.
- Tobler, W., U. Deichmann, J. Gottsegen and K. Maloy (1995), *The global demography project*, [Technical Report 95-6](#), National Center for Geographic Information and Analysis, Santa Barbara.
- Tobler, Waldo, Uwe Deichmann, Jon Gottsegen and Kelly Maloy. 1997. "World Population in a Grid of Spherical Quadrilaterals," *International Journal of Population Geography*, 3:203-225.
- United Nations, 1999. *World Population Prospects: The 1998 Revision. Volume 1: Comprehensive Tables*. NY: United Nations.
- U.S. Census Bureau, 2000. *TIGER/Line® Files: Redistricting Census 2000 TIGER/Line® Files Technical Documentation*. Washington: U.S. Census Bureau.