Documentation for

Total Population

for countries and territories

Gapminder Documentation constitutes work in stepwise progress.
We welcome all sorts of comments, corrections and suggestions through e-mail to the author.

Gapminder Documentation 003
Version 2
Uploaded: 2011-12-12
Published by: The Gapminder Foundation,
Sweden, Stockholm, 2008

Author: Mattias Lindgren
E-mail: mattias.lindgren (at) Gapminder.org
Gapminder is a non-profit foundation
promoting sustainable global development
by increased use and understanding of statistics.
www.gapminder.org

1. Introduction

This documentation is for the Gapminder compilation "Total Population" and "Population with projections". Total population is, by default, used for the size of the bubbles in Gapminder World: www.gapminder.org/world

An excel-sheet accompanies this documentation. It contains the data as well as the detailed meta-data for each observation. This excel-sheet can be found at:

http://www.gapminder.org/downloads/documentation/gd003

Our goal is to have a complete data set, from 1800 to 2011, for "all" countries and territories. In previous versions (version 1) we had the ambition to collect as much as possible for each observation. The purpose was to do a quality grading for each observation. However, we have dropped this ambition, so the quality grading is no longer included in the excel-sheet. [1]

The main purpose of the data is to produce graphical presentations in the Gapminder World Graph. Since population is normally used for the size of the bubbles in Gapminder World, a missing observation means that no bubble is shown, irrespectively of whichever other indicator we have in the graph. Therefore, we have included very rough estimates for countries and territories for which reliable data was not available, and we have also, for a few historical observations, simply "guessed".

Furthermore, we have not been able to make sure that every single observation is based on the best estimates available. Hence we discourage the use of this data set for statistical analysis and advise those who require more exact data to investigate the available data more carefully and look for additional sources, when appropriate.

For a discussion on what countries and territories we try to cover, and how we try to handle border changes and the like, see the document "Countries and Territories in Gapminder World". However, it should be noted that the inclusion or exclusion of any area in this data set does not, in any way, imply a stated opinion of Gapminder as to the legal status of the area.

It is not always clear to what extent certain semi-autonomous or disputed territories are also included in the data for their "mother country", so there may be some "double counting" in these instances. Thus, it is not feasible to simply summarize all the countries and territories in this data set to get the total population of the world since the "double counting" would overestimate the world population.

2. What we have done

We started by utilizing existing compilations of international data, where we gave precedent according to how well documented they were (on an observation-by-observation basis) and how good their coverage was. We then turned directly to national sources. Next we turned

[1] The indicator for "data quality for population" is still available in Gapminder World (under "for advanced users"). However, the quality grades refer to the first version of teh data set.

to “undocumented” sources, such as Wikipedia, the CIA fact book and journalistic accounts. After that we used various extrapolation techniques, and finally after that, we guessed.

In the first version we tried, for each observation, to include as much meta-data as we could, on an observation-by-observation basis. Most observations have travelled quite a long way from the original primary data, through a long chain of citations and manipulations, where the manipulations by Gapminder, if any, are only the last ones in this chain.

We have tried to track these chains as far back as we could. This has meant that we had to look up the sources used by our sources, and then the “third layer” of sources used by those “second layer sources” etc. This work has only started. So far, in only a few cases, have we managed to find the primary source. Many international compilations do not give meta-data or sources on an observation-by-observation basis, only the general principles they’ve used.

All the meta-data we found, including the “chain of citation” and the various manipulations done by them along the way, are documented in the excel sheet.

3. Sources used by Gapminder

The sources below are roughly listed in the order of priority we used, i.e. we only turned to the next source in the list when there was no data in any of the previous ones. However, since this dataset has grown in an iterative process, it is not guaranteed that the order of priority has been followed to the letter. Rather, the guiding principle has been a combination of convenience (i.e. coverage), on the one hand, and quality and transparency, on the other. The full references to each source is found in the excel-sheet.

3.1 World Population Prospects 2010 revision

World Population prospect (WPP) was the main source for observations after 1950. It also included projections going up to 2100. In some cases the WPP displayed discrepancies with other sources that we wanted to link to. In most cases we tried to make any adjustment in the other sources, and keep the WPP intact. In a few cases (documented in the excel-sheet) we made some adjustment to handle implausible discrepancies (e.g. we chose another year than 1950 to make the linking).

In few cases we choose, for pragmatically reasons, to use the 2006 revision of the WPP (the revision that was used in the version 1 of the gapminder dataset). This is documented as a separate source in the excel-sheet.

3.2 The dataset of Angus Maddison

The monumental work of Angus Maddison was the starting point for the observations before 1950. His dataset has the best historical coverage (with some series going very far back in history), is available digitally and has a very good written documentation for individual

observations. Even though most of the data were available digitally from his homepage, there were also some additional data in his written documentations (e.g. Maddison 2001).

To avoid making our dataset too big, we have not included his earliest observations (e.g. those from the first millennium), but that might be included in future updates.

3.3 Mitchell "International historical statistics"

To fill the gaps in the data from the above source we used Mitchell (1998 a & 1998 b).

Mitchell has, in three publications, compiled historical data for most countries of the world. Mitchell provides two sets of tables for the total population of countries: one which only includes observations based on censuses or similar and one that also includes estimates. We used the first set of tables. We were able to find footnotes on an observation-by-observation basis.

3.4 UN statistics division

To fill the gaps in the data from the above sources, we used data from the UN statistical division. They have yearly data for 1970-2006. We have been unable to find meta-data on an observation-by-observation basis for this compilation.

3.5 U.S. Census Bureau, International Data Base.

To fill the gaps in the data from the above sources, we used data available from the International Data Base at the U.S. Census Bureau. We were unable to find meta-data on an observation-by-observation basis for this compilation.

3.6 "The population of Oceania in the second millennium" by Caldwell et al (2001)

To fill the gaps in the data for the Pacific from the above sources, we used an article of Caldwell et al (2001). We were able to find meta-data on an observation-by-observation basis in this article.

3.7 National sources

To fill the gaps in the data from the above sources, we turned to more direct sources on a country-by-country basis, e.g. the statistical bureaus of individual countries. In many cases this meant that we went directly to a "primary source", e.g. census reports and the like.

3.8 "Undocumented" ad-hoc sources

To fill the gaps in the data from the above sources we turned to a variety of more undocumented sources such as Wikipedia, the CIA fact book and journalistic accounts. We have in general been unable to assess the quality of the data in these sources.

This does not, in itself, imply a low precision, but rather that we are not sure about what the quality is (see the discussion on the "meta-meta-data" below). We have tried to cross-check some of the observations in some of these series with other sources, whenever possible.

3.9 Gap-filling using the sources indirectly

The sources described so far were the only ones used by us. However, there were still some gaps. For these gaps we used various indirect data, based on the sources above, estimated with various interpolation and extrapolation techniques (e.g. what we have labeled geographical interpolation or extrapolation, temporal interpolation etc). See section 6, "Modifications", for more details.

3.10 Guesstimates

When all of resources listed above had been exhausted there still remained a few missing observations to get the full data-set we wanted (especially the desire to have a data-set starting in 1800). For these observations we simply made an arbitrary guess, i.e. setting the observation equal to the earliest observation available. These observations were solely included to enable display of all countries in the graph; they do not add any information what-so-ever.

4. Type of primary data

The "sources" listed above are simply where *we* took our data, i.e. normally one of the major compilations of international data. All that data originate from one primary source or another, and we have occasionally tried to track down what that primary data is. Most primary sources can be grouped into four categories, which are described below, with the sources with the supposedly best quality first.

4.1 Census or equivalent

These include censuses, administrative enumerations or registration records executed with an elaborate and documented methodology according to modern standards.

4.2 Informal census

By informal censuses we mean an actual counting of people, or at least something very similar (e.g. households, taxpayers), but a counting that was not done according to modern standards. This includes pre-modern censuses, where there were some important deviations from modern standards of censuses. This can also include more informal impressions on the number of people living in an area made by contemporary eye-witnesses, but where the conditions were such that the observation could be considered somewhat reliable (e.g. the observation concerns a very small area such as a small island).

4.3 Indirect estimate

Indirect estimates do not entail any actual counting of people. Rather, they are based on other indirect information, combined with a more extensive set of assumptions. This could, for example, be based on tax records, the size of armies, archeological evidence of settlement patterns and size of cities combined with assumptions on how all this could relate to the size of the population. Typically, a number of different kinds of information are combined.

4.4 Arbitrary guess

When no information was available, however indirect, we arbitrarily set the population as being the same as the earliest observation we had. This is only done to get a full data set for graphical display against other indicators and is only done for the first relevant year (since the Gapminder graph interpolates all other years).

5. Modifications

In many cases the primary data has been modified in one way or another to get the final figure. We have tried to classify the modifications done to the primary data into a limited number of categories.

5.1 Summations of parts

Sometimes we have data for all the constituent parts of a territory, e.g. we have data for Guernsey and Jersey that constitutes "the Channel Islands". Then it is simply a matter of adding up these observations to get the new one.

5.2 Larger area minus non-included parts

Sometimes we have the total population for a larger area, as well as the population of some of the constituent parts, but we lack data for the other constituent parts. For example, we might have data for "Serbia and Montenegro" and for "Serbia", but no data for "Montenegro". Then we could simply calculate the population of Montenegro as "Serbia and Montenegro" minus "Serbia".

Ideally this should not pose any problems, just as for "summation of parts". However, inconsistencies between the observations can pose a much more serious problem than for "summations of parts", especially if the area we are looking for is a very small part of the larger area we do have data for. For example, the population of Timor Leste is less than 1% of Indonesia, which is actually smaller than the deviations between some of the sources for Indonesia. Hence, if the population of Timor Leste were estimated as "Indonesia including

Timor Leste" minus "Indonesia excluding Timor Leste" (it is not), and the two population figures came from different sources, then we might, in the worst case, even end up with a negative population for Timor Leste.

5.3 Geographical interpolation

With "geographical interpolation" we assume that the country's population has had the same population growth as a larger area of which it is a part. This could be used when we have data for former countries that have now split up, while at the same time, we have data for the new countries, but only for a few years. Some of the sources (e.g. Maddison & Caldwell) supply more aggregated data, e.g. total population for a whole region or for a group of countries.

If we assume that the share of a specific country in the larger area was the same in earlier years then we can use the regional total to estimate the population of the country in the earlier years for which only regional data is available. As an example:

$$\begin{aligned} Population\ of\ Slovenia\ \ (1910) = \\ = Population\ of\ Yugoslavia\ (1910) \\ \times \frac{Population\ of\ Slovenia\ (1950)}{Population\ of\ Yugoslavia\ (1950)} \end{aligned}$$

In the above case we consider the population of Yugoslavia (1910) as being the "source observation".[2] Hence, the information about "Type of data" and the like would, in this case, refer to the population of Yugoslavia in 1910.

The term "interpolation" (which we use in a rather new way) refers to the fact that our country of interest is a constituent part of the larger area we are using. Therefore, we at least know that the population of our country is less than the population of the larger area (assuming that the population of the larger area is correct, of course). This means that geographical interpolation is a somewhat better method than the geographical extrapolation discussed below.

5.4 Geographical extrapolation

When doing a "geographical extrapolation" we assume that the country's population has had the same population growth as a neighboring country. In principle we mean the same thing as with "geographical interpolation", the only difference being that our country is not a part of the area we used for our estimation. As an example:

[2] Since that is an observation for the same year as our observation of interest.

$$Population\ of\ Estonia\ \ (1820) =$$

$$= Population\ of\ Lithuania\ (1820) \times \frac{Population\ of Estonia\ (1860)}{Population\ of\ Lithuania\ (1860)}$$

In the above case we consider the population of Lithuania (1820) as being the "source observation"[3]. Hence, the information about "Type of data" and the like would, in this case, refer to the population of Lithuania in 1820.

The term "extrapolation" refers to the fact that we are using data from an area that is not overlapping our country of interest. This means that even if the population data for our "source country" is absolutely accurate, we do not have a "guaranteed maximum" population. This would only be a significant problem if one of the countries has gone through some major migration movements or wars, and the other has not. We tried to minimize this risk by doing some quick review of the country's history. One case in point is 19th century southern Africa where there has been a great deal of population movement; for the time being, this is something for which we are occurred for which we are lacking even tentative information.

5.5 Temporal interpolation

"Temporal interpolation" is the term we use for what is normally meant by interpolation, i.e. drawing a straight line between two points. This can be done in a variety of ways, e.g. either assuming a constant growth rate between two years with data, or simply assuming that the population changed with a fixed absolute number.

Both observations used for the interpolation are considered as the "source data" in this case.

We have not done any temporal-interpolations, with some very few exceptions, since the graphing software does temporal interpolations automatically. However, some of the data conveyors have done temporal interpolations. Furthermore, in the Excel file called "Population data & working notes (xlsx)" we have included observations for all years, with the missing years filled with interpolated values using constant growth.

5.6 Temporal extrapolation

By "temporal extrapolation" we mean all other methods of extending a series outside its time range. This can be done in a number of ways; the roughest is to just assume that the population is the same as the closest observation (similar to our "arbitrary guess"). A slightly more sophisticated method is to extend the growth rate of some adjacent period backward

[3] Since that is an observation for the same year as our observation of interest.

or forward. There are also more sophisticated modeled projections, based on various assumptions of fertility and the like.

Generally speaking, this manipulation is rougher than temporal interpolation since we only have either a starting or ending year, while in temporal interpolation we have both.

In principle, Gapminder has not done any temporal extrapolations. The few exceptions are cases for some very short time spans, when we were not sure about what year the data were referring to. Furthermore, “arbitrary guess” could, as noted, be considered as being a very rough form of extrapolation.

However, the other data conveyors have occasionally done temporal extrapolations. We have not always managed to find information on exactly what observations are based on this method, and exactly how that was done. Observations for future years can of course be assumed to be based on this method.

5.7 Adjustments for under-enumeration

Sometimes the data fail to cover certain groups, e.g. nomads, or have been judged, for various reasons, that a census had less than ideal coverage. In those cases efforts have been made to remedy these shortcomings by adjusting the figures. Sometimes, when specific groups have been missed (such as our friends, the nomads) attempts have been made to estimate the size of these groups, e.g. by doing some kind of interpolation or utilizing other data. Other times more ad-hoc adjustments have been made.

Note that when the excluded groups are geographically based, we consider any adjustments to be “recalculations to fit present borders”, which is described below.

5.8 Recalculated to fit present borders

As can be noted in the document “Countries & Territories in Gapminder World” we try to make the historical data refer to the present borders of a territory, even when there have been substantial border changes in the past. Hence, there is often a need to do some recalculations, e.g. to subtract the population in areas that are no longer part of the country, and to add population for areas that were not part of the country in the past.

Ideally we have the data for the relevant sub-regions (that constitute the difference from the present borders). In such cases it is a simple matter of adding or subtracting the sub-regions in question. This should ideally not have any major impact on the uncertainty range rating for the resulting observations.

However, in most cases we do not have direct data for some of the relevant areas, so we have to get the needed data in more indirect ways by utilizing some of the of the methods described above, e.g. geographical interpolation.