Modern population censuses and progress in counting the world population

Abstract

BACKGROUND

Population censuses have been a fundamental source of demographic data since the first modern enumeration was conducted in Iceland in 1703. Yet, the information on where and when censuses were carried out, and what population count did they return is very scattered, especially for the pre-1945 era.

OBJECTIVE

This paper documents a database listing all modern population censuses since 1703 for all territories (independent or not) in a standardized and readily-reusable format, indicating the population count, census type, date, territorial coverage and other relevant information on the enumeration procedure or accuracy. It further explores how the census coverage developed over time and what is the uncertainty in population returns.

METHODS

The database draws from a variety of United Nations publications and a wide array of national and international resources. An uncertainty analysis of census enumerations was done using economic data and bivariate copulas.

RESULTS

More than 3,200 population censuses has been carried out so far. The coverage has been steadily increasing, reaching 93% of estimated world population. It was estimated that census underenumeration has declined globally from around 5% in the 1950s to 3% in the 2010s. In total, censuses lag behind the actual population of the world by nine years (at present demographic growth), which is underestimated by UN by about one year of population growth.

CONTRIBUTION

The study provides a database that allows studying spatial and temporal developments in enumerating the world population, and innovatively quantifies the uncertainty in the world population estimates.

1. Introduction

Population censuses are a fundamental source of statistical data. Indeed, one of the United Nations' *Demographic Yearbook* explains that "the uses of census data are too numerous and varied to be listed in their entirety" (United Nations 1955). As a complete count of a population and a recording of its demographic characteristics, it underlies research and administrative procedures (e.g. appropriation of voting districts and funding to local governments). Further, the data from censuses is necessary to ensure sample surveys can properly estimate the values of the phenomena in question for the whole population.

The first "modern" census was carried out in Iceland in 1703, returning a population of 50,358 (McEvedy and Jones 1978, Statistics Iceland 2018). The "modern" moniker signifies the difference between censuses being simple recordings of the number of people, or registration made for administrative purposes, and which were made already in antiquity, with the a proper census recording demographic data of each individual separately, pertaining to a particular moment in time and for the complete population of a territory. Information on when and where censuses where carried out is important for researchers. Presently, United Nations (UN) is a major source of information, and have been gathering the data from all countries in the *Demographic Yearbook* since 1948. However, there are some limitations of this source:

- Online databases only provide information on censuses at present-day political divisions, and only for the period after 1945;
- Information for censuses carried out between 1850 and 1945 are only available in the scanned yearbooks, and usually not for countries which didn't exist at the time of publication.
- No data for pre-1850 censuses were included.

Having a full historical overview of census-taking would therefore require consulting national sources of many countries. An interesting research question related to this is what percentage of the world population have been actually enumerated, presently and in the past? It was partially answered in one of the UN yearbooks, which found that only 17% of the

estimated world population was counted during the 1860 census round[1], but as much as 78% in the 1950 round (United Nations 1962). The incomplete coverage of the world results in uncertainty over the size of the global population, both past and present. This includes errors created by the inaccuracies of census counts, which are typically underenumerated to an extent that varies significantly between countries.

The objectives of this paper are threefold. First, to document a database listing all modern population censuses in a standardized and readily-reusable format. Second, analyse the body of censuses in terms of their basic characteristics (population count, date, type) and contrast the population figures from censuses with modern estimates to track the historical progress in counting the world population. Finally, estimate the extent to which population censuses underenumerate the world population and compare those results with UN estimates. The study is based on data available as of 1 August 2018.

2. Materials and methods

2.1. Building the database of census enumerations

The work presented in this paper was done in two parts. In the first phase, a database of polities, i.e. societies with an organized government, was compiled. This was necessary as censuses are organized by governments, either of sovereign states or dependent territories, and therefore pertain to a certain polity in its political organization and territorial extent in the census year. As many polities have been divided, merged or reshaped over time, it was essential to track their evolution. This enables identification of all censuses made by defunct polities, or situations where the census population total was no longer valid due to change in the territorial composition of a polity. A search from available datasets of present or historical polities revealed that none is comprehensive enough for this analysis, mostly because they focus on sovereign states.

A new list of polities was constructed merging multiple sources, using as a starting point a list by Russett et al. (1968), amended with Correlates of War (2014, 2016), Gleditsch (2013) and Marshall et al. (2017). Correlates of War (2014) dataset was further used to identify changes in the territorial composition of polities. The list was compiled back to 1800 due to availability of sources and the very small number of censuses made before that year. It also

[1] A census "round" pertains to censuses carried out in a period of 10 years centered around a reference year, e.g. 1860 round includes censuses conducted during 1855–1864.

generally omits pre-colonial polities of non-European territories, as none of those is known to have carried out a census. A cross-reference table with four alternate lists of polities was created, so that its easy to connect the database of censuses with other coding systems used in social sciences. The total number of polities in the list used for this study is 406, including 129 defunct as of 2017.

In the second phase, the database of census enumerations was constructed, on country-by-country basis. Various editions of the United Nations' *Demographic Yearbook* were a fundamental source together with other UN datasets and publications (United Nations 2018). Additionally, a large number of country-specific sources was consulted including statistical yearbooks and websites, as well as publications by the US Census Bureau. Counting multiple editions of the same national statistical yearbook as a single source, the total number of sources included in the database is 73. In some cases, national and UN sources (and UN publications between themselves) disagreed on the census date, type or population figure. The national source was generally preferred, as were more recent UN publications over older ones. The database used information available up to 1 August 2018.

Inclusion of censuses in the database was subject to certain limitations. In principle all should satisfy the principles of modern censuses as laid out by United Nations (1955): universality (all members of community are covered), simultaneity (all facts refer to one point in time), individual units (of recording data), defined territory (of the census operation) and compilation (including publication of results). In practice, whether a given satisfies all conditions might be difficult to assess. As a consequence, some censuses do not have precise dates, which might imply that collection of data was not simultaneous; for others the spatial extent is not precisely known. They were nonetheless included if they were covered by UN yearbooks. The rule of universality was more strictly was enforced, therefore the database excludes:

- Colonial censuses where the non-European population was not enumerated, or only summary statistics were collected on the non-European population. It is unclear, however, in case of some censuses whether the non-European population was actually enumerated in the field; those censuses were generally included in the database. Also, all censuses of Australia and New Zealand are included despite omitting the indigenous population from most enumerations.

- Sample censuses, except situations where the census was carried out through a combination of sample field surveys and use of administrative registers. The French "rolling census", which is in effect a large annual sample survey, is also included.

The censuses were assumed to be population and housing censuses, unless it was known that they were only population censuses. Housing or agricultural censuses were excluded from the database, except in situations where such census enumerated the population as well.

2.2. Contents of the database

The contents of the database is presented in Table 1. First four fields identify the polity's name and its numeric or alphabetic code according to three different coding systems. This allows for quick cross-referencing, though for complete details the user should consult the full cross-reference table provided with the dataset. In the next field, the principal component of the database in recorded: the census date and enumerated population. In a few cases, the census was reported to had been carried out, but no population figure was found. Such censuses were kept in the database as population data might be traced in the future, but excluded from the analysis. In a few cases of islands becoming uninhabited since their last census (South Georgia and some US territories in the Pacific), a "dummy" census was added with population 0 at a date when the census would have been conducted (Falkland Islands and US censuses, respectively). Such "censuses" were not included when analyzing the summary statistics of censuses, though were needed for computing the total enumerated population in the world.

The dataset also records the area over which the census was carried out, albeit this information was not always available. Another field contains the type of census, which could be *de jure* (people habitually resident in an area) or *de facto* (people actually present at the time the census). For some censuses both *de jure* and *de facto* population figures were available. In general, only one figure is shown according to the most frequent type of census in the body of censuses in a given country, but both are shown when the census type was changed at some point, allowing for better interpretation of historical time series. All countries have specific rules who falls inside the scope of the census, related e.g. to armed forces, diplomats, seamen or refugees who are either nationals outside, or foreigners inside, a country. Major population groups included or excluded from the census (with estimated number, if available) were identified in 'census_note' field, usually concerning exclusion of indigenous, tribal, nomadic, refugee or foreign population.

The territory of the census is described in another field, identifying border changes of polities compared with the present situation. If case the territory changed between censuses, data was presented under both territorial compositions, as long as data was available. In principle, this can only work if a polity loses some of its territory. Addition of a territory does not modify the census population unless the added territory had a census as well. An entry will therefore show the combined population, noting if the other census was carried out on a different date (which would usually be the case). For instance, to account for the unification of Germany, two entries were created for the 1987 census: one for its original spatial extent (West Germany), valid for 1987–1989; and one with the population of the East German census of 1981 added to the total, valid for 1990–2010 (i.e. until the first post-unification census of 2011). For some countries, some regions would be enumerated on different dates, therefore combined entries for particular periods of time were created. For example, the 1931 census in the United Kingdom (UK) did not include Northern Ireland, which was only enumerated in 1937. Therefore, the entry for the UK 1931 census shows the population enumerated in the UK sans Northern Ireland combined with the enumerated population of Northern Ireland from the previous census (1926). Then, the entry for the UK 1937 census added Norther Ireland population of 1937 with the 1931 population for the remainder of the UK (valid until a "unified" census of 1951).

The "validity" of a census entry mentioned in the previous paragraph refers to the period of time during which the population number indicated by a census is valid given the actual territorial composition of a polity. For example, due to the cession of Alsace-Lorraine in 1871 to Germany, the French 1866 census has two entries, one valid for 1866–1870 (pertaining to the French 1861–1871 territory and indicating 38,067,064 persons) and one valid for 1871 (pertaining to the French 1871–1918 territory and indicating 36,470,866 persons). The next census of 1872 (36,102,921 persons) is then valid for 1872–1875 as no territorial had occurred until the next census of 1876 (36,905,788 persons), which itself is valid for 1876–1880 until another census and so on (INSEE 2018, United Nations 1955). The years of validity always pertain to the situation in the end of the year (31 December).

Finally, the database contains a field with additional descriptions relevant for analyzing the results, a list of sources and a special "dummy" variable 'census_count'. The dummy variable identifies "independent" enumerations, i.e. a single census carried out in a specific polity. If multiple entries pertain to the same census, 'census_count' will have a value of "1" only for one entry, which is the territorial composition at the time of the census. For example,

the entry for the Soviet 1989 census appears 16 times as it was first valid for the USSR during 1989–1990 and then was valid its 15 successor states until they carried their own post-independence censuses. Therefore, only the entry for "Russia" under 'census_territory' of "USSR, 1945-1991 territory" valid for 1989–1990 has 'census_count' of "1", and all other have 'census_count' of "0".

Table 1. Variables in the database of census enumerations.

Variable	**Description**	**Mandatory field**
Country_code	Numeric code used for uniquely distinguishing all polities in the database	yes
Country_RSS_code	Numeric code based on Russett et al. (1968), updated with Correlates of War project files "State System Membership (v2016)" and "Territorial Change (v5)"	no
Country_ISO_code	Three-letter code from the International Organization for Standardization's (2017) standard ISO 3166, including discontinued codes for historical polities	no
Country_name_EN	Name of the polity. A single, most recent name is used throughout, hence it is not necessarily valid at each census year	yes
Census_year	Year of the census	yes
Census_month	Month of the census. If the month is unknown, "0" value is assigned	yes
Census_day	Day of the census. If the day is unknown, "0" value is assigned	yes
Census_population	Population enumerated during the census. The figure does not include adjustment for underenumeration unless only an adjusted figure was available (such cases are indicated in the 'Census_note' field)	no
Census_area	Area (in km^2) of the census. In a few cases (indicated in the 'Census_note' field) this value is not equal to the total area of the polity. This figure is generally based on the 2016, 1990, 1970 and 1955 UN yearbooks, and national sources, and often changes due to remeasurements, rather than actual change in size	no
Census_type	Type of the census: • DJ – de jure. • DF – de facto. • RDJ – de jure, register-based. • RDF – de facto, register-based.	no
Census_territory	Information on the territorial coverage of the census in general, excluding special cases for undercoverage of certain parts of the polity	yes

	(which is indicated in the 'Census_note' field). Recurring explanations include: • Current – the territory during the census was the same as the current (as of 2018) territory of the polity. • Territory of the census – refers to the territory within borders of the polity as they were during the census year. It is used in cases where the exact description of the territory is not available or not feasible, or the territory was changing very frequently. • Pre-WW1 territory – refers to the territory as it was before the start of World War I in 1914. • Interwar territory – refers to the territory as it was as the result of redrawing of borders after the end of World War I in 1918, but before the start of World War II in 1939. • Post-WW2 – refers to the territory as it was as the result of redrawing of borders after the end of World War II in 1945.	
Census_note	Other relevant information on the census, particularly on: • the dates of field enumeration, if it was particularly long, thus affecting the accuracy of the population total; • estimates of census underenumeration or overenumeration; • which major groups of population, or parts of the polity, were excluded from the census; • whether population of other polities was included in the population figure; • relevant major changes in the territory since the last census not described in 'Census_territory'; • major deviations from the de facto/de jure concepts; • different census dates for parts of the polity; • whether the population figure is provisional, rounded, adjusted or based on incomplete processing of census returns; • whether the census was only for population (no housing enumeration) or a housing/agricultural census with a population count; • whether the census used other sources of information apart from field enumeration (except censuses indicated as register-based); • methodological differences affecting the census area figure; • population of certain enumerated population groups or parts of the polity, if such information is helpful for interpreting the time series.	no
Census_valid_start	First year of validity of the census population figure for calculating total enumerated population in the world. Value of "0" is assigned in the	yes

	particular census is not used	
Census_valid_end	Last year of validity of the census population figure for calculating total enumerated population in the world. Value of "0" is assigned in the particular census is not used	yes
Census_count	Dummy variable for the purpose of counting the total number of censuses taken, with value of "1" identifying entries that should be counted (value of "0" is assigned otherwise)	yes
Source	List of sources of data used to generate the entry	yes

2.3. Summarizing the census enumerations and uncertainty analysis

After the databases of censuses and polities were prepared, summary statistics of the body of censuses were computed. The task was carried out with a Matlab script (included in the online dataset). Total enumerated population per year was calculated for the whole world and per six continents: Europe (including all former USSR states), North America (including Hawaii), South America, Africa, Asia (including all of present-day Turkey and Indonesia but without the Asian parts of the former USSR) and Oceania (without Hawaii). For comparison, population estimates were collected per continent from two sources, HYDE 3.2 decennial estimates for 1700–1940 (Klein Goldewijk et al. 2017) and World Population Prospects: the 2017 Revision annual estimates for 1950–2017 (United Nations 2017a).

The uncertainty analysis was carried out using estimates of census errors (under- or overenumeration) collected while building the database of censuses. A total of 186 estimates from 83 different polities were gathered, spanning from 1769 to 2012. Typically, such estimates are based on post-census surveys, and sometimes include estimates for population groups excluded from the census (tribal, nomadic, refugee etc.). This information can be used to undertake a global uncertainty analysis of world population estimates from the aforementioned World Population Prospects. As one can expect that countries that are more developed economically will be able to organize a more accurate census, a statistical model was built combining estimates of census errors with gross domestic product (GDP) per capita.

The model utilizes a bivariate copula, which is a joint distribution on the unit hypercube with uniform [0,1] margins. Copulas allow to model the dependency between variables probabilistically using many possible parametric distributions. They have many applications (Morales Nápoles et al. 2017) and are described comprehensively by Joe (2014). The dependency between GDP per capita and census error was computed with a Gaussian copula,

which fitted the data better than alternate models such as Gumbel, Clayton and Frank copulas. The cumulative distribution function of the Gaussian copula C is as follows:

$$C(u,v) = \Phi_\rho\left(\Phi^{-1}(u), \Phi^{-1}(v)\right), (u,v) \in [0,1]^2 \quad (1)$$

where Φ is the bivariate Gaussian cumulative distribution and ρ is the (conditional) product moment correlation between the two marginal probability distributions u and v in the interval [0,1]. The goodness-of-fit was analysed with the Cramèr–von Mises test statistic M discussed by Genest et al. (2009), which is computed by comparing the parametric and empirical copulas:

$$M_n(\boldsymbol{u}) = n \sum_{|\boldsymbol{u}|} \left\{C_{\widehat{\theta}_n}(\boldsymbol{u}) - B(\boldsymbol{u})\right\}^2, \boldsymbol{u} \in [0,1]^2 \quad (2)$$

where $B(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(U_i \leq \boldsymbol{u})$ is the empirical copula, $C_{\widehat{\theta}_n}(\boldsymbol{u})$ is a parametric copula with parameter $\hat{\theta}_n$ estimated from the sample, and n is the sample length. It should be noted that the marginal distribution of census errors in this analysis is the absolute value of the relative error (unadjusted count divided by the adjusted count), so without the $\pm$ sign. One estimate from the sample was excluded from the analysis, i.e. the error of the 1969 Bhutanese census, which is estimated to have been overenumerated by an astonishing 258%[2]. It is a large outlier, therefore it was omitted from the analysis, even if it was found to have only minor influence on the results.

After the statistical model was prepared, a database of population estimates and GDP per capita for all countries for 1950-2017 was compiled. The World Population Prospects: the 2017 Revision provided the necessary demographic data, while GDP per capita in constant 2011 dollars and purchasing power parity was mostly taken from the Maddison Project Database (Bolt et al. 2018), with gaps amended using data from Maddison (2010), World Economic Outlook (IMF 2018) and National Accounts Main Aggregates Database (United Nations 2017b). Information on GDP per capita was missing for some small countries and dependent territories, but altogether the polities with missing information constituted less than 0.1% of world population. Those territories are mostly highly-developed economically, hence it was assumed that there is no uncertainty in their population counts. The influence of missing uncertainty estimates on the global results would be negligible. Finally, the census database was reworked so that it represented present-day political divisions and borders (for

[2] The census enumerated 1,034,774 persons (United Nations 1992), whereas the World Population Prospects: The 2017 Revision (United Nations 2017a) put the actual number at around 290,000.

all censuses valid for 1950 and after), so that it would match the dataset of population estimates and GDP.

With all data in place, the copula model was sampled 10,000 times for each polity and year to provide 10,000 estimates of possible census errors, which were then applied to correct the census return for a given polity and year (with a defined probability that the population was underenumerated and not overenumerated). By summing up the corrected estimates for all countries in a given year, 10,000 possible corrected world population estimates were obtained, thus providing possible uncertainty ranges of world population. Further, the UN estimates of population corresponding to census years were also summed up and compared with the results. Of course, this exercise uses a number of assumptions:

- The census errors depends only on GDP per capita (while many factors could be important);
- There is no uncertainty in GDP per capita estimates (which in reality can be very large);
- The census errors are independent of each other (which is almost certainly true);
- The census returns are not adjusted for underenumeration (which is not always the case).

Additionally, adjusting to present-day borders was not possible for all years for some countries, hence the world total in section 3.3. (mainly during the 1950s and 1960s) is lower than reported in section 3.2. Also, 1950 population estimates and GDP per capita was used for censuses carried out before 1950, but valid in some period starting in 1950. This affects the quality of the uncertainty estimate during the 1950s, but has little influence later on, as only countries constituting altogether less than 1% of world population did not replace pre-1950 censuses with newer enumerations by 1960.

3. Results

The Population Enumerations Database is freely accessible online (Paprotny 2018). The repository contains the main dataset of censuses (in different formats) combined with additional files and Matlab code that were used to carry out the analysis described hereafter.

3.1. Main characteristics of censuses

The database contains 3,872 entries, of which 3,225 are unique censuses of sovereign states and dependent territories with a published population count. They were carried out in 334 different polities, with the highest number (42) recorded in France since 1801. Yet, this number is partly due to the annual publication of results from the "rolling census" conducted continuously since 2006. Second-highest number of enumerations belongs to Germany if one includes the pre-unification censuses of Prussia. Luxembourg, New Zealand, Sweden, Maldives and some French overseas dependencies also have had at least 30 enumerations. China has had the largest census by number of people counted (almost 1.34 billion in 2010), but India (including the larger pre-independence India) had so far counted more persons than any other polity: 7.5 billion out of 41.8 billion counted in all censuses in the world. On the other end of scale, the 2000 census in the US dependency of Wake Island counted only one person falling within the scope of the census. Majority of censuses counted less than a million persons, with one-third counting less than 100,000 (Fig. 1a).

The number of censuses has been steadily growing since the 18th century (Fig. 2). An exceptionally large number of enumerations from the 1820s to the 1860s is a consequence of intense (usually 3-yearly) census-taking by 30-odd countries in present-day Germany, which ended with the unification of Germany in 1871. Most of the censuses (59%) were carried out on *de facto* basis, though for some of those counts a *de jure* figure is also available. The same is true in a historical perspective, though the census-taking of pre-unification German states was on done on *de jure* basis, warping the figures for 1820s–1860s. Also, there has been slight increase in the share of *de jure* enumerations in the past few decades.

Countries have various preferences in the time of the year to carry out a census (Fig. 1b). In 10% of cases, the month of the census was not identified, or was not defined. December has been most popular in total, though again mainly due to pre-unification German censuses. Otherwise, April has been most common month for censuses, as both the United States and United Kingdom prefer this date, and so it is also applied in many present and former dependencies of those states. Few censuses are carried out in the summer as the movement of population would disturb figures especially under *de facto* method. Recently, the rise of register-based censuses and the French "rolling census" increased the number of enumerations dated 1 January or 31 December.

Figure 1. Number of censuses (a) by population enumerated and (b) by month.

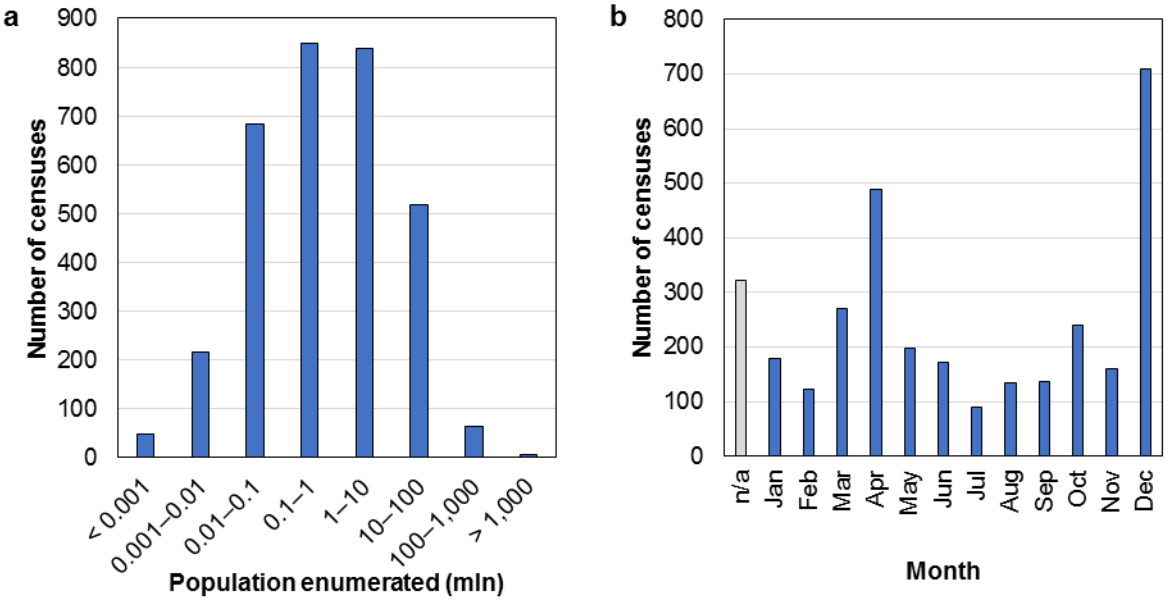
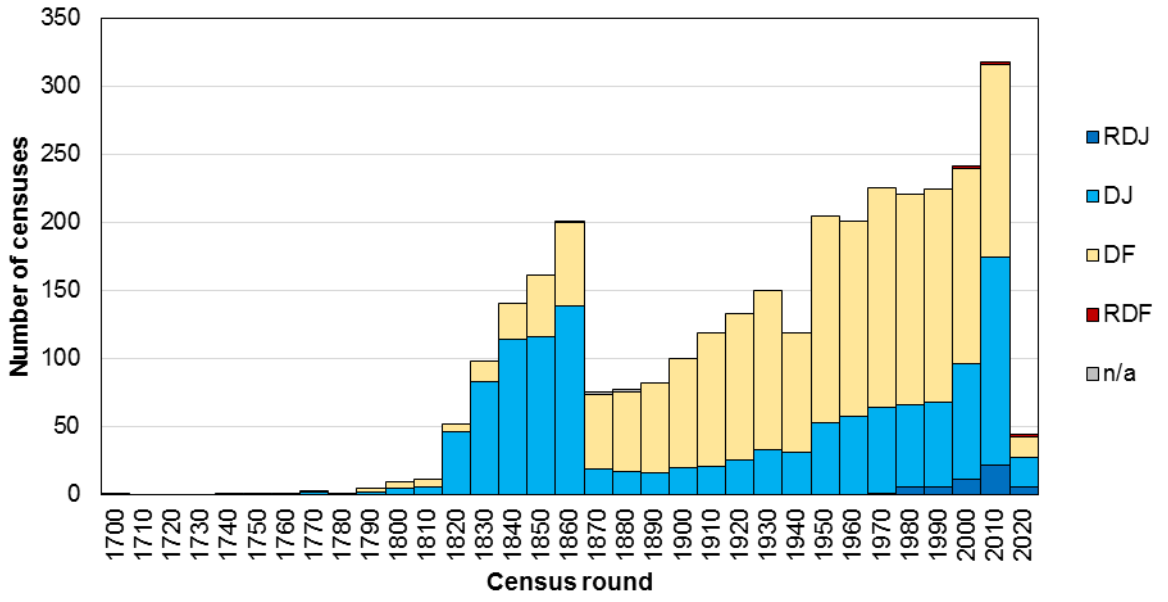


Figure 2. Number of censuses by type and census round.



Note: Years refer to the middle year of a census round, e.g. 2000 refers to 1995=2004, 2010 to 2005–2014 etc.

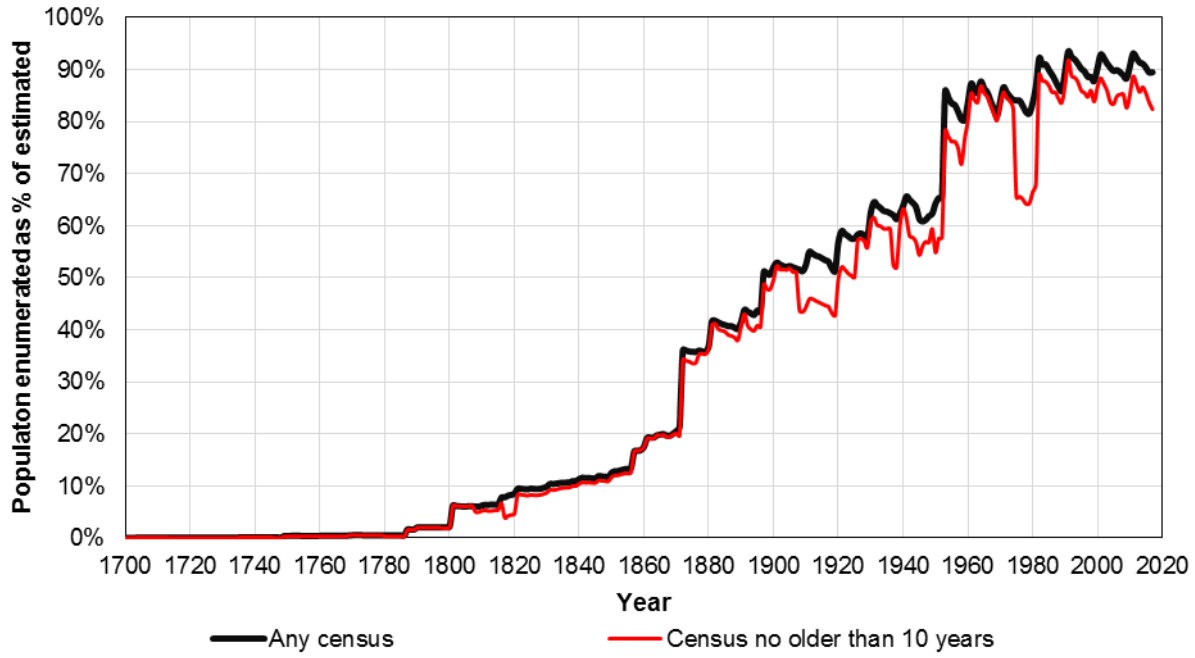
3.2. Progress in census-taking and enumerating the world population

The first modern census, carried out in 1703 in Iceland, enumerated only 50,358 persons out of a global population of around 600 million. Only when larger countries begun enumerations, visible progress was made in counting the global population (Fig. 3). The total population count exceeded 5% of estimated population in 1801, when France and United Kingdom did their first censuses (Table 2). 25% threshold was reached with landmark enumerations of India (under British colonial rule) and Brazil in 1872. Half of the global population have become counted with the Russian census of 1897, and three-quarters when China made its first tally in 1953. Highest-ever percentage was reached during the 1990 census round (93.5% in 1991), with 2000 and 2010 rounds only marginally less successful (maximum of 92.9% and 93.1%, respectively). Due to the effort by the United Nations to concentrate census-taking around years ending with 0 or 1, a certain "cyclicality" is noticeable in recent decades (Fig. 3). Yet, the availability of up-to-date censuses, i.e. no older than 10 years, has been declining somewhat recently, reaching only 82.4% of world estimated population in 2017, the worst result since 1981. However, the 2017 figure is preliminary as not all countries which carried out censuses recently have yet released the results of their counts.

At continental level, there are vast differences (Figs. 4 and 5). Europe led the world in the 18th century, but since the US census of 1790 North America had better enumeration progress, and the two have been largely on par since around 1900. South America also achieved good results in the late 19th century, but the frequency of censuses had been rather low, resulting in sometimes poor coverage of up-to-date censuses (in the late 1930s below 10% of estimated population). South America was followed closely by Oceania, where thanks to frequent census-taking in Australia and New Zealand the coverage of censuses no older than 10 years has been very high in recent decades. Asia had much less success most of the time, due to the difficulty of counting the large population such of China and the Indian subcontinent. But, since this has been overcome, coverage of Asian censuses is similar to other continents except Africa. This continent has had the highest rate of population growth for several decades (almost 3% per year), hence censuses quickly become outdated. Half of African population has only become counted with the 1952 Nigerian census, and as of 2017 the enumeration progress stands at 74%. Yet, the delayed follow-up to 2006 Nigerian and Egyptian censuses reduced the coverage of up-to-date enumeration to 50% of estimated population.

Spatially, almost all independent states and all inhabited dependencies had a census at some point in time. Half of the world's independent states had already done an enumeration by 1834 (Fig. 6), and more than 80% by 1923. Presently, Lebanon is the only country that never have had a census[3]. However, Eritrea, South Sudan and Uzbekistan had made no enumeration since independence – they are covered by censuses of the countries they were previously part of: Ethiopia, Sudan and the Soviet Union, respectively (Fig. 7). Still, at national level the availability of up-to-date censuses has been increasing steadily, reaching a peak of 92% in 2011. It can be added that some territories now publish census results annually: France with some its dependent territories ("rolling census" since 2006), Guernsey ("rolling electronic census" since 2014) and Switzerland (a register-based census supplemented by annual sample surveys since 2010).

Figure 3. Population enumerated worldwide as a % of estimated population.



[3] Though 1921 and 1932 censuses are being mentioned (Maktabi 1999), they seem to have been a process of administrative registration of the population in context of citizenship acquisition without recording even basic demographic characteristics. This is especially likely given that the results of the 1932 census were published immediately after the supposed count was made, whereas processing the census returns always consumes many months. UN yearbooks make no mention of those 'censuses', and while some League of Nations yearbooks (starting with League of Nations 1927) do mention them, the reference was dropped in the later editions. Also, McEvedy and Jones (1978) write: "Lebanon has never had a census at all: attempts to hold one in 1921, 1932 and 1941 all foundered".

Figure 4. Population enumerated by continent as a % of estimated population.

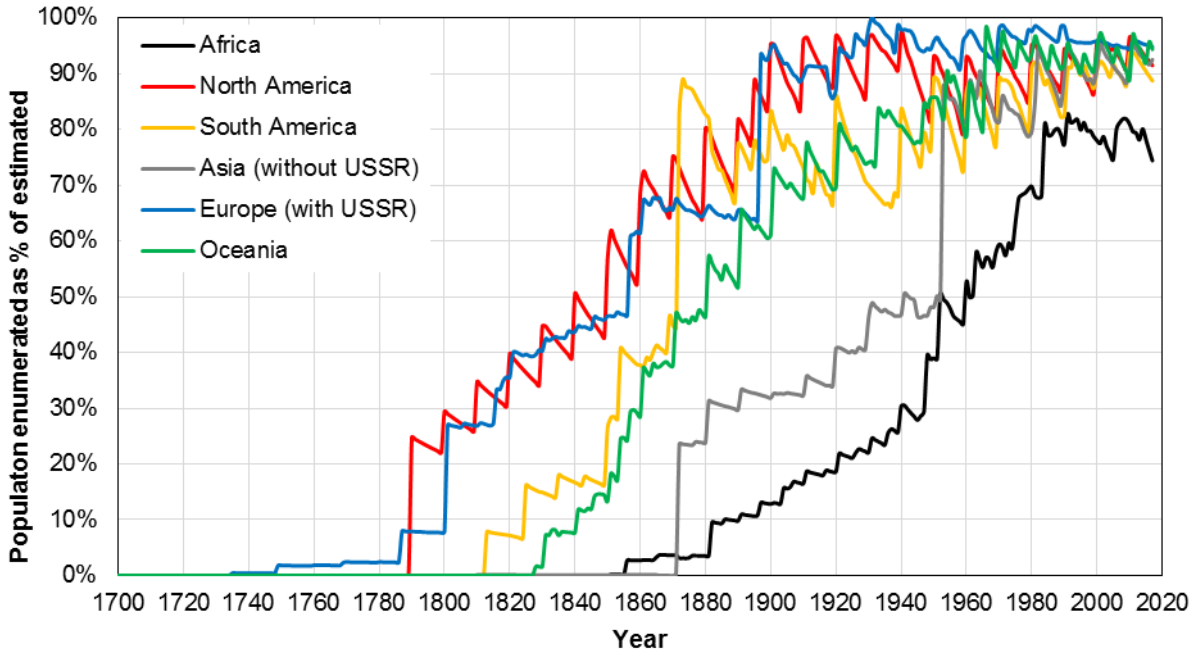


Figure 5. Population enumerated by continent as a % of estimated population, including only censuses no older than 10 years.

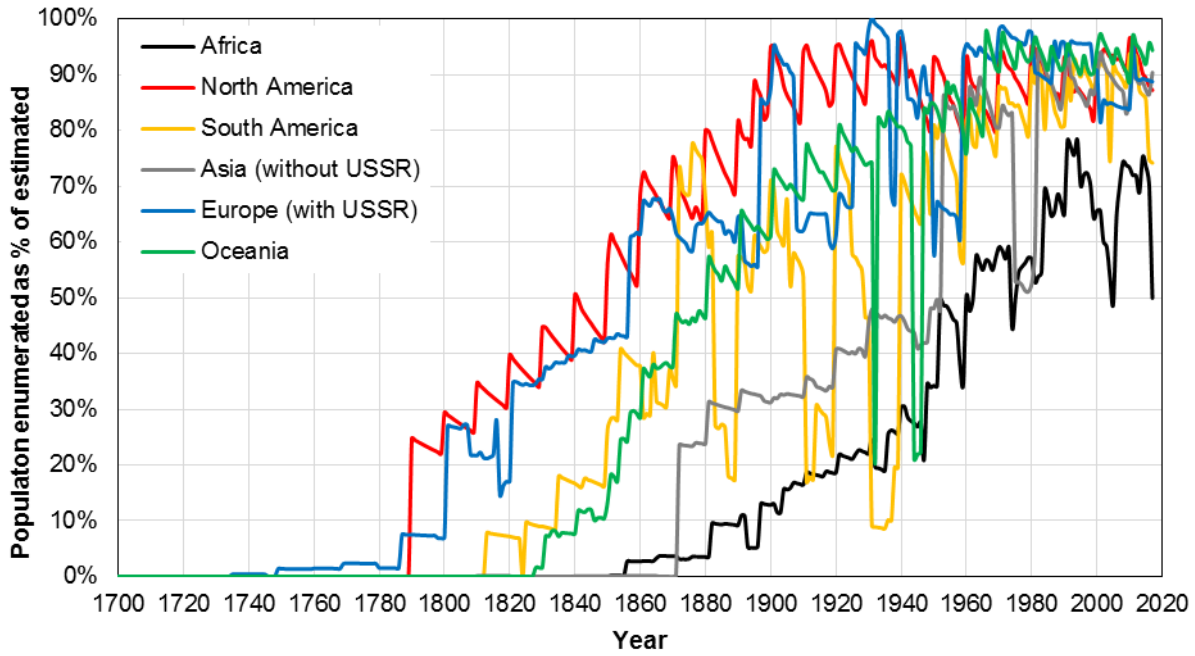


Figure 6. Sovereign states by availability of census population figures, 1815–2017.

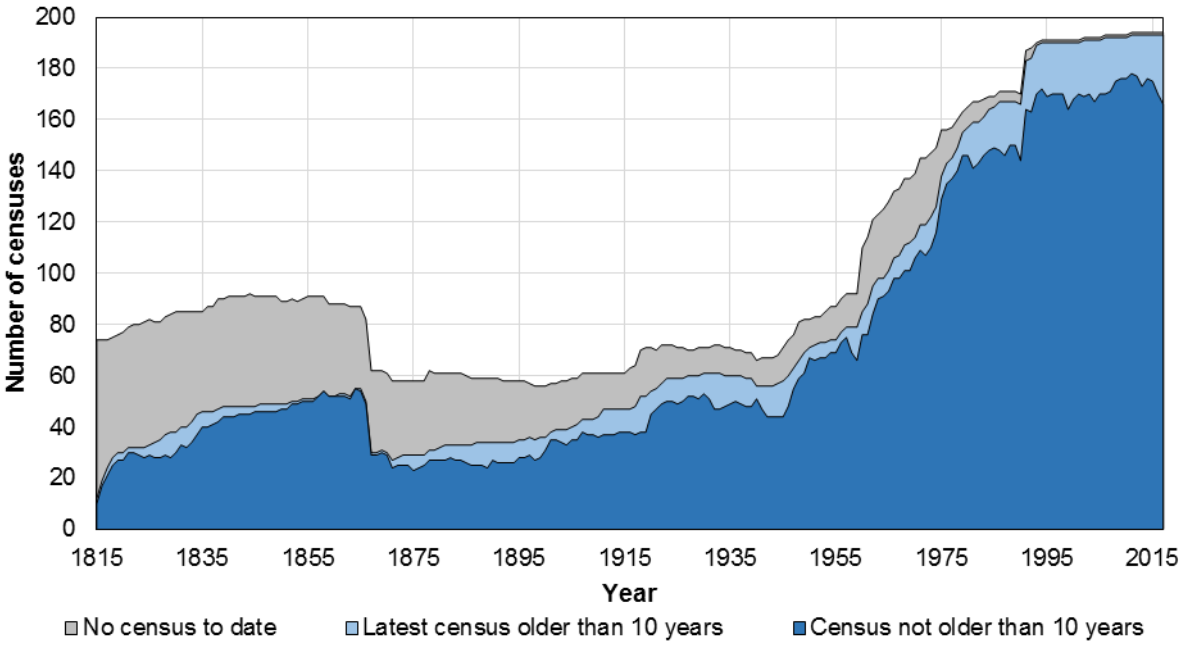


Figure 7. Present-day polities by date of first census covering at least part of the territory in question.

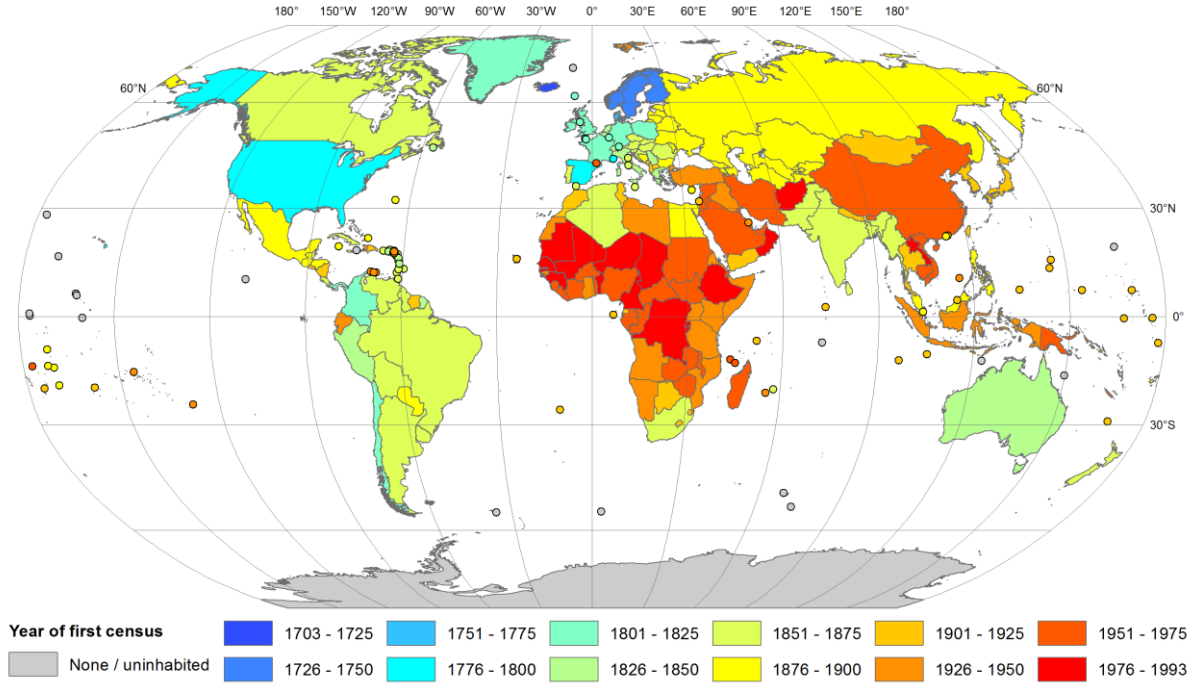


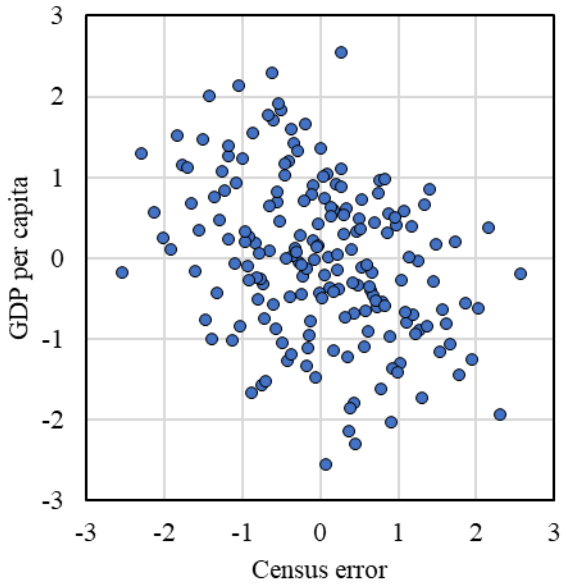
Table 2. Year of first achieving 5, 25, 50, 75 and 95% coverage of estimated population with a census, and population enumerated as of 2017 as % of estimated.

Continent	5%	25%	50%	75%	95%	2017
Africa	1882	1936	1952	1984	-	74%
North America	1790	1800	1840	1870	1900	91%
South America	1813	1850	1872	1872	-	89%
Asia (without USSR)	1872	1881	1941	1953	1982	93%
Europe (with USSR)	1787	1801	1857	1897	1901	95%
Oceania	1831	1857	1881	1911	1966	94%
World	**1801**	**1872**	**1897**	**1953**	**-**	**89%**

3.3. Uncertainty of census population figures

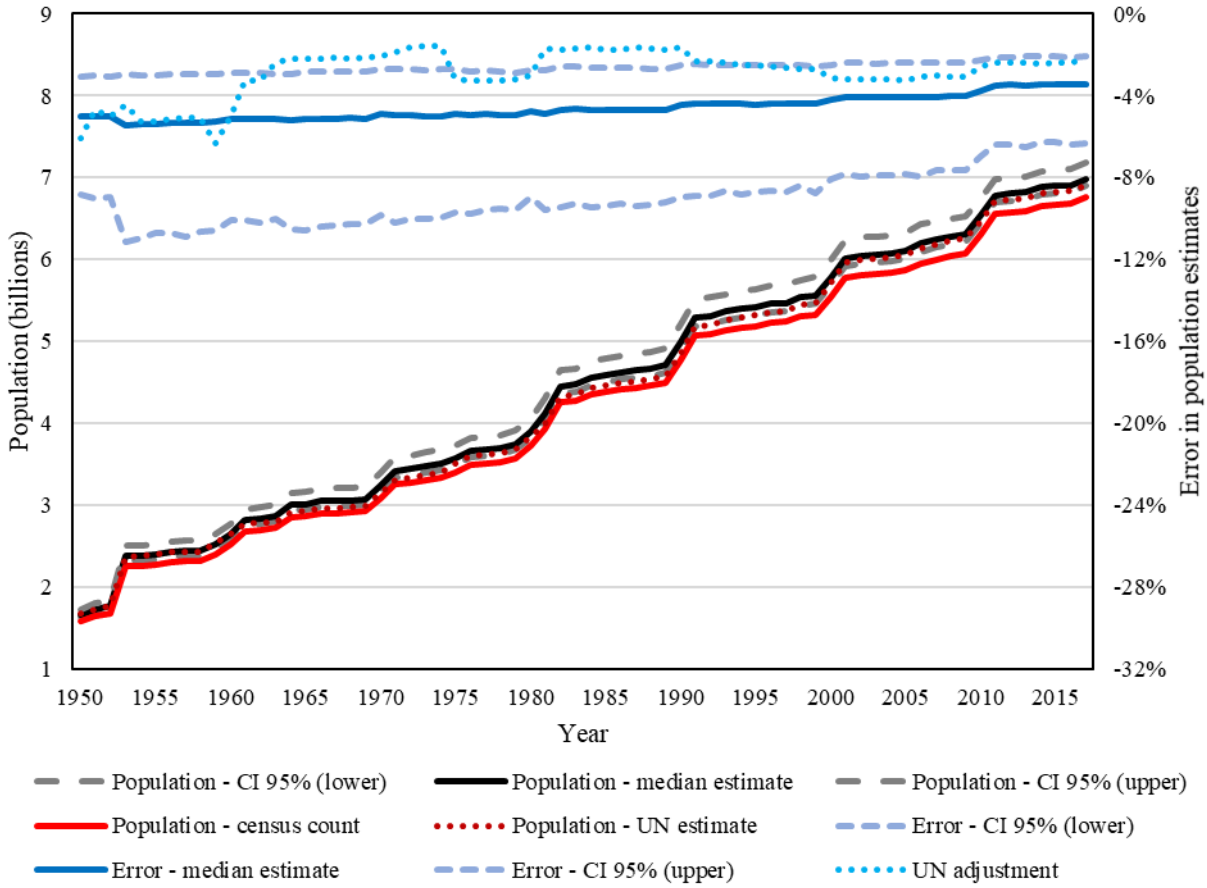
A total of 185 estimates of census errors were collected in this study, excluding the extreme overenumeration of the 1969 Bhutanese census. In 181 cases (97.8%), the censuses were underenumerated, with the largest errors of around 26% identified for the 1918 Mongolian and 1981 Namibian censuses, while the 1979 Afghan census missed an estimated 23.1% of the population. On the other end of scale, the 1940 Guatemalan census overenumerated the population by almost 44%, followed by the 1920 Brazilian census, which counted an extra 11.4% of the population. Overall, there has been a negative correlation between the magnitude of the error and GDP per capita (Fig. 8), with Spearman's rank correlation given as -0.35. This in line with the expectation that higher economic development is correlated with smaller errors in counting the population.

Figure 8. Distribution of census error and GDP per capita transformed to standard normal distribution.



Overall, the underestimation of world population in censuses has increased, but only in absolute terms, from around 120 million in 1953 to more than 230 million in 2017. This is partially due to improvement in census coverage. The median estimate of the relative error is -5.4% for 1953 and -3.4% for 2017 (Fig. 9). The upper bound of the 95% confidence interval is around three times higher than the lower bound throughout the time series. Similarly, the compared to annual demographic growth, the error is estimated to have been rather stable throughout, and equivalent of almost three years' worth of world population growth. It was closer to two years during the 1970s and 1980s. Given that the lag between the actual world population and the aggregated census count was about seven years on average in recent decades, underenumeration during censuses explains a third of the difference. Adjustment of census figures in UN's estimates is remarkably close to the median estimate for the 1950s, but then oscillates around the lower bound of the 95% confidence interval. In effect, the UN's adjustment should have been higher by a half, or approximately 90 million people. This corresponds to around one year of world population growth (Table 3).

Figure 9. World census returns, 1950–2017 with and without adjustment for census under- and overenumeration.



Note: the adjustment is shown as estimated from this study and as computed by the UN. The estimates are shown for the median and 95% confidence interval (CI).

Table 3. Lag between census counts and estimated world population, in years of demographic growth, averaged by census round.

Census round	Lag resulting from…			Total lag
	frequency of censuses	under-enumeration (accounted for by UN)	further under-enumeration	
1960	7.36	1.97	0.42	9.76
1970	7.24	0.82	1.30	9.36
1980	6.80	1.24	1.10	9.14
1990	4.97	1.08	1.37	7.42

2000	5.54	2.03	0.92	8.49
2010	5.80	2.06	0.77	8.62

4. Discussion

The database described herein is an attempt to gather information on population censuses in a reusable format, drawing from 70 years of efforts by the United Nations in collecting demographic data. Still, more work is needed to validate and enhance the information contained in the database. In particular, the following aspects should be addressed in the future:

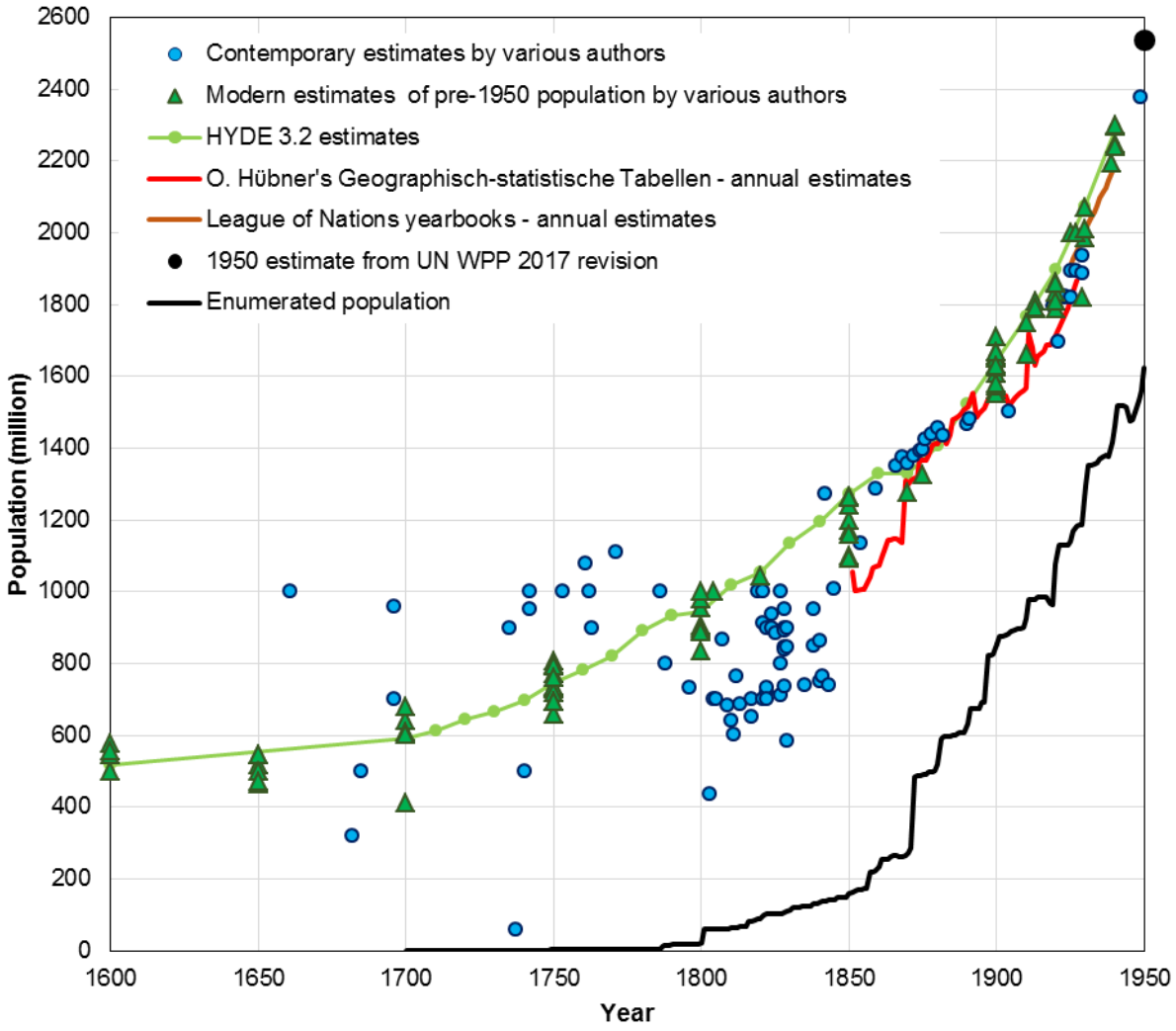
- Fill gaps in some records, e.g. concerning census type, territory or area, also specifying the vague 'territory of the census' with more accurate description;
- Resolve problem of contradicting information between sources regarding some censuses, including different UN publications;
- Investigate closer the colonial-era, and some old European/American censuses in regard to their eligibility for inclusion in the database. Also, consider whether to include censuses that counted only the non-indigenous population by showing only the small number of people actually enumerated with appropriate notes;
- Link the database with original census publications that are available online;
- Similarly to UN online databases, possibly also record basic demographic breakdowns of the enumerated population (sex, age, urban/rural residence) and official population estimates, as the UN presently only records information at present political divisions and from 1946 onwards.

In this study, world population was found to be underestimated, but the uncertainty has been declining in the last few decades. But what about the pre-1950 population figures? A similar analysis to the one described in section 3.3 would be difficult to perform for two reasons. First, census coverage before 1950 is lower, therefore making the analysis less universal. Second, the pre-1950 estimates of GDP per capita are probably even more uncertain than the population figures for most countries (Maddison 2003). Hence, the topic was approached differently, i.e. by analyzing the span of past population estimates. Historical developments in estimating the world population have been described before, with some further juxtaposing estimates from different sources (e.g. Willcox 1931, Maddison 2003, Caldwell and Schindlmayr 2002, Klein Goldewijk et al. 2010). In a similar vein, Fig. 10 was devised extending Diagram 2 from Willcox (1931). It includes 86 contemporary population

estimates together with 66 annual estimates from Otto Hübner's *Geographisch-statistische Tabellen* gathered by Willcox. Those data points are supplemented by 85 modern estimates of pre-1950 population, 16 annual estimates from the League of Nations yearbooks and HYDE 3.2 time series (27 data points). Finally, the total enumerated population is shown with the UN's 1950 estimate from the World Population Prospects: The 2017 Revision. The full list together with sources of information is included in the online dataset.

In the period before 1850, the world population estimates vary considerably, by 20–25% in case of modern sources, and far more in case of contemporary ones (a 1702 estimate of 4 billion is not shown). Only after several American and Asian countries had been enumerated in the second half of the 19th century, the span of estimated narrowed considerably to about 10%. At the turn of the century, more than half of the world population was counted and the uncertainty declined further to around 5% in the 1930s. Similarly, the UN has re-estimated the 1950 population many times, with the highest estimate being 5.1% above its lowest published figure. Not quite incidentally this equals the adjustment of global census returns for underenumeration.

Figure 10. Contemporary and modern estimates of world population, 1600–1950 and total enumerated population.



Source: Extended from Willcox (1931) with various sources (see online dataset for a full list).

5. Conclusions

The paper documented a compilation of basic information on population censuses from around the world. Its standardized format together with appropriate code and auxiliary datasets on political divisions and population estimates provides a tool for analysing spatial and temporal developments in census-taking activities. More than 3,200 censuses were identified, some including results at different political divisions and territorial compositions. It was found that all but one polity have had at least one census, and enumeration in the last census round reached 93% of estimated world population, or 89% if only up-to-date (i.e. no more than 10 years old) censuses are taken into consideration. *De facto* method has been more common than *de jure*, though the popularity of the latter slightly increased over time.

December and April have been the most frequent months of reference in the whole body of censuses.

The uncertainty analysis based on the dependency between estimated census errors and GDP per capita, modelled using a Gaussian copula, provided a global overview of population underenumeration. It was estimated that the error has declined slightly from around 5% in the 1950s to 3% in the 2010s. Further, United Nations estimates were found to be mostly in the vicinity of lower bound of the underenumeration's confidence interval. As a consequence it was estimated that in the 2010 census round the difference between the actual world population and combined census figures was equivalent to around nine years of demographic growth. According to the calculations made in this study, six of those nine years are caused by the low frequency of census-taking activities (typically decennial); two years are caused by underenumeration during censuses, an error which is corrected in UN estimates; and one year is due to further underenumeration, i.e. not taken into account by the UN, but likely to exist in light of the results of the uncertainty analysis.

6. Acknowledgments

The author would like thank Oswaldo Morales-Nápoles of Delft University of Technology for his help with the uncertainty analysis.

References

Bolt, J., Inklaar, R., de Jong, H., and van Zanden, J. L. (2018). Rebasing 'Maddison': new income comparisons and the shape of long-run economic development. Groningen: University of Groningen (Maddison Project Working Paper 10).

Caldwell, J. C. and Schindlmayr, T. (2002). Historical population estimates: unraveling the consensus. *Population and Development Review* 28(2): 183–204.

Correlates of War Project (2014). Territorial Change (v5) [electronic resource]. http://correlatesofwar.org/data-sets/territorial-change.

Correlates of War Project (2016). State System Membership List, v2016 [electronic resource]. http://correlatesofwar.org/data-sets/state-system-membership.

Genest, C., Rémillard, B., and Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics* 44: 199–213.

Gleditsch, K. S. (2013). List of Independent States [electronic resource]. Colchester: University of Essex. http://ksgleditsch.com/statelist.html.

IMF (2018). World Economic Outlook Database [electronic resource]. Washington DC: International Monetary Fund. https://www.imf.org/external/pubs/ft/weo/2018/01/weodata/index.aspx.

INSEE (2018). Statistiques [electronic resource]. Paris: Institut national de la statistique et des études économiques. https://www.insee.fr/fr/statistiques.

International Organization for Standardization (2017). ISO 3166 "Codes for the representation of names of countries and their subdivisions" [electronic resource]. Geneva: ISO. https://www.iso.org/obp/ui/#search.

Joe, H. (2014). *Dependence Modeling with Copulas*. London: Chapman & Hall/CRC.

Klein Goldewijk, K., Beusen, A., and 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(4): 565–573.

Klein Goldewijk, K., Beusen, A., Doelman, J. and Stehfest, E. (2017). Anthropogenic land use estimates for the Holocene; HYDE 3.2. *Earth System Science Data* 9: 927–953.

League of Nations (1927). *International Statistical Year-Book 1926.* Geneva: Economic and Financial Section, League of Nations.

Maddison, A. (2003). *The World Economy: Historical Statistics*. Paris: OECD.

Maddison, A. (2010). Historical Statistics of the World Economy: 1-2008 AD [electronic resource]. Groningen: University of Groningen. http://www.ggdc.net/maddison/Historical_Statistics/horizontal-file_02-2010.xls.

Maktabi, R. (1999) The Lebanese Census of 1932 Revisited. Who Are the Lebanese?. *British Journal of Middle Eastern Studies* 26(2): 219–241.

Marshall, M. G., Gurr, T. R., and Jaggers, K. (2017). Polity IV Project - Political Regime Characteristics and Transitions, 1800-2016, Dataset Users' Manual [electronic resource]. Vienna, VA, USA: Center for Systemic Peace. http://www.systemicpeace.org/inscrdata.html.

McEvedy, J. and Jones, R. (1978). *Atlas of World Population History.* Penguin.

Morales Nápoles, O., Paprotny, D., Worm, D., Abspoel-Bukman, L., Courage, W. (2017). Characterization of precipitation through copulas and expert judgement for risk assessment of infrastructure. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering* 3(4): 04017012.

Paprotny, D. (2018). Population Enumerations Database [electronic resource]. Figshare. https://figshare.com/articles/Population_Enumerations_Database/6949334.

Russett, B. M., Singer, J. D., and Small, M. (1968). National Political Units in the Twentieth Century: A Standardized List. *The American Political Science Review* 62(3): 932–951.

Statistics Iceland (2018). *Inhabitants overview* [electronic resource]. Reykjavik: Statistics Iceland. https://statice.is/statistics/population/inhabitants/overview/.

United Nations (1955). *Demographic Yearbook 1955*. New York: Statistical Office of the United Nations.

United Nations (1962). *Demographic Yearbook 1962*. New York: Statistical Office of the United Nations.

United Nations (1992). *Demographic Yearbook 1990*. New York: Statistical Office, United Nations.

United Nations (2017a). World Population Prospects: the 2017 Revision [electronic resource]. New York: UN Statistics Division. http://esa.un.org/unpd/wpp/.

United Nations (2017b). National Accounts Main Aggregates Database [electronic resource]. New York: UN Statistics Division. https://unstats.un.org/unsd/snaama/Introduction.asp.

United Nations (2018). Demographic Yearbook System [electronic resource]. New York: UN Statistics Division. https://unstats.un.org/unsd/demographic-social/products/dyb/#overview.

Willcox, W.F. (1931). Increase in the Population of the Earth and of the Continents since 1650. In: Willcox, W.F. (ed.). *International Migrations, Vol. II*. New York: National Bureau of Economic Research: 33–82.