National Accounting Review

NAR, 5(3): 282–297.
DOI: 10.3934/NAR.2023017
Received: 06 June 2023
Revised: 14 September 2023
Accepted: 19 September 2023
Published: 27 September 2023

http://www.aimspress.com/journal/NAR

Research article

Population data quality checks: Romanian adult deaths and lives, an evaluation

Iulia Toropoc*

Independent Researcher, 172 Holland Park Avenue, London, United Kingdom

*** Correspondence:** Email: iulia_toropoc@yahoo.co.uk.

Abstract: This article investigates the quality of the Romanian population data for the years 1966–2018. The following age heaping measures are being used, each with its recommended population age span: Whipple, Myers, Bachi, UNASA, Kannisto, Coale and Li. In addition, the Kannisto measure of old age overstatement is used. Our analysis corroborates the results of our measures with the demographic characteristic of the Romanian population. We aim to establish whether; (a) the data spanning the time interval 1966–1989 is more accurate than the data collected during the 1990–2018 interval, (b) the Romanian data is at least of comparable quality with the tier 2 HMD data and/or (c) the Romanian data is not grossly inaccurate and therefore suitable for research. We find support for (a) and, cautioning against the live data for the time interval 2002–2018, for (b) and (c). Limitations aside (no through overstatement or cohort analysis), this is the first study to attempt such evaluation and hopefully not the last.

Keywords: demography; human geography; age heaping; age overstatement

JEL Codes: J11, J13, J14, Y10

1. Introduction

Country level annual population size estimates and death counts are produced by national statistics institutes and can be accessed by both institutional entities and private individuals. Population size estimates and death counts are of utmost importance for both governmental and private planning. While deaths counts are considered to be, with the exception of those recorded at advanced ages,

generally reliable, the population estimates and the census data that they are derived from are generally considered less reliable. There is also great variability in the quality of data, its reliability being conditioned by a country's level of development.

For both population and death counts, errors in the data arise at the point of collection and at the point of entry. A good indication of data quality and reliability is obtained with the help of standardised data testing procedures. As any other type of collected data, census data is prone to reporting or response errors and entry errors (United Nations, 1956, Shryock and Siegel, 1976). Point of collection and entry errors may arise when census dedicated resources are either insufficient or inadequate. The response errors that are most likely to occur are those related to age and sex reporting. According to the United Nations (1956), these errors can be the result of deliberate misreporting, for example the understatement of age for young and middle-aged women in an effort to appear younger, or involuntary, for example the overstatement of age for adolescent girls in some African cultures. There is also a possibility that the overstatement of age for adolescent girls in traditional communities is voluntary, serving the purpose of circumventing the legal restrictions around the minimum age of marriage. As factors that contribute to census misreporting, Byerle and Terera (1981) mention the age of the respondent, the level of formal education of the respondent and the method by which the age is assessed. Interestingly, Byerle and Terera (1981) do not find gender to be a significant factor. Influenced by Mokyr and O Grada (1982), who linked age heaping to numeracy skills and recommended age heaping as a proxy for population sophistication, A'Hearn et al. (2009) identified a correlation between age heaping and literacy (and between numeracy and literacy) and proposed age heaping as a proxy for illiteracy and, more widely, as a measure of human capital. A decade later, A'Hearn et al. (2022), reconsidered the connection between age heaping and literacy and choose to explain age heaping as representative of cultural, economic and institutional development.

Lee and Zhang (2017) identify two sources of age heaping in census data, the interviewer and the interviewee, with the errors originating from these two individuals as representative of a weak state capacity. The two researchers found consistent correlations between age heaping and state capacity indicators, such as government efficiency, corruption and political stability. Particularly problematic are the very old age counts, these raising reliability issues regardless of the level of country development. Indeed, very old age death counts have been found to be artificially high for populations as diverse as those of Latin America (developing countries) (Dechter and Preston, 1991) or Germany (developed country) (Jdanov et al., 2016). Irrespective of age, census data has been found to be particularly unreliable in African countries (mostly classed as least developed), where traditional norms and customs circumvent the conventions of age and adulthood that are established by western societies (Caldwell and Igun, 1971, Byerle and Terera, 1981). However, there is indication that the situation in Africa has improved recently, as the data for Botswana, year of census 2011 confirms (Bainame and Letano, 2015). Similarly unreliable are the population size estimates supplied by countries that are heavily affected by migration, such as the countries of the former Eastern Bloc (classed as developing and developed countries) (Penina et al., 2015).

To maximize the success rate in the identification of age and sex reporting errors in demographic data, particularly census data, the United Nations put forward a four-step framework: data inspection, data mapping, ratio analysis and index measures (United Nations, 1956). This framework is to be accompanied by intimate knowledge of the associate historic demographic trends so that human errors

are separated from true demographic peculiarities, such as the low numbers of males produced by census data from conflict or emigration zones (United Nations, 1956). For example, referring to the census data of Bangladesh, Fajardo-Gonzales et al. (2014) suggested country specific high emigration rates rather than data error as explanation for the persistent underreporting of young males.

Regarding the advanced age population, there is general consensus among researchers that the number of individuals reported either alive in censuses or registered in death statistics is unnaturally high (Preston et al., 1997), indicating coverage errors. At the same time, there is less consensus regarding the type of error that prevails in old age reporting, with researchers opting for both over and understatement. Gibril (1975) identified overstatement at advanced ages in Gambian census data, while Myers (1966) found evidence of exaggeration at ages over 100 in the US census data. Conversely, Caldwell and Igun (1971) found that Nigerian females over the age of 60 were more likely to underreport their age than males, while Preston et al. (1996) found that African Americans over the age of 60 were underreported in regard to age on their death certificates. Kannisto (1999), pointed out that gender bias was associated with age overstatement at advanced ages of death, with overestimation only occurring in the male data. Another source of error in mortality data at advanced ages is age heaping, particularly on the 0 and 5 digits (Ewbank, 1981). If these errors are selective, they may result in sex bias (Kannisto, 1999). Coale and Kisker (1986) and Kannisto (1999) concluded that age overstatement and age heaping explained the low mortality rates observed in populations with poor quality data, with Kannisto (1999) further denying the possibility of age understatement among the old. As for the separate detection of age heaping and age overstatement at advanced ages, Kannisto (1999) cautions that while age heaping is easily detected, age overstatement is far more difficult to detect, requiring rigorous birth registration practices that have been in place for at least a century. This last caveat significantly restricts the countries that supply good data at advanced ages. Indeed, Kannisto identifies only 13 countries with good quality data (Kannisto, 1999).

In this paper, we attempted to evaluate the accuracy of the Romanian population live and death counts for the years 1966–2018, using a number of standard tests for age heaping and age overstatement, as suggested by the United Nations (1956), Shryock and Siegel (1976), Coale and Li (1991), Kannisto (1999) and Yi and Gu (2008). Inspired by these sources, we used the following measures to address age heaping: Whipple's, Myers', Bachi's, UN age-sex accuracy (UNASA), Kannisto's and Coale & Li's indices and Kannisto's (1999) overstatement ratio for the oldest old. Based on the hermetic population conditions (very little and tightly controlled migration flows) that characterised Romania between the years 1947 and 1989, we hypothesised that the data for the 1966–1989 time interval were more accurate than that for the 1990–2018 interval. We found support for our hypothesis in Jdanov et al. (2016), who pointed out major errors in the Moldavian demographic data, largely due to the phenomenon of labour migration, phenomenon specific to the countries of the former Eastern Bloc. However, we did not exclude data irregularities for the first interval, mainly but not solely, because of the irregularities likely to be carried on from the pre-1947 era. We also suspect that the high fertility rates, although somehow tempered by the high mortality rates, might have further contributed to the demographic peculiarities of our dataset. We also hypothesised that at least part of our data was of comparable quality with the human mortality database (HMD) data of acceptable quality (HMD2). To test this hypothesis, we compared Kannisto and Coale and Li index values for the Romanian and HMD2 data. Overall, we hypothesised that at least part of our data did not present major errors and that it was adequate for scientific research.

Concluding our introduction, we would like to make a few points about the framework of this paper. The knowledge that we use and produce has always been imperfect. The best we can do is to be aware of this and willingly produce imperfect, yet proficient knowledge artefacts. Abbot's (2018) canonical and legalist working paradigm seems to offer the best solution for dealing with the imperfect knowledge that we procure and produce. While the canonical approach constricts us to a temporary fixed space of established texts and methods, the legalist approach challenges us by exposing us to an ever-changing set of problems. For this paper, we referred ourselves to established methodologies and accessed secondary information sources related to our data. While we approached this knowledge deductively, we approached our problem inductively, questioning the suitability of each of the measures employed for our data.

2. Data and methodology

2.1. Data

We obtained Romanian population and death counts by age and sex from Eurostat. The data was available in both single year and 5-year format. This was fortuitous because it allowed us to compile all the indices we settled upon. The death counts were available for the time interval 1966–2018, while the live counts only covered the interval 1968–2018. Both counts were available up to age 84 for the period 1966–1988 and up to age 99 for the remaining period, with an open interval for 85+ and 100+, respectively. The data contained complete population counts covering four census years: 1966, 1977, 1992, 2002 and incomplete counts for the year 2011, this giving us the opportunity to split the data into the five census groups that we used for our HMD comparisons, with the caveat that the most recent census group is incomplete.

We decided for the time interval 1966–2018, as this was the nearest longest interval with data available. We also believed that an interval of this length allowed for a comfortable split with 1989 the cut-off point and finally, as smallest analysis units, for census data intervals. As pointed out by Simandan (2010), a longer time interval greatly diminishes the chances that short term noise is mistaken for long term trend. Naturally, we could have used only the most complete interval, 1990–2018. However, we felt that we might have fallen prey to the fallacy of the law of small numbers (Tversky and Kahneman, as cited by Simandan, 2010), i.e., we might have succumbed to the temptation of frequent information, one of the two pitfalls of our current knowledge economy. We believe that if we restricted our analysis to a shorter time interval than the one used, our analysis would have offered a rather incomplete tableau of the Romanian population data.

2.2. Methodology

We commenced our analysis with the calculation of six indices: Whipple's, Myers', Bachi's, UNASA, Coale and Li's and its Yi and Gu extension and Kannisto's indices. These indices are standard protocol for the evaluation of age heaping. For HMD2 comparisons, we split our data into five census intervals, four of which contained complete data.

Evidence for age heaping has been found for both very young and old ages. Research conducted on various African and Bangladeshi populations established that at younger ages there is a clear preference for ages 4 and 5 over the age of 3 (Bairagi et al., 1982) in the Bangladeshi rural population, while for all populations studied, there is clear misreporting of the ages 0 to 4 (Bairagi et al., 1982, Caldwell, 1966, Caldwell and Igun, 1971). As for advanced ages, Caldwell and Igun (1971) found that Nigerian females over 50 tend to have their age underreported, while Preston et al. (1997) confirmed misreporting, predominantly in the form of underreporting of ages, particularly for females, at advanced ages.

The detailed, step by step procedure for the calculation of the first three indices can be found in the United Nations population manual (1956) and in Shryock (1976). It is important to mention that all three indices assume rectangularity of the true distribution of the population by age (Shryock and Siegel, 1976) and that, as is usually the case with composite measures, their capacity to measure correctly and exactly the heaping on individual ages or specific digits will always be questionable.

Whipple's index is designed to tackle age misreporting due to preferences for the digits 0 and 5. It is therefore a measure of digit preference. Originally applied to the 23–62 age group, it can be applied to any age group, with the caveat that it is not an accurate evaluator for age misreporting at very young and very old ages (Shryock and Siegel, 1976). Although, a very efficient and simple measure of age heaping, the index is restricted only to two digits (United Nations, 1976). The range for Whipple's index is from 100–500, where values under 105 indicate data accuracy and values over 125 indicate considerable age heaping in the data. Values lower than 100 but very close to 100 still indicate accuracy, and other values lower than 100 indicate heaping at other digits. The index is derived as follows:

$$WI = \frac{\sum(P_{25} + P_{30} + P_{35} + \cdots P_{60})}{\frac{1}{5}\sum(P_{23} + P_{24} + \cdots P_{61} + P_{62})} * 100 \quad (1)$$

Myers' index extends Whipple's index to all 10 digits. It is therefore a measure of digit preference. To avoid the bias towards digits other than 0, Myers' index uses weighted totals for each digit so that, once averaged, a frequency for each digit is obtained that is close to 10% of the total. The index is derived by taking the half of the total of the deviations from 10% for all the 10 digits. The range for Myers' index is from 0–90, where 0 indicates no inaccuracies regarding age and 90 indicates the presence of only 1 digit for all the data. Myers' Index can be applied to all ages, starting at age 10.

Bachi's index applies the Whipple method iteratively and successfully avoids the theoretical shortfalls of the previous two indices (United Nations, 1976). Similar to Myers' index, it computes a frequency of about 10% for each digit. The method is computationally time consuming and therefore inferior in popularity to Whipple's and Myers' indices. The range for Bachi's index is from 0–90. The values for Bachi's index should be close to those obtained with Myers' index. Similar to Myer's index, Bachi's index can be applied to all ages, starting at 10. Using the two indices on the same data should therefore indorse the accuracy of the findings.

The UNASA index uses sex and age-ratios for 5-year age groups, stopping at age 70 to compute averages of the differences between adjacent age groups. For the sex ratios the standard formula is $SR = \frac{M}{F} * 100$. For the age ratios the standard formula is $AR = \frac{5\_P_a}{\frac{1}{2}(5\_P_{a-5} + 5\_P_{a+5})} * 100$. The index is

calculated as three times the mean difference in SR plus the mean deviation for male and female AR. The index does not account for real irregularity in age distribution and for the age-progressive decline in age ratios (Shryock and Siegel, 1976). As it uses an AR formula that lacks a central group, the index is prone to upward bias (i.e., overestimation) (Shryock and Siegel, 1976). In addition, the index gives considerable importance to the sex-ratios. Critics of the index, accuse it of inability to capture naturally occurring irregularities. Nevertheless, the UNASA index is a valuable measuring tool, particularly when country comparisons are made (Shryock and Siegel, 1976) and when single-year age data is not available. Unlike the other three indices, the index captures age misstatement, digit-preference and differential omission (United Nations, 1956). The index values classify the data as accurate for values smaller than 20, inaccurate for values between 20 and 40 or highly inaccurate for everything higher than 40. The UN ASA Index is recommended for the 0–74 ages.

The Coale and Li index is a measure of the deviation of the reported age in census data for both live and death counts from an ideal smooth sequence using a two-stage moving average (Coale and Li, 1991). This index is a natural follow-up to the solution proposed by the United Nations Statistics Division in 1956 for the evaluation of misreporting at advanced ages. The United Nations recommended the comparison of individual age data with data obtained via graduation or smoothing and suggested a 5-year moving average as satisfactory procedure (United Nations, 1956). In their original paper, Coale and Li (1991) focused on ages divisible with 10 within an age range of 40–90. As a follow-up, Yi and Gu (2008) constructed indices based on all individual ages within an age range of 85–105. We opted for two versions of the Coale and Li index, one restricted to 0 digits and one extended to all digits, in the manner of Yi and Gu (2008) but covering the original Coale and Li ages. The range of the Coale and Li index is between 0 and 5, with indicators close to 1 suggesting no heaping.

The Kannisto index procedure is described in detail in Kannisto (1999). The Kannisto index is in fact a measure of age heaping for individual ages ending in 0 or 5 above the age of 80. To be able to compute the index for a particular age, one needs individual years of age at adjacent ages: $H_x = \frac{D_x}{\widehat{D}_x}$, where $\widehat{D}_x = \exp\left(\frac{1}{5}\sum_{y=x-2}^{x+2} \ln D_y\right)$. Indicators should be ideally higher than 1 due to the effect of the logarithm on the mortality curve and lower than 1.05 for the age 90 (Kannisto, 1999).

The Kannisto good country data index is based on data for the following countries: Austria, Denmark, England & Wales, Finland, France, Iceland, Italy, Japan, Netherlands, Norway, Sweden and Switzerland.

The Kannisto measure of age overstatement is obtained with the help of the ratio between the +100 and the +85 death counts. A high ratio questions the quality of the data both for the older old and the younger old (Kannisto, 1999). Good country data presents ratios of 7.0 for males and 15.5 for females.

The Human Mortality Database (HMD) provides open access mortality and population data for 41 countries. The database is the joint effort of Max Plank Institute and the University of California, Berkeley and, to date, is the most comprehensive and accurate database of its kind. The country data that is compiled by the HMD is corrected for processing errors but not for age misreporting (age heaping and age exaggeration) and over or under-enumeration. Regarding age misreporting, the age heaping (on digits 0 and 5) that is present in the data is considered to be of a satisfactory level, while age exaggeration at old ages is addressed via the extinct generation method (population data derived

from death data). The country data is split in series of 10 and classed as: best data (no problems across the entire period), acceptable data (some data problems), conditionally acceptable data (consistent and moderate data problems) and weak data (major data problems) (Jdanov et al., 2008). Based on the level of data accuracy, a country can be part of one group only or of more than one group (the majority).

3. Results

3.1. Lives

We used the following indices for our live counts: Whipple, Myers, Bachi, UNASA, Coale and Li.

3.1.1. Whipple's index

We produced indices for the age range 23–62, total population counts and males and females separately. For all age ranges and for all three population count types the index indicated that as far as age heaping at 0 and 5 was concerned, the data was highly accurate. The majority of the values were slightly below 100, indicating very high accuracy. The highest value obtained was 117 for females, year 2018. Overall, the data was very accurate. The values of the Whipple index are shown in Table Set 3 of Appendix 2.

3.1.2. Myers' index

We produced indices for total population counts and males and females separately. We used the following age ranges: 10–79 for the 1968–1988 data and 10–99 for the 1989–2018 data. For all age ranges and for all three exposure types (male, female, total) the index values indicated that, as far as age heaping at 0 to 9 was concerned, the data was very accurate. Values ranged from 1 to 4. The values of the Myers index are shown in Table Set 3 of Appendix 2.

3.1.3. Bachi's index

We used the same rationale as for the Myers' Index. Values ranged between 1 and 3, rendering the data very accurate. The small differences between Bachi's and Myers' indices further endorsed the accuracy of our data. The values of the Bachi index are shown in Table Set 3 of Appendix 2.

3.1.4. The UN age sex accuracy index

We calculated the UNASA index for our total population counts, age range 0–74, following the procedure detailed in Shryock (1976). The values ranged between 21 for 2018 and 36 for 1985, rendering the Romanian population data inaccurate. For comparison, we calculated indices for Lithuania and Latvia, two former Eastern Bloc countries that were included in our composite HMD2 index. The data for the two countries covered the period 1959–2019. For Lithuania, the index ranged

from 22–38, while for Latvia, the values were between 27 and 39. None showed a clear value descending trend. The values for the UNASA Index are shown in Table Set 3 of Appendix 2.

3.1.5. Coale and Li index

We divided our data into 5 intervals, each with a census year as start point. We calculated the Coale and Li index for live counts, males and females separately, at 30–80 for the first two intervals and at 40–90 for the last three intervals.

Using the HMD2 data, we calculated country indices and then an overall HMD2 index. We then compared the values of the Romanian indices to those of the HMD2 index. We produced indices both for 0 digit only, as originally calculated by Coale an Li (1991), and for all digits, as subsequently calculated by Yi and Gu (2008). The HMD2 index methodology is described in detail in our Kannisto Index paragraph, as Kannisto (1999) was our inspiration for our index.

Coale and Li index 0-digit index values for male and female lives across all intervals were either lower or non-significantly higher than their HMD counterparts.

The all-digit index values showed a rather complicated picture. For the age range 30–80, male and female lives, the two intervals spanning the time period 1968–1991 produced values that were smaller or significantly not different from their composite counterparts. For the age range 40–90, male and female lives, two of the three intervals produced indices that were higher and significantly different from their HMD2 counterparts, both at composite and individual level. The problematic intervals were 2002–2010 and 2011–2018. However, all Coale and Li values were close to 1. The values of the Coale and Li index for live data are shown in Table Set 1 of the Appendix 2.

3.2. Deaths

3.2.1. Kannisto index

We calculated the Kannisto index for the ages 80, 85, 90, 95 for the years 1966–2018. We split the data in five groups to cover the time periods between censuses: 1966–1976, 1977–1991, 1992–2001, 2002–2010 and 2011–2018. We used HMD data to calculate the Kannisto Index for 13 countries with good data, as suggested by Kannisto (1999). We also calculated a conditional data Kannisto index based on tier 2 HMD data, HMD2. For our index, we selected six out of the eight countries classed by Jdanov et al. (2008) as conditionally acceptable in terms of their death data (tier 2). Our selection criteria were a time series of length 10, a similar life expectancy to that of the Romanian population, time proximity to our census interval, the absence of extreme events. We settled for the following countries and time periods: Canada (1961–1970), Latvia (1961–1970), Lithuania (1981–1990), Luxembourg (1961–1970) New Zealand Non-Māori (1951–1960) and USA (1961–1970). We calculated country indices and then an overall tier 2 index (HMD2).

The Romanian index values were either lower or non-significantly higher than those of the composite good data Kannisto index for all ages and time periods, apart for males 1992–2001 and females 196–1976 of age 85. The HMD2 Kannisto index singled out the male data for the age of 85,

time period 1992–2001. All our values were close to 1. None of the values for the ages 90 and 95 exceeded the value of 1.5. The values of the Kannisto index are shown in Table Set 2 of Appendix 2.

3.2.2. Kannisto death ratios

We calculated ratios for 100+/85+ for the time interval 1966–2018. We split the interval in five subintervals based on intercensal periods. Our female ratios were: 6 (1966–1976), 6.5 (1977–1991), 4.7 (1992–2001), 7.4 (2002–2010) and 9.1 (2011–2018), and for the same time periods, our male ratios were 4, 4.5, 3.3, 6.6 and 7.2. For comparison, the ratios obtained for the thirteen good data countries for the time period 1980–1990 by Kannisto (1999) were 7.0 for males and 15.5 for females. Ratios below this norm are indicative of higher than standard mortality, while ratios above this norm indicate potential age overstatement. Our female ratios were considerably lower than the norm, indicating higher mortality. For the complete intercensal intervals, our male highest ratio was only slightly lower than the norm, all others being considerably lower than the norm. As decennial ratios increase with improved mortality, there are no improvements in mortality during the interval 1966–2001. In contradistinction, the interval 2002–2018 shows improvements in both female and male mortality, to the extent that the male mortality matches that of the benchmark. Among the countries used by Kannisto (1999), we found three countries that we included in our benchmark index. The ratios for Canada (17.3 and 30.3) and USA (11.8 and 22.8) were extremely high, clearly indicating age overstatement, while the data for New Zealand Non-Māori was only slightly elevated (8.6 and 17.6). Based on these comparisons we conclude that the oldest old Romanian population data is unlikely to suffer from overstatement. The values of the Kannisto ratios are shown in Table Set 2 of the Appendix.

3.2.3. Coale and Li Index

We proceeded in an identical manner as for the live data. For the age range 40–90, male and female deaths, all five intervals produced Coale and Li 0-digit indices that were either lower or non-significantly higher than their HMD counterparts. For the all-digit Coale and Li version, the overall picture is rather complicated. As far as female deaths were concerned, all five intervals produced indices that were either lower or non-significantly higher than their HMD counterparts. The male deaths, on the other hand, varied in quality. All death data in the 1966–1991 interval was of acceptable quality. In contradistinction, the interval 1992–2018 produced indices that were significantly higher from their HMD composite counterparts. However, those in the first part of the interval covering the 1992–2001 period, were not significantly different from the country indices of USA and Canada, while those reported during the 2011–2018 interval, were significantly lower than those for USA and Canada. Nevertheless, all Romanian values were close to 1. The values of the Coale and Li index for live data are shown in Table Set 1 of the Appendix 2.

4. Discussion

The primary aim of this paper was to evaluate the Romanian population data, death and live counts, from the point of view of its suitability for scientific research. For this purpose, we formulated three

hypotheses: (1) The data spanning the time interval 1966–1989 is more accurate than the data collected during the 1990–2018 interval. (2) The Romanian data is at least of comparable quality with the tier 2 HMD data. (3) The Romanian data is not grossly inaccurate and therefore suitable for research.

We would like to start our discussion by making a few points about the internal environment characteristic of the Romanian population during the period under investigation. We consider these points necessary, as, similar to all former Eastern Bloc countries, Romania experienced a seismic shift at the end of the 1980's and is on an abrupt and in many ways dissonant westernisation course ever since. During the period 1966–1989, Romania had a controlled and monitored external and internal migration, a stable internal environment (economic hardships of the 1980's aside), meticulous population monitoring and registration, a very high literacy rate and only one significant natural disaster. This period was covered by two population censuses, the 1966 and the 1977 censuses. These censuses are considered by specialists as remarkable in their completeness (coverage) and quality of data and relied on electronic computations (Mihaescu et al. 2018). There is therefore little evidence for age heaping resulting from interviewer/interviewee, input or methodology errors. In contradistinction, the post 89 period is characterised by low fertility rates, an aging and declining population, strong internal migration and steady temporary emigration, with permanent external migration peaking in the early 1990's and more recently, following Romania's EU accession, by immigration. Considering the numerous sources that had to be aggregated to produce population data, the data is open to both under and overestimation (Sandu, 2018). In a study conducted on the population of Moldova, a country with high emigration rates and data recording procedures that are prone to overestimation, Penina et al. (2015) found that the official population size was 18% higher than the actual size. We would like to suggest a further difficulty arising from the use and abuse of Romanian identity cards, a requirement for Romanian residency. We do not exclude respondent, interviewer and input errors during the time interval 1990-2018 and therefore an element of state capacity (or rather state weakness), however, this is unlikely to be of such magnitude, as to seriously impact the quality of the data. Nevertheless, this latter period is more likely to produce problematic population data. See Appendix 1 for a more complete picture of Romanian demographic characteristics.

Next, we would like to make some points about our methodology and how this might have impacted on our results. We used the standard procedure for the Whipple, Myers, Bachi and UN ASA indices. For the Kannisto index, we followed the standard procedure, including the comparison with the 13 countries with good data, with one major exception, the calculation of the Index at age 80. As our data allowed us to extend the calculation for older ages to age 80, we proceeded with the calculation of the index at age 80 in an identical manner with that of the indices for the older ages. We did not find this approach in the existing literature and therefore we cannot endorse it based on precedent. As we suspected that our data would not match the quality of that included in the original Kannisto index, we created a similar index based on six countries with data of conditionally acceptable quality from the HMD (HMD2). We compared the values of our Romanian indices with those of the composite data indices and, where necessary, individual country data. We did not find this approach in the existing literature and therefore we cannot endorse it based on precedent. We venture to suggest that we allowed ourselves a degree of creativity, as far as the calculations for this index are concerned, extending the approach for older ages to age 80 and compiling a new benchmark index more suitable for our purposes. For the Coale and Li index (1991), we followed the standard procedure detailed in Coale and Li (1991)

and extended our calculations to include all digits for both death and live counts as proposed in Yi and Gu (2008). For comparison purposes, we used the composite index based on the six countries with acceptable data quality that we compiled for the Kannisto index. We compared the values of our Romanian indices with those of the composite data indices and individual country data where necessary. We did not find this approach in the existing literature and therefore we cannot endorse it based on precedent. Again, as with the previous index, we approached this index creatively, keeping the original calculations and yet extending the applicability to all digits and also compiling a benchmark index for comparison.

We will now be revisiting our live data results and comment on them. To detect age heaping in our live data, we employed a number of well-regarded, established methods: the Whipple, Myers, Bachi, UNASA, Coale and Li and Kannisto indices. While the Whipple, Myers and Bachi indices established that our entire interval live data was very accurate as far as age heaping was concerned, the UNASA index classed our data as inaccurate, with inaccuracy more pronounced for the interval 1968–1989 than for the interval 1990–2018. Our Whipple index included values in their 90s, slightly smaller than 100, i.e., the index's starting point. Similar values were obtained by Chemhaka et al. (2016) and were considered as indicators of high data accuracy. The Coal and Li index rendered the data for the time interval 2002–2018 as being of lower quality than that of the HMD2. Individual country comparisons did not establish data quality compatibility. We venture to suggest emigration as a potential disturbance factor in our data for the 2002–2018 period. In January 2002 Romanian citizens were granted free movement in the Schengen area and in January 2007 Romanian citizens received the right to free movement and work within the European Area. These two events opened the door to unrestricted migration, either temporary or permanent. A further supporting factor for our explanation is that the index calculations found age heaping in both male and female data. Unlike countries where gender is a significant determinant in the level of migration, for example the predominance of males among the Bangladeshi emigrants, the internal and external economic conditions and the level of skill within the adult population, favoured an almost identical number of male and female emigrants. Sandu (2018) found no significant gender differences in Romanian temporary emigration data covering the time period 1990–2016. The findings of Sandu (2018) also exclude the possibility of gender misreporting, contrary to the findings of Fajardo-Gonzales et al. (2014) regarding the Bangladeshi census. Against the results of the UNASA index, we concluded that our 1968–2001 data passed the research suitability test, i.e., that it was of compatible quality with the HMD2 data and therefore contained no major discrepancies and was suitable for research purposes. Considering the failure of the more recent data to pass the Coale and Li test, we also conclude that the Romanian live counts up to the year 1989 are more accurate than those after. As for the exclusion of the UNASA index results from our supporting evidence, we relied on precedent, index's limitations, peculiarities of Romanian demographics and comparisons with selected HMD2 data. Bainame and Letano (2015) were confronted with the same unusual situation in their evaluation of the 2011 Botswana Census data, albeit their index value being lower than ours. They suggested that the UN ASA index was unable to account for the high mortality of Botswana males and concluded that their data was accurate. Indeed, the UNASA index is sensitive to both normal and therefore expected fluctuations in births and deaths and unexpected fluctuations in population numbers such as those caused by manmade or natural disasters (Shryock and Siegel, 1976). In addition, the exclusion of a mid-point from their age-ratio calculation,

renders the index prone to overestimation (Shryock and Siegel, 1976). Although the UNASA index calculations produced high values for the entire time interval, to our surprise, it is during the first, pre-89 interval that the UNASA index produced the highest values. We suspect that these values reflect the inability of the index to incorporate the data inconsistencies that resulted from the five wars that ravaged the Romanian population during the first half of the 20th century, the unusual pattern of the Romanian fertility rates during the period 1966–1989 and last but not least the precarious health situation of the Romanian population during the first half of the 20th century. We also do not exclude age registration errors for the population born previous to the year 1947. The population movements that characterised the post-89 interval might have impacted on the Index calculations for the second interval. These events might have initiated what the UN calls demographic peculiarities. We compared the Romanian UNASA index values with the values of Latvia and Lithuania, two countries with data of acceptable quality within the HMD. We selected these countries based on the similarity of demographic phenomena and environmental conditions: high number of casualties during WW2, relative high mortality, including infant mortality, controlled and documented population movements before 1991, high permanent and temporary migration and low fertility rates after 1991 (unlike Romania, neither of the two countries implemented a birth control policy but on the contrary). The UNASA indices rendered the data of these two countries inaccurate. This comparison reinforces our reservations about the compatibility of the UN ASA index with our dataset.

We will now be revisiting our death data results and comment on them. To detect age heaping in our death data, we employed the Coale and Li and the Kannisto indices. We used Coale and Li for death counts for the age group 30–80, time periods 1966–1976 and 1977–1991 and for the age group 40–90, the remaining time periods, each time period meant to approximate one census interval. For deaths over the age of 80 or the oldest old, we used the Kannisto index. For age overstatement in our oldest old, we used the Kannisto measure of overstatement. Our age ratios indicated little to no overstatement. As we did not possess individual age data for ages above 99, we were not able to perform any centenarian analysis, for example Jdanov's old age heaping index, this relying on deaths counts for the ages 95, 100 and 105. All heaping measures indicated that the female data spanning the entire time interval was comparable with HMD2 data, did not contain major inaccuracies and was suitable for research purposes. A different result was obtained for the male data, where the intervals 1992–2001 and 2011–2018, ages 40–90 were of inferior quality to the HMD2 data as far as age heaping was concerned. Individual country comparisons erased the discrepancies between the Canada and USA data used in our composite index and our data. The Romanian data for 85 years of age females 1966–1976 and males 1992–2001 and 2002–2018 failed to comply with the standards of Kannisto's good quality data countries. However, when we downgraded our comparison to that with the HMD2 index, only the male data of 1992–2001 remained problematic. All other data was of comparable quality to Kannisto's good data countries. However, a limitation of our old age data analysis still stands: the data was not rigorously tested for overestimation, where we only relied on a simple age overstatement ratio for the oldest old. According to Kannisto, irregularities in old age are fuelled by male overestimation. We would therefore welcome attempts at performing such analysis, with the caveat that rigorous analysis of this sort requires accurate births registration for at least a century, a demographic performance that is achieved by only a handful of countries (Kannisto, 1999). Cohort analysis at least for the populations found problematic by the Kannisto index might prove a welcoming avenue for

future research. It is worth pointing out at this stage that the 85-year-old male population of 1992–2001, would have been born between 1901–1911. Therefore, the corresponding data might have been impacted by the potential effect of pre 1947 data inaccuracies and population characteristics. Based on the results of our analysis, we conclude that the death data for the entire time period passes the research suitability test, i.e., that it was of compatible quality with the HMD2 data and therefore contained no major discrepancies and was suitable for research purposes. As the data covering the interval 1992–2018 required additional country specific comparisons to establish its suitability, we conclude that the data spanning the pre-89 interval is more accurate than that of the post-89 interval.

Interestingly, and in accordance with the opinion held by the majority of the population researchers to date, the Romanian death counts proved more accurate than the live counts, at least as far as the 2002–2018 period is concerned. We venture to offer migration as the driving force behind the data inaccuracy for live counts belonging to this interval. Our explanation finds support in the fact that there were no major inaccuracies in the data older than 2002, UNASA index aside. While we do not exclude respondent and interviewer error, and therefore an element of state capacity (or rather state weakness), as long as census procedure was adhered to in the post-89 time interval, we struggle to figure out how respondent preference for particular ages or interviewer error can influence the quality of the data to such extent in a country where birth certificates and identification cards are compulsory for all citizens and are used by interviewers at the point of census data collection. The post-89 abusive and erroneous use of identity cards might on the other hand be a source of serious error.

A final point that we would like to make as to the limitations and strengths of our paper, is that we are aware that both this article and the sources that we used for it are incomplete and situated. Indeed, the information that we source and sift through and our interpretation of it are affected by four epistemic gaps: possible worlds versus realised world, realised world versus witnessed situation, witnessed versus remembered situation and remembered versus confessed situation (Simandan, 2019). We are aware that the third and the fourth gaps are prominent in the case of monographies and interviews and also in research produced under a certain political regime or even a research paradigm. This subjects all our sources and even our paper to these two epistemic gaps. We have tried as much as possible to maintain objectivity in our inferences to minimise the impact of these two gaps. Of the four epistemic gaps, we believe that the second gap is particularly pronounced in situations when western methodologies are applied in non-western spaces. Although Romania has been through a course of rapid and alert westernisation since 1989, there is still little to suggest that the geographic, political and cultural characteristics of Romania define it as such. We hope that we managed to bridge, yet imperfectly, this gap with the help of locally sourced population information. As Abbot (2006) beautifully stated, we are all explorers in Borges' library of Babel, where we laboriously sift, process and interpret the imperfect information through which we make sense of our surrounding world.

5. Conclusions

Based on the Whipple (ages 23–62), Myers and Bachi (ages 0–99) Indices, the live counts for the entire 1968–2018 are accurate. The same data, was evaluated as inaccurate by the UNASA index (ages 0–74). The Coale and Li index (ages 40–90) pointed towards inaccuracies greater than those present in the HMD2 data, as far as the data for males and females, 2002–2018, are concerned. We conclude

that the live data for the interval 1968–2001 does not present major inaccuracies, is of comparable quality with the tier 2 HMD data and is suitable for research purposes. We also acknowledge the superiority of the 1968–1989 data to that of the 1990–2018 data.

The Coale and Li index renders the male death counts for 1992–2018 as of inferior quality to that of HMD2 composite data. However, the inaccuracies are not greater than those for USA and Canada, two of the HMD2 countries included in our composite index. The remaining data is either superior to or comparable in quality to the HMD2 data. When the data is benchmarked against the good country data, the Kannisto index detects heaping at age 85 for males, for the time interval 1992–2001, and for females, for the time interval 1966–1976. However, when the HMD2 data is used, only the male data is classed as of inferior quality to its benchmark. All other ages, at all time periods produce indices that are either superior to or not significantly different from the good country data index. We conclude that the death data for the entire time interval is of comparable quality to that of HMD2 countries and suitable for research purposes. Deaths and live counts taken into account, the data spanning the interval 1966–1989 is of superior quality to the data belonging to the 1990–2018 interval.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

Abbott A (2006) Reconceptualizing knowledge accumulation in sociology. *Am Soc* 37: 57–66. https://doi.org/10.1007/s12108-006-1005-9

Abbott A (2018) Varieties of normative inquiry: Moral alternatives to politicization in sociology. *Am Soc* 49: 158–180. https://doi.org/10.1007/s12108-017-9367-8

A'Hearn B, Baten J, Crayen D (2009) Quantifying quantitative literacy: Age heaping and the history of human capital. *J Eco Hist* 69: 783–808. https://doi.org/10.1017/S0022050709001120

A'Hearn, B, Delfino A, Nuvolari A (2021) Rethinking age heaping: a cautionary tale from nineteenth-century Italy. *Eco Hist Rev* 75: 111–137. https://doi.org/10.1111/ehr.13087

Bainame K, Letamo G (2015) Evaluation of Data Quality of the Botswana 2011 Population and Housing Census. *Botswana Notes and Records* 46.

Bairagi R, Aziz KMA, Chowdhury MK, et al. (1982) Age misstatement for young children in rural Bangladesh. *Demography* 19: 447–458. https://doi.org/10.2307/2061012

Bucur B (2016) Population Health in Interwar Romania Reflected in the Sociological School of Bucharest's Research and Publications, In: Marinescu, V., Mitu, B., *Valentina Marinescu and Bianca Mitu Health and the Media: Essays on the Effects of Mass Communication*, Jefferson [NC, USA]: McFarland & Co, 215–240.

Byerle D, Terera G (1981) Factors affecting reliability in age estimation in rural West Africa: A statistical analysis. *Popul Stud* 35: 455–465. https://doi.org/10.1080/00324728.1981.11878517

Caldwell JC (1966) Study of age misstatement among young children in Ghana. *Demography* 3: 477–490. https://doi.org/10.2307/2060173

Caldwell JC, Igun AA (1971) An experiment with census-type age enumeration in Nigeria. *Popul Stud* 25: 287–302. https://doi.org/10.1080/00324728.1971.10405804

Coale AJ, Kisker EE (1986) Mortality crossovers: Reality or bad data? *Popu stud* 40: 389–401. https://doi.org/10.1080/0032472031000142316

Coale AJ, Li S (1991) The effect of age misreporting in China on the calculation of mortality rates at very high ages. *Demography* 28: 293–301. https://doi.org/10.2307/2061281

Chemhaka GB, Odimegwu C, Zwane EN (2016) Is Swaziland census data suitable for fertility measurement? *Genus* 72. https://doi.org/10.1186/s41118-016-0010-2

Dechter AR, Preston SH (1991) Age misreporting and its effects on adult mortality estimates in Latin America. *Popul Bull UN*, 1–16.

Fajardo-González J, Attanasio L, Trang Ha J (2014) An Assessment of the Age Reporting in the IPUMS-I Microdata. Available from: https://paa2014.populationassociation.org/papers/140099.

Golopentia A, Grigorescu DC (1948) Populatia Republicii Populare Romane la 25 Ianuarie 1948. Rezultatele provizorii ale recensamantului. Extras din Probleme economice, nr 2. Institutul National de Statistica.

Jdanov DA, Jasilionis D, Soroko EL, et al. (2008) Beyond the Kannisto-Thatcher database on old age mortality: An assessment of data quality at advanced ages. Available from: https://www.demogr.mpg.de/papers/working/wp-2008-013.pdf.

Kannisto V (1999) Assessing the Information on Age at Death of Old Persons in National Vital Statistics, In: *Validation of Exceptional Longevity*, Odense: Odense University Press, 6: 240–249.

Lee MM, Zhang N (2017) Legibility and the informational foundations of state capacity. *J Polit* 79: 118–32. https://doi.org/10.1086/688053

Mihaescu C, Dumitrescu I, Mirica A (2018) Romania: un secol de istorie. Date statistice. Andrei Tudorel (coord). Bucuresti: Editura Institutului National de Statistica, 2018. Populatia, 12–48.

Mokyr J, Gráda CÓ (1982) Emigration and poverty in prefamine Ireland. *Explor Econ Hist* 19: 360–384. https://doi.org/10.1016/0014-4983(82)90008-0

Penina O, Jdanov D, Grigoriev P (2015) Producing reliable mortality estimates in the context of distorted population statistics: the case of Moldova. Available from: https://www.demogr.mpg.de/papers/working/wp-2015-011.pdf.

Pisica S, Murgescu B, Sora FA, et al. (2018) Romania: un secol de istorie. Date statistice. Andrei Tudorel (coord). Bucuresti: Editura Institutului National de Statistica, 2018. Forta de munca, 49–65.

Preston SH, Elo IT, PRESTON SH (1999) Effects of age misreporting on mortality estimates at older ages. *Popul Stud* 53: 165–177. https://doi.org/10.1080/00324720308075

Pripoaie R, Cretu CM, Turtureanu AG, et al. (2022) A Statistical Analysis of the Migration Process: A Case Study—Romania. *Sustainability* 14. https://doi.org/10.3390/su14052784

Sandu D (2018) Migratia temporara in strainatate (1990–2016), In: Gheţău, V., *Demografia Romaniei*, Editura Academiei Române, 245–278.

Shryock HS, Siegel JS (1975) *The methods and materials of demography*.

Simandan D (2019) Revisiting positionality and the thesis of situated knowledge. *Dialogues Hum Geogr* 9: 129–149. https://doi.org/10.1177/20438206198500

Simandan D (2010) Roads to perdition in the knowledge economy. *Environ Plan* 42: 1519–1520. https://doi.org/10.1068/a4324

Yi Z, Gu D (2008) Reliability of age reporting among the Chinese oldest-old in the CLHLS datasets. In: *Healthy longevity in China*, 61–78, Springer, Dordrecht.

United Nations (1955) Manual II: Methods of appraisal of quality of basic data for population estimates (United Nations Publications, Sales No. 56.XIII.2). Chapter 3: The Accuracy of Age and Sex Statistics.

United Nations, Department of Economic and Social Affairs Population Division. World Prospects 2022. Romania Demographic Profiles, Line Charts, 2022. Available from: https://population.un.org/wpp/Graphs/DemographicProfiles/Line/642.

AIMS Press