# Data Quality Checklist

Country: Australia Range of Years Covered: 1921-2016

Reported Submitted by: Magali Barbieri Date: February 25, 2019

Data Quality Charts at:

https://www.mortality.org/HMDWORKV6/AUS/CHECKS/HMDQC.html

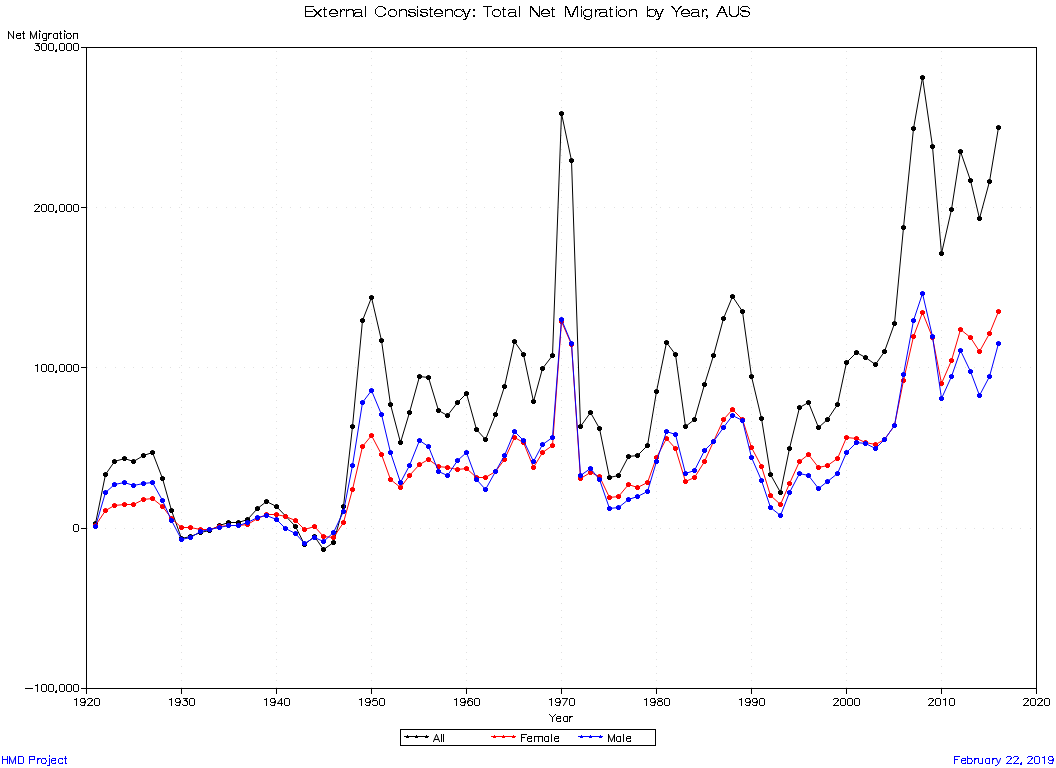
|  |  |  |
| --- | --- | --- |
| *Updated to include 2015-2016* | **Date**  **Completed** | **Comments** |
| **Checks for Input DB** | | |
| * *Internal Errors*: Check that rows and columns add up correctly | February 22, 2019 | There are small differences between births by sex (BBS) and births by month (BBM) due to difference in definitions: BBS are tabulated by year of registration while BBM refer to year of occurrence. |
| * *Gross Errors*: Note any corrections or problems in published numbers in the “Notes” file | February 22, 2019 | No gross errors were found |
| **Internal Consistency (InputDB vs. LexisDB)** | | |
| * Check outliers identified during procedures for splitting open age interval. [see Chart 4] | February 22, 2019 | No Outliers were detected |
| * Check deaths and population estimates in LexisDB against the raw data in the InputDB [see Charts 5-8] | February 22, 2019 | Deaths match. There are small differences in populations, especially in ages 80+, as expected. |
| **Internal Plausibility** |  |  |
| * Inspect plots of death rates by age and time for any unusual patterns [see Charts 9-10] | February 22, 2019 | Checked. |
| * Check for evidence of age heaping [see Chart 11] | February 22, 2019 | Age heaping appears to be a cohort phenomenon in Australia: cohorts born before 1900 heap strongly on 0s, males worse than females. Males also heap some on 5s. Not among the worst in the HMD, however, and this will not introduce measurable bias in e0, e60 or e80. |
| * Check implied migration for unexpected patterns [see Chart 12] | February 22, 2019 | See Note 1 below |
| **External Plausibility** |  |  |
| * Check total annual (implied) migration and compare with external estimates if available [see Chart 13] | February 22, 2019 | See Note 1 below |
| * Check e0, e65, and e80 compared to Sweden. Are estimates plausible? [see Chart 14] | February 22, 2019 | Estimates are plausible. |
| * *External Comparisons:* Compare e0 and q0 estimates to some external source   External source: ABS historical statistics and more recent ABS estimates  Range of years: 1921-2015 | February 22, 2019 | See Note 2 below |

|  |  |  |
| --- | --- | --- |
|  | **Date**  **Completed** | **Comments** |
| ***IF DATA WAS UPDATED*: Old vs. New** |  |  |
| * Check ratio of “new” to “old” death rates.Investigate any suspicious cases (e.g., ratio <0.95 or >1.05). [see Chart 15] | February 22, 2019 | Checked – expected differences due to the more detailed death counts (by Lexis triangle) used for 2012-2016 for the current update compared to the previous one + the substitution of births by year of occurrence to the previous births by year of registration for all years since 1975 (which has a small impact on the allocation of deaths below one between the two triangles and, thus, on the infant mortality rates) + to the updated population estimates for years 2012-2017. |
| * *If any big differences between old and new death rates*: Check ratio of deaths and ratio of population estimates. What is the source of the differences? [see Charts 16 & 17] | February 22, 2019 | See above. Nothing unexpected. |

# Additional Notes

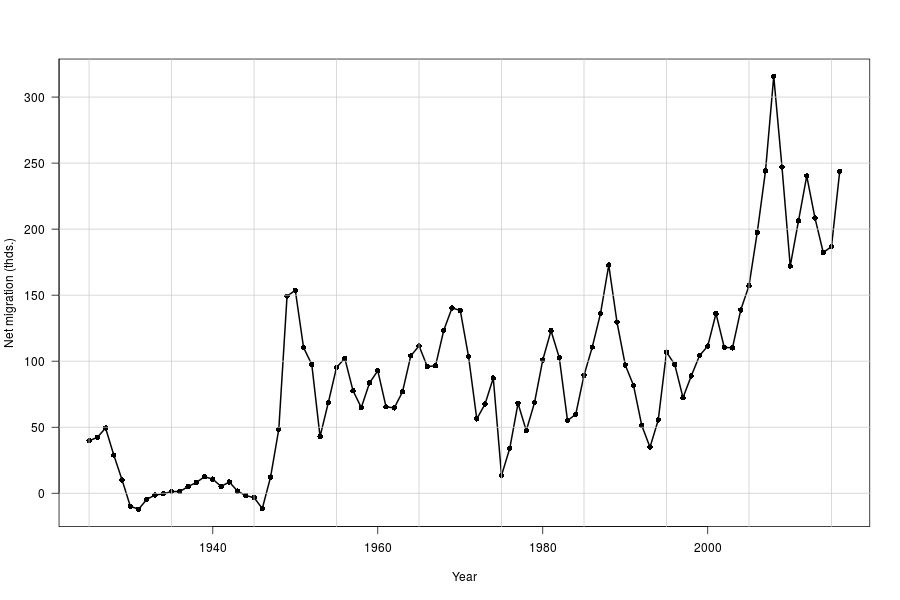
***NOTE 1 – External plausibility - implied migration***

Charts below show that implied migration calculated by HMD match official estimates, except for the peak in 1971, which is due to definitional and methodological changes. Starting in 1971, population estimates started to be based on de jure census counts adjusted for net undercount and to include Australians temporarily overseas (See B&D file).



The migration pattern continues to be very similar for the most recent years.

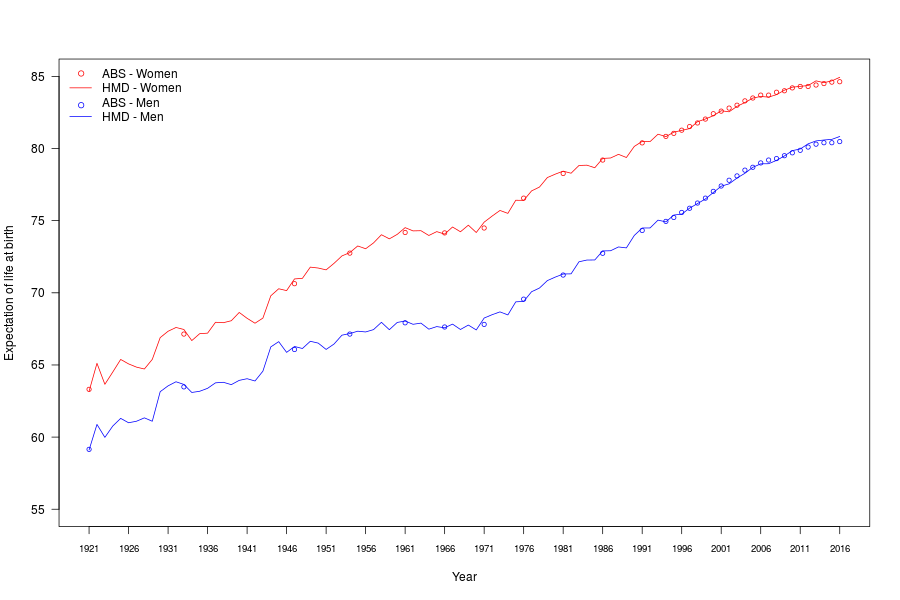
**Net migration, 1925 - 2016 – Australia (ABS)**



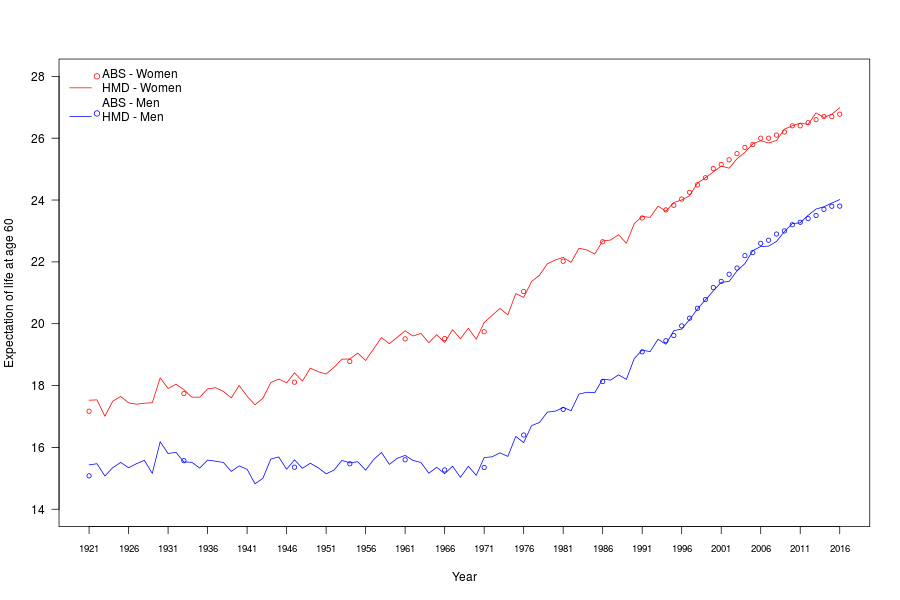
***NOTE 2 – External comparisons: ex and q0***

The charts below show the comparison between HMD estimates of life expectancy at ages 0, 60 and 80 with the official figure published by ABS. ABS publishes 3-year average life tables, which account for most of the differences in e0 and e65. For recent years, HMD estimates are slightly higher than ABS ones, particularly for males. The comparison for e80 shows a more consistent difference between the two estimates, probably due to differences in the methods used to estimate old age mortality. HMD estimates tend to be higher for years before 1980 and lower for the 1990s and 2000s decades but the differences has become negligible since 2001.

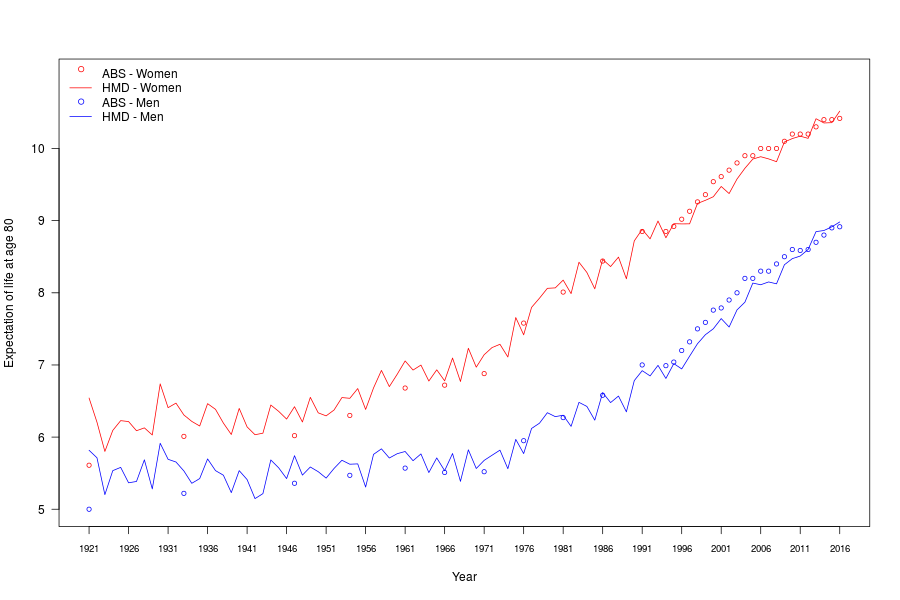
**Expectation of life at birth by calendar year and by sex, 1921-2016**



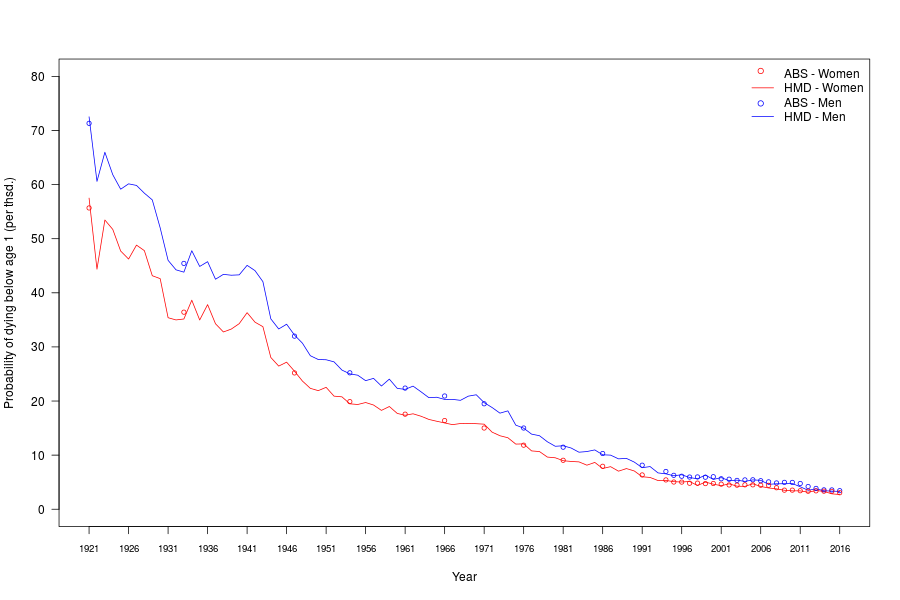
**Expectation of life at age 60 years by calendar year and by sex, 1921-2016**



**Expectation of life at age 80 years by calendar year and by sex, 1921-2016**



HMD infant mortality estimates also match ABS estimates, as shown by the figures below.



# Notes from previous updates

***NOTE 3 – Comparisons between estimates with and without Lexis triangles (2016 update)***

In the 2014 update, death counts by year of occurrence from 1964 to 2011, by sex, age and birth cohort, were purchased from ABS and replaced the previous series by year of registration.

In June of 2016, new death count data became available for free on the ABS website. These data are less detailed than the previously purchased data as the death counts are provided by single years of age but not by year of birth [MB: this has become obsolete as we have just bought the Lexis triangle deaths for years 2012-2016 from ABS but since this analysis might become relevant again later on, I leave it here]. We compared the results of lifetable estimations using the two sets of data for years 2004-2011 to figure out whether the difference warranted spending a non-negligible amount to purchase the more detailed data. The charts below show the results of this comparison for life expectancies at ages 0, 65 and 80 for the whole period 2004-2011 and for the Mx by sex and age for the years 2004, 2008 and 2011. These graphs indicate that reconstructing Lexis triangle counts from single-year-of-age data does not significantly modifies life table estimates compared to using the detailed data. We thus decided to update the AUS HMD series using the free data (death counts by single years of age) for years 2012-2014.













# Notes from Tim (2014)

1. (1911-1920 omitted due to pending WWI adjustments/analysis) The Spanish flu came to Australia in 1919, and caused ca 11,000 deaths in 1919 and 1920 (http://www.migrationheritage.nsw.gov.au/objects-through-time/essays/1918-1939/), a lower and slightly later impact than elsewhere in the world. If the ABS description of the timing is correct, then it appears these deaths were registered in the correct years. It could also be the case that the ABS is basing their statement on the same registration data that was late registered, in which case most of the excess deaths would have actually been in 1919 (doubtless, because this is when the soldiers returned). These data cannot answer the question.
2. 1968-1971 migration event? The Australian Department of Immigration and Border Protection publishes annual migration statistics in a spreadsheet here: <https://www.immi.gov.au/media/statistics/historical-migration-stats.htm> . These show elevated numbers of immigrants in years 1968-1970, but lack out-migration to compare. The apparent spike in HMD data is, however in 1970-1972. Another different series on historical migration is found here: <http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/3105.0.65.0012008?OpenDocument>

In this spreadsheet if we isolate 'permanent' movements and take the difference between in and out migration, there is indeed a spike in net inflows in the years 1969 and 1970, still a year earlier than the HMD anomaly.

Mila suggests below that there was a change in universe at the 1971 census, which included Australians temporarily abroad and which was better adjusted for undercount. This count certainly account for the bulk of the jump, especially the broad age pattern, but the entire apparent influx is likely a mix of both actual migration and universe change.

1. There were major changes in inputs to the Australia series over the entire period covered. These include:
   1. Death counts in single ages for years 1911-1963 (1911-1920 captured, but not included in series), and replace previous counts by 5-year age groups for years 1921-1963. This is original data capture work from scanned pdfs of the original statistical series. The new single age counts go up to the highest observed age at death until 1941, and until 100+ until 1963. The previous 5-year counts had an open age group of 85+. The new data likely improve our old-age estimates. These data are by year of registration, as was this portion of the old HMD series. The switch to single ages introduces acceptable age heaping in these years for cohorts born before 1900 as they pass through the data. As before, the effect of using registration year data is on average to smooth the series. This smoothing is slight, however, as 90%-95% of deaths are typically registered in the year they occurred, and it only introduces a bias in the event of a strong year-to-year trend and/or change in seasonality or a mortality event timed at the end of a year (see comparison below).

The difference between old and new HMD estimates is trivial.

* 1. Death counts for years 1964-1989 have been switched from registration to occurrence, but are otherwise in the same shape as before (age-period, single ages up to the highest recorded age). These data came from a file that was originally sent to the HMD in a 2001 email from the ABS, but it was decided by the CS at that time to keep the whole series by year of registration. Probably this decision was not due to the joining of 1963 registration with 1964 occurrence, but rather because it is easier to get new data from the ABS by year of registration, so this decision avoided flopping back and forth between registration and occurrence data. Mila mentioned in the previous B&D that the decision was also in order to make HMD estimates comparable with other external sources, but I disagree that this should be a motivation. Based on both principle and quality of the recent part of the HMD series, I have opted to reverse this policy, and to use occurrence year data from the earliest availability, which is 1964. Prior to 1964 there is no occurrence data available.

This new input entails changes in death rates for ages less than 50 and greater than 100, and only minor changes between ages 50 and 100. The most common reasons for a late registration are because of a coroner or police investigation of a death, which seems most likely for these younger ages, and this is a plausible explanation for why rates change in younger ages more than in older ages. I do not know why ages over 100 would have a different registration pattern, but I take the new occurrence estimates as superior. Figure 2 shows the percent difference in counts (100 \* (Occ-Reg)/Occ ) as Lexis surface, for 1964 to 2010.

There is also an apparent increase in rates for a wide range of ages in 1984, and I can only posit that this is due to having more late-registered deaths than other years, but I don't know why this would happen in 1984 especially. The effect was to decrease e0 in 1984 by .36 years for females and 0.39 years for males. This is greater than the typical difference in e0 estimates for years that switched from registration to occurrence. On inspecting the overall trend in e0 in the old and new series, both estimates are plausible, but the new estimate entails a more consistent sign of the difference at this point in the series (the old estimate was an uptick). I will not over-analyze this deviation, as it is like reading tea leaves, but I do think the new estimate is probably more reliable for 1984. To get an overall sense of the deviations in this period, see Figure 1 (gray highlighted area):

Figure 1(blue dots male, red dots female)

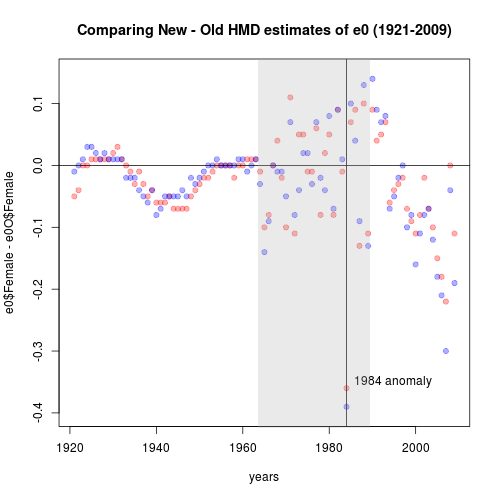
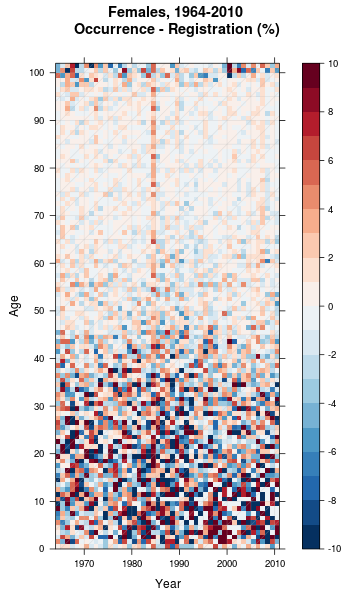
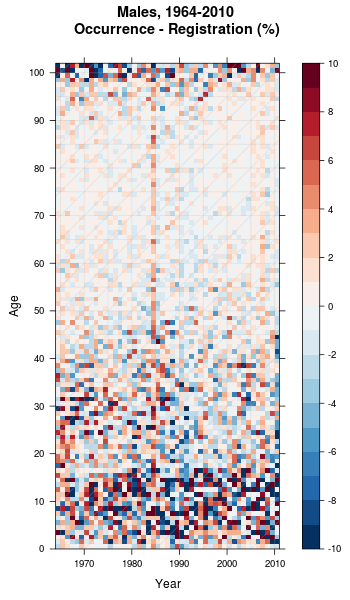


Figure 2: Percent difference in registration versus occurrence year by Lexis squares, 1964-2010.

Absolute changes in e0 between 1964 and 1989 are mostly random noise, with no clear trend, and only 1984 as an outlier (discussed above).

* 1. Years 1990-2009 were replaced with newly tabulated data by ABS (and years 2010-2011 added). The new data are also year of occurrence data, but these only reach age 100+. For this reason, age-period data is taken from the old series for ages 100-maximum observed age for years 1990-1999, while years 2000-present have an open age group of 100+ per the new ABS policy. The main difference between old and new data for these years is the switch from registration to occurrence data, but another important aspect of these data is that *recent* late registrations are now incorporated into the series. The change in e0 due to the change in numerators is actually random noise (see Figure 2), as it is for years 1964-1989 (except 1984), and the downward correction is due to a change in denominators, described in the following note.
  2. In this range of years there was also an important change in denominators, as new intercensal estimates were carried out by the ABS between the 1991 and 2001 censuses and again between the 2001 and 2011 censuses. The first intercensal period involved a change in methodology by the ABS, and the second intercensal period involved a switch from post to inter-censal estimates, which decreased denominators, especially in the most recent period. This of course had the effect of increasing rates, and decreasing e0, which is visible in the recent section of Figure 1. These denominators will not change in forthcoming HMD updates, unless ABS decides to redo a longer series of population counts that we deem of higher quality. The current change constitutes improvement.

# Notes from Mila (2012):

# Comparisons with external sources

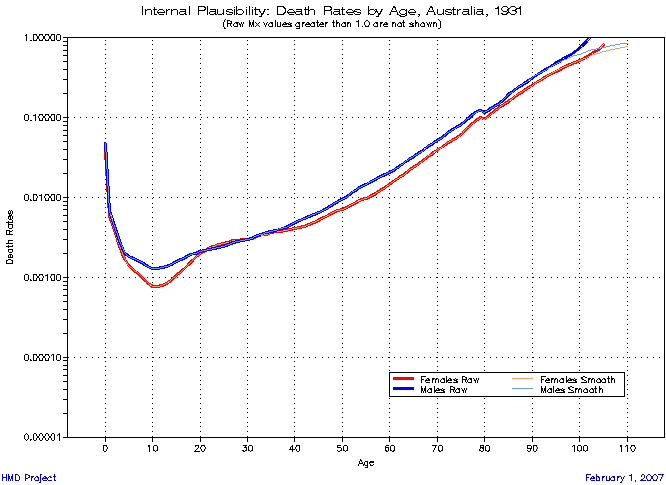
HMD mortality estimates have been compared with the following independently-produced mortality estimates. External mortality data have been compiled from the official statistical publications. Complete reference information on sources of the external mortality estimates is included in the internal database.

|  |  |
| --- | --- |
| **type** | **coverage** |
| "Life Tables" | "1881-1890, 1891-1900, 1901-1910, 1920-1922, 1932-1934, 1946-1948, 1953-1955, 1960-1962, 1965-1967, 1970-1972, 1975-1977, 1980-1982, 1985-1987, 1990-1992, 1994-1996, 1995-1997, 1996-1998, 1997-1999, 1998-2000, 1999-2001, 2000-2002, 2001-2003, 2002-2004, 2003-2005, 2005-2009" |
| "Life Expectancy" | "1932, 1946, 1953, 1960, 1965, 1970, 1975, 1980, 1990, 1994-1997, 2001-2003, 2005, 2006, 2007, 2008, 2009" |
| "IMR" | "1881, 1891, 1901, 1920, 1932, 1946, 1953, 1960, 1965, 1970, 1975, 1980, 1985, 1990, 1994-2003, 2008" |

*[Run dqrexternalavail.m and update the in-text tables with its output]*

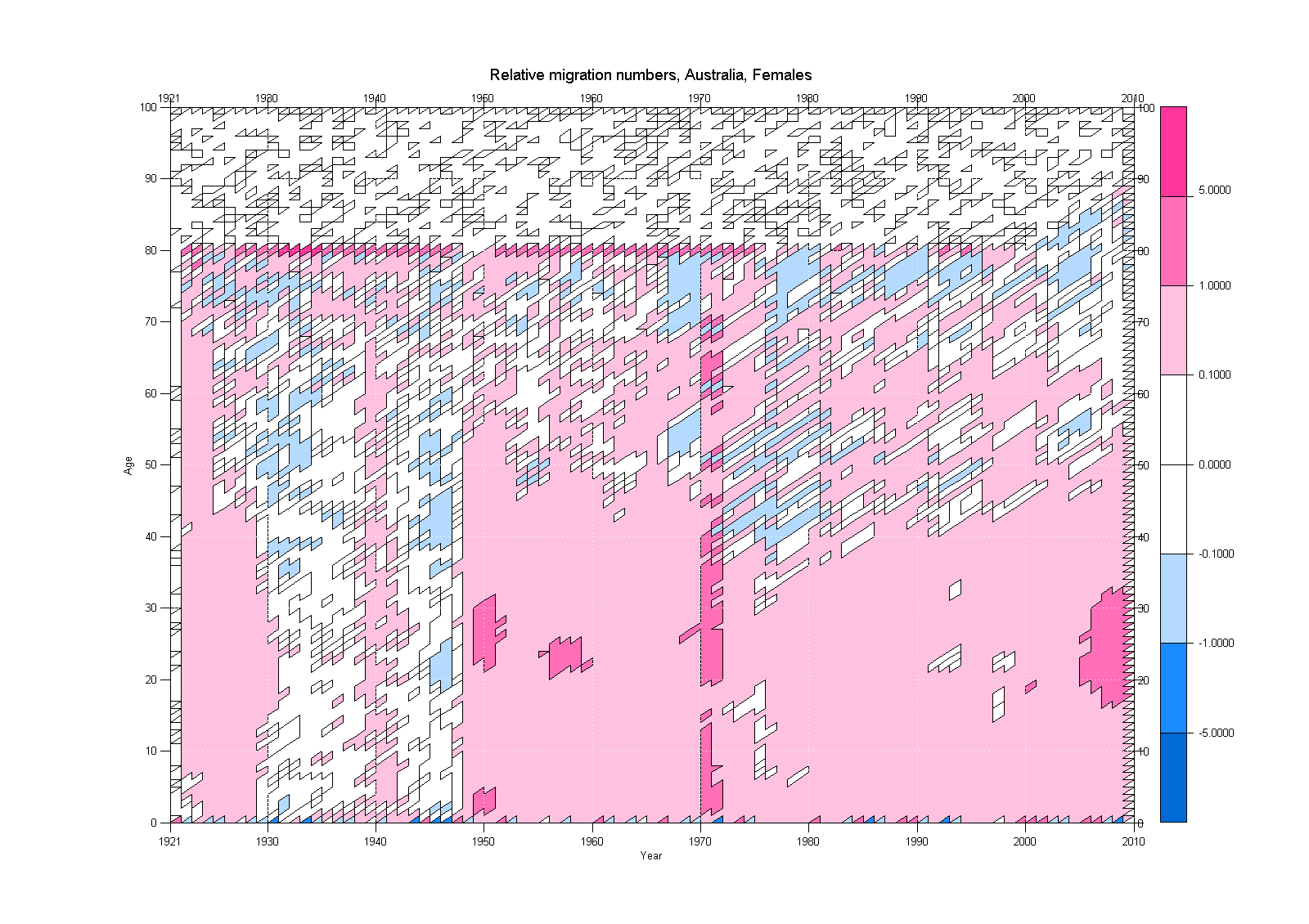
## Jump in death rates at age 80

For several years, usually in the beginning of the observation period, there is a jump in death rates at age 80. At this age extinct cohort population joins official population estimates.



## Implausible migration pattern

The horizontal migration line is due to joining of extinct cohort population and official population estimates at older ages. The vertical migration line is due to changes in population estimation procedure.



# Custom Data Quality Checks

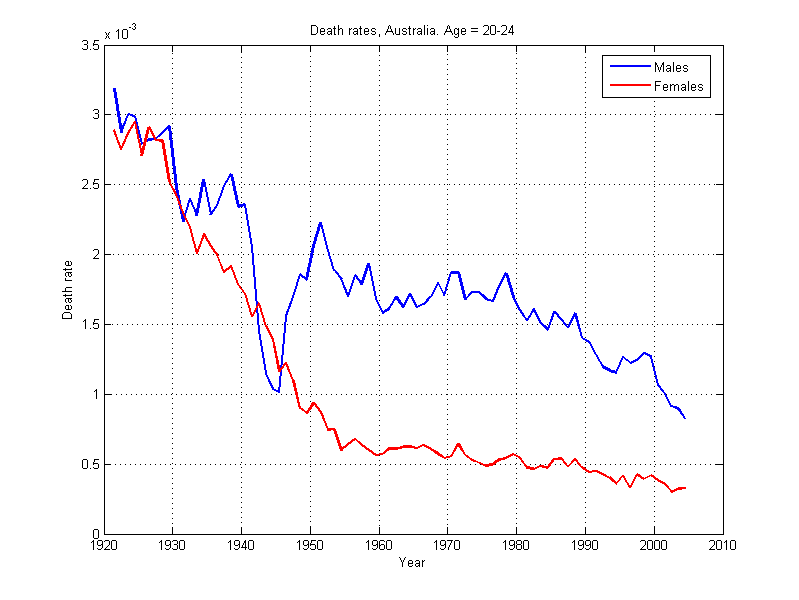
In addition to the UCB HMD data quality checks I am also conducting my own data quality checks at CUNY which are different from the data quality checks computed at Berkeley. The results are available in the form of HTML report and they are stored in the CHECKS folder for each population. There is some overlap between the two procedures, and there are some additional checks carried out. In particular we are checking for

1. consistency between cohort and period life tables;
2. published and HMD estimates of life expectancy at birth;
3. published and HMD estimates of IMR;
4. published and HMD period and cohort life tables.

This report is available on request. Sections below highlight possible problems with the data.

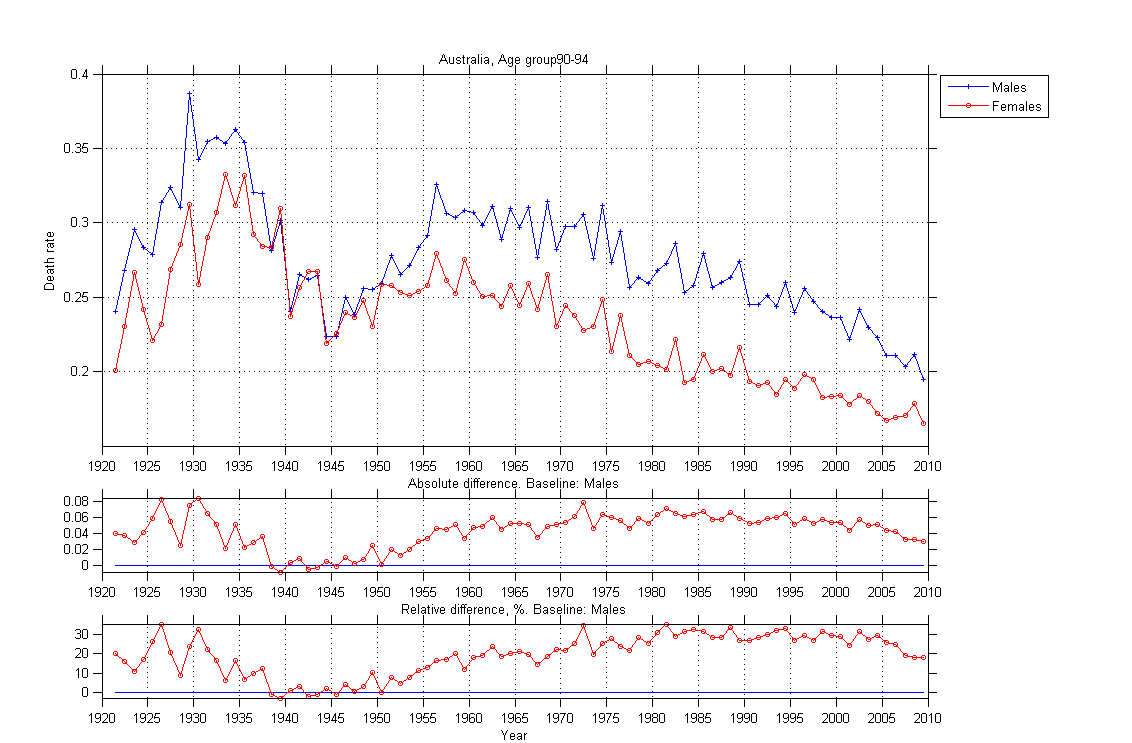
## Death rates during WWII

Due to inconsistencies between population and death series, death rates for males 20-24 years old drop during WWII. This problem requires further investigation. It’s on my To do list.



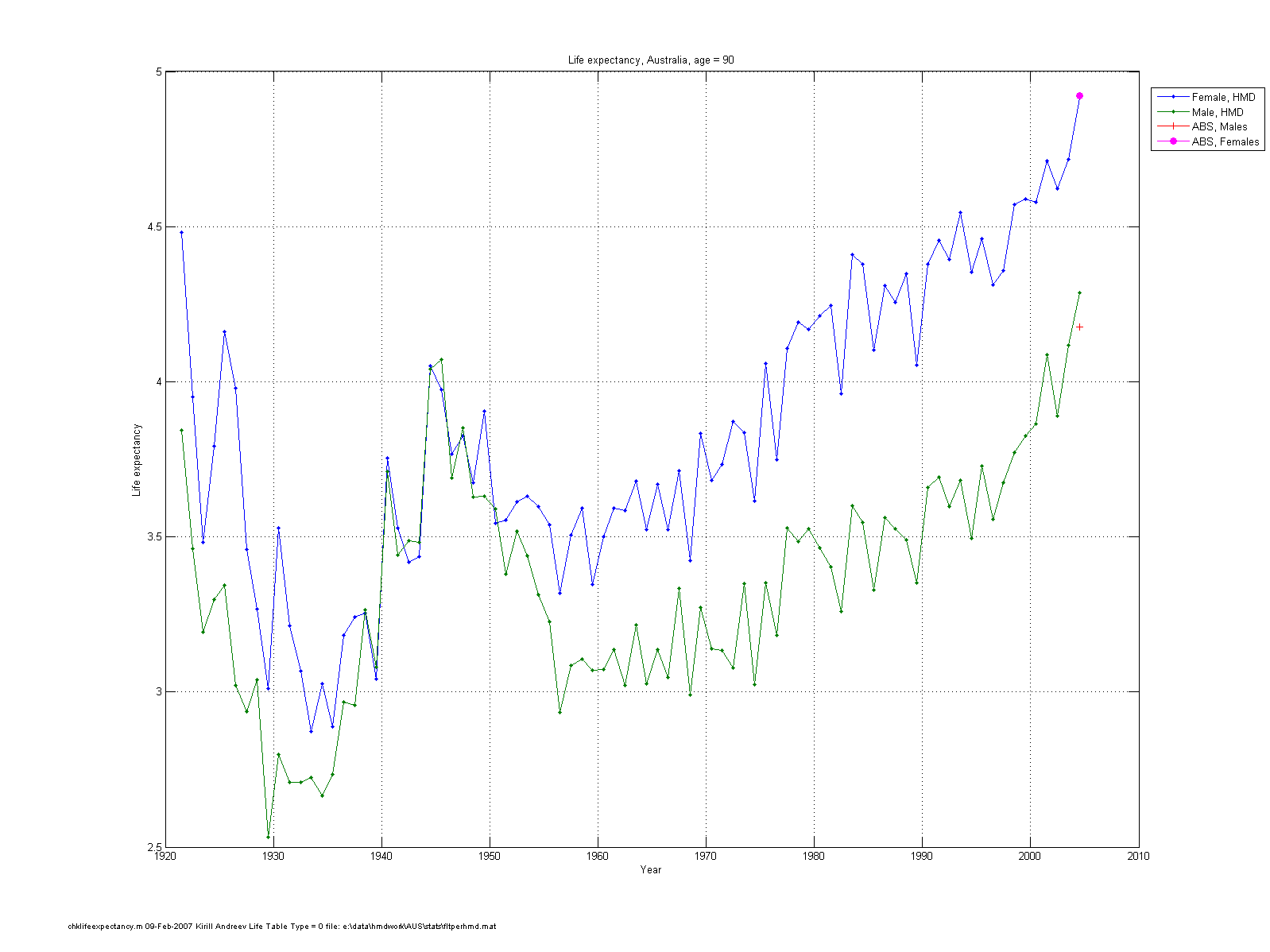
## Death rates at ages 90 and over, 1920-1960

Death rates at these ages and years exhibit implausible pattern likely attributed to the data quality problems.



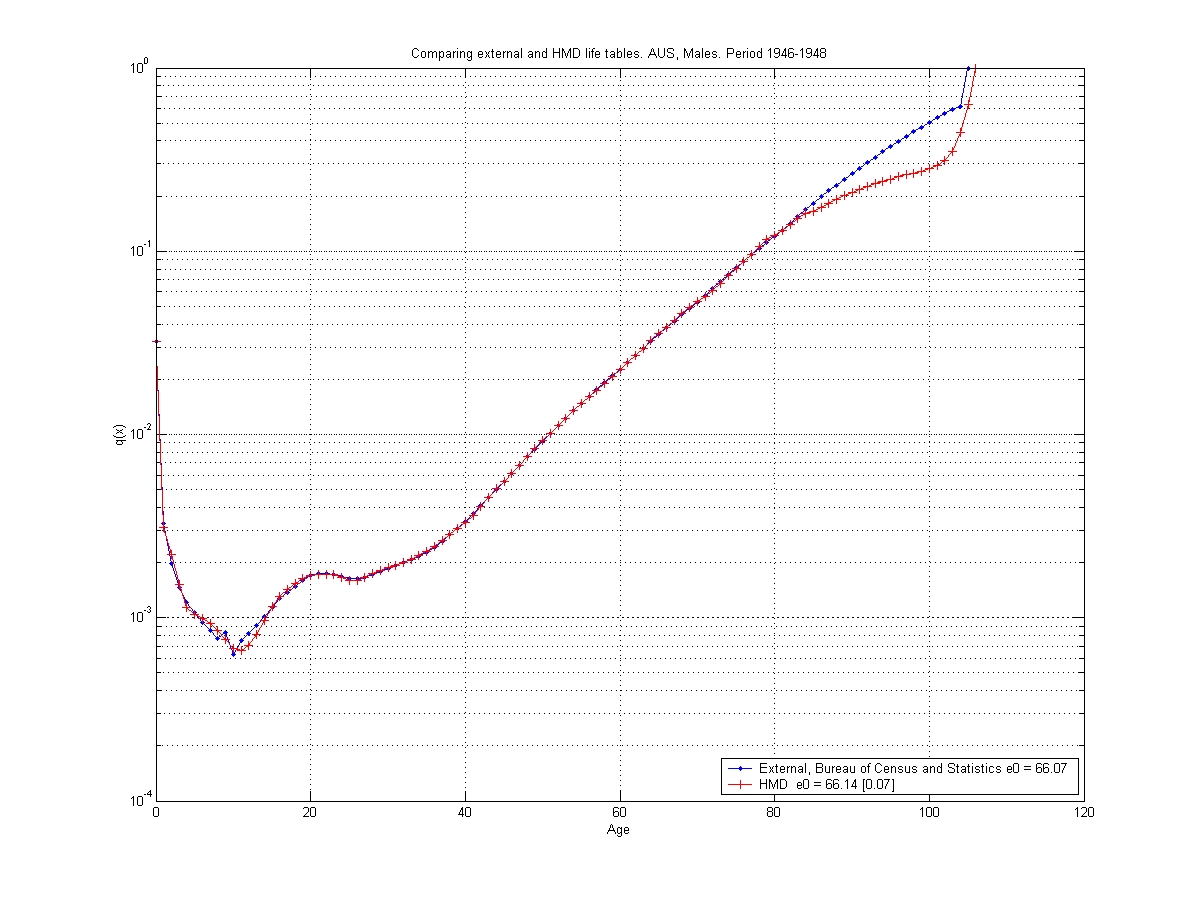
## Life expectancy at age 90 and 100, 1920-1960

Life expectancy trends are highly irregular because of the data problems. We don’t have data by single age over 90. We have only aggregated data which have been distributed by HMD procedures.



## Comparisons with official life tables

There are some disagreements between HMD death rates and official Australian life tables. For earlier periods 1946-1948 HMD death rates at ages above 80 are lower than death rates in the official Australian life tables. The Australian statisticians might have adjusted the actual rates due to concerns about age misreporting or small population sizes but there is no information on the methods used to produced official estimates so we don’t know for sure. This figure illustrates the difference.



In 1995-1997 see plot below the pattern is reverse: official death rates are below the HMD rates, again there has obviously been some adjustment carried out by the Australian National Statistics Office that would produce a systematic bias but there is no possibility to confirm this hypothesis considering the lack of documentation.

