# Traditional and newly emerging data quality problems in countries with functioning vital statistics: experience of the Human Mortality Database

*Dmitri A. Jdanov*
*Domantas Jasilionis*
*Vladimir M. Shkolnikov*
on behalf of
the HMD team

**United Nations Expert Group Meeting on the methodology and lessons learned to evaluate the completeness and quality of vital statistics data from civil registration**

*New York, November 3-4, 2016*

MAX PLANCK INSTITUTE FOR DEMOGRAPHIC RESEARCH
MAX-PLANCK-INSTITUT FÜR DEMOGRAFISCHE FORSCHUNG

NES New Economic School

# The Human Mortality Database

- Joint project of the Department of Demography at the University of California at Berkeley (USA) and the Max Planck Institute for Demographic Research in Rostock (Germany)
- Work began in autumn 2000, launched online in May 2002 with 17 country series
- Now: a leading data resource on mortality in developed countries.
- Includes 38 countries and 8 regions, 30,000+ users

***Main advantages:***

- Comparability across time and space
- Continuous, long-term series without gaps or ruptures
- Data by age, year, cohort, in age-time formats 1x1, 5x1, 1x5, 5x5 etc.
- Detailed documentation on origins and quality of the data

*However, one of the main principles of the HMD is to include countries with reliable population statistics, especially requiring a full coverage of registration of vital events.*

# Reliable population data

**First questions:**

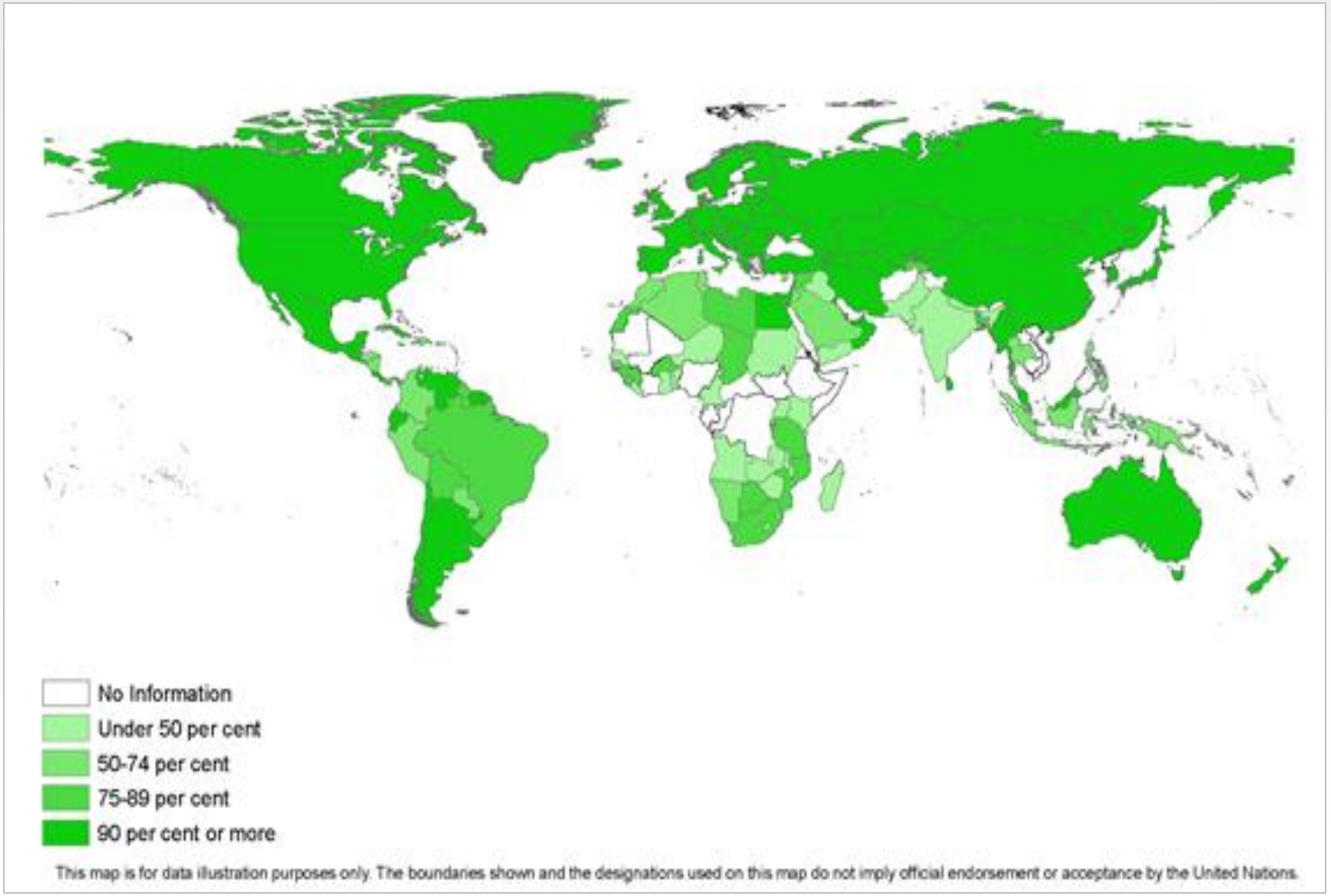
*Is there civil registration system and reliable vital statistics?*
*Is there reliable population estimates?*
*Is it possible to get all these data?*

**Preliminary quality checks in the HMD:**

- Accuracy: coverage, completeness, proportion of missing data
- Availability and relevance: access to detailed enough tabular data
- Comparability across space and time

For example, using UN and WHO sources a naïve user might conclude that data for Moldova, Chile, Costa-Rica are O.K. (last two censuses – complete, death and births - coverage >=90%).
This might be not correct according to the HMD criteria.

# UN assessment of coverage by death registration (Dec 2014)




Source: UN Population Division
(http://unstats.un.org/unsd/demographic/CRVS/CR_coverage.htm)

# Censuses and assessment of the population denominator

# Censuses and inter-censal population estimates

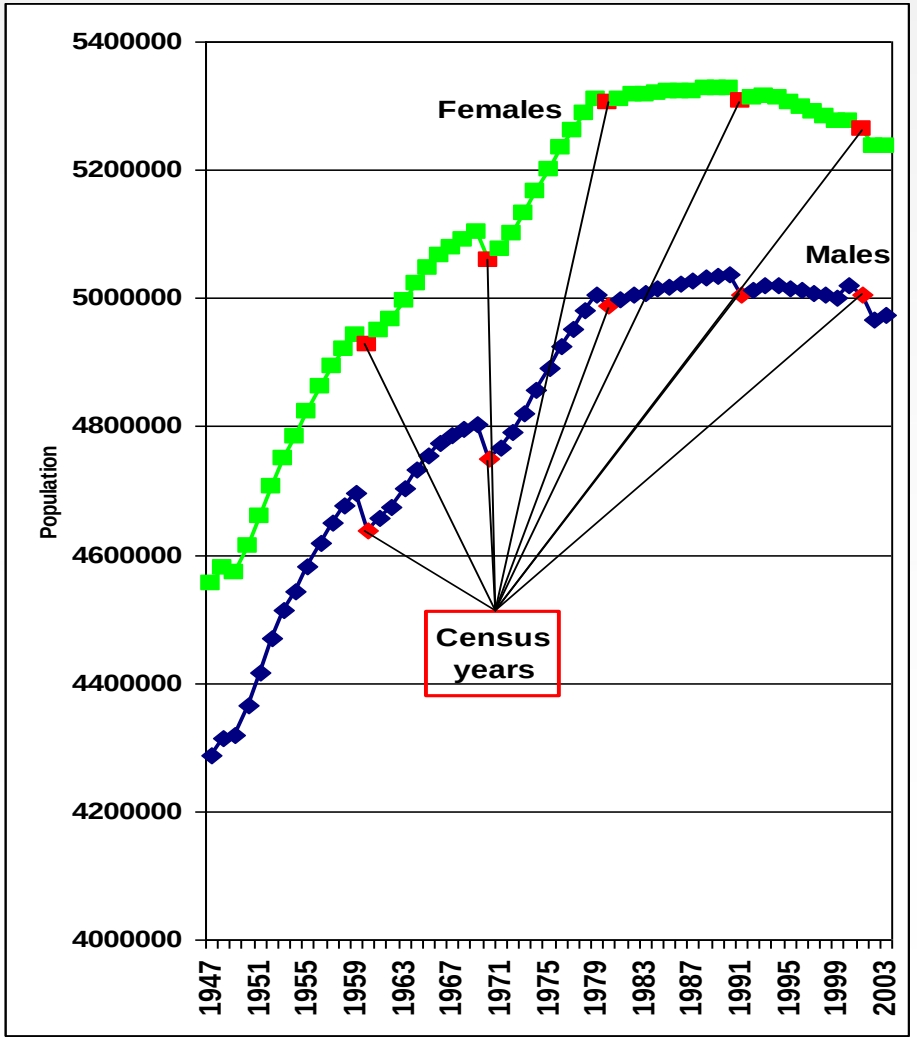
Assuming good quality of census data:

After a new census the post-censal population estimates should be replaced by inter-censal estimates (backward from this census). Four components:

- Census counts
- Death counts
- Births
- Migration

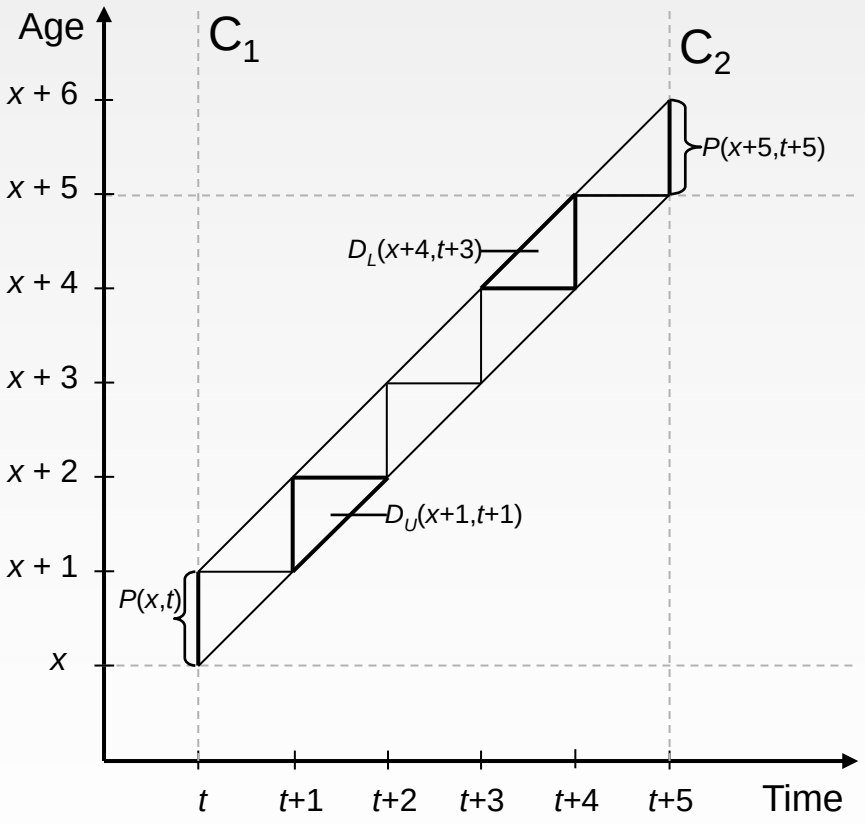
Developed countries with high quality vital registration system which do NOT produce inter-censal estimates: Germany, Italy, Czech Republic, ….

*Czech Republic, Official Population Estimates as of December 31*



# Inter-censal population estimates: the HMD approach

The standard HMD methodology (Wilmoth et al., 2007) for the cases when inter-censal population estimates are either not available or unreliable is based on the assumption of *uniform distribution* of migration across the entire inter-censal period. This assumption works well in many conventional situations, but may be violated in the case of special events (e.g. the collapse of the USSR and abrupt social-economic changes in Eastern Europe, the EU enlargement in 2004, the financial crisis in 2008-2009).

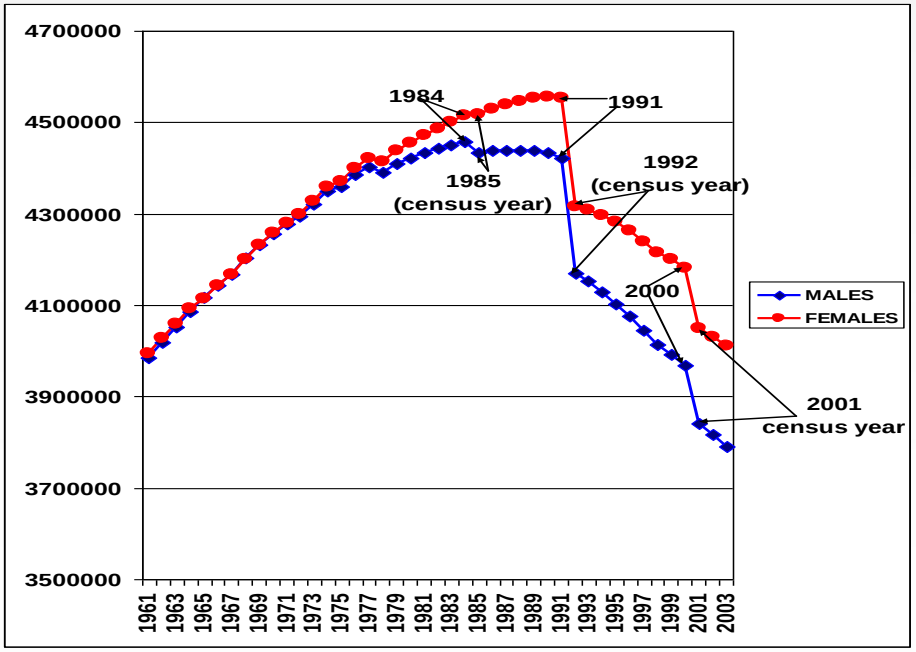


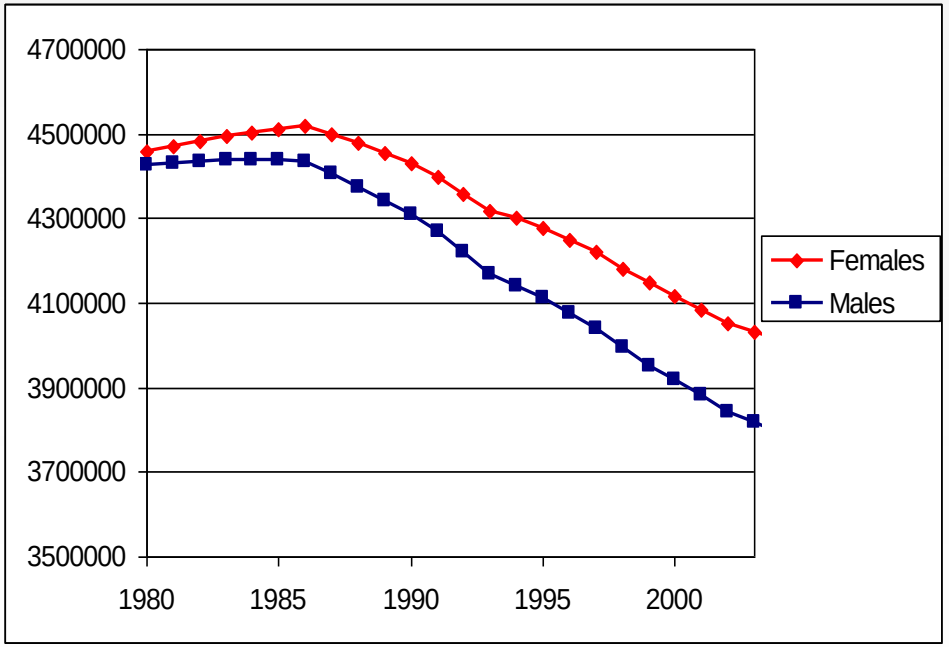

$$\Delta_x = C_2(x+5) - C_1(x) + \sum_{i=0}^{4}\left[D_U(x+i,t+i) + D_L(x+i+1,t+i)\right]$$

$$P(x+n,t+n) = C_1(x) - \sum_{i=0}^{n-1}\left[D_U(x+i,t+i) + D_L(x+i+1,t+i)\right] + \frac{n}{5}\Delta_x$$

# Bulgaria: correction of population data (inter-censal estimates)

The standard HMD inter-censal method is not applicable to the period 1985-1992 because of an irregular pattern of out-migration. In 1985-8, international migration was very restricted in Bulgaria. After the collapse of communism in 1989 - mass emigration (mostly of the Turkish minority) over the next several years.




**Trends in the total number of males and females. Bulgaria, 1961-2003. Official population estimates (left) and HMD data (right)**. Source: Jasilionis D., Jdanov D.A. Human Mortality Database: Background and Documentation for Bulgaria

HMD Solution: official population estimates were used for 1985-8, but new population estimates were calculated for the latter period. The year 1988 was treated as a "pseudo-census point" as the beginning of the inter-censal interval.

# Germany: three decades between censuses

Before the 2011 census, East Germany had a census 30 years ago and West Germany - 24 years ago. Whereas before the 2011 census Germany's population was estimated to be 81.7 million, the census corrected this down to 80.2 millions, a difference of 1.5 million people (~ 1.8%). *The statistical office of Germany decided not to produce adjusted inter-censal population estimates by age.*

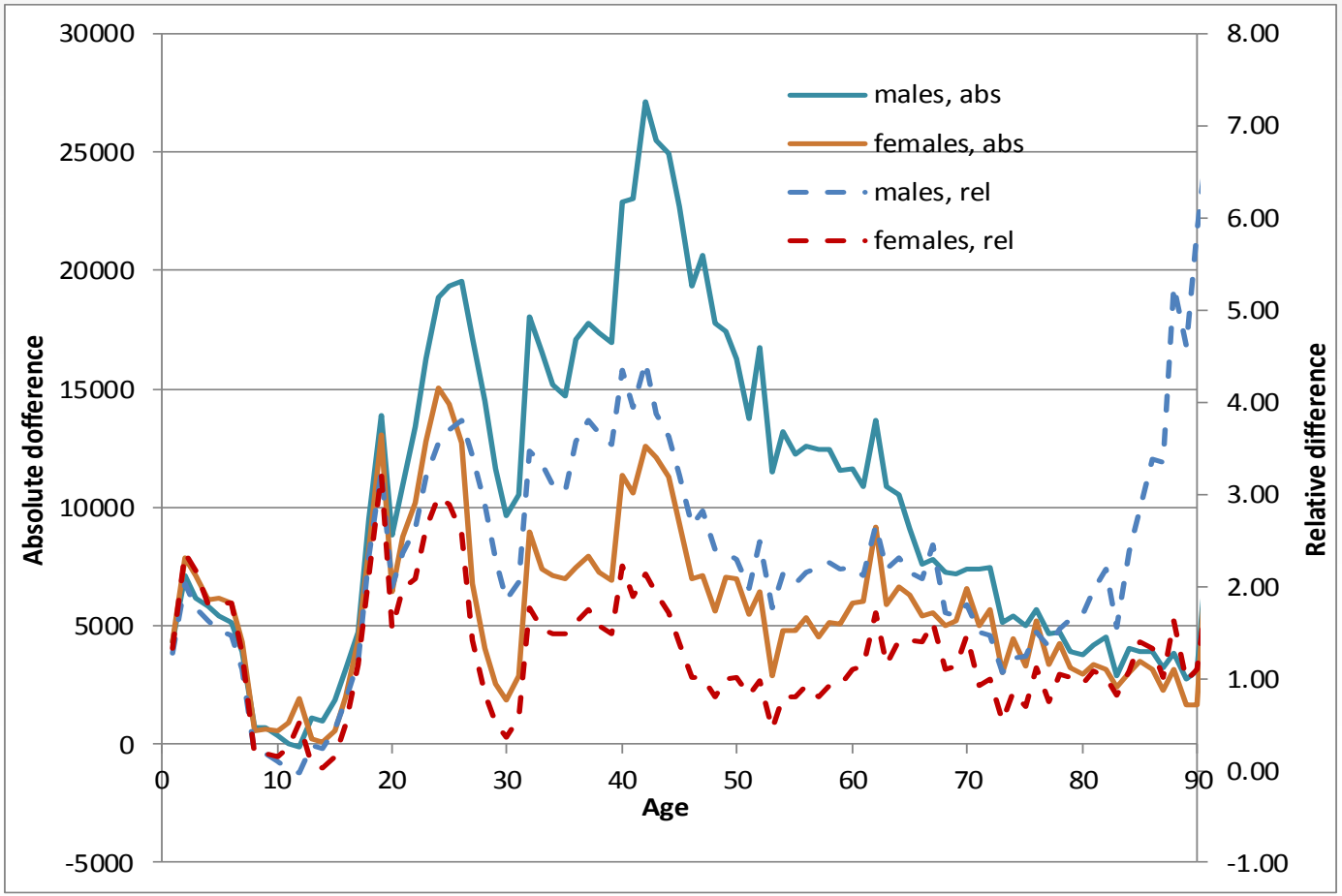
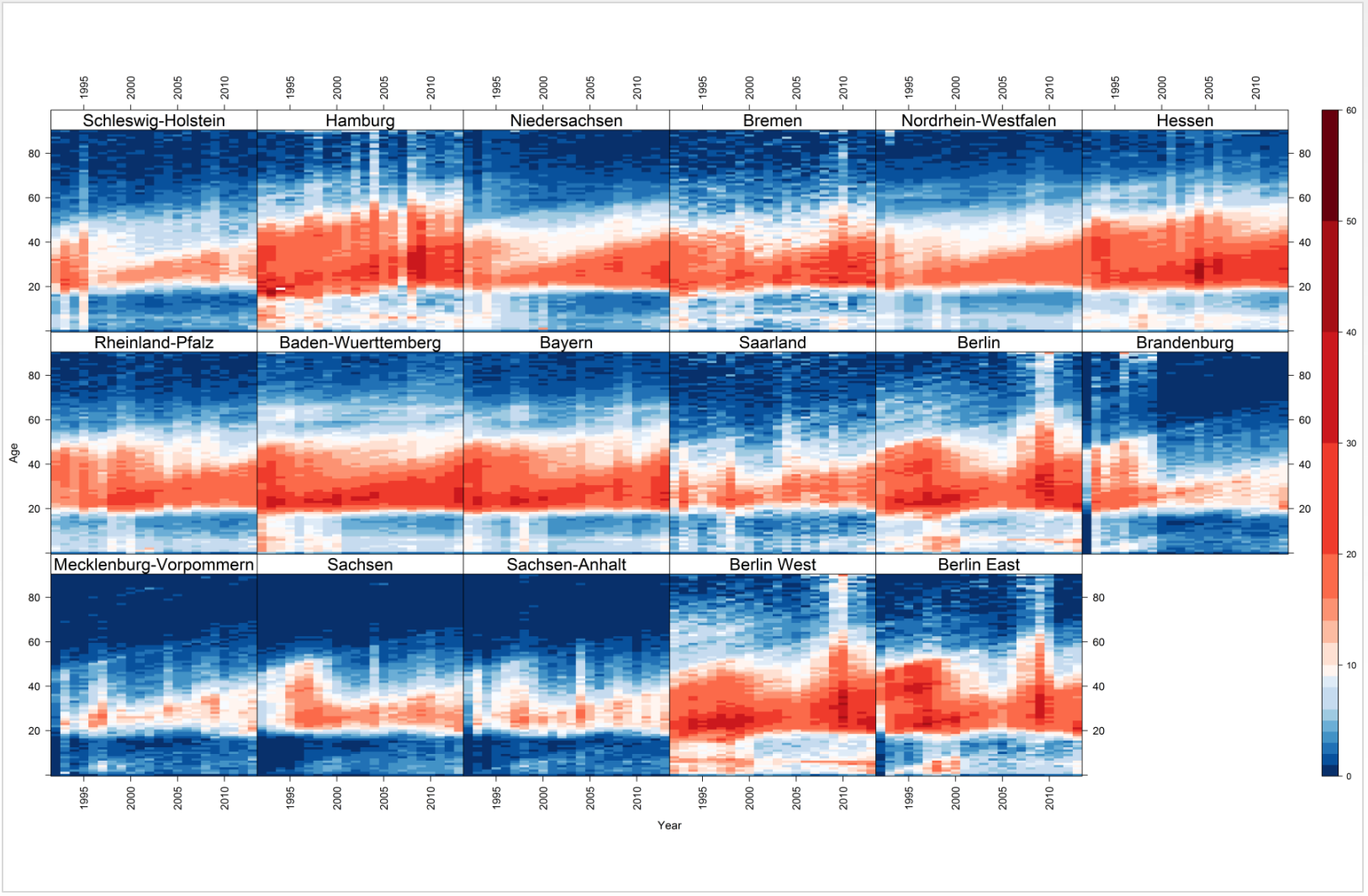



**Figure: the difference between current population estimates and the census counts of 2011**

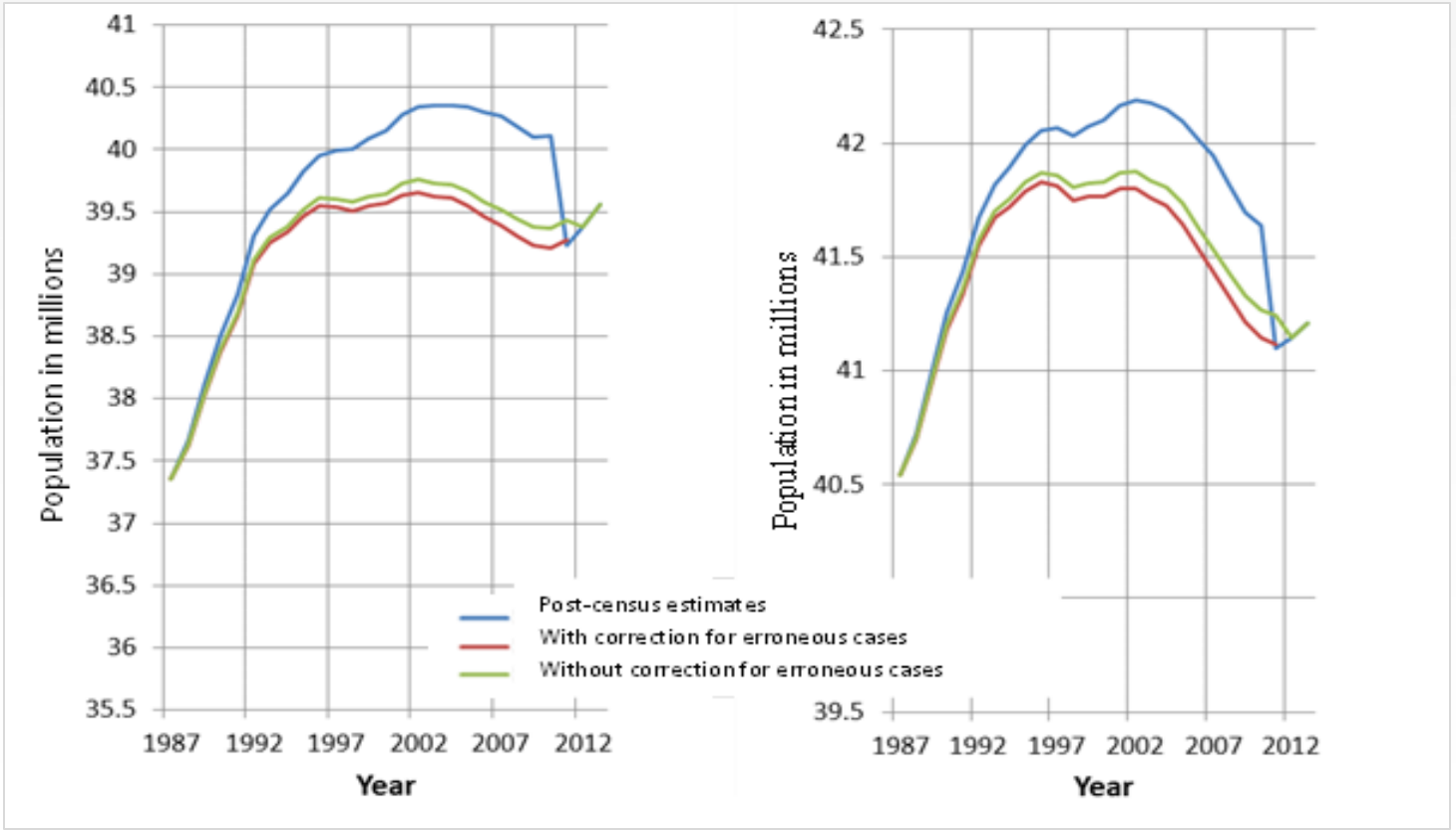
# External outmigration intensities by age for the German lands, males




In the 2000s statistical offices of many German lands have made efforts to eliminate from their populations non-existent residents (erroneous cases). Using information from local tax offices they removed erroneous cases by creating external outmigration events in the year of “cleaning”.

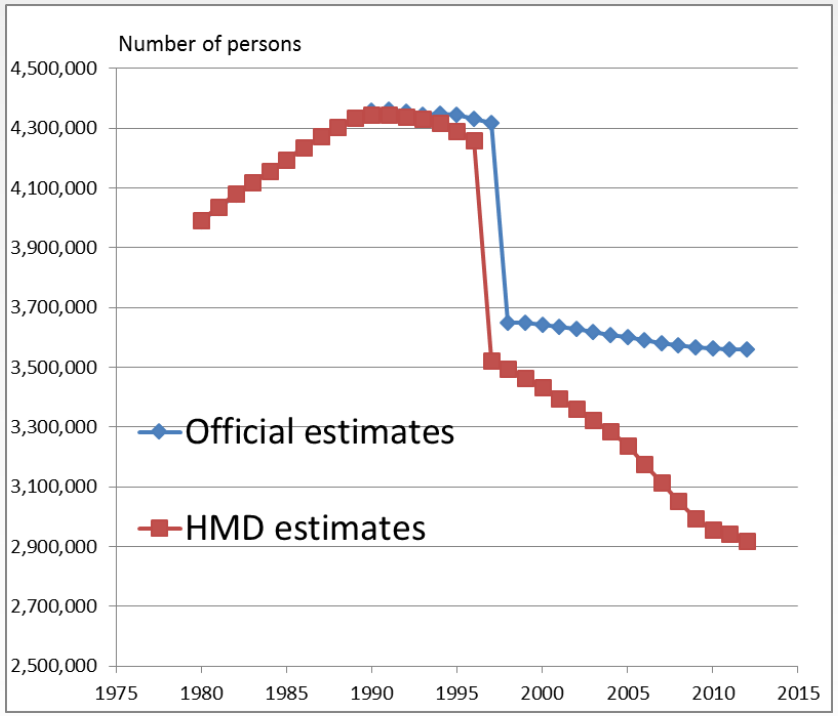
# The HMD inter-censal estimates for Gerrmany

1) Using additional migration data and cubic spline interpolation for migration trends across cohorts we removed the population changes due to the earlier “cleaning” by the statistical offices.
2) We distributed the accumulated error (not the net migration!) uniformly over the adjustment period of 24 years (30 years for East German lands):



# Changeable population definitions across time

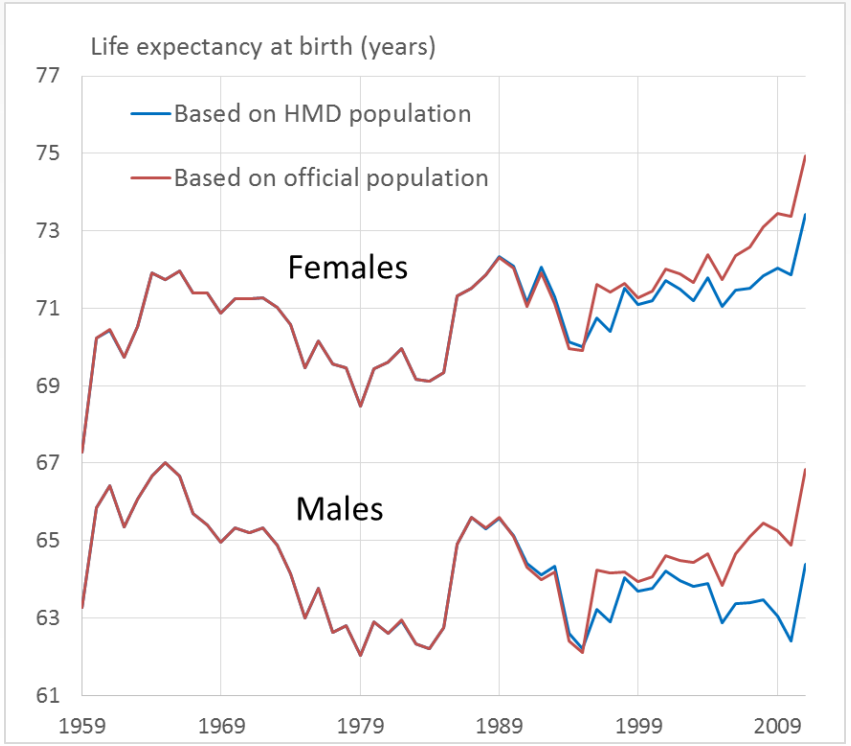
# Numerator-denominator bias: case of Moldova




* Since 1998 official population counts do not include Transnistria region

**The problem:** systematic bias (deaths and births refer to the *de facto* population, (i.e. occurred within the country, while population estimates also include long-term emigrants - Moldavian citizens living abroad). Results in under-estimation of mortality and fertility.

**The solution**: population estimates were corrected using data on border crossings and additional data collected at the census of 2004




Source: Penina, Jdanov, Grigoriev (2015)

# Changes in the definition of population: Poland

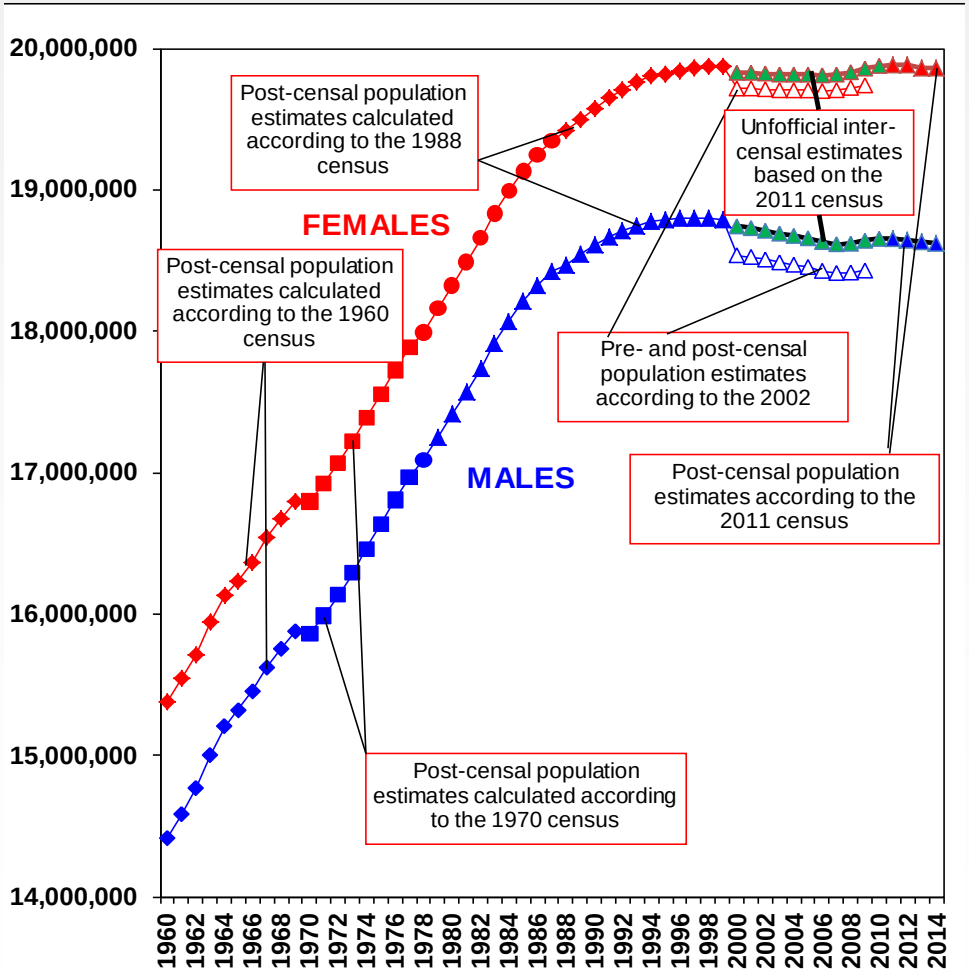



Figure: Official and adjusted (Tymicki et al. , 2015) estimates of population of Poland

In the 2000s, Poland faced a massive out-migration that followed the EU enlargement of 2004. It was expected that the population counts will be corrected downward after the next population census of 2011. But Statistics Poland has unexpectedly decided to change the official definition of the population status from the permanently resident (acting in 2010 and earlier) to the usually resident (from 2011 onward). *Statistics Poland did not re-estimate age-specific population counts back to previous census.* Due to irregular migration pattern the standard HMD inter-censal method for reconstruction of annual population estimates is not applicable.

# Change in the definition of ethnicity: New Zealand Māori

For New Zealand, HMD has separate data series for non-Māori and Māori populations. Change in definition of Māori in the census of 1991 from the one based on ethnicity of parents to the one based on self-identification. The new definition caused a jump in Māori population, but the death counts were not corrected simultaneously. The respective change in definition of ethnicity for death and birth was introduced only in September 1995.

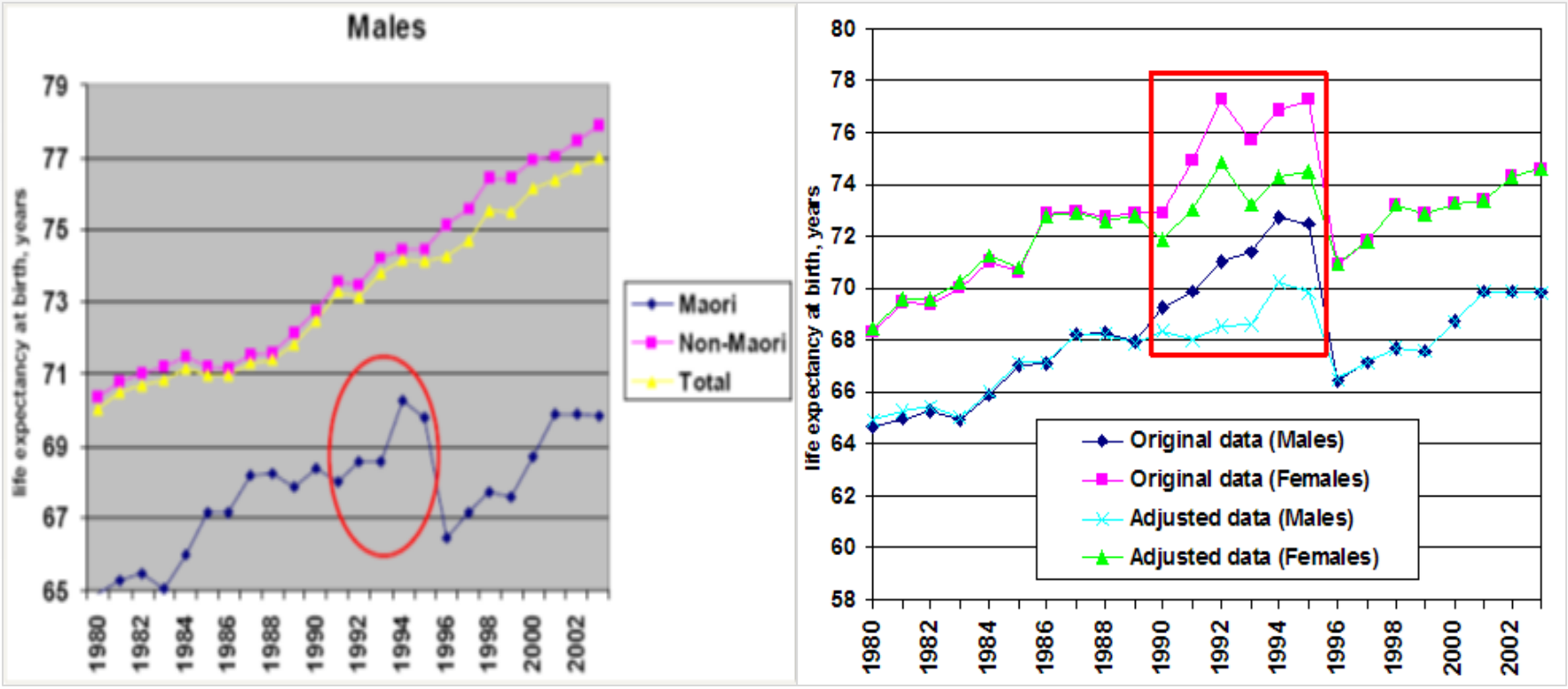
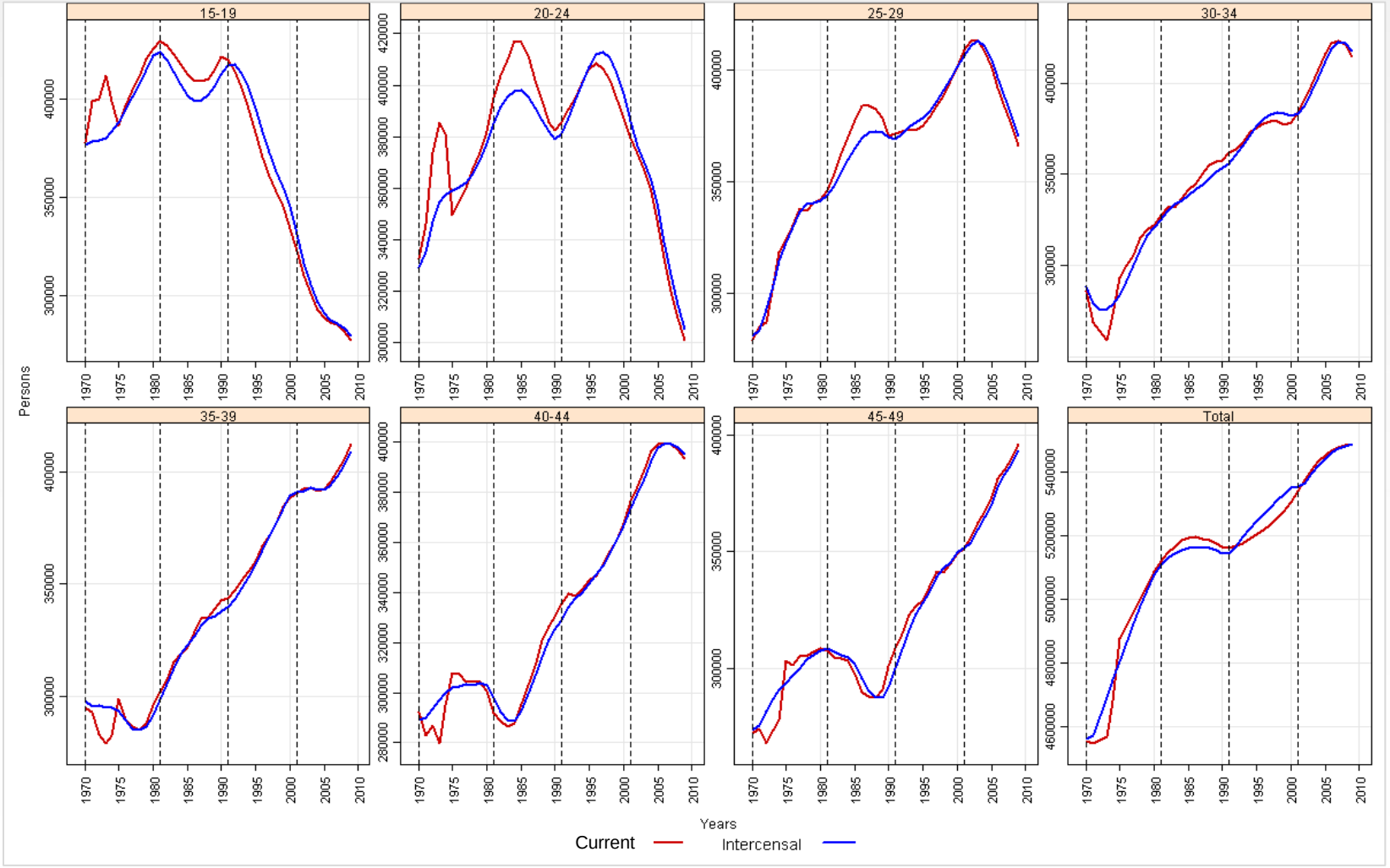



**Figure:** Life expectancy at birth of Māori, Non-Māori, and total population of New Zealand calculated from the official (unadjusted) data (left panel) and adjusted HMD data (right panel). Source: (Jasilionis et al., 2015)

# Impact of migration at working/reproductive ages on mortality and fertility estimates

*Population exposures by age group, Portugal, females*

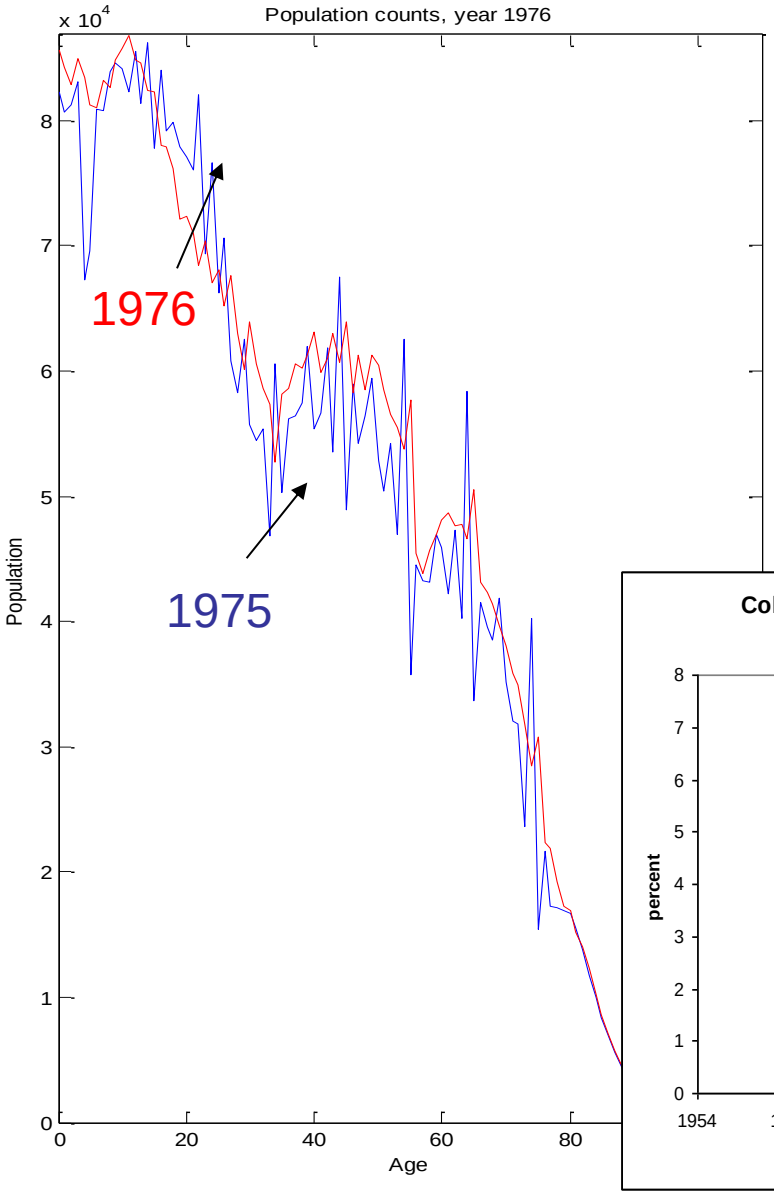
*Red line – official current estimates, blue line – HMD inter-censal*



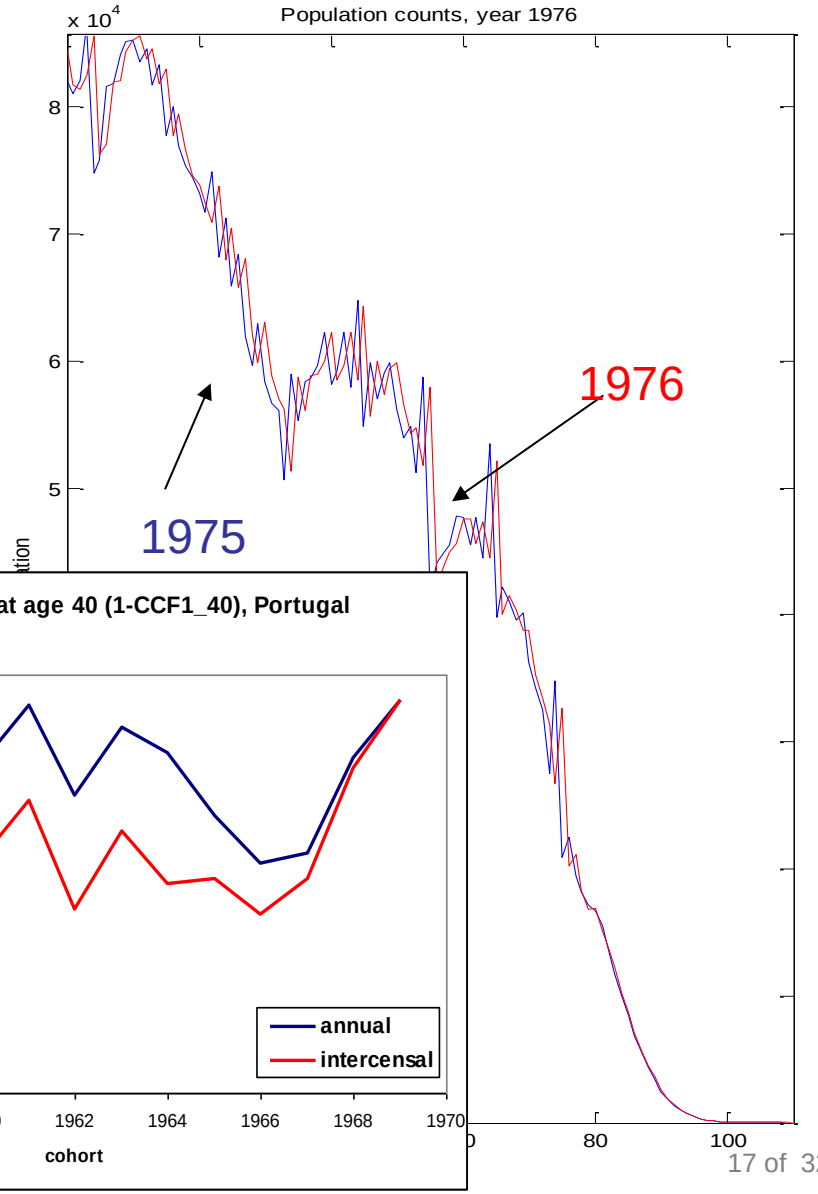

In 1974 Portugal lost its African colonies, 500 000-700 000 returned to Portugal in the second half of the 1974 and in 1975

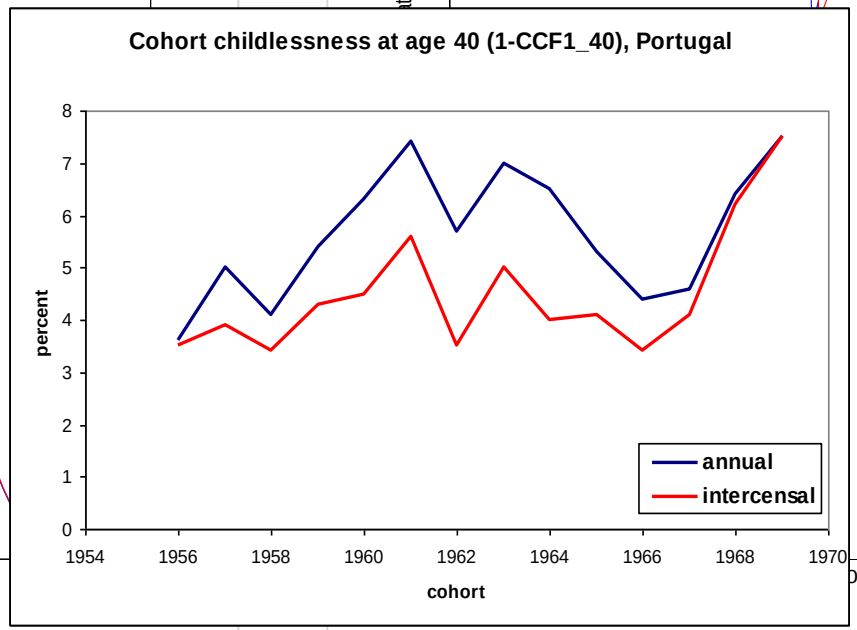
# Inter-censal population estimates and fertility indicators: Portugal in 1975-76

## Current population estimates




## HMD inter-censal estimates



# Mortality at advanced ages

# Mortality estimates at old ages

- Internationally comparable high quality demographic data on old-age populations remain insufficient.
- The HMD is the only major demographic database which provides such data. Population estimates for ages 80+ in the HMD are recalculated using extinct/almost extinct cohort and survival ratio methods.

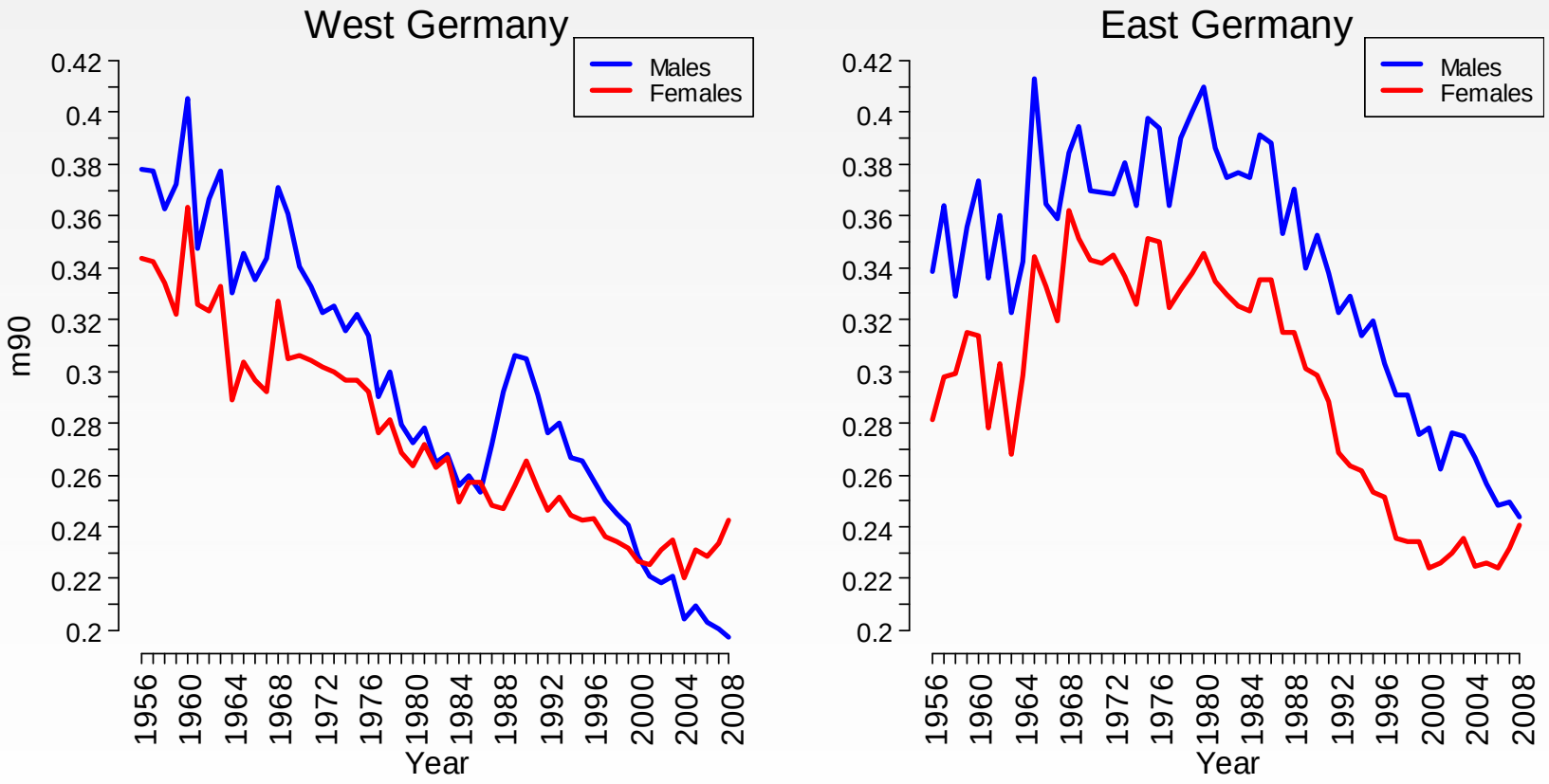
**Evaluation of data quality at old ages**

The standard set of methods includes tests on

- age overstatement (e.g. the ratio of the total person-years lived above age 100 to the total person-years lived above age 80);
- precision of age reporting with the UN age-sex accuracy index;
- age heaping with the Whipple's Index of age accuracy.

The comparison to other countries with reliable statistics may be also used for evaluation of data quality.

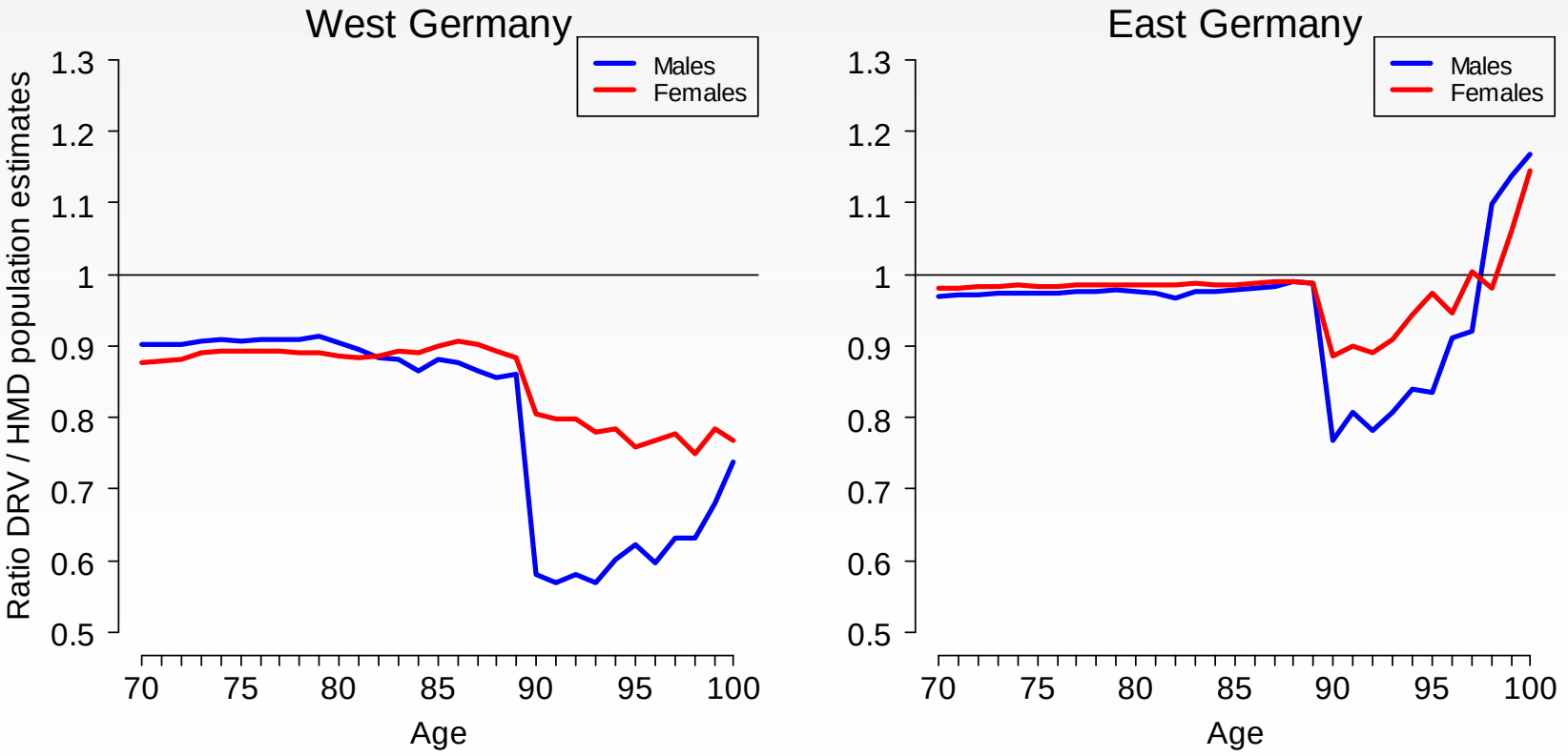
# Germany: old ages




Trends in death rates at age 90+, calculated from the official population estimates, for the West and East Germany, males and females, 1956-2008.

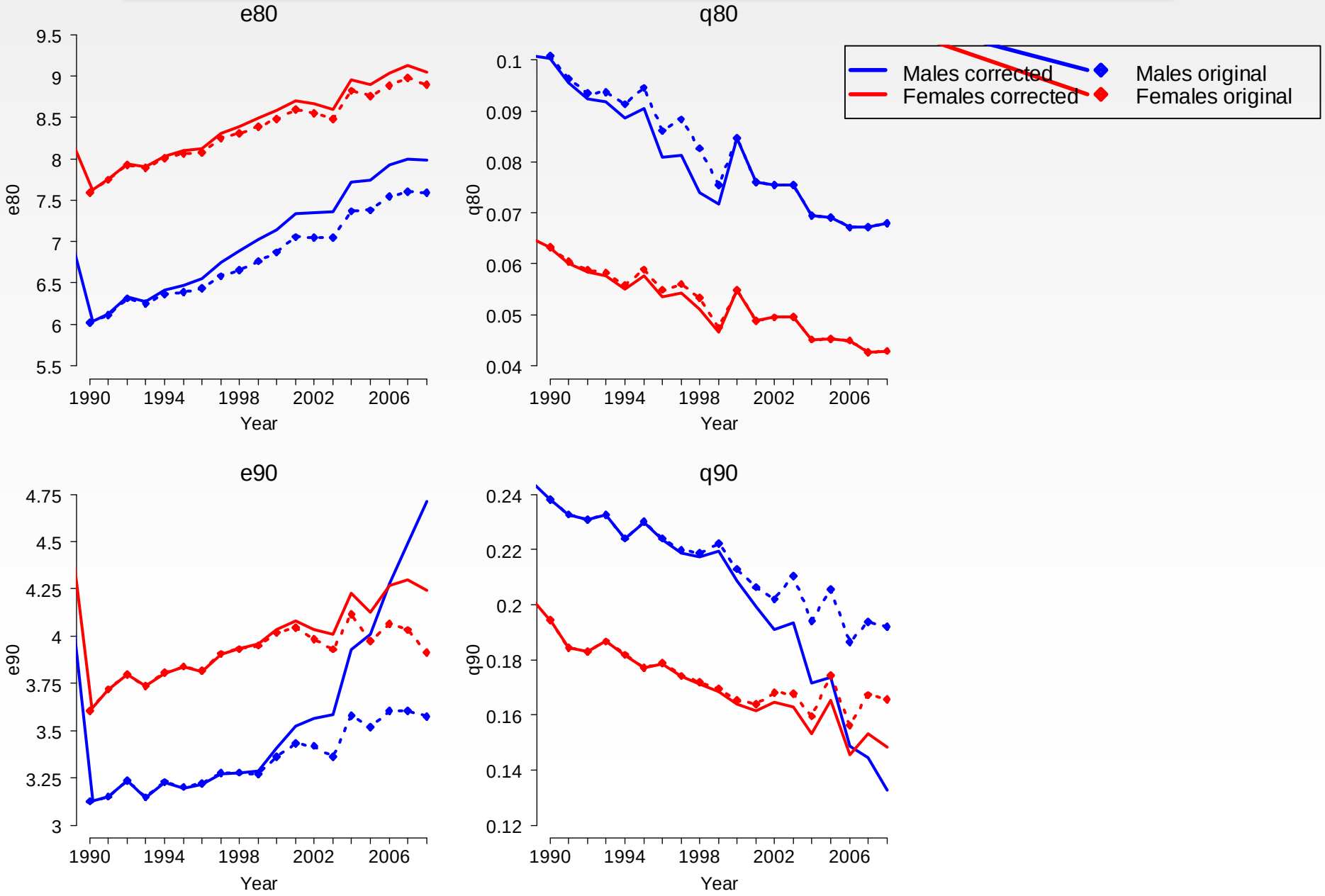
# Germany: old ages (cont.)

To correct population estimates for West Germans at older ages in 2010, the HMD team used data by the *Deutscher Rentenversicherung Bund* (DRV), the German Pension Scheme.

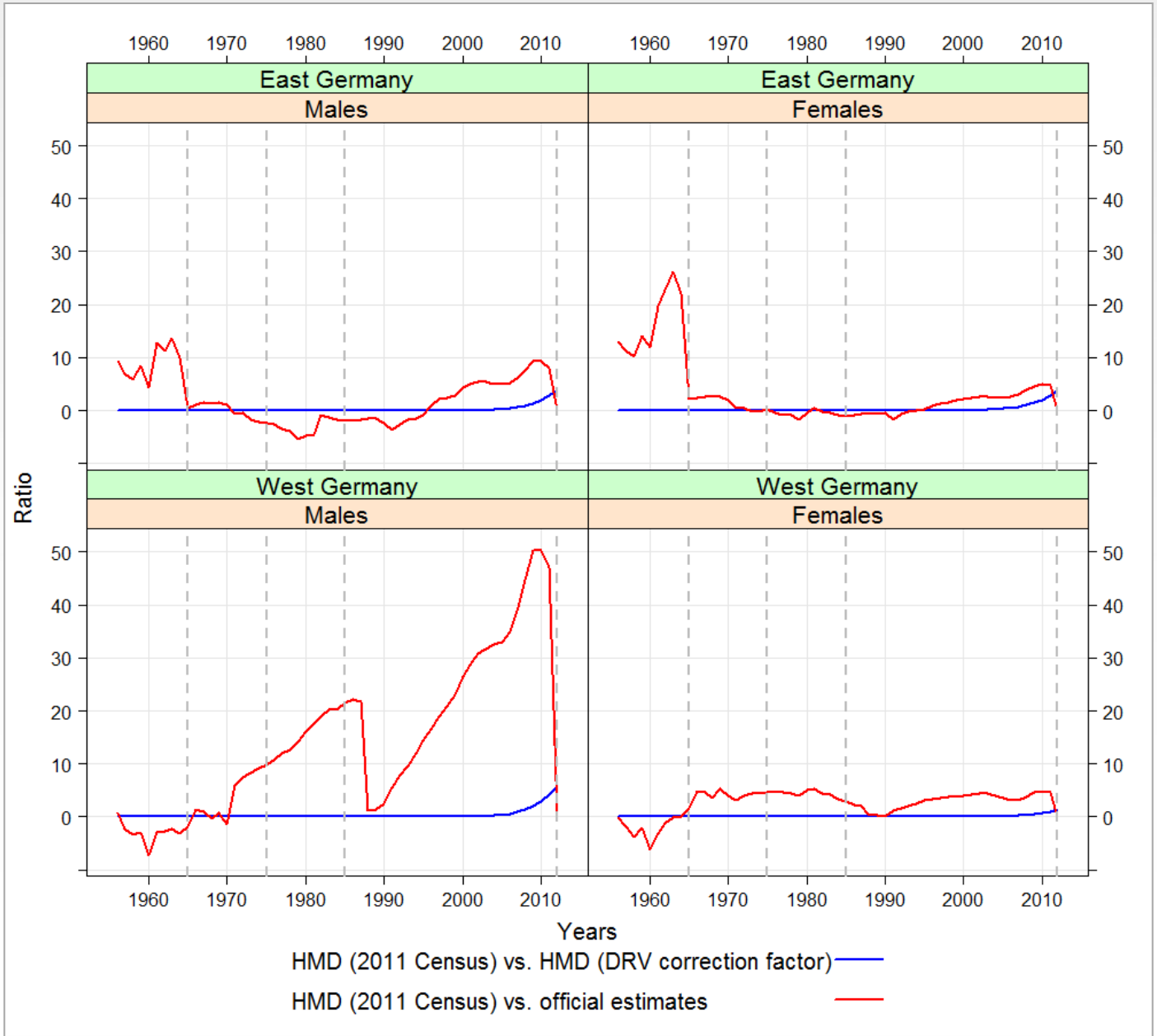



Ratio of DRV records by age based on own pensions to estimates based on official data, West and East Germany, 2009.

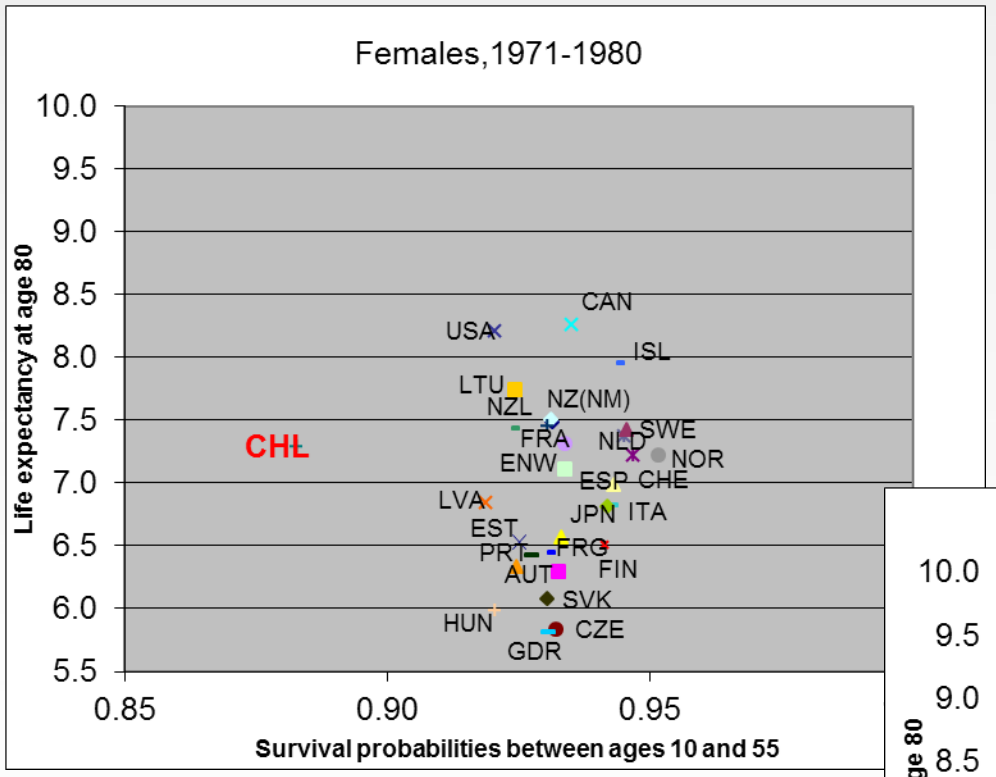
# Life expectancy and probability of death for the corrected and the original data, West Germany, 1990-2008

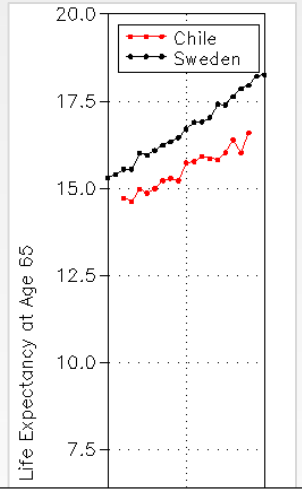


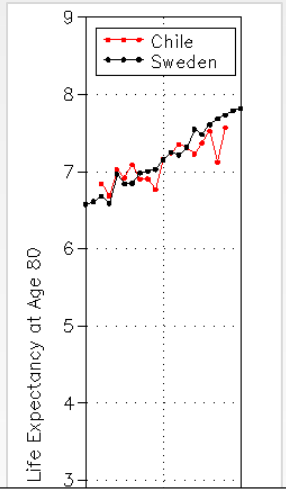
# Relative difference (per cent): HMD (2011 Census) vs. official population estimates and HMD (2011 Census) vs. HMD (DRV correction at ages 90+)

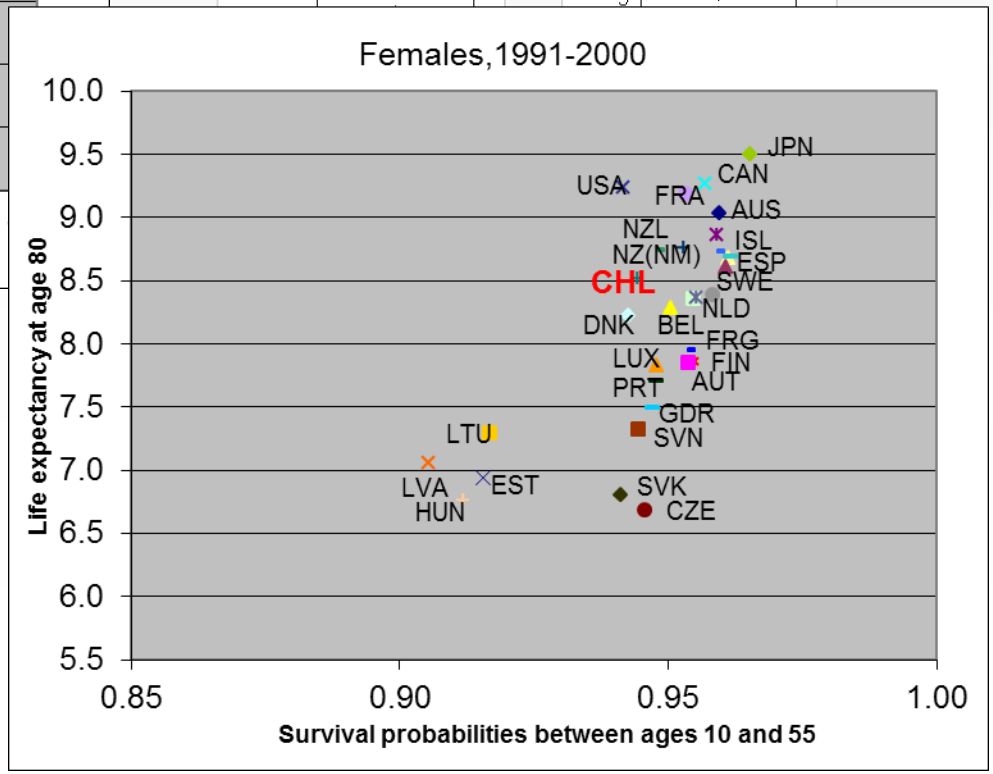


# Chile: old ages



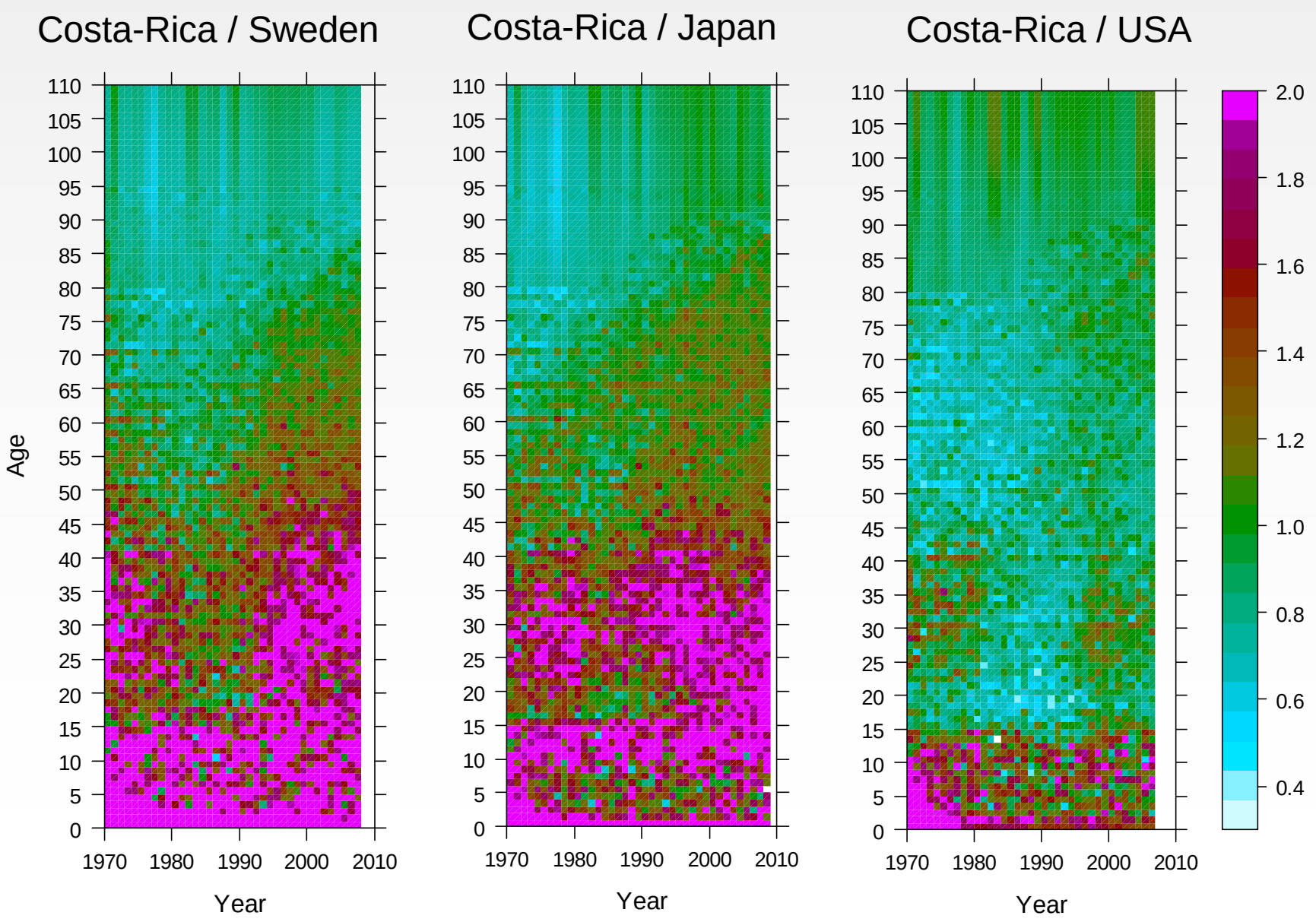



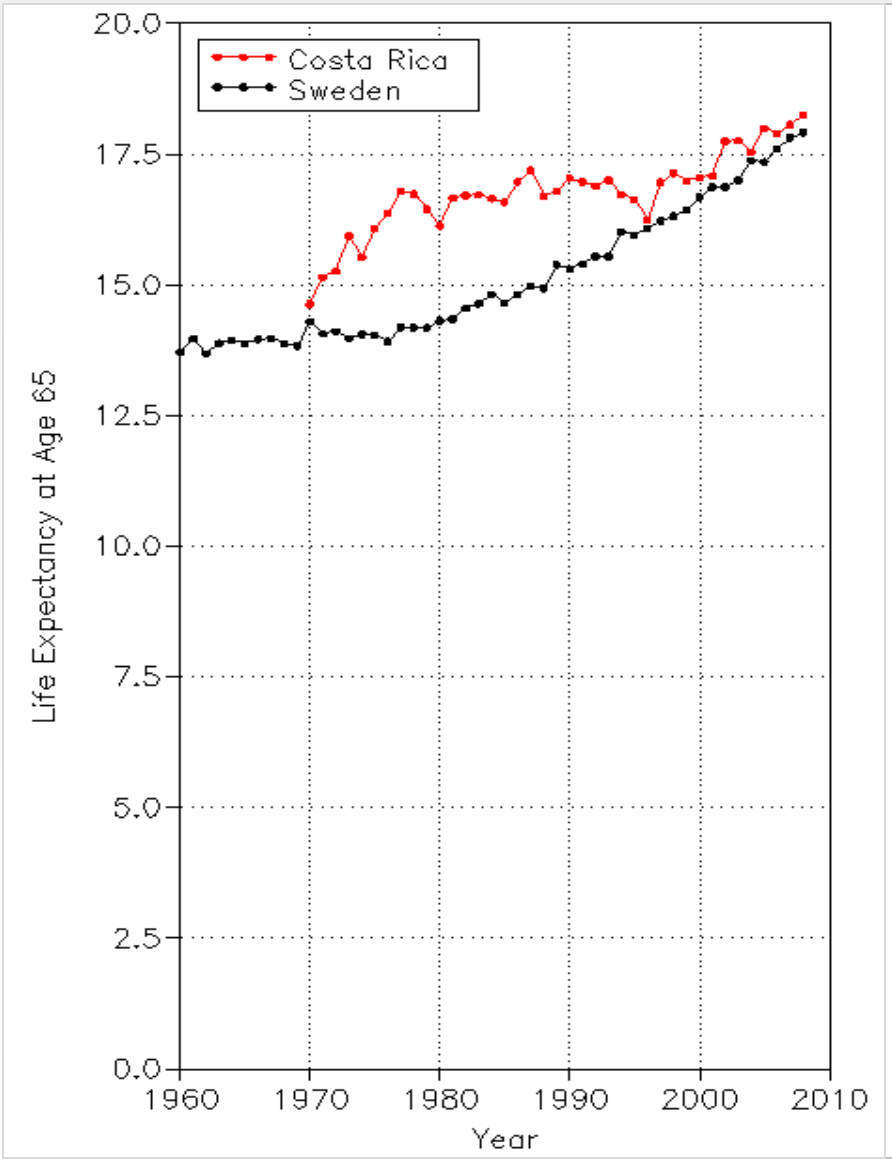



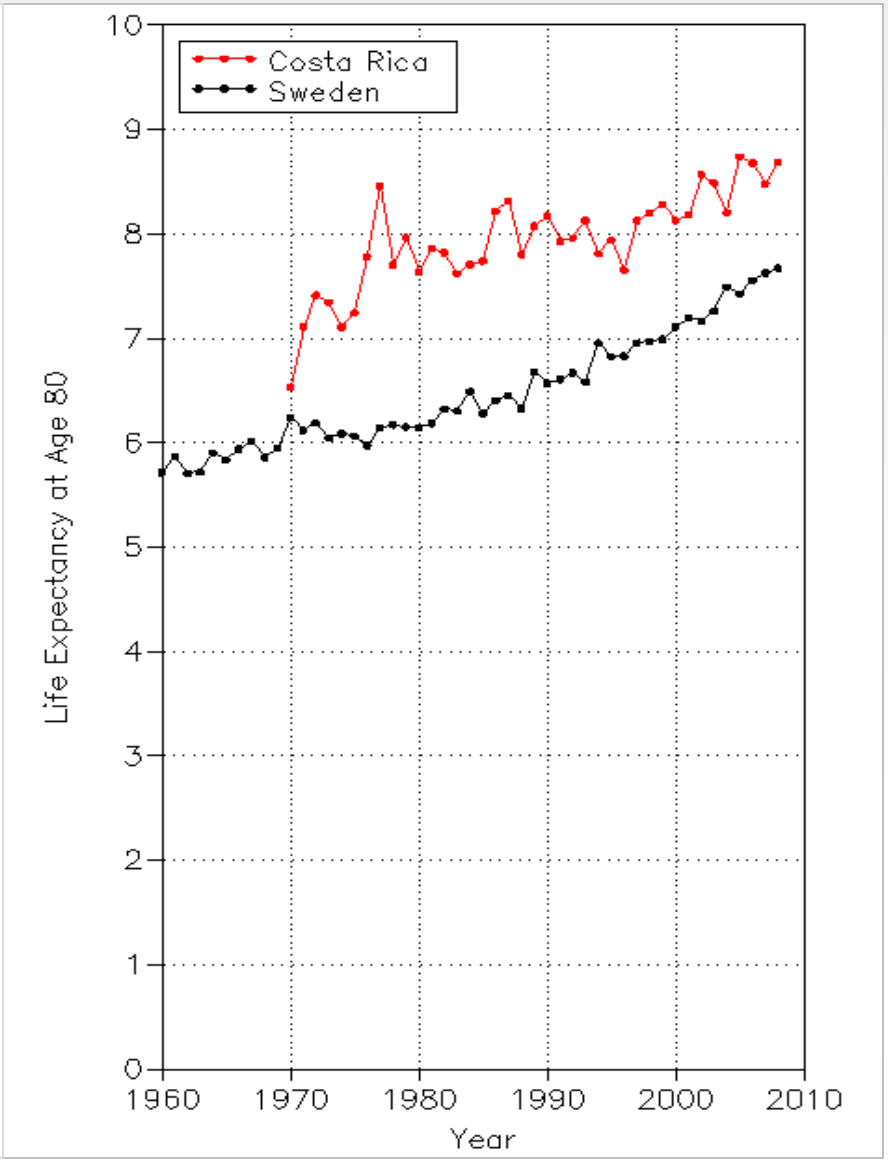

In the HMD/HFD data series for Chile starts in 1992

# Costa-Rica: death rate ratios, males



# Costa-Rica: life expectancy at advanced ages



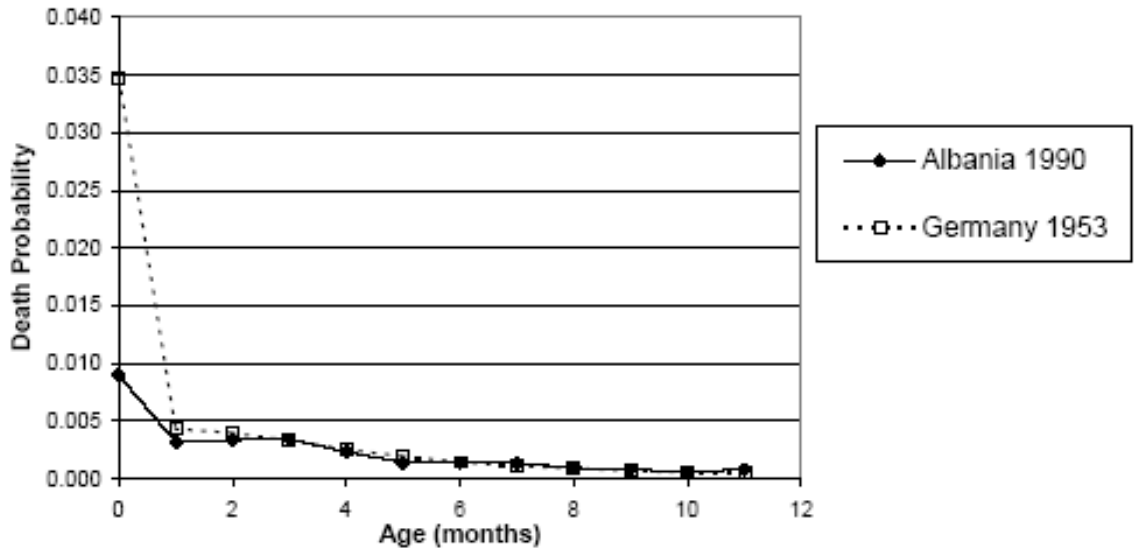
# Infant mortality

# Underestimation of infant mortality due to a restrictive definition of live birth and death undercount

| | Average official infant mortality rate 1987-2000 | Adjustment factor (per cent) | Adjusted infant mortality rate 1987-2000 |
|---|---|---|---|
| Albania (4) | 28.4 | +110.9 | 59.83 |
| Bulgaria (5) | 15.1 | +56.6 | 23.64 |
| Croatia (6) | 9.8 | +0.8 | 9.90 |
| FYR Macedonia (8) | 27.9 | +32.9 | 37.05 |
| Romania (9) | 23.5 | +81.2 | 42.57 |

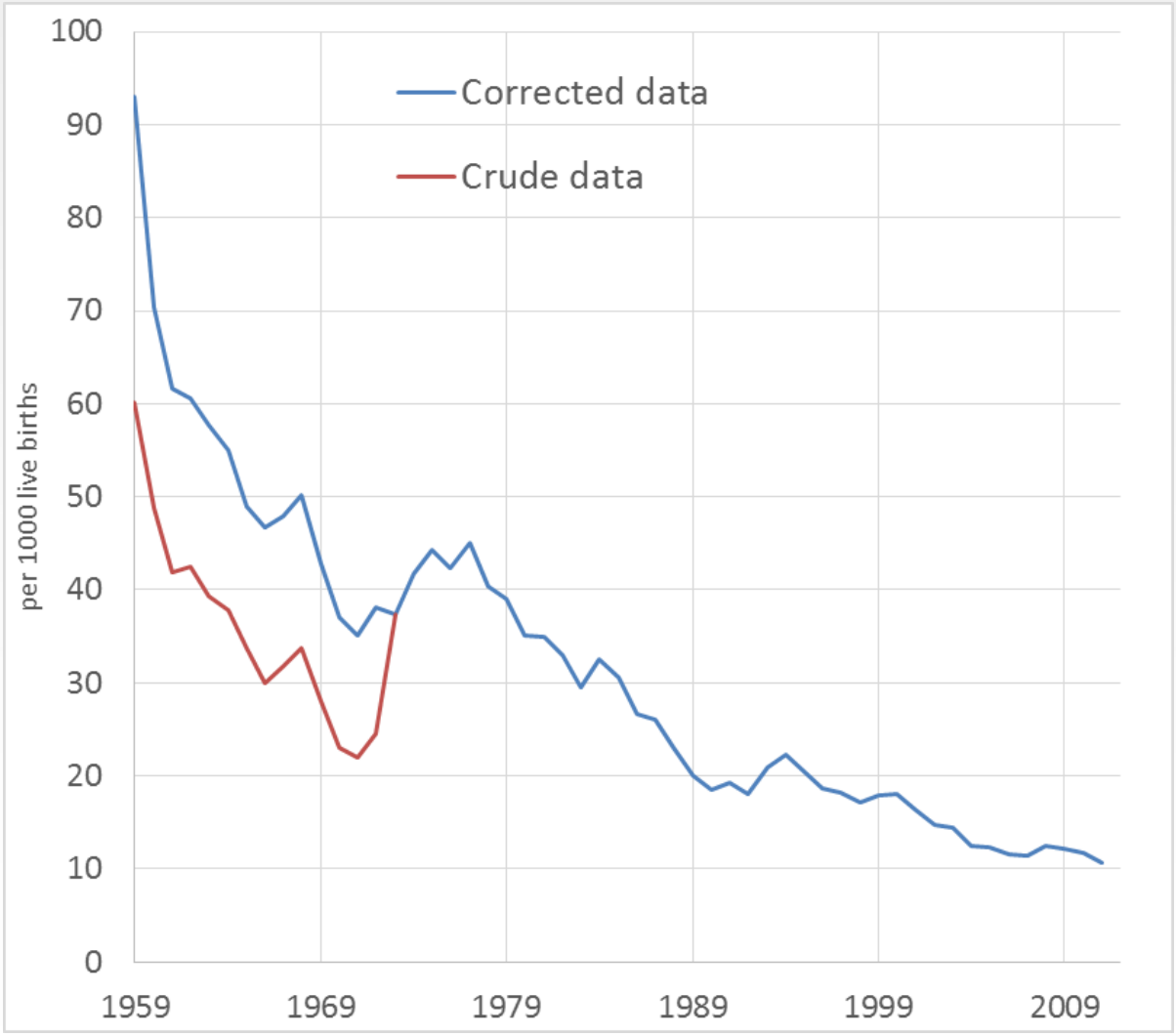
**Adjustment is made by correction of the monthly mortality curves. The adjustment brings these curves to certain "golden-standard" curve.**

Figure 3. Monthly Probabilities of Dying in First Year of Life
Albania vs Germany




Source:
Kingkade & Sawyer, 2001.

# Underestimation of infant mortality: adjustment by mortality trend



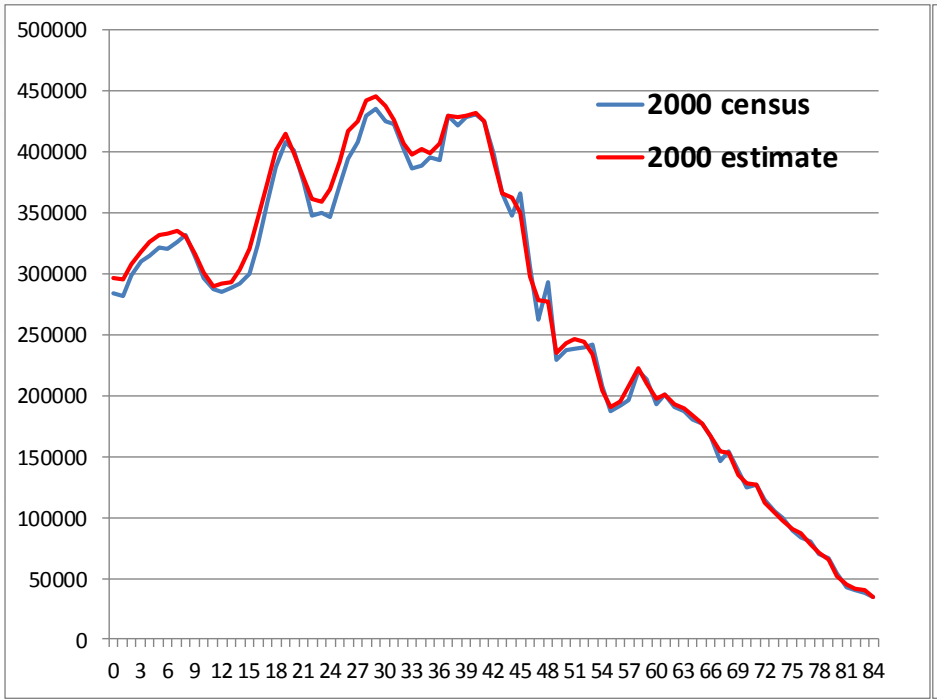

An abrupt increase in the infant mortality that occurred in all of the Soviet republics at the beginning of the 1970s was interpreted by Anderson and Silver (1986) as a result of improvements in the registration rather than a real deterioration in survival of the newborns.

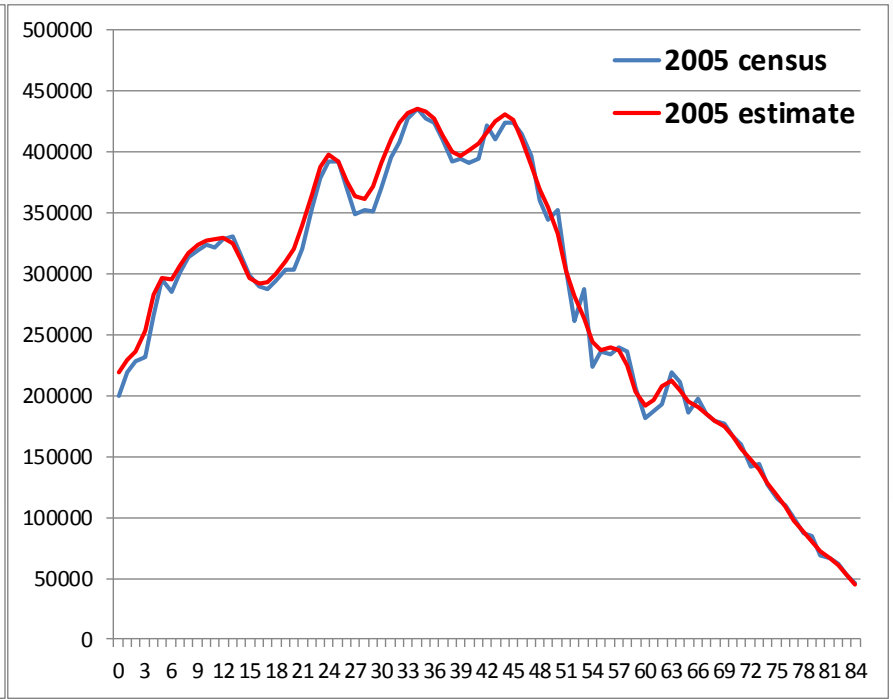
Figure: Infant mortality rate in Moldova before and after correction prior to 1973, both sexes, Moldova, 1959–2014. Source: (Penina et al., 2015)

# South Korea: preliminary analysis

Infant mortality before 2000 is not reliable

After 2000:
Substantial differences between pop estimates for 2000 and 2005 and census counts for the same years. Perhaps census counts excludes foreigners? These differences are important for the ages 0,1 and also for other ages (including adult and old ages). It can be also seen that some smoothing was used to produce pop estimates (fluctuations observed at some ages in census data are absent in pop estimates - this cannot be explained by a simple exclusion of foreigners).



# Conclusion

- Data are of high quality if they are “Fit for Use” in their intended operational, decision-making and other roles (Juran and Godfrey, 1999). This is why the understanding of problems hidden in the data is important in any demographic estimation, forecast or study.

- We discussed several approaches which allow us to increase significantly utility of the data even if data quality is problematic.

- Standard demographic methods which work well with data from developing countries or historical data series are often not applicable to problematic data from countries with functioning statistical systems.

- Country-specific approaches in combination with usage of additional and alternative data sources are needed. They should be combined with certain general principles that are applied in all countries to ensure comparability of HMD data series across time and space.

**Thank you!**