Population Census Microdata Time Machine: Social History and Data Quality Assessment

Griffith Feeney <feeney@gfeeney.com>

IPUMS-International "Big Census Microdata" Workshop

Saturday 1 March 2014

1

My aim is that 15 minutes from now you will understand why a population census can be a "time machine" and how censuses can tell us about long term demographic and social trends and data quality.

Introduction

Time-plotting is …

- a way of displaying census data that provides indicators of (a) long term historical trends and (b) accuracy of census statistics
- I'm going to show you three examples of time-plots using IPUMS-International data

2

Example 1

Mean Children Ever Born
Argentine Censuses of 1970, 1980, 1991, 2001 and 2010

3

Ryder's Observation

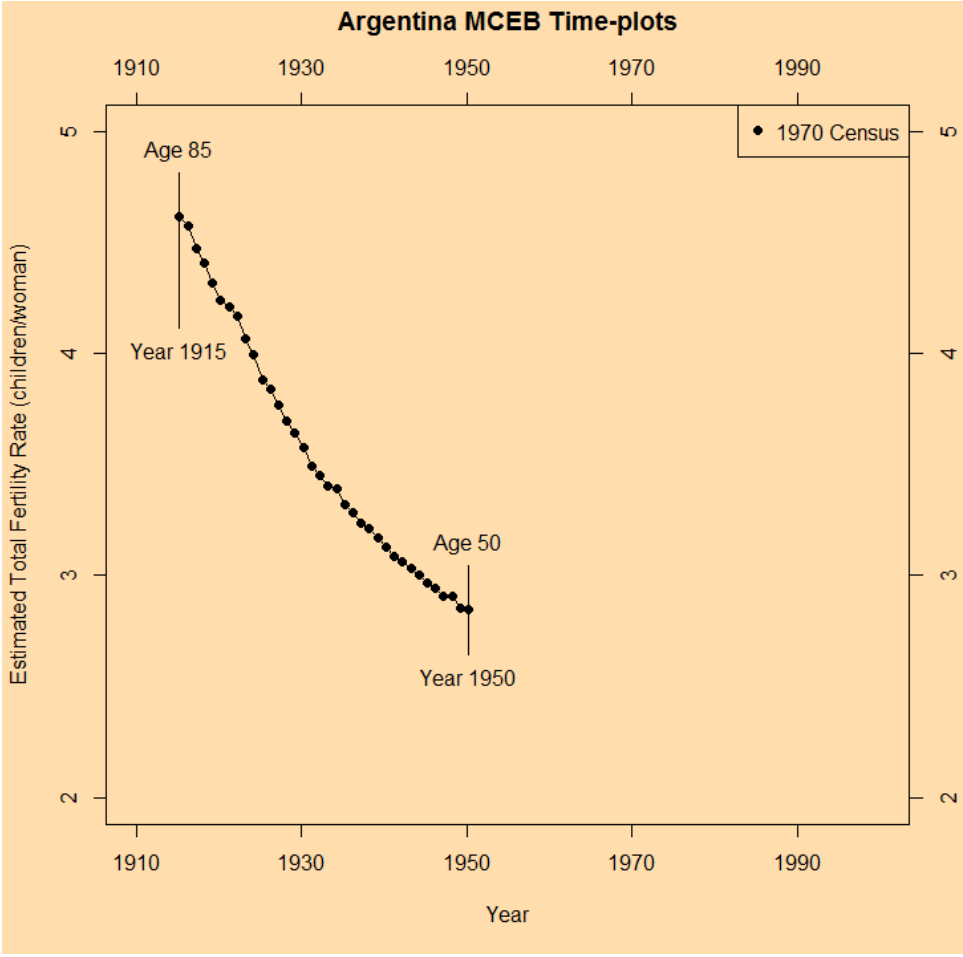
- Completed fertility of a birth cohort ≈
- Period total fertility rate
- At the time the cohort reached its mean age at childbearing (MAC)
- MAC ≈ 30 years

4

Mean age at childbearing tends to be about 30 years and tends not to vary much over time or between countries.

If you're not a demographer, that slide is probably completely unintelligible. Let's look at the example and see if we can make it clear.

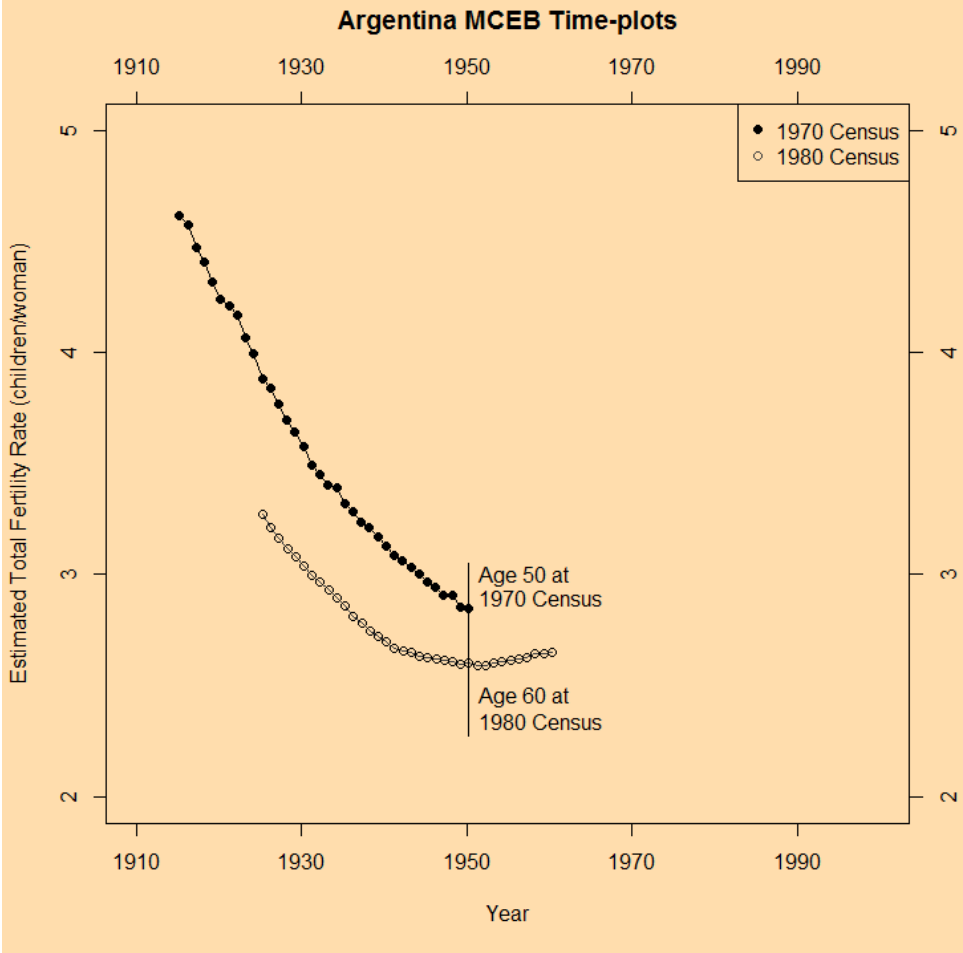
Here's the time-plot for the 1970 census.



The far right point shows mean children ever born of 50 year old women in the 1970 census. They had their children at age 30, on the average, which is 20 years before the census, which is 1950. This is the time at which the value for women age 50 is plotted.

Moving up one year in age corresponds to moving back one year in time, so the points to the left are for women age 51, 52, and so on.

Let's add the time-plot for the 1980 census.

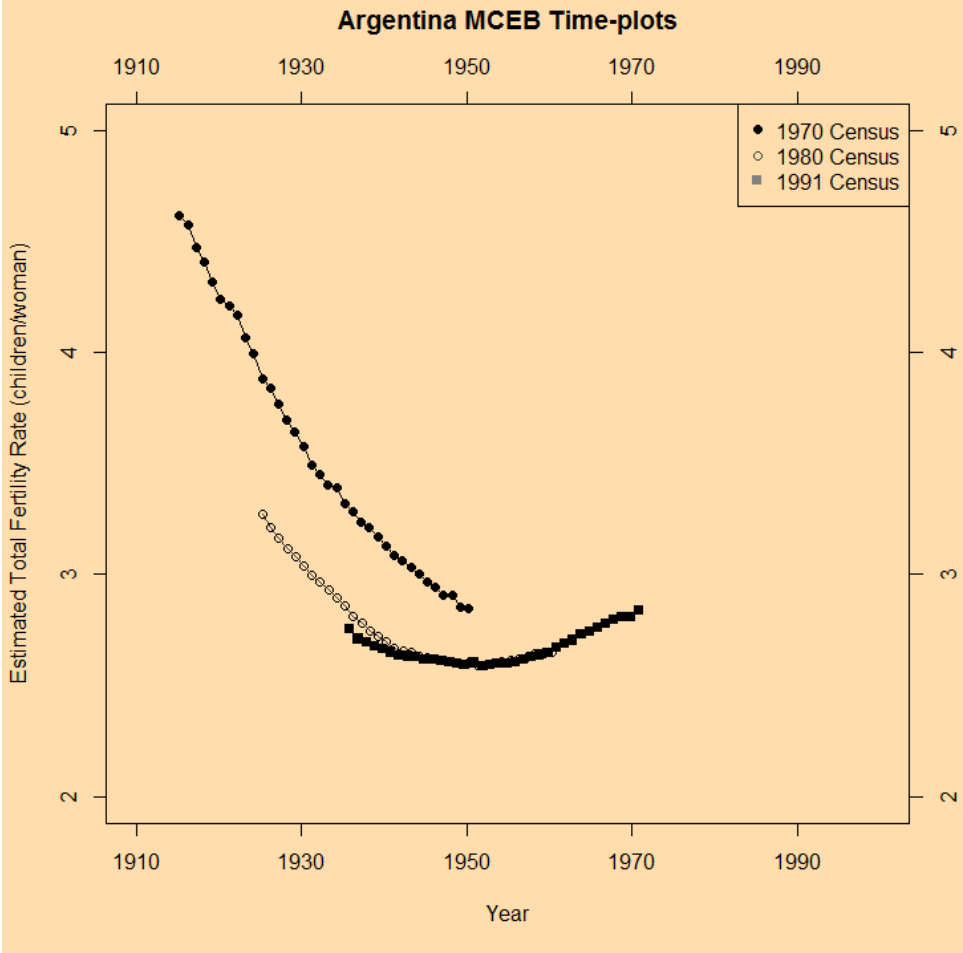


Well, we would like to see perfect agreement, but we don't always get what we want.

What's interesting here is that the *pattern* of decline shown by the two censuses is very similar, though the level indication is different.

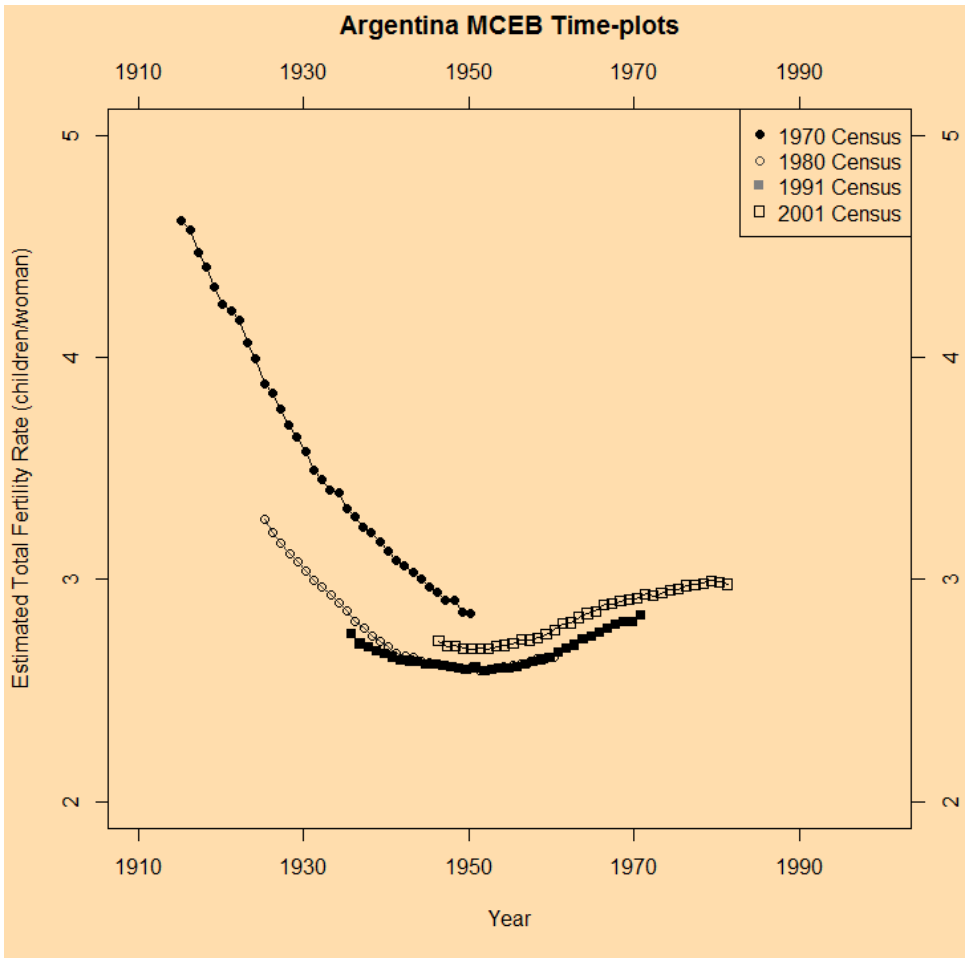
Incidentally, when we move down the vertical line, we are following a cohort from one census to the next.

Now add the 1991 census.



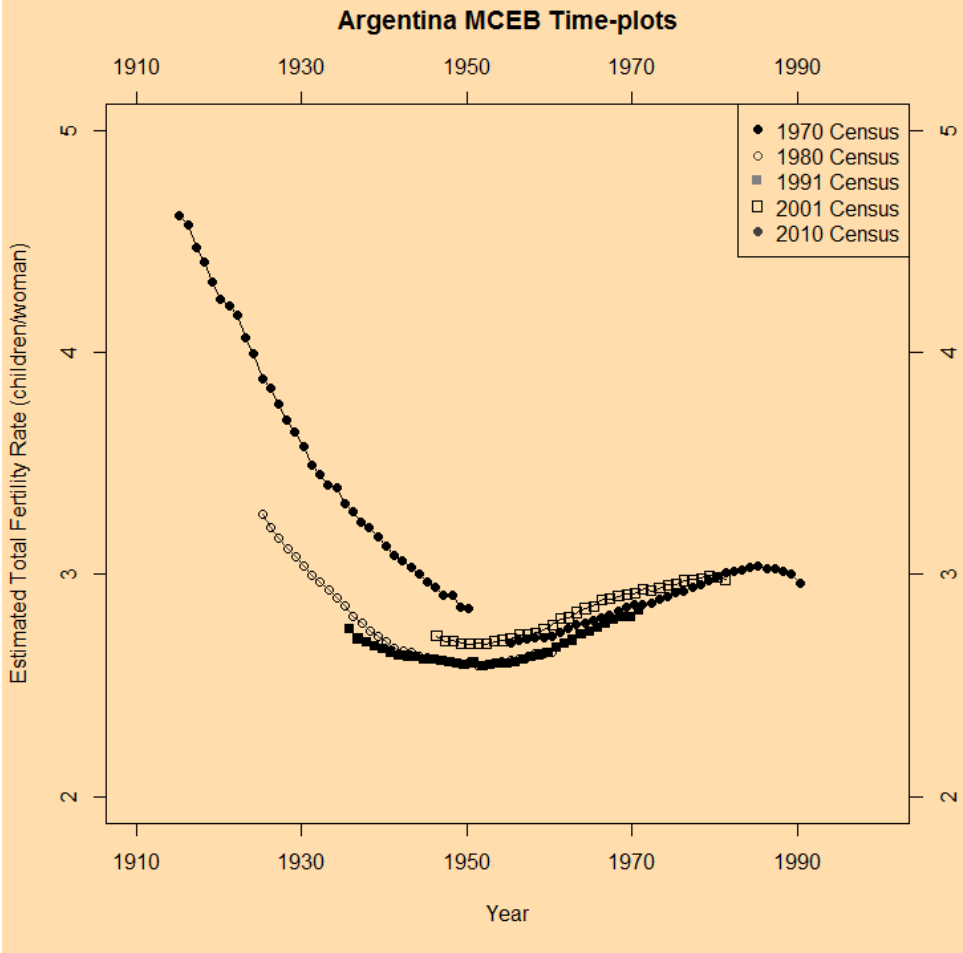
Wow! Almost perfect agreement here.

Let's see what happens when we add the 2001 census.



Interesting. We have another case of what looks like the same trend at a different level. But in this case the level difference is only about 1/10th of a child per woman.

Finally, the 2010 census.



Interesting again, in a different way. The 2010 census series moves back and forth between the 2001 and 1991 censuses.

Overall, though, consistency for the last four censuses is good.

1. Conclusion on Argentina MCEB series

[PAUSE]

Let's step back and see what the children ever born data from the five census seem to be telling us.

Note that we are looking at 75 years of changing fertility—most of the 20th century.

Here's what we see.

1. Rapid decline under way in 1915
2. Slowing decline around 1930
3. Bottom and begin slow increase around 1950
4. Top and resumed decline around 1985
5. The trend is consistently indicated by all five censuses
6. Good level consistency for the last four censuses
7. The 1970 census is a level outlier

It's unclear how we should deal with the level difference between 1970 and latter censuses.

On the one hand, we tend to accept higher over lower mean children ever born values because the usual worry is under reporting of children ever born.

On the other hand, we tend to accept values consistently indicated by multiple sources over outliers.

This is a puzzle yet to be solved.

Example 2

Literacy

Malaysia Censuses of
2000, 1980 and 1970

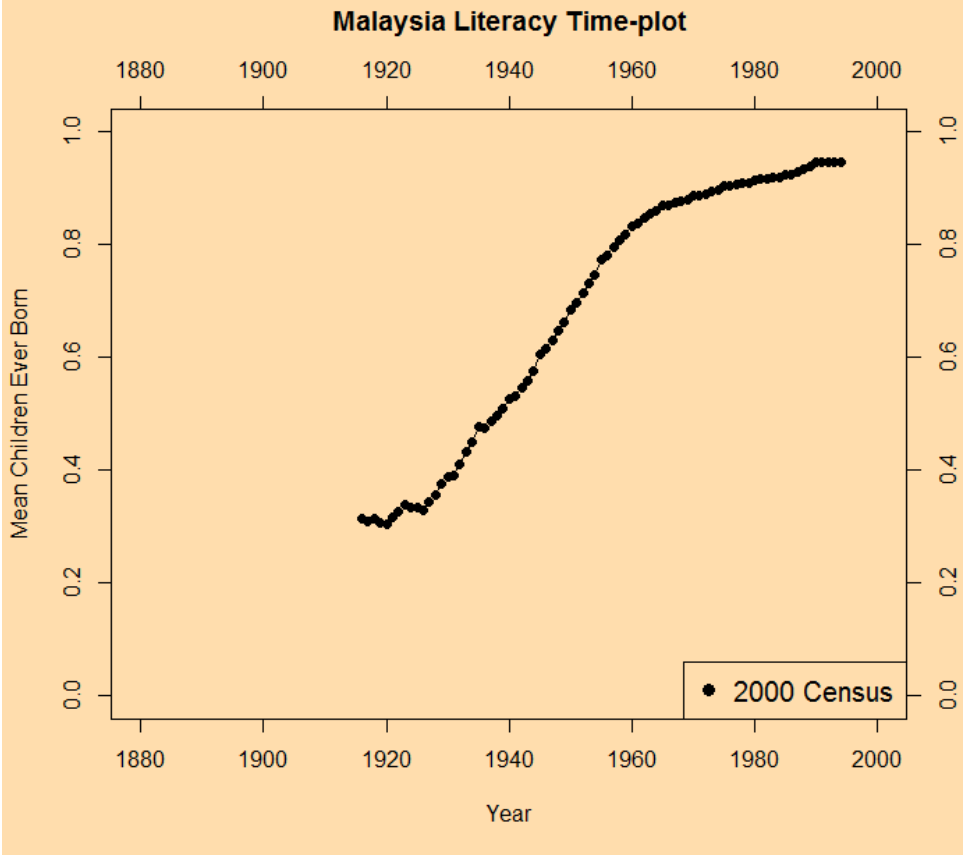
10

Method

- Mean age at attaining literacy ≈ 10 years
- Persons age *x* now became literate x – 10 years ago on the average
- Proportion literate at age *x* reflects society's production of literate persons *x* - 10 years ago

11

Here's the literacy plot for the 2000 census . . .

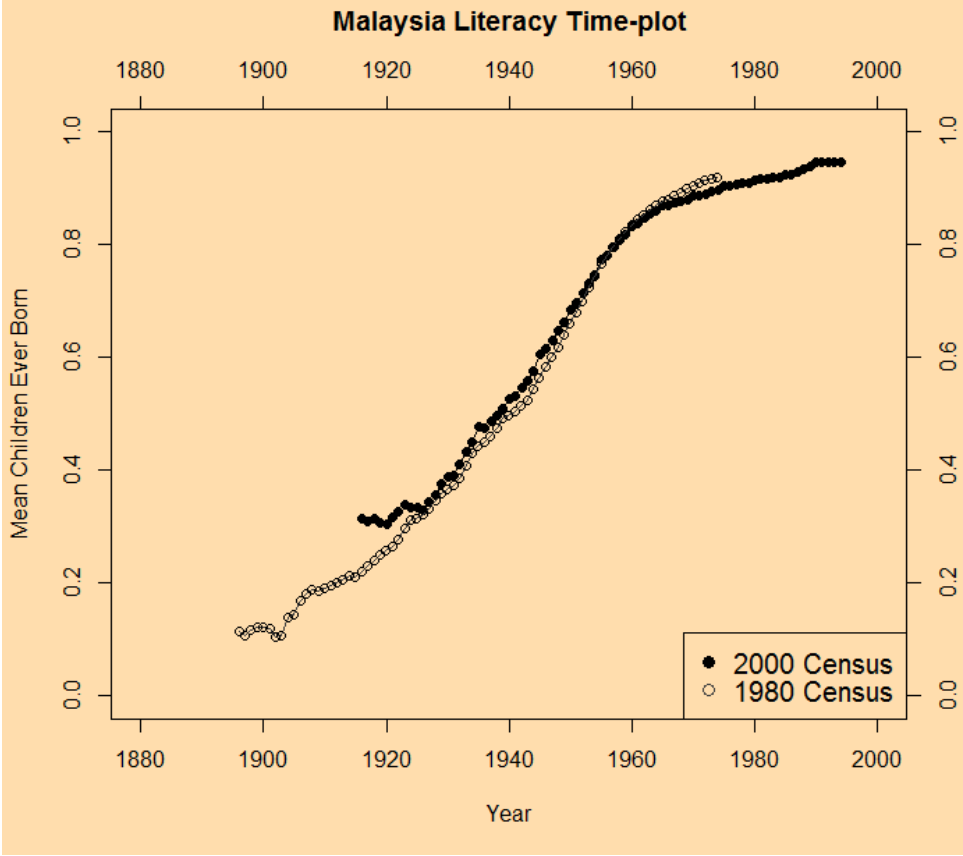


The far right point is the proportion literate for persons age 16 at the 2000 census. They became literate at age 10, on the average, which means 6 years before the 2000 census, which is 1994. This is the time at which they are plotted.

And similarly for proportions literate at older ages.

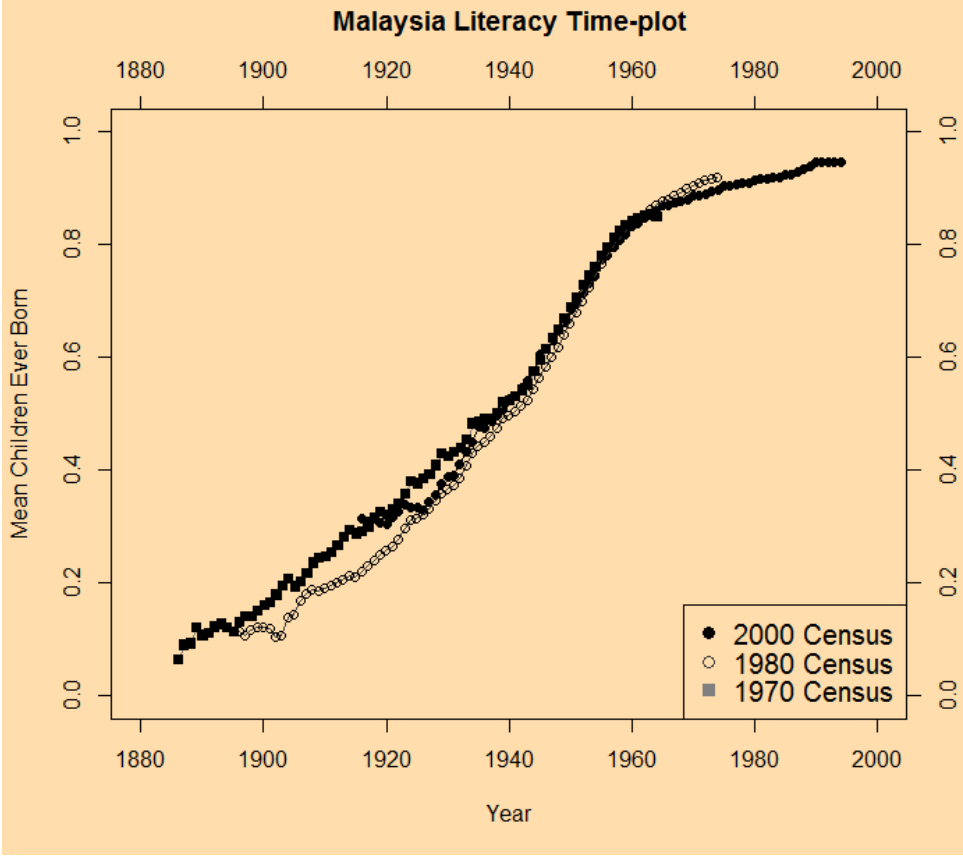
The far left point, incidentally, is for persons age 94 at the the 2000 census.

Now add the 1980 census ...



Quite good consistency, suggesting high data quality in both censuses.

Now the 1970 census ...



[PAUSE]

We have here a series representing society's production of literate persons over more than a century. The level rises from 5-10 percent in the mid-1880s to 95 percent in the mid-1990s.

The rate of increase speeds up slightly after 1940.

It slows down sharply in the late 1950s.

We might be concerned about what appears to be a stagnation at 95 percent in the 1990s.

Consistency back to about 1940 is generally good, suggesting good data quality.

Before 1940, the 1980 census series falls substantially below the 1970 series.

The censuses agree, however, on the rising trend.

The third and last example . . .

Example 3

Primary School Completion
Uganda Censuses of
2002 and 1991

15

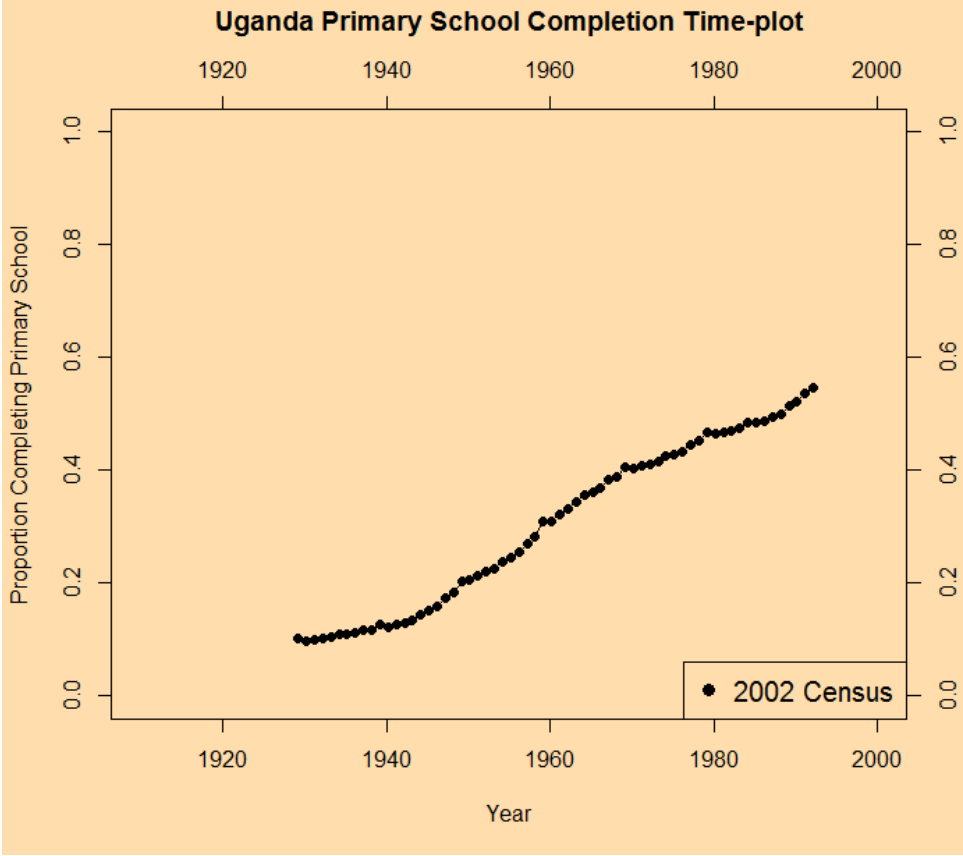
Method

- Mean age at completion ≈ 12 years
- People age *x* completed primary school *x* - 12 years ago on the average
- Age *x* proportion completing primary school reflects educational system's production of primary school graduates *x* - 10 years ago

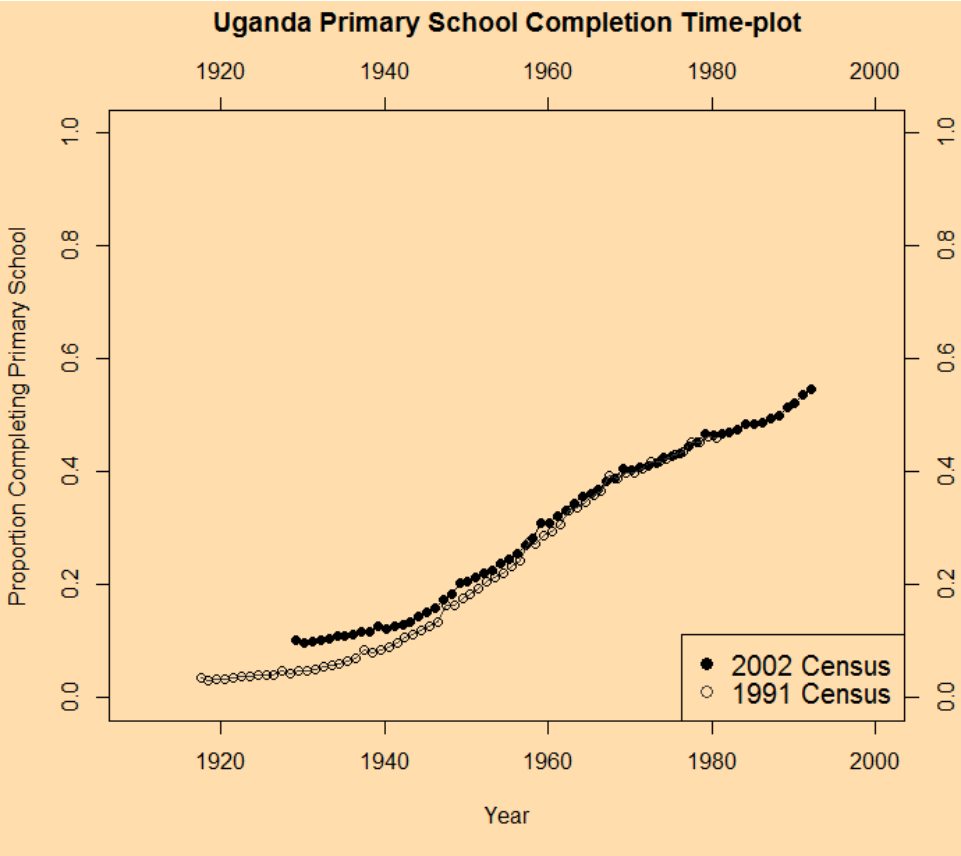
16

The method is the same. We don't need to go through it.

Here's the plot from the 2002 census.


Uganda Primary School Completion Time-plot
1920
1940
1960
1980
2000
0.0
0.2
0.4
0.6
0.8
1.0
Proportion Completing Primary School
2002 Census
Year

Add the plot from the 1991 census.



We have here a series of the educational system's production of primary school graduates from the mid-1910s to the mid-1990s, most of the 20th century.

Proportion of primary school graduates rises from close to zero at the beginning of the century to about 55 percent in the early 1990s.

Consistency is quite good back to 1950, suggesting good data quality.

Consistency is less good before 1950, but this doesn't *necessarily* mean poor data quality.

For people this old, higher values from the later census might reflect mortality selection in favour of primary school graduates.

Discussion

19

The Power of *Microdata* + *Information Technology*

- Using the IPUMS-International database I have produced mean CEB time-plots for 55 countries, literacy time-plots for 49 countries, and primary school completion time-plots for 61 countries
- *Microdata* + *Information Technology* made this a work of weeks rather than years

20

Why Microdata is Important

- Number of possible tables from a modern census is effectively infinite
- No tabulation plan can anticipate more than a fraction what is useful—now, much less in the future
- With microdata, we can generate any table we need, quickly, easily, inexpensively
- Past censuses can be as available as the most recent census

21

Benefits of Microdata

- Seamless transition from tabulation to analysis—get results faster
- Superior analysis of census results
- Superior census products for users
- Microdata is an important resource *within the NSO, w*here you can work with complete count data as well as samples

22

Benefits of IPUMS-International

- Compare your census results with those of other countries
- Harmonization of variable metadata makes this relatively easy to do
- A microdata resource management model, for microdata used within your office as well as for microdata products for users

23

Conclusion

Time-plotting presentation by an Eminent Statistician at a Distinguished Venue

24

Statistician General Pali Lehohla presents 2011 census results, 30 October 2012

25

For More Information

- Email me at feeney@gfeeney.com
- Visit my website at www.gfeeney.com

26