

Reconstruction of Populations by Age, Sex and Level of Educational Attainment for 120 Countries for 1970-2000: A Summary

Ricostruzione della popolazione per età, sesso e livello di istruzione di 120 Paesi dal 1970 al 2000: una sintesi

Wolfgang Lutz, Director of the Vienna Institute of Demography (VID); Leader of the World Population Program at the International Institute for Applied Systems Analysis (IIASA), lutz@iiasa.ac.at

Anne Goujon, Research Scholar, VID and IIASA, goujon@iiasa.ac.at

Samir K.C., Research Scholar, IIASA, kc@iiasa.ac.at

Warren Sanderson, Departments of Economics and History, State University of New York at Stony Brook and IIASA, wsanderson@notes.cc.sunysb.edu

Riassunto

Facendo ricorso a metodi multistato che consentono di retro-proiettare le popolazioni per età, sesso e livello di istruzione di 120 Paesi, dal 2000 al 1970, in questo contributo ci si propone di ricostruire dati sul capitale umano essenziali per studiare empiricamente, a livello aggregato, i ritorni nel sistema di istruzione. Questo dataset, prodotto congiuntamente dallo IIASA e dal VID, fornisce informazioni dettagliate sul livello di istruzione della popolazione per classi quinquennali di età, basandosi su definizioni assolutamente comparabili nel corso del tempo. Sulla base delle distribuzioni per titolo di studio secondo l'età e il sesso, relative all'anno 2000, il suddetto metodo ci consente di andare indietro per generazione, tenendo esplicitamente conto dei livelli di mortalità differenziali per genere e livello di istruzione.

Keywords: back-projection, human capital, education

1. Introduction

In this paper we summarize an article aimed at describing the methods used for reconstructing the educational attainment distributions for 120 countries using the methods of multi-state demographic modelling (see Lutz *et al.* 2007). This article is part of an ambitious, multiphase project whose objectives include the production of a new national level dataset on educational attainment by age and sex for as many countries in the world as possible over the period 1970-2000, the analysis of these new data, the making of projections of educational attainment by age and sex for those countries through 2050, and the assessment of the likely effects of future changes in educational structure.

For many years economists interested in the determinants of economic growth have been puzzled by the lack of consistent empirical evidence on the macro-level returns to

education, in stark contrast to the strong evidence on the individual level where it is well established that more education on average leads to higher income. This unsatisfactory situation has led to the suspicion that the problem may not lie with the theory or the models used but rather with the aggregate level education data themselves that could be solved by using more accurate, consistent and detailed education data. These data can also be useful in the study of other important and policy-relevant areas as well. This information needs to be consistent in terms of the definition of educational categories across countries and over time. Since the effects of educational attainment can also be expected to differ by age (e.g., one might expect that the education of 25-34 year olds should be more important for economic growth than that of persons beyond retirement age) as well as by sex, having full age details for men and women can be considered a great asset for a comprehensive analysis. In addition, only the explicit consideration of distinct levels of educational attainment allows for the analysis of the relative importance of primary versus secondary or tertiary education (and different mixes of the three) which should be key to the development of relevant education policy plans at national and international levels. Such consistent information by age, sex and level of education has not been available so far for a large set of countries, including both industrialized and developing countries and over several decades of time, although some partial efforts at reconstructing levels of educational attainment have been developed at a more aggregated level.

Unlike earlier reconstruction efforts that mostly used economic capital accumulation models, this effort is based on demographic multi-state methods that allow vital rates in different educational categories to differ. Starting with only one empirical dataset for each country for the year 2000, we go backward in time and reconstruct earlier distributions by level of education along cohort lines. Since the overall size and age distribution for each country and point in time is given by the population estimates of the United Nations Population Division, the task of this reconstruction effort essentially boiled down to estimating the proportions with different educational attainment for each given five-year age group of men and women over the period 2000 back to 1970¹.

This paper is divided in two sections. The first presents the main features of the back-projections in the frame of the available data and reconstruction exercises implemented by others. The second section details the methodology and the adjustments that had to be made.

2. The Approach of Demographic Back Projections, Existing Data and Previous Reconstruction Efforts

The concept of projecting populations backward in time is not new. Applications have mostly been in historical demography for reconstructing population size and structure for early periods for which no such information was otherwise available. Wrigley and Schofield (1982) developed a specific back-projection method to provide new

¹ This reconstruction exercise focuses strictly on levels of educational attainment, which are measures of the quantity and formal level of education received. Educational quality also has an important effect on human capital. However, datasets such as PISA, PIRLS or IALS based on direct testing of skills are so far only available for a small number of (mostly OECD) countries.

demographic estimates for England for the period 1541-1871 A.D. A method of ‘inverse projection’ had also been developed by Lee (1978 and 1985) to estimate demographic structures in the past. In a different context, the method of demographic back projection has been used widely to estimate HIV incidence from AIDS incidences data (De Angelis, Gilks, and Day 1998, cited in Law et al. 2001) and to estimate the number of dependent heroin users from the observed numbers of opioid deaths and new entrants to methadone treatment (Law et al. 2001).

When trying to collect empirical international data on educational attainment by age and sex over time, it is at first surprising to see how little consistent time series data on levels of educational attainment exist both for developing and developed countries. The main problem hindering the availability of a database that is consistent over time is the definition of the categories for which data has been collected. Several efforts have been made to construct such series using whatever exists in terms of available empirical data. The problem is that the official data from censuses such as those collected by the United Nations Statistical Office and UNESCO are only fragmentary and scattered over time and countries. Despite intensive efforts by UNESCO in terms of harmonizing the data, data collection is still a national responsibility with censuses carried out at different points in time, and the data in the UNESCO databases suffers from all the problems present in the original data as mentioned above. Another difficulty is added by the fact that for the sake of consistency, national data are further classified according to UNESCO’s predefined categories for all countries and the allocation of the census data to the UNESCO categories may have caused some of the observed inconsistency problems. This is complicated by the fact that UNESCO has incorporated changes in their definition of categories according to the changes made by the international standard classification ISCED.

Because of the inconsistent and fragmentary nature of the purely empirical dataset collected from national census information, several attempts have been made in the past to estimate complete, comprehensive and consistent datasets for large numbers of countries. The first and most often used dataset was developed by Barro and Lee (1993, 1996, 2001) who complement the existing attainment data with the somewhat more consistent time series of national school enrolment data at different levels using perpetual inventory methods which help transform accumulated education flows (enrolment) into human capital stocks. This resulted in a widely-used dataset that gives the proportion of the population by highest level attained and mean years of schooling of the entire adult population (by sex but without age details) for 142 economies, of which 107 have complete information at five-year time intervals from 1960 to 2000. The main drawback of the Barro and Lee methodology is that the authors used existing real data and interpolated gaps based on enrolment rates, making the data very sensitive to inconsistencies in the educational categories used, as mentioned above. Similar independent efforts have been made by Kyriacou (1991) Lau et al. (1991), Nehru et al. (1995), De la Fuente and Doménech (2002), and by Cohen and Soto (2001), which in many cases result in quite different estimates of mean years of schooling, with most of the estimates being significantly higher than Barro and Lee. A recent summary of available educational datasets can be found in Cohen et al. (2007) and Bloom (2006). None of the listed reconstruction efforts give the desirable age detail cross classified with the distribution over different educational attainment categories. They also

disregard in their calculations the well-established fact that people with higher education have lower mortality rates, which can have quite significant effects on the educational composition of the older adult population. One common disadvantage of all these exercises (with the notable exception of Barro and Lee and De la Fuente and Doménech for OECD countries) is that the main indicator used is mean years of schooling (MYS). This indicator is used in most of the numerous economic growth regression models that have been produced over the past years. The calculation of MYS is very difficult, however and hides the potentially important effect of educational attainment distributions.

While all these previous reconstruction attempts have made important contributions to the discussion, only our new reconstruction provides full age detail (five-year age groups) cross-classified with the educational attainment distribution for a large number of developing and industrialized countries. Moreover our method is insensitive to the problem of changing educational classifications over time because we only use the classification given for the empirical data in 2000 and project those backward in time. Of course, this does not come without certain assumptions, which we will discuss in shortly in this paper.

3. Our Method

The basic idea of back projection in the context of reconstructing the educational distribution is rather simple: Assuming that the educational attainment of a person remains invariant after a certain age, we can derive, e.g., the proportion of women without any formal education aged 50-54 in 1995 directly from the proportion of women without formal education aged 55-59 in 2000. Assuming that this proportion is constant along cohort lines, it directly gives us the proportion of women without education aged 25-29 in 1970. In a similar manner, the proportions for each educational category and each age group of men and women can simply be moved to the next younger five-year age group as one move back in time in five-year steps. It is important to see that these are not arbitrary assumptions, but truisms under certain conditions. In the above example, the proportions of women without schooling aged 25-29 in 1970 and 55-59 in 2000 must be identical if nobody moves to the category with primary education after the age of 25 and if mortality and migration do not differ by levels of education. This follows directly from the fact that the size of a birth cohort as it ages over time can only change through mortality and migration. In reality we know, however, that mortality tends to strongly vary with the level of education in every country of the world and that migration can do so as well in specific cases. That is why we—unlike earlier reconstruction efforts—will make special adjustments for these differentials.

It is worth noting that we do not have to worry about the level of fertility. In a backward projection, the population increases along cohort lines by accounting for mortality and migration. The level of fertility can be indirectly inferred from the size of the youngest age group but does not enter as a component of change when going backward in time. Beside, we only project the population down to a minimum age of 15 (because we focus

on educational attainment) and also because the age and sex structure (without the education detail) is not reconstructed but directly taken from the UN estimates.

Formally our model can be summarized as follows (see also Box 1):

Starting with $t = 2000$ as the jump-off year for our back projection for which we have a full distribution of the population by age (five-year age groups), sex and level of education (four categories), when there are no transitions between education levels, we go back in time in five-year intervals calculating the same full distribution for year $t-5$ according to

$$N(\text{age} - 5, \text{educ}, t - 5, \text{sex}) = \frac{N(\text{age}, \text{educ}, t, \text{sex})}{\text{SurvivalRatio}(\text{age} - 5, \text{educ}, t - 5, \text{sex})} \quad (1)$$

where

$N(.)$ refers to the number of people in the group defined by $(.)$,
 age refers to the five-year age group starting with age a (e.g., $a=20$ refers to the age group 20-24),
 educ refers to the educational attainment category (see definition below),
 t refers to calendar year t and $t-5$ to five years earlier,
 sex refers to the gender of individuals,
and
 $\text{SurvivalRatio}(.)$, refers to the proportion of people surviving for five years in the country (i.e., combining mortality and migration) in each age-, sex- and education-specific group over the period $t-5$ to t .

The aim of the back projection is to obtain a dataset with the population distributed by five-year age groups, starting at age 15 and with the highest age group 65+, by sex, and by four levels of educational attainment over a period of 30 years from 2000 (base year) back to 1970 in five-year intervals.

The four educational attainment states (ISCED refers to the International Classification of Education) are defined as:

- No education: those who have never been to school and have received no formal education (No Education)
- Primary: those with uncompleted primary to uncompleted lower secondary (ISCED 1)
- Secondary: those with completed lower secondary to uncompleted first level of tertiary (ISCED 2,3,4)
- Tertiary: those with at least completed first level of tertiary (ISCED 5,6).

Box 1: *Summary of the key steps taken in producing our reconstruction results for all 120 countries*

Methodology

Step 1: Find reliable empirical information on the proportions of population by levels of educational attainment for men and women for 5-year age groups for the base year around 2000).

Step 2: Adjust the educational categories, if necessary, to make them comparable across countries.

Step 3: Apply the empirical proportions to the age structure as given by the United Nations Population Division (UN 2005) for the corresponding country for the year 2000.

Step 4: Obtain the *period life expectancy at age 15* for all men and women from the UN general model life table as used for the corresponding country for the period 1995-2000, i.e., the five-year period preceding t (Source: UN 2005).

Step 5: Calculate the corresponding *education-specific period life expectancy at age 15* by using standard education differentials in life expectancy.

Step 6: Obtain *survival ratios* for all five-year age groups above 15 corresponding to each education-sex-specific period life expectancy at age 15 (using the UN general model life table).

Step 7: If there is no empirical information for the closed age interval 65-69 but only for the open interval 65+, the information for 65-69 must be estimated through trend extrapolation.

Step 8: Calculate the number of people N (*age, educ, sex, 1995*) by age (*age* going from 15-19 to 60-64), sex and education living five years earlier (in 1995).

Step 9: Adjust for the transitions to secondary and tertiary education that happen after the age of 15.

Step 10: Convert the number of people by age and education calculated for 1995 ($t-5$) into age- and sex-specific proportions and apply to the UN (2005) estimates of population structure for this year in order to assure full consistency (including adjustments for migration).

.... Go back to Step 4 and repeat the procedure until the year 1970 is reached.

Step 11: Compare the reconstructed results to existing historical data (where available) and identify significant discrepancies and their source.

Step 12: Resolve discrepancies either by finding that the definition of educational categories differed in the other source (the most common problem) or that our assumption of no significant education differentials in migration was violated and make a plausible correction of this assumption.

We chose 2000 as the base year, since the data for 2005 were not available for a vast majority of countries. Our method completely depends on the educational input in the base year. This makes the baseline education-related data very important, since no other inputs on education are introduced during the back projection, unlike earlier reconstruction efforts that often used school enrolment rates. It has also the great

advantage that the educational attainment categories by definition cannot change over time, which has been the main stumbling block for using the empirical UNESCO data and earlier reconstruction efforts. Since our empirical baseline data is always standardized in terms of the age and sex distribution to exactly match the UN data, only the part of the empirical information that refers to the education distributions is of critical importance.

The method as described above seems rather straightforward although several assumptions had to be made since not all the data necessary for the reconstruction was available. The different adjustments and assumptions are presented shortly below. More details can be found in Lutz *et al.* (2007). We categorize the challenges in five main groups.

3.1. Raw Data and their Adjustments

We searched for the initial distribution of the population by sex and age (by five-year age groups starting at age 15 to at least the age groups 60-64 and 65+) in 2000 data and were able to collect the data for 120 countries. Our main sources were national censuses mostly from UNESCO (UIS, 35 countries), but also directly from national statistical agencies (NSA, 28 countries), Demographic and Health Surveys (DHS, 33 countries), Labour Force Surveys (LFS, 8 countries), and Eurostat (16 countries). But even these data were not always in the form we needed. The main irregularities stem from data referring to years slightly different from the year 2000, data that have only 10-year age groups, data where the last age group was lower than 65+, and data with differing educational attainment categories. Detailed adjustments procedures can be found in Lutz *et al.* (2007).

3.2. Estimation of Mortality Differentials by education in life expectancy at age 15

Because the direct measurement of mortality by level of education requires a reliable and comprehensive death registration system, together with information on the education of the deceased and the corresponding risk populations, such empirical data are limited to a few industrialized countries and are virtually absent from the developing world. Hence, based on a sequence of censuses for Brazil, China, France, Kenya, Malawi, Mexico, Uganda, and Vietnam, we found that with reference to the secondary educational category, the average difference in e_{15} was three years less in the no-educational category, two years less in the primary category, and two years more in the tertiary category. This implies a differential in life expectancy at age 15 of one year between the lowest two categories and of two years each between the highest two categories, i.e., five years between the no education and the tertiary educational categories.

3.3. Dealing with the Open-Ended Age Group

One problem that is common to all back-projection efforts is the fact that in all empirical datasets, the highest age group is usually an open one, such as 65+ as is the standard in our baseline data. We took advantage when countries had more information about the older age groups, such as India, which has information up to the age group 80+. At every back-projection step, the task is to estimate from the given open interval 65+ the proportions in different educational attainment groups for the age group 65-69, which, after this step, will become the age group 60-64. For other countries, we used a procedure to derive these proportions for the 65-69 age group based on the extrapolation of the trend as derived from the proportions in the younger age groups. This procedure is done in several iterations to make sure the estimates are consistent with the known education proportions for the highest open age group (65+). While doing so, we also consider that proportions always lie between 0 and 1, and that the sum of the proportions in each age group must equal unity.

3.4. Age at Progressing to Higher Attainment Category

When going back in time and to younger age groups of this same cohort, the only question that we need to worry about is the timing of transition. Since in this reconstruction effort we only go down to the 15-19 age group as the lowest age group for which we reconstruct the data, transitions that typically happen before this age need not be of concern here: the transition from the category no formal education (E1) to that with some primary education (E2). But the issue already becomes more problematic for transitions from primary (E2) to the completed lower secondary education (E3) and completed tertiary categories, where a certain proportion is expected to still happen between ages 15 and 19. For the transition to completed lower secondary (E3) which in most countries typically happens around the age of 14, it is assumed that three-quarters of the transitions happen before the age of 15, but one-fourth of the transitions happen in the 15-19 age group. The transitions to completed tertiary (E4) clearly can happen in a broad range of age groups although typically, most of the transitions to the tertiary educational attainment category that occur by completing the first level of tertiary (generally called a Bachelor degree) happen around the age of 22. However, due to repetition and late or delayed entry into the education system, significant proportions of people complete the first level of tertiary at later ages. First, we look at the tertiary category in 2000 across the age groups to find the age group with the highest proportion. If the peak is in the age group 20-24, we assume that all the people who had tertiary education in the age group 20-24 had secondary education in the age group 15-19. If the peak is in the age group 25-29, we assume that two-thirds of those who have tertiary education in this age group were in the secondary category at age 20-24, i.e., had not yet completed their education at this age. By the age group 15-19, all are assumed to be in the secondary category. If the peak turns out to be in the age group 30-34, we assume that one-third of them were still in the secondary category at age 25-29, two-thirds were in the secondary category at age 20-24, and all of them were in the secondary category at age 15-19.

3.5. Validation of Results

Finally, one activity that has only started under the project described here is the validation of our reconstructed results against all the empirical data that are given by old censuses (mostly from the UIS database) as well as older surveys and national series of school enrolment rates at different levels. In this first round of validation we applied two clear criteria to identify significant discrepancies: If our reconstructed proportions, at any level, age group or point in time, deviated by more than 5 percentage points or by more than 20 percent on a relative scale from the other data source, it was classified as an outlier that needed further attention. We then made an in-depth analysis for all the outliers to try to determine the source of the discrepancy. In many cases we could resolve the problem either by finding that the definition of educational categories differed in the other source (the most common problem) or that our assumption of no significant education differentials in migration was violated and we could make a plausible correction of this assumption. A handful of cases remained unresolved and since the discrepancies were significant, we decided to remove these countries from our dataset. These countries are not part of the 120 countries presented here, which still represent 93 percent of the world population.

4. Outlook

This paper gave an overview of the demographic back projection method that was used to estimate a new comprehensive and detailed dataset on human capital by age and sex. This reconstruction exercise resulted in an unprecedented amount of detailed and consistent information for levels of education by age (for five-year age groups) and sex for 120 countries over three decades, from 1970 to 2000.

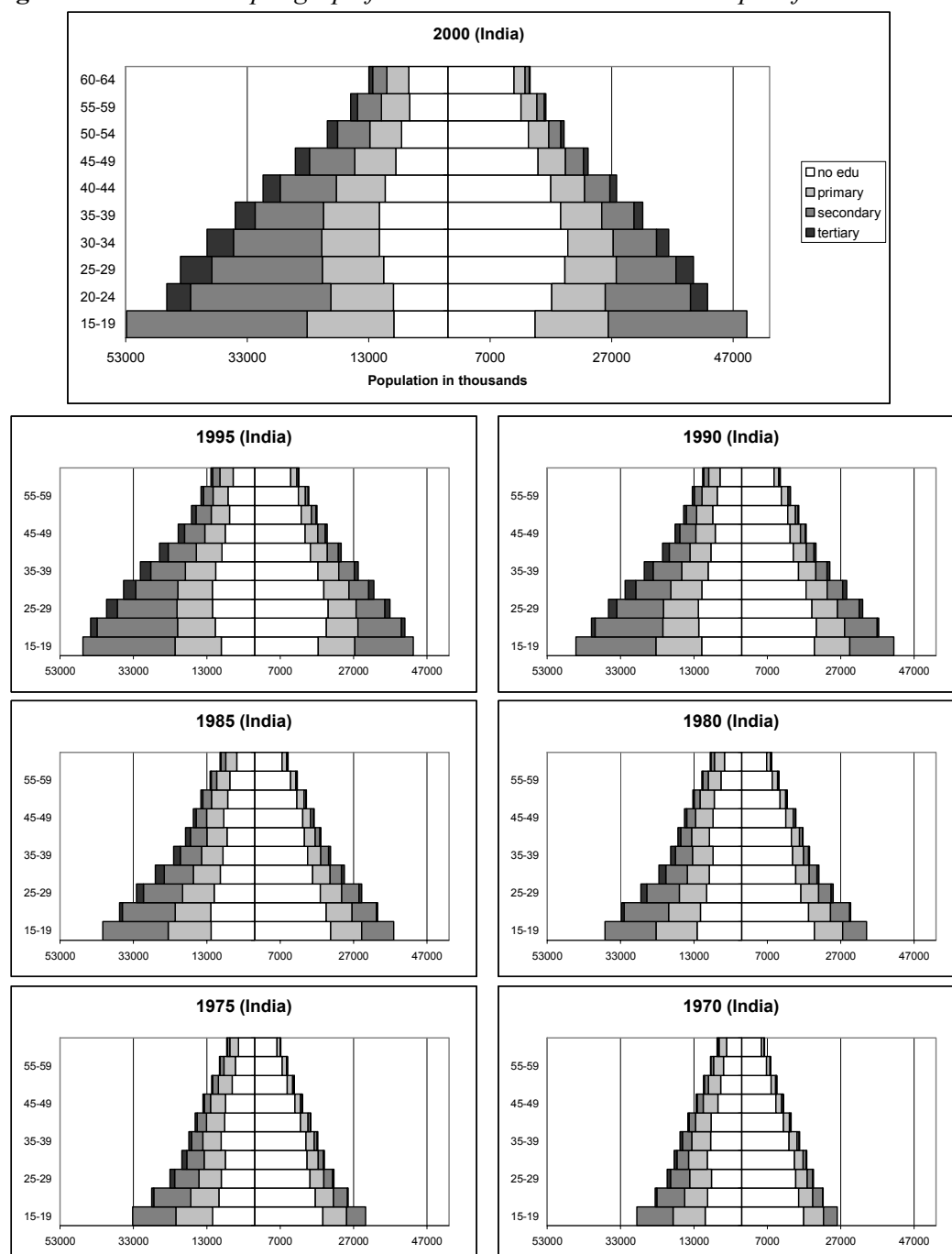
Table 1: *Standard output matrix for the example of the absolute number of Egyptian women in the year 1980*

Females	No Edu.	Primary	Secondary	Tertiary	MYS
1980 15-19	1056.7	451.1	695.0	0.0	4.2
20-24	972.4	561.6	444.5	19.1	3.8
25-29	946.7	440.4	254.6	39.6	3.2
30-34	840.1	260.2	164.3	32.9	2.7
35-39	714.8	189.2	86.5	29.5	2.2
40-44	728.7	159.0	77.9	13.6	1.8
45-49	722.5	121.7	54.8	9.3	1.4
50-54	653.6	84.7	34.9	5.7	1.1
55-59	557.8	55.4	20.7	3.3	0.8
60-64	449.0	34.0	11.5	1.7	0.6
65+	898.1	40.9	11.6	1.6	0.3
15+	8540.4	2398.3	1856.5	156.4	2.5
25+	6511.3	1385.6	717.0	137.3	1.8

Table 1 gives an example of one of these standard matrices for Egyptian women in 1980. Each of these matrices has the age dimension along the vertical axis going in five-year age groups from the population aged 15-19 to the highest category 65+. At the very bottom, it lists the sum of the population 15+. In order to make it directly comparable to the Barro and Lee data, we also included the aggregate age group 25+. Along the horizontal axes, it lists the four educational attainment categories considered. At the right margin, as a summary measure across educational categories, the matrix lists the age-specific mean years of schooling. In the lower right corner, we find the summary

measure along both dimensions, which are the mean years of schooling for the entire population above age 15 (and above age 25). Figure 1 presents an example of reconstructed data in the form of population pyramids for India.

Figure 1: *Standard output graph for reconstructed data. The example of India*



This paper gave an overview of the demographic back projection method that was used to estimate a new comprehensive and detailed dataset on human capital by age and sex. Together with the first round of the validation exercise as described in the previous

section, this constitutes Version 1 of our dataset. This version will not yet be widely and “officially” disseminated—before this we plan to conduct further validation exercises—but it is ready to serve as an input for all kinds of studies, particularly for the study of issues that education is assumed to influence positively e.g. Health and survival, Fertility levels. The greatest immediate interest in these education data clearly comes from scholars of economic growth. Most economists interested in these issues long thought that they simply have to live with the highly unsatisfactory data situation and there cannot be any further improvements in the available database because they are of a historical nature and one cannot go back in time and collect new empirical data for these past periods. The fact that certain demographic methods (unknown to most economists working in this field) are now able to reconstruct such detailed historical data is a good example of the benefits of interdisciplinary collaboration and cross-fertilization.

5. References

- Barro, R. J., Lee J.W. (1993) International comparisons of educational attainment, *Journal of Monetary Economics*, 32(3): 363-394.
- Barro, R. J., Lee J.W. (1996) International measures of schooling years and schooling quality, *American Economic Review*, 86(2): 218-223.
- Barro, R. J., Lee J.W. (2001) International data on educational attainment: Updates and implications, *Oxford Economic Papers*, 53(3): 541-563.
- Bloom, D. E. (2006) *Measuring global educational progress*, American Academy of Arts and Sciences, Cambridge, MA.
- Cohen, D., Soto, M. (2001). Growth and human capital: Good data, good results, *CEPR Discussion Paper No. 3025*, The Centre for Economic Policy Research, London.
- Cohen, J., Bloom D. E., Malin, M. B (Eds) (2007) *Educating all Children: A global agenda*, American Academy of Arts and Sciences, Cambridge, MA.
- De Angelis, D., Gilks, W. R., Day, N. E. (1998) Bayesian projection of the acquired immune deficiency syndrome epidemic, *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 47(4): 449-498.
- De la Fuente, A., Doménech, R. (2002) Human capital in growth regressions: How much difference does data quality make? An update and further results, *CEPR Discussion Paper No. 3587*, The Centre for Economic Policy Research, London.
- Kyriacou, G. (1991) Level and growth effects of human capital: A cross-country study of the convergence hypothesis, *Working Paper 91-26*. C.V. Starr Center for Applied Economics, Department of Economics, New York University, New York, NY.

- Lau, L. J., Jamison, D. T., Louat, F. F. (1991) Education and productivity in developing countries: An aggregate production function approach, *Policy Research Working Papers, WPS 612*, The World Bank, Washington, D.C.
- Law, M., Lynskey, M., Ross, J., Hall, W. (2001) Back projection estimates of the number of dependent heroin users in Australia, *Addiction*, 96(3): 433-443.
- Lee, R. D. (1978) *Econometric studies of topics in demographic history. A dissertation in economics*, Harvard University, Arno Press, Washington, D.C.
- Lee, R. D. (1985) Inverse projection and back projection: A critical appraisal, and comparative results for England, 1539 to 1871, *Population Studies*, 39: 233-248.
- Lutz, W., Goujon, A., K.C., S., Sanderson, W. (2007) Reconstruction of Populations by Age, Sex and Level of Educational Attainment for 120 Countries for 1970-2000, *IIASA Interim Report IR-07-002*, International Institute for Applied Systems Analysis, Laxenburg, Austria.
- Nehru V., Swanson E., Dubey A. (1995) A new database on human capital stock in developing and industrial countries: Sources, methodology and results, *Journal of Development Economics*, 46: 379-401.
- UN (2005) *World population prospects: The 2004 revision*, United Nations, Department of Economic and Social Affairs, Population Division, New York.
- Wrigley, E. A., Schofield, R. S. (1982) *The population history of England, 1541-1871: A reconstruction*, Harvard University Press, Cambridge, MA.