Educational attainment in the 20th century.

Using data from historical censuses and statistical yearbooks to reconstruct and validate a new global dataset on education.

Jakob Eder*, Markus Speringer*, Anne Goujon*, Samir K.C.*,
Michaela Potančoková*, and Ramon Bauer*+

*Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/ÖAW, WU),
Vienna Institute of Demography (Austrian Academy of Sciences); +University of Vienna, Department of Geography and Regional Research

Corresponding author: Jakob Eder

Address: Welthandelsplatz 2/Level 2, 1020 Vienna, Austria

Email: jakob.eder@oeaw.ac.at

Phone: +43 1 31336 7731

Subject: Extended abstract for a paper to be presented at the 2016 European Population Conference (EPC 2014), Mainz/Germany, 31st of August – 3rd of September 2016

Introduction

Today, the quality of global data on educational attainment is substantial and most countries ensure data collection in a comprehensive way. The data compiled through censuses or representative surveys can be mapped according to a categorization scheme that is internationally recognized and widely used, i.e. the International Standard Classification of Education (ISCED). However, this has not always been the case and the collection of comparable data on educational attainment has been hindered by many factors: different measurements of educational attainment, different schooling cycles, different years of censuses taken, changing definitions, and also by changing political borders. Hence, comparisons across countries, time and cohorts are often difficult or at worst meaningless.

In 2015, the EDU20C project was funded[1] (starting in October 2015) to reconstruct past levels of educational attainment of the population of 30 countries from the beginning of the 20th century up to the present so as to have a complete and harmonized times series of the levels of education of the

[1] Jubilee Fund of the City of Vienna for the Austrian Academy of Sciences (2014-2015 Call)

population by age and sex from 1900 to 2010. It is the first step of a larger endeavour to reconstruct levels of education for most world countries in the 20th century.

Why is this valuable? It serves two main purposes that are interlinked. First, educational attainment is used as a proxy for socio-economic development in many models for which education matters: econometric, demographic, environmental, and technological. Historical data are often used to validate past trends. Hence, these models are decisive as they may influence decisions in terms of future public or international investments. However, the usability of the existing demographic data remains limited in an international perspective, especially in terms of educational attainment levels.

Moreover, and this is be the second purpose of the reconstruction, the 20th century was a century of change and many of the technological and sociological revolutions could not have happened without the massive increases in educational attainment which took place during this time. The available data do not allow for an exploration of the role played by education in the century, which is important from the perspective of overall historical and scientific knowledge but also to fully understand the dynamics of educational development.

State of research

Several attempts have been made to solve the lack of consistent time series on education. They can be classified into two main categories. The first is the collection of datasets on educational attainment from censuses, as carried out by UNESCO (United Nations Educational, Scientific and Cultural Organization) or by IPUMS (Integrated Public Use Microdata Series). Although these exercises often include harmonization measures, they are quite limited and tend to mostly take census data at face value.

The second is closer to what we propose to do here, that is at the reconstruction of time series on past levels of educational attainment of the adult population; the most well-known and cited efforts being that of BARRO and LEE (several versions and 2015). Their latest work published in 2015 also aims at reconstructing educational levels for the first half 20th century (they actually cover the period 1870 to 1945 for 89 countries) on top of the period 1950 up to now. However, BARRO and LEE often rely on data taken at face value without harmonization and/or validation efforts.

On the other hand, LUTZ et al. reconstructed in 2007 levels of education following a new methodology of back projections, for the reconstruction of the levels of educational attainment of some 120 countries with four levels of education. It was extended again in 2015 to include 171 countries and six education categories (WIC 2015). In both cases, the back projections only cover the period until 1970. However, for the latest attempt, the database for the base-year was fully harmonized, selecting the most accurate dataset, correcting for changes in the education system across time, allocating country specific categories to the ISCED 1997 scheme, and other required adjustments as explained in BAUER et al. (2012).

The main drawback of the 2014 dataset (WIC 2015) is that it only covers the period 1970 to 2010 because most of the information on education of the population is reliable until the age of 65 (before the mortality differentials become visible and distort the education composition), hence going further back in time from the original base year dataset becomes challenging. This database, which provides one data point for each country (usually the educational distribution during the period 1998-2010), is the first base for the work envisaged under this paper. Consequently EDU20C proposes to extend the time frame of the projections by harmonizing and validating historical data, thereby providing more base-year data points for the back projections. It is the intention to maintain the six education categories introduced in the WIC 2015 dataset until 1950. For the period of 1900 to 1950 they will be collapsed into four categories: no education, primary education, secondary education, and tertiary education.

Spatial Focus

EDU20C focuses on 30 countries in Asia, the Americas, Europe, and Oceania (see Table 1). Those selected countries have a long history in taking censuses and in general data on education are available after World War II. Further, the countries are representative of different streams of educational development on different continents. The presentation will highlight the main challenges dealing with the different data sources and types of data and will present the first results for available reconstructed countries at the time of the EPC 2016.

Table 1 - Countries of study of EDU20C

Continent	**Countries**	**LIT***	**EDU****	**Countries**	**LIT***	**EDU****
Asia	Japan	NA	1960-2010	Republic of Korea	NA	1966-2010
	Philippines	1903-1918	1990-2010			
Latin America	Argentina	1895-1960	1970-2011	Cuba	1899-1943	1953-2012
	Bolivia	NA	1950-2010	Ecuador	NA	1950-2010
	Brazil	NA	1940-2010	Guatemala	1921	1950-2012
	Chile	1907-1940	1952-2012	Mexico	1900-1940	1950-2010
	Costa Rica	1892-1927	1950-2010	Puerto Rico	1899-1940	1950-2010
	Colombia	1928-1938	1951-2005			
Europe	Austria	1900-1910	1951-2011	Ireland	1901-1911	1966-2011
	Finland	1900-1930	1940-2010	Italy	1901-1931	1951-2011
	France	1901-1954	1962-2011	Norway	NA	1950-2011
	Greece	1907-1928	1951-2011	Portugal	1900-1930	1940-2011
	Great Britain	1901	1971-2011	Spain	1900-1950	1960-2011
	Hungary	NA	1920-2011	Sweden	1930	1950-2011
North America	Canada	1901-1931	1941-2011	USA	1900-1930	1940-2010

Oceania	Australia	1911-1921	1966-2011	New Zealand	1901-1921	1961-2013

Methodology

The harmonization of the collected data requires several simple techniques to deal mainly with problems of allocation of the different education categories to our education grouping, and splitting large age groups into 5-year age groups. Bauer et al. (2012) provide examples of difficulties that can arise and how to deal with those.

The methodology of back-projections of education was developed in 2007 by Lutz et al. and further applied in 2014. It basically relies on a given population by age and sex and reconstructs the share by education taking advantage of one main characteristics of education that the bulk of education is acquired at young ages, mostly until the age of 25 and that education characteristics do not change afterwards. Hence, if person *p* aged 55 had achieved lower secondary education in 1995, we can assume logically that most likely *p* in 1985 at the age of 45 had the same level of education. At the level of the population, it is not so simple because one has to discount for depletion due to mortality and migration differentiated by education. For instance, if out of 100 persons aged 85 in 1995, 50 had more than a primary education and 50 had less than a primary education, it does not mean automatically that the 50-50 share apply to the same group in 1985 when they were 75 years of age. They were more numerous, but their education composition was different because they experienced differential mortality depending on their education – the more educated living longer than the less educated. Hence, the 75 year-olds in 1985 are more likely to be less educated than the 85-year olds in 1995. Something similar can happen with migration in case or large education differentials. In most countries, where migration rates are low, this will not affect the educational composition of the population but in some particular countries it will have to be taken into account.

Since the data on mortality and migration differentials by education most of the time do not exist, we used standard depletion factors by age and sex based on the most recent data (Lutz et al., 2014). That is why the validation is the key element of the present work, where we need to compare the resulting back projections to any harmonized data point where educational attainment was observed (Riosmena et al. 2008; Speringer et al. 2015) and to other reconstruction exercises (Barro and Lee, 2015; de la Fuente and Domenech, 2006; Morrison and Murtin, 2009). When differences are found, they have to be explained and most often the depletion factors will have to be adjusted.

Until the 1950s, the collection of census – and other – data on educational attainment was very limited. Hence, we need to resort to other methods to validate the reconstruction data. We will rely on other indicators that were collected such as illiteracy rates to approximate for instance the proportion without or with little education, as well as enrolment numbers to be translated into attainment using a simplified version of the perpetual inventory method.

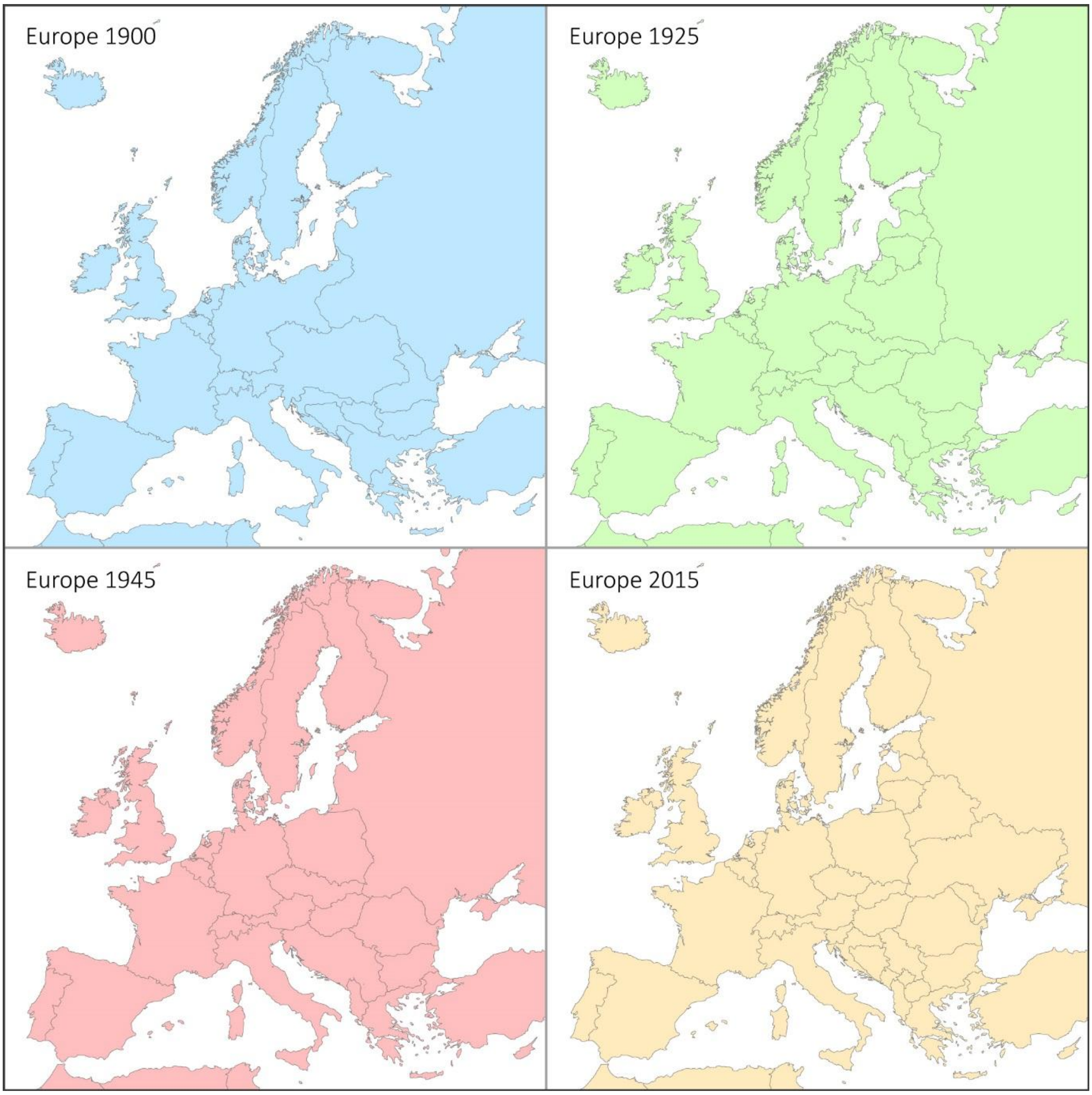

Figure 1: Changing Borders in Europe 1900 – 2015 – Source: EDU20C

Since we will be reconstructing countries from the present time with their modern borders, in cases (see Figure 1) when borders have changed during the century (e.g. Austria, France, Hungary, Italy and Ireland), we will use past regional level data (when it is available) and aggregate them according to present borders.

Data requirements

To reconstruct populations by age, sex, and education we need various demographic indicators which will be used in our back projection model to calculate the time series. The reconstruction only concerns the shares of the population by levels of educational attainment. Once those are estimated, they are applied to the known population. Hence the information on the population by **age** (5-year age groups) and **sex** at all points in time is fundamental. For the period 1950-2010, these data are estimated by the United Nations (latest edition, UNITED NATIONS 2015). For earlier years, we will need other sources, like historical census or statistical yearbook data and furthermore compilations such as the MITCHELL series (2003a, 2003b, 2007) and the publications by FLORA (1983) and FLORA et al. (1987). We will aggregate data (if available by 1-year steps) and split data (if only available by broad age groups) according to regional averages.

In the very best case data on **education** are available by highest level of educational attainment for 5-year age groups and by sex, and with information on the number of grades completed at each level. If these data exist, we will harmonize them to our standards and include them into our model. However, data on educational attainment are usually only available from the 1950s or 1960s onwards and only for more developed regions. These missing data points will be projected by our back-projection model. To be able to validate data points when educational attainment data are only available to a limited degree, we will look into other education indicators instead, that are more readily available from historical censuses and reconstruction exercises, namely enrolment numbers and literacy.

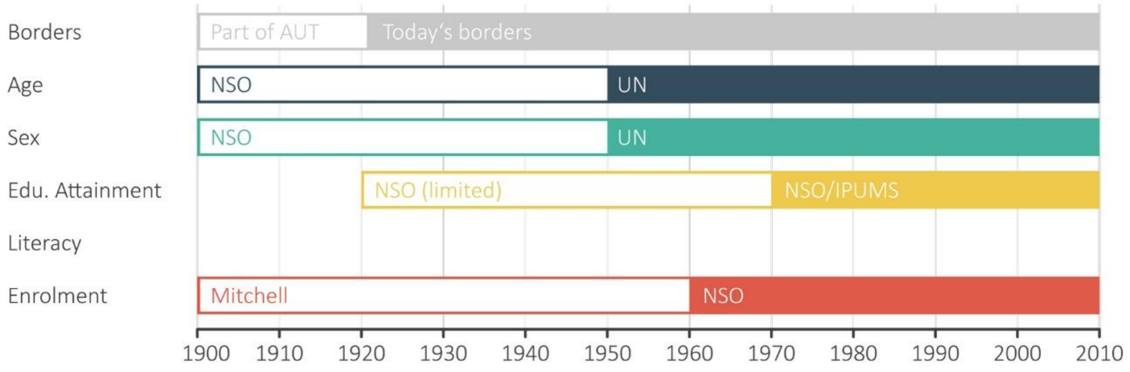

Figure 2: Challenges for back projections by education in Hungary

Figure 2 illustrates the data availability and challenges for a back projection of educational attainment for Hungary. Since 1921 the country exists in its present borders, before it was part of the Austrian Empire and the Hungarian part encompassed areas that belong today to Croatia, Romania, Serbia, Slovakia, and Ukraine. Hungary already carried out independent censuses in that period, but the results do not correspond to present borders. Nevertheless, if available, we will use education data from these censuses as validation points for our back projections.

For the age and sex structure of the population before 1950 we will use reconstructed data by *Statistics Hungary*, which corresponds to today's borders. Data on educational attainment exists since 1920, but until 1970 only four broad education categories are available. Data on enrolment are available from the MITCHELL (2003a) series, which will be used to validate our back projection results.

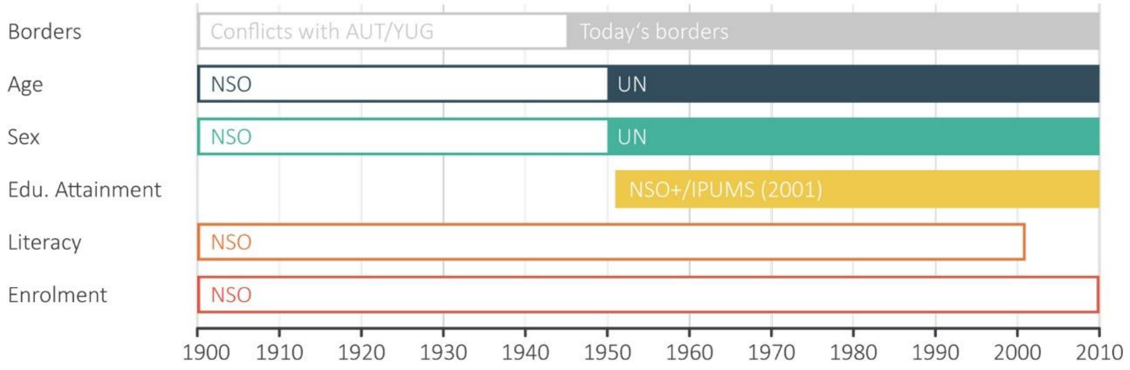

Figure 3: Challenges for back projections by education in Italy

The case of Italy is slightly different from Hungary (Figure 3). Indeed we can rely on census data by age and sex from 1901 to 1951. However, the Austrian census has to be used to estimate the Italian population before World War I. Furthermore, Italy also occupied parts of the later Yugoslavia in the interwar period, which is why the data for 1921 and 1931 also need to be adjusted. Information on educational attainment exists from 1951 onwards, in terms of validation we can rely on a comprehensive time series of literacy and enrolment, available from the censuses and statistical yearbooks of Italy.

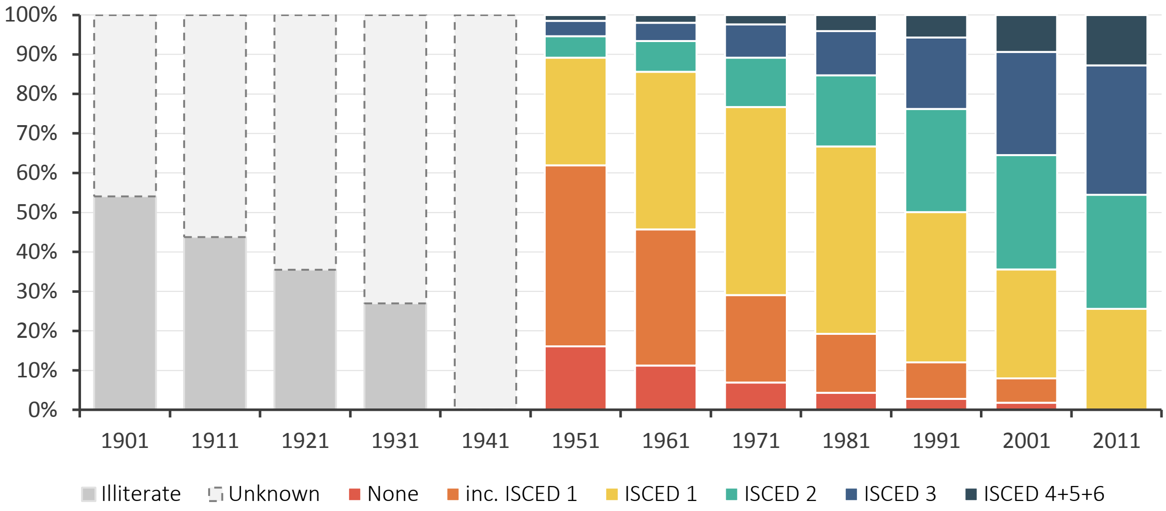

Figure 4: Literacy and Educational Attainment in Italy 1901-2011

As Figure 4 shows for Italy, the second half of the 20th century is for some countries well covered with census data. The crucial point is to use this data, to harmonize it, to add missing education categories, and to extent the time series back through until 1900, when education systems started to expand.

Besides data on age, sex and education, we also need data on overall **mortality**. They will be obtained from the human mortality database for country specific life tables or from regional model life tables, if no specific country data can be found.

To obtain as most historical data points as possible for a wide range of countries, we will use all data sources available. These include national statistical offices as well as other national and supranational organisations. Furthermore, since only a small fraction of historical censuses are available online, we will use printed census reports and digitalize the necessary information on sex, age, and education.

Expected outcomes

The goal of the EDU20C project is to construct a new, detailed, harmonized and comparable dataset on age, sex and education for 30 countries, which is publicly available online. However, we will keep up our efforts also after EDU20C and we plan to expand the dataset to more countries over the next years.

At the time of the EPC 2016 we aim to present results for the countries that are ready providing detailed information on changing political borders, data availability, harmonization efforts, and – most important – results.

References

BARRO, R. J. and J. W. LEE. 2010. A New Data Set of Educational Attainment in the World, 1950–2010. Working Paper 15902, National Bureau of Economic Research. Available at: http://www.nber.org/papers/w15902

BARRO, R. J. and J.W. LEE. 2015. Education Matters: Global Schooling Gains from the 19th to the 21st Century, 1st Edition. Oxford University Press.

BAUER, R., POTANČOKOVÁ, M., K.C., S. and GOUJON, A. 2012. Populations for 171 Countries by Age, Sex, and Level of Education around 2010: Harmonized Estimates of the Baseline Data for the Wittgenstein Centre Projections. IIASA Internal Report (IR-12-016), Laxenburg.

DE LA FUENTE, A., and DOMÉNECH, R. 2006. ”Human Capital in Growth Regressions: How Much Difference Does Data Quality Make?” Journal of the European Economic Association 4(1):1–36.

FLORA, P. 1983. State, economy, and society in Western Europe: 1815 – 1975: The growth of mass democracies and welfare states, I part. Frankfurt am Main: Campus-Verlag.

FLORA, P, KRAUS. F. and PFENNING W. 1987. State, economy, and society in Western Europe: 1815 – 1975: The growth of industrial societies and capitalist economies, II part. Frankfurt am Main: Campus-Verlag.

GOLDIN, C. and Katz, L. F. 2009. The Race between Education and Technology. Harvard University Press.

K.C., S, POTANČOKOVÁ, M., BAUER, R., GOUJON, A. and STRIESSING E. 2013. Summary of Data, Assumptions and Methods for New Wittgenstein Centre for Demography and Global Human Capital (WIC) Population Projections by Age, Sex and Level of Education for 195 Countries to 2100. IIASA Internal Report 13-018.

LUTZ, W., GOUJON, A., K.C., S. and SANDERSON W. 2007. Reconstruction of Populations by Age, Sex and Level of Educational Attainment for 120 Countries for 1970-2000. IIASA Interim Report 07-002.

LUTZ, W., BUTZ, W. and KC, S. 2014. World Population and Human Capital in the Twenty-First Century. Oxford.

MITCHELL, B. R. 2003a. International historical statistics - Europe: 1750 – 2000, 5th edition. Basingstoke: Palgrave Macmillan.

MITCHELL, B. R. 2003b. International historical statistics - The Americas: 1750 – 2000, 5th edition. Basingstoke: Palgrave Macmillan.

MITCHELL, B. R. 2007. International historical statistics - Africa, Asia, Oceania, 1750 - 2005. 5th edition. Basingstoke: Palgrave Macmillan.

MORRISON C. and MURTIN, F. 2009. The Century of Education. Journal of Human Capital 3 (1): 1-42.

RIOSMENA, F., PROMMER I., GOUJON A. and K.C., S. An Evaluation of the IIASA/VID Education-Specific Back Projections. IIASA Interim Report 08-019.

SPERINGER, M., S. K.C, J. EDER, A. GOUJON, M. POTANČOKOVÁ, and R. BAUER. 2015. Validation of the Wittgenstein Centre Back-projections for Populations by Age, Sex, and Level of Education from 1970 to 2010. IIASA Interim Report 15-008.

UNITED NATIONS POPULATION DIVISION. 2015 Revision of World Population Proespects.

WIC. 2015. Data Explorer. www.wittgensteincentre.org/dataexplorer