mosaic
http://www.censusmosaic.org

MOSAIC WORKING PAPER WP2017-001
NOVEMBER 2017

Age heaping patterns in Mosaic data

Mikolaj Szoltysek, Radek Poniat, and Siegfried Gruber

This working paper has been approved for release by:

Mikolaj Szoltysek (szoltysek@eth.mpg.de).

Mikołaj Szołtysek[*], Radosław Poniat[^], Siegfried Gruber[^^]

Age heaping patterns in Mosaic data[**]

Abstract

This paper analyzes the extent and nature of age-misreporting in the Mosaic data, currently one of the largest historical census microdata infrastructures for continental Europe. We use demographic measures known as the age heaping indexes to explore regional, periodic and sex-specific patterns of age misreporting across 115 Mosaic regional datafiles, from Catalonia to Moscow, during Europe's demographic *ancien régime* and thereafter. The paper's second significant contribution is the comparison of Mosaic-based results to those derived from two other big census data projects – IPUMS and NAPP. Beyond this exploratory data analysis, we also investigate possible sources of variation in age heaping across Mosaic data by examining how it relates to variability in socioeconomic, institutional, and environmental conditions. Overall, our systematic inquiry into quality of age reporting in Mosaic consolidates the project's potentially transformative role in comparative historical family demography and suggests some avenues for future research.

[*] Corresponding author; Max Planck for Social Anthropology, Halle (Saale), Germany (szoltysek@eth.mpg.de)

[^] University of Białystok, Poland (r.poniat@gmail.com).

[^^] Karl-Franzens-Universität Graz, Austria (si.gruber@uni-graz.at).

[**] Forthcoming in *Historical Methods: A Journal of Quantitative and Interdisciplinary History*. Sebastian Kluesener (Max Planck Institute for Demographic Research in Rostock, Germany) is acknowledged for deriving GIS-covariates for Mosaic data. We also thank Julia Szołtysek for language editing and three anonymous reviewers for their comments on the earlier version of the paper.

Introduction

Thanks to increased availability of big historical census microdata historical demography has recently witnessed unprecedented expansion of its data infrastructure (Ruggles 2012, 2014, 2016; also Szołtysek 2016). A recent contribution to these initiatives has been the Mosaic project which started in 2009 at the Max Planck Institute for Demographic Research (MPIDR) in Rostock. Drawing upon experiences of a global community of researchers – particularly of giant endeavors such as Integrated Public Use Microdata Samples (IPUMS) and the North Atlantic Population Project (NAPP), Mosaic has recently been recognized as one of the most promising historical census microdata initiatives in the field of historical family demography (see Ruggles 2012; 2016, 151). At present, it contains 115 machine-readable harmonized samples of historical census and census-like microdata derived from early national censuses, as well as from a wide range of individual-level population listings of varying provenience. Mosaic data stretch over a large area of continental Europe from Catalonia to Moscow, between 1700 and 1918, and include almost a million individual records which can be used to compute comparable indicators of co-residence patterns and living arrangements, as well as a range of other demographic measures across multiple locations (Szołtysek and Gruber 2016; Szołtysek et al. 2017b).

Yet, as with many other types of historical data, those included in Mosaic are likely to raise questions about their quality. More frequently than in other sectors of demography, researchers studying past populations (and especially populations of the so-called "prestatistical age") have to use data that are often rough, imprecise, or fragmentary. This problem has led demographic historians to pay special attention to the tasks of data assessment and checking, and to consider these practices "the cornerstone of research in historical demography" (Henry 1968; Del Panta et al. 2006, 597–598).

Among the many ways of examining the quality of population censuses (or other types of census-like microdata) the one that seems particularly adequate is assessment of the extent and nature of deficiencies stemming from age heaping (the rounding of ages) (Szołtysek 2015b, vol. 2)[1]. This specific type of age misreporting constitutes "one of demography's most frustrating problems" (Ewbank 1981, 88). It represents an insidious obstacle in census enumeration because these digit preferences are difficult or even impossible to detect at an individual level (Steckel 1991, 581–82). Given that age constitutes one of the most crucial demographic variables, the presence of age heaping in census microdata inevitably impacts the precision of various demographic estimates, including those related to family and co-residence patterns. The extent to which these distortions occurred can be measured by means of age heaping indices, which assess the tendency in a population to round ages using certain digits. Not unusually, these checks reveal other more general data quality problems, since digit preferences are often linked to other sources of inaccuracy in age statements, and to a general lack of reliability of the age distribution (United Nations 1990, 20). Not surprisingly, then, the updated United Nations' recommendations presented in the "Tools for demographic estimation" (or Manual XI), advise that those assessments are "carried out as a matter of course before embarking on a process of demographic analysis."[2]

Taking account of the prospective cumulative research based on Mosaic a systematic inquiry into quality of its age statistics presents an important and timely task. The better researchers understand patterns of error, omission and bias that stem from age reporting in Mosaic data, the more accurately they will be able to describe the populations they want to analyze with this infrastructure tool. Furthermore, such an exercise seems particularly timely given a recent interest in global assessment of age reporting in the Integrated Public Use

[1] Measuring age heaping is recommended by the UN Manual XI as the first out of three diagnostic steps for targeting "suspect" patterns in the age and sex data distributions in census microdata (see http://demographicestimation.iussp.org/content/get-pdf-book-website).
[2] http://demographicestimation.iussp.org/content/general-assessment-age-and-sex-data

Microdata Series-International (IPUMS-I), thus offering some interesting vantage points for comparisons of historical and contemporary patterns (see Fajardo-González et al. 2014; Sobek 2016, 165-166).

Against this backdrop, a twofold purpose of this paper can be formulated as, first, assessing the accuracy of age reporting and the patterns of digit preference in the Mosaic data in a comparative perspective; and, second, exploring possible sources of variation in age heaping patterns by examining how these relate to variability in socioeconomic, institutional, and environmental conditions across our data. With these goals in mind we organize our paper into five major parts. We start by presenting Mosaic data. Then, we explore these data with the use of age heaping methodology to reveal regional and periodic patterns of age misreporting across continental Europe. In the third step, we compare these patterns to those obtained from IPUMS and NAPP data. Next we discuss the most widespread age heaping patterns in Mosaic and show how these manifest between the sexes. In the penultimate section we present spatially sensitive regression models of the relationship between age heaping patterns and broad variations in institutional, socioeconomic, and locational characteristics across Mosaic locations. We conclude by summarizing our findings and discussing some research agendas for the future.

To the best of our knowledge, this paper is the first attempt at assessing age heaping patterns for such a large corpus of historical census microdata from continental Europe[3]. A second significant contribution of our article is that we explore the relevance of specific contextual characteristics for explanations of historical meso-level variation in age heaping

[3] A'Hearn et. al. (2009) developed a "European age heaping data set" covering over 130 locations in 16 European countries, based on over 300 data samples with the median sample size of about 900 individuals. However, their collection was very heterogeneous, merging eight different source types, e.g. census and census-like data with conscription and muster lists, as well as passenger lists, and death and marriage registers. More recently, Hippe and Baten (2012) developed an age heaping dataset for 570 regions in Europe based on aggregate information derived mostly from published census returns, mainly from the 19th century.

across the continent[4], thus adding to the existing body of literature on the determinants of digit preference (e.g. Crayen and Baten 2010; Baten, Szołtysek, and Campestrini 2017; also Hippe and Baten 2012) based on a unique dataset.

Data

The primary data used in this paper come from the Mosaic project (www.censusmosaic.org). Mosaic started in 2009 at the Max Planck Institute for Demographic Research (MPIDR) in Rostock, drawing upon the experiences of a global community of researchers involved in international data infrastructure projects like IPUMS and NAPP. The major stimulus for Mosaic was a deficiency of existing comparative family history data, which – it was felt – should be overcome were the most pertinent research questions of historical family demography to be systematically answered (Szołtysek and Gruber 2016; Szołtysek 2016). While IPUMS and NAPP projects brought about the unprecedented expansion of census microdata, their coverage remains either confined to the populations of the North Atlantic region or embraces mainly the late 19th and the twentieth centuries (Ruggles et al. 2011). Such a situation poses certain challenges to recovering and understanding the population and family history of continental Europe during the demographic *ancien régime* and early phases of the demographic transition.

Starting from these premises, the Laboratory of Historical Demography at MPIDR has harnessed the energies of a large number of historians, demographers and archivists, who jointly committed themselves to recover surviving census records of historical Europe, including all kinds of historical census-like materials, and not only those with full-count data or samples of national censuses (e.g., church lists of parishioners, tax lists, local estate

[4] A pioneering study in this respect was that of Nagi, Stockwell and Snavley (1973), in which the authors sought to identify social and economic characteristics related to age heaping in census statistics across a range of African populations. Complementary research was done by Stockwell and Wicks (1974), using a sample of sixty-four countries from around the world.

inventories). Pursuing that mission fostered the creation of historical microdata samples for countries available in neither NAPP nor IPUMS, at the same time making researchable a wide range of miscellaneous historical enumerations which remain beyond the scope of these projects. [5]

Figure 1: Spatial distribution of Mosaic data by European regions

Data which have eventually found its way onto Mosaic turned out to be unexpectedly abundant (Szołtysek and Gruber 2016). Table 1A (see Appendix 1), and Figure 1 above show the distribution of the Mosaic regional data across Europe[6]. The Mosaic database currently includes machine-readable harmonized census microdata samples for 115 regions of continental Europe from Catalonia to Moscow, between 17th and early 20th centuries. It consists of individual records for 932,000 persons living in 186,000 family households based on which comparable indicators of co-residence and living arrangements, as well as a range of other demographic measures, can be computed across many previously under-researched areas. The sheer volume, spatial coverage and public accessibility of historical microdata provided within Mosaic means that – compared to early data infrastructure efforts in family history, such as those of the Cambridge Group for the History of Population and Social Structure or the Vienna Database on European Family History – this new initiative offers unprecedented opportunities for comparative analysis of historical family patterns[7].

[5] Prospects of folding Mosaic data into NAPP and/or IPUMS have been discussed at several meetings and workshops, e.g. at the 2nd Mosaic Conference "Residence patterns of the elderly", Hungarian Central Statistical Office, Budapest, Hungary (September 6th – 7th, 2012) and the North Atlantic Population Project Meeting in Copenhagen, Denmark (14-17 April 2016). However, no decisions have been made as yet. Though the general likeness of the Mosaic data structure to both IPUMS and NAPP could make such a data collapse theoretically feasible, challenges remain (e.g. small Mosaic samples compared to IPUMS/NAPP full-count census data and the former varying level of representativeness; see more in Szołtysek and Gruber 2016, 42-44).

[6] Note that several most recent Mosaic datafiles are not reported there.

[7] Data created by these two antecedent initiatives relate to only some parts of Europe, or - as in the English case – they have never been fully computerized, nor made publicly available (e.g. Wall et al. 2004).

While running across many important fault lines in the European geography of demographic regimes (Szołtysek 2015a)[8], the Mosaic database also captures a large share of variation across Europe in terms of environmental features, cultures (including kinship regimes), and socio-economic geography, as well as patterns of economic growth in the early modern and modern times. About two fifths of the 115 datasets contain data collected after 1850, including data from the early 20th century (41.7 per cent); 40.9 per cent of the datasets cover the period 1800-1850, while 17.4 per cent predate 1800. The collection contains both rural and urban regions, although rural regions clearly predominate.

While targeted to assemble a broad set of comparable familial and demographic information across multiple sites, Mosaic demands only minimal data requirements. A data file can be included in Mosaic if: 1) the data source lists individual persons, preferably by name; 2) the data source enlists all of the individuals in a settlement or area, not just the household heads, men, or adults; and 3) it enumerates all of the individuals by clearly delineated residence units (houses, hearths, domestic groups, or households). Moreover, the file can be included in the Mosaic project only if the individual's age, his/her relationship to household head, sex, and marital status were provided by primary sources (the latter two, either explicitly or implicitly). Accordingly, all Mosaic samples have exactly the same content, structure, and organization. In each case, they describe the characteristics of all persons in a locality grouped into co-resident domestic groups, providing a core set of common variables, including the relationship of each individual to the household head, and each inhabitant's age, sex and marital status, which are harmonized across space and time using international standards (Szołtysek and Gruber 2016).

[8] The current scope of Mosaic does not cover the main Iberian and Mediterranean countries, like Portugal, Spain (except for Catalonia), Italy, and Greece, but there are prospects for correcting these deficiencies in the near future.

One of the benefits of such harmonization is that it allows a seamless pooling of data at the micro level from different data sources and time periods into Mosaic, providing that these meet the basic requirements discussed above. At the time when this paper is being completed, a number of new Mosaic datasets is under preparation, covering over one hundred thousand individuals across such diverse locations as the area of Coimbra in 1801 Portugal, the Mediterranean island of Kythera in 1724, the Swiss Canton of Zurich in the 17th and 18th centuries, the 1897 census of Berdichiv county in the present-day Ukraine, and the Russian North of the 1926/27 Soviet Polar census.

Individual- and household-level observations in Mosaic are hierarchical and multilevel, and the recorded micro-level evidence can be aggregated and taken as evidence for lower- as well as larger-scale "structures". The present analysis is anchored at the meso-level and deals with individual Mosaic datafiles agglomerated into 115 "regions", each of them being geo-referenced and linked to a range of detailed GIS-derived covariates.[9] These regions have been further grouped into five larger European territorial clusters meant to capture the varying institutional and socioeconomic characteristics at the time of the census – into "Germany" (to cover German-speaking areas outside of Habsburg territories), "West" (west and southwest of Germany), "Habsburg", "East" (the area of East-Central and Eastern Europe, i.e. the former Polish-Lithuanian Commonwealth, as well as Russia), and Balkans (south of Croatia and Hungary).

The trans-cultural and cross-temporal information contained in Mosaic makes this database particularly suitable for comparative historical demographic research. Such analyses may involve comparative study on the residential arrangements of the elderly (e.g. Szołtysek and Gruber 2014; Gehrmann 2014), explorations of the determinants of spatial variation of

[9] These "regions" are either administrative units used in the respective census, or geographical clusters in the absence of applicable administrative units. As a rough guideline, one "region" should have at least 2,000 inhabitants, and include only urban or rural settlements.

family systems (Gruber and Szołtysek 2012; Ori and Levente 2014; Szołtysek 2016), or inquiries into the complex associations between different elements of family systems in space and time (Szołtysek et al. 2016). A range of harmonized variables from Mosaic was also used to develop a composite measure of differences in sex- and age-related inequalities (the Patriarchy Index) across all Mosaic populations (Gruber and Szołtysek 2016). The majority of recent advancements in this regard include pooling Mosaic datasets with samples of the NAPP data for the exploration of trans-cultural and cross-temporal variation in historical patriarchy levels in Europe (Szołtysek et al. 2017a). The present study offers yet another case for the exploration of this combined data infrastructure.

General patterns of age misreporting

Age structures represent the starting point for any population study. Obtaining information on age structures and plotting it on a graph is often the first step in seeking to understand the nature of processes affecting populations. It also provides an essential guide to considering potential drawbacks and deficiencies in census coverage. Following the established practice, our analysis of age heaping patterns in Mosaic considers age reporting over the age range 23–62 years (Hobbs 2008, 138), which in demographic terms is the most stable population group. Thus delimited, the Mosaic database allows to scrutinize age heaping patterns based on information for 413.000 men and women, between 1700 and 1918.

Figure 2: Reported age by single years in Mosaic data (pooled cross-sections; sexes combined)

Our elucidation of age heaping patterns in the Mosaic data starts with Figure 2 which presents the distribution of reported ages by single years. This distribution reveals that certain

numbers in the Mosaic listings had a powerful attraction. However, the selection of declared ages in the enumerations does not seem to have been entirely arbitrary, since rounding generally occurred in consistent patterns yielding pronounced spikes at the decadal years and secondary spikes at ages ending in five[10]. Single-year age groups one or two digits apart may show significant variations in size. The most "crowded" age was 30, followed by 40 (although there was regional variation in this pattern; see below). This stress on even ages persisted in older age groups, while the preference for reporting ages ending in a five and in other digits declined. Signs of other types of preferential age reporting (although of a much smaller magnitude) - such as the even-numbered terminal digits two, four, six, and eight over those ending in one, three, seven, and nine, can also be spotted in the figure.

A more insightful test of the general reliability of our age statistics can be provided by referring to the most commonly used age heaping index, which measures the degree of preference for or avoidance of ages ending in zero and five (the so-called Whipple's Index). The original index is calculated as the number of individuals between the ages of 23 and 62 whose reported age ends in zero or five, over the expected number of individuals whose ages should end in zero or five in the 23-62 age group, multiplied by 100. The formula for computing the W_h is the following:

$$Wh = \left(\frac{\sum (Age25 + Age30 + ...Age60)}{1/5 \times \sum Age23 + Age24 + Age25 + .. + Age62)} \right) \times 100$$

The United Nations has stated that if the values of Whipple's Index are less than 105, then the age distribution is deemed "highly accurate." If the index values oscillate between

[10] In the overwhelming majority of Mosaic listings people were asked about their age, not about a birth date. Note, however, that even in censuses with a birth date question people not seldom started with age and then counted years backward to obtain year of birth (Buławski 1930).

105 and 109.9, the age distribution is considered "fairly accurate." Meanwhile, values of between 110 and 124.9 are deemed "approximate;" values of between 125 and 174.9 are considered "rough;" and values of 175 or higher are deemed "very rough" (United Nations 1990, 18–19). Below, we examine Mosaic populations through those lenses[11].

Figure 3: Whipple's Indexes for 115 Mosaic regional populations, by the UN typology

Inspection of Figure 3 reveals striking variability in the volume of age heaping in the Mosaic data, which ranges from almost 400 index points to values near (or slightly below) one hundred. Generally, Mosaic data split fairly equally between listings of worse and better quality (62 to 53 listings, respectively). Altogether, 36.5% of Mosaic regional datasets can be considered "very rough" by modern demographic standards and the next 17.4% as "rough", with the former category representing by far the largest relative share of all listings. Meanwhile, the other half of Mosaic datasets scores much better on the quality scale, with approximately one-fifth of all censuses reflecting the expected age structure of the population either "approximately" or "fairly accurately" (15.7% and 5.2%, respectively), and the other 25

[11] Age heaping indexes used in this paper assume linearity and rectangularity in a 5-year range in a population. Any departure from these assumptions may be due either to actual data errors (i.e. age misreporting) or it can be related to historically skewed age patterns caused by a serious fertility decline or high infant mortality associated with single crisis years, or selective migration preceding the census. The extent to which either of the factors was driving age heaping cannot be ascertained directly from our data, because this would require knowing the demographic history of the previous sixty years or so for all 115 populations under study, which is beyond our reach. Still, there are at least two indirect arguments against the deciding role of demographic variation. First, migration effects of young people do not affect the age heaping patterns presented in this paper because the indexes we use rely on ages 23-62 years, and migration generally does not affect only one single year of age, but a broader age group. Furthermore, the general pattern of the saw tooth fluctuations between ages at the decadal years and secondary smaller spikes at ages ending in five reported in Figure 2 has been found across nearly all Mosaic populations with "rough" and worse data. Th consistency with which this trend appears across those 62 populations (see below) means it is unlikely that the heaping patterns in Mosaic might be caused primarily by random variations in demographic events. However, since some of Mosaic datasets are relatively small samples potentially more prone to stochastic variation, the presence of such effects cannot be entirely ruled out, and should be kept in mind when interpreting comparisons attempted in further sections (see also the modeling part).

percent of them indicating little or no age heaping in the data whatsoever ("Highly accurate").[12]

Regional and period-specific clustering of age heaping

Once we move to a regional distribution of age misreporting a complex pattern can be discerned. On the one hand, as Table 1 suggests, data from western and central Europe seem to be much less prone to age misreporting than the enumerations from eastern and southeastern part of the continent. The fact that the upper decile of the Whipple's Index distribution (the very "roughest" values in Figure 3 above) consists entirely of the Balkan and eastern European censuses (with the hot spots of massive age heaping in Albanian, Romanian and some Polish-Lithuanian data), nothing but adds to the impression of a strong eastern cleavage in patterns of age misreporting[13].

Table 1: Whipple's Index for broad territorial groupings of the Mosaic data (sample means)

Figure 4: Age heaping patterns in Mosaic data by macro-regions and quality thresholds

However, Figure 4 shows that such a generalization must be subject to some qualifications. While it is indeed the case that the Balkan (Albanian and Romanian) data are by far outstanding in severe preference for terminal digits 0 and 5, these data also display an

[12] Upon a closer inspection, however, we notice that the latter group is, at least partly, composed of listings in which the Whipple Index' values are below 100. Part of this phenomenon may arise just from random variation in birth cohorts (and hence these values can be set to 100), but it may also indicate the index's difficulty in capturing the forms of heaping concentrated at ages that do not end in 0 or 5. We will come back to this issue below.

[13] At a most general level, these regularities are reminiscent of numeracy patterns established in earlier studies (see Hippe and Baten 2012).

important variation spanning nearly all quality thresholds, from "very rough" to "very accurate" data. Listings from the German speaking areas (outside the Habsburg monarchy) provide yet another case for a similar variegation. Although the German data constitute by far the largest share in the highest quality group of the Mosaic censuses (50% percent of the German listings are in this category), some of the listings from that group (generally the oldest) may occasionally display "very rough" characteristics (though not the "roughest"). Compared to Germany, census listings from the "West" (Catalonia, France, the Netherlands and Belgium) locate somewhat down the quality scale; though they do not display the extremities in age rounding, they also only rarely yield information that would indicate no heaping at all. In consequence, the majority of the censuses in the "West" (12 out of 14 datasets) represent "approximate" to "rough" quality. Meanwhile, as expected, "Eastern" listings lean towards strong age heaping, yet with some rare exceptions illustrating an unusually good quality of age registration. Finally, data from the "Habsburg" area seem to be more polarized than the others between better and worse quality data, with almost no appearance in the intermediate categories.

There are also some noteworthy differences in the age heaping patterns between datasets drawn from societies of one specific historical-geographic area, sometimes even neighboring in space and time. Listings from what nowadays is central Ukraine are exemplary in this regard. The 1791 enumeration from the Zhytomyr County (west of Kiev) yields the Whipple's Index of 131. This stands in a sharp contrast to data from not so far away Podolia at the turn of the 18th century and from the Braclav Governorate in 1795 (both regions some 300 km to the south and southwest from Zhytomyr), for which the corresponding values are as high as 274 and 265, respectively.

While explaining such inter- and intra-regional differences might be a formidable challenge (see below; also ft. 11), at least part of that variation is likely to arise from

differences in the time of enumeration. Figure 5 shows some of the ways along which the relationship between the quality of age reporting and the period of enumeration can be explored.

Figure 5: Age heaping patterns in Mosaic data by time-period and quality thresholds

The figure suggests that as far as the Mosaic datasets are concerned, the increase in the quality of age reporting may not have been unequivocally continuous or linear. First, the median values of the Whipple's Index for the three time periods considered in Figure 5 amount to a curvilinear form, with their maxima among the listings from the earliest and the latest date (265/159 index points, respectively), and a minimum at the intermediate period (108 index points)[14]. One reason for that is the unequal distribution of Mosaic data among time periods, in particular the strong concentration of the "very rough" Albanian data in the later period, as well as the clustering of the high-quality German data in the first half of the 19th century. Compared to the central period, earlier and later data are also more dispersed, though listings of "bad" or "very bad" quality of age reporting can be found in all three time schedules. Interestingly, however, datasets from before 1800 are generally free from the absolute extremities in age heaping exhibited by some later data, including some 20th century enumerations.

Mosaic age heaping patterns in comparison

Whereas age heaping measures presented above constitute a basic tool for assessing the structural-*cum*-quality features of the Mosaic data, their relative positioning in a wider assembly of comparable data might be useful to actually decide about how "bad" or how

[14] This pattern has been confirmed with local weighted regression (LOESS) which fits a smooth curve through points in a scatter plot of Mosaic data arranged by time and the Whipple's Index (available on request).

"good" Mosaic data are. Such a comparison might be meaningful also because it can help determining to what extent patterns similar to those observed in Mosaic datasets are visible in contemporary and historical data which are already widely in use. In order to facilitate such comparisons we proceed in two steps by looking at Mosaic data set side by side with two similar, though bigger, data collections: first, with that of IPUMS, and then NAPP[15]. It needs to be emphasized that while the time dimension is central to this endeavour, it is only insofar as it helps elucidate Mosaic's comparative advantage or disadvantage over other data of a similar structure but more recent provenience, without attempting to reconstruct a long-term evolution of numeracy. For the latter task to be accomplished, longitudinal or time-series data for the same time and place would be needed, something that the mere combination of Mosaic, NAPP and the selected IPUMS data does not allow for.

Looking at Figure 6 we notice an apparently much larger dispersion of Mosaic data over the Whipple's Index distribution compared to both IPUMS and NAPP. That the absolute range of Mosaic data is far greater compared to IPUMS is only partly due to a higher overall fraction of the "very rough" listings in the former dataset, since in this matter the difference between the two datasets is not particularly large (26.9% to 36.5%); it is rather a startling overrepresentation of the age heaping extremities in the Mosaic data that is at stake here. Scores above 300 index points (14% of Mosaic data) are only reached in two IPUMS listings: that of Bangladesh (1991) and Pakistan (1973). On the other hand, Figure 6 clearly shows that the inter-quartile range is fairly similar in both Mosaic and IPUMS data (the median for the

[15] From the IPUMS database we have selected countries with potential for age heaping, basically including all countries from Latin America, Africa and Asia, while omitting the majority of OECD member countries (in each case we used the oldest available country census). A similar approach was followed with regards to NAPP, again giving preference to the oldest available microdata. Accordingly, it was possible to obtain the 18th or early 19th century data for Iceland, Denmark, and Norway for; while for Sweden (1880) and Great Britain (1881) we were forced to use NAPP data from the late 19th century (the data for Great Britain in 1851 were highly clustered, and therefore were not considered). Except for England, where we employed a 10-percent sample, we used 100-percent samples. All of the other data from Great Britain represent 100-percent samples. To ensure better comparability with lower-scale Mosaic datasets, NAPP data were divided into 151 regional populations, following administrative units that were used in the respective censuses and were considered by NAPP.

Whipple's Index is 131 in Mosaic and 125 in IPUMS; the *p*-value statistic for Mood's Median test is insignificant), and that the bulk of index values in both datasets is concentrated in the "approximate" and "rough" categories, though with a tendency to go above those thresholds in Mosaic. Chronological decomposition of the Mosaic data generally confirms that picture by showing that significant quality advantage can be attributed to IPUMS only in relation to the oldest (pre-1800) historical listings (p=0.002). However, the comparison of the "medium old" (1800-1850; 45 datasets) and more recent (post-1850) Mosaic data with IPUMS no longer shows any clear disadvantage of historical listings over contemporary censuses (in both cases the difference between medians is not significant; p=0.259 and 0.132, respectively).

Figure 6: Dispersion of the Whipple's Index values in Mosaic, IPUMS and NAPP data

Overall, these findings suggest that inaccurately reported ages are common to both Mosaic and the 20th century IPUMS data. Although the intensity of age heaping may not have been as pronounced in IPUMS censuses as it was in some historical enumerations, the comparisons above also make clear that at least some of the Mosaic listings report individual ages in a more precise manner than contemporary censuses from developing countries. Such was the case of western and central European Mosaic data from the 1800-1850 period (altogether 38 datasets; $\tilde{x}$ = 109; IQR = 16.4), which indicate a clear superiority of the quality of their age statistics compared to IPUMS data (population medians are significantly different at p=0.043).

Meanwhile, a confrontation with the NAPP data seems overly unfavorable to Mosaic. NAPP data are strongly concentrated over a very narrow range of the Whipple's Index values, and none of the NAPP regional populations considered here exceeded the Index value of 156 ($\tilde{x}$ =110; IQR = 15.6; the *p*-value statistic for Mood's Median test is significant at 0.000). One

reason for generally lower age heaping and its small variation in NAPP is that nearly three quarters of its regional populations used in this comparison came from late nineteenth-century national censuses (1880-1881), with presumably better overall quality of survey-taking, and above all, more consistent surveying practices. Since Mosaic represents a miscellaneous collection of listings emerging from different time periods and different local contexts, covering a stunning cultural variety of historical populations, its internal variation in age reporting cannot be surprising. Accordingly, an attempt at reducing that inner variety of Mosaic data by focusing on listings from the 1800-1850 period, and specifically on those confined to western and central Europe (38 datasets already referred to above), results in measures of dispersion in age heaping much closer to those of NAPP (medians are not significantly different).

Of course, some early NAPP censuses score notably better than the majority of the Mosaic listings. First to mention in this regard is the 1703 census of Iceland, with the W_h value as low as 112. Two other early NAPP datasets which seem to outscore Mosaic in the quality of age reporting include the census of Denmark from 1787 (N=838,623) and the Norwegian census of 1801 (N=878,073). All 21 regional populations from Denmark except for one reveal the Whipple's Index of 100, basically indicating no heaping whatsoever on 0 or 5, and hence a "highly accurate" data quality according to the UN criteria. Twenty Norwegian regions are nearly identical in this respect, again suggesting an absolutely "accurate" age reporting by the same standards. Meanwhile, there are only two Mosaic datasets from the 18th century (out of 23) of a roughly comparable quality: the listings from the Prussian Silesia from the 1750s (N=12,265) with W_h of 106; and the 1700 *Status Animarum* of the county of Vechta in the north-western corner of Germany (N=10.987), with W_h just below 100 (98).

However, in gauging this apparently outright quality of the early NAPP censuses when compared to Mosaic it is important to remember that the original Whipple's Index is a fair

and reliable measure of the quality of age returns only when the attractions on ages ending by 0 and 5 are important. By the same token, this measure cannot reflect completely the quality of age reporting once the attractions on 0 and 5 age-digits reduce, or are accompanied by different kinds of digit preference (Spoorenberg 2007).

Figure 7: Patterns of digit preference in the NAPP data, by means of Whipple's and the Total Modified Whipple's indexes

In order to explore this issue further, we have relied on a recent refinement of the original Whipple's Index, known as the Total Modified Whipple's Index, designed by Spoorenberg (henceforth W_{tot}; see Spoorenberg 2007). This measure, which builds on earlier work of Noumbissi (1992), takes account of preference for and avoidance of all ten digits rather than only those based on rounding one's age to a number ending with a five or a zero, while retaining the same basic principles of the original Whipple's Index (linearity and rectangularity over a 5-year age range and the 23-62 age interval). The total modified Whipple's Index is written as follows:

$$W_{tot} = \sum_{i=0}^{9} (|W_i - 1|)$$

- basically taking the sum of the absolute differences between *Wi (*the digit-specific modified Whipple's Index for each of the ten digits 0-9 developed by Noumbissi) and 1, counting all differences as positive. If there is no age preference, then W_{tot} equals zero. The theoretical maximum of W_{tot} is 16 which will be reached if all persons report their age with the same final digit (for example, all ages ending in 4 (24, 34, 44, etc.). However, such a possibility will never be found in any population.

A visual representation of the application of this measure to the NAPP data is provided in Figure 7. The figure demonstrates clearly that the Danish and Norwegian censuses stand apart from the rest of the NAPP data by revealing the presence of major heaping on digits other than 0 and 5. With values of the Whipple's Index set approximately at 100, i.e. indicating "highly accurate" age distribution, most of the Danish and Norwegian regional datasets still display relatively elevated values of $W_{tot,}$ which upon closer inspection turned out to be caused by over-reporting the ages ending with 9, 5, and 3. The reasons for this atypical pattern can be very complex and their exploration is beyond the scope of this paper. It must suffice to stress that the same pattern prevailed among men and women, so any reasoning referring to the specificities of the conscription system alone cannot be validated. Furthermore, neither in the Danish nor in the Norwegian case are there grounds to interpret a significant over-representation of ages ending in 9 (even less so on 5 and 3) as reflecting preferences not for the digit, but for a birth year ending in '0' or '5', since those censuses were taken in 1787 and 1801, respectively. Whatever the sources of this bias, it is only by the application of W_{tot} that its presence in the NAPP data can be detected. This, naturally, must lead us away from the otherwise positive assessment of these data achieved only through the Whipple's Index, and suggest that assessing the quality of their age reporting merely through the application of that latter measure may be misleading.

Correlation between different measures of age heaping

This leads to the question of what kind of age heaping was most widespread in the Mosaic data, and whether – apart from over-reporting the ages with terminal digits 0 and 5 – other types of digit preference could be identified in our data. The scatterplot in Figure 8 shows for Mosaic data the relationship between two respective measures of age reporting: the original W_h and its refinement in the form of $W_{tot.}$

Figure 8: Age heaping in Mosaic data by type of digit preference measure

Figure 8 delivers two important messages. The first is that, unlike in the NAPP data, the scoring of the Mosaic populations on W_h seems to be a good overall measure of the quality of age reporting, as in general this measure is in a very good compliance with the values obtained through the more sensitive measurement (the two measures considered here are nearly perfectly correlated across Mosaic datasets, with Pearson's *r*=.989; significant at the 0.01 level, 2-tailed). This means that a lion's share of age misreporting in Mosaic data indeed results from heaping on terminal digits 0 and 5. Secondly, Figure 8 suggests that all Mosaic listings, perhaps except for one (Istanbul city in 1885), scoring "highly accurate" on the Whipple's Index (below 105) are generally devoid of any accompanying signs of significant heaping on terminal digits other than 0 and 5, and hence can be considered of absolutely good quality as far as age reporting is concerned[16].

Sex differences in age heaping at the regional level

So far we have relied primarily on age heaping estimates for both sexes combined. However, within the general trend of an increase in the number of age heaping studies in recent years, gender-specific studies have also become increasingly common (Manzel and Baten 2009; De Moor and Van Zanden 2010; Földvári, van Leeuwen and van Leeuwen-Li 2012). Against a general wisdom that in traditional societies women were generally well behind men in numeracy (including their ability to report ages properly), these studies found mixed evidence. While we cannot consider all aspects of this debate in this brief section, the

[16] Against this general pattern caution might still be required when assessing age-reporting quality with the W_h with regards to particular Mosaic data files, since a few of them may be prone towards other forms of digit preference than merely those centered on digits 0 and 5 (e.g. Szołtysek 2015b, vol. 2, 853 ff.). Overall, however, such atypical patterns of digit preference are confined to a very small number of Mosaic listings.

exploration of Mosaic data offers an interesting vantage point from which more comprehensive contribution to this discussion might be offered in the future.

Figure 9: Male versus female age heaping patterns across Mosaic populations

Figure 9 summarizes the three key insights of our exploration. First, it shows striking variation across Mosaic datasets with regards to gender gap in age heaping (measured by W_{tot}). Secondly, it reveals what seems to be a predominant pattern across the dataset, namely the female general disadvantage in age heaping: in 67 out of 115 Mosaic datafiles, women rounded their ages more strongly than men. Thirdly, it unravels a clear tendency for the gender gap to increase with the overall intensity of age heaping.

The difficulty with taking these observations at their face value resides, however, in assessing the statistical significance of the posited differences[17]. It is well-known that in small populations some irregularities in age structure are bound to exist (along with the stochastic variations in demographic behavior), thus potentially impacting estimates of age heaping measures. Given the absence of conventional ways of calculating *p*-values or confidence intervals for our preferred age heaping measure (W_{tot}), one way to circumvent the problem was to use the bootstrap methods. The procedure involved drawing 1,000 samples (by sampling with replacement) for each of Mosaic 115 populations and for both sexes within. The results obtained were then used to calculate for each datafile 1,000 W_{tot} values for men and 1,000 W_{tot} values for women from which estimates of gender differences were derived. Standard evaluating technique of creating pseudo confidence intervals was then applied by computing 0.025 and 0.975 percentile. Results are considered significant if their lower and

[17] Földvári, van Leeuwen and van Leeuwen-Li (2012) used data for 4 distinct historical populations to test various hypotheses regarding gender gap in age heaping, but were able to reach statistically significant results only for bigger populations.

upper boundaries are both below or above 0. The outcome of this procedure is visualized in Figure 10[18].

Figure 10: Bootstrapped 95% confidence intervals for gender differences in age heaping across Mosaic populations

Based on the results presented in Figure 10 we can estimate that out of 115 Mosaic populations significant gender differences in age heaping can be ascribed to only 36 of them (see lower and upper parts of the curve in Figure 10). This amounts to 27 cases where females had significantly higher age heaping than men and to 9 cases where the opposite is true.

Using the same procedure, in Figure 11 the intensity of gender inequality in age heaping was assessed for five broad clusters of Mosaic data. As expected, the overall pattern is to a large extent driven by the Balkan data, of which 80 percent exhibit significant gender gaps in age misreporting, notably with a powerful tendency towards female disadvantage. Eastern European populations, though having certain demographic, socioeconomic and cultural commonalities with Southeastern Europe, are very different from the Balkans, if only due to a much more pronounced share of datasets in which male ages were rounded more often than those of women. Interestingly, the "western" and "Habsburg" clusters are – while containing a relatively smaller number of datasets with significant gender differences – clearly skewed towards female disadvantage in age heaping. Meanwhile, the German data illustrate a reverse situation and at the same time represent that fraction of Mosaic data in which gender gap is the least pronounced (only 4.5% percent of these populations revealed significant gender differences).

[18] Significant results were obtained if (1) gender gap tended to be large, and (2) populations were sufficiently big, even after splitting along gender lines. Some Mosaic datasets could exhibit substantial differences between men and women, but their sex-specific populations were too small to exclude the possibility of random variation.

Figure 11: Macro-regional differences in gender inequality in age heaping across Mosaic populations

In summary, it may be posited that in the Mosaic regions characterized with more stringent hierarchical and gender biased organization of the domestic sphere (like in the Balkans), one is likely to find a stronger gender gap in age heaping, though this observation cannot be taken as an iron rule applicable to all Mosaic data. It is notable, for example, that this patterned relationship is much less pronounced in Eastern European populations otherwise often no less patriarchal than the Balkan ones (Szołtysek et al. 2017a).

Overall, we found that gender differences in age heaping across the Mosaic datasets is smaller than might be expected from readings of economic history and demographic literature (e.g. Földvári, van Leeuwen and van Leeuwen-Li 2012; Manzel and Baten 2009). Though it is true that in the majority of Mosaic populations (67 out of 115) women on average heaped their ages more strongly than men, we also obtained proofs that a reversed pattern was far from rare. Focusing on statistically significant results nuances the above observation, but it does not invalidate it. In the group of populations which passed the tests for the significance of difference (36), female disadvantage in age heaping predominates, but it is still far from universal. Whether these findings warrant a revision to the accepted popular image of men evincing a (much) higher numeracy and age awareness cannot be decided at present. A full understanding of the observed conundrum in Mosaic historical data would require a more careful investigation and a different methodology, which go beyond the scope of this paper (see, however, Szołtysek 2015b, vol. 2, 866-880; Szołtysek et al. 2017b; cf. Földvári, van Leeuwen and van Leeuwen-Li 2012).

Age heaping in context

Why did some regional populations have higher levels of age heaping than others? Theoretically, the extent of age heaping can be explained in both "individualistic" and "contextualist" ways (A'Hearn et al. 2016, 3). In the narrower, former sense, the proximate cause of age heaping can be a respondent's ignorance of his/her exact age or his/her lack of numeric discipline. People with lower mental capacities who were not able to accurately determine or remember their ages (and in practice had almost no recourse to written baptismal records), or who lacked numerical discipline, could only give a rough estimate of their age. In making such estimates they likely used the ages of close relatives as points of reference, and would be prone to round their ages, to deliberately understate or exaggerate them, or to fail to understand whether current age meant years completed or the year currently underway (Szołtysek 2015b, vol. 2, 847; A'Hearn et al. 2016, 21).

Practically, however, the variety of ultimate contextual factors is likely to affect such patterns. In their pioneering studies, Nagi, Stockwell and Snavley (1973), and Stockwell and Wicks (1974), found that the magnitude of error in age reporting is closely related to the level of modernization as measured by a host of socio-economic indicators (e.g. the proportion of population economically active, the percentage literate and the proportion of population working in the non-agricultural sector). More recent research has testified to the importance of at least one specific component of that "modernization package", i.e. the degree of schooling a person can receive. It was pointed out that the formal schooling system is likely to improve children's structural thinking skills in general, which in turn might enhance their numeric knowledge and discipline later in life (Crayen and Baten 2010; earlier Ambannavar and Visaria 1975). However, as posited by long-run growth theories (e.g. Galor 2005), expansion in human capital via educational attainment might be mitigated by fertility levels prevailing in a society, since the latter two tend to be negatively associated. Recent studies of

Becker et al. (2010), Klemp and Weisdorf (2017) and Fernihough (2017) have confirmed this relationship in historical populations, showing that areas with higher fertility also had lower levels of school enrollment and that children of parents with lower fecundity were more likely to become literate.

Nevertheless, suggestive examples have also been put forward where variations in age heaping have been partly independent of literacy rates in the population, thus fixing responsibility for inaccuracies in age reporting on statistical bureaucracy and inadequate survey taking procedures (Ewbank 1981, 15). Rowney and Stockwell (1978) showed that illiteracy does not account for all of the skewing of the age data in the Russian Census of 1897, and raised the issue of poor performance of inadequately trained or manipulative enumerators. Szołtysek (2015b, vol. 2, 866-880), using a wide range of census microdata from early modern Poland-Lithuania (now in Mosaic), found that a significant portion of inter-regional differences in age heaping patterns in that region can be explained by different organizing principles of the enumeration process inherent to different types of listings and by the variable ability to monitor, gather and process accurate information by the political-administrative organisms which commissioned them. Recent perceptive analysis of the age heaping variation in nineteenth-century Italy also points in the same direction (A'Hearn et al. 2016).

Other local or regional political-economic factors may also interfere, such as the degree to which regional populations were subjected to rigid tributary forms of governance. Baten, Szołtysek, and Campestrini (2017) argued that in East European areas dominated in the early modern era by the so-called second serfdom (e.g., Cerman 2012), large landowners prevented the establishment of tax-financed public schooling. They also found a statistically significant and positive relationship between the strength of serfdom and the intensity of age heaping (also Szołtysek et al. 2017b).

Furthermore, environmental characteristics, such as rugged terrain, geographical isolation or low population density, may pose significant costs to state or governmental intervention, including constraints on the efficiency of conducting population surveys (Jimenez-Ayora & Ulubaşoğlu 2015). Szołtysek, for example, found rich evidence of the long history of environmental constraints detrimental to proper population counting in what is now southern Belarus. Without neglecting the independent effect of human capital shortages on parts of the local population, he suggested that massive age heaping in that area could be succesfully explained by a failure of the responsible overseers to collect data properly due to hostile biogeographic conditions (Szołtysek and Zuber 2009, 22ff; also Szołtysek 2015b). Consistently with priors, rugged topography frequently represents an obstacle to the construction of transportation infrastructure, while sparse population and the lack of transport system can make establishing and maintaining effective schooling more difficult and costlier. Moreover, as areas with rugged topography and isolated ones may have been more prone than other regions to have maintained their cultural anomalies due to constraints on congregation, communication, and interaction/diffusion, investments in human capital and skill acquisition may have been inhibited in these areas (see Jimenez-Ayora & Ulubaşoğlu 2015; Goldin 2016, 59).

Finally, age heaping could accelerate when age information was not supplied by the responding individuals, but rather by a second party (such as some "significant others", like husband, father, etc.) (Tollnek and Baten 2016); or due to deliberate misreporting caused by distrustful attitudes on the part of the lay people towards the census personnel or their outright resistence to the very surveying process. While at least the latter behaviour seemed to have been part of a repertoire of political resistance tools in many traditional populations (Szołtysek 2015b, 830-831), there are reasons to believe that both attitudes were likely to be more pronounced in more strongly hierarchical societies. As stressed by Putnam (1993) and

earlier by Banfield (1958), in such societies codes of good conduct and honest behavior are often confined to small circles of related people (members of the family, or of the lineage), while outside of this small network, opportunistic and highly selfish behavior is regarded as natural and morally acceptable. By applying the principles of good and evil inside the family or kin group only, such "amoral familism" encourages dealing distrustfully and deceitfully with non-family members, and treating outside politicians, the central bureaucracy, tax collectors and census takers with particular distrust (see A'Hearn et al. 2016 and Patriarca 1996, 88-95, for the Italian case). Furthermore, given an often rigid male- and senior-centred social hierarchy in such societies, the omnipotent position of the oldest male family member in the household could dictate that he provided enumerators with information on the ages of all people living in the household, thus causing additional age reporting inacccuracies (see Szołtysek et al. 2017a). To test some of these hypotheses with the Mosaic data, we calculated a series of OLS regression models with the W_{tot} as our dependent variable at the meso-level of 115 Mosaic regions, jointly for men and women. To meet the regression assumptions, we decided to log-transform our W_{tot} measure. Since our interest resides in describing a pan-European panorama of age-heaping patterns, in all cases we employed regressions with regional weights that help to reduce the influence of the populations that are overrepresented in our dataset (e.g. Germany).[19]

As spatial data are used in these models, the model estimates may potentially be distorted by spatial autocorrelation problems (Anselin 1988; Bivand et al. 2013). One of the underlying assumptions of an OLS regression model is that the sample consists of independently drawn observations. This assumption is often violated in spatial analyses of regional data, as adjacent spatial units are likely to share many similarities. Nevertheless,

[19] These regional weights were computed by dividing the number of populations from each macro region in our database by the number of all researched populations (e.g., number of populations from "Germany" divided by 115). Apart from Germany, these main European regions include "Balkans", "East", "Habsburg", and "West", as in Figure 1.

standard regression models treat these adjacent observations as independent, which could lead to biases in coefficient estimates and derived significance levels.

In order to account for those potential problems, Moran's *I* tests were performed[20]. Since our regressions include regional weights, we decided not to derive the Moran's *I* for the dependent variable, but instead to calculate a base model that includes the dependent variable, the intercept, and the weights. For the residuals of the base model, we then generated the Moran's *I* on the residuals. The obtained Moran's *I* for the model residuals amounts to 0.62 (p= 0.000), which is indicative of high positive spatial autocorrelation. This finding provides confirmation that it is important to control for spatial autocorrelation in our model diagnostics. To determine whether the model is able to account for the spatial autocorrelation pattern present in the dependent variable, we decided to perform for each model Moran's *I* tests on the unexplained model residuals. If these tests report insignificant results, this provides reassurance that the specific model estimates are not substantially biased by spatial autocorrelation.

The independent variables used in the regressions were grouped into institutional, environmental, socio-economic and cultural categories, in that order. To assess potential determinants of the variation in age heaping across Mosaic data we first consider the characteristics of each Mosaic listing using the criteria suggested in the resume of the Statistical Congress of 1853 (Levi 1854)[21] and the rich contextual information from Mosaic data inventories as guidance. Two decisive markers were deployed to classify our listings. The involvement of "special agents, or enumerators" in the census taking (point 4 of the

[20] The Moran's *I* index is very similar to Pearson's product moment correlation coefficient, except that instead of assessing the correlation between the values of two variables *x* and *y* by each unit *i*, it measures the correlation between the values of a variable *x* in each region *i*, with the (weighted) mean value of the same variable *x* in the regions *j* that are adjacent to region *i*. In calculating the Moran's *I*, we considered the five nearest neighbouring regions, derived by calculating the spherical distances between the regions' coordinates. As the regions' coordinates for the Mosaic dataset, we used the population-weighted coordinates obtained from our 1692 Mosaic locations. The Moran's *I* Index can take on values from –1 (strong negative spatial autocorrelation) through zero (no spatial autocorrelation) to +1 (strong positive spatial autocorrelation).

[21] The congress made recommendations and first principal requirements for census taking.

resume) allowed to isolate listings in the conduct of which a clerical or semi-clerical staff had been involved. This group was further divided according to the second distinction referring to a clearly formulated rule of collecting information on a set of individual characteristics (point 5), of which place of birth, date of birth, and occupation were assigned particular importance. Application of these criteria made it possible to capture a gradation of advancement in census management across our data by means of a tripartite division into: 1) "modern state censuses" (Type 2); 2) "semi-modern censuses (Type 1); 3) "premodern censuses" (Type 0; reference category). This division yields a fair distribution of the Mosaic data, with Type 2 covering 38.3% of Mosaic data (44 datasets); Type 1 including 37.4% percent of them (43 datasets); and Type 0 covering 23.5% of Mosaic data (27 items).

Two spatial control variables were included following suggestions made in recent economic geography studies, which argued that an unfavourable geographic location and/or spatial isolation may represent a penalty that provides disincentives to successful administrative operations, including census taking (Diebolt and Hippe 2016; Lopez-Rodriguez et al. 2007). The first of these covariates is terrain ruggedness (Wilson et al. 2007)[22]. The second geographic variable is population potential (Stewart 1942), which accounts for the centrality and the accessibility of a region by determining the size of the population living close to the location of a region. To calculate this variable, we applied spatial weights that give the population living near a given location more weight than a

[22] Information on the ruggedness of the terrain has been derived from elevation data from the GTOPO30 dataset, which is a global digital elevation model (DEM) with a horizontal grid spacing of 30 arc seconds (downloaded 30 and 31 August 2016 from http://earthexplorer.usgs.gov/; files: gt30e020n40, gt30e020n90, gt30w020n40, gt30w020n90, gt30w060n90). We use the Terrain Ruggedness Index as applied by Wilson et al. (2007) by employing the focal function in the R package *raster* (formula provided in the help function of "terrain" in the raster package). The calculation is performed separately for each of the 1692 Mosaic locations that form our 115 Mosaic regional populations. Around each location we included all raster points within a diameter of 7.5km centred on the location coordinates for obtaining our ruggedness measure. Based on the data for the 1692 locations, we derived population-weighted values for our 115 Mosaic regions.

population living farther away[23]. We expect to find a positive relationship between our measures of ruggedness and age heaping, and a negative one for our population potential variable.

To capture potential modernization effects (broadly understood), the proportion of the elderly (aged 65+) in each regional population was chosen as a crude approximation of the living standards in line with the assertion of demographic literature (Rosset, 1964, 209-210, 231), and hence to serve as a proxy for medical progress and improvements in the public health system (in fact, the latter have been closely tied to the rise of statistical thinking and the propagation of numerical skills; see Porter 1999). We thus expect this share to be negatively related with the prevalence of age heaping[24].

Secondly, we use the child-woman ratio (CWR) which is a net (*effective*) fertility measure computed by dividing the number of children under the age of five by the number of women aged 15-49 (see Willigan and Lynch 1982, 102-104; also Haines 1979). Although CWR is a rather rough-and-ready measure[25], it captures fairly well what can be named the

[23] A working example of how the population potential measure is derived, which is based on the same library and commands as used in this paper, can be found at https://cran.r-project.org/web/packages/SpatialPosition/vignettes/StewartExample.html. Since we used raster data, our outcome is closer to the second presented outcome based on grids.
To calculate the population potential, we used population data derived from the History Database of the Global Environment (HYDE), Version 3.2. These are available in 10-year intervals from 1700-2000, and we took the data for 1800: http://themasites.pbl.nl/tridion/en/themasites/hyde/index.html. In obtaining the population potential, we restricted ourselves to areas located between a longitude of 60° west and 60° east, and latitude of 20° and 80° north. We calculated the population potential using the stewart-command in the R library *SpatialPosition* with the following specifications: span=100,000; b=2; typefct= exponential. This operation was done for each coordinate of our 1692 Mosaic locations that form our 115 Mosaic regions. After a series of consistency checks we found that the HYDE population data, although they are often estimates based on variety of assumptions (Goldewijk et al. 2010; see Klüsener et al. 2014 for details on this dataset), are of sufficient quality to allow us to estimate at a European scale whether a Mosaic regional population was located in close proximity to important population centres or in a rather peripheral location.

[24] Naturally, the proportion of elderly might also be influenced by other factors than just human development, of which outmigration of younger individuals and families could be of prime importance. While older people generally tend to report their ages less accurately - thus potentially implying a "mechanical" relationship of this variable with the volume of age-heaping in our data, the fact that the age range on which W_{tot} is based ends on 62 years makes it unlikely that this issue has an impact on our models.

[25] Given the scope and nature of information provided in the Mosaic listings, CWR is the only fertility measure which can be estimated for all our populations without a heavy computing input and parsimonious assumptions about the underlying mortality patterns. Despite certain caveats associated with the use of CWR as a fertility measure , in the absence of more direct information it may be an advisable and efficient index of reproductive behaviour (e.g. Moore 1990; Haines and Hacker 2011; also Gauvreau et al. 2000; Scalone and Dribe 2016).

"burden of children" in a population and hence it can be used to account for a possible "quantity-quality trade-off" in human capital investments (including numerical skills and literacy; see Becker 1960), and as an indication of the constraints on women's mobility based on the sexual division of labour in the household. Furthermore, given that modernization processes and fertility behavior tend to be functionally (though not absolutely) interrelated whereby modernization mitigates high fertility and large family size by encouraging recalculation of the socio-economic values of children, it may stand to reason to consider lower fertility as one of the defining features of economic development, along with higher household income, better physical infrastructure, more advanced technology, and a larger share of the economy from services (e.g. Jayachandran 2015). Given the above mentioned, we assume that the CWR should be positively related to the strength of age heaping[26].

To account for the fact that the data for our 115 regions come from different points in time, we control for the period in which all or most of the data of a regional population were collected. To do so, the following categories are considered: pre-1800, 1800-1850, and after 1850 (reference category). We assume that age heaping would decline over time (Hippe and Baten 2012). Next to this, our regional datasets were distinguished as either urban or rural, and based on whether their respective populations were subjected to serfdom.

In order to account for hierarchical social structures across Mosaic societies we used a recent composite measure known as the Patriarchy Index (henceforth: PI) that reflects varying degrees of sex- and age-related social inequality across different family settings (Szołtysek et al. 2017a; Gruber and Szołtysek 2016). The index combines ten variables grouped in four "domains" – the domination of men over women, the domination of the older generation over the younger generation, the extent of patrilocality, and the preference for sons, into a

[26] In the child–woman ratio (CWR), the relationship between the number of children and the number of potential mothers is usually multiplied by 1,000. However, to avoid small coefficient values in our regression results, we decided to use this ratio without such a multiplication.

composite measure constructed on the basis of information contained in Mosaic data and at the level of resolution of regions as defined in this paper[27]. It has been shown that in the absence of comparative qualitative information PI can be used to account for the strength of familism and is a good measure of strong/weak family ties in historical populations (Szołtysek and Poniat 2017). Previous research has documented a strong, positive and robust relationship between the PI and age heaping across the Mosaic data (Szołtysek et al. 2017b).

Finally, we included dummies for 5 regions of Europe (as depicted in Figure 1) in an attempt to account for unobserved developmental effects, such as the efficiency of bureaucracy, the role of the labor markets, the legal system or the extent of compulsory schooling (with Germany used as a reference category).[28] Descriptive statistics of the variables used in the models are presented in Table 2, while Table 3 presents regression results.

Table 2: Descriptive statistics

Table 3: Regression results

Regression presented in column 1 of Table 3 predicts the age heaping levels by only one independent variable – the type of the listing. In accordance with expectations, it reveals a strong unconditional negative relationship between the volume of age heaping and the advancement in census management. In comparison to more traditional listings ("premodern"; reference category), "semi-modern" and "modern" censuses display lower age heaping. Although "semi-modern" censuses have slightly stronger beta coefficient than modern

[27] Cronbach's alpha for 11 components of the Index equals 0.83 (CI: 0.77-0.83), suggesting that the items have relatively high internal consistency.

[28] We are using these regions as general umbrella terms that allow us to control for some other factors for which detailed historical and place-specific information is hard to get or completely unobtainable. We are aware that these macro-regional dummies are rather crude measures, but we consider this approach justified as we use them simply as controls to explore how their introduction affects the association between the set of other predictors and our dependent variable.

censuses, this difference is not statistically significant. However, Moran's *I* test on the residuals indicates that the estimates for this model might be biased due to positive spatial autocorrelation.

In Model 2, we augmented the analysis with geographical controls and broad socioeconomic measures. As expected, the relative isolation of a population represented by higher terrain ruggedness and lower population potential was associated with stronger age heaping. Specifically, the OLS estimate, which is siginificant at 0.05 level, implies that the decrease of population potential by one percent is associated with 0.8% increase of Modified Whipple's Index. At the same time, a one-per cent increase of ruggedness caused an increase of age heaping by 0.1%. Meanwhile, results for both rurality and serfdom are insignificant. This model, however, also raises concerns about potential violation of the regression assumptions due to spatial autocorrelation.

The importance of geographical factors was reduced and made insignificant in Model 3, in which additional socioeconomic and institutional (socio-cultural) variables were introduced. Especially the Patriarchy Index has now become the most important predictor of age heaping. With each additional point of the PI, the Modified Whipple's Index increased by 9%, which is not only a strongly significant, but also an economically substantial effect. Of the two indicators of the socio-demographic transformation, also the CWR shows a strong positive and quite sizeable association with age heaping. Increase of CWR by 1 results in 190% higher Modified Whipple's Index. Yet we still have to remember that, as it is shown by the standardised coefficients, of those two predictors PI has a greater impact on the age heaping levels in our data. Importantly, reduced values of the Moran's *I* on the residuals reassure that the outcomes of Model 3 may not be affected by spatial autocorrelation.

Model 4 employs the same set of variables as previously, with only one modification. Instead of census-type we have decided to use dummy variables for the time period (these two

variables are highly correlated). The model's results are similar to the results obtained in Model 3. Lower age heaping levels could be observed in more recent censuses, whereas populations with stronger patriarchy and elevated CWR were still associated with higher W_{tot}. Overall, regressions from Models 3-4 explain a substantial part of variations in age heaping across the Mosaic data. The advantage of Model 4 is the further reduction of the spatial autocorrelation, which now no longer poses much of a concern.

These results were further controlled in Model 5, in which dummy variables for European regions were included. Although only in the case of Balkans the new variable was significant, indicating that the censuses from this region had considerably stronger age-misreporting than their counterparts in Germany, controlling for the unobserved regional characteristics has proven the importance of period, patriarchy and CWR for age heaping prediction. The direction of the effect of those variables remains the same as in the previous models.

Finally, Model 6 tests the robustness of our findings with additional regressions based on MM-estimators[29]. The results show only slight and insignificant differences from Model 5, and can therefore be interpreted as indication of a general soundness of the full model (5)[30].

Overall, our regression results reaffirm earlier intuitions about the possible determinants of age heaping, but they also expand them significantly. Age heaping patterns tended to be higher in areas that were more remote, less well integrated, and socio-demographically less developed, other things being equal. They were particularly high in

[29] We used the MM-type regression estimator described by Yohai (1987) and Koller and Stahel (2011), which is implemented in the R library *robustbase* (http://projecteuclid.org/euclid.aos/1176350366). Robust regression is less affected by violations of linear regression assumptions, such as those caused by the presence of outliers.

[30] The potential multicollinearity between predictors was tested with the variance inflation factors (VIF). For all our models and all independent variables considered, the results were below 5 which indicates that our predictors are only moderately correlated. Finally, following suggestions of one of the reviewers we have re-run all the models with the size of regional population as an additional control variable (results available on request). This variable, however, turned out to be highly insignificant in all models which, above all, retained generally the same effect sizes and siginificance levels as when population size was not included. This reassures us that the models presented above are not driven by small populations which may not verify the rectangularity assumption behind the age heaping indexes due to random variation in their age structures.

regions characterized by stronger patriarchal features in domestic organization, and those shaped by attitudes collectively serving to maximise fertility. Familial (cultural), demographic and environmental factors remained significant even after controlling for censuses' institutional framework.

Although our results are mere prolegomena to a more comprehensive understanding of the variation in patterns of age misreporting, they seem to be indicating that elevated age heaping was a corollary of low state penetration, weak institutions, absence of public investments in education (especially in peripheral regions), and poor access to public services and infrastructure; and hence they might be taken to represent yet another aspect of what developmental scholarship termed as "spatial poverty traps" (Bird, Higgins, & Harris, 2010) – territorial conglomerates where "geographic capital" (the physical, natural, social, political and human capital of an area) is low, partly as a result of environmental disadvantage.

Though potential more specific transmission channels may be hard to disentangle with the use our models, the detection of a significant positive association between the Patriarchy Index and fertility, on the one hand, and regional age-heaping patterns, on the other, deserves a more careful consideration. Given that the PI accounts for familistic and hierarchical societies likely to be characterized with the vicious propensity to put family or lineage interests first, it stands to reason that the major channel through which this variable could have impacted age heaping was indeed by creating a favorable *milieu* for distrustful and deceitful dealing with local administration, and with the census personnel in particular. Such reasoning is even more plausible given that the effect of PI remains robust also after controlling for factors potentially influencing the efficacy of the data gathering procedures[31].

[31] In this case, more plausible than ignorance of numbers would rather be reluctance to make an effort to report them accurately.

Nevertheless, given the PI's relevance for social dimensions beyond those associated with distrust and lack of associative life (Szołtysek et al. 2017a), other channels may also be suggested. Since family is a primary arena for socialization, economic cooperation, and transmission of values, patriarchal more stringent hierarchical organization of the domestic sphere may be disadvantageous to the accumulation of human capital (including numerical skills) by some of its members, especially the young and females, by disincentivizing the resource allocation to their education (which might be considered a threat to parental or spousal authority) and by placing powerful constraints on individual agency and mobility. Furthermore, by emphasising loyalty to family, lineage, and kin, a patriarchal family structure may discourage family members from forming cooperative relationships with non-relatives, and thus limits potentially stimulating "peer group effects" on human capital acquisition (Acemoglu, 2002; also Whyte, 1996, 3-4). All these features of patriarchy are likely to negatively affect the accumulation of human capital, including numerical skills crucial for accurate age reporting.

The strong positive relationship between age heaping and fertility can be interpreted along similar lines, at least insofar as piatriarchal structures tend to create incentives for female "overspecialisation" in reproductive, child-rearing, and domestic work at the cost of accumulating other forms of human capital. This female-specific inhibiting factor (Galor and Klemp 2014) – and female age heaping represents a substantial part of our story – may also have its wider societal implications across Mosaic populations. Given that the child-woman ratio is significant in all models, and holding other factors constant, it provides confirmation (if only indirectly and partially) of Becker's "quality-quantity trade-off" hypothesis, by suggesting that the extent to which a society might have been endowed with numerical skills necessary for accurate age reporting was dependent on the prevailing "burden of children" in a population and the sexual division of labour in the household.

Considering that CWR is associated strongly and positively with the volume of age heaping in Mosaic populations, an additional research question might be asked about the relation between the quality of the age statement and the location of a population on the path of its demographic transition. This question, though interesting, is beyond the scope of this paper and must be left for further research. It should be noted, though, that using CWR to interpret fertility change rather than differentials with Mosaic data might be difficult because mortality and fertility effects are confounding in the cross-sectional listings pooled from different times and places (cf. Moore 1990, 33).

General conclusions

The twofold purpose of this paper was to comparatively assess the accuracy of age reporting and the patterns of digit preference in the Mosaic data; and, secondly, to explore the possible sources of variation in age heaping patterns at the meso-level of Mosaic locations.

The paper employs demographic methods to identify and quantify deficiencies in census age reporting that were developed and mostly applied to age data from statistically less developed contexts, i.e. mostly developing countries. What lessons can we draw from applying these methods to Europe's demographic *ancien régime*? Many experts on early modern statistics would be inclined to admit that it represents "a sheer jungle of uncertainties and traps" (Kula 1951, 96), and that the statistical materials of the early modern era differ substantially from those of later ages in that they were collected haphazardly and organized without skill (Henry 1968; Del Panta et al. 2006, 597–598). The accuracy of these records has been commonly viewed as varying depending on the individual predispositions and inclinations of the priests and estate managers responsible for maintaining them, as well as on the attitudes of the respondents themselves, many of whom were illiterate, and who may not have been always keen to disclose their personal information. As a result, the problems that

led to omissions and misreports – e.g., faulty census administration, low levels of education, inaccessible places of residence, reluctance to reveal personal information, and extended enumeration periods – must have been much more severe in the early modern times than they were in modern enumerations (Ruggles and Brower 2003).

Set against those presuppositions, a careful examination of the Mosaic census and census-like listings presents a more nuanced, and generally more optimistic, picture of historical data. We found that Mosaic data split fairly equally between these of worse and better quality according to contemporary UN criteria, and that a substantial portion of Mosaic regional datasets (nearly 46% of them) should not present major obstacles for demographic analysis. In fact, between listings in which heaping at terminal digits 0 and 5 does not appear to be an issue and those characterized by severe age misreporting, there is a substantial fraction of Mosaic data in which reported ages, though not completely accurate, provide a fair approximation of the expected age distribution.

At the same time, our analysis revealed a large inter- and intra-regional, as well as time-wise variation in age heaping in the Mosaic data. This property of Mosaic made its listings also more heterogeneous in terms of the quality of age statistics than both the NAPP and the IPUMS data recalled here. Despite the fact that the intensity of age heaping may not be as pronounced in contemporary censuses from Latin America, Africa and Asia as it was in some historical enumerations from our collection, we found that inaccurately reported ages are common to both Mosaic and the 20th century IPUMS data from developing countries; better still, we established that a substantial portion of the Mosaic listings exhibit a higher quality of age reporting than these contemporary enumerations. This is an important finding which urges us not to overgeneralize the presumed disadvantage of historical census microdata in demographic analysis.

While the confrontation with the NAPP data seems unfavorable to the Mosaic collection, the comparison of more homogenous subsets of the latter clearly withstands the comparative quality tests. Furthermore, we have shown that the superiority of some early NAPP censuses over Mosaic data is more apparent than real, since those listings are still affected by age heaping patterns not accountable for with the most popular age heaping measures.

Our exploration of specific contexts in which historical variation in age heaping might have arisen, though representing mere prolegomena to a truly comprehensive tackle on the subject, contributes to the existing body of literature by pointing out the importance of familial characteristics (or the strength of family ties) as potential determinants of digit preference, and by linking the prevalence of age heaping to female "overspecialization" in reproductive tasks and the sexual division of labor. Future research should strive to identify more specific transmission channels between social, familial and demographic spheres, on the one hand, and the extent of age misreporting, on the other.

The discussion presented in the paper opens the door to several additional research questions, of which the possibility to study age heaping as an indicator of individual numeracy and human capital across Mosaic populations seems particularly attractive. While in this paper age heaping is treated as a source of distortion in age statistics which scholars have to be aware of before embarking on data analysis, new institutional economic historians (e.g. Tollnek and Baten 2016; A'Hearn et al. 2009) have been increasingly interested in using age heaping in self-reported age data as an indicator of basic numeracy. It has been argued that the tendency to round off their ages can serve as a proxy for the degree to which people could count and calculate, and that these *basic numeracy* assessments can provide a potential link to levels of human capital in the past (A'Hearn et al. 2009, 805–806; cf. A'Hearn et al. 2016). Given its scope and timeframe, Mosaic data could potentially offer a very attractive

corpus to examine these issues and to test the underlying methodology of numeracy studies. Such an empirical test based on a large battery of historical census microdata from continental Europe is still due.

References:

Acemoglu, D. 2002. The Theory of Human Capital Investments. Lecture Notes for Graduate Labor Economics, 14.662, Part 1, Chapter 1, MIT.

A'Hearn, B., Crayen, D. and J. Baten. 2009. Quantifying Quantitative Literacy: Age Heaping and the History of Human Capital. *Journal of Economic History* 69: 783-808.

A'Hearn, B., Delfino, A., and A. Nuvolari. 2016. Rethinking age heaping: a cautionary tale from nineteenth century Italy. LEM Working Paper Series, 2016/35, October 2016.

Ambannavar, J.P and Visaria, P. 1975. Influence of Literacy and Education on the Quality of Age Returns. *Demography India* 4: 11-15.

Anselin, L. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht et al.: Kluwer Academic Publications.

Banfield, E. C. (1958). *The Moral Basis of a Backward Society*. Glencoe: The Free Press.

Baten, J., Szołtysek, M., and M. Campestrini. 2017. 'Girl power' in eastern Europe? The human capital development of central-eastern and eastern Europe in the seventeenth to nineteenth centuries and its determinants. *European Review of Economic History* 21(1): 29-63.

Becker, G. S. 1960. An Economic Analysis of Fertility. In *Demographic and Economic Change in Developed Countries*, ed. A. J. Coale, 209–40. Princeton, NJ: Princeton University Press.

Becker, S. O., Cinnirella, F., and L.Woessmann. 2010. The trade-off between fertility and education: evidence

from before the demographic transition. *Journal of Economic Growth* 15: 177–204.

Bird, K., Higgins, K., and H. Harris. 2010. Spatial poverty traps: An overview. CPRC Working Paper 161. London: Chronic Poverty Research Centre.

Bivand, R. S., Pebesma, E., and V. Gomez-Rubio. 2013. *Applied spatial data analysis with R* (2nd ed.). New York, NY: Springer.

Buławski, R. 1930. Projekt drugiego powszechnego spisu powszechnego na tle doświadczeń spis 1921 r. oraz praktyki zagranicznej. *Kwartalnik Statystyczny* 7: 17–151.

Cerman, M. 2012. *Villagers and Lords in Eastern Europe, 1300-1800*. Basingstoke: Palgrave Macmillan.

Crayen, D., and J. Baten. 2010. Global trends in numeracy 1820–1949 and its implications for long-term growth. *Explorations in Economic History* 47: 82–99.

De Moor, T., and J.L. van Zanden. 2010. 'Every Woman Counts': A Gender-Analysis of Numeracy in the Low Countries during the Early Modern Period. *Journal of Interdisciplinary History* 41(2): 179-208.

Del Panta, L., Rettaroli, R., and P. A. Rosental. 2006. Methods of historical demography. In *Demography: Analysis and Synthesis. A Treatise in Population. Volume* 4, ed. G. Caselli, J. Vallin and G. Wunsch, pp. 597-618. Elsevier: Academic Press.

Diebolt, C., and R. Hippe. 2016. Remoteness equals backwardness? Human capital and market access in the European regions: insights from the long run. Working Papers of BETA 2016-32, Bureau d'Economie Théorique et Appliquée, UDS, Strasbourg.

Ewbank, D. C. 1981. *Age misreporting and age-selective underenumeration: sources, patterns, and consequences for demographic analysis*. Washington, D.C.: National Academy Press.

Fajardo-González, J., Attanasio, L., and J.Trang Ha. 2014. An Assessment of the Age Reporting in the IPUMS-I Microdata. Paper submitted for presentation at the 2014 Annual Meeting of the Population Association of America.

Fernihough, A. 2017. Human capital and the quantity–quality trade-off during the demographic transition. *Journal of Economic Growth* 22(1): 35–65.

Földvári, P., van Leeuwen, B., and J. van Leeuwen-Li. 2012. How did women count? A note on gender specific age heaping differences in the 16th-19th century. *Economic History Review* 65(1): 304-313.

Galor, O. (2005). From Stagnation to Growth: Unified Growth Theory. In *Handbook of Economic Growth*, ed. P. Aghion and S. N. Durlauf, 171-293. Amsterdam: Elsevier.

Gauvreau, D., Gossage, P. and L. Gingras. 2000. Measuring Fertility with the 1901 Canadian Census: A Critical Assessment. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 33(4):219-228.

Goldewijk, K. K., Beusen, A., and P. Janssen. 2010. Long-term dynamic modeling of global population and built-up area in a spatially explicit way: HYDE 3.1. *The Holocene* 20(4): 565-573.

Goldin, C. 2016. Human Capital. In: *Handbook of Cliometrics*, ed. C. Diebolt and M. Haupert, 55-86. Heidelberg, Germany: Springer Verlag.

Gruber, S., and M. Szołtysek. 2012. Stem Families, Joint Families, and the European Pattern. What Kind of a Reconsideration Do We Need? *Journal of Family History* 37(1): 105-125.

Gruber, S., and M. Szołtysek. 2016. The patriarchy index: a comparative study of power relations across historical Europe. *The History of the Family* 21(2): 133-174.

Haines, M. 1979. *Fertility and Occupation: Population Patterns in Industrialization*. New York: Academic Press.

Haines, M., and D. J. Hacker. 2011. Spatial aspects of the American fertility transition in the nineteenth century. In Navigating Time and Space in Population Studies, ed. M.P. Gutmann, G.D. Deane, E.R. Merchant, K.M. Sylvester, 37–64, New York: Springer.

Henry, L. 1968. The Verification of Data in Historical Demography. *Population Studies* 22(1): 61-81.

Hippe, R. and J. Baten. 2012. Regional inequality in human capital formation in Europe 1790-1880. *Scandinavian Economic History Review* 60: 254-289.

Hobbs, F. 2008. Age and sex composition. In *The methods and materials of demography (2nd edition)*, ed. J. S.Siegel and D. A. Swanson, 125-173. Bingley: Emerald.

Jayachandran, S. 2015. The Roots of Gender Inequality in Developing Countries. *Annual Review of Economics* 7(1): 63-88.

Jimenez-Ayora, P., and M. Ali Ulubaşoğlu. 2015. What underlies weak states? The role of terrain ruggedness. *European Journal of Political Economy* 39: 167-183.

Klüsener, S., Devos, I., Ekamper, P., Gregory, I., Gruber, S., Martí-Henneberg, J., van Poppel, F., da Silveira, L.E., and A. Solli. 2014. Spatial inequalities in infant survival at an early stage of the longevity revolution: A pan-European view across 5000+ regions and localities in 1910. *Demographic Research* 30: 1849-1864.

Klemp, M., and J. Weisdorf. 2017. Fecundity, Fertility and the Formation of Human Capital. *Economic Journal* (forthcoming).

Koller, M., and W. A. Stahel. 2011. Sharpening wald-type inference in robust regression for small samples. *Computational Statistics & Data Analysis* 55(8): 2504-2515.

Kula, W. 1951. Stan i potrzeby badań nad demografią historyczną dawnej Polski (do początków XIX wieku). *Roczniki Dziejów Społecznych i Gospodarczych* 13: 23-109.

Levi, L. 1854. Resume of the Statistical Congress, held at Brussels, September 11th, 1853, for the Purpose of Introducing Unity in the Statistical Documents of all Countries. *Journal of the Statistical Society of London* 17(1): 1-14.

Lopez-Rodriguez, J., Faina, J., and Jesus Lopez-Rodriguez. 2007. Human Capital Accumulation and Geography: Empirical Evidence from the European Union. *Regional Studies* 41(2):217-234.

Manzel, K., and J. Baten. 2009. Gender Equality and Inequality in Numeracy: The Case of Latin America and the Caribbean, 1880–1949. *Revista de Historia Económica (Second Series)* 27: 37-73.

Moore, E.G. 1990. Fertility decline in three Ontario cities, 1861-1881. *Canadian Studies in Population* 17(1):25-47.

Nagi, M. H., Stockwell, E. G., and L. M. Snavley. 1973. Digit Preference and Avoidance in the Age Statistics of Some Recent African Censuses: Some Patterns and Correlates. *International Statistical Review* 41(2): 165-174.

Noumbissi, A. 1992. L'indiee de Whipple modifie: une application aux donnees du Cameroun, de la Suede et de la Belgique. *Population* 47(4): 1038-1041.

Ori, P., and L. Pakot. 2014. Residence patterns in nineteenth century Hungary: Evidence from the Hungarian Mosaic sample. Working Papers on Population, Family and Welfare, No. 20. Budapest: Hungarian Demographic Research Institute.

Patriarca, S. 1996. *Numbers and Nationhood. Writing statistics in nineteenth-century Italy*. Cambridge, Cambridge University Press.

Porter, D. 1999. *Health, Civilization and the State: A History of Public Health* from. *Ancient* to Modern Times. London: Routledge.

Putnam, R. D. 1993. *Making Democracy Work. Civic Traditions in Modern Italy*. Princeton University Press.

Rowney, D.K., and E.G. Stockwell. 1978. The Russian Census of 1897: Some Observations on the Age Data. *Slavic Review* 37(2): 217-227.

Ruggles, S. 2012. The Future of Historical Family Demography. *Annual Review of Sociology* 38, 423-441.

Ruggles, S. 2014. Big Microdata for Population Research. *Demography* 51: 287-297.

Ruggles, S. 2016. Data sharing in historical demography. In *The future of historical demography: Upside down and inside out*, ed. K. Matthijs, S. Hin, J. Kok and H. Matsuo, 99-102. Leuven: Acco.

Ruggles, S., and S.Brower. 2003. The Measurement of Family and Household Composition in the United States, 1850-1999. *Population and Development Review* 29: 73-101.

Ruggles, S., Roberts, E., Sarkar, S., and M. Sobek. 2011. The North Atlantic Population Project: Progress and prospects. *Historical Methods* 44: 1-6.

Scalone, F., and M. Dribe. 2017.Testing child-woman ratios and the own-children method on the 1900 Sweden census: Examples of indirect fertility estimates by socioeconomic status in a historical population. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 50(1): 16-29.

Sobek, M. 2016. Data Prospects: IPUMS-International. In *International Handbook of Migration and Population Distribution*, ed. Michael J. White, 157-174. Dordrecht-Heidelberg-New York-London: Springer.

Spoorenberg, T. 2007. Quality of age reporting: Extension and application of the modified Whipple's index. *Population-E* 62(4):729-742.

Steckel, R. H. 1991. The quality of census data for historical inquiry: a research agenda. *Social Science History* 15(4): 579-599.

Stockwell, E.G., and J.W. Wicks. 1974. Age Heaping in recent National Censuses. *Social Biology* 21(2): 163-167.

Szołtysek, M. 2015a. Households and family systems in early modern Europe. In *The Oxford Handbook of Early Modern European History, 1350-1750, Volume I: Peoples and Place*, ed. H. Scott, 313-341. Oxford: Oxford University Press.

Szołtysek, M. 2015b. *Rethinking East-central Europe: family systems and co-residence in the Polish-Lithuanian Commonwealth* (2 vols). Bern: Peter Lang.

Szołtysek, M. 2016. Historical family systems and European inequalities: a way forward for the future. In *The future of historical demography: Upside down and inside out*, ed. K. Matthijs, S. Hin, J. Kok and H. Matsuo, 59-62. Leuven: Acco.

Szołtysek, M., and S. Gruber. 2014. Living arrangements of the elderly in two Eastern European joint-family societies: Poland-Lithuania around 1800 and Albania in 1918. *The Hungarian Historical Review* 3(1): 101-140.

Szołtysek, M., and S. Gruber. 2016. Mosaic: recovering surviving census records and reconstructing the familial history of Europe. *The History of the Family* 21(1): 38-60.

Szołtysek, M., Gruber, S., and R. Poniat. 2016. Household Formation and Postmarital Residence: Historical Cross-cultural Perspectives using Mosaic Data. Paper presented at the 11th European Social Science History conference, Valencia, 30 March - 2 April 2016.

Szołtysek, M. and R. Poniat. 2017. Historical family patterns and European cultural and developmental disparities: persistence of the past? Paper presented at the Gender-Governance Link (GGL) conference, Goettingen, July 2017.

Szołtysek, M., Klüsener, S., Poniat, R., and S. Gruber. 2017a. The Patriarchy Index: A New Measure of Gender and Generational Inequalities in the Past. *Cross-Cultural Research* 51 (3): 228 – 262.

Szołtysek, M., Poniat, R., Klüsener, S., S. Gruber. 2017b. Family organisation and human capital inequalities in historic Europe: testing the association anew. MPIDR Working Paper WP-2017-012.

Szołtysek, M., and B. Zuber-Goldstein. 2009. Historical family systems and the great European divide: the invention of the Slavic East. *Demográfia (English edition)* 52(5): 5-47.

Tollnek, F., and J. Baten. 2016. Age heaping-Based Human Capital Estimates. In *Handbook of Cliometrics*, ed. C. Diebolt and M. Haupert, 1-20. Heidelberg, Germany: Springer Verlag.

United Nations. 1990. *1988 Demographic Yearbook*. New York: United Nations.

Wall, R., Woollard, M., and B. Moring. 2004. Census schedules and listings, 1801-1831: an introduction and guide. Colchester: University of Essex, Department of History, <https://www.academia.edu/619532/Census_Schedules_and_Listings_1801-1831_An_Introduction_and_Guide> accessed 24 March 2014.

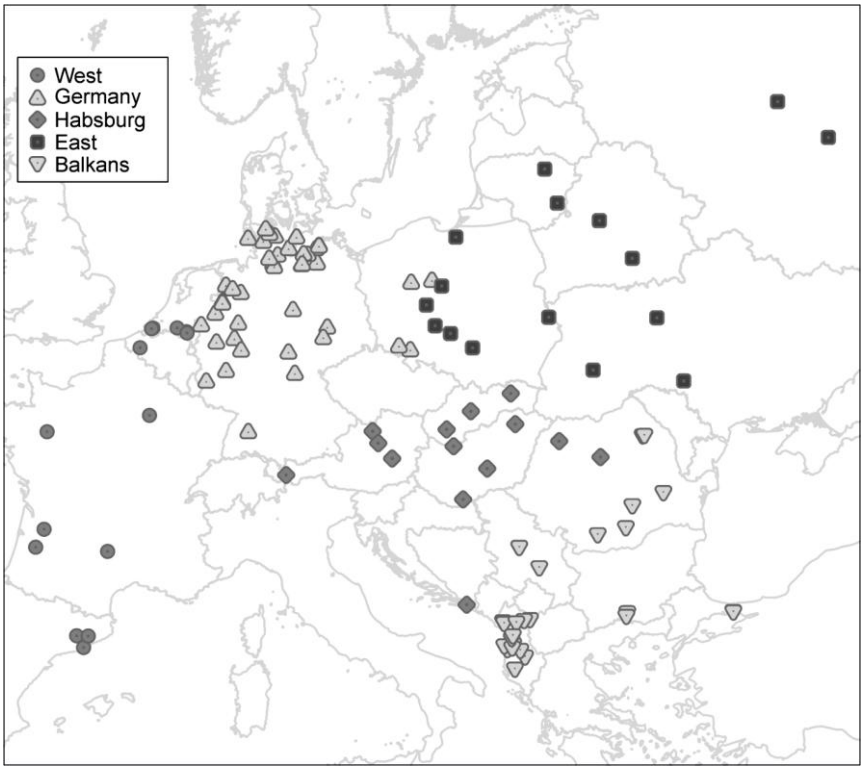
Whyte, M.K. 1996. The Chinese Family and Economic Development: Obstacle or Engine. *Economic Development and Cultural Change* 45(1): 1-30.

Wilson, M.F.J., O'Connell, B., Brown, C., Guinan, J.C., and A.J. Grehan. 2007. Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope. *Marine Geodesy* 30: 3-35.

Yohai, V. J. 1987. High Breakdown-Point and High Efficiency Robust Estimates for Regression. *Ann. Statist.* 15(2): 642-656.

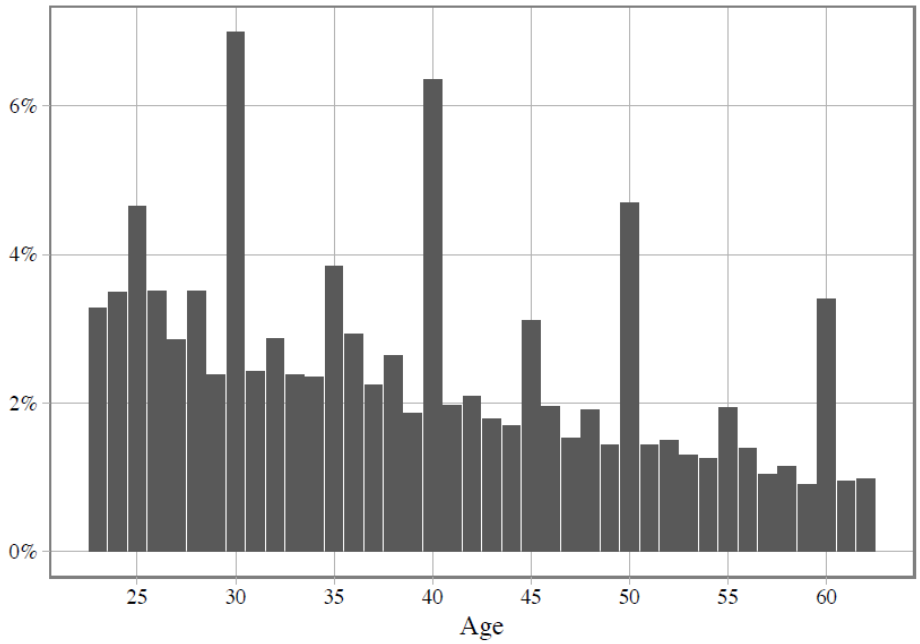
Figures and tables

Figure 1: Spatial distribution of Mosaic data by European regions



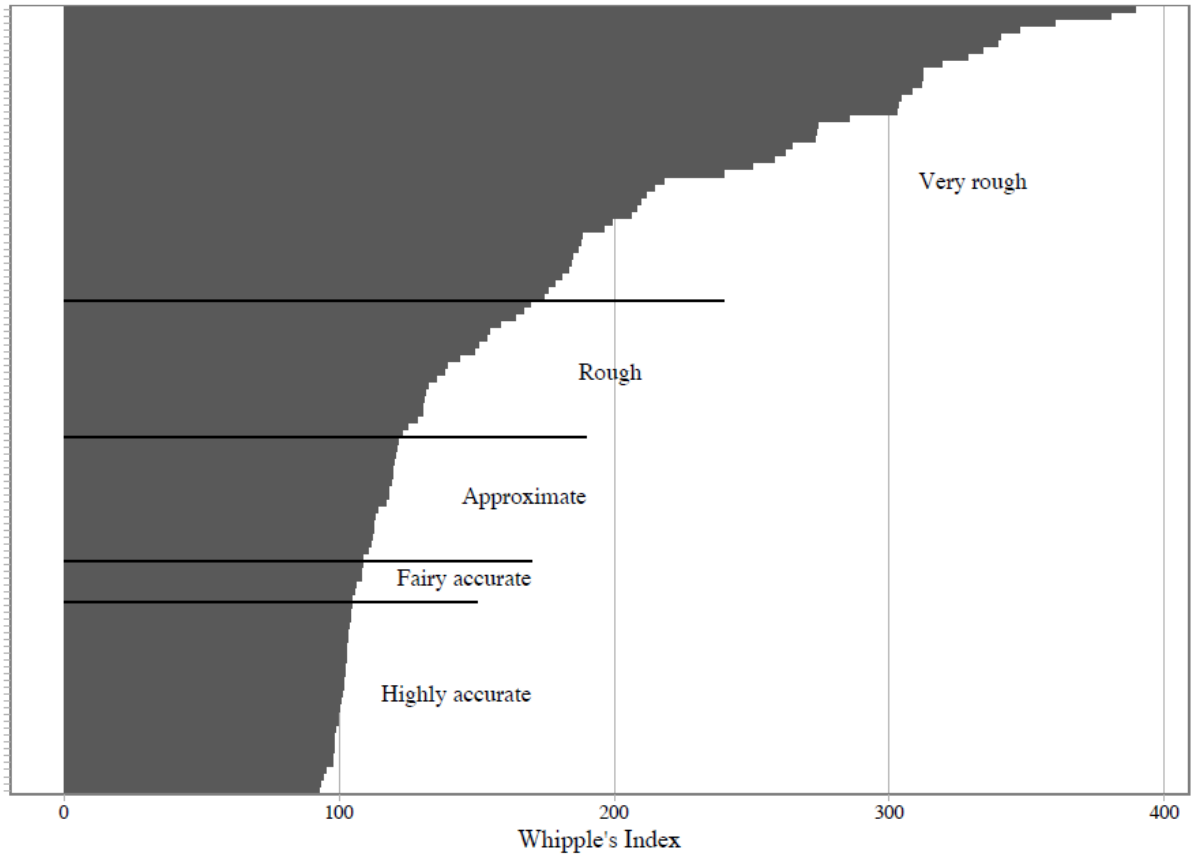
Source: Mosaic.
Note: one icon represents a regional Mosaic datafile.
Map design: S. Kluesener (MPIDR).

Figure 2: Reported age by single years in Mosaic data (pooled cross-sections; sexes combined)



Source: Mosaic datafiles. Data for 207,857 females, and 205,342 men (unweighted data)

Figure 3: Whipple's Indexes for 115 Mosaic regional populations, by the UN typology



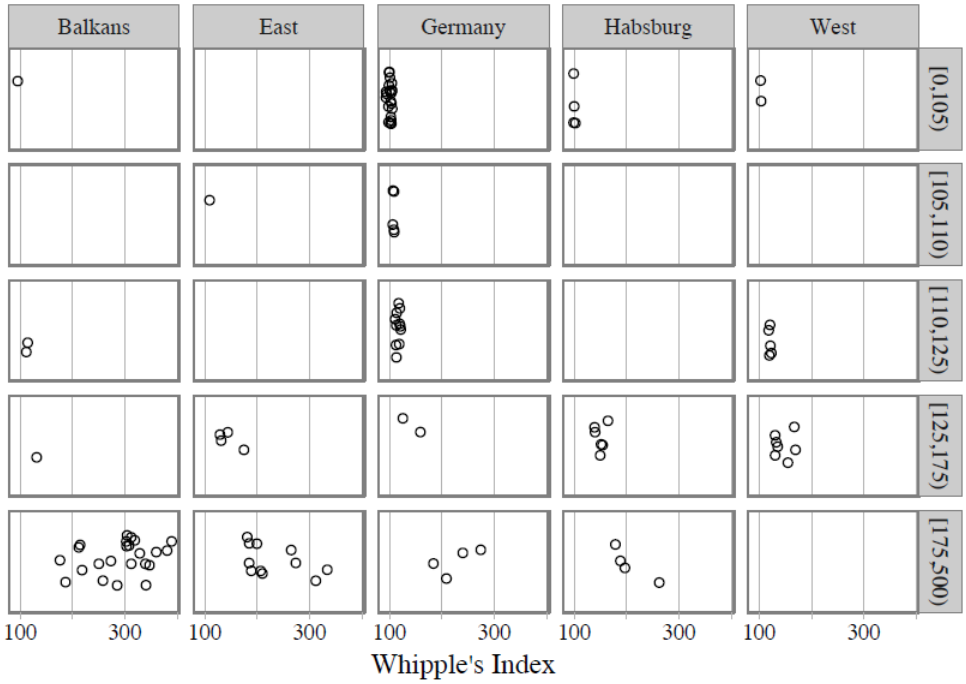
Source: as in Fig. 1.

Table 1: Whipple's Index for broad territorial groupings of the Mosaic data (sample means)

Macro region	N	Mean	Std. Deviation
Balkans	27	265,7	85,2
East	16	201,5	64,9
Germany	44	118,4	37,7
Habsburg	14	151,3	46,1
West	14	130,6	20,5
All	115	170,0	81,3

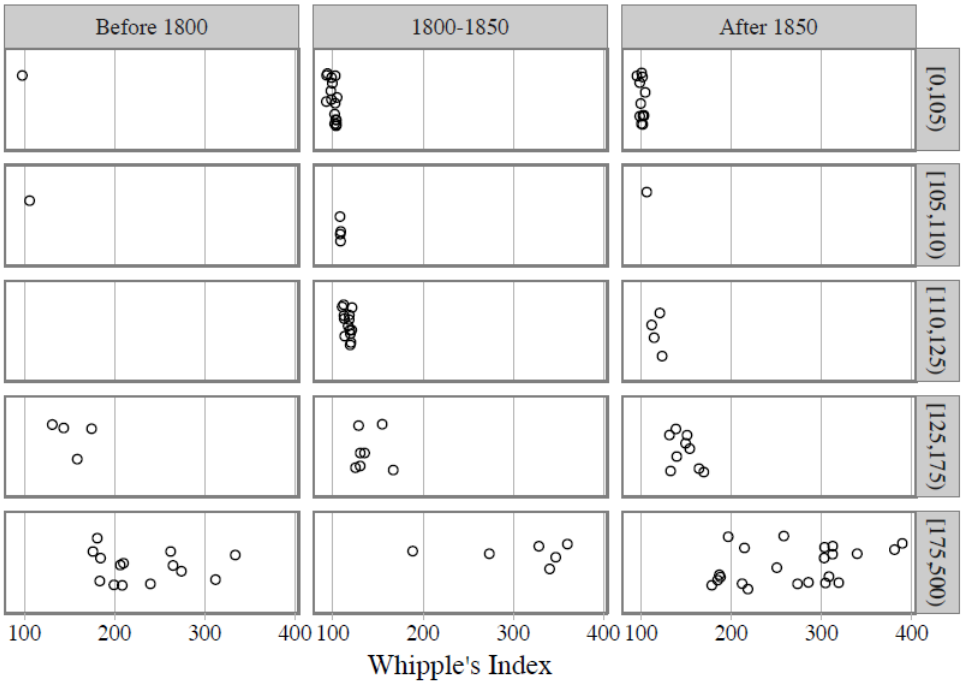
Source: as in Fig. 1.

Figure 4: Age heaping patterns in Mosaic data by macro-regions and quality thresholds



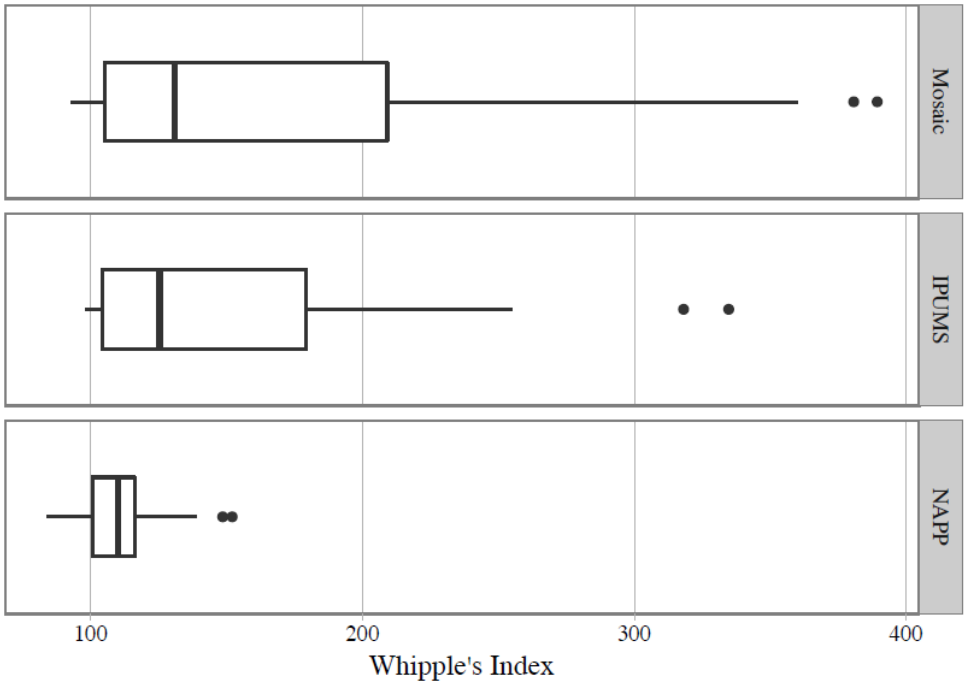
Source: as in Fig. 1

Figure 5: Age heaping patterns in Mosaic data by time-period and quality thresholds



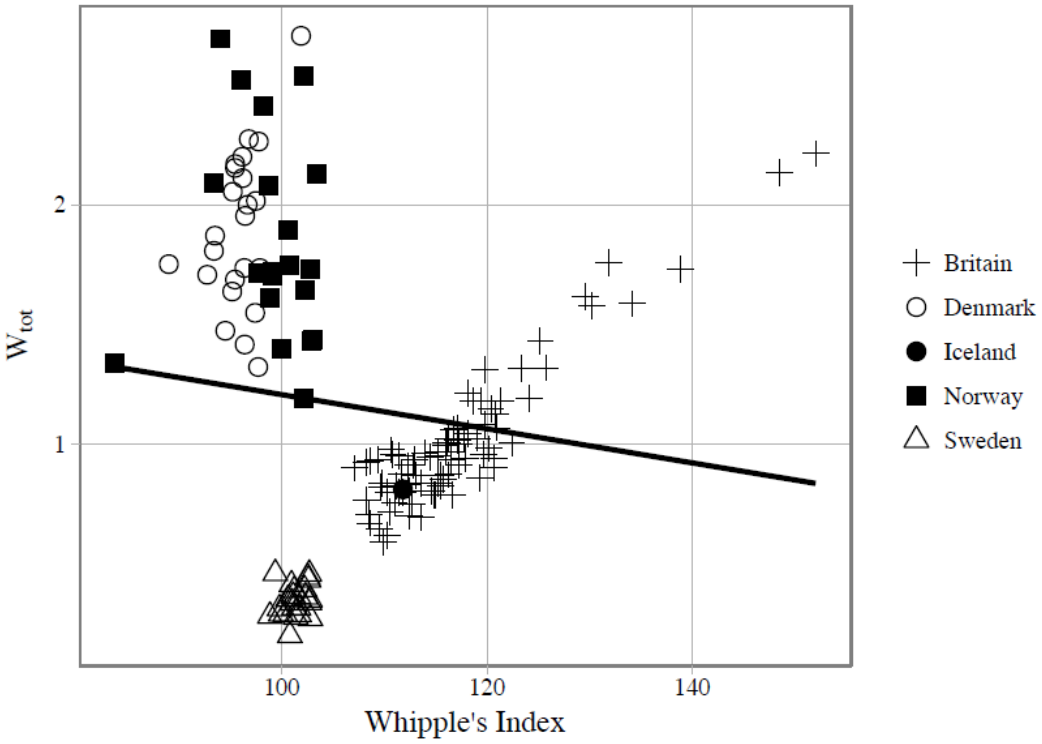
Source: as in Fig. 1

Figure 6: Dispersion of the Whipple's Index values in Mosaic, IPUMS and NAPP data



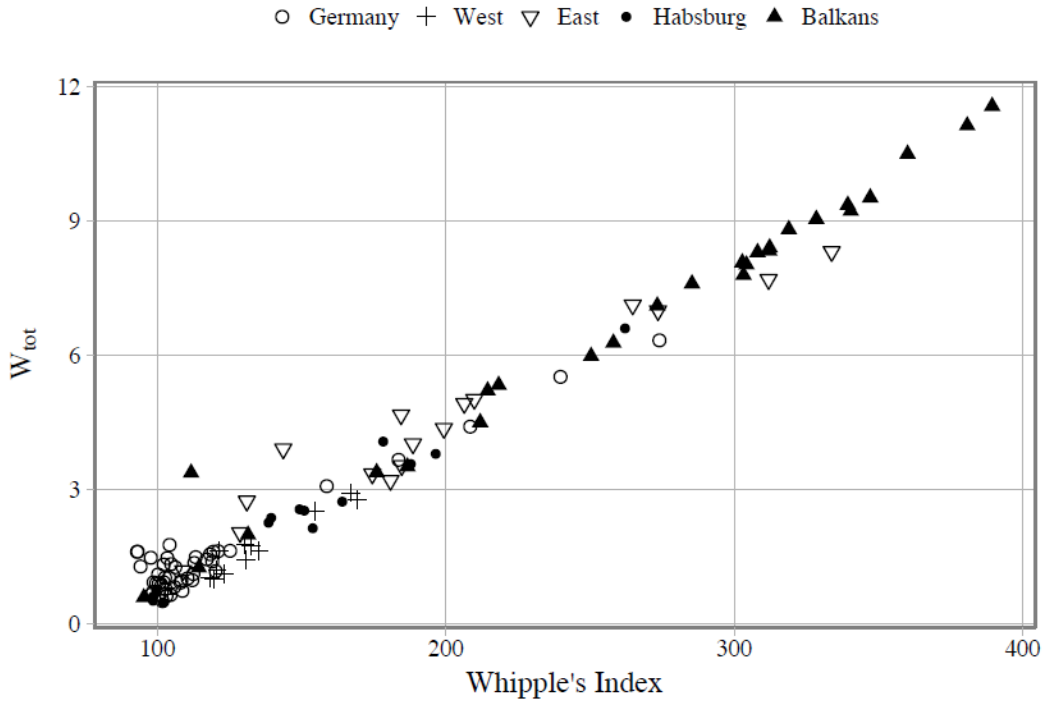
Source: for Mosaic - as in Fig. 1. For IPUMs and NAPP, see Appendix 2.

Figure 7: Patterns of digit preference in the NAPP data, by means of Whipple's and the Total Modified Whipple's indexes



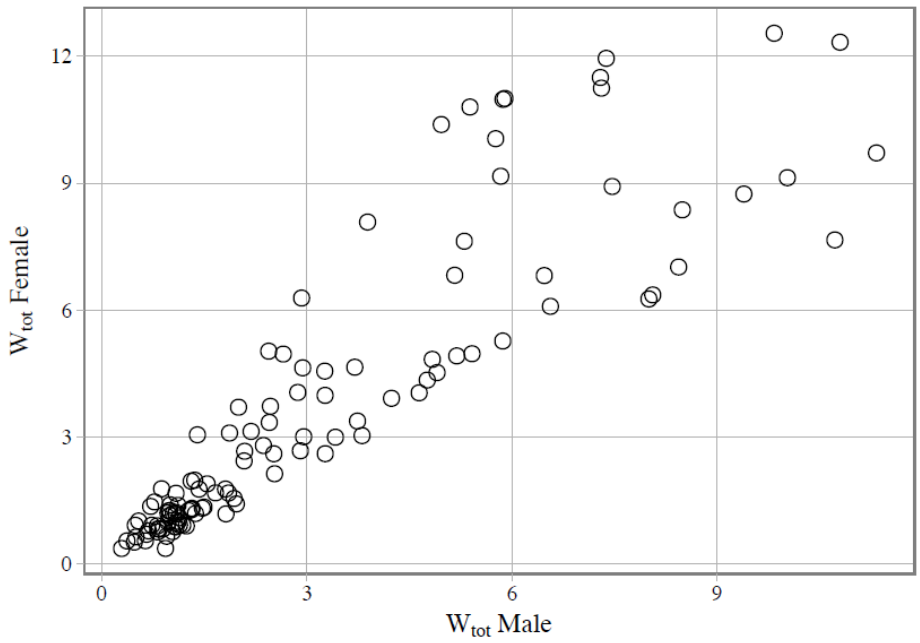
Note: Line represents the linear regression calculated for all populations. *Source*: see Fig. 6.

Figure 8: Age heaping in Mosaic data by type of digit preference measure



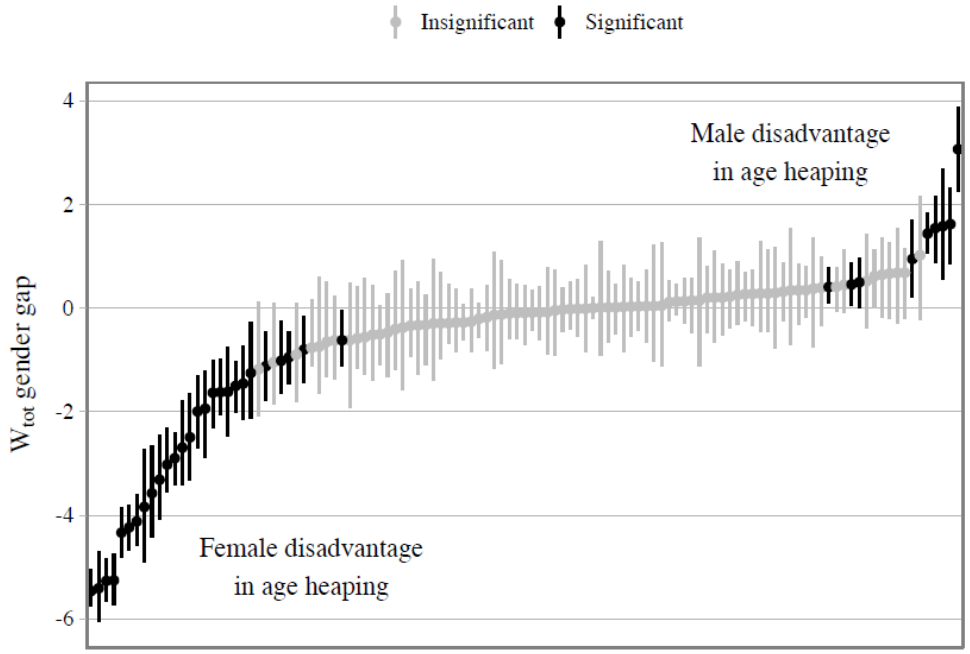
Source: as in Fig. 1

Figure 9: Male *versus* female age-heaping patterns across Mosaic populations



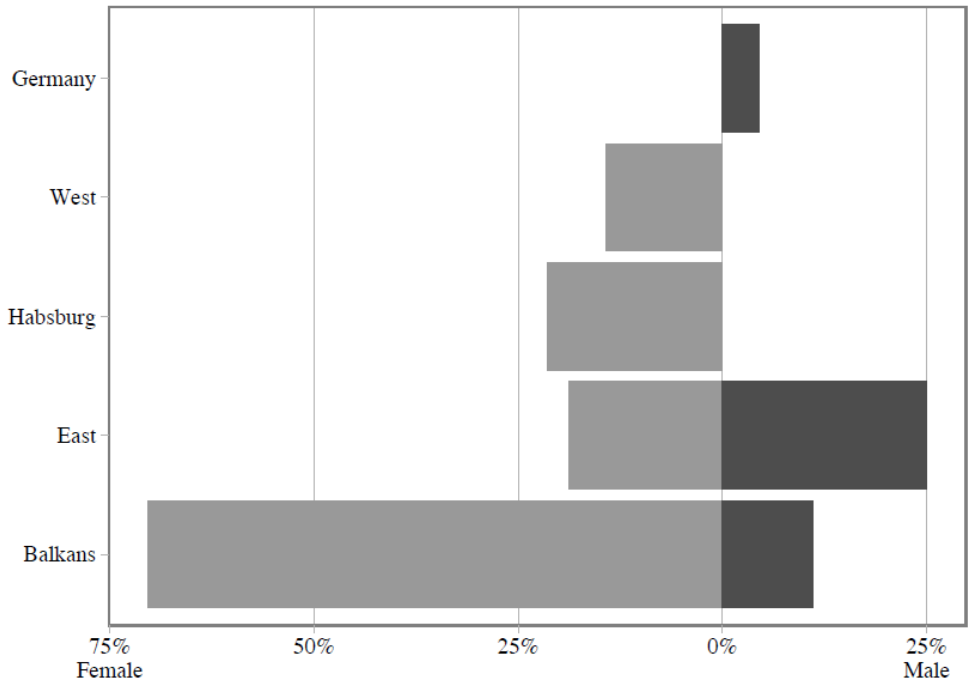
Source: as in Fig. 1

Figure 10: Bootstrapped 95% confidence intervals for gender differences in age-heaping across Mosaic populations



Source: as in Fig. 1

Figure 11: Macro-regional differences in gender inequality in age-heaping across Mosaic populations



Source: as in Fig. 1

Table 2: Descriptive statistics

	N	Mean	Median	sd	min	max
Wtot (ln)	115	0.33	0.16	0.73	-0.27	1.76
Premodern census	27					
Modern census	44					
Semi-modern census	44					
Period before 1800	20					
1800-1850	47					
after 1800	48					
Serfdom	18					
Rural	89					
Ruggedness (ln)	115	2.33	2.12	1.45	-1.83	5.38
Population pot. (ln)	115	16.19	16.20	0.26	15.48	16.63
Patriarchy Index	115	18.45	18	5.54	8	35
CWR	115	0.50	0.49	0.13	0.26	0.92
Proportion 65+	115	4.52	4.77	1.81	0.57	10.81
Germany	44					
West	14					
East	16					
Habsburg	14					
Balkans	27					

Source: as in Fig. 1

Table 3: Regression results

	model 1			model 2			model 3			model 4			model 5			model 6 (robust)	
	B	std. Beta	p	B	std. Beta	p	B	std. Beta	p	B	std. Beta	p	B	std. Beta	p	B	p
(Intercept)	0,81		**<.001**	13,7		**0,001**	-7,65		**0,05**	-3,53		0,352	-7,53		0,078	-3.829725	0.438503
Census type																	
Modern census	-0,55	-0,37	**<.001**	-0,64	-0,43	**0,001**	-0,55	-0,37	**<.001**								
Semi-modern census	-0,78	-0,52	**<.001**	-0,53	-0,36	**0,011**	-0,36	-0,24	**0,025**								
Period																	
1800-1850										-0,58	-0,39	**<.001**	-0,52	-0,35	**<.001**	-0.693211	**0.000624**
after 1800										-0,78	-0,53	**<.001**	-0,72	-0,49	**<.001**	-1.005822	**0.000231**
Serfdom				-0,09	-0,04	0,667	0,06	0,03	0,705	-0,06	-0,03	0,679	-0,1	-0,05	0,617	-0.231542	0.408087
Rural				0,08	0,05	0,572	-0,17	-0,09	0,142	-0,19	-0,11	0,082	-0,1	-0,06	0,377	-0.047408	0.646821
Ruggedness (ln)				0,1	0,21	**0,017**	-0,01	-0,01	0,853	0,01	0,02	0,779	-0,01	-0,01	0,865	0.021097	0.611387
Population pot. (ln)				-0,82	-0,29	**0,002**	0,35	0,12	0,132	0,12	0,04	0,59	0,38	0,14	0,13	0.177768	0.528377
Patriarchy Index							0,09	0,69	**<.001**	0,09	0,65	**<.001**	0,07	0,51	**<.001**	0.057028	**0.000618**
CWR							1,9	0,34	**<.001**	1,78	0,32	**<.001**	1,35	0,24	**0,002**	1.585814	**0.0000**
Proportion 65+							0,04	0,1	0,169	0,03	0,07	0,321	0,04	0,1	0,167	-0.009886	0.866443
West													-0,03	-0,01	0,82	0.004731	0.968105
East													0,4	0,19	0,085	0.255490	0.396217
Habsburg													0,14	0,06	0,392	0.151649	0.266603
Balkans													0,59	0,34	**0,006**	0.694107	**0.000653**
Observations	115			115			115			115			115			115	
R^2 / adj. R^2	.189 / .175			.317 / .279			.636 / .605			.669 / .640			.700 / .661			.752 / .720	
Moran I	0.494		0.000	0.387		0.000	0.023		0.063	0.007		0.131	-0.003		0.077	-0.009	0.064

Appendix 1

Table 1A: Mosaic datafiles (as of April 2017).

census	regions	N (=pop.)
Mosaic data:		
Albania, 1918 census	8 rural regions, 6 cities	140,611
Austria-Hungary, 1869 census	9 rural regions from Hungary, Romania, Slovakia	31,406
Austria-Hungary, 1910 census	3 rural regions and 1 city from Austria	20,036
Belgium 1814 census	1 rural region from Western Flanders	13,666
Bulgaria, 1877-1947 household registers	1 rural region and 1 city from the Rhodope area	8,373
Dubrovnik, 1674 status animarum	1 rural region from Dalmatia	1,880
Denmark, 1803 census	9 rural regions and 2 urban regions from Schleswig and Holstein	107,861
France, 1846 census	3 rural regions	16,967
France, 1831-1901 census	1 rural region from South-Western France	5,109
France, 1846-1856 census	1 city from South-Western France	5,669
German Customs Union, 1846 census	10 rural regions and 4 urban regions	36,760
German Customs Union, 1858 census	1 rural region from the East	3,468
German Customs Union, 1861 census	1 rural region from the Southwest	6,541
German Customs Union, 1867 census	4 rural regions and 1 city in Mecklenburg-Schwerin	66,938
Germany, 1900 census	1 city	55,705
Mecklenburg-Schwerin, 1819 census	3 rural regions and 1 city	37,332
Münster, around 1700 status animarum	3 rural regions in North-Western Germany	23,010

Münster, 1749 status animarum	3 rural regions in North-Western Germany	34,169
Netherlands, census 1810-1811	2 rural regions and 3 cities in the south	40,037
Poland-Lithuania, 1768-1804 listings	12 rural regions	155,818
Moldavia, 1781-1879 status animarum	2 rural regions	5,291
Wallachia, 1838 census	4 rural regions	21,546
Russia, 1795 revision lists	1 rural region in Ukraine	8,050
Russia, 1814 private enumeration	1 region in Central Russia	2,955
Russia, 1847 enumeration	2 rural regions in Lithuania and Belarus	19,917
Russia, 1897 census	1 rural region around Moscow	11,559
Serbia, 1863 census	1 rural region and 1 city	9,746
Serbia, 1884 census	1 rural region	9,434
Spain, 1880-1890 local census	1 rural and 2 urban regions in Catalonia	23,997
Ottoman Empire, 1885 census	Istanbul	3,408
Ottoman Empire, 1907 census	Istanbul	4,946
Mosaic data overall	115 regions (89 rural and 26 urban)	932,205

Appendix 2

Table 2A: References to the data

Mosaic data:

Karl Kaser, Siegfried Gruber, Gentiana Kera, Enriketa Pandelejmoni. *1918 Census of Albania, Version 0.1* [SPSS file]. Graz, 2011.

Laboratory of Historical Demography (MPIDR). *1869 Census of Hungary, Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

Laboratory of Historical Demography (MPIDR). *1910 Census of Austria, Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

Familiekunde Vlaanderen and Laboratory of Historical Demography (MPIDR). *1814 Census of Western Flanders, Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

Ulf Brunnbauer. *Household registers of Rhodope region, Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

Laboratory of Historical Demography (MPIDR). Status Animarum for Lisac and Pridvorje, *Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2015.

Danish Data Archive. *1803 Census of Schleswig and Holstein, Version 1.1* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2012.

Laboratory of Historical Demography (MPIDR). *1846 Census of France, Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

University of Bordeaux. *1831 Census of Sallespisse, Version 1.2* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2013.

University of Bordeaux. *1836 Census of Boulazac, Version 1.1* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2012.

University of Bordeaux. *1841 Census of St. Jean de Luz, Version 1.1* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2012.

University of Bordeaux. *1841 Census of Targon, Version 1.1* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2012.

University of Bordeaux. *1876 Census of Boulazac, Version 1.1* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2012.

University of Bordeaux. *1901 Census of Sauternes, Version 1.1* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2012.

University of Bordeaux. 1846 Census of Saint-Émilion, Version 1.2 [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

University of Bordeaux. 1856 Census of Saint-Émilion, Version 1.2 [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

Laboratory of Historical Demography (MPIDR). *1846 German Customs Union Census, Version 2.1* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

Laboratory of Historical Demography (MPIDR). *1846 Census of Höhscheid, Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

Laboratory of Historical Demography (MPIDR). *1858 German Customs Union Census, Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

Laboratory of Historical Demography (MPIDR). *1861 Census of Haigerloch, Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

State Main Archive Schwerin, Laboratory of Historical Demography (MPIDR), and Department of Multimedia and Data Processing, University of Rostock. *1819 Census of Mecklenburg-Schwerin, Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2016.

State Main Archive Schwerin, Laboratory of Historical Demography (MPIDR), and Department of Multimedia and Data Processing, University of Rostock. 1819 Census of Rostock, Version 1.0 [Mosaic Historical Microdata File]. www.censusmosaic.org, 2015.

State Main Archive Schwerin, Laboratory of Historical Demography (MPIDR), and Department of Multimedia and Data Processing, University of Rostock. 1867 Census of Mecklenburg-Schwerin, Version 1.0 [Mosaic Historical Microdata File]. www.censusmosaic.org, 2016.

State Main Archive Schwerin, Laboratory of Historical Demography (MPIDR), and Department of Multimedia and Data Processing, University of Rostock. 1867 Census of Rostock, Version 1.0 [Mosaic Historical Microdata File]. www.censusmosaic.org, 2015.

State Main Archive Schwerin, Laboratory of Historical Demography (MPIDR), and Department of Multimedia and Data Processing, University of Rostock. *1900 Census of Rostock, Version 1.0* [Mosaic Historical Microdata File]. Rostock, Germany: www.censusmosaic.org, 2013.

Laboratory of Historical Demography (MPIDR). *1749 Status Animarum of Münster, Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

Laboratory of Historical Demography (MPIDR). *1690-1713 Status Animarum of Oldenburger Münsterland, Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

Laboratory of Historical Demography (MPIDR). *Status Animarum for Oggelshausen, Dischingen, Gögglingen, Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

Laboratory of Historical Demography (MPIDR). 1847 Lithuanian Estate Household Listings, Version 1.0 [Mosaic Historical Microdata File]. www.censusmosaic.org, 2015.

Laboratory of Historical Demography (MPIDR). 1811 Census of Zeeland, Version 1.0 [Mosaic Historical Microdata File]. www.censusmosaic.org, 2015.

Laboratory of Historical Demography (MPIDR). 1810 Census of North Brabant, Version 1.0 [Mosaic Historical Microdata File]. www.censusmosaic.org, 2015.

Mikołaj Szołtysek (2012) CEURFAMFORM database, Version 23 [SPSS file]. Rostock.

Laboratory of Historical Demography (MPIDR). 1781-1879 Status Animarum in Moldavia, Version 1.0 [Mosaic Historical Microdata File]. www.censusmosaic.org, 2015.

Laboratory of Historical Demography (MPIDR). *1838 Census of Wallachia, Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

Laboratory of Historical Demography (MPIDR). *1897 Russian Census, Moscow Region, Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org,2014.

Laboratory of Historical Demography (MPIDR). *1795 Braclav Region Revision Lists, Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

Laboratory of Historical Demography (MPIDR). *1814 Russian list of inhabitants, Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

Joel M. Halpern and Siegfried Gruber. *1863 Census of Jasenički srez, Serbia, Version 1.1* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2012.

Joel M. Halpern and Siegfried Gruber. *1884 Census of Jasenički srez, Serbia, Version 1.1* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2012.

Laboratory of Historical Demography (MPIDR). 1880-1890 Local Censuses in Catalonia, Version 1.0 [Mosaic Historical Microdata File]. www.censusmosaic.org, 2015.

Alan Duben. *1885 Census of Istanbul, Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

Alan Duben. *1907 Census of Istanbul, Version 1.0* [Mosaic Historical Microdata File]. www.censusmosaic.org, 2014.

IPUMS Data

Minnesota Population Center. *Integrated Public Use Microdata Series, International: Version 6.5* [Argentina 1970 – National Institute of Statistics and Censuses, Bangladesh 1991 – Bureau of Statistics, Armenia 2001 – National Statistical Service, Bolivia 1976 – National Institute of Statistics, Botswana 1981 – Central Statistics Office, Brazil 1960 – Institute of Geography and Statistics, Burkina Faso 1985 – National Institute of Statistics and Demography, Cambodia 1998 – National Institute of Statistics, Cameroon 1976 – Central Bureau of Census and Population Studies, Chile 1960 – National Institute of Statistics, China 1982 – National Bureau of Statistics, Colombia 1964 – National Administrative Department of Statistics, Costa Rica 1963 – National Institute of Statistics and Censuses, Cuba 2002 – Office of National Statistics, Dominican Republic 1960 – National Statistics Office, Ecuador 1962 – National Institute of Statistics and Censuses, Egypt 1986 – Central Agency for Public Mobilization and Statistics, El Salvador 1992 – General Directorate of Statistics and Censuses, Ethiopia 1984 – Central Statistical Agency, Fiji 1966 – Bureau of Statistics, Ghana 1984 – Ghana Statistical Services, Guinea 1983 – National Statistics Directorate, Haiti 1971 – Institute of Statistics and Informatics, India 1983 – Ministry of Statistics and Programme Implementation, Indonesia 1971 – Statistics Indonesia, Iran 2006 – Statistical Centre of Iran, Iraq 1997 – Central Statistical Office, Jamaica 1982 – Statistical Institute, Jordan 2004 – Department of Statistics, Kenya 1969 – National Bureau of Statistics, Kyrgyz Republic 1999 – National Statistical Committee, Liberia 1974 – Institute of Statistics and Geo-Information Systems, Malawi 1987 – National Statistical Office, Malaysia 1970 – Department of Statistics, Mali 1987 – National Directorate of Statistics and Informatics, Mexico 1960 – National Institute of Statistics, Geography, and Informatics, Mongolia 1989 – National Statistical Office, Morocco 1982 – High Commission of Planning, Mozambique 1997 – National Institute of Statistics, Nicaragua 1971 – National Institute of Statistics and Censuses,

Nigeria 2006 – National Bureau of Statistics, Pakistan 1973 – Statistics Division, Panama 1960 – Census and Statistics Directorate, Paraguay 1962 – General Directorate of Statistics, Surveys, and Censuses, Peru 1993 – National Institute of Statistics and Informatics, Philippines 1990 – National Statistics Office, Puerto Rico 1970 – U.S. Bureau of the Census, Rwanda 1991 – National Institute of Statistics, Saint Lucia 1980 – Government Statistics Department, Senegal 1988 – National Agency of Statistics and Demography, Sierra Leone 2004 – Statistics Sierra Leone, South Africa 1996 – Statistics South Africa, South Sudan 2008 – National Bureau of Statistics, Sudan 2008 – Central Bureau of Statistics, Tanzania 1988 – National Bureau of Statistics, Thailand 1970 – National Statistical Office, Trinidad & Tobago 1970 – Central Statistical Office, Turkey 1985 – Turkish Statistical Institute, Uganda 1991 – Bureau of Statistics, Uruguay 1963 – National Institute of Statistics, Venezuela 1971 – National Institute of Statistics, Vietnam 1989 – General Statistics Office, Zambia 1990 – Central Statistical Office]. Minneapolis: University of Minnesota, 2017. http://doi.org/10.18128/D020.V6.5.

NAPP data:

Minnesota Population Center. *North Atlantic Population Project: Complete Count Microdata. Version 2.3* [Machine-readable database]. Minneapolis: Minnesota Population Center, 2016.

- England and Wales 1881: K. Schürer and M. Woollard, National Sample from the 1881 Census of Great Britain [computer file], Colchester, Essex: History Data Service, UK Data Archive [distributor], 2003
- Scotland 1881: K. Schürer and M. Woollard, National Sample from the 1881 Census of Great Britain [computer file], Colchester, Essex: History Data Service, UK Data Archive [distributor], 2003.
- Denmark 1787: Nanna Floor Clausen, Danish National Archives. 1787 Census of Denmark, Version 1.0
- Iceland 1703: Ólöf Garðarsdóttir (University of Iceland) and National Archives of Iceland (NAI). 1703 Census of Iceland, Version 1.0.
- Norway 1801: The Digital Archive (The National Archive), University of Bergen, and the Minnesota Population Center. Census of Norway 1801, Version 1.0. Bergen, Norway: University of Bergen, 2011.
- Sweden 1880: The Swedish National Archives, Umeå University, and the Minnesota Population Center. National Sample of the 1880 Census of Sweden, Version 1.0. Minneapolis: Minnesota Population Center [distributor], 2014.

Geodata:

Mapfile of historical Denmark (derived from the following sources):

- Danish National Archives (originally from the University of Southern Denmark)
- Dansk Center for Byhistorie [Danish Centre for Urban History]. (2008). Danmarks lokaladministration 1660-2007 [The local administration of Denmark 1660-2007]. Århus: Dansk Center for Byhistorie. Retrieved March 15, 2015 from http://dendigitalebyport.byhistorie.dk/kommuner

Mapfile of historical Europe:

MPIDR [Max Planck Institute for Demographic Research] and CGG [Chair for Geodesy and Geoinformatics, University of Rostock] (2016): MPIDR Population History GIS Collection – Europe (partly based on © EuroGeographics for the administrative boundaries). Rostock: MPIDR. Retrieved August 31, 2016 from http://www.censusmosaic.org/data/historical-gis-files (file: europe19002003.zip)

Historical population and land use data (estimates):
NEAA [Netherlands Environmental Assessment Agency] (2016): HYDE [History Database of the Global Environment], Version 3.2. (beta). Bilthoven: NEAA. Retrieved September 11, 2016 from ftp://ftp.pbl.nl/hyde/hyde3.2/2016_beta_release/zip/ (files: 1800AD_pop.zip, 1800AD_lu.zip)

Elevation data:
USGS [U.S. Geological Survey Center for Earth Resources Observation and Science] (2016): GTOPO30 Global 30 Arc-second Elevation. Sioux Falls SD: USGS. Retrieved August 31, 2016 from http://earthexplorer.usgs.gov/ (files: gt30e020n40, gt30e020n90, gt30w020n40, gt30w020n90, gt30w060n90)