Social Science History (2025), **49**, 254–289
doi:10.1017/ssh.2024.18

ADVANCES IN DATA AND METHODS

Mosaic Database: Consolidation, Innovation, and Challenges in the Comparative Family Demography of Historical Europe‡

Mikołaj Szołtysek[1], Bartosz Ogórek[2], Siegfried Gruber[3], and Radosław Poniat[4]

[1]The Cardinal Wyszyński University, Warsaw, Poland, [2]Institute of History, Polish Academy of Sciences, Warsaw, Poland, [3]University of Graz, Graz, Austria, and [4]University of Białystok, Białystok, Poland
Corresponding author: Mikołaj Szołtysek; Email: mszoltis@gmail.com

(Received 12 September 2023; revised 3 April 2024; accepted 5 April 2024; first published online 11 September 2024)

Abstract
This paper looks at the progress that the Mosaic database has enabled in the study of family structures in continental Europe in the past. Our main argument is that the combination of comprehensive archival research, digitization and computation, data mining, and open-access dissemination that is at the core of the Mosaic project is bringing about an important shift in the fundamental principles that have driven European family history research to date. These transformative features of Mosaic go beyond mere data infrastructural developments, as scaling up to much larger datasets leads to qualitative differences in measurements, methods, and questions. Integrating these perspectives can lead to an important incremental shift in both the scale and the scope of knowledge about historical European family systems.

Keywords: Historical demography; census microdata; household and family systems

Introduction

The family has been the focus of interest for generations of scholars convinced that studying it offers significant insights into populations, societies, and even entire nations (Le Play 1877–1879). One of the most powerful ideas that fired the imaginations of countless researchers was the shared belief that the characteristics of historical family patterns in Europe could be identified, recorded, and analyzed in structural and numerical terms and that these patterns could be understood by paying attention to the phenomenon of household co-residence (Anderson 1980; Hammel and Laslett 1974; Ruggles 2012; Wall 1995).[1] As part of the broader agenda

‡The original version of this article was published with a funding section omitted. A notice detailing with this has been published and the error rectified in the online PDF and HTML copies.

[1]The term refers to a task-oriented residence unit consisting of a group of people who share the same physical space for eating, sleeping, resting, recreating, growing up, raising children, and reproducing.

of unearthing past demographic regimes, since the mid-1960s a number of scholars have undertaken an unprecedented effort to study historical co-residence patterns comparatively, analyzing archival documents that contained enumerations of people by residence units, and employing measurement tools from the social sciences and demography, first within pre-industrial England (Laslett 1965), then within Europe, and eventually beyond (Hajnal 1982; Laslett 1977; Ruggles 2010; VDEFH 1998; Wall 2001). As well as spatially classifying and taxonomizing European societies based on family characteristics, these scholars recognized that the way historical European families were organized in the past could spill over to higher levels of organization as societies evolve (Laslett 1983; Reher 1998), leading to fruitful reflections on the relationship between the past and the present.

Despite their enthusiasm, however, the protracted efforts of these scholars have so far failed to provide a comprehensive reconstruction of historical European family structures. A key reason why is that all these initiatives have had to cope with a lack of reliable, large-scale historical data on family patterns representing the rich diversity of family structures on the continent. Not only was there no "pan-European" data infrastructure, but new data for comparative historical family demography generally proved difficult to obtain and time-consuming and costly to compile, analyze, and interpret within the technological limitations of the time, forcing scholars to rely on informal data sharing and painstaking efforts to compile/compare data collected by others (Wall 2001; also Bohon 2018; VDEFH 1998). For many areas in Europe, data remained scarce, and even where datasets were available, they were rarely in a machine-readable and standardized format, which made them difficult to process when seeking to account for the complexity of family organization or to conduct replication analyses (VDEFH 1998; Viazzo 2003; cf. Kitchin 2014: 32). Although the need for multi-layered analyses of family systems became apparent early on (Laslett 1983; Wall 1995), the successful implementation of such analyses required the use of tools and methods for data management and processing that were out of reach for family historians (cf. Boonstra et al. 2006). Thus, the decomposition of data into easily usable but small parts (i.e., individual communities or a small group of communities), which were not infrequently far apart in space and in time, has long been the main approach applied in the debate on the geography of historical family systems in Europe (see Ruggles 2012). Nevertheless, these diverse and sparse comparative data collections have often served as building blocks for the development of the most far-reaching models of the geography of European family systems.

Older practices of data collection and management were placed on a completely new footing in the 1990s, when the IPUMS and NAPP projects revealed the possibilities for mobilizing new historical demographic data, including for historical north-western Europe, through extensive digitization and transcription initiatives. Researchers of historical family structures who were used to working in "data deserts" now faced an avalanche of information. Thanks to the development of new

Relatives and non-relatives living under the same roof and sharing the same hearth, and servants and lodgers participating in some common activities, are all considered members of the household or domestic group. This definition clearly distinguishes the household from the biological family, whose members are related but do not necessarily live together. Following this clarification, the terms "family," "household," and "domestic group" are used synonymously in this paper and always have the above meaning.

computer technologies and the availability of the internet, rapid data processing and analysis, as well as unlimited data sharing and dissemination, became possible (see e.g., Ruggles 2014; Ruggles et al. 2011; Sobek et al. 2011).

Yet for all the enthusiasm generated by the IPUMS/NAPP revolutionaries, there were ambivalent feelings about the extent to which the emerging "data boosterism" would actually fulfill the longstanding dream of a pan-European reconstruction of family patterns. This was in part because those recent advances were limited to the population of the North Atlantic region, and focused mainly on the second part of the 19th century (Szołtysek and Gruber 2016). At the start of the 21st century, large parts of continental Europe (for an exception, see VDEFH 1998) were still lacking the necessary data infrastructure for conducting systematic comparative historical family research. Thus, researchers in these regions did not even attempt to formulate their arguments based on the analysis of large-scale and harmonized census microdata (e.g., Burguière et al. 1996; Kertzer and Barbagli 2001; Wall et al. 2001). By the late 2000s, it was suggested that an extensive pool of census or census-like material should be developed for as broad a territorial spectrum of continental Europe as possible, as had been previously done for the North Atlantic region. The Mosaic project (Szołtysek and Gruber 2016), building on the experiences of the IPUMS and NAPP initiatives, took up this challenge by extending the collection and distribution of historical census and census-like microdata to the regions of continental Europe.[2]

This paper is concerned with the changes that Mosaic has enabled in the study of historical European family patterns. Our main argument is that the combination of comprehensive archival search, digitization and computation, data mining, and open-access dissemination that is at the core of the Mosaic project is bringing about an important shift in the fundamental principles that have driven research on European family history to date. We also contend that these transformative features of Mosaic can lead to a significant shift in the scale and the scope of knowledge about historical European family systems (cf. Borgman 2015).

Accordingly, we argue that the transformation heralded by Mosaic has changed the ways data are sought, acquired, stored, processed, and made available for analysis. The availability of this unprecedented amount of computationally manipulable data is creating new options for expanding historical knowledge about past family systems (cf. Emigh and Hernández-Pérez 2022). As we will show, the sheer volume of Mosaic data now offers researchers opportunities to gain insights that were not previously possible, encompassing many areas that were either barely explored or entirely unknown before. Moreover, these advances can propel this field of research into new areas.

However, we also reiterate that the proposed vision of Mosaic-induced change goes beyond data infrastructure developments, as scaling up to much larger datasets leads to qualitative differences in the measurements, methods, and questions that

[2]Two other approaches to overcoming the "data desert" are Todd's (1985) attempt to create a "global" reconstruction of historical family structures in Europe by defining the "cultural ideal" of supposed family organization using anthropological evidence, and Dennison and Ogilvie's (2014) meta-analysis of 365 papers on European historical demography. These approaches are not discussed in detail, as we do not believe that either of them is particularly well-suited to overcoming the impasse in which research on historical family structures has found itself (but, see Szołtysek and Poniat 2018).

are used (see Bohon 2018; Borgman 2015). In addition to breaking with the "data desert" paradigm, these new directions in family history research are dependent on applied computer-based innovations and techniques for combining data (cf. Boonstra et al. 2006; Schürer 1986; Schürer and Wall 1986) that allow multiple censuses to be analyzed as a single dataset; comparative analyses to be conducted at different geographical levels; and different characteristics of family systems to be effectively measured with metrics tailored to a particular place, time, and level of aggregation. Finally, we argue that the historical census microdata in Mosaic, rich and informative though they may be, come with their own challenges and limitations, some of which can be mitigated, and some of which cannot. This has resulted in a certain dialectic in the overall assessment of the data discussed here, which can be seen as either "great and rich" or "poor and uninformative," depending on the research question and the epistemological standpoint.

These concerns shape the structure of the paper. After providing an overview of the genesis of the Mosaic project, and noting that the discussion of the "new" is always linked to the "old" (Aronova et al. 2017), we present these themes along the main axes mentioned in the title: i.e., as advances that have revealed new ways of embodying the main concerns of an earlier tradition of family history; and, accordingly, as improvements that have enabled innovations in concepts and approaches that are indeed capable of changing the ways in which research on historical family patterns will be shaped in the years to come. These two perspectives are complemented by a discussion of the main challenges that may arise in using Mosaic.

It should be noted that to understand the nature of the changes brought about by Mosaic, we must at least briefly consider the broader developments in historical-comparative family demography. We will not, however, deal with these developments in their entirety here. We are also aware that while Mosaic plays an important role within these broader trends, it is not the only recent project of its kind. In particular, we must be careful not to regard many of the features of the Mosaic project – especially the infrastructural and computational advances – as stand-alone achievements, as many of them, stem from several parallel knowledge infrastructure projects that are actually older and much larger than Mosaic, such as IPUMS and NAPP.[3] In many ways, Mosaic "stands on the shoulders" of its larger predecessors. There have, after all, been many parallel achievements in the development of longitudinal databases in recent decades (Mandemakers et al. 2023). While a number of these studies have provided real innovations in family history in recent decades (e.g., Tsuya et al. 2010), their contributions to the continental European and the pan-European geography of family patterns have been rather limited (e.g., Dillon and Roberts 2002).

The emergence of Mosaic

Mosaic grew out of two census microdata infrastructure developments that took place almost simultaneously in the late 2000s. The first compilation was the

[3]From spring 2024, the management of Mosaic data and metadata was transferred to IPUMS-International. See mosaic.ipums.org (formerly at censusmosaic.org) to find out about the status of Mosaic and the options for downloading data.

CEURFAMFORM database, which contained information on the inhabitants of more than 20,000 rural households belonging to 236 parishes and 900 settlements in late 18th-century Poland–Lithuania. The data came from various types of population registers that were meticulously excavated from historical archives in Poland, Belarus, Ukraine, and Lithuania, and were then transcribed into a computer file (Szołtysek 2015). The other database was made up of the rich surviving material from the 1918 Albanian census, which covered most of the country, and contained transcribed information on 140,611 persons out of the 524,217 people who were living in some 1800 villages, towns, and cities in the Austro-Hungarian administered territory during the First World War (Kaser et al. 2011).

Simply due to the sheer amount of information they amassed, these two databases were unprecedented endeavors in the history of demographic studies of past populations. However, the innovative features of these databases did not end there. Although they covered great expanses of space and time and originated from different institutional contexts, both datasets followed similar core surveying principles. In particular, they both described the characteristics of all the individuals in a given locality by grouping them into co-resident domestic groups and provided information on each person's age, sex, marital status, and relationship to the household head. In addition, in both datasets, such units consisted not only of the head's core family, but also of his relatives, co-resident servants, and lodgers. Third, all of this information was harmonized across both datasets using the international coding structure of IPUMS (Sobek and Kennedy 2009).

These similarities made it possible to combine the two databases (Szołtysek and Gruber 2014) while ensuring that they could be analyzed as a single dataset in which the same variables could be coded, and standardized queries could be made. Consequently, the Albanian-Polish project established the "prototype" for future Mosaic-type datasets in terms of the database structure and the rules for data inclusion, and in terms of the particular research framework in which they were embedded. Further data developments occurred quite rapidly (see Figure 1) due to the strong and coordinated financial and infrastructural support from the Max Planck Institute, the help of a pan-European network of researchers, and internet access. The Mosaic team and their partners were thus able to identify, sample, and digitize vast amounts of previously unknown census and census-like microdata from many areas of continental Europe.[4]

These advances in data collection were accompanied throughout by a commitment to thoroughly examine the preconditions for data inclusion and to trace how and with which categories each population survey was conducted in a given context to ensure comparability (cf. VDEFH 1998: 115). Finally, to facilitate data transformation and dissemination, the common harmonization scheme was applied to all data collections.

Figures 2 and 3 show the spatial distribution of the most recent Mosaic data by location and region, including forthcoming data releases. While covering the entire

[4]In some cases, data acquisition was achieved through donations or the identification of existing machine-readable data. By census and census-like materials we mean the parts of censuses with full enumeration, as well as local/regional census fragments, church lists of parishioners, tax lists, and local estate inventories (for details, see Szołtysek and Gruber 2016).

https://doi.org/10.1017/ssh.2024.18

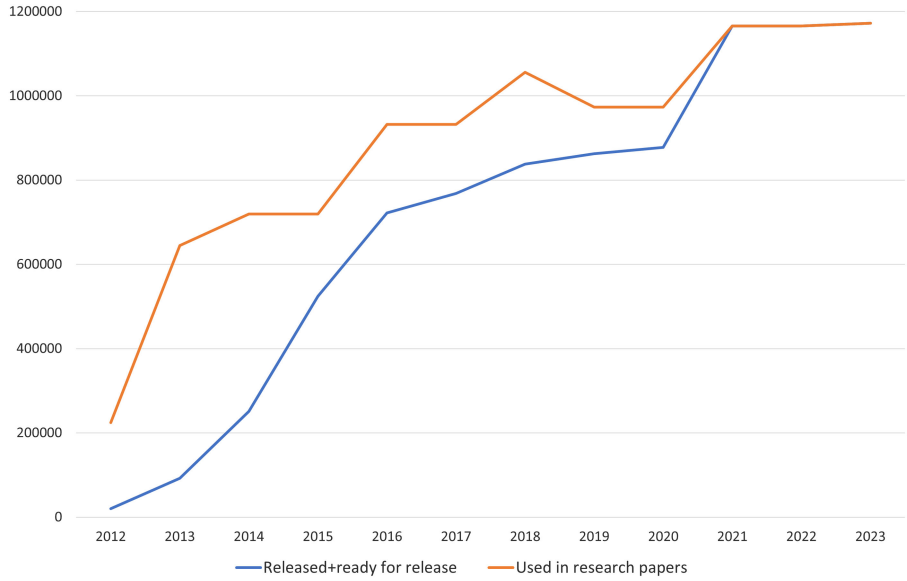

Figure 1. Changes in the volume of the Mosaic data over time (in population totals).
Source: Gruber, Siegfried, Mikołaj Szołtysek, and Bartosz Ogórek (2023) Mosaic datafile, 2023 [machine-readable dataset]. IPUMS-International (mosaic.ipums.org).
Note: The discrepancy between the two curves is related to the fact that the release of new data in the early years of Mosaic was somewhat delayed by the requirements of several ongoing research projects. Currently, all datasets ever researched are also publicly available as part of Mosaic.

territory of continental Europe with historical census microdata remains a dream we may never achieve, Mosaic's current data scope represents an unprecedented expansion in the volume and the spatial breadth of data for the study of historical family patterns. In total, Mosaic contains information on 4364 settlements (villages, hamlets, parishes, estates) with 1,172,241 people living in over 200,000 family households across societies stretching from Navarre and Vizcaya in the west to western Siberia in the east, and from the "far north" of Europe via Saint Petersburg to Almeria and Kythera in the south.[5] These Mosaic sites are also grouped into 161 regions, which correspond either to the respective administrative units (usually also counties), or, in the absence of administrative units, to geographical clusters to facilitate meso-level analysis.[6] As a rule of thumb, efforts were made to ensure that each Mosaic region has at least 2000 inhabitants and that urban and rural

[5]Each individual harmonized Mosaic data file contains 30 variables related to three different levels of information defining, respectively, the dataset, the household, and the person. In general, most variables are designed according to the standards of IPUMS International (also NAPP). Region files are created from the individual-level data files. Each of these data files can be analyzed separately or in combination. See the forthcoming updated Mosaic website at IPUMS-International for download options.

[6]Meso-level entities have the advantage of representing smaller scale, lower level social units with specific sets of guidelines for societal organization and institutional profiles shaping interactions between kin. Thus, these entities allow us to capture aggregate contextual factors that influence demographic outcomes, and are likely to be shared within a community.

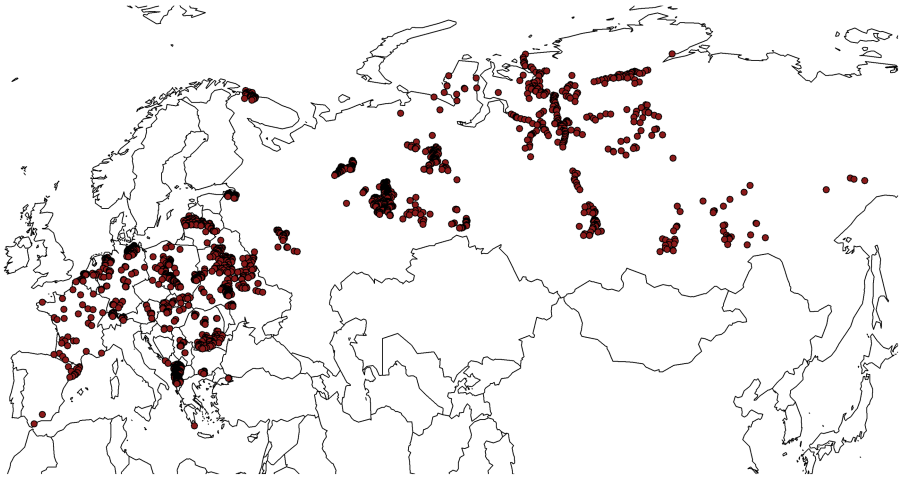

Figure 2. Spatial distribution of Mosaic data by settlement points.
Source: Gruber, Siegfried, Mikołaj Szołtysek, and Bartosz Ogórek (2023) Mosaic datafile, 2023 [machine-readable dataset]. IPUMS-International (mosaic.ipums.org).

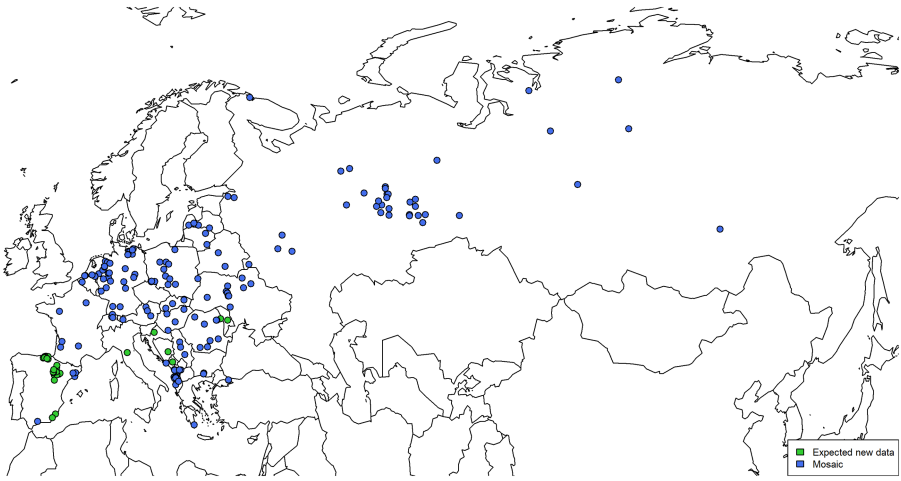

Figure 3. Spatial distribution of Mosaic data by regions.
Source: Gruber, Siegfried, Mikołaj Szołtysek, and Bartosz Ogórek (2023) Mosaic datafile, 2023 [machine-readable dataset]. IPUMS-International (mosaic.ipums.org).
Note: each point on the map (B) represents the centroid of one Mosaic regional population as defined in the text. "Expected new data" include: (1) census samples Bessarabia 1850, Zagreb 1857, Serbia 1863, Montenegro 1879, Armenians in Istanbul 1907, Sarajevo 1910, about (200,000 persons in total) (see S. Gruber "Demography and society in historical Southeastern Europe" (FWF No. P 34285); (2) a selection of 12 localities from the Spanish census of 1887 (Censo de la Población de España) (ca. 60k persons) (University of Zaragoza); (3) a selection of 33 localities from the 1860 census in the province of Zaragoza in Spain (ca. 25k people) (University of Zaragoza); (4) the Florentine Catasto of 1427 based on D. Herlihy and C. Klapisch-Zuber's original datafile (ca. 270,000 persons).

https://doi.org/10.1017/ssh.2024.18

settlements are separated.[7] An important extension of the current version of the dataset is the inclusion of historical census microdata from western Siberia that cover a large proportion of the indigenous peoples of Russia's circumpolar north. This marks the first attempt to study the populations of north-west Asia using integrated census microdata structures.

The Mosaic samples come from different types of historical census and census-like materials (see Szołtysek and Gruber 2016; also ft. 4). Despite the rigorous data pre-selection procedures, this diversity can affect both the nature and the quality of particular listings. To capture this institutional variability, our metadata were used to categorize all regional censuses into three groups according to their varying degrees of control over census administration (i.e., more direct and more intensive involvement of trained personnel in the census process) (see more in Szołtysek et al. 2018).

All these data are geo-referenced (both as location points and as regional centroids), which makes it possible to link them to various covariates derived from geographic information system (GIS) and other location attributes (see below). While the total area covered by the Mosaic data is extremely large, spanning 6345 km from west to east and 3687 km from north to south, the relevant data points are mostly noncontiguous (see discussion below). The database crosses many important fault lines in the European geography of demographic regimes. However, it also captures much of the variation across the continent in environmental characteristics, cultures (including kinship regimes), and socioeconomic geography, and in patterns of economic growth in the early modern and modern periods.

In total, the database covers 22 European countries, and most of these data – with the exception of the Croatian, Bulgarian, Belgian, Turkish, and Spanish data – come from census collections covering very large populations from multiple localities and wide geographical areas, and therefore provide a reasonably adequate representation of historical diversity in these areas, even if they are not nationally representative in a statistical sense. Most of the Mosaic samples also remain the best samples that are currently available for the regions or countries they cover, and it is likely that for some areas (e.g., Poland–Lithuania), better samples will never be obtained (Szołtysek and Gruber 2016: 44; also Szołtysek 2015).

Consolidations

Mapping variation

One of the most tangible implications of the Mosaic project in relation to the main concerns and interests of the older family history tradition is its potential to map family characteristics in geographical space. Thanks to the geo-referenced nature of all the data, it is possible to display a large number of elements related to family organization at the meso (regional) or local level in cartographic (digital) form, and thus to make instant comparisons. For example, for the first time since the appearance of the seminal works of the 1960s and 1970s, we can map quite

[7]In the absolute majority of Mosaic records, a "place-based" approach was used: i.e., each settlement, village, or parish had been classified by the census takers as "rural" or as a "town" (more rarely as part of a city) based on the legal status of the particular place at the time.

accurately many European regions in terms of the three variables that Hajnal (1982), Laslett (1977), and many others have considered crucial to the study of historical family organization: marriage patterns, household structure, and the incidence of service (Figure 4) (see below on more sophisticated variables).

In addition to illustrating the patterns that once existed in Europe, this approach can serve important analytical purposes. It can, for example, show the role that geographical proximity played in patterns of family organization, and can thus improve our understanding of how aspects of family organization in one area differed from those in other areas. Rather than relying on simplistic notions of dividing lines, "transition zones," and/or "ideal family systems" (Hajnal 1982; Mitterauer 2003; Reher 1998; Therborn 2004; Todd 1985), the analysis of Mosaic data can result in a more sensitive description of the geography of family patterns, and may lead to the discovery of more complex patterns, including those reflecting the ways in which family and demographic boundaries were crossed and spread, both spatially and temporally. These new geographies may still be incomplete, changeable, or contestable. However, compared to the ways these issues were managed in the "pre-Mosaic world," this approach represents a major breakthrough. Take, for example, Laslett's famous regional "sets of familial tendencies" (Laslett 1983), which can now be discussed not only on the basis of a few local case studies (e.g., Wall 1995, 2001), but also on the basis of a large pool of regionally differentiated data on households, families, and individuals.

By mobilizing spatially organized, large-scale information at different levels of aggregation, the Mosaic database can not only better address the question of what the most important variations in European family organization were, it can also move the problem of variability in family characteristics to the center of inquiry (cf. Smith 1984).

Figure 5 illustrates this point by showing the distribution of the values of the shares of nuclear and multifamily households for two sub-datasets of the Mosaic collection from the historical German territories and the Polish-Lithuanian Commonwealth. Despite its simplicity, this type of "compositional" data representation provides several important insights. For example, it shows that the extent of variation observed in Poland-Lithuania is not comparable to that found in the German data, and that none of the standard population units are homogeneous. It also shows that the identification and the sorting of sub-populations are indeed necessary to understand the family history of any area, because these are the only ways to capture real differences in local or regional conditions that make certain family patterns "thinkable" in particular contexts (cf. Plakans and Wetherell 2005). Accordingly, Mosaic allows for populations to be compared not only in terms of the mean values of certain indicators but also in terms of how much variation in certain family characteristics they can include.

In addition, the approach illustrated in Figure 5 alludes to the possibility of investigating the extent to which the size of localities can lead to random variations in the distribution of certain indicators. For example, a simple permutation test conducted for the two "country" populations in Figure 5 shows that if two German villages were randomly selected and the average of the simple family households was

https://doi.org/10.1017/ssh.2024.18

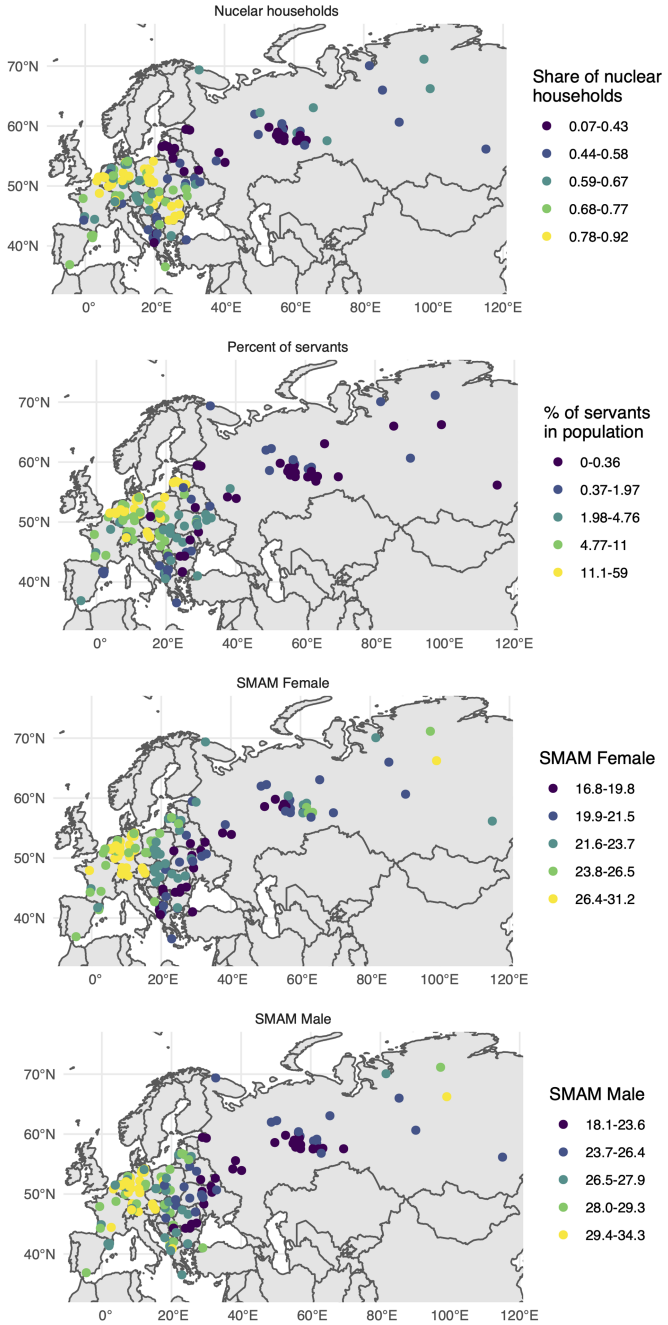

Figure 4. Spatial distribution of the selected demographic parameters across Mosaic data.
Source: Gruber, Siegfried, Mikołaj Szołtysek, and Bartosz Ogórek (2023) Mosaic datafile, 2023 [machine-readable dataset]. IPUMS-International (mosaic.ipums.org).

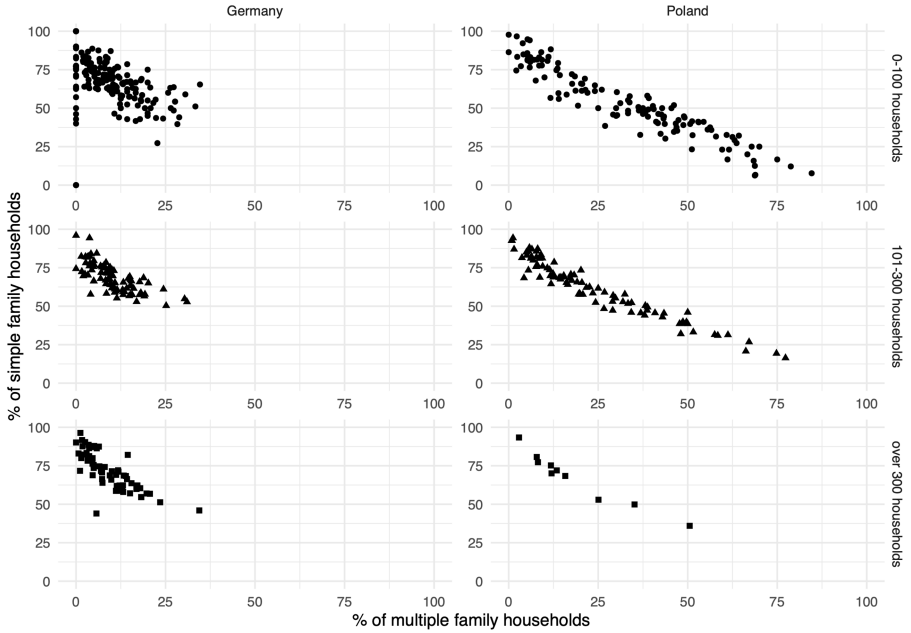

Figure 5. The share of nuclear and multifamily households for two sub-datasets of the Mosaic collection (by number of households per region).
Source: Gruber, Siegfried, Mikołaj Szołtysek, and Bartosz Ogórek (2023) Mosaic datafile, 2023 [machine-readable dataset]. IPUMS-International (mosaic.ipums.org).
Note: On the left side of the scatter diagram - the 363 localities from the German territories from the period 1690-1867; on the right side - the 234 parishes from the Polish-Lithuanian Commonwealth (the late 18th and early 19th centuries).

calculated from 1000 draws, 95 percent of the results would range from 44.2 to 85.1, and from 27.6 to 84.2 in Poland.[8] Thus, we observe a lot of differentiation each time, and see no significant differences between countries that we intuitively know are very different. In this respect, the agglomeration of Mosaic data can be more robust and rewarding, in part because the use of larger populations (of regions or macro-regions) can help to compensate for random errors due to stochastic fluctuations, allowing for more accurate and parsimonious estimates of many parameters than those obtained in earlier comparisons (cf. Burguière and Lebrun 1996: 36).

Because it offers large-scale data integrated across different levels of aggregation, Mosaic can easily be used to place local patterns in a larger meso- or macro-level context of which they are a part, and can thus better distinguish the particular from the general than scattered case studies could (see, e.g., Flandrin 1979; Todorova 1996; cf. Kurosu 2016). How the particular can be systematically distinguished from the general and assessed on the basis of the scalable and multi-layered geographical structure of the dataset is shown in Figure 6 using the example of the proportion of female servants in a small community in Poland in 1791 and the corresponding

[8]A similar exercise for complex households yields a range of 1.66–22.6 for Germany and 6.30–60.8 for Poland.

https://doi.org/10.1017/ssh.2024.18

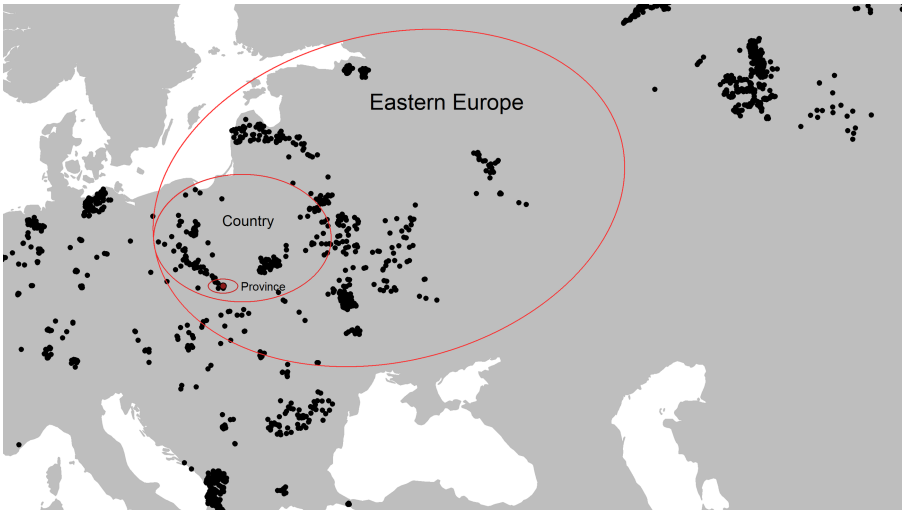

Figure 6. The example of multilevel embodiment of a single locality from the Mosaic collection (the parish of Kaziemirza Wielka in Poland, 1791).
Source: Gruber, Siegfried, Mikołaj Szołtysek, and Bartosz Ogórek (2023) Mosaic datafile, 2023 [machine-readable dataset]. IPUMS-International (mosaic.ipums.org).
Note: the eclipses from the smaller to the larger indicate, accordingly: the province, the country, the macro-region.

scaling of the Mosaic data. This simple exercise shows that Kazimierza Wielka was only slightly different on the measure in question in the province to which it belonged (38.4 percent to 32.8 percent), but that it was definitely exceptional at the level of the country (12.4 percent) and the entire Eastern European region included in the database (14.8 percent). Such programmatic comparisons can be made for most Mosaic sites with a large collection of regional data and for a long list of variables.

Measurements

Due to the prevailing paradigm of research and data organization in the past, and given the technological limitations at that time, many important dimensions of family organization could not be effectively quantified and compared, let alone visualized.

Take, for example, a comparative analysis of the relationship between the age-specific proportion of men who had ever been married and the proportion of men who were heads of households, which has been advocated as a measure of the extent to which marriage signified the creation of an independent residential and economic unit. Such analysis has rarely been undertaken (and then with limited information content), because it was extremely difficult in the past to generate the necessary comparative data on age-specific marriage and household headship rates *en masse* (Hajnal 1982; cf. Smith 1993: 396–399), and it was even more difficult to process these data. Today, by contrast, historical microdata infrastructures such as Mosaic allow us to calculate these parameters simultaneously for multiple datasets and populations.

https://doi.org/10.1017/ssh.2024.18

Figure 7 illustrates how such an investigation could be carried out for all Mosaic records. Because of the agglomeration of local censuses and technological capacities for data processing, what had been seen as a scarce commodity in earlier studies can now be easily transformed with Mosaic into a veritable "flood" of fine-grained information that can be sorted, sifted, and scaled for specific analyses. This information can be further used to investigate variations, spatial groupings, and central tendencies, generating potential discoveries on topics that – although central to family history research – could not be fully captured before (cf. Smith 1993; also Szołtysek and Ogórek 2020).

By relying on synthetic cohort methods (as in Figure 7), we can compensate to some extent for missing longitudinal cohort data and obtain reasonable surrogate measures of the timing, magnitude, and pace of certain life course changes, especially for populations clustered around the same census period (see Watkins 1980). Figure 8 shows how this might be done for a section of the Mosaic data, and illustrates the differences in the timing of key life course transitions for three regions in 18th-century Poland-Lithuania. New studies of the life course (e.g., the impact of service, early marriage, living with grandparents) can use such (or similar) Mosaic results to assess the relative importance of particular historical demographic contexts.

The above examples have shown how a combination of the sheer volume of data can enable advances in measurement that were previously only possible with "low-hanging fruit." While having more data does not always result in better research (e.g., Borgman 2015), another example of how Mosaic's drive to assemble much larger datasets can increase the chances of gaining important research insights is the application of machine learning.[9] Because of its scale, content, and coverage, Mosaic is particularly well-suited to harnessing the power of unsupervised machine learning or cluster analysis techniques to infer optimal natural groupings in multidimensional data, which allow complex patterns to be identified with high levels of efficiency and low costs (e.g., Han et al. 2011; Hastie et al. 2009). This quality could prove crucial, as many classical models of family patterns are in fact sets of interrelated variables or elements (Hajnal 1982; Laslett 1983), but have seldom been formally "tested" (e.g., Barbagli 1991). The application of machine learning tools could provide new insights by answering previously unresolved questions, such as whether historical European populations form natural groupings based on how similar or dissimilar they are with respect to certain family demographic markers, and if so, how many such groupings can plausibly be identified. Such approaches can be particularly helpful in replacing the *ad hoc* deductive typologies prevalent in previous studies with formal methods of automatic pattern recognition.[10]

For example, using the Partitioning Around Medoids algorithm and careful optimization criteria, Szołtysek and Ogórek (2020) have shown (see Figure 9) that

[9]For earlier examples of using the various machine learning techniques in family history, see Bohon (2018); also Schürer and Penkova (2015) and Pujades-Mora et al. (2022).

[10]It is, of course, undeniable that domain experts will continue to play an important role, especially when it comes to the danger of automated approaches recognizing patterns in random or meaningless data (so-called apophenia) (Boyd and Crawford 2012).

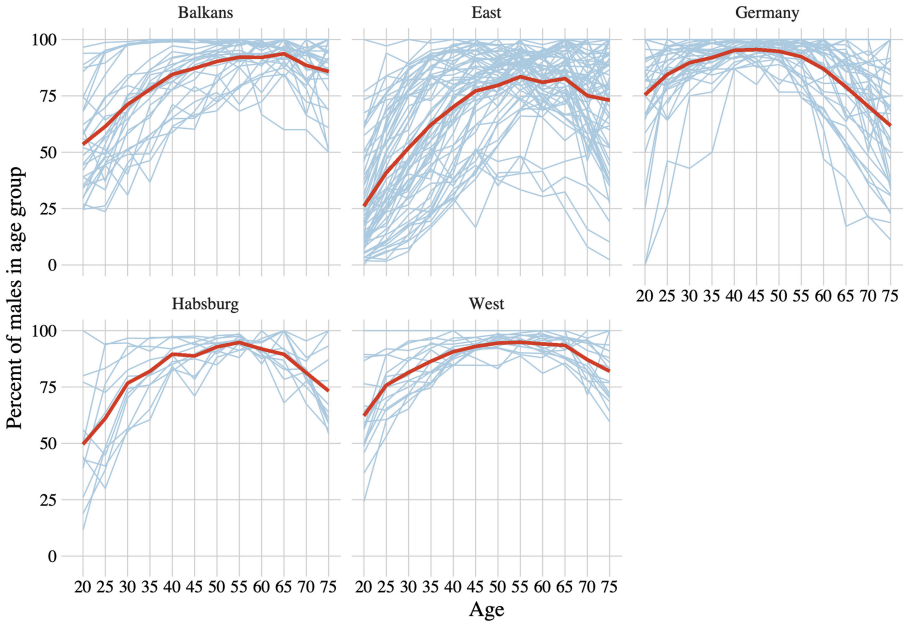

Figure 7. The share of householders among ever-married men, all Mosaic regions.
Source: Gruber, Siegfried, Mikołaj Szołtysek, and Bartosz Ogórek (2023) Mosaic datafile, 2023 [machine-readable dataset]. IPUMS-International (mosaic.ipums.org).
Note: the proportion of households among ever-married men in certain age groups was used as a proxy for the relationship between entering marriage and becoming head of household.
Five bigger territorial groupings followed major institutional and socioeconomic distinction across historic Europe. "Germany": German-dominated areas other than the Habsburg territories; "West": areas west and south-west of Germany; "Habsburg": Austrian, Hungarian, Croatian, as well as Slovakian data; "East": east-central and eastern Europe, including the former Polish-Lithuanian Commonwealth and Russia (including Siberian territories geographically in Asia); "Balkans": areas south and/or east of Croatia and Hungary.

partitioning household formation systems in historical populations into four clusters is a far more reasonable way to capture variation in the Mosaic data (merged here with NAPP; see below) than the dual partition model proposed by Hajnal (1982). The proposed clustering solution yielded several other intriguing results. A similar approach can be applied to many other historical demographic problems.

It is noteworthy that most of the above measurements can be broken down by urban-rural differences. However, the validity of such comparisons is compromised by the overwhelming dominance of rural regions (80 percent) and the uneven spatial distribution of the urban population in the Mosaic database.

Finally, it should be mentioned that Mosaic can ultimately facilitate partial analyses of the impact of the socioeconomic status of household heads on various domestic group characteristics. Three-quarters of the Mosaic regions, which account for 73.8 percent of the population in the Mosaic database, include occupational information. Only 48 regions (26.2 percent) with 28.2 percent of the database population do not contain information on occupational titles. Currently, however, only 69 regions (37.7 percent) with 47.0 percent of the database population

https://doi.org/10.1017/ssh.2024.18

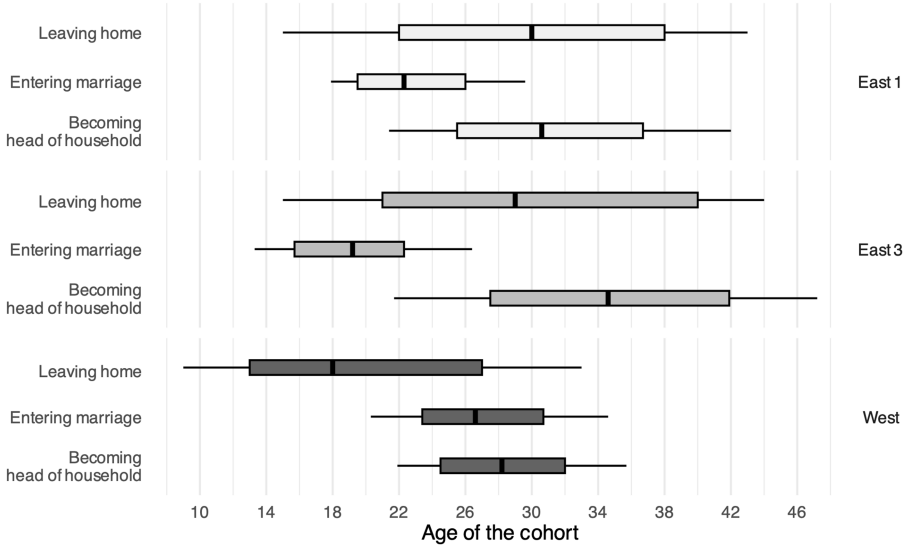

Figure 8. Sequences of main life course transitions in Poland-Lithuania based on synthetic cohorts.
Source: Gruber, Siegfried, Mikołaj Szołtysek, and Bartosz Ogórek (2023) Mosaic datafile, 2023 [machine-readable dataset]. IPUMS-International (mosaic.ipums.org).
Note: Only male population. Entry into marriage is measured with the singulate mean age at marriage (Hajnal 1982); for leaving home - singulate mean age at leaving home (see Schürer 2004; Szołtysek 2015: 282–83); for household formation – singulate mean age at household headship (see Szołtysek 2015: 512–13). Home leaving data based on estimates of parental co-residence taken from the listings adjusted for the availability of parents assessed through CAMSIM microsimulation (Szołtysek 2015: 284–85). "West," "East1," and "East 3" stand, respectively, for: western and central Kingdom of Poland (including Silesia); central Belarussian part of the Grand Duchy of Lithuania; and, the southern Belarussian part of the Grand Duchy of Lithuania.

(536,214 persons) have their occupational titles coded, and further work to improve this situation is in progress.[11]

Innovations

While Mosaic can help to consolidate the field of comparative historical family demography by providing better answers to many critical questions that have long been asked, it also provides fertile ground for innovations in the ways historical family demographic research is conducted in general. The following section describes some of these new elements, focusing on the issues of measurement, analysis, and data merging.

Measurements

As early as the 1980s, it was recognized that classifications of co-residence at the household level are limited and that such measures must be combined with measures

[11]The absolute majority of Mosaic data are not suitable for studying differences in the family forms of natives and non-natives in a consistent manner, as information on place of birth is only available in a very small share of the data.

https://doi.org/10.1017/ssh.2024.18

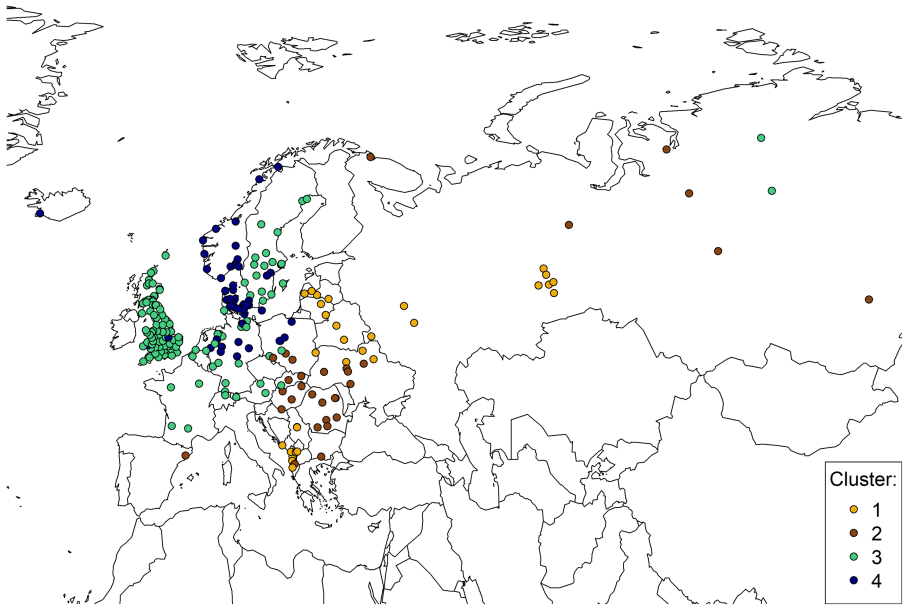

Figure 9. Four-cluster structure of Hajnal's household formation markers on the geographic coordinates, Mosaic, and NAPP datasets combined.

Sources: Gruber, Siegfried, Mikołaj Szołtysek, and Bartosz Ogórek (2023) Mosaic datafile, 2023 [machine-readable dataset]. IPUMS-International (mosaic.ipums.org); Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 7.2 [dataset]. Minneapolis, MN: IPUMS, 2019. https://doi.org/10.18128/D020.V7.2.

Note: SMAM – Singulate Mean Age at Marriage; Service – the share of unmarried women in the age group that is determined by the value of the female SMAM in each of populations under study; CMHD – Cumulative Marriage Headship Difference (for details, Szołtysek and Ogórek 2020: 56–57).

Characteristics of cluster medoids: *k*1 - 20.8 (female SMAM), 3.9 (proportion female servants), 7.4 (cmhd), 0.41 (share nuclear households); *k*2 (respectively)- 20.7, 10.6, 2.4, 0.7; *k*3 – 26.3, 32, 1.1, 0.7; *k*4 – 27.2, 59.6, 0.7, 0.8.

of family composition at the individual level to capture the complexity of living arrangements (Ruggles 1987; Schürer and Wall 1986). Because of the structure of its core variables, Mosaic can enable such analyses by applying a common coding scheme for housing units based on the commonly used classification schemes, while also representing the individual relationships between the people included in the database through distinct but linked and compatible classification pointers (to be further broken down by sex, age, and marital status) (see Ruggles 1995).

Table 1 shows an example of a combination of coding variables of different orders applied to the census list of members of a domestic group from an exemplary parish in the Mosaic collection. First, the relationships of these individuals to the main reference person on the list, a household head, are determined. Then, each person is assigned a common code for the residential structure in which they live. This is supplemented by the codes that capture the conjugal-family relationships of all individual household members (Wall 1998), and finally by a set of dyadic variables that identify the marital, parental, sibling, and other kinship relationships between all persons living under the same roof (only a subset of the actual dyads available is given in the table). When analyzed in combination (either cross-sectionally or by age group), the different dyads

https://doi.org/10.1017/ssh.2024.18

Table 1. Application of various locator variables to the encoding of the members of a domestic group in the Mosaic data

Original entry	Age	Sex	Marital status	Relationship to head	Household type membership	Relationship to head (enhanced)	CFU	Lives in CFU	Lives with a spouse	Lives with at least one parent	Lives with at least one child	Lives with at least one married child	Lives with at least one sibling	Lives with other kin only
Tomasz Piątek	30	M	M	HEAD	multiple-family (type 5)	head	spouse	yes	yes	yes	yes	no	yes	no
Anna wife	44	F	M	SPOUSE	multiple-family (type 5)	wife	spouse	yes	yes	no	yes	no	no	no
Marianna daughter	18	F	S	CHILD	multiple-family (type 5)	daughter	child	yes	no	yes	no	no	yes	no
Chieronim son	15	M	S	CHILD	multiple-family (type 5)	son	child	yes	no	yes	no	no	yes	no
Marta daughter	8	F	S	CHILD	multiple-family (type 5)	daughter	child	yes	no	yes	no	no	yes	no
Salomea daughter	1	F	S	CHILD	multiple-family (type 5)	daughter	child	yes	no	yes	no	no	yes	no
Wojciech Piątek brother	22	M	S	OTHER KIN	multiple-family (type 5)	brother	child	yes	no	yes	no	no	yes	no
Piotr Piątek father	60	M	M	PARENT	multiple-family (type 5)	father	spouse	yes	yes	no	yes	yes	no	no
Barbara his wife	59	F	M	PARENT	multiple-family (type 5)	mother	spouse	yes	yes	no	yes	yes	no	no
Tomasz Krzyczan servant	21	M	S	SERVANT	multiple-family (type 5)	servant	non-kin	no	no	no	no	no	no	no
Anastazja Siwkowo	35	F	W	LODGER	multiple-family (type 5)	lodger	lone parent	yes	no	no	yes	no	no	no

(*Continued*)

https://doi.org/10.1017/ssh.2024.18

Table 1. (*Continued*)

Original entry	Age	Sex	Marital status	Relationship to head	Household type membership	Relationship to head (enhanced)	CFU	Lives in CFU	Lives with a spouse	Lives with at least one parent	Lives with at least one child	Lives with at least one married child	Lives with at least one sibling	Lives with other kin only
lodger, a widow														
Helena daughter	7	F	S	LODGER	multiple-family (type 5)	lodger's daughter	child	yes	no	yes	no	no	yes	no
Piotr son	3	M	S	LODGER	multiple-family (type 5)	lodger's son	child	yes	no	yes	no	no	yes	no
Andrzej Siwek, her relative	33	M	W	LODGER	multiple-family (type 5)	lodger's other relative	other kin	no	no	no	no	no	no	yes

Source: Szołtysek, Mikołaj (2015).
Note: the data are for the census from Słupia parish in Greater Poland province of Poland in 1791.

https://doi.org/10.1017/ssh.2024.18

can provide information on the simultaneous presence (or absence) of several kinship ties at certain stages of the person's life in the domestic sphere. This can foster various in-depth research approaches focusing on the residential circumstances of older people, on age-specific changes in "micro-networks" ("roles") in domestic groups, or on empirical considerations of the advantages and disadvantages of using individual-level versus household-level measures in specific contexts (Szołtysek 2015: 684–89; Szołtysek et al. 2020; cf. Ruggles 2012).

Second, by combining household- and individual-level variables that are harmonized across multiple datasets, Mosaic allows researchers to develop measures tailored to specific research problems without having to rely on predefined schemes (cf. Ruggles 2012: 341). The main example in this regard concerns the use of Mosaic data to construct the Patriarchy Index (hereafter PI) to quantify the social and ideological construct of familial patriarchy (see Gruber and Szołtysek 2016). For this index to be useful, it was first necessary to identify clearly defined items for cross-cultural comparisons in the multifaceted manifestations of the patriarchal order. Accordingly, the operationalisability of these items had to be tested using the information available in the Mosaic data, which inevitably led to the omission of aspects that, although theoretically important, are hardly reflected in the historical sources (e.g., domestic violence). Furthermore, given the open-ended, cross-cultural, and cross-temporal structure of Mosaic, it was necessary to walk a tightrope between specificity and generality in compiling the index to ensure that all its potential components had equal chances of occurring in populations from different regions and time periods. The aim of this approach was to ensure the greatest possible effectiveness with a minimum of information content.[12]

The result was a composite measure consisting of four sub-indices to capture inter-generational and inter-gender relations: dominance of men over women, dominance of the older generation over the younger generation, patrilocality, and preference for sons. All 11 (earlier 12) variables that made up these sub-indices could easily be calculated from routine individual-level censuses or census-like microdata that had been widely used in Europe since the early modern period.[13]

The PI can serve several purposes in the study of family history: (1) It can be used to measure the intensity of patriarchy in family systems across cultures (see Figure 10), and to assess whether the clustering of PI elements on particular dimensions differs across populations; (2) it can be used as a composite measure of family systems, and as a measure of strong/weak family ties in historical populations (cf. Reher 1998); and, finally (3), it can serve as a predictor variable in modeling different demographic behaviors, also in comparison to other similar measures (see Szołtysek and Poniat 2018; Szołtysek, Beltran Tapia, et al. 2022).

[12]As the PI was tested incrementally as the Mosaic dataset grew, it was also important to keep all the original variables in the index, rather than relying on the dimensionality reduction techniques that are commonly used for index construction.

[13]The PI relies on demographic indicators of marital practices (including age hypergamy); family structure and roles by age, sex of offspring; and power relations within the household.

https://doi.org/10.1017/ssh.2024.18

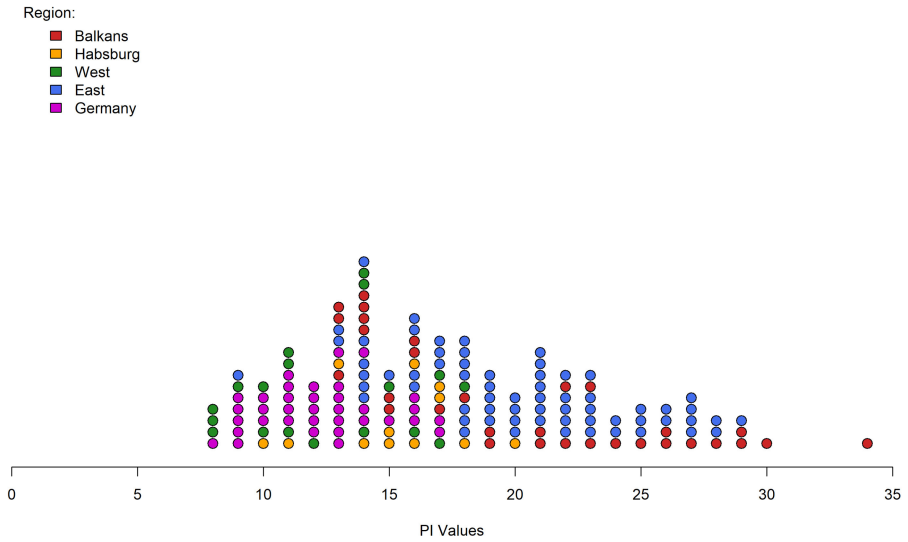

Figure 10. Distribution of Mosaic regions by the value of the Patriarchy Index, by five bigger territorial groupings.
Source: Gruber, Siegfried, Mikołaj Szołtysek, and Bartosz Ogórek (2023) Mosaic datafile, 2023 [machine-readable dataset]. IPUMS-International (mosaic.ipums.org).
Note: The groupings as in Figure 7.

Spatial analyses

Because the Mosaic data are geo-referenced (see above), a wide range of family demographic characteristics contained in the database can be projected onto geographic coordinates of specific populations and at different geographic levels. Thus, in addition to enabling descriptive mapping (see above), it is possible to take advantage of rapid advances in spatial computing technology (Gutmann et al. 2011) to examine more explicitly the local spatial patterns of particular aspects of family systems, and to identify and understand their spatial variability (Anselin 1995; Fotheringham 1997). Thus, analyses based on Mosaic data have considerable potential for improving on the findings of previous scholarship. This is because much of the research to date on the historical family demography of the continent has been conducted without spatially structured data or even basic forms of spatial modeling (e.g., Alter 2013; Ruggles 2010), despite the recognition that "place really did matter" (Goodchild 2008: 200) when it came to the evolution of family structures in historical Europe. The fact that only small quantities of data were collected for many areas of continental Europe in the "pre-Mosaic era" was obviously one of the factors that hindered the development of spatial models.[14]

Apart from the compilation of a vast collection of data, a prerequisite for moving forward in this area is having an appropriate definition of a network structure that reflects the idea of locality and connectivity (Anselin 1988). In the context of the

[14]Most of the data on family patterns used in previous research meet the definition of spatial data, i.e., data that can be decomposed into pairs of the form where x denotes a point in space-time and z denotes one or more properties of that point (e.g., household structure) (Goodchild 2008: 201).

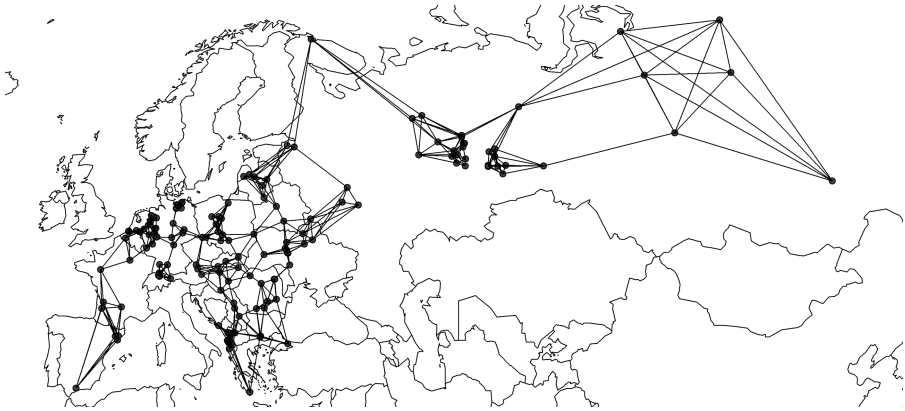

Figure 11. The connectivity graph showing the spatial weight matrix for Mosaic data.
Source: Gruber, Siegfried, Mikołaj Szołtysek, and Bartosz Ogórek (2023) Mosaic datafile, 2023 [machine-readable dataset]. IPUMS-International (mosaic.ipums.org).
Note: the five nearest neighbors (based on great circle distances) with a row-standardized inverse distance weight matrix.

spatial dispersion of Mosaic data points (and regions' centroids) and their uneven density in many parts of Europe, the network structure of the five nearest neighbors (based on great circle distances) with a row-standardized inverse distance weight matrix (Anselin 1988) seemed to be the most optimal solution (see Figure 11). With this approach, each spatial point in our data has exactly the same number of neighbors, but the relative importance (weight) of each neighborhood attribute is proportional to its inverse distance (Getis and Aldstadt 2004). This implies that the structure of our data can take into account spatial relationships and proximity, as expressed in the so-called first law of geography, which states that patches that are close to each other are generally more similar than those that are further apart (Tobler 1970). By applying this matrix to Mosaic data, we can formally regionalize the many demographic variables stored in the database and locate boundaries between areas, flagging areas with anomalous values within regions, or identifying local patterns that deviate from regional patterns.

Figure 12 uses the example of the "proportion of older people living in stem families" (Szołtysek et al. 2020) to produce what is known as the Moran scatter plot (Anselin 1995), which illustrates the relationship between the values of the focal attribute at each of the Mosaic sites and the average value of the same attribute at neighboring sites in the matrix. In this case, we see that the majority of the Mosaic data fall in the upper-right quadrant and the lower-left quadrant in Figure 12, corresponding to positive spatial autocorrelation (similar values are observed at neighboring sites, either as high-high or low-low spatial autocorrelation). This pattern is also confirmed by a global indicator of spatial autocorrelation (Moran's Global I), which is 0.43 ($p < .001$).[15] Using this scatter plot, we can also determine which areas of the Mosaic data map are most responsible for the observed high or low spatial autocorrelation, and which locations, if any, run counter to the overall

[15]The p value obtained from the Bonferroni correction.

https://doi.org/10.1017/ssh.2024.18

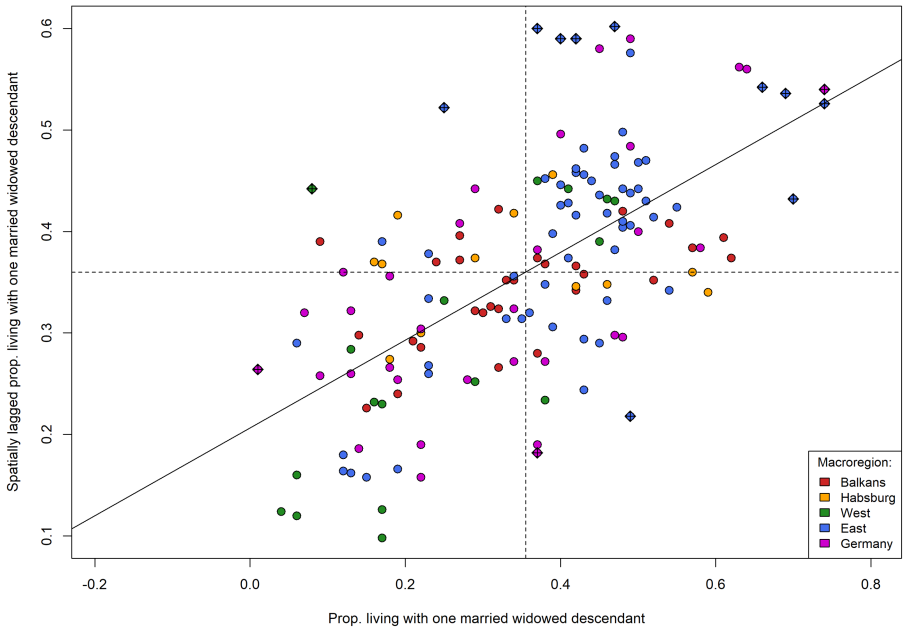

Figure 12. Moran scatter plot for the proportion of elderly living in stem family configurations.
Source: Gruber, Siegfried, Mikołaj Szołtysek, and Bartosz Ogórek (2023) Mosaic datafile, 2023 [machine-readable dataset]. IPUMS-International (mosaic.ipums.org).
Note: spatially lagged variable is equal to the average value of the variable of interest among the neighbors of each datapoint.

pattern. This allows to capture the variation better than the older approaches based on more fragmentary data and less rigorous (non-spatial) comparisons could.[16]

Data merging

The final area susceptible to innovation is related to the possibilities for expanding Mosaic data both vertically (in terms of content) and horizontally (in terms of scope). The former efforts stem from the motivation to increase the self-contained explanatory power of the database, and to move from describing to explaining patterns in the Mosaic data by embedding them in relevant sociocultural, demographic, and ecological/environmental contexts. In the "pre-Mosaic" studies, such gaps could occasionally be filled by intensive case studies or small subsystem studies and data triangulation (e.g., Mitterauer 1992). In large "surface" studies with multiple censuses, such a goal could only be achieved by mobilizing exogenous information from different sources and areas, which was then linked to the demographic/family data in Mosaic through geographical linkage and spatial overlay.

[16]Note that the global Moran's I ignores the potential instability in space, and therefore suggests the use of other spatial analysis tools, such as local indicators of spatial association (LISA), and, in particular, the local Moran's *I* (Anselin 1995).

https://doi.org/10.1017/ssh.2024.18

First, the regional Mosaic populations were linked to information on the prevailing infant mortality rate (hereafter IMR) and life expectancy at birth (e0), based on the assumption that both parameters had an important influence on living arrangements (Ruggles 1987). Despite the heterogeneity of the procedures used to obtain this information (both data fusion and top-down/bottom-up extrapolations had to be used), a total of 160 Mosaic regions were assigned IMR values, and 145 regions were assigned the corresponding e0 values (Szołtysek, Ogórek, et al. 2022). The data collected were generally consistent with the spatial distribution and evolution of infant mortality and life expectancy in historical Europe. Both variables also showed an expected mutual correlation (Pearson $r = -.68$ ($p < 0.001$).

In addition, for each regional population included in our database, the stage of demographic development was approximated by matching the respective data with the corresponding provincial-level estimates of the onset of fertility decline from the European Princeton Fertility Project (Coale and Watkins 1986). Accordingly, a dummy variable was created for each regional population that indicated whether the respective population belonged to a province that had already experienced monotonic fertility decline at the time of the census. Overall, the three variables discussed above could be used as moderately coarse control variables in modeling various family demographic processes operationalized with the Mosaic data, along with some variables that could be derived from the data itself (e.g., SMAM or child-women ratios) (e.g., Szołtysek, Beltran Tapia, et al. 2022).

The next example of the vertical extension of the data concerns the possibilities for including environmental variables, either to use them as explananda of the European family patterns recorded in Mosaic or to include them as control variables in specific studies.[17] Again, such enrichment efforts can be done by collecting information from various increasingly available Big Data repositories on environmental features and biogeographical conditions.[18]

Figure 13 shows some of the existing possibilities in which Mosaic regional data are overlaid and directly linked to specific contemporary geo-environmental raster data or to existing areal and raster top-down reconstructions of land-use patterns at the global scale.

For example, the measure of terrain ruggedness can be calculated separately for each of the Mosaic sites by weighting the gridded elevation data by the population size of the regions. This measure, which is perhaps the least controversial of all the geo-variates considered here, has already been shown to be a good and robust predictor of a range of family demographic characteristics drawn from the Mosaic data (e.g., Szołtysek, Beltran Tapia, et al. 2022). Similarly, measures of the suitability of land for agriculture and the proportion of land under cultivation, either separately or in combination, can be used as rough proxy measures for the impact of geographical characteristics on the ecological endowment and historical role of agriculture in a given region. It is noteworthy, for example, that the three measures alone explain 11.5 percent of the variation in the proportion of multiple-family

[17]The idea that family patterns can be a reaction to certain environmental influences goes back to F. Le Play.

[18]In the absence of large-scale (global) historical land use data, most of the datasets considered here consisted of a combination of real data and modeling (see section below).

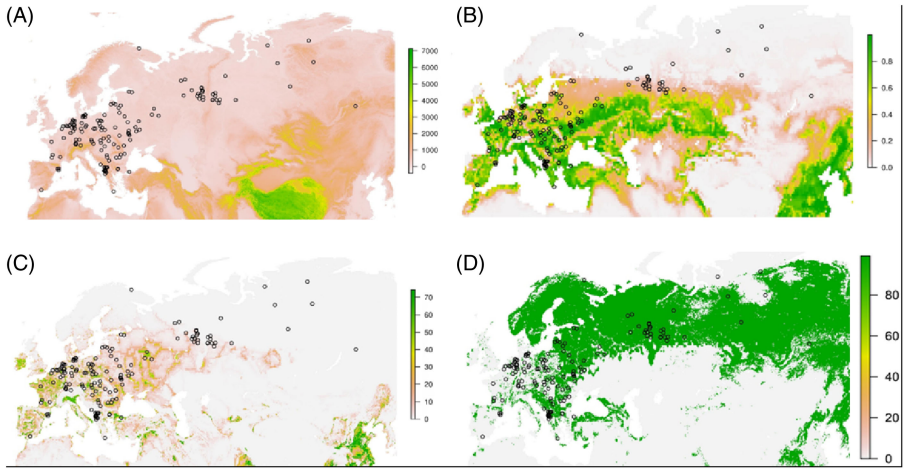

Figure 13. Mosaic regions' centroids overlaid over selected geocovariates.
Source: for Mosaic - Gruber, Siegfried, Mikołaj Szołtysek, and Bartosz Ogórek (2023) Mosaic datafile, 2023 [machine-readable dataset]. IPUMS-International (mosaic.ipums.org).
For (**A**) - ***Gridded elevation***: the GTOPO30 dataset (downloaded 30 and 31 August 2016 from http://earthexplorer.usgs.gov/; files: gt30e020n40, gt30e020n90, gt30w020n40, gt30w020n90, gt30w060n90). For (**B**) – ***Land suitability for agriculture;*** Ramankutty et al. (2002), data download - https://sage.nelson.wisc.edu/data-and-models/atlas-of-the-biosphere/mapping-the-biosphere/land-use/suitability-for-agriculture/. For (**C**) - ***Share croplands (1800):*** the History Database of the Global Environment (HYDE 3.1; https://public.yoda.uu.nl/geo/UU01/G4HO5I.html). For (**D**) – ***Share forests (1800):*** Ellis et al. (2010).
Notes: All this information, over which the Mosaic data is superimposed, has been converted into numerical data on an interval scale linked to the centroids of the regions.
To derive the information on the terrain ruggedness, we used the Terrain Ruggedness Index (TRI) (Wilson et al. 2007). For this, we applied the focal function in the R- library raster (the TRI formula can be found in the help function of "terrain" in the raster library). For Mosaic sites, we generated the information for all sites by looking at the raster data within a circle with a diameter of 7.5 km around the site coordinates. Based on this data, we derived the population-weighted values for all regions.
The Land suitability for agriculture (LSA) index was extracted from the corresponding raster files using the "extract" function in the R package "raster." For each region in our database (residential points for Mosaic), a population-weighted centroid was first derived so that the value of the variable reflected the mean around that point (with the buffer size set at 50 km to better capture local variation in the environment). The same procedures were applied to cropland. However, since the proportion of cropland is available from HYDE 3.1 for each decade after 1700, we can use the raster for the date closest to each census date for each region in our collection. The afforestation data is available in four scenarios (1700, 1800, 1900, and 2000). Again, we can use the grid data closest to the respective census date.

households in the Mosaic dataset (results of the ordinary least squared regression [OLS]).[19]

One of the greatest benefits of data harmonization is that it allows data collected in different cultural contexts and over long periods of time to be brought together (see Borgman 2015; Kitchin 2014). In the context of Mosaic, this created an interoperability that made it possible for the first time to place the large-scale family

[19]Ultimately, using this information along with information on urban-rural differences at the regional level, as well as on the SES of a household head within a framework of a multilevel modeling, can help us better understand the extent to which non-nuclear family forms were intrinsically connected to agriculture.

https://doi.org/10.1017/ssh.2024.18

demographic patterns of historical continental Europe into a much broader comparative framework than ever before.

In the first instance, Mosaic could be easily integrated into the largest collection of nationally representative historical European census microdata compiled by the North Atlantic Population Project (distributed by IPUMS-International; Ruggles et al. 2011). As the Mosaic data tend to be chronologically biased towards earlier periods, to ensure comparability, preference was given to the oldest available censuses when selecting NAPP data (i.e., for Iceland, Denmark, England and Wales, and Sweden), using complete censuses in each case (or samples thereof).[20] To achieve a relative balance in the number of regions between the two data corpora, the microdata from the NAPP were aggregated into 156 administrative units used in the respective census, and were included in the NAPP (generally counties).

This combination of the Mosaic and NAPP datasets created a real critical mass of data that has already led to a number of unexpected discoveries. It revealed for the first time the full range of family patterns across Europe, from the simplest to the most complex. It also showed that many assumed features of the north-west type of family organization were present in parts of Europe where they had not been expected, and often with intensities greater than those in the alleged "core" areas (Szołtysek and Ogórek 2020; Szołtysek, Ogórek et al. 2020; cf. Dennison and Ogilvie 2014; Ruggles 2010). Finally, in the Mosaic data, the highest levels of agreement in terms of mutual associations between the four household formation traits advocated by Hajnal were found outside the north-western "heartlands" in different central European populations (Szołtysek et al. 2021).

However, even more global accounts could be created by merging historical and current data, as the harmonized structure of Mosaic and NAPP is fairly closely aligned with IPUMS-International's global data. With such a goal in mind, a "global" patriarchal dataset has recently been created that combines Mosaic/NAPP data on 311 regions with 29 million people of historical Europe and North America with IPUMS-I data on 22 countries with 65 million people for the 1970–2014 period, and projects 546 territorial units (Figure 14). Such a comparative dataset can serve various purposes, including to map the concentration of patriarchal family systems in a "global" regional perspective by confronting the alleged European "uniqueness" using a Eurasian mirror; to examine the differences between historical Europe and its North Atlantic offshore territories in the past; or to assess how differences in basic historical and structural conditions (while also taking into account the factors discussed in Figure 13) have conditioned the emergence of various patriarchal formations (Szołtysek et al. 2022).[21]

[20]The Scottish data suggested for integration are from the 1881 census rather than from the 1851 census, as it was not possible in the latter census to derive a rural dataset and infant mortality estimates close to the census date. All UK data are from the censuses provided to NAPP/IPUMS-I by the I-CeM project: https://icem.data-archive.ac.uk/#step1

[21]These issues are the subject of our ongoing work in which we use Bayesian hierarchical modeling to assess variation in family patriarchy patterns in the Mosaic, NAPP, and IPUMS data, controlling for historical and current GDP levels, rural population share, prevalent religion, fertility and mortality (life expectancy), and a range of environmental covariates.

https://doi.org/10.1017/ssh.2024.18

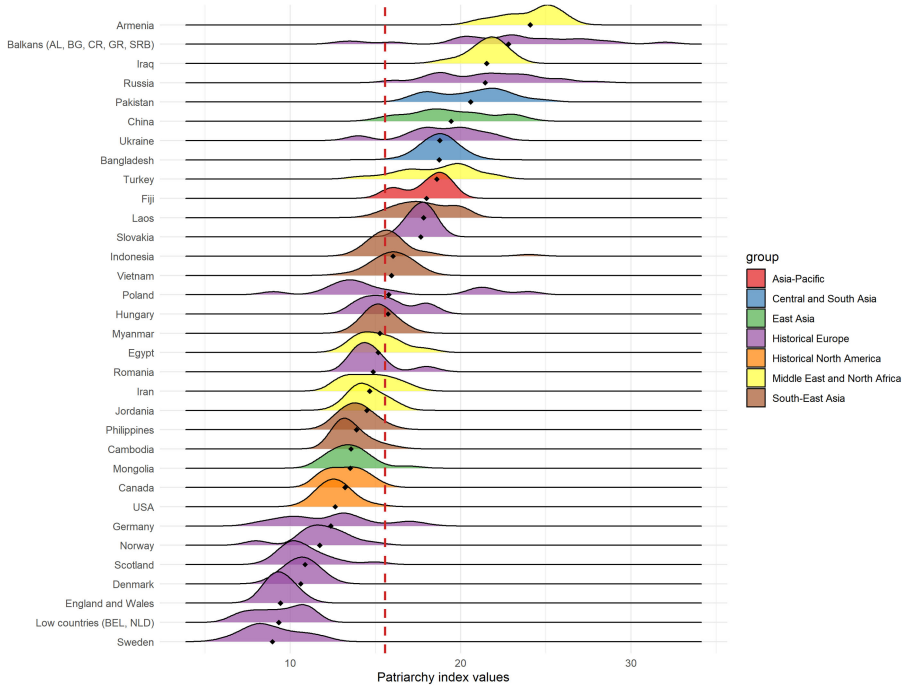

Figure 14. Within-country regional distribution of the Patriarchy Index across Eurasia.
Sources: For Mosaic – Gruber, Siegfried, Mikołaj Szołtysek, and Bartosz Ogórek (2023) Mosaic datafile, 2023 [machine-readable dataset]. IPUMS-International (mosaic.ipums.org).
For NAPP and IPUMS-I: Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 7.2 [dataset]. Minneapolis, MN: IPUMS, 2019, https://doi.org/10.18128/D020.V7.2
Notes: The data used for the figure comprised 311 regional populations from 1700 to 1926 with 29 million individuals. The contemporary data included 546 regions from 21 countries with 65 million individuals. Each of the stacked histograms refers to the distribution of regional PI values within a country and contains the mean PI value for a particular region. The dashed vertical line shows the mean PI value for the entire data set.

Challenges

Given their scope and coverage, the Mosaic data surpass all previous efforts to create an infrastructure for family history data in continental Europe and offer many promising research opportunities. However, the use of these data comes with certain challenges.

Italy and the Iberian Peninsula are either not included or insufficiently included in the current dataset. This data gap limits our ability to explore the north-south dimension of variation in family systems in Europe (Reher 1998) and may represent a missing element in the development of a "new" geography of family patterns based on machine-learning technologies.

Broad cross-cultural and cross-temporal comparisons using Mosaic data could pose epistemic risks in terms of the ontological status of the basic census units "unearthed" from historical censuses or census-like registers. These may arise if there is too little cross-cultural overlap in census definitions (which risks occidentalisation); if the term "household," as defined by survey statisticians to

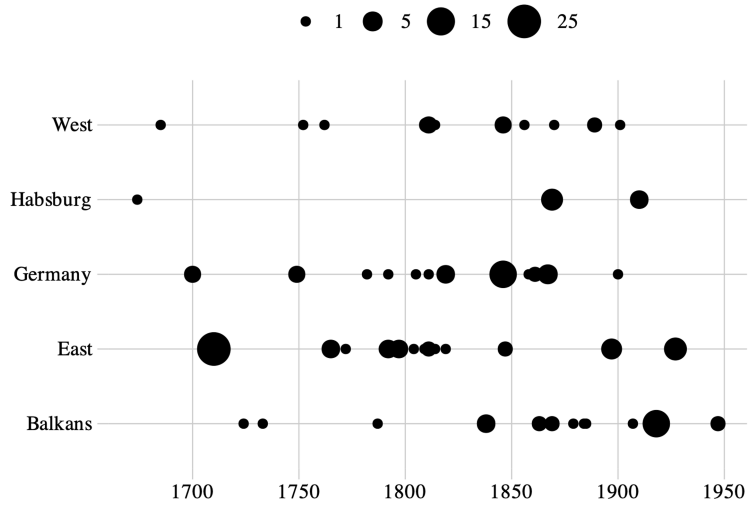

Figure 15. Spatio-temporal variation in the Mosaic data.
Source: Gruber, Siegfried, Mikołaj Szołtysek, and Bartosz Ogórek (2023) Mosaic datafile, 2023 [machine-readable dataset]. IPUMS-International (mosaic.ipums.org).
Notes: The size of the circles indicates the number of regions in each period and region. Five bigger territorial groupings as in Figure 7.

ensure complete coverage, is not consistent with particular economic or social concepts (requiring a distinction between "etic" and "emic" ways of grouping people; see Szreter et al. 2004); and if census "units" are taken out of context by overly mechanistic standardisation (requiring careful cross-cultural translation of the source material) (see Szołtysek 2023).

Because Mosaic captures populations that are unevenly distributed across time and space, each time window of the dataset literally contains different populations, even within broad macro-regions (Figure 15). As well as severely limiting the analysis of family change (although some broad temporal trends can certainly be identified), this seems to contradict the idea of comparing elements of different temporal sequences without a clear idea of the extent to which they might change over time (Wawro and Katznelson 2022; for similar examples in earlier studies, see Barbagli 1991; Hajnal 1982; Laslett 1977; Smith 1993; Wall 2001; cf. Dennison and Ogilvie 2014).

Since this mixing of time periods is virtually unavoidable with such extensive data (and an ideal data structure to mitigate this problem is unrealistic), we make four practical suggestions to ensure that analyses based on Mosaic data are justified even with this caveat in mind. First, 146 of the 161 Mosaic populations (90 percent of the current regions) represent populations that have not yet experienced a fertility transition, and, with the exception of France, most regions without this characteristic are widely dispersed without changing the overall picture. This narrows the gap between the Mosaic populations, at least in terms of the general demographic characteristics that most of them have long exhibited (Del Panta et al. 2006). Second, the two largest data collections of "recent" populations in Mosaic (the 1918 census for Albania and the 1926 Polar Census) represent not only

https://doi.org/10.1017/ssh.2024.18

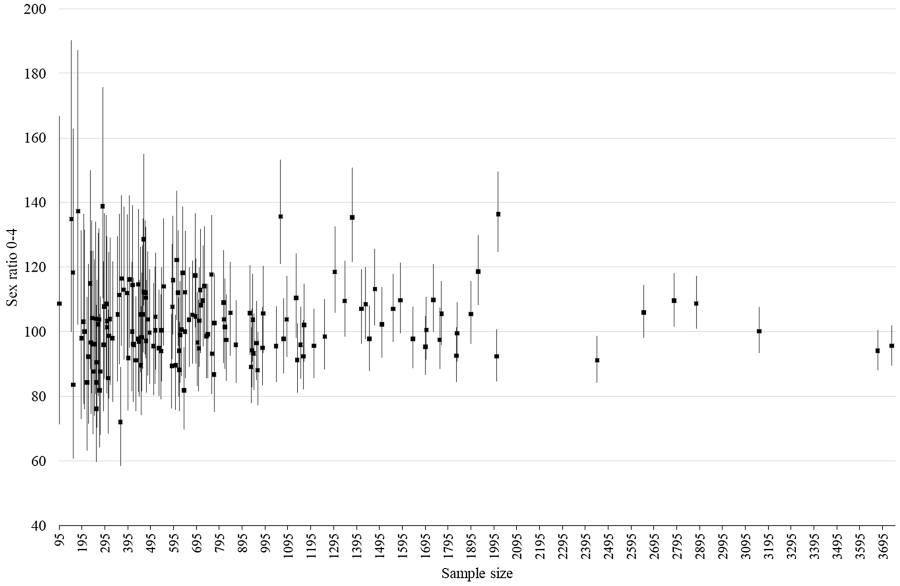

Figure 16. Bootstrapped sex ratios and their corresponding 95% confidence intervals by sample size, Mosaic data.
Source: Gruber, Siegfried, Mikołaj Szołtysek, and Bartosz Ogórek (2023) Mosaic datafile, 2023 [machine-readable dataset]. IPUMS-International (mosaic.ipums.org).
Notes: sample size refers to the number of children 0–4 in particular region. Confidence intervals based on resampling with replacement (5,000).

pre-transitional populations but also quite archaic family organizations, further reducing the seemingly huge time span of the data. Third, we suggest that all multivariate analyses of Mosaic data always include the period or other time-varying covariates (census quality, onset of fertility decline, IMR, or e0) as control variables. Finally, the pooled time cross-sections from Mosaic should ideally be cross-checked with other place-specific evidence before they can be assumed to represent family patterns that are durable beyond the specific time window covered by the data (e.g., Reher 1998; Schürer et al. 2018; Therborn 2004).

The fact that the Mosaic dataset has a huge overall volume does not necessarily mean that all its variables are free from noise generated by small Ns. Although Mosaic has tried to minimize these potential effects by creating regions that are "large enough" (see above), and thus allow the random fluctuations to become smaller as the sample size increases, population size can still be an issue if the calculation of certain variables requires a large reduction in the denominator (e.g., for age-specific measures).

An exemplary variable of this type is the child sex ratio, i.e., the number of males per 100 females in the 0–4 age group, which is commonly used as a cumulative measure of sex-specific mortality around birth, in infancy, and in childhood (Szołtysek, Ogórek, et al. 2022). In Figure 16, the original sex ratios of the original samples (represented by filled squares) are overlaid with the 2.5 and 97.5 percentiles of the distribution of sex ratios resulting from the bootstrapping procedure using

https://doi.org/10.1017/ssh.2024.18

individual-level information from the Mosaic data (5000 sex ratio values were resampled for each region). This exercise clearly shows that the uncertainty of the calculated measure (sex ratio) increases dramatically as the sample size (the number of children 0–4 in the region) decreases. A practical lesson that can be drawn from this example is that researchers using Mosaic data should always be mindful of which at-risk population is being considered for particular demographic indicators, and should take every precaution when proceeding with the analysis. It is recommended that researchers apply resampling methods that use individual-level information from the Mosaic data file attached to the regional file to gain more confidence in specific measures.

As was mentioned above, Mosaic's core data are relatively weak semantically, and linking them to additional contextual information (see above) leads to insurmountable limitations. For many potentially critical intervening factors (e.g., the socioeconomic structures and the labor, inheritance, and kinship patterns), creating relevant variables based on information from the secondary literature or from the original data providers would be extremely tedious, unproductive, and most likely impossible for the entire dataset. Many hindcast reconstructions of historical land-use patterns (see above) are clearly not "data" in the sense of measured quantities, but are, rather, good guesses about what happened (e.g., Klein Goldewijk and Verburg 2013). For many of these areas, it would only be possible to obtain meaningful information in the context of high-resolution local case studies (Hedefalk et al. 2017), which, once again, is not feasible for all Mosaic data points. These limitations should be kept in mind when developing multivariate models with the Mosaic data.

Furthermore, Mosaic data are not particularly useful for individual life course analyses, and their linkage/integration with longitudinal databases is actually quite cumbersome (Mandemakers et al. 2023). This apparent lack of synergy is in fact reciprocal, as transforming the latter into a cross-sectional matrix of the NAPP/Mosaic data structure would require generalized solutions that are currently difficult or impossible to obtain (cf. Alter et al. 2009). Nevertheless, both the Mosaic data and the longitudinal data can serve the common goal of charting and explaining demographic dynamics (cf. Dillon and Roberts 2002). First, the existing longitudinal databases could become a source of additional information for Mosaic-like large "surface" studies, at least for some areas of historical Europe (and even beyond). Moreover, as many of the longitudinal data sources are highly localized (e.g., Matthijs and Moreels 2010), they could benefit from the use of Mosaic data to assess the relative importance of particular family demographic contexts. This is particularly true with regard to the Mosaic project's potential to outline broad regional patterns of life course transitions across cohorts (as mentioned above).

Last but not least, the successful management of a project like Mosaic requires a combination of different practices, skills, and technologies, and necessitates interdisciplinary conversations between scientists who do not always communicate directly with each other. Such collaboration can be very difficult without long-term and flexible institutional support, a long-term vision, and a commitment to manage and be accountable for the content on behalf of the data curators (Borgman 2015; Kitchin 2014: 40). Strong institutional support is also crucial to continue the long-term task of digitizing and curating additional microdata samples for many parts of

Europe in the future (cf. Emigh and Hernández-Pérez 2022), especially as such efforts often require international interactions and collaboration across large distances.

Conclusions

The main motivation for initiating the Mosaic project was a lack of existing comparative family history data, which, it was argued, had to be overcome to answer more systematically many important research questions related to our understanding of the population and family history of continental Europe. In this paper, we have explored the opportunities and challenges associated with filling this gap by developing and exploring a specifically European data infrastructure on historical family patterns.

The changes that Mosaic has ushered in reshape some of the fundamental principles of family history research in the data domain. For most of its history, historical family demography has operated in a data-poor environment in which measurements of many aspects of family organization have been difficult or inaccessible, or have been expensive and cumbersome to obtain, purchase, and process. Thanks to the Mosaic database, scholars interested in researching family history now have access to an unprecedented amount of fine-grained data on populations and societies, regions, and small areas and places, with a large share of these data referenced in geo-space and time.

The proposed vision of change goes beyond the purely technical aspects. Scaling from traditionally small data infrastructures to much larger data infrastructures leads to the introduction of new approaches to data processing and analysis that enable older questions to be answered and new questions to be asked in a more efficient way. By enabling them to shift from a data-poor to a data-rich approach to analyzing historical family systems, Mosaic provides researchers with opportunities to move from coarse aggregations to high resolutions, from simple descriptions to complex modeling, and from tentative observations to formal pattern recognition. These advances should, in turn, lead to a much broader, deeper, and more comprehensive understanding of past family patterns. A fuller history of European family organization can now be provided using a range of approaches, from sharpening and developing insights that have often been marginalized, obscured, or only secondarily addressed; to engaging in Big Data-like data *dredging* to comprehensively examine relationships between a large number of variables for which data are available.

Mosaic also raises fundamental questions about the organization and practice of historical family research (Borgman 2015). Efforts like the project discussed here offer new possibilities for fostering interdisciplinary collaborations beyond the lone-scholar model that has long dominated family history research. The complexity of research practices and the possible ways to explore Mosaic-like data inevitably encourage more (network) collaborations ("crowdsourcing of minds"), especially (but not only) between "computationally literate social scientists and socially literate computer scientists" (Kitchin 2014: 137). The usage of Mosaic data may also improve the levels of research productivity within the field (especially in the context

of public data sharing), the possibilities for further data re-use, and the provision of test-bed data for teaching and student projects.

Although the large volume of data collected by Mosaic may produce important innovations and improvements on previous studies of historical family systems based on more limited data, there are also strong continuities and potential synergies between the Mosaic project and the older practices of historical family demographers. For example, Mosaic does not advocate entirely replacing older studies with small datasets with large datasets analyzed using automated approaches. While the Mosaic database offers opportunities for conducting large-scale "surface" studies, it can also support more traditional approaches that focus on in-depth analysis of smaller entities, be it a community or a village. Small-scale studies can answer more finely tailored research questions or specifically formulated comparisons, telling individual, nuanced, and contextual stories, while also being less resource-intensive (cf. Kitchin 2014: 29 ff). At the same time, the Mosaic database can help the authors of such studies develop better micro-stories (i.e., embedded in larger structures).

The "deluge" of Mosaic-like structured information on historical family patterns notwithstanding, some important areas are still not yet covered by the dataset. Thus, the organizers of the project are eager for it to grow bigger. The Mosaic project's ability to generalize about the European familial past would definitely improve if more data on the Iberian, Mediterranean, Russian, and perhaps also French areas could be included. Moreover, the project's ability to generalize about the place of Europe in world family systems would be enhanced if historical census and census-like microdata from Asia could be combined with its data (e.g., Dong et al. 2015; Ochiai and Hirai 2023); an opportunity that so far has not been taken up by any Asian colleagues. Although the scope of the data that could be usefully included in the database is not infinite, Mosaic is still far from the point beyond which further data would not add any (new) information (cf. Succi and Coveney 2019). For such data expansions to happen in the future, large amounts of funding, institutional support, and cooperation of the broader research community for data curation would be necessary.

Databases are now more widespread than microscopes, voltmeters, and test tubes. The increasing amount of data has led to major changes in research practices, and historical family demography is no exception to this general trend. While the Mosaic project is probably not a prime example of the use of "Big Data" (although the data might be called "biggish"), its transformative capabilities should not be ignored. With "data is the new oil" as the motto of our *Zeitgeist,* we are challenged to remember that mining new horizons of data can indeed yield scientifically useful insights even within the confines of historical family demography. Despite the potential outcry from parts of the family history community over such practices (e.g., Dennison 2021; Devos 2016), it is unlikely that the trend of adopting large-scale data solutions in historical family demography will be slowed down and reshaped. We argue that social science historians of the family should recognize and face the challenges associated with large-scale data projects. The price of missing out on such opportunities may be high, given that family historians have already lost some of their previous standing as the primary interpreters of the panoramic worlds of historical family (see, e.g., Bertocchi and Bozzano 2019; Duranton et al. 2009;

Gutman and Voigt 2022). After all, if the avalanche of data is here, shouldn't we be digging?

Acknowledgments. We thank Joshua Goldstein for his encouragement and thorough support in the development of the Mosaic database.

Funding. Mikolaj Szoltysek disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research has been funded in whole by the National Science Centre (Poland) under the grant scheme OPUS (no. 2022/47/B/HS3/00004).

References

Alter, George, Kees Mandemakers, and Myron P. Gutmann (2009) "Defining and distributing longitudinal historical data in a general way through an intermediate structure." Historical Social Research **34** (3): 78–114.

Alter, George (2013) "Generation to generation: Life course, family, and community." Social Science History **37** (1): 1–26.

Anderson, Michael (1980) Approaches to the History of the Western Family 1500–1914. London: Macmillan.

Anselin, Luc (1988) Spatial Econometrics: Methods and Models. Dordrecht: Springer.

—— (1995) "Local indicators of spatial association—LISA." Geographical Analysis **27**: 93–115.

Aronova, Elena, Christine von Oertzen, and David Sepkoski (2017) "Introduction: Historicizing Big Data." Osiris **32** (1): 1–17.

Barbagli, Marzio (1991) "Three household formation systems in eighteenth- and nineteenth-century Italy," in David I. Kertzer and Richard. P. Saller (eds.) The Family in Italy from Antiquity to the Present. New Haven, CT: Yale University Press: 255–69.

Bertocchi, Graziella, and Bozzano, Monica (2019) "Origins and implications of family structure across Italian provinces in historical perspective," in Claude Diebolt, Auke Rijpma, Sarah Carmichael, Sellin Dilli, and Charlotte Störmer (eds.) Cliometrics of the Family. Studies in Economic History. Cham: Springer: 121–47.

Bohon, Stephanie A. (2018) "Demography in the big data revolution: changing the culture to forge new frontiers." Population Research and Policy Review **37**: 323–41.

Boonstra, Onno, Breure Leen, and Peter Doorn (2006) "Past, present and future of historical information science." Historical Social Research **29**: 4–132.

Borgman, Christine L. (2015) Big Data, Little Data, No Data: Scholarship in the Networked World. Cambridge, MA: The MIT Press.

Boyd Danah, and Crawford, Kate (2012) "Critical questions for big data." Information, Communication & Society **15** (5): 662–79.

Burguière, André, Christiane Klapisch-Zuber, Martine Segalen, and Françoise Zonabend (eds.) (1996) A History of the Family, Vol. 2, The Impact of Modernity. Cambridge: Polity Press.

Burguière, André, and François Lebrun (1996) "The one hundred and one families of Europe," in André Burguière, Christiane Klapisch-Zuber, Martine Segalen, and Françoise Zonabend (eds.) A History of the Family, Vol. 2, The Impact of Modernity. Cambridge: Polity Press: 11–94.

Coale, Ansley J., and Watkins, Susan C. (eds.) (1986) The Decline of Fertility in Europe. Princeton: Princeton University Press.

Del Panta, Lorenzo, Rosella Rettaroli, and Paul-Andre Rosental (2006) "Methods of historical demography," in Graziella Caselli, Jacques Vallin, and Guillaume Wunsch (eds.) Demography: Analysis and Synthesis. A treatise in Population. Vol. 4. Elsevier: Academic Press: 597–618.

Dennison, Tracy, and Sheilagh Ogilvie (2014) "Does the European marriage pattern explain economic growth?" The Journal of Economic History **74:** 651–93.

Dennison, Tracy (2021) "Context is everything: the problem of history in quantitative social science." Journal of Historical Political Economy **1** (1): 105–26.

Devos, Isabelle (2016) "Not everything that counts can be counted, and not everything that can be counted counts," in Koen Matthijs and Jan Kok (eds.) The Future of Historical Demography. Leuven: Acco: 156–60.

Dillon, Lisa, and Evan Roberts (2002) "Introduction: Longitudinal and cross-sectional historical data: Intersections and opportunities." History and Computing **4** (1–2): 1–7.

Dong, Hao, Cameron Campbell, Satomi Kurosu, Wenshan Yang, and James Z. Lee (2015) "New sources for comparative social science: historical population panel data from East Asia." Demography **52** (3): 1061–88.

Duranton, Gilles, Andrés Rodríguez-Pose, and Richard Sandall (2009) "Family types and the persistence of regional disparities in Europe." Economic Geography **85** (1): 23–47.

Ellis, Erle C., Kees Klein Goldewijk, Stefan Siebert, Deborah Lightman, and Navin Ramankutty (2010) "Anthropogenic transformation of the biomes, 1700 to 2000." Global Ecology and Biogeography **19**: 589–606.

Emigh, Rebecca J., and Johanna Hernández-Pérez (2022) "The present of the past: a sociotechnological framework for understanding the availability of research materials." IEEE Annals of the History of Computing **44** (4): 16–27.

Flandrin, Jean-Louis (1979) Families in Former Times. Kinship, Household and Sexuality in Early Modern France. trans. by Richard Southern. Cambridge: Cambridge University Press.

Fotheringham, Alexander Stewart (1997) "Trends in quantitative methods I: Stressing the local." Progress in Human Geography **21** (1): 88–96.

Getis, Arthur, and Jared Aldstadt (2004) "Constructing the spatial weights matrix using a local statistic." Geographical Analysis **36**: 90–104.

Goodchild, Michael F. (2008) "Spatial data analysis," in Shashi Shekhar and Hui Xiong (eds.) Encyclopedia of GIS. New York: Springer: 200–203.

Gruber, Siegfried, and Mikołaj Szołtysek (2016) "The patriarchy index: a comparative study of power relations across historical Europe." History of the Family **21**: 133–74.

Gutmann, Myron P., Glenn D. Deane, Emily R. Merchant, and Kenneth M. Sylvester (2011) "Introduction," in Emily R. Merchant, Glenn D. Deane, Myron P. Gutmann, and Kenneth M. Sylvester (eds.) Navigating Time and Space in Population Studies. International Studies in Population, Vol. 9. Springer: Dordrecht: 1–17.

Gutmann, Jerg, and Stefan Voigt (2022) "Testing Todd: family types and development." Journal of Institutional Economics **18** (1): 101–18.

Hajnal, John (1982) "Two kinds of preindustrial household formation system." Population and Development Review **8**: 449–94.

Hammel, Eugene A., and Peter Laslett (1974) "Comparing household structure over time and between cultures." Comparative Studies in Society and History **16** (1): 73–109.

Han, Jiawei, Micheline Kamber, and Jian Pei (2012) Data Mining Concepts and Techniques. 3rd ed. Morgan Kaufmann Publishers: Waltham.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer: New York.

Hedefalk, Finn, Patrick Svensson, and Lars Harrie (2017) "Spatiotemporal historical datasets at micro-level for geocoded individuals in five Swedish parishes, 1813–1914." Scientific Data **4**: 170046, https://doi.org/10.1038/sdata.2017.46.

Kaser, Karl, Siegfried Gruber, Gentiana Kera, and Enriketa Pandelejmoni (2011) 1918 census of Albania, Version 0.1 [SPSS file]. Graz.

Kertzer, David I., and Marzio Barbagli (eds.) (2001) Family Life in Early Modern Times 1500–1789. The History of the European Family, Vol. **1**. New Haven and London: Yale University Press.

Kitchin, Robert (2014) The Data Revolution: Big Data, Open Data, Data Infrastructures & their Consequences. Los Angeles: SAGE Publications Ltd.

Klein Goldewijk, Kees, and Peter H. Verburg (2013) "Uncertainties in global-scale reconstructions of historical land use: an illustration using the HYDE data set." Landscape Ecology **28**: 861–77.

Kurosu, Satomi (2016) "Historical demography going 'glocal': Eurasia project and Japan," in Koen Matthijs, Saskia Hin, Jan Kok, and Hideko Matsuo (eds.) The Future of Historical Demography: Upside Down and Inside Out. Leuven/Den Haag: Acco: 60–62.

Laslett, Peter (1965) The World We Have Lost, 1st ed. London: Methuen.

—— (1977) "Characteristics of the western family considered over time." Journal of Family History **2**: 89–115.

https://doi.org/10.1017/ssh.2024.18

—— (1983) "Family and household as work group and kin group: areas of traditional Europe compared," in Richard Wall and Jean Robin (eds.) Family Forms in Historic Europe. Cambridge: Cambridge University Press: 513–63.

Le Play, Frédéric (1877–1879) Les ouvriers européens. 6 vols. Tours: A. Mame et fils.

Mandemakers, Kees, George Alter, Hélène Vézina, and Paul Puschmann (eds.) (2023) Sowing: The Construction of Historical Longitudinal Population Databases. Nijmegen: Radboud University Press.

Matthijs, Koen, and Sarah Moreels (2010) "The Antwerp COR*-database: A unique flemish source for historical-demographic research." The History of the Family **15** (1): 109–15.

Mitterauer, Michael (1992) "Peasant and non-peasant family forms in relation to the physical environment and the local economy." Journal of Family History **17**: 139–59.

—— (2003) "European kinship systems and household structures: medieval origins," in Patrick Heady and Hannes Grandits (eds.) Distinct inheritances. Property, family and community in a changing Europe. Münster: Lit Verlag: 35–52.

Ochiai, Emiko, and Hirai Shoko (eds.) (2023) Japanizing Japanese Families. Leiden, The Netherlands: Brill.

Plakans, Andrejs, and Charles Wetherell (2005) "The Hajnal line and Eastern Europe," in Theo Engelen and Arthur P. Wolf (eds.) Marriage and the Family in Eurasia. Perspectives on the Hajnal Hypothesis. Amsterdam: Aksant: 105–26.

Pujadas-Mora, Joana, Alicia Fornés, Terrades O. Ramos, Josep Lladós, Jialuo Chen, Miquel Valls-Fígols, and Anna Cabré (2022) "The Barcelona Historical Marriage Database and the Baix Llobregat Demographic Database. From algorithms for handwriting recognition to individual-level demographic and socioeconomic data." Historical Life Course Studies **12**: 99–132, https://doi.org/10.51964/hlcs11971.

Ramankutty, Navin, Jonathana A. Foley, John Norman, and Kevin McSweeney (2002) "The global distribution of cultivable lands: current patterns and sensitivity to possible climate change." Global Ecology and Biogeography **29** (11): 377–392.

Reher, David S. (1998) "Family ties in Western Europe: persistent contrasts." Population and Development Review **4** (2): 203–34.

Ruggles, Steven (1987) Prolonged Connections: The Rise of the Extended Family in Nineteenth-Century England and America. Madison: University of Wisconsin Press.

—— (1995) "Family interrelationships." Historical Methods: A Journal of Quantitative and Interdisciplinary History **28**: 52–8.

—— (2010) "Stem families and joint families in comparative historical perspective." Population and Development Review **36**: 563–77.

—— (2012) "The future of historical family demography." Annual Review of Sociology **38**: 423–41.

—— (2014) "Big microdata for population research." Demography **51** (1): 287–97.

Ruggles, Steven, Evan Roberts, Sula Sarkar, and Matthew Sobek (2011) "The north Atlantic population project: progress and prospects." Historical Methods: A Journal of Quantitative and Interdisciplinary History **44** (1): 1–6.

Schürer, Kevin (1986) "Historical databases and the researcher," in Manfred Thaller (ed.) Datenbanken und Datenverwaltungssysteme als Werkzeuge historischer Forschung. St. Katharinen: Scripta Mercaturae Verl.: 145–57.

—— (2004) "Leaving home in England and Wales 1850–1920," in Frans Van Poppel, Michel Oris, and James Lee (eds.) The road to independence. Leaving home in Eastern and Western societies, 16th–20th centuries. Bern-Bruxelles: Peter Lang: 33–84.

Schürer, Kevin, and Richard Wall (1986) "Computing the history of the family: a question of standards," in Manfred Thaller (ed.) Datenbanken und Datenverwaltungssysteme als Werkzeuge historischer Forschung. St. Katharinen: Scripta Mercaturae Verl.: 159–74.

Schürer, Kevin, and Tatiana Penkova (2015) "Creating a typology of parishes in England and Wales: Mining 1881 census data." Historical Life Course Studies **2**: 38–57.

Schürer, Kevin, Eilidh Garrett, Hannalis Jaadla, and Alice Reid (2018) "Household and family structure in England and Wales (1851–1911)." Continuity and Change **33** (3): 365–411.

Smith, Daniel Scott (1984) "A mean and random past: the implications of variance for history." Historical Methods: A Journal of Quantitative and Interdisciplinary History **17** (3): 141–48.

—— (1993) "American family and demographic patterns and the Northwest European Model." Continuity and Change **8** (3): 389–415.

Sobek, Matthew, and Sheela Kennedy (2009) The Development of Family Interrelationship Variables for International Census Data. Minneapolis, MN: University of Minnesota Working Paper No. 2009-02, https://doi.org/10.18128/MPC2009-02.

Sobek, Matthew, Lara Cleveland, Sarah Flood, Patricia Kelly Hall, Miriam L. King, Steven Ruggles, and Matthew Schroeder (2011) "Big data: large-scale historical infrastructure from the Minnesota Population Center." Historical Methods: A Journal of Quantitative and Interdisciplinary History **44** (2): 61–8.

Succi, Sauro, and Peter V. Coveney (2019) "Big data: The end of the scientific method?" Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **377** (2142): 20180145, https://doi.org/10.1098/rsta.2018.0145.

Szołtysek, Mikołaj (2015) Rethinking East-Central Europe: Family Systems and Co-Residence in the Polish-Lithuanian Commonwealth (2 vols). Bern: Peter Lang.

—— (2023) "Categories and contexts: the fiddly concept of 'household' and the Ukrainian family pattern in a comparative perspective." Presentation delivered to 5th Conference of The European Society of Historical Demography Radboud University Nijmegen, https://doi.org/10.13140/RG.2.2.29312.40960

Szołtysek, Mikołaj, and Siegfried 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): 101–40.

—— (2016) "Mosaic: recovering surviving census records and reconstructing the familial history of Europe." History of the Family **21** (1): 38–60.

Szołtysek, Mikołaj, and Radosław Poniat (2018) "Historical family systems and contemporary developmental outcomes: What is to be gained from the historical census microdata revolution?" The History of the Family **23** (3): 466–92.

Szołtysek, Mikołaj, Radosław Poniat, and Siegfried Gruber (2018) "Age heaping patterns in Mosaic data." Historical Methods: A Journal of Quantitative and Interdisciplinary History **51** (1): 13–38.

Szołtysek, Mikołaj, and Bartosz Ogórek (2020) "How many household formation systems were there in historic Europe? A view across 256 regions using partitioning clustering methods." Historical Methods: A Journal of Quantitative and Interdisciplinary History **53** (1): 53–76.

Szołtysek, Mikołaj, Bartosz Ogórek, Radosław Poniat, and Siegfried Gruber (2020) "Making a place for space: a demographic spatial perspective on living arrangements among the elderly in historical Europe." European Journal of Population **36**: 85–117.

Szołtysek, Mikołaj, Bartosz Ogórek, and Siegfried Gruber (2021) "Global and local correlations of Hajnal's household formation markers in historical Europe: A cautionary tale." Population Studies **75** (1): 67–89.

Szołtysek, Mikołaj, Francisco J. Beltrán Tapia, Bartosz Ogórek, and Siegfried Gruber (2022) "Family patriarchy and child sex ratios in historical Europe." The History of the Family **27** (4): 702–35.

Szołtysek, Mikołaj, Bartosz Ogórek, Siegfried Gruber, and Francisco J. Beltrán Tapia (2022) "Inferring 'missing girls' from child sex ratios in historical census data." Historical Methods: A Journal of Quantitative and Interdisciplinary History **55** (2): 98–121.

Szołtysek, Mikołaj, Siegfried Gruber, and Bartosz Ogórek (2022) "World-wide patterns of family patriarchy: Putting European uniqueness on trial." Paper presented at the European Society of Historical Demography conference, Madrid, March 3.

Szreter, Simon, Hania Sholkamy, and Aruna Dharmalingam (eds.) (2004) Categories and Contexts: Anthropological and Historical Studies in Critical Demography. Oxford: Oxford University Press.

Therborn, Goran (2004) Between Sex and Power: Family in the World, 1900–2000. Routledge: London, New York.

Tobler, Waldo R. (1970) "A computer movie simulating urban growth in the Detroit region." Economic Geography **46**: 234–40.

Todd, Emmanual (1985) The Explanation of Ideology. Family Structures and Social Systems. Oxford: Blackwell.

Todorova, Maria (1996) "Situating the family of Bulgaria within the European pattern." History of the Family **1**: 443–59.

Tsuya, Noriko O., Wang Feng, George Alter, and James Z. Lee (2010) Prudence and Pressure, 1700–1900. Cambridge/London: The MIT Press.

Viazzo, Pier P. (2003) "What's so special about the Mediterranean? Thirty years of research on household and family in Italy." Continuity and Change **18** (1): 111–37.

VDEFH [Vienna Database on European Family History] (1998) "Vienna Database on European Family History." Historical Social Research/Historische Sozialforschung 23 (4): 113–21.

Wall, Richard (1995) "Historical development of the household in Europe," in Evert van Imhoff, Anton Kuijsten, Pieter C. Hooimeijer, and Leo J.C. van Wissen (eds.) Household Demography and Household Modeling. New York: Plenum Press: 19–52.

—— (1998) "Characteristics of European family and household systems." Historical Social Research **23**: 44–66.

—— (2001) "Transformation of the European family across the centuries," in Richard Wall, Tamara K. Hareven, Josef Ehmer, and Markus Cerman (eds.) Family History Revisited. Comparative Perspectives. Newark: University of Delaware Press: 217–41.

Wall, Richard, Tamara K. Hareven, Josef Ehmer, and Markus Cerman (eds.) (2001) Family History Revisited. Comparative Perspectives. Newark: University of Delaware Press.

Wawro, Gregory J., and Ira Katznelson (2022) Time Counts: Quantitative Analysis for Historical Social Science. Princeton, NJ: Princeton University Press.

Watkins, Susan C. (1980) "On measuring transitions and turning points." Historical Methods: A Journal of Quantitative and Interdisciplinary History **13** (3): 181–86.

Wilson, Margaret F.J., Brian O'Connell, Colin Brown, Janine C. Guinan, and Anthony J. Grehan (2007) "Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope." Marine Geodesy **30** (1–2): 3–35.

Mikołaj Szołtysek is Professor of Sociology at the Cardinal Wyszynski University in Warsaw. He is interested in the global family variation, especially in historical Eurasia. He is a co-founder of "Mosaic," the big data infrastructure project for researching historical family structures in Europe. His recent research focuses on cross-cultural variation in family patriarchy and the long-term effects of the historical family on developmental inequalities. His publications include "Rethinking East Central Europe: Family Systems and Co-Residence in the Polish-Lithuanian Commonwealth" (2015).

Bartosz Ogórek is a postdoctoral researcher at the Institute of History of the Polish Academy of Sciences (Warsaw). His interests lie at the intersection of social and quantitative history. A long-time collaborator of the Mosaic Project, he is currently working on his own project reconstructing the position of different Polish population groups in the historical decline of fertility in Europe. His most recent publications include *Inferring "missing girls" from child sex ratios in historical census data*. Historical Methods, 55(2), 98–121 (2022; with M. Szoltysek et al.).

Siegfried Gruber is a researcher at the section of Southeast European History and Anthropology, Institute of History, University of Graz, where he received his doctoral degree in 2004. His main research topics are historical demography, ageing, family history, patriarchal structures within Southeastern Europe, and European comparative studies. His recent publications include Siegfried Gruber et al., eds. (2020) From the Highlands to Hollywood. Multidisciplinary Perspectives on Southeastern Europe. Festschrift for Karl Kaser and SEEHA (Studies on South East Europe 25). Wien, Zürich.

Radosław Poniat, PhD, assistant professor at the University of Bialystok (Poland). Research area: historical demography, environmental and economic history. Recent publications: Izdebski, A., Guzowski, P., Poniat, R. et al. (2022) Paleoecological data indicates land-use changes across Europe linked to spatial heterogeneity in mortality during the Black Death pandemic. *Nat Ecol Evol* **6**, 297–306; Związek, T., Guzowski, P. Poniat, R. et al. (2022) On the economic impact of droughts in central Europe: the decade from 1531 to 1540 from the Polish perspective. *Climate of the Past* **18**, 1541–61.

Cite this article: Szołtysek, Mikołaj, Bartosz Ogórek, Siegfried Gruber, and Radosław Poniat (2025) "Mosaic Database: Consolidation, Innovation, and Challenges in the Comparative Family Demography of Historical Europe," *Social Science History* **49**:254–289. doi:10.1017/ssh.2024.18