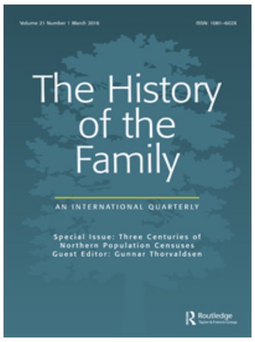
The History of the Family
AN INTERNATIONAL QUARTERLY
Special Issue: Three Centuries of Northern Population Censuses
Guest Editor: Gunnar Thorvaldsen

The History of the Family

Routledge
Taylor & Francis Group

ISSN: 1081-602X (Print) 1873-5398 (Online) Journal homepage: https://www.tandfonline.com/loi/rhof20

Mosaic: recovering surviving census records and reconstructing the familial history of Europe

Mikołaj Szołtysek & Siegfried Gruber

The History of the Family, 2016
Vol. 21, No. 1, 38–60, http://dx.doi.org/10.1080/1081602X.2015.1006655

Mosaic: recovering surviving census records and reconstructing the familial history of Europe

Mikołaj Szołtysek[a]* and Siegfried Gruber[b]

[a]*Department of Resilience and Transformation in Eurasia, Max Planck Institute for Social Anthropology, Halle (Saale), Germany;* [b]*Department of History, University of Graz, Graz, Austria*

(*Received 23 October 2014; accepted 8 January 2015*)

In recent years, there has been a marked increase in the demand for global data on historical family systems, both in the social sciences and in the humanities. Until lately, however, scholars interested in historical global family variation had to rely on simplified and often ahistorical world-scale classifications of family systems by world geographic regions. This article communicates Mosaic to the scholarly community – one of the largest infrastructural projects in the history of historical demography and family sociology. The article provides a brief history of the project, a discussion of the main issues involved in creating the database (including sampling and representativeness), and Mosaic's data structure and coverage. In the remainder of the article, the authors provide an overview of methodological and research opportunities that the project can offer to scholars, showing how the most pertinent problems of historical family demography can be tackled in more systematic ways than previously.

Keywords: microdata; household and family systems; regional demography

1. Introduction and rationale

In recent years, there has been a marked increase in the demand for global data on historical family systems, both in the social sciences and in the humanities (Dennison & Ogilvie, 2013; Duranton, Rodríguez-Pose, & Sandall, 2009; Klüsener, Szołtysek, & Goldstein, 2012; Rijpma & Carmichael, 2013; Ruggles, 2010; Therborn, 2004; Todd, 2011). Until lately, however, scholars interested in large-scale patterns of family variation in the past had to rely on simplified and often ahistorical world-scale classifications of family systems by world geographic regions (Kok, 2009; Rijpma & Carmichael, 2013; Therborn, 2004; Todd, 1985), or were forced to use 'meta-databases', which combined dozens of individual studies into one common scheme (Dennison & Ogilvie, 2013; Kaser, 2002; Wall, 2001). The Integrated Public Use Microdata Series (IPUMS) and the North Atlantic Population Project (NAPP),[1] the largest systematic census microdata efforts to date, are either confined to the populations of the North Atlantic region or cover mainly the twentieth century (Ruggles, Roberts, Sarkar, & Sobek, 2011). Early infrastructure efforts, which started in the 1970s at the Cambridge Group for the History of Population and Social Structure and in Vienna, are no longer able to fill in this vacuum because they relate to only some parts of Europe, or for reasons of data accessibility.[2]

According to conventional wisdom, major hindrances to furthering the development of 'global' accounts of the family history of Europe have stemmed from constraints on data availability. For example, while the current large-scale initiatives like IPUMS International or NAPP are based on existing complete census material for entire

*Corresponding author. Email: szoltysek@eth.mpg.de countries, for many parts of historical Europe such complete census material is extremely difficult to obtain or non-existent. This might foster a tacit assumption that the materials of relevance to the study of family and co-residence in the more remote past are scarce, or that acquiring such data would be extremely difficult. Because the historical listings have not survived or never existed, or their exploration is risky, it could appear that some aspects of the regional patterning of social and family structures in the historical landmass of Europe will never be fully explored.

Over the course of the past several years, however, the assumption that future scholars will need to conduct their research ‘below the data poverty line’ has been shattered (see Ruggles, 2012), as has the ‘image of limited good’, whereby the most desired items (i.e. the surviving and accessible population listings from historical Europe) exist in a finite quantity and are always in short supply. Whereas a census microdata initiative that covers the entire territory of historical Europe may never be fully realized, the use of scientific harmonized samples of census microdata comparable across time and space has already begun to revolutionize historical demography. This article elucidates the importance of the current disciplinary moment by communicating Mosaic to the scholarly community – one of the largest infrastructural projects in the history of historical demography and family sociology. The article provides a brief history of the project, a discussion of the main issues involved in creating the database (including sampling and representativeness), and Mosaic’s data structure and coverage (with plans for its future expansion). In the second part of the article, we provide an overview of the methodological and research opportunities that the project can offer to scholars, pointing out the ways in which the most pertinent research questions of historical family demography might be answered in more systematic ways than previously.

2. Starting Mosaic

The Mosaic project originated in the research and data-infrastructure activities of the Laboratory of Historical Demography at the Max Planck Institute for Demographic Research (MPIDR) in Rostock, Germany. Since 2008, the Laboratory has been involved in several large data-transcription projects, as well as in depositing and inventorying various microdata collections from historical eastern Europe. At the same time, comparative analysis of family systems has been actively pursued by the Laboratory’s team members. The ultimate stimulus came from Joshua R. Goldstein, the then director of the MPIDR, who has since taken the lead in an effort to create a database for historical census and census-like material for the whole of continental Europe and beyond.

In May 2011, the Mosaic project was launched with a conference on Reconstructing the Population History of Continental Europe by Recovering Surviving Census Records, held at the MPIDR in Rostock. The 2011 conference helped to establish and cement the international collaboration necessary to enforce the new agenda, and to promote the idea of the Mosaic project among wider scholarly circles. It harnessed the energies of a large number of historians, demographers and archivists, who jointly committed themselves to the shared purpose of recovering surviving census records of historical Europe (see Figure 1).

The Mosaic project was funded by the MPIDR until mid 2013, when the then director, Joshua R. Goldstein, left the institute to take a chair in demography at the University of California, Berkeley. Currently, the project is headed by a managing board consisting of Siegfried Gruber (University of Graz), Mikołaj Szołtysek (Max Planck Institute for Social Anthropology, Halle/Saale), Joshua R. Goldstein (University of California, Berkeley), Kees Mandemakers (International Institute of Social History, Amsterdam), Péter Őri

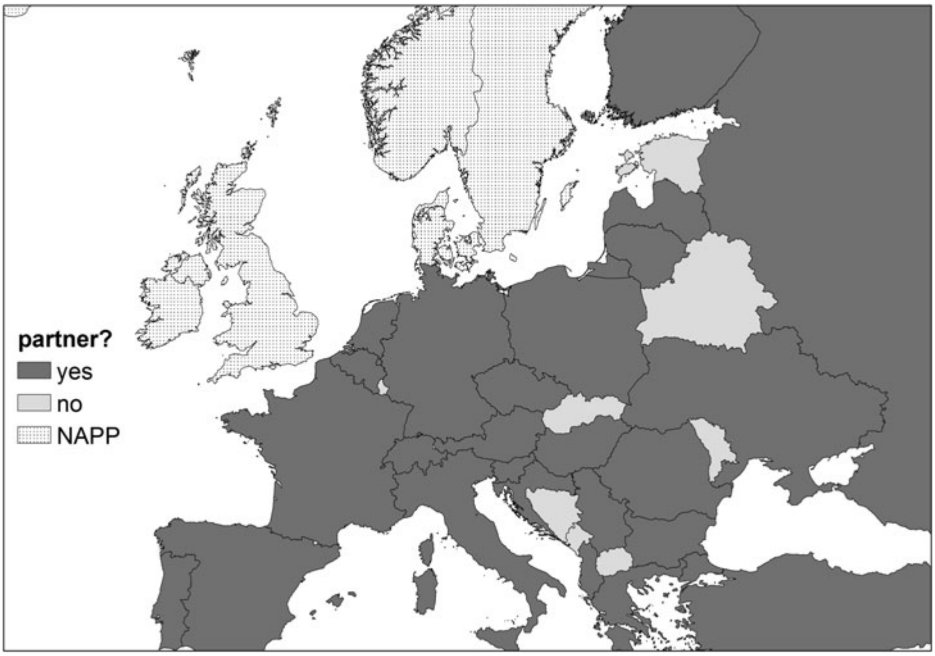

Figure 1. Current Mosaic partnerships. Source: Mosaic database. Design: S. Gruber.

(Demographic Research Institute, Budapest) and Steven Ruggles (Minnesota Population Center, Minneapolis).

In the years following the launch conference, the work continued along several primary paths – i.e. the identification, digitization, transcription, coding and harmonization of surviving census or census-like materials from different parts of Europe.

3. Country inventories and recovering surviving census records

Data for the Mosaic project was derived from two main sources: either existing data was donated or new data was identified and transcribed using the MPIDR's resources. The original database for the Mosaic collection consisted mainly of donated files for Albania, Mecklenburg-Schwerin (including Rostock) and part of Poland-Lithuania, and the Vienna database. Consequently, the Albanian and Polish-Lithuanian projects (the latter in its wider extension) became the 'prototype' for further Mosaic-type data sets, both in terms of database structure and with regard to the particular research framework they were embedded in (see Kaser, Gruber, Kera, & Pandelejmoni, 2011; Szołtysek, 2012, 2014).

Next, the search for already existing machine-readable data which would fit into the Mosaic format (see below) was undertaken. It targeted not only historians or historical demographers, but also genealogists and their respective associations, especially in Germany, but also elsewhere.

Second, we developed a long-term strategy for searching for surviving census and census-like microdata in various parts of Europe. In each case, the first step was the creation of an inventory of local surviving census and census-like material for a given historical region, an administrative area or the whole country (either within its contemporary or historical boundaries). This was necessary because the majority of countries under consideration lacked any sort of overview of such material, or knowledge of it had long since been forgotten, while most historians and historical demographers continued mistakenly to assume that only a few manuscripts survived.

The inventories were made by local historians visiting and communicating with local archives, backed up by back-and-forth consultations with the Mosaic core team. In some cases, it turned out to be a quite straightforward, if still very laborious, task.[3] However, in most countries, data identification was a complicated and thorny process because the archival materials of interest were not centrally stored, but scattered in many regional or even local archives. Altogether, such inventories have been made for Austria, Catalonia, Germany, Hungary, Lithuania (including parts of Belarus), the Netherlands, Poland, Romania, Serbia, Slovakia and Western Ukraine (see Figure 2; most of them are available for download as part of the Mosaic Working Paper series at http://censusmosaic.org/web/publications). They demonstrate that, even though enumeration forms for national censuses taken before the early twentieth century were preserved to drastically varying degrees (from large proportions to almost nothing), in virtually every country, census manuscripts or various individual-level census-like listings (church lists of parishioners, tax lists, local estate inventories) survived in great quantities.

Inventorying the surviving census and census-like records from different cultural milieus in Europe was seen as crucial to sustaining not only their long-term access, management and dissemination across the worldwide scientific community, but also to safeguarding their preservation. Many historical data sets (especially in eastern Europe) exist only as manuscripts, deteriorating Xerox copies or disordered scans. If they were to be properly identified, preserved and computerized, the risk of a large part of the world's historical heritage literally going to waste would be significantly diminished. In addition to recording detailed information on the European ancestry of past centuries, the historical

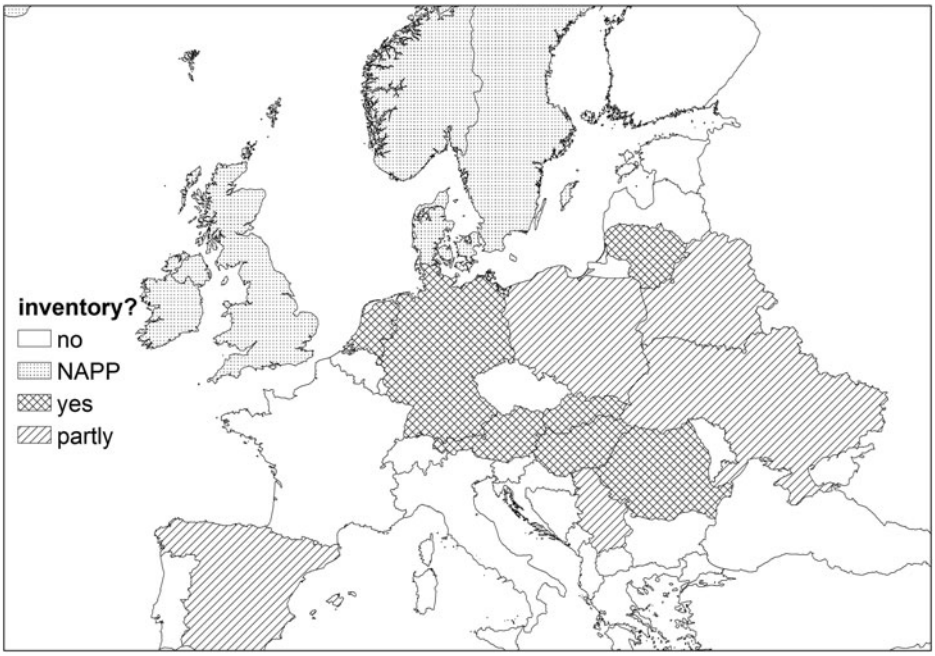

Figure 2. Countries inventoried within the Mosaic project for the availability of historical census and census-like materials. Source: Mosaic database. Design: S. Gruber.

listings at stake recount important social practices carried out in various spatial and cultural contexts, thus forming a vital part of Europe's joint historical and sociocultural heritage.

Furthermore, the inventories provide necessary regional contextual information about the census-taking practices which led to the production of a given material, often including discussion of the categories and indicators employed during the information-gathering. In the future, standardized accounts of which data sets are comparable to each other regarding certain specific research topics should be incorporated into the existing body of inventories.

4. Sampling and representativeness

Finally, country-specific inventories of surviving census microdata proved indispensable in providing selection frames for strategically digitizing relatively small numbers of records which would be useful for research, and which would make their way into the final database upon transcription.

One of the unique points of Mosaic when juxtaposed with existing large-scale data-infrastructure efforts is that it relaxes the requirement of full-count census data. In order for the project to foster expansion of historical demographic knowledge beyond the well-known confines, it had to deal not only with full-count data or representative samples of surviving complete census material (like in NAPP), but also with incomplete censuses or census-like materials. This, in fact, applied to the majority of European countries and regions for which no complete historical census coverage survived or ever existed, and which were therefore omitted from previous projects.

This very feature of the Mosaic data introduces some vital concerns about the data representativeness. Before full advantage is to be made of the new collection by its prospective users, a solid idea of the representativeness of the data has to be clearly spelled out.

Before the foray of full-count census data in historical demographic research associated with the IPUMS and NAPP projects, the problem of data representativity has made only a rare, if not whimsical, appearance in family history literature. In the past, it was customary for scholars to examine the material or populations they had chosen to study by deliberately picking up on what appeared to be the most 'representative' cases. For Le Play – the doyen of family sociology and an early advocate of the typologization of European family systems – it was sufficient to seek out what he considered a 'typical' or 'representative' family of a particular region (and some of his 'regions' could spread over thousands of kilometres) by inspecting selected cases believed to prove a particular manifestation of a given social trait, either in its 'average' or 'extreme' forms (Zimmerman & Frampton, 1935, quoted in Kruskal & Mosteller, 1979b, pp. 253–254). But Le Play could not prove that his families were representative of the groups which he studied (see Le Play, 1877–1879).

No more scrutinizing in this respect was the approach of the Cambridge 'School' of historical family sociology. At no point does Laslett's discussion of the well-known 'English sample of 64 settlements' used for the comparative study of domestic groups feature a proper exploration of the sampling scheme or representativeness of the entire collection. Referring to the 64 listings included in the collection, which dated from 1574 to 1821 and stretched from Devonshire to Durham and Westmorland to Kent, Laslett wrote:

> It would be idle to expect 64 settlements to be representative of the whole of England, or to be effective for indicating change over time. Nevertheless, 25 of the 40 counties of the country appear with one or more settlements, and only 1 county … can be said to occur much too frequently for its relative importance. (Laslett, Wachter, & Laslett, 1978, pp. 73–74; cf. also Laslett, 1969)

Referring to the 'one hundred most informative listings' used for the analysis of the mean household size, Laslett (1969) wrote: 'It was decided in 1967 to analyze the hundred most informative listings; where there was a choice some respect was paid to geographical distribution, but for the most part selection had to be arbitrary as regards time and place' (p. 199).

Mosaic provides a more thoughtful consideration of these issues not only by making imperfect samples more readily comparable, but also by scrutinizing more closely what particular data sets stand for, while making statistical inferences from them. Table 1 shows the basic characteristics of various Mosaic samples, providing information about their size, their creator and their release date. Table 2 is more detailed, classifying the countries and regions contained in Mosaic by the proportion of the surviving manuscript material,

Table 1. Mosaic country and region samples: basic characteristics.

Country/region	Year	Creator	Sample size	Release date
Albania	1918	Project team: Karl Kaser, Siegfried Gruber, Gentiana Kera and Enriketa Pandelejmoni	140,612	2015–2016
Austria	1910	Peter Teibenbacher	20,036	2014
Belgium (western Flanders)	1814	Familiekunde Vlaanderen	13,666	2014
Catalonia	1880–1890	Centre d'Estudis Demogràfics	23,997	2015–2016
Courland (Latvia)	1797	MPIDR	37,104	2015–2016
France	1846	Rolf Gehrmann	16,967	2014
Germany	1846	Rolf Gehrmann	30,751	2014
Mecklenburg-Schwerin	1867	Research project: State Main Archive Schwerin, MPIDR and Department of Multimedia and Data Processing, University of Rostock	75,577	2015–2016
Schleswig and Holstein	1803	Danish Data Archive	217,951	2015–2016
Hungary/Slovakia/Transylvania	1869	Péter Őri, Levente Pakot	31,406	2014
Lithuania (including some parts of Belarus)	1847	Dalia Leinarte	19,917	2015–2016
Grand Duchy of Lithuania (the Jewish census)	1764	Center for Studies of the Culture and History of East European Jews	21,370	2015
Netherlands (Zeeland and Northern Brabant)	1810–1811	Kees Mandemakers	40,130	2015–2016
Poland-Lithuania	1768–1804	Mikołaj Szołtysek	155,818	2015
Russia (Moscow region)	1897	Irina Troitskaya	11,559	2015–2016
Ukraine (eastern Ukraine, Cossack region)	1765	Yuriy Voloshyn	18,395	2015
Ukraine (western Ukraine)	1863	Piotr Guzowski	9836	2015–2016
Wallachia (Romania)	1838	Bogdan Mateescu	21,546	2015
Overall			904,500	

Source: Mosaic database.

the kind of sampling applied and the assumed representativeness of the sample (representative of the country, representative of the region or representative of particular social strata of the region). In addition, it gives information about whether urban populations are included in the samples or not.

To this end, based on Table 2 , we distinguish between several major groups of country or region samples within Mosaic:

1. Country samples based on random sampling: these samples are based on complete or almost complete coverage of the census area, in most cases covering the entire territory of the respective country, either within its historical or contemporary boundaries. These samples are representative of the census area. The strata for sampling are generally administrative or geographic areas within a country. One example of such a sample is the Albanian census of 1918, where most of the material has been preserved and a random sampling based on seven regional strata could be applied. In addition, the complete urban and minority populations were added to the sample. Another example is a sample of the 1838 census of the Principality of Wallachia (fourteenth century–1859), which covers the southern part of present-day Romania. It is a representative sample of the rural population based on four regional strata (east, north, south and south-west).
2. Country or region samples based on known criteria: these samples are based either on low shares of surviving material or sampling schemes which were restricted by financial resources or the accessibility of manuscripts. Sampling should ensure regional coverage or different ecotypes/legal statuses of the peasant population. We can assume that these samples are somehow representative of the sampled regions, but we cannot be sure about it. The data of the 1867 census of Mecklenburg-Schwerin, for instance, is based on clusters of villages and cities covering different regions and different legal statuses of the peasant population.
3. Country or region samples based on other sampling schemes: these country samples were, for the most part, donated to Mosaic, and sampling was based on criteria other than representativeness. Ultimately, these country samples could cover the whole country because all of the manuscript material is available and the data transcription is still ongoing. The censuses of western Flanders in 1814 and of Schleswig and Holstein in 1803 have been transcribed by genealogists, and part of this material is already available for research.
4. Country or region samples based on complete surviving material: as no other material has survived, the representativeness of these samples cannot be directly verified, but can only be gauged through an informed speculation based on a thorough knowledge of the respective local or regional conditions and socio-economic or environmental properties. The Central European Family Forms Database (CEURFAMFORM), covering the peasant population of historical Poland-Lithuania, is such a country sample (see Szołtysek, 2014, ch. 2), as well as the data set for western Ukraine in 1863.

While in some cases the relative scarcity of primary materials fostered an inclusive approach (category 4 above – i.e. all of the primary materials for a particular area found in the archives or in printed editions were included in the database), in the main, sampling schemes were necessary either due to convenience (there was too much data to work on and process by a single team) or financial constraints (see categories 1–3). Sampling has generally been done by complete settlements (villages), with subdivisions of countries (or regions) used as strata for sampling, but in each case it heavily depended on the extent of

Table 2. Mosaic country and region samples: sampling and representativeness.

Country/Region	Year	proportion surviving material	random sampling	no. of strata	sampling based on known criteria	other sampling	all surviving material included in sample	assumed representativeness	urban population included	remarks
Albania	1918	83%	X	7				population of census area	yes	oversampling of urban and minority population
Austria	1910	9%			different ecotypes			rural population of Western Tyrol/ sample regions	yes	
Belgium (Western Flanders)	1814	complete				X		southeast of Western Flanders	yes	
Catalonia	1880–1890	< 10 %	X	2				Catalonia	yes	
Courland (Latvia)	1797	85%	X	15				rural Courland	no	
France	1846	53%			regional coverage			rural France	no	
Germany	1846	< 1 %			regional coverage			German Customs Union	yes	
Mecklenburg-Schwerin	1867	complete			different regions and legal status			sample regions	yes	
Schleswig and Holstein	1803	complete				X		population of Schleswig and Holstein	yes	early releases not representative, finally complete coverage
Hungary/Slovakia/ Transylvania	1869	15%	X	9				rural population	no	in some regions all surviving material had to be used
Lithuania (incl. some parts of Belarus):	1847									only peasant population on estates
Kovno gub.		1.5 %					X	rural population	no	
Vilnius gub.		24%	X	1				rural population	no	

Grand Duchy of Lithuania	1764	complete	X	5			Jewish population	yes	only manuscripts of best quality transcribed
Netherlands: Zeeland	1810/11	60%			completeness		population of Zeeland	yes	
Northern Brabant						X	population of North Brabant	yes	
Poland–Lithuania	1768–1804	very low				X	peasant population of Poland–Lithuania	no	only peasant population
Russia (Moscow region)	1897	unknown			regional coverage		rural population of Moscow region	no	
Ukraine (Cossack state)	1765	complete	X	10			rural population of Cossack region	no	
Ukraine (Western Ukraine)	1863	low				X	rural population of the region	no	only Catholics
Wallachia (Romania)	1838	72%	X	4			rural population	no	

Source: Mosaic database.

the data coverage of the surviving material. Sample sizes vary (Table 1), depending primarily on the expense of primary data collection and the financial resources available. Some of the country samples have already been released, and most of the others are scheduled for release within the next two years.

Since, in the majority of the surveyed countries, the surviving material was not distributed evenly, truly representative (in statistical terms) country samples could be obtained only for a few countries. The Mosaic samples are therefore generally more representative of the surviving census material than of the population of the concerned region or country, but they remain the best samples for the concerned regions or countries available for now, and for some areas better samples may never be obtained (for example, Poland-Lithuania). In order to better indicate the differential status of the Mosaic data, a column called 'assumed representativeness' was added to Table 2, which gives information about the different levels of representativeness we assume for particular populations.

Despite this shortcoming, the data gathered in the Mosaic database seems better safeguarded than that used in pre-Mosaic family history research. One of the definitions of 'representative sample' is the absence of selective forces (Kruskal & Mosteller, 1979a). In this sense Mosaic decidedly departs from earlier studies in which cases presented and popularized in literature were hardly ever discussed in terms of how they should be dealt with as regards representative or selection bias; they were simply either the most appealing to the scholarly imagination (for various reasons) or the easiest to get (see, for example, Laslett, 1983; for Czap's Mishino estate data, see Czap, 1983). In this respect, Mosaic has no selective forces involved, except for the 'random' forces of history which delimited the very survival of the material.[4] Some areas might have been prioritized for archival scrutiny in the early phase of the project, but this has since changed.

Another feature of 'representative sampling' that is not met in our case is that the majority of the Mosaic samples (except for those in category 1 above) cannot be ascertained to have the same distribution as the respective country population from which they were drawn – in other words, it cannot be asserted that they are made up of typical units of that general population (Kruskal & Mosteller, 1979b). Although it is not unlikely that several more sizeable 'samples' in fact represented the average tendencies of the much bigger spatial entities to which they belonged, it is not possible to know exactly to what extent this is true because more encompassing data collections do not exist for those areas. At present, there is simply no way of knowing how representative of Spain the data from Catalonia is, or how representative of the whole of France the data from the 1846 census is.

Still, it would not, however, do to say that this data just happened to come to hand and we have no notion whatsoever of the relation between what would statistically be called the 'target' population and the 'sample' population. First, this data may still be representative of certain areas or administrative regions, some of which are quite large (for example, Schleswig and Holstein or Catalonia). Whereas it might be difficult to argue that the collection of data from Münsterland may be representative of the entire German population of that time (see Table 3), much better grounds exist to assume that it exhibits on a smaller scale the relevant familial characteristics of the population of the Prince-Bishopric of Münster (a large administrative area in the northern part of today's German states of North Rhine-Westphalia and western Lower Saxony) from which it came. Furthermore, some of the regions included in the collection have data coverage which successfully represents the given region's – or even country's – population and socio-economic heterogeneity (for example, Poland-Lithuania and the Kingdom of Hungary), which tallies well with the generally acclaimed sense of representativeness (Kruskal & Mosteller, 1979a).

Table 3. Current data coverage of Mosaic.

Census	Country/region	Person records (N unweighted)
Albania, 1918 census		140,611
Austria-Hungary, 1869 census	Hungary/Slovakia/Transylvania	31,406
Austria-Hungary, 1910 census	Austria	20,036
Belgium, 1814 census	Western Flanders	13,666
Bulgaria, 1877–1947 household registers	Rhodope region	8373
Denmark, 1803 census (German territories)	Schleswig and Holstein	107,861
France, 1846 census		16,967
France, 1831–1901 census	South-western France	5109
German Customs Union, 1846 census		36,760
German Customs Union, 1858 census	Government Districts of Danzig and Posen	3468
German Customs Union, 1861 census	Government District of Sigmaringen	6541
Germany, 1900 census	Rostock (city)	55,705
Prince-Bishopric of Münster, around 1700 status animarum		23,010
Prince-Bishopric of Münster, 1749–1750 status animarum		34,169
Bishopric of Constance, 1749–1811 status animarum		2480
Italy, 1430 enumeration	Legnano	2101
Poland-Lithuania, 1768–1804 listings		155,818
Wallachia, 1838 census		21,546
Russia, 1795 revision lists	Braclav Governorate	8050
Russia, 1814 private enumeration	Estates of the Gagarin family (among others, Mishino)	2955
Russia, 1897 census	Moscow Governorate	11,559
Serbia, 1863 census	District of Jasenica	7128
Serbia, 1884 census	District of Jasenica	9434
Istanbul, 1885 census		3408
Istanbul, 1907 census		4946
Overall		733,107

Source: Mosaic database.

In addition to country or regional samples, Mosaic also includes a large collection of data of single villages (or cities) or regional clusters of villages (for example, Istanbul or Italy and some locations from European Russia – see Table 3). This data is usually the outcome of various research projects of different scholars which have been donated to Mosaic, most of them in the form of case studies. The caveat of unrepresentativeness is particularly relevant in this regard, as what we have for those areas is so unlikely to represent the grander ‘whole’ fairly that every small addition of empirical data can be expected to change the patterns revealed by this data rather dramatically. Special caution needs to be exercised when using this data in a larger context. Nevertheless, this data might be valuable for some comparative research exercise, and Mosaic provides a portal for making it available to the international scientific community.

5. Mosaic’s data structure

The crucial component for the success of Mosaic is the integrated character of the transcultural and cross-temporal data it contains. There are three primary requirements for a data file to fulfil in order to be included in the Mosaic project: (1) the data source should list individual persons, preferably by name; (2) the data source should list all individuals in a settlement or an area, not just the household heads, men or adults; and (3) the data source should list all individuals by clearly delineated residence units (houses, hearths, domestic groups or households). The file cannot be included in the Mosaic project unless the primary sources provide important details either explicitly or implicitly – i.e. the individual's age, sex, relationship to the household head, marital status, place and year of enumeration, and first and last names for the whole population.

Sex might only be available as a variable derived from first names or the relationship to the head of the household. Marital status might be accessible only in the case of married people and possibly widowed people. Age should be available for the whole population, not only for adults or children. Occasionally, though, missing ages do not pose a problem. The relationship to the head of the household is essential for analyzing co-residence patterns, and thus can only be missing for a few persons (Ruggles & Heggeness, 2008). More marginal characteristics, such as occupational title or status, literacy, religion and language, are of interest too, but are not obligatory for the source to be included in the Mosaic framework. We are aiming at pre-1900 data from all over Europe: state censuses, church listings (for example, lists of souls), local enumerations, tax lists, etc. Longitudinal material – i.e. referring to periods of time (for example, population registers) or vital events (for example, church books) – remains outside the scope of the Mosaic project, in which preference is given to cross-sectional data (cf. Dillon & Roberts, 2002).

The starting point for getting all of the material properly organized – in line with some of the best practices for handling historical nominal sources for the purposes of creating large databases (Kelly Hall, McCaa, & Thorvaldsen, 2000; Mandemakers & Dillon, 2004; Thorvaldsen, 1994) – was the full transcription of original manuscripts or published sources. Information on all individuals from selected listings was entered into a computerized database according to the domestic groups they inhabited. The process progressed with full conformity to the original manuscripts, retaining the spellings for first and last names, occupational terms and interpersonal relationships, as well as the original order of appearance of individuals within a residential group. Thus transcribed, the original data was then subjected to standardization and, subsequently, to various coding decisions (cf. Mandemakers & Dillon, 2004, p. 36).[5]

The task of transcribing the census-like material was undertaken by student helpers in Rostock or was commissioned to partners within or outside Germany (several partner teams responsible for specific country inventories subsequently became involved in the data transcription for their particular country-specific 'samples'). Harmonization processes (also known as data 'integration'), on the other hand, were conducted mostly by student helpers in Rostock. This was also the case for many of the small databases donated to the project, which had often been designed according to various research needs and varied database practices.[6] The generous support of the MPIDR made it possible to employ up to 2 research assistants and 24 student helpers at its peak, in addition to the core members of the Laboratory of Historical Demography. On a few occasions, however, it proved much more convenient for harmonization tasks to be performed by the local researchers who were responsible for data collection, based on careful instructions received from Mosaic's core team.

Each harmonized Mosaic data file contains 30 variables, which refer to three different levels of information: 4 variables defining the data set, 8 variables defining the household (such as household size and group-quarter status), and 18 variables defining the person (for example, age, sex, marital status, occupation and relationship to the head of the household). In general, most of the variables are designed according to their IPUMS International and/or NAPP counterparts, and can therefore be readily used in comparative research with data from these databases. Occupational titles are coded into the OCCHISCO codes also used in NAPP. Those who are interested may wish to consult the documentation file of all variables with all values, available at the Mosaic website (http://www.censusmosaic.org). First and last names, relationship terms and occupational titles are available in a non-harmonized form, unless the data was donated in an already harmonized way, or sometimes only in a coded version. Under no circumstances are scholars wishing to work with the Mosaic data expected to harmonize the data by themselves. Should weighting the data prove necessary, weighting variables are provided. Basic checks with regard to the completeness of the codes and the internal coherence of the data are performed before the data is released. Some of the data needed correction (for example, female sons or parents being younger than their children), and the changes that were introduced are flagged in quality flag variables.

Another component of integration is variable documentation, the aim of which is to highlight important comparability issues that are not self-evident from the coding structure for the variable. A general comparability discussion has been integrated into the documentation file on the project's website, emphasizing concerns for inter-regional comparisons, which can be called on when making intra-regional comparisons.

The harmonized Mosaic data files are distributed free of charge for scholarly purposes via the data section of the Mosaic website. Every individual wishing to download Mosaic data files has first to register as a user and then to confirm his/her acceptance of the terms (no misuse and improper citation).

Each data set comes in a zipped file, containing comma-separated values files and a readme file with the appropriate citation of the respective data file. Geographic Information System (GIS) files are also available for download. The MPIDR Population History GIS Collection supports demographic and socio-economic research by filling in the gaps in the European GIS data infrastructure on historical national and regional administrative boundaries and historical place names. Currently, there are GIS files with administrative boundaries available for Albania, Austria-Hungary, Courland, Germany, Poland and Serbia, and maps for the whole of Europe with the borders in 1900, 1930, 1960, 1990 and 2003. All of the settlements of the released Mosaic data files are geo-referenced. A table with this information will soon be available for download directly from the project's website.

6. Current data coverage

Table 3 presents a detailed list of the data sets included in the most current version of Mosaic. Table 4 shows the distribution of those regions across time and urban and rural contexts, and Figure 3 reveals the spatial patterning in the distribution of the data across Europe.

The current data covers 92 regions/locations in Europe, stretching from the Atlantic Ocean to the Ural Mountains. A slight majority of the included locations comes from the nineteenth century (56%), with more or less equal numbers covering earlier and later periods. The collection contains both rural and urban sites, although rural societies clearly predominate (82%).

As of now (see Figure 3), the data is largely concentrated in the central continental belt, providing quite a good coverage of the French, German, Austro-Hungarian, Polish and Balkan areas. However, some areas that are important for the investigation of the European geography of family systems, for example, are not yet included, or the coverage

Table 4. Mosaic data coverage by period and urban and rural populations.

Period	Rural regions	Urban regions	Rural persons	Urban persons
Until 1800	19	1	215,531	1898
1801–1850	30	5	171,796	39,553
1851–1900	15	2	73,174	59,521
1901–1947	12	8	102,194	6944
Overall	76	16	562,695	170,412

Source: Mosaic database.

of these areas in the database is very limited. These areas include the Low Countries, which are often assumed to have encompassed the essential features of 'north-west European' family systems (De Moor & Van Zanden, 2010); the Italian territories, which, according to some scholars, exemplify the 'Mediterranean' zone (Smith, 1981); and Russia. As the Mosaic database expands, some of these deficits will soon be addressed.

However, in addition to covering both urban and rural communities, the current database runs across many important fault lines in the European geography of family systems, including places located

1. eastward of the Hajnal-Mitterauer line (parts of Poland, Russia, Ukraine, Belarus, Hungary and Latvia);
2. within the south-eastern Europe zone (Albania, Serbia, Turkey, Romania and Bulgaria); and
3. in the 'intermediary central European zone' of Laslett's (1983), e.g. Austro-Hungarian and German areas, as well as parts of historical Poland and western Europe – France).

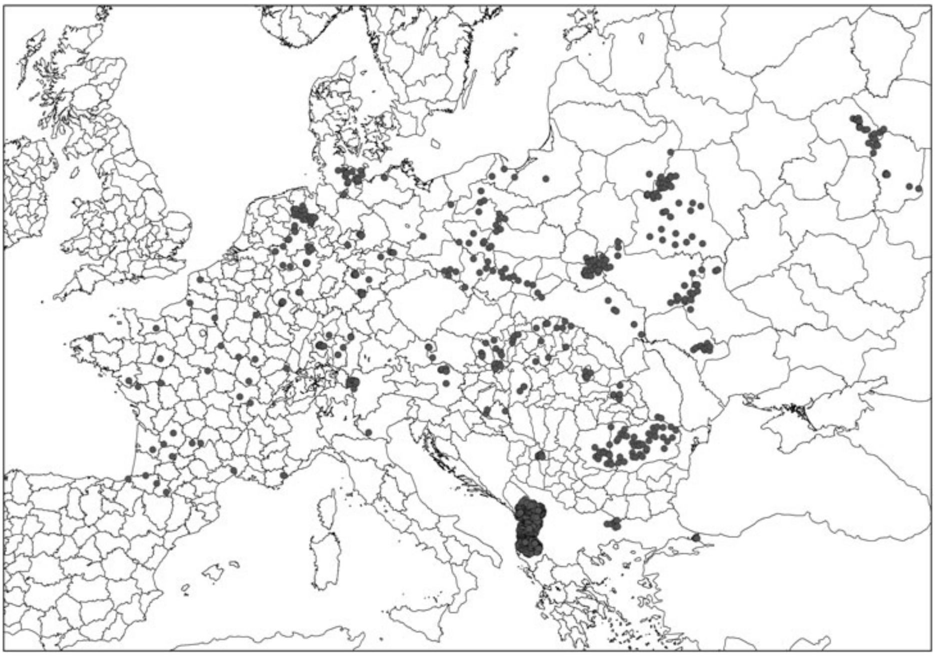

Figure 3. Spatial distribution of current Mosaic data. Source: Mosaic database. Design: S. Gruber.

The collection encompasses societies which varied significantly in terms of basic principles of family and household organization, including strictly nuclear and neolocal populations (like urban Rostock, but also southern Ukraine, the Braclav area and Podolia); stem-family societies (like those in the area of Münster in Germany, in south-western France and in parts of western Poland); complex societies exhibiting a 'classic' eastern European joint-family pattern (like those in Mishino near Moscow studied by Czap (1983) or in Polesia in eastern Poland-Lithuania) or Balkan versions of this pattern (Albania and Serbia); and a range of intermediate patterns with varying degrees of intermingling of nuclear- and stem-family organization (Poland proper, Germany in 1846 and Austria around 1910), or stem- and joint-family patterns (Red Ruthenia in Poland and fifteenth-century Italy). Furthermore, even in its present scope, the database already covers much of European variability in terms of geographical features, populations, cultures and socio-economic geography (Jordan-Bychkov & Bychkova-Jordan, 2002): i.e., plains, mountains, and coastal areas; free and unfree peasantries; a range of ethnicities and religions; and a range of regional patterns of economic growth in the early modern and modern eras.

Millions of records that have already been identified by our team and a European network of collaborators, arranged according to Mosaic's structural requirements, currently remain outside the scope of the project, although samples from them are potentially obtainable. The primary future goal of the Mosaic initiative is, thus, to expand the existing data infrastructure into further regions of Europe, and ultimately to extend the reach of comparative analysis into the landmasses of Asia. Spain, Italy and Russia need to be targeted first, as it is indispensable to include these territories if major European tendencies in family organization are to be captured by the Mosaic database (eastern Europe, the Balkans, the Mediterranean and north-west Europe).

The prospects for further expansion are particularly strong in eastern Europe, especially Russia, and in Siberia, where hundreds of digitized nineteenth-century individual-level taxation lists (so-called 'revision lists') are known to exist in the Urals and surrounding area, including also some parts of central Asia (for example, for the Astrakhan Governorate, which borders with Kazakhstan).[7]

In addition, nationally representative samples of census microdata from Canada, Great Britain, Germany, Iceland, Norway, Sweden and the USA from 1801 to 1910 can also be easily combined with the Mosaic data; initial attempts have already been made (see below). Finally, it may be possible to harmonize historical representative country samples from the expanded Mosaic collection with contemporary European microdata samples available through the Integrated European Census Microdata project.[8] All these data sets have generally comparable formats and can yield comparable basic information on co-residence patterns.

7. Research significance

Space limits our possibilities to discuss this issue at length, but it has to be stressed that apart from its major focus on the development of data infrastructure, Mosaic has primarily been a research-driven initiative, and the information it produces can become a powerful tool in transforming the historical demography of Europe. It can make substantial contributions to other research fields as well.

Despite the lack of consistency of the sampling schemes and concerns about the rigid statistical representativeness of the data, the sheer amount of data assembled and its spatial regional distribution make Mosaic capable of putting the study of historical family forms on a new footing. The vast majority of the quantitative evidence used in the debate on the geography of historical family and household composition has consisted of studies of a single community or a small group of communities (see, for example, Fauve-Chamoux & Ochiai, 2009; Laslett, 1977, 1983; Wall, 2001; cf. Ruggles, 2012), raising questions about the extent to which their results can be generalized for larger populations and for larger territorial units. Apart from that, these studies have often relied on a range of different methodologies (for example, *ad hoc* coding schemes or non-transparent operationalization of the variables), posing obvious challenges in terms of systematic data comparability. Although Mosaic is definitely not immune to the first set of flaws just mentioned, there are at least two respects in which the project might steer the study of family and residence patterns far away from its usual drawbacks.

First, Mosaic is better equipped to tackle the methodological challenges involved in studying small populations, such as the stochastic variations in demographic behaviour at a micro level, which may themselves affect the observable co-residence and other demographic patterns (Ruggles, 1987). Unless it is done with careful sampling considerations (which, as already mentioned, has hardly ever been accomplished in pre-Mosaic research on family history), focusing on local micro-populations runs the risk of being affected by what is known as the 'law of large numbers', which posits that the smaller the number of observations, the greater the problem of inaccuracy of parameter estimates due to large measurement errors.[9] This statistical principle has long been known to demographers, who have assumed that the size of a population affects the pattern of relationships between the elements of the natural movements within that population (Ruggles, 1999, pp. 126–127; Vallin, 2006, p. 6; Wachter, 1978). Using larger populations (of regions or macro-regions) usually causes random errors stemming from stochastic variation to average each other out, and thus yields more accurate and more parsimonious estimates. To us, the difference between assuming regularities based on a single study (Laslett, 1983) and inferring them by averaging a few dozen or a few hundred observations boils down to yielding incomparably more robust estimates.

The second major asset of Mosaic in this regard is that all of the microdata samples discussed above are very similar to one another in terms of their structure, organization and available information, and manifest a core set of common variables. By this token, Mosaic makes imperfect samples more readily comparable, allowing researchers to measure family systems or other concepts systematically across space. Recently, Gruber and Szołtysek (2015) capitalized on this feature and used a range of harmonized Mosaic variables related to familial behaviour – including nuptiality and age at marriage, living arrangements, post-marital residence, power relations within domestic groups, the position of the aged and the sex of the offspring – to develop a composite measure of differences in sex- and age-related social inequality in 91 regions of historical Europe, covering more than 700,000 individuals from the Atlantic to Moscow. Based on a large set of harmonized variables, this index has the potential to become the new 'master variable' for cross-cultural studies of family organization and relations.

Another significant issue is Mosaic's data coverage and its spatial distribution. Within the general aim of collecting surviving microdata from different areas of Europe, the initial focus of the project has been on gathering data from central and eastern Europe. While much data on western Europe was already available through NAPP, in countries like Poland, Lithuania, Hungary, Russia, etc., although a wealth of data has been preserved, it has barely been collected and analyzed. These countries present a particular challenge to historians and social scientists – central and eastern European societies have been long believed to have had different demographic regimes from those of western European societies (for example, different marriage and fertility patterns, more complex family systems, etc.). The fact that Mosaic contains a large amount of previously unavailable data from these under-researched areas makes the possibility of future comparative research in historical family demography especially promising.

Accordingly, one can reasonably argue that, by the very amount of material amassed, Mosaic might elevate the ongoing discussions of the geography of European family forms to new heights. Szołtysek (2014), for example, in his prototype Mosaic study, used measures of co-residence and marriage patterns to reveal that, at the end of the eighteenth century, three household and family patterns with substantial numerical and qualitative differences existed in the Polish-Lithuanian territories, thus challenging the notion of a homogenous family system in what used to be one of the largest regions of east-central Europe. Similarly, Őri and Pakot (2014), working on evidence from the Hungarian Mosaic sample, revealed a great deal of patchiness in the patterns of marriage and household formation across pre-industrial Hungary, Slovakia and Transylvania, which called for a move beyond the stereotypical and artificial divisions of Europe into 'western' and 'eastern'. Pre-industrial Germany, too, has been discovered to have been only slightly less variegated (Szołtysek & Gruber, 2012; Szołtysek, Gruber, Klüsener, & Goldstein, 2014).

The outcomes of these and future investigations might call into question the homogeneity of family forms in any region of Europe, thus turning it into a dubious endeavour to further promote a spatial construction of European family systems by branding major areas. At the same time, however, it may lead to compelling discoveries of broader, perhaps more complex, regularities, and thus yield refined spatial classifications of family types. These new geographies may still be incomplete, subject to change or partly disputable – either because of white spots, which are likely to remain on the map of the surviving microdata recovered by Mosaic, or because of the fragmented nature of many of these pieces of evidence (even if assembled in great numbers). Nevertheless, in comparison with how these things were handled in the pre-Mosaic world, this project represents a major breakthrough. Providing that future scholars will maintain their interest in regionalizing family systems, these attempts may very well profit from discussing regional 'sets of familial tendencies' not based on just a few local case studies (as Laslett did – see Laslett, 1983; see also Polla, 2006; Wall, 1995, 2001), but on a large pool of regionally differentiated data on households and families. In addition to regional 'averages', these future investigations could also incorporate various measures of variation within the spatial clusters involved.

At present, in the case of several countries – Poland-Lithuania, historical Hungary, most of the German-speaking areas and the Balkans, and to some extent also Ukraine – a spatially diverse data coverage will facilitate an exploration of the extent to which meso-level regional demographic regularities corresponded to these countries' internal long-term socio-economic and/or cultural divisions. Szołtysek (2014), for example, found that the broad regional variation of family and marriage patterns across Poland-Lithuania tallied neatly enough with the country's historical socio-economic and cultural fault lines discovered by sociocultural anthropologists of the 1920s and 1930s, thus prompting serious questions about the relationship between the familial sphere and other domains of social life. A preliminary study of household structure in Ukraine revealed a striking correspondence to the country's east–west sociocultural and (later) political divides, which have recently resurfaced as some of the underlying factors of the harsh military conflicts (Szołtysek, 2014). On the other hand, however, the variation in residence and marriage patterns within Germany and Hungary precludes any straightforward classification in terms of long-presumed socio-economic or cultural frontiers (Szołtysek et al., 2014).

These examples invite us to look at the Mosaic data through the prism of more than just statistical representativeness. Almost 100 European regions currently stored in the Mosaic database encourage meso-local investigations of a large number of topics that, thus far, have been poorly explored from a wider territorial perspective. A study of leaving home may serve as an example. Currently, the most comprehensive account of historical nest-leaving patterns looked at differences in home-leaving in the nuclear- and stem-family societies of Eurasia (Van Poppel, Oris, & Lee, 2004; Great Britain, Scania, two provinces in The Netherlands, two Japanese villages, two local sites in Italy, 27 villages in the Pyrenees and 1 rural site from Belgium were analyzed), but did not cover joint-family societies or any historical sites in eastern Europe. Instead of dealing with dispersed cases through variable methodologies, a thoughtful Mosaic-based study of leaving-home patterns could grapple with the problem for dozens (if not hundreds) of regional populations, for all of which harmonized variables could be designed, quantified and analyzed with common statistical methods.

The same holds true for many other issues, such as the residential patterns of the aged, household formation, etc. Integrated across space, the Mosaic microdata has already spawned a comparative study on the residential arrangements of the elderly in different joint-family societies (Szołtysek & Gruber, 2014); other comparisons are equally possible. Such efforts could provide important added value to our understanding of the complexity and variability in family, life-course and residential situations in the past. Even if providing definite answers to questions, along the lines of how much a particular observation tells us about historical Germany, Spain or France, may not always be possible with the Mosaic data, the indisputable gain would be revealing variability in human agency that is not artefactual to data structure or different variable operationalization.

This is particularly true should attempts be made to understand the meso-level diversity of demographic behaviour captured in the Mosaic data by linking it to potential environmental, socio-economic and cultural factors. Though still a largely unoccupied terrain, were it approached properly, it could bypass the project's representativeness problems by shifting scholarly attention away from universal claims to classic anthropological-like concerns about determinants of cross-cultural variation (see, for example, Levinson & Malone, 1980; Whyte, 1978). Understanding the determinants of historical family systems has potentially enormous importance for contemporary demography, if only by providing understanding of the extent to which certain demographic patterns have been persistent and universal, and to what degree they were moulded by diverse economic, environmental and social conditions.

Finally, regional clusters of data assembled within Mosaic – which are more or less representative of their constitutive larger wholes – through offering a grid of information on households and families in hundreds of localities, might facilitate better micro-level research in the future. First, they would provide a natural – and necessary – benchmark against which the more in-depth investigations of family historians could be compared. It is in this realm that the cooperation and synergy between Mosaic and the European Historical Population Samples Network (EHPS-Net) with its Intermediate Data Structure (IDS) is expected to be particularly fruitful (see http://www.ehps-net.eu/; Alter & Mandemakers, 2014), at least in a threefold way: either by facilitating the development of joint relational databases relying on the record linkage of Mosaic cross-sectional data with vital statistics, along the path followed earlier by the Eurasia Project in Population and Family History (see Tsuya, Feng, Alter, & Lee, 2010); by providing a test bed against which longitudinal and cross-sectional observations about life course can be mutually compared; and by providing meso-regional quantifiers of the demographic and familial domains to be used as contextual variables in the investigation of other topics (for example, fertility).

Another advantage of the historical Mosaic data is that, due to its harmonized structure, it can easily be compared with other large data-infrastructure efforts, such as IPUMS International (worldwide contemporary data) and NAPP (historical data, but generally of later provenance than in Mosaic). Gruber and Szołtysek's (2012) research serves as a good example of a fruitful combination of these data sets, which materialized as a reaction to the study by Ruggles (2010), who, comparing 87 censuses from 34 countries around the world, argued that there is virtually no cultural variation in the living situations of the elderly with regard to stem-family arrangements; rather, it is economic and demographic change within particular fractions of populations employed in agriculture that acts as the main driver of observed differences. On the other hand, Gruber and Szołtysek (2012) have established that the incorporation of the first set of Mosaic data from eastern and central Europe reveals distinct differences, which go beyond the economic development explanation. The problem with earlier analyses is that the available data from North America and northern Europe was too limited in time and space to reveal much diversity.

Other applications of the Mosaic data beyond demography and family history should also be possible. We shall highlight only two examples from the neighbouring field of economic history. Historical demography has attracted new attention in recent years, as economists have begun to argue that certain patterns of marriage, female celibacy, individual life course and household structure might have been more conducive to economic growth than others (De Moor & Van Zanden, 2010; Duranton et al., 2009; Greif, 2006; Kick, Davis, Lehtinen, & Wang, 2000). The Mosaic data can be of immense value in testing global-level hypotheses about how family types affect different societal and economic outcomes by providing a large number of localized indicators of demographic and familial, as well as gender and age relations, which can be used as independent variables in regression models, provided that at least partial compliance between Mosaic and other data sets is ensured.

Secondly, the Mosaic project can be of great assistance to economic history in that severe data limitations have forced researchers to resort to proxy indicators in measuring the human capital of historical populations. This concerns especially the project's capacity to equip economic historians with, to date, the most detailed geographical coverage of numeracy (quantitative literacy) patterns in historical societies, by focusing on the phenomenon of 'heaping' in the self-reported age data contained in all of Mosaic's micro-level censuses (see A'Hearn, Baten, & Crayen, 2009; also Baten & Szołtysek, 2014).

Mosaic also offers a partial possibility to analyze change over time: there are already a few places or regions for which Mosaic holds data from more than one census. For other places, there is microdata for different points in time, but it has not yet been transcribed. Future scholars might be interested in transcribing additional census microdata (information on existing microdata may be obtained from the country inventories mentioned above) and linking it to already existing microdata within Mosaic.

Research with Mosaic's microdata is not restricted to analyses of household structures and similar research questions; other demographic questions can also be addressed. One is fertility, which could be analyzed by using the own-child method or the child–woman ratio (see Breschi, Kurosu, & Oris, 2003). Studies of societal differentiation could be conducted with the use of information about occupations, which could then be further applied to investigations about the proportion of the servant population. Naming patterns and their change over time could be juxtaposed with and analyzed against the respective patterns of different – older – generations, including the change from more religiously based names to more secular ones.

Last but not least, some data sets have the potential to utilize information not included in the harmonized variables, such as place of birth, language spoken, legal status or property. Although spatial coverage of these listings is far from comprehensive, they may open a window onto studies of migration patterns, multiethnic societies and historical patterns of societal diversity.

Acknowledgements

Earlier versions of this article were presented at the European Social Science History Conference, Vienna, 23–24 April 2014, and later at the Centre for Population Studies, Umeå University, 17 June 2014. We would like to thank Gunnar Thorvaldsen and the two anonymous reviewers for their comments on an earlier draft of this article, as well as Julia Szołtysek for language editing.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes

1. See https://usa.ipums.org/usa/ and; https://www.nappdata.org/napp/intro.shtml.
2. A large collection of nominative census lists of pre-industrial English communities, to our knowledge, has never been fully computerized, nor made publicly available (see Laslett, 1972, p. 133; Wall, Woollard, & Moring, 2004). The Vienna Database on European Family History contains census lists for at least 50 settlements in Austria (30,644 households with a population of 175,429 individuals), as well as some scattered nominative listings from various parts of Europe (246,000 individuals overall). Unfortunately, no transcriptions of names, relationship terms or occupational titles were stored in the database, but only codes, which were unique to each respective project. See http://www.univie.ac.at/Wirtschaftsgeschichte/famdat/index-gr.html and also Mitterauer (1986, pp. 194–197).
3. Like for example in Lithuania where all records were stored in the central historical archive in Vilnius, but in the sections which have never been recatalogued since the maelstrom of the Second World War.
4. On a metahistorical level, it could be argued that the available relics of historical listings are just a selection of all that could have survived – a random sample, so to speak, of a total universe of relics. The range of the listings' actual preservation was the outcome of stochastic processes affecting the material's chances of survival (with the exclusion of the presumably rare cases of the purposeful destruction of particular listing units) through wars, dislocations, robberies and natural disasters of all kinds.
5. Naturally, the two versions of the database – i.e. the one that functioned merely as a form of storage of the original transcripts and the other, which was amended using various technical and methodological enhancements, like standardization and coding – were kept separate during the entire process.
6. For example, although most of Mosaic microdata samples ask about marital status, they may differ in their classification schemes. In order to create an integrated variable for marital status, the marital status variable from each listing was recoded into a unified coding scheme that we have designed.
7. Courtesy of Elena Glavatskaya, Faculty of History, Institute of Archaeology and Ethnology, Ural Federal University, Ekaterinburg, Russia.
8. See http://www.iecm-project.org/index.php.
9. These measurement errors may stem from either 'demographic' randomness (which enters at the level of the the individual person or family) or 'environmental' randomness (which enters at the level of population due to unequal exposure to social, economic, climatic and other influences).

References

A’Hearn, B., Baten, J., & Crayen, D. (2009). Quantifying quantitative literacy: Age heaping and the history of human capital. *The Journal of Economic History*, *69*, 783–808. 10.1017/S0022050709001120

Alter, G., & Mandemakers, K. (26-05-2014). The intermediate data structure (IDS) for longitudinal historical microdata. version 4. *Historical Life Course Studies*, *1*, 1–26. Retrieved from http://hdl.handle.net/10622/23526343-2014-0001?locatt=view:master

Baten, J., & Szołtysek, M. (2012). A golden age before serfdom? The human capital of Central-Eastern and Eastern Europe in the 17th–19th centuries. MPIDR Working Paper WP 2014-008 AUGUST 2014.

Breschi, M., Kurosu, S., & Oris, M. (2003). *The own-children method of fertility estimation. Applications in historical demography*. Udine: Forum.

Czap, P. (1983). ‘A large family: The peasant’s greatest wealth’: Serf households in Mishino, Russia, 1814–1858. In R. Wall, J. Robin, & P. Laslett (Eds.), *Family forms in historic Europe* (pp. 105–151). Cambridge: Cambridge University Press.

De Moor, T., & Van Zanden, J. L. (2010). Girl power: The European marriage pattern and labour markets in the North Sea region in the late medieval and early modern period. *The Economic History Review*, *63*(1), 1–33. 10.1111/j.1468-0289.2009.00483.x

Dennison, T. K., & Ogilvie, S. (2013). Does the European marriage pattern explain economic growth? CESifo Working Paper Series, 4244 (May 31, 2013).

Dillon, L., & Roberts, E. (2002). Introduction: Longitudinal and cross-sectional historical data: Intersections and opportunities. *History and Computing*, *14*(1–2), 1–7. 10.3366/hac.2002.14.1-2.1

Duranton, G., Rodríguez-Pose, A., & Sandall, R. (2009). Family types and the persistence of regional disparities in Europe. *Economic Geography*, *85*, 23–47. 10.1111/j.1944-8287.2008.01002.x

Fauve-Chamoux, A., & Ochiai, E. (Eds.). (2009). *The stem family in Eurasian perspective. Revisiting house societies, 17th–20th centuries*. Bern: Peter Lang.

Greif, A. (2006). Family structure, institutions, and growth: The origins and implications of Western corporations. *American Economic Review*, *96*, 308–312. 10.1257/000282806777212602

Gruber, S., & Szołtysek, M. (2012). Stem families, joint families, and the European pattern: What kind of a reconsideration do we need?. *Journal of Family History*, *37*, 105–125. 10.1177/0363199011428124

Gruber, S., & Szołtysek, M. (2015). The Patriarchy Index: A comparative study of power relations across historic Europe. *The History of the Family*, in press. (http://dx.doi.org/10.1080/1081602X.2014.1001769)

Jordan-Bychkov, T. G., & Bychkova-Jordan, B. (2002). *The European cultural area: A systematic geography* (4th ed.). New York, NY: Rownam & Littlefield.

Kaser, K. (2002). Power and inheritance: Male domination, property and family in Eastern Europe, 1500–1900. *The History of the Family*, *7*, 375–395. 10.1016/S1081-602X(02)00109-4

Kaser, K., Gruber, S., Kera, G., & Pandelejmoni, E. (2011). *1918 census of Albania, Version 0.1* [SPSS file]. Graz: University of Graz.

Kelly Hall, P., McCaa, R., & Thorvaldsen, G. (Eds.). (2000). *Handbook of international historical microdata for population research*. Minneapolis: Minnesota Population Center.

Kick, E., Davis, B., Lehtinen, M., & Wang, L. (2000). Family and economic growth: A world-system approach and a cross-national analysis. *International Journal of Comparative Sociology*, *41*, 225–244. 10.1177/002071520004100203

Klüsener, S., Szołtysek, M., & Goldstein, J. R. (2012). Towards an integrated understanding of demographic change and its Spatio-Temporal dimensions: Concepts, data needs, and example case studies. *Die Erde: Zeitschrift der Gesellschaft für Erdkunde*, *143*, 75–104.

Kok, J. (2009). Family systems as frameworks for understanding variation in extra-marital births, Europe 1900–2000. In A. Fauve-Chamoux & I. Bolovan (Eds.), *Families in Europe between the 19th and the 21st centuries: From the traditional model to contemporary PACS* (pp. 13–38). Cluj-Napoca: University Press.

Kruskal, W., & Mosteller, F. (1979a). Representative sampling, II: Scientific literature, excluding statistics. *International Statistical Review / Revue Internationale de Statistique*, *47*, 111–127.

Kruskal, W., & Mosteller, F. (1979b). Representative sampling, III: The current statistical literature. *International Statistical Review / Revue Internationale de Statistique*, *47*, 245–265.

Laslett, P. (1969). Size and structure of the household in England over three centuries. *Population Studies*, *23*, 199–223. 10.1080/00324728.1969.10405278

Laslett, P. (1972). Mean household size in England since the sixteenth century. In P. Laslett & R. Wall (Eds.), *Household and family in past time* (pp. 125–158). Cambridge: Cambridge University Press.

Laslett, P. (1977). Characteristics of the western family considered over time. *Journal of Family History*, *2*, 89–115. 10.1177/036319907700200201

Laslett, P. (1983). Family and household as work group and kin group: Areas of traditional Europe compared. In R. Wall & J. Robin (Eds.), *Family forms in historic Europe* (pp. 513–563). Cambridge: Cambridge University Press.

Laslett, P., Wachter, K. W., & Laslett, R. (1978). The English evidence on household structure compared with the outcomes of microsimulation. In K. W. Wachter, E. A. Hammel, & P. Laslett (Eds.), *Statistical studies of historical social structure*. New York, NY: Academic Press.

Le Play, F. (1877–1879). *Les Ouvriers Européens* [The European workers]. (6 vols). Tours: A. Mame et fils.

Levinson, D., & Malone, M. J. (1980). *Toward explaining human culture: A critical review of the findings of worldwide cross-cultural research*. New Haven, CT: HRAF Press.

Mandemakers, K., & Dillon, L. (2004). Best practices with large databases on historical populations. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *37*, 34–38. 10.3200/HMTS.37.1.34-38

Mitterauer, M. (1986). Formen ländlicher Familienwirtschaft. Historische Ökotypen und familiale Arbeitsorganisation im österreichischen Raum [Forms of rural household economy. Historical eco-types and organisation of work within the family in Austria]. In J. Ehmer & M. Mitterauer (Eds.), *Familienstruktur und Arbeitsorganisation in ländlichen Gesellschaften* [Family structure and organisation of work in rural societies]. (pp. 185–323). Wien: Böhlau.

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

Polla, M. (2006). Family systems in central Russia in the 1830s and 1890s. *The History of the Family*, *11*, 27–44. 10.1016/j.hisfam.2005.11.001

Rijpma, A., & Carmichael, S. (2013). Testing Todd: Global data on family characteristics. Paper presented at the Workshop Workshop "Agency, Gender, and Economic Development in the World Economy 1850–2000", Utrecht University.

Ruggles, S. (1987). *Prolonged connections: The rise of the extended family in nineteenth. Century England and America*. Madison: University of Wisconsin Press.

Ruggles, S. (1999). The limitations of English family reconstitution: English population history from family reconstitution 1580–1837. *Continuity and Change*, *14*, 105–130. 10.1017/S0268416099003288

Ruggles, S. (2010). Stem families and joint families in comparative historical perspective. *Population and Development Review*, *36*, 563–577. 10.1111/j.1728-4457.2010.00346.x

Ruggles, S. (2012). The future of historical family demography. *Annual Review of Sociology*, *38*, 423–441. 10.1146/annurev-soc-071811-145533

Ruggles, S., & Heggeness, M. (2008). Intergenerational coresidence in developing countries. *Population and Development Review*, *34*, 253–281. 10.1111/j.1728-4457.2008.00219.x

Ruggles, S., Roberts, E., Sarkar, S., & Sobek, M. (2011). The North Atlantic population project: Progress and prospects. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *44*(1), 1–6. 10.1080/01615440.2010.515377

Smith, R. M. (1981). The people of Tuscany and their families in the fifteenth century: Medieval or Mediterranean? *Journal of Family History*, *6*, 107–128. 10.1177/036319908100600111

Szołtysek, M. (2012). *CEURFAMFORM database, Version 23 [SPSS file]*. Rostock: MPIDR.

Szołtysek, M. (2014). Rethinking East-central Europe: Family systems and co-residence in the Polish-Lithuanian Commonwealth. Habilitation thesis submitted to the Philosophische Fakultät I der Martin-Luther-Universität Halle-Wittenberg.

Szołtysek, M., & Gruber, S. (2012). Family systems in preindustrial Germany: Comparative spatial perspectives. Paper presented at the Workshop "Family, property and markets: comparative perspectives on Japan and Europe". University of Münster, Department of History, Germany, February 2012.

Szołtysek, M., & Gruber, S. (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*, 101–140.

Szołtysek, M., Gruber, S., Klüsener, S., & Goldstein, J. R. (2014). Spatial variation in household structure in 19th-century Germany. *Population-E*, *69*, 55–80.

Therborn, G. (2004). *Between sex and power: Family in the World 1900–2000*. London: Routledge.

Thorvaldsen, G. (1994). The encoding of highly structured historical sources. *Computers and the Humanities*, *28*, 301–305. 10.1007/BF01830278

Todd, E. (1985). *The explanation of ideology. Family structures and social systems*. Oxford: Basil Blackwell.

Todd, E. (2011). *L'origine des systèmes familiaux. Tome 1: l'Eurasie* [The origin of family systems. Vol. 1: Eurasia]. Paris: Gallimard.

Tsuya, N. O., Feng, W., Alter, G., & Lee, J. Z. et al. (2010). *Prudence and pressure. Reproduction and human. Agency in Europe and Asia, 1700–1900*. Cambridge, MA: The MIT Press.

Vallin, J. (2006). Populations and individuals. In G. Caselli, J. Vallin, & G. Wunsch (Eds.), *Demography: Analysis and synthesis. A treatise in population* (Vol. Vol. 1, pp. 5–8). Elsevier: Academic Press.

Van Poppel, F., Oris, M., & Lee, J. (Eds.). (2004). *The road to independence: Leaving home in Western and Eastern societies, 16th–19th centuries*. Bern: Peter Lang.

Wachter, K. W. (1978). Age pyramid variances. In K. W. Wachter, E. A. Hammel, & P. Laslett (Eds.), *Statistical studies of historical social structure* (pp. 189–217). New York, NY: Academic Press.

Wall, R. (1994). Historical development of the household in Europe. In E. van Imhoff, A. C. Kuijsten, P. Hooimeijer, & L. J. C. van Wissen (Eds.), *Household demography and household modeling* (pp. 19–52). New York, NY: Plenum Press.

Wall, R. (2001). Transformation of the European family across the centuries. In R. Wall, T. K. Hareven, J. Ehmer, & M. Cerman (Eds.), *Family history revisited. Comparative perspectives* (pp. 217–241). Newark: University of Delaware Press.

Wall, R., Woollard, M., & Moring, B. (2004). *Census schedules and listings, 1801–1831: An introduction and guide*. Colchester: University of Essex, Department of History. Retrieved March 24, 2014, from <https://www.academia.edu/619532/Census_Schedules_and_Listings_1801-1831_An_Introduction_and_Guide>.

Whyte, M. K. (1978). *The status of women in pre-industrial societies*. Princeton, NJ: Princeton University Press.

Zimmerman, C. C., & Frampton, M. E. (1935). *Family and society: A study of the sociology of reconstruction*. New York, NY: D. Van Nostrand Company.