Articles

Does Parents' Position Persist?

Measuring Multigenerational Persistence of Socioeconomic Status in Genealogical Databases on the Example of Germany, 1600-1900

Jan Michael Goldberg
Martin Luther University, Halle-Wittenberg

Abstract

Using a large public genealogical database, this study presents new indicators to measure multigenerational mobility. They are examined on the example of Germany between 1600 and 1900, specifically analyzing up to six contiguous generations to determine the degree of multigenerational stability of the socioeconomic status. An AR(1) model is employed to assess the socioeconomic status, using occupational information. The findings reveal lower mobility in historical Germany, but stable with the number of observed generations and with variations across different regions and time periods. This study presents an innovative approach to incorporating crowdsourced genealogical data into institutional scientific research, as the data was gathered and linked by private individuals.

1. Introduction

Mobility research usually focuses on social mobility between two generations. While mobility refers to a weak dependence on the previous generation, persistence describes a strong connection of the socioeconomic status.[1] However, one limitation of this bi-generational approach is that it only examines short periods. This

[1] Unless otherwise stated, the term "mobility" refers to intergenerational upward or downward mobility. The terms social and socioeconomic mobility as well as social and socioeconomic status are only differentiated linguistically and not in terms of content.

means that no conclusions can be made about the long-term development within families. Nevertheless, Hällsten and Kolk identify a research gap in studies about several contiguous generations (2021, p. 3). Recently some studies have emerged that deal with more than two connected generations. Studies that illuminate structures over three or more generations are even given the term "multigenerational" rather than "intergenerational" (Mare, 2011; Solon, 2018, p. F340). Such multigenerational studies often conclude that a long-term view results in a higher persistence than that of two-generational families (Lindahl et al., 2015, p. 25; Braun and Stuhler, 2018, p. 607; e.g. Adermon, Lindahl and Palme, 2021, pp. 1540–1541; Hällsten and Kolk, 2021, p. 10). Some of the results contradict each other, for example in the intensity of dependence on previous generations. This is another reason why Solon shows that there is a need to better understand the different results of multigenerational studies (2018, p. F351).[2]

The starting point of this study is Clark's "law of social mobility." Clark describes a context in which socioeconomic status is consistently passed on across all societies and time periods with a factor of about 0.75 (Clark et al., 2014, pp. 108-109) or above 0.80 (Clark, Cummins and Curtis, 2022, p. 11). This invariability of social mobility implies that policy measures do not influence the shape of social mobility. Clark et al. use surname studies to compare the distribution of rare surnames over time. If so, these results must also be reproducible in individual linkages.

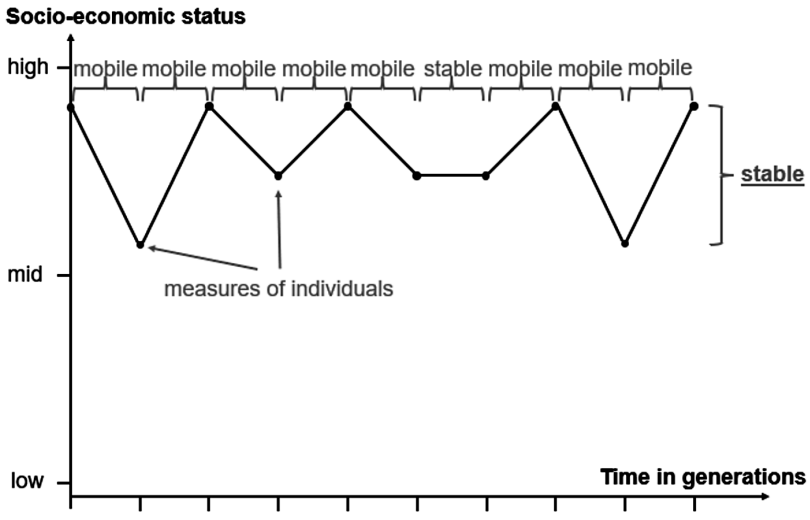
This work focusses on a new way of measuring multigenerational mobility. In doing so, genealogical networks are examined without the assumption of a relationship based on the same surname. These new methods are applied to data sets comprising many generations. For this purpose, combined data over as many genera-

[2] In addition to the differences that arise from inter- and multigenerational perspectives, different results are also found between the mobility of groups and individual families or even between regions. Group effects in particular offer a starting point for further research (Güell, Rodríguez Mora and Solon, 2018, pp. F337-F338).

tions as possible is used. In the following, we will check whether the non-surname-based methods determine the persistence of the status at a similar level as the surname-based ones. Then, temporal and spatial differences or similarities are considered. It is argued that considering several related generations leads to lower mobility. Considering only two generations in classical studies thus underestimates the effect of the family on the socioeconomic status: if only two generations are compared, more mobility is observed than if many generations are compared together (Figure 1). Even if mobility is low, as in Clark's case, differences in time and space can be observed.

This study defines the socioeconomic status as access to social and economic resources. It becomes visible through a variety of measurable characteristics. Vosters, for example, recommends income, education and occupation (2018, p. 419). Collecting the necessary data sets for multigenerational mobility studies is challenging (Barone and Mocetti, 2021, p. 1864). Therefore, crowdsourced ge-

Figure 1
Mobility as a question of perspective;
evolution of socioeconomic status in a fictitious family



nealogical data is applied in this study: family tree files (GEDCOM) uploaded by private individuals. The data contained in these online databases represents a largely untapped potential for science.[3] One challenge in dealing with this data is determining the socioeconomic status. According to Clark et al. many studies arrive at high mobility because the measures are subject to error. Comparisons between countries are thus "suspect" (Clark, Cummins and Curtis, 2022, p. 16). To test this, this study is carried out on the example of Germany, which has a diverse history due to its federal structure, but at the same time also features wide similar cultural areas. Due to its historically fragmented structure, Germany is a suitable example for regionally differentiated observations. Furthermore, the author is not aware of any studies other than Braun/Stuhler (2018) that use Germany as an example to examine the long-term persistence of socioeconomic statuses over many successive generations. In the absence of multigenerational studies about Germany, this study thus opens another chapter by examining long-term socioeconomic mobility, from the Thirty Years' War to the 20th century. The main points are the following: (1.) lower mobility is indeed measured when looking at many generations but (2.) differences can be found between regions and periods.

The (predominantly German-language) database "Genealogische Datenbasis" (GEDBAS, Verein für Computergenealogie, 2020) is selected for this study. The database is managed by the Verein für Computergenealogie e. V. (Association for Computer Genealogy). The GEDCOM files to be found on the platform are created by many individuals. Only those files that have been made available for free download are used. Many records are based on data from church registers, which in most German parishes were not created until the 17th century. In terms of time, GEDBAS records often cover the period from the 16th to the 20th century, so this is the period taken under consideration.

[3] A counterexample that makes use of Family Search data can be found in Price et al. (2021).

The occupational information in the database is used as an indicator of socioeconomic status. Examples of intergenerational studies based on occupations are offered by various authors (e.g. Long and Ferrie, 2013, 2018; Dribe and Helgertz, 2016; Modalsli and Vosters, 2024). It is operationalized according to different classification systems (HISCO, KldB 2010, see in the appendix "Classification of occupational data"). From this, various indicators of social mobility are formed and calculated for each family. A comparison is made as to whether different mobility can be measured for families over two generations than for families over more generations. In this way, it can be determined whether the measured persistence increases with an increasing multigenerational consideration.

The next chapter opens describing data and its processing, followed by modeling multigenerational social mobility indicators. Building on this, the results are presented and interpreted before a conclusion is drawn.

2. Data

Multigenerational studies face a significant challenge due to the lack of appropriate data. To overcome this obstacle, new methods must be developed to both collect and process extensive data with available resources (Mare, 2011, p. 2). Recent research has explored innovative approaches to this problem, as seen in the studies conducted by Clark et al., who present an innovative approach to multigenerational studies, which involves analyzing rare surname groups instead of direct kinship relations (2015, pp. 6-7). This approach has also been adopted by other researchers, including Barone and Mocetti (2021) and Häner and Schaltegger (2022). The study of surname groups over time has revealed that social mobility may be lower than previously thought. Clark and Cummins' research on England demonstrates that intergenerational status persistence is higher for surname groups than for individual families (2015, p. 78). They propose that this difference from bi-generational studies may be due to

the transmission of underlying social skills that are not directly measurable through observable characteristics such as earnings, wealth, residence, education, occupation, health, and longevity (Clark et al., 2014, pp. 8-9). Clark argues that measurement errors in individual variables may cancel each other out on average, thus validating their approach (Clark et al., 2014, pp. 110-111). In general, Ward also finds out that "accounting for […] measurement error can double the estimates of intergenerational persistence" (Ward, 2023, p. 3213), although the approach he uses is not surname-based.

However, the approach via the surname group also shows significant weaknesses. On the one hand, some depict only marginal groups such as the upper or lower echelons of society (e.g. Clark and Cummins, 2015, p. 66). On the other hand, the selection of surnames only refers to the socioeconomic position at a defined starting point (Clark et al., 2015, pp. 12-13). Additionally, the assumption that people with the same surnames belong to the same families may approximate actual kinship structures well, but individual linkages would be more suitable to avoid group effects. Vosters criticizes Clark's result as based on a measurement error, as he only uses one unit of measurement (2018, p. F419). There is also criticism that group mobility cannot be compared with individual mobility (Torche and Corvalan, 2018, pp. 11-12). As a general criticism of Clark's approach, Solon's work is also recommended (Solon, 2018, pp. F347-F350). Moreover, this picture would have to be confirmed not only among (elite) fringe groups but in the broader population. Testing this would require a substantial amount of data. Hällsten and Kolk uses linked public registers from Sweden with data from church records to examine up to seven generations. Their research compared distant cousins who share common ancestors dating back to the 18th century (Hällsten and Kolk, 2021, p. 3). Similarly, Adermon, Lindahl, and Palme took a dynastic perspective in their research, utilizing public Swedish data from 1932 onwards to illuminate the 20th century across four generations (Adermon, Lindahl and Palme, 2021, p. 1523 u. 1530). Stuhler and Braun also utilized innovative methods by analyzing data from the German

National Educational Panel on four generations spanning between the 19th and 20th centuries (2018, pp. 576-577). These studies provide valuable insights into multigenerational research and demonstrate the potential for innovative approaches in multigenerational linkage to overcome data limitations.

One innovation in this study is the use of crowdsourced genealogical databases for analysis. However, these databases may suffer from various biases that need to be considered. Firstly, individuals with more time and resources may be more likely to engage in this type of research, potentially biasing the sample towards those with higher socioeconomic status in the present. Secondly, genealogical information from certain population groups may be more accessible on the internet due to their higher socioeconomic status. Thirdly, individuals with a higher socioeconomic status may be easier to trace, potentially leading to under-representation of those with a lower socioeconomic status. Fourthly, individuals who are more skilled in using archives are likely to have more success in assembling family histories, potentially leading to over-representation of those with certain skills or professions. Additionally, individuals with a family history of migration may face language barriers and distance to archives, making it less likely for them to trace their family history and so on. In principle, this behavior means that people with a higher socioeconomic status tend to be overrepresented in genealogical databases. It is crucial to be aware of these potential distortions in the data when analyzing it.

Various genealogical databases with different structures can be found online, but GEDBAS has been selected for further investigation. It is a significant database in German-speaking countries where genealogical data can be uploaded and viewed without any financial barriers. The underlying data of the genealogies is based mainly on church books (parish registers) which are often available in Germany from the middle of the 17th century and contain information about baptisms, marriages and funerals occurred in each parish. Excluding losses, it is almost a documentation of the entire Christian population. Private individuals upload their records in a standardized

GEDCOM format, which is the common standard used in the exchange of genealogical information (Gellatly, 2015, p. 112; Harviainen and Björk, 2018, p. 4). Users can share their files for download, and the GEDCOM files consist of text divided into lines, with tags defining the content of each line (such as names and occupational details). One further advantage of using GEDBAS is the availability of data over multiple generations, and in this study up to six generations have been used. The choice of six generations is based on the availability of sufficient, stationary data up to this number of consecutive generations (see Appendix, Tables 12 and 13). There are few studies that cover such a long period. Barone and Mocetti claim to be the first to present a study with linked individuals over a span of 600 years (2021, p. 1866), but their linkage is also based solely on surnames.

The indexing of GEDBAS is described in the appendix (refer to "Indexing GEDBAS data"), where the genealogical structures are divided into a panel data set ("data sets") for each descent sequence, resulting in a total of 58,000 data sets (or families) being indexed. The individual data sets are divided into cohorts of 30 years, which roughly corresponds to the interval of one generation. The appendix also explains the effects of dividing genealogical structures into these panel data sets (see "Effects of dividing genealogical structures into panel data sets").

Using online crowdsourced genealogical databases like GEDBAS, poses the relevant issue of representativeness (see Stelter and Alburez-Gutierrez, 2022, p. 1), especially in terms of geography, time period, and socioeconomic factors. Based on this, the following is divided into three sections, namely temporal, geographical, and socioeconomic representativeness. Summarized, the data accurately reflects relevant conditions for a practical analysis, despite a few exceptions. Temporally, data shows a notable drop after 1900 due to data protection limits. Before that, there is a strong correlation with historical statistics of births and deaths. Geographically, the data aligns well with 19th-century Prussian provinces, though some regions are underrepresented (especially the former German eastern

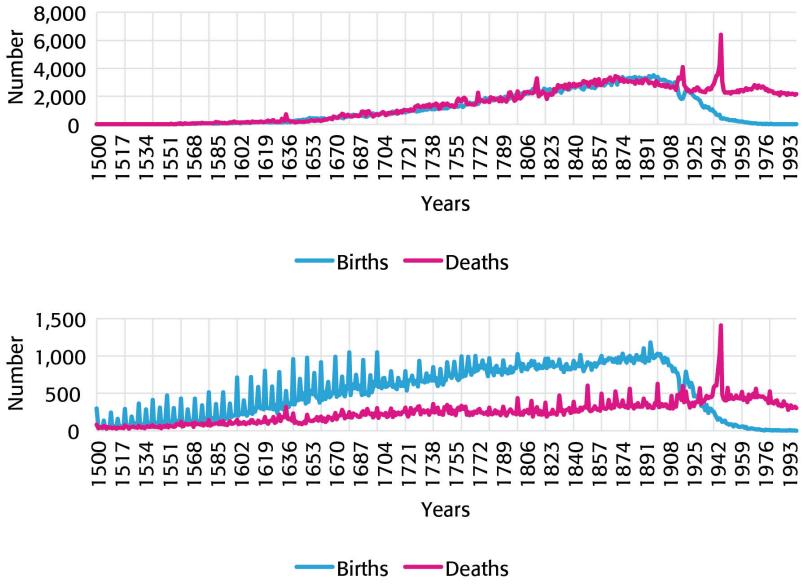
territories and small regions, with are excluded from this analysis). Socioeconomically, the GEDBAS data aligns reasonably with the 1867 Prussian census, though agricultural occupations are underrepresented, indicating potential biases in genealogical interest and data collection. All in all, the genealogical data in GEBDAS is very similar to other data sources of historical demography, which is in line with the high correlation of crowdsourced genealogical data with representative data as shown by Blanc (2024a, p. 8, 2024b, p. 11).

2.1 Temporal Representativeness

Figure 2 (top) depicts an exponential increase in the number of people born between 1500 and 1800. This growth trend continued until the second half of the 19th century, after which the number of recorded births sharply declined. The drop in births around 1900 is not surprising, as the German Civil Status Act mandates a retention period of 110 years for birth records. GEDBAS adheres to this rule and automatically deletes data for potentially living individuals upon upload (Verein für Computergenealogie, 2021). In contrast, the number of recorded deaths remained relatively low compared to births, with a local maximum observed in the second half of the 19th century. This discrepancy can be attributed to genealogical practice, where death records are less commonly recorded due a greater difficulty of determination or lesser relevance for genealogical research beyond the immediate family. However, notable fluctuations in the number of deaths, such as increased mortality during the Thirty Years' War or the World Wars, can be observed. Unlike births, death records are not subject to automatic deletion and the graph does not show a leveling off after 1900, as there is no legal regulation for their retention period. Overall, while the data isn't a representative sample, it will be shown that it is still considered reliable for analysis purposes regarding the periods before 1900.

Genealogical interpretations and missing data make birth dates particularly prone to distortions, referred to as artefacts. Birth dates

Figure 2
Temporal distribution of births and deaths
in the downloadable data of GEDBAS



Above: exact dates, below: Year data (artefacts) and date data on 1 January.

are often estimated based on age information in marriage or burial records, or even on the birth year of the first child, resulting in rounded years of birth (e.g. 1700 instead of 1702). Often, only the year is available, which must be converted to a date format for processing, typically set to 1 January. This makes the artefacts indistinguishable from genuine events that occurred on 1 January, which are filtered out for display (see Figure 2, bottom). A closer inspection of the data reveals larger peaks at the beginning of each decade, and smaller ones five years later, especially before 1800, where larger peaks occur in the 10s and smaller ones in the 5s. These artefacts indicate a clear and consistent decline in the quality of data from the more distant past up to 1800. This situation changed in the 19th century, presumably due to a more precise data collection and more accurate birth information in marriage and death documents.

To assess the temporal representativeness of the data, it is useful to compare it with existing historical statistics. Pfister and Fertig have published data for German-speaking areas from 1729 to 1869 (see online attachment of Pfister and Fertig, 2020). By comparing the relative development of births and deaths in GEDBAS with this data (see Figure 3 for deaths and Figure 4 for births), we can see a strong correlation between the two time series, normalized to the year 1830. However, the data suggests that deaths before the 19th century are underrepresented in GEDBAS. This gap appears to increase until 1729, and it is likely to have increased even further in earlier periods.

Figure 3
Comparison of the relative development of deaths, normalized by 1869 to 1.0

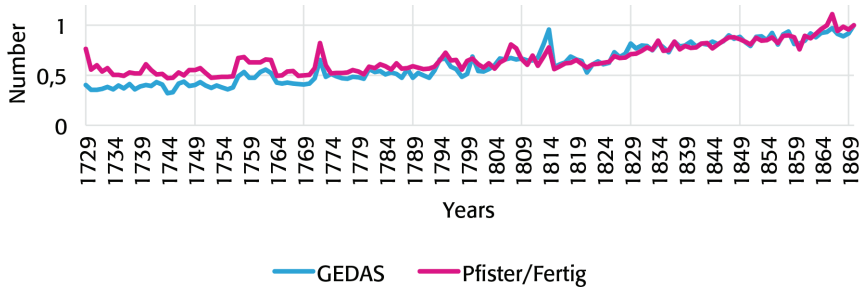
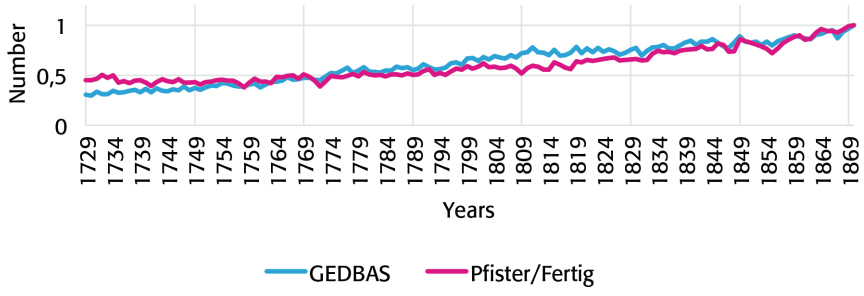


Figure 4
Comparison of the relative development of births, normalized to 1.0 around 1869.987



The birth time series also exhibits a strong correlation with the historical statistics data, but it is noteworthy that GEDBAS shows a relative birth surplus from 1800 to 1830 when normalized to 1869. This discrepancy can be explained by genealogical practice, as most databases tend to have more records of ancestors further back in time, since every new added parent expands the number of relatives. The 19th century is well-documented, but sources become scarcer and less reliable the further back in time we go. Therefore, the surplus disappears in the 18th century and, as with the death data, there are signs of under-representation in the GEDBAS records before 1750. This gap is likely to increase in the preceding periods.

The GEDBAS data is not a complete representation of the historical development over time. As demonstrated, there is a clear quantitative temporal imbalance in the data. The underrepresentation in GEDBAS becomes probably more pronounced in the period before 1729, as there is scarce data available for the years around 1500, despite an estimated nine million inhabitants living in the territory of Germany at the time (Pfister, 2010, pp. 10-11). According to Pfister, the population of the same area was around 16.2 million in 1600, but decreased to 14.1 million by 1700 due to the Thirty Years' War (2010, pp. 10, 14). However, this population reduction is not reflected in the GEDBAS data. While the correspondence between historical statistics and GEDBAS are relatively close in the 19th and 18th centuries, this is not the case for the 17th century and earlier periods. This fact is considered in further analysis.

2.2 Geographical Representativeness

Ensuring geographical representativeness is crucial in evaluating the GEDBAS data, as it could potentially be skewed by regional biases. To address this issue, researchers have utilized data on births and deaths from 38 former states of Germany or Prussian provinces spanning from 1815 to 1870 (Fertig et al., 2018). Although this period is not entirely congruent with the one covered by the GEDBAS data, the comparison of the geographical distribution in the 19th century

implies that there is a similar correspondence in the preceding period because of limited variation in the geographical makeup. This is because the pre-modern population had low spatial mobility, making it unlikely that the relationship between the Thirty Years' War and the beginning of the 19th century would significantly change (Imhof, 1981, pp. 35-36). While some scholars like Hochstadt argue against the notion of an immobile population in the pre-industrial era, their objections do not significantly impact this observation, as inter-territorial mobility was low (1983, pp. 198, 214). By comparing cumulative births and deaths in the 38 provinces, 66 percent of the people could also be assigned to a historical province.

The data cleaning and supplementation process has revealed that the relative proportions of birth numbers between 1815 and 1870 do not exactly match in most provinces (Table 1). The proportions of regions reveal which areas are well represented genealogically (e.g., Westphalia) and which are not (e.g., East Prussia). Regions are not considered further for four reasons: (1) fewer than 900,000 births in the period under consideration according to Fertig et al., (2) fewer than 900 births in the GEDBAS-based data set, (3) relation of both shares outside the range 0.5-3.0, (4) former eastern German provinces due to high probability of asymmetric war losses and difficult access to sources.

Therefore, only eleven regions are considered, namely the Province of Saxony (A 09), the Province of Westphalia (A 11), the Rhine Province (A 13), the Grand Duchy of Baden (B 10), Hesse (B 11), the Kingdom of Württemberg (B 18), the Kingdom of Bavaria (B 19), the Kingdom of Hanover (B 20), the Kingdom of Saxony (B 21), the Electorate of Hesse (B 22), and the Thuringian States (B 24). These regions represent a significant portion of the current territory of Germany, while other regions such as the eastern Prussian provinces, city-states, and small principalities, as well as Brandenburg, Holstein, and Mecklenburg, are not considered (Figure 5). The regions with an A in their abbreviation are Prussian, while those with a B are not.

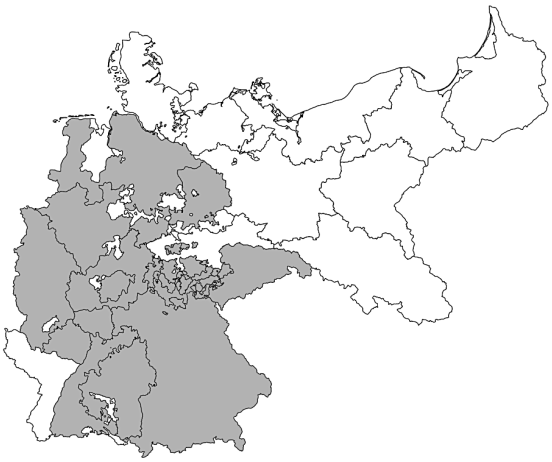
TABLE 1
Births and deaths by administrative-territorial units in the 19th century

Province	Cumulative birth rate, 1815-1870				Relation of both shares
	Fertig et al.		GEDBAS[1]		
	Abs.	Rel. (%)	Abs.	Rel. (%)	
A01 Province of Holstein	638,505	1.26	442	0.49	0.39
A02 Province of Lauenburg	58,236	0.11	225	0.25	2.28
A03 Province of Brandenburg (without Berlin)	2,332,076	4.59	1,421	1.58	0.35
A04 Province of Hesse-Nassau	228,163	0.45	715	0.80	1.77
A05 Province of Hohenzollern	111,717	0.22	68	0.08	0.34
A06 Province of East Prussia	2,640,944	5.19	474	0.53	0.10
A07 Province of Pomerania	1,600,273	3.15	2,022	2.26	0.72
A08 Province of Poznan	2,348,625	4.62	2,514	2.80	0.61
A09 Province of Saxony	2,582,290	5.08	2,392	2.67	0.53
A10 Province of Silesia	5,138,716	10.11	1,788	1.99	0.20
A11 Province of Westphalia	2,111,943	4.15	7,935	8.85	2.13
A12 Province of West Prussia	1,835,365	3.61	3,051	3.40	0.94
A13 Rhine Province	4,012,234	7.89	10,680	11.91	1.51
A14 Province of Berlin	650,786	1.28	0	0.00	–
B01 Amt Bergedorf	16,724	0.03	33	0.04	1.23
B02 Hanseatic City of Bremen	112,733	0.22	33	0.04	0.17
B03 City of Hamburg	341,006	0.67	416	0.46	0.69
B04 City of Lübeck	59,961	0.12	45	0.05	0.42
B05 City of Frankfurt am Main	77,693	0.15	12	0.01	0.09
B06 Principality of Lippe-Detmold	204,862	0.40	328	0.37	0.91
B07 Principality of Schaumburg-Lippe	33,284	0.07	34	0.04	0.54
B08 Principality of Waldeck-Pyrmont	82,710	0.16	1,646	1.84	11.47
B09 Grand Duchy of Oldenburg	370,059	0.73	1,035	1.15	1.58
B10 Grand Duchy of Baden	1,990,123	3.91	7,976	8.90	2.28
B11 Hesse	1,120,179	2.20	5,684	6.34	2.88
B12 Grand Duchy of Mecklenburg-Schwerin	657,785	1.29	1,240	1.38	1.07
B13 Grand Duchy of Mecklenburg-Strelitz	112,572	0.22	99	0.11	0.50
B14 Duchy of Anhalt	209,093	0.41	180	0.20	0.49
B15 Duchy of Brunswick	380,589	0.75	684	0.76	1.02
B16 Duchy of Nassau	491,064	0.97	43	0.05	0.05
B17 Duchy of Schleswig	426,297	0.84	5	0.01	0.01
B18 Kingdom of Württemberg	2,983,465	5.87	12,174	13.58	2.31
B19 Kingdom of Bavaria	7,240,840	14.24	14,127	15.76	1.11
B20 Kingdom of Hanover	2,361,113	4.64	4,879	5.44	1.17
B21 Kingdom of Saxony	3,064,011	6.03	2,817	3.14	0.52
B22 Electorate of Hesse	926,936	1.82	1,403	1.56	0.86
B23 Landgraviate of Hesse-Homburg	29,145	0.06	43	0.05	0.80
B24 Thuringian States	1,257,952	2.47	996	1.11	0.45
Sum	**57,362,234**	**100.00**	**89,659**	**100.00**	–

[1] Data basis: Table "entriescleaned.csv", filtering the birth years 1815 to 1870 as well as by provinces.

Source: GEDBAS and Fertig et al. in comparison.

Figure 5
Map of Germany (1871), considered areas in grey



2.3 Socioeconomic Representativeness

It is important to consider the possibility of a distorted social structure of the data when interpreting the GEDBAS records. People with a higher socioeconomic status may be more likely to have an interest in their family history, resulting in a disproportionate representation of people with higher-status occupations in the data. To assess the potential impact of this, a comparison was made with Prussia in 1867, as many of the provinces covered were part of Prussia at that time. Data from a census providing information on status and occupation, broken down by industry (Engel, 1869, pp. 108-177, 1871), was used for this purpose.

The occupational distributions found in GEDBAS were manually assigned to the corresponding industry sectors and compared with the corresponding figures from the Prussian census. The resulting matches show that the distribution of occupations in GEDBAS tends to align with that of the Prussian census (Table 2). Only GEDBAS data from Prussian provinces were used in the comparison, and a temporal restriction was also applied, including only persons born between 1800 and 1855 and excluding persons who died before 1867.

TABLE 2
Occupations of the Prussian statistics of 1867 (Engel, 1869, 1871) and an excerpt of the GEDBAS data in comparison

Occupations	Prussia		GEDBAS (Prussian people, born between 1800 and 1855)		Relation of the shares
	Absolute	Relative (%)	Absolute	Relative (%)	
1. Agriculture, livestock, viticulture, horticulture	11.497.411	48,76	610	36.1	0.74
2. Forestry and hunting	115.333	0,49	15	0.9	1.81
3. Fishing	53,243	0,23	29	1.7	7.45
4. Mining and metallurgy	628,284	2,67	54	3.2	1.20
5. Large and small industries incl. construction	5,438,243	23,06	476	28.1	1.22
6. Trade	830,494	3,52	75	4.4	1.26
7a. Land transport	433,825	1,84	45	2.7	1.45
7b. Water traffic	161,816	0,69	23	1.4	1.97
7c. Refreshment and accommodation	387,251	1,64	18	1.1	0.65
8. Personal services	2,097,192	8,89	120	7.1	0.80
9. Health care and sick service	87,386	0,37	20	1.2	3.19
10. Education and teaching	206.421	0,88	36	2.1	2.42
11. Arts, Literature, Press	60.658	0,26	17	1.0	3.86
12. Church and worship, the burial of the dead	95.444	0,40	3	0.2	0.44
13. Royal household and court	4.154	0,02	8	0.5	23.64
14. State administration	135.127	0,57	77	4.6	7.98
15. Justice	92.144	0,39	24	1.4	3.64
16. Army	291.716	1,24	24	1.4	1.14
17. Battle Fleet	3.482	0,01	0	0.0	–
18. Municipal and Corporate Administration	147.440	0,63	18	1.1	1.69
19. Persons not practicing a profession	812,668	3,45	0	0.0	–
Sum	23.579.732	100,00	1,692	100.00	–

Only the GEDBAS data of the Prussian provinces are used, only persons born between 1800 and 1855 are included, persons who died before 1867 are removed, overall 1,931 occupational statements but only 1,692 can be assigned to a category.

However, the data is underrepresented in agriculture (36 percent in GEDBAS compared to 49 percent in the Engel statistics), which may be due to agricultural occupations being recorded less frequently in the primary sources underlying the data and as a consequence these are less often found in GEDCOM files. Conversely, occupations in large and small industries, including construction, are slightly overrepresented (28 percent in GEDBAS compared to 23 percent in the statistics). It should be noted that the Prussian trade statistics are of limited informative value (Hoffmann, 2012), so there is a possibility that the figures in this statistics are not exactly correct. The deviation does not appear to be so strong that reweighting seems necessary. Nevertheless, the lack of farmers and its impact on the results should be briefly interpreted at this point.

The proportion of farmers can impact the outcomes, as Xie and Killewald highlight differing mobility patterns compared to non-farmers (2013, pp. 2, 12). However, the specific influence of the farmer's proportion on outcomes remains unclear. There is a potential underestimation of social mobility, as less movement between farmers and non-agricultural workers may be measured. Indeed, a significant portion of actual mobility occurred due to individuals leaving agriculture (Song et al., 2020, p. 255). If the data underrepresents the agricultural proportion, mobility may be underestimated. Conversely, overestimation can occur if the remaining data shows more upward mobility than the one of the general populations owing to greater occupational change options in non-agricultural professions. The exact offsetting effects cannot be determined precisely, introducing an uncertainty that persists as a source of measurement error. Nonetheless, given (1) the caution required with Prussian statistics and (2) the overall alignment of conditions, the data is considered as highly representative.

Another potential sampling bias in the GEDBAS data is the difficulty in tracing the life histories of individuals with high spatial mobility. Probably, these individuals had less landownership and a lower socioeconomic status, which may have made it more challenging to track their family histories. As a result, they, and their ances-

tors, may be underrepresented in the GEDBAS data. While this fact should be acknowledged, it cannot be precisely quantified in the data. It is unclear why an ascendancy-oriented genealogist did not dive deeper into the past at certain points, so the extent of this bias cannot be accurately determined.

As demonstrated, socioeconomic status is multifaceted. In GEDBAS, the socioeconomic status of individuals is primarily determined by their occupation. Status research in the 1960s focused on occupation and was subsequently criticized for its one-dimensionality (Hradil, 2005, p. 383f.). However, given that the data cannot be altered, this limitation must be acknowledged. Income or educational level, on the other hand, are often only determinable indirectly (and imprecisely). Nevertheless, education is often a prerequisite for certain occupations, while income is a consequence. While occupation alone cannot fully capture an individual's socioeconomic status, it remains a crucial and appropriate indicator.[4] Therefore, the occupational data is classified using HISCO and the Classification of Occupations 2010 (KldB). The latter is specifically designed for German-speaking countries. Further details about the classification systems can be found in the appendix (see "Classification of occupational data"). HISCO codes are linked to other scales: HISCLASS and HISCAM. HISCLASS can be used to examine class mobility, while HISCAM can be used to analyze occupational prestige. HIS-

[4] However, there are specific challenges in comparing occupational information over an extended period. Occupational profiles are not consistent over several hundred years: the content of the activities described by an occupational description can change over time (as well as between different places). An occupational description can, therefore, vary in different dimensions (occupational title, time, place), which can result in a different socioeconomic status for the same occupational description. The profession of elementary or secondary school teachers can serve as an example of this. Schüren notes that this profession changed in the 19th and 20th centuries, so he placed it in the lower middle class before 1920 but in the upper middle class afterwards (1989, p. 314). The valuation of the profession has, thus, changed. However, it remains open whether the changes would be relevant to such an extent that the relative comparison to other occupations would lead to a different result. In the following, this effect will be neglected.

CAM is widely used for mapping social status according to Clark, Cummins and Curtis (2022, p. 3, 2024, p. 42). KldB provides information about the skill level required for a particular occupation. Therefore, occupational data is utilized in different ways to indicate socioeconomic status. Finally, data available for analysis includes HISCAM, HISCLASS, and KldB variables, which were used to develop and model new parameters.

3. Modelling

Clark's method is based on the assumption that the status of grandparents and earlier generations does not provide independent information about the likely outcomes for their descendants (Clark et al., 2014, pp. 292–293). Previous generation status correlations only appear relevant because they offer further information about underlying competence, but ultimately, all determining information is contained within the genetic code of the parents (Clark et al., 2014, p. 119). Clark's gene-biological consideration led to the choice of AR(1) modeling, which has received criticism (see e.g. Long and Ferrie, 2018). AR(1) is not optimal for adequately describing multigenerational mobility (Long and Ferrie, 2018, pp. F442-F444; Hällsten and Kolk, 2021, p. 9; Häner and Schaltegger, 2021, p. 111). Nevertheless, the use of an AR(1) model is justified by the intention to reproduce the results of the surname studies (equation 1).

Socioeconomic status is denoted by y_t for an individual and y_{t-1} for the previous generation, with t representing generation. The dependence on the previous generation is denoted by β, with this variable being able to assume a value of 0 to 1. The error term is represented by u_t.

$$y_t = \beta y_{t-1} + u_t \tag{1}$$

This formula can only be used to calculate β over two generations. If several generations are integrated, an AR(n) model is required, which also generates several β_n. Since one β is to be

generated over several generations here, a methodological extension takes place. So, the coefficient β is estimated for each data set (in this case, patrilineal lineage) by evaluating how well the observed values y_t fit an AR(1) process for a given β, based on the transformation of residuals using the cumulative distribution function of a normal distribution (see equation 2).[5] The variable n denotes the number of individuals in a data set, which must be reduced by 1 to account for the relevant number of generation changes. The expected value and standard deviation of y are denoted by μ and σ, respectively. β is varied from 0 to 1, and the residual sum $S(\beta)$ measures the average deviation of the normalized residuals from the center of the normal distribution. A lower $S(\beta)$ indicates a better fit of the AR(1) model to the data with the given β. A more detailed derivation of this approach is provided in the appendix (see "Derivation of β").

$$S(\beta) = \frac{1}{n-1} \sum_{t=2}^{n} \sqrt{\left(2\Phi\left(Z_t\right) - 1\right)^2} \tag{2}$$

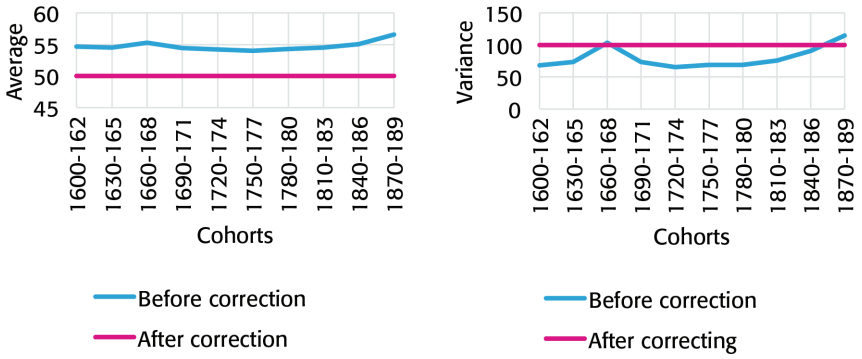
The coefficient $\hat{\beta}$ is determined as the value that minimizes the residual sum $S(\beta)$, reflecting the best alignment between the observed data and the assumed AR(1) model. To calculate an overall $\bar{\hat{\beta}}$, the individual (unweighted) $\hat{\beta}$ values from multiple data sets can be combined, for instance by clustering families regionally and computing the arithmetic mean across the resulting $\hat{\beta}$ values. This is the first indicator presented and can only be used for HISCAM time series.

The arithmetic mean of the HISCAM values within the cohorts (Figure 6, left) indicates a general small downward mobility trend in the data up to the 1750-1779 cohort, followed by periods of stronger upward mobility. However, these fluctuations pose a challenge for calculating β as stationary data is required. The underlying model assumes a constant expected value of 50 for HISCAM scores, indicating bias in the data that must be corrected. This bias is addressed by mean value correction per cohort, adjusting individual

[5] This also rests on the assumption that status is normally distributed.

HISCAM values. This process is explained in the appendix (see "HISCAM values"). Moreover, variance (Figure 6, right) is also a relevant factor for multigenerational HISCAM analysis, requiring balancing out fluctuations. The model requires a constant variance of 100, given the standard deviation of 10. To this end, both mean and variance corrections are carried out, with necessary values provided in the appendix (Table 11). Following the correction, the mean remains at 50, and variance at 100, resulting in a stationary data after a unit root test (Table 12 and Table 13, in the appendix), a prerequisite for the next section.

Figure 6
Average and variance of HISCAM across all data sets by cohort



Besides the coefficient β, there are other indicators that can help draw conclusions about stability. Secondly, a low standard deviation of y within a family ($\sigma_{y_{family}}$) indicates high persistence, which means the values remain the same or similar across generations. Thirdly, the arithmetic mean of all values ($\overline{y_{family}}$) is also an indicator of mobility / stability: if this mean value is higher or lower than expected (i.e. the general average value), the socioeconomic status is intergenerationally persistent. However, if the mean of a family is close to the expected value, this indicates mobility.

Only the variable HISCAM is suitable for determining β, σ or $\overline{y}$, and it is standardized to a level from –1 to 1 using equation 3. This transformation adjusts the normal distribution parameters from before ($\sigma = 10$, $\mu = 50$)[6] to after ($\sigma = 0.2020202020$, $\mu = 0$).

$$y_{after} = \frac{\left(y_{before} - 50\right) \bullet 2}{99} \tag{3}$$

For HISCLASS and KldB 2010, the method described earlier for determining intergenerational persistence is not suitable due to differences in the scale levels. However, changes in social class can still be used as an indicator of mobility. The proportion of individuals who have experienced a change in social class is calculated by dividing the number of class changes by the number of generation changes (van Leeuwen et al., 2016, p. 601). Fourthly, a change in social class is recorded when a child reaches an occupational class different from that of their father in the case of HISCLASS. Fifthly, KldB 2010 measures the level of qualifications required for an occupation. The relative frequency of class changes (m) is used as an indicator of the persistence or mobility of a family (see equation 4). The number of generations in a data set is denoted by n, while c represents the number of changes in social class or demand levels. Higher values of m indicate lower stability.

$$m = \frac{c}{n-1} \tag{4}$$

So, this chapter describes the five indicators of high persistence of socioeconomic status that can be identified across multiple generations in individual data sets. The first indicator is when the estimated $\hat{\beta}$ is close to 1. The second indicator is a mean value of a family ($\overline{y_{family}}$) that deviates from the mean HISCAM value of the respective

[6] There are different figures reported for the standard deviation of HISCAM, and indeed both values – 15 (van Leeuwen et al., 2013; Clark, Cummins and Curtis, 2024, p. 43) and 10 (Lambert et al., 2013, p. 78; Rosenbaum-Feldbrügge, 2019, p. 1833; Baranyi et al., 2023, p. 4; Dalman, 2025, p. 17) – are used in recent research. In this study, a standard deviation of 10 is chosen because it appears to be used more frequently.

province, while the third indicator is a low value of the standard deviation ($\sigma_{y_{family}}$) if it is less than the standard deviation of the total data ($\sigma_{y_{all}}$). The first three indicators are related to HISCAM. A high persistence in HISCLASS is expressed as a low proportion of class changes in a family ($m_{hisclass}$, fourth indicator), while a high persistence in KldB 2010 is expressed as a low proportion at the requirement level (m_{kldb}, fifth indicator). If a person has multiple occupational data, the respective average is determined for HISCAM and rounded to the nearest whole number. For the other variables, the first mentioned occupation is used, assuming the more relevant occupation is named first.

The study includes all cohorts simultaneously in the determination of values, making it a longitudinal observation. While none of the five indicators can be tracked over time, they can be assessed based on the regional breakdown or the length of the data sets examined (i.e., number of generations/cohorts). To gain insight into the extent to which the indicators depend on the number of connected generations studied, the data sets are grouped according to length and the indicators are calculated separately for each group. Additionally, a regional breakdown is performed to determine if a differentiated development of the regions is reflected in the intergenerational persistence of socioeconomic status. The regional clustering is carried out based on the historical provinces of Germany within the borders of 1870. For instance, earlier industrialization processes in a province could have led to earlier mobility and influenced persistence over a longer period.

4. Result and Interpretation

The results chapter is divided into three sections "Multigenerational Mobility," "Spatial Differences" and "Temporal Differences," each beginning with stylized facts that summarize the statement of the section. To perform the multigenerational analysis, the previous section's five indicators are computed.

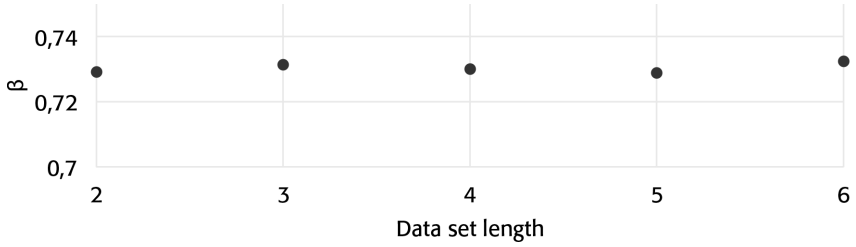
4.1 Multigenerational Mobility

Stylized fact # 1: Multigenerational mobility is at a low level and contrary to the hypothesis, is stable the more connected generations are considered.

First indicator β

The multigenerational persistence is observed in β, as shown in Figure 7. The average value of β across all data is 0.73, which is roughly in line with the value of 0.75 predicted by Clark and Cummins. Interestingly, the value of β is stable at a length of six generations, indicating no greater persistence with several contiguous generation sequences. Based on this variable, it can be concluded that analyzing many generations does not offer any decisive added value compared to analyzing two generations.

Figure 7
Comparison of HISCAM beta and the length of the data sets

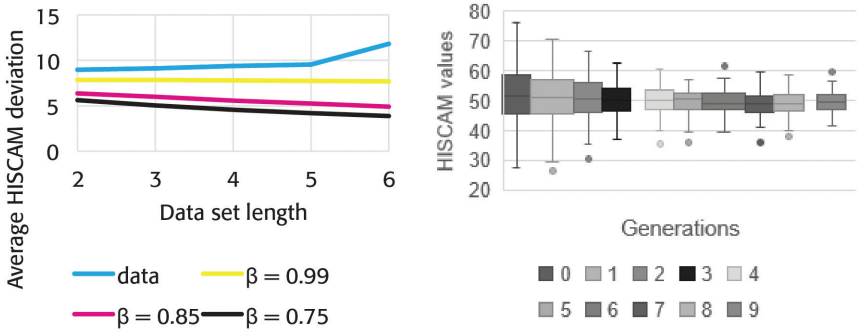


Second indicator $|\bar{y}|$

Initially, considering the second HISCAM indicator $\bar{y}$ with all data would lead to an average of 0 since positive and negative values cancel each other out. However, when considering the amount, the average decreases slightly over time. The decreasing trend is due to the regression towards the mean and is shown in Figure 8 (right). Families are unlikely to hold a very high or low status over many generations, even with a β of 0.75. Individual values rarely even make it beyond the $\bar{y} \pm \sigma$ limit (here 40 to 60) after a few generations.

Figure 8 (left) shows that the measured decreasing trend is initially due to the regression towards the mean. It is noticeable that the GEDBAS data is above the values simulated at a β of 0.99, which at first sounds unintuitive. This is due to a disproportionately large number of high values in the HISCAM distribution, which lead to a higher deviation from the mean. In fact, we see here that successor generations not only maintain the trend (both positive and negative) but expand it: starting advantages/disadvantages lead to further advantages or disadvantages.

Figure 8
Development of $|\bar{y}|$



In relation to the length of the data sets as well as the simulations for $\beta = 0.99$, $\beta = 0.85$ and $\beta = 0.75$, right: regression to the mean for 100 normally distributed random variables in an AR(1) process.

Third indicator $\bar{\sigma}$

The described standard deviation of the data sets shows a degressive trend as the length of the data sets increases, as observed in Figure 9. An absolutely persistent society would have a standard deviation tending towards 0, and an absolutely mobile society towards 10. To interpret this finding meaningfully, it needs to be compared to simulated data with differing values of β. The AR(1) process with a $\hat{\beta}$ of 0.75 and 0.50 and the values from Figure 7 (approx.) are compared to the data. The values of the 0.75 line are very similar to the approx. line because of similar values. The data shows greater

Figure 9
Ratio of standard deviation and the length of the data sets

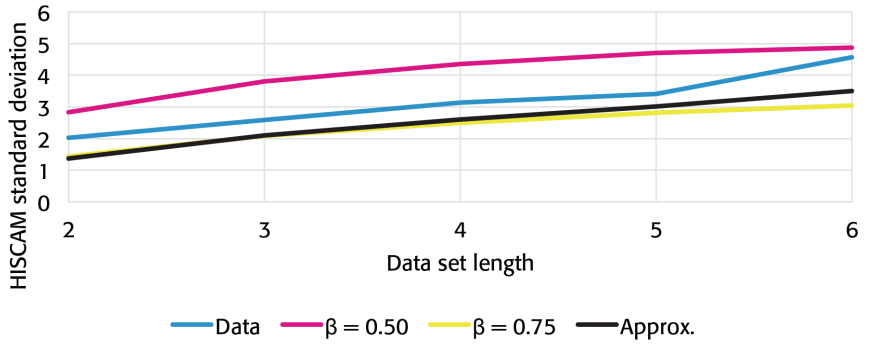
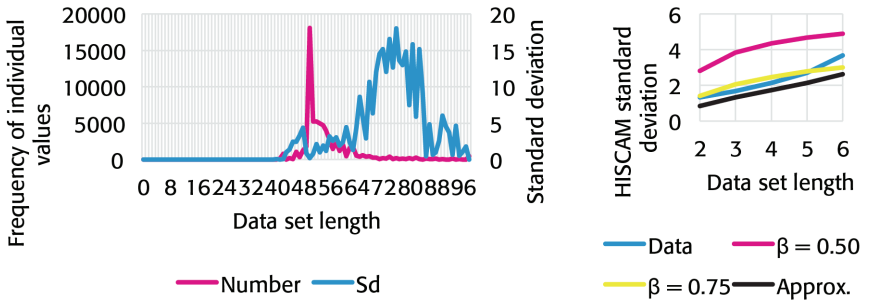


Figure 10



Left: Comparison of the percentage frequency of expressions and the standard deviation of the data sets on the HISCAM scale (average of the individual data sets), right: Development of the standard deviation according to the length of the data set disregarding the HISCAM values below 40 and above 60.

mobility than the data simulated at 0.75, but less than the 0.50 line. Apart from the value regarding six generations, which shows a tendency towards a mobile value, there is no clear change over the number of generations analyzed.

In analyzing the data, a unique feature is observed in the disproportionate frequency of higher HISCAM values. Figure 10 (left) displays the distribution of values on the HISCAM scale compared to the respective standard deviations per HISCAM value. Notably, the

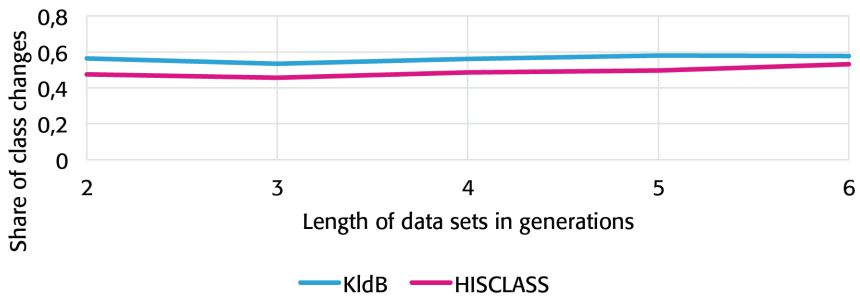
standard deviation is low at the point featuring the densest distribution (around 50), indicating a high persistence in the broad population. However, a high standard deviation is found for the range between 65 and 90, primarily due to the possibility of large jumps across different occupations on the scale. By removing the distortions from the values above 60 and only considering the values between 40 and 60, a higher multigenerational persistence is observed compared to Figure 10 (right). In contrast to the previous figure, we see a stability greater than 0.75, which decreases with increasing length of the data records – i.e. contrary to the hypothesis of increasing stability.

Fourth indicator $m_{hisclass}$ and fifth indicator m_{kldb}

Then, the number of HISCLASS class changes in the individual data sets is considered. A change of class here describes the fact that a son has a different characteristic from his father. Figure 11 shows, too, that the values are relatively stable over the observation of different numbers of generations. Between 2 and 3 generations a slight decrease can be recognized (more stability); between 3 and 6 generations a slight increase (more mobility).

However, this picture can be distorted if the data sets are not equally distributed in time. For instance, if the two-generational data sets were predominantly from a more recent period in which mobility is structurally higher, this would lead to a bias in the analysis. To ac-

Figure 11
Share of class/requirement level changes per generation regarding the length of data sets



count for this, the temporal distribution of the data sets is considered. With a completely independent temporal distribution for each additional generation in the length of the data set, the average cohort of the data contained in this group would become larger by half a cohort (15 years), since a longer data set would tend to have a later average cohort. Thus, from two generations to six generations, an increase of two cohorts (60 years) is expected. Figure 12 (left) shows that the difference is slightly less than two cohorts (~54 years), indicating only a slight temporal bias, which can be neglected further on.

Since the short data sets are more flexible in their temporal classification, it is also expected that the standard deviation of the average cohort is larger and decreases with increasing length. This can be observed in the data: the standard deviation of the average cohort decreases to an expected degree with increasing length (see Figure 12, right). Therefore, higher persistence can be measured here when many generations are considered together.

In Table 3, correlations between the calculated indicators are presented, and several noteworthy correlations can be observed. For instance, a low value of $\hat{\beta}$ (suggesting high mobility or low persistence) is positively correlated with a high standard deviation (suggesting high mobility or low persistence). However, only the correlations of HISCLASS/KldB exceed a value of 0.50, indicating that there are no very strong correlations between the indicators, so that the various parameters illuminate mobility from different angles.

Figure 12
Relationship between the average (left) and standard deviation (right) of the average cohort of data sets and the length of the data sets

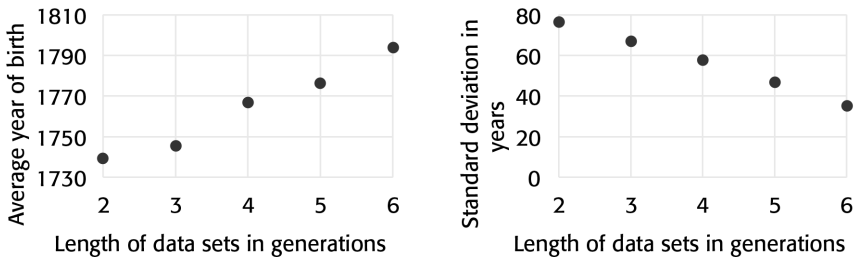


Table 3
Correlation of persistence indicators across all data sets, n = 4,508 (valid values for all parameters)

	β	$\lvert\bar{y}\rvert$	$\bar{\sigma}$	$\overline{HISCLASS_{rel}}$	$\overline{KldB_{rel}}$
β	1.00	-0.11	-0.48	-0,43	-0,30
$\lvert\bar{y}\rvert$		1.00	0.39	0.11	0.10
$\bar{\sigma}$			1.00	0.45	0,32
$\overline{HISCLASS_{rel}}$				1.00	0.53
$\overline{KldB_{rel}}$					1.00

4.2 Spatial Differences

Stylized fact # 2: Multigenerational mobility is different in different regions and can therefore be influenced by the environment.

To examine the influence of "Clark's Law" on the indicators, a regionally differentiated analysis is required, too. Table 4 displays the five indicators of multigenerational persistence for eleven provinces, ranked from most mobile to most persistent (average of the ranks of the five indicators in the bold column). The provinces of the former Hesse and the Thuringian states are the most mobile, while the Prussian Kingdom of Hanover is characterized by higher multigenerational stability. The regional differences in the mean values of the individual parameters are predominantly significant, as shown in the appendix's Table 14.

The multigenerational parameters for socioeconomic mobility reflect a long-term mobility practice in the respective regions. It is conceivable that this practice is associated with other developments, which could become apparent through a comparison of socioeconomic variables at the end of the observation period. To achieve this, the identified parameters are compared with various socioeconomic variables. The "Deutsche Reichsstatistik" (Statistics of the Germann Empire) of the 1870s serve as the basis for this comparison.[7] The un-

[7] Available online at http://www.digitalereichsstatistik.de/. A comparison of current socioeconomic variables at the level of today's federal states is not pursued. Between the end of the observation period and the beginning of the 21st century, so much has transpired that calculating correlations between present socioeconomic parameters and mobility a few hundred years ago seems speculative.

Table 4
Multigenerational persistence of socioeconomic status and temporal distribution by province

Province	Number of data sets	Parameters of multigenerational mobility					Rank of the parameters (⌀)
		HISCAM			KldB stability	HISCLASS stability	
		$\hat{\beta}_{province}$	$\lvert\bar{y}_{all} - \bar{y}^{province}\rvert$	$\sigma_{y_{province}}$			
B11 Hesse	3.354	0.68	2.66	3.86	0.58	0.64	3.2
B24 Thuringian States	1.308	0.65	2.17	4.56	0.53	0.64	4.0
A13 Rhine Province	9.061	0.72	1.58	4.44	0.47	0.60	4.6
A11 Province of Westphalia	3.425	0.70	1.93	4.55	0.46	0.55	5.4
A09 Province of Saxony	2.217	0.71	2.66	4.27	0.48	0.61	5.8
B19 Kingdom of Bavaria	8.216	0.74	1.62	4.57	0.50	0.54	6.0
B21 Kingdom of Saxony	2.667	0.72	1.89	4.57	0.54	0.51	6.0
B22 Electorate of Hesse	1.478	0.73	2.02	3.74	0.51	0.49	6.2
B18 Kingdom of Württemberg	3.660	0.74	1.77	4.66	0.49	0.51	7.4
B10 Grand Duchy of Baden	7.104	0.71	2.45	4.57	0.44	0.49	8.4
B20 Kingdom of Hanover	3.004	0.75	1.99	4.73	0.38	0.54	9.0

derlying data of the Reichsstatistik can be found in the appendix (see Table 6 to Table 9)[8] and are compared with the mobility rankings determined in Table 4 for historical provinces. The lower the value of the ranking, the more mobile the families are in the observed area in the long term.

In the study by Güell et al., high social mobility correlates with variables of economic prosperity (2018, pp. F364-F365). Measures of economic prosperity in the Reichsstatistik include figures on employees in industrial enterprises ("Gewerbebetrieb").[9] The percent-

[8] In the captions of the tables in the appendix, the source references for the respective volumes of the Deutschen Reichsstatistik are also included.
[9] As "Gewerbebetrieb", the following sectors are classified: ornamental and commercial horticulture, fisheries, mining, metallurgy and salt extraction, industry of stones and earth, metal processing, machinery, tools, chemical industry, industry of heating and lighting materials, textile industry, paper and leather, industry of wood and carving

age of industrially employed individuals in the considered regions in 1875 ranges from 25 percent (Kingdom of Saxony) to 13 percent (Kingdom of Hanover). It correlates negatively (-0.49) with the mobility rank (see Table 5). More mobility therefore goes hand in hand with a higher proportion of industrial employees – but here the proportion of industrial employees (and the faster deagrarianization) could also have led to more mobility. The same picture emerges when dividing enterprises by the number of employees. Regions with a high proportion of enterprises with more than five assistants exhibit a more mobile society (-0.36). The opposite is observed for the share of enterprises with less than five assistants (0.36). Regions with larger industrial companies tend to be more mobile.

High tax revenues have only a small correlation of 0.15 to the rank. At that time, there were no more suitable central revenue sources in the German Empire for this comparison, primarily relying on consumption taxes.

Table 5
Overview of the correlation between the rank and various socioeconomic parameters

Year	1875			1874		1873			1875	1873	
	A	B	C	D	E	F	G	H	I	J	K
	Share of industrial employed individuals	# of enterprises with ≤5 employees / # enterp. empl.	# of enterprises with > 5 employees / # enterp. empl.	Number of employees per enterprise	Customs revenue per inhabitant	Share of urban population	Share of rural population	Population density per km²	Share of illiteracy	Share of Protestant population	Share of Catholic population
Rank	-0.49	-0.31	-0.45	-0.37	0.15	-0.28	0.32	-0.35	-0.02	-0.02	0.01

materials, food and beverages, clothing and cleaning, construction industry, polygraphy enterprises, artistic enterprises for commercial purposes, commercial enterprises, transportation enterprises, accommodation, and refreshment (Kaiserliches Statistisches Amt, 1879, p. 518).

In addition, there are further socioeconomic variables that can be compared with social mobility. A positive correlation (-0.29) is observed in the proportion of the population living in cities. Mobility therefore happens more often in cities. In the context of the Reichsstatistik, these are places with more than 2,000 inhabitants. Also, population density per square kilometer correlates in the same order of magnitude with the rank (-0.35). To make a more valid statement here, a subdivision of the data into urban/rural categories would be desirable.

The education of the enlisted personnel in the German Army and Navy can serve as an indicator of the quality of the educational system in different regions. Better education might be associated with higher mobility, as individuals are intellectually more capable of pursuing activities based on their interests and abilities. Indeed, the percentage of enlisted soldiers and sailors in 1875/76 who are not literate (ranging from 0.00 percent in the Thuringian States to 1.79 percent in Bavaria) almost does not correlate with the rank (-0.02). Noticeably, there is a strong correlation between the proportion of Catholics and the proportion of individuals with illiteracy (0.58, not shown in the table), and vice versa for Protestants as Becker and Woessmann have already shown (2009).

Religious affiliation itself and the moral values conveyed therein can also be related to social mobility. But here, too, the rank does not correlate with the proportion of the Protestant population (-0.02) and positively with the proportion of Catholics (0.01).[10]

In summary, the data do show a connection between social mobility and the proportion of industrially employed individuals, as

[10] However, since many regions have a mixed religious composition, a denominational separation of the data would be more suitable and provides grounds for further studies. In subsequent research, the influence of religious denomination on social mobility could be further examined. For instance, a higher number of children in Catholic families can be observed (Westoff and Jones, 1979, p. 209; Eggen and Rupp, 2007, p. 9; Peri-Rotem, 2016, p. 234). If there are more children, there is a higher risk of social descent (Van Bavel et al., 2011, pp. 314, 338), making the number of children a driver of mobility. There may be additional influences on social mobility that lead to contrary results.

well as between the proportion of urban/rural populations. Correlations arise in comparison of economic structure: bigger enterprises are more prevalent in regions with higher mobility. However, these comparisons provide fewer explanations for the differences and lead to new hypotheses for further research. To examine the effects of individual exogenous shocks or trends (such as local conflicts, introduction of compulsory education, introduction of economic freedom, etc.), a more detailed temporal and spatial analysis is necessary. This is also the case since many effects are presumably overlapping at the present time.

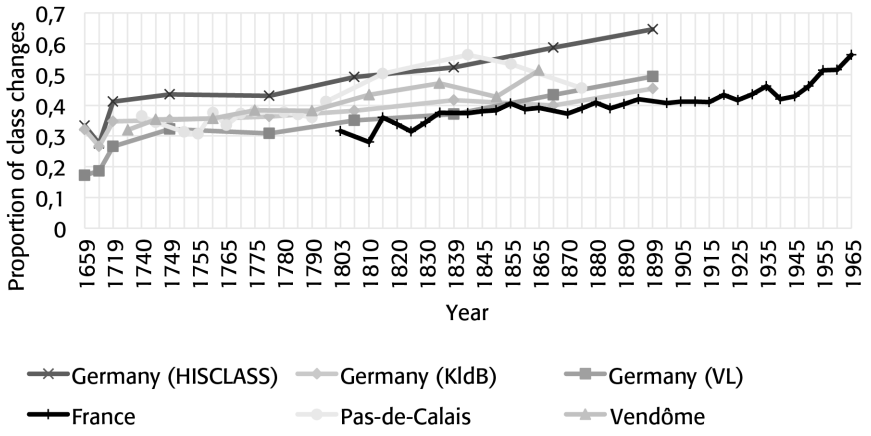
4.3 Temporal Differences

Stylized fact # 3: Mobility changes over time and can therefore be influenced by the environment.

Delving finally into the temporal differences, Figure 13 shows that there has been a significant increase in the two-generational HISCLASS mobility. For instance, while in the first cohort, over 70 percent of children's occupations matched those of their fathers, in the last cohort, almost 70 percent of children had occupations that differed from their fathers'. Similar trends have been observed in international studies such as van Leeuwen et al.'s research on France (2016, p. 601). However, it should be noted that different scales can yield different results, as demonstrated by the different outcomes when comparing individual French regions, as shown by Pas-de-Calais and Vendôme. Attempts to reproduce the structure of van Leeuwen's classes (called "VL", combine the HISCLASS classes 1-5, 6-7, 8, 9 and 11 as well as 11 and 12 into five classes) reveal similar values for Germany and France in the 19th century. Additionally, it is important to note that van Leeuwen et al. use occupations at the time of marriage, while the GEDBAS values used in this analysis refer to the time of birth and have been adjusted accordingly.

To better explain the social change observed, a more detailed analysis is needed. Therefore, it is important to understand the structural mobility in the data. For this purpose, it is important to note

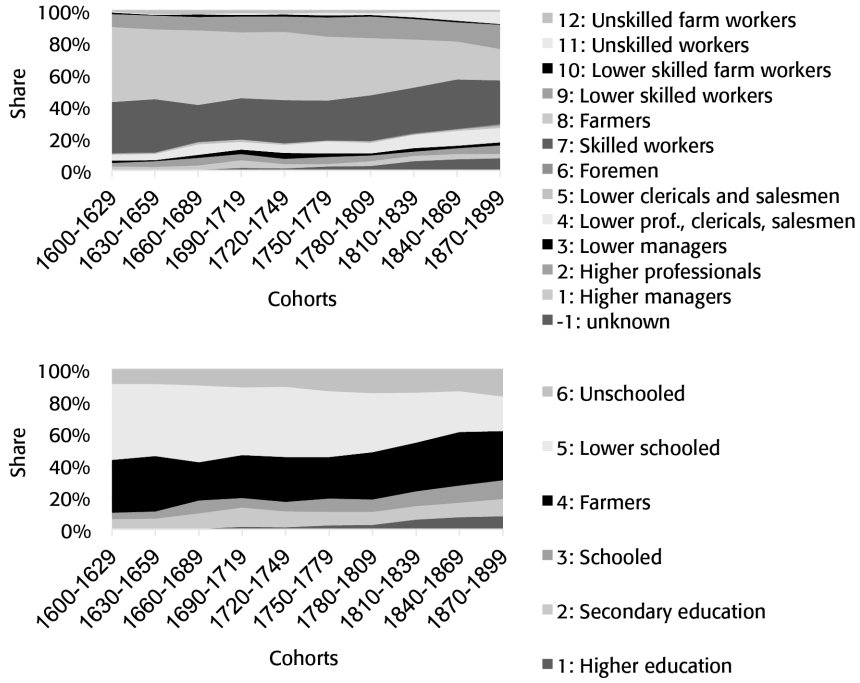
Figure 13
Class mobility over time across all data sets, own representation in conjunction with van Leeuwen et al. (2016, pp. 601, 609-610)



that the data in GEDBAS is skewed towards the industrial sector, so the trends shown represent shifts rather than actual shares at a specific point in time. However, across all cohorts, there is a clear trend away from agricultural production (see Figure 14, top) and a corresponding decrease in the proportion of the "farmers" class (8). Such a change in the economic structure leads to a shift in social classes. From the 1780-1809 cohort onwards, there is an increasing proportion of higher socioeconomic status occupations, indicating a trend towards more senior positions. However, there is also an increase in unskilled workers (11, 12) during this period, leading to a stronger polarization. To make the correlations clearer, individual HISCLASS categories are grouped into "occupational classes" in Figure 14 (bottom).[11] The KldB requirement level is not shown here, as the low number of variables means that around 80 percent of the people are in the group of "specialized activities."

[11] 1: higher education, HISCLASS 1/2/3, 2: secondary education, HISCLASS 4/5, 3: schooled, HISCLASS 6/7, 4: farmers, HISCLASS 8, 5: lower schooled, HISCLASS 9/10, 6: unschooled, HISCLASS 11/12.

Figure 14
Proportions of occupations in the population by cohort, HISCLASS on the top, the summarized "occupational classes" on the bottom



In summary, mobility has also changed over time and during the observation period it has increased. Throughout the entire period, structural mobility can be observed, especially away from agricultural professions. However, this is not decisive in explaining the increase in mobility over time – it happens mainly very slowly. Thus, a change is noticeable over time – as it was already the case with the regions. Social mobility is therefore not the same across all times and regions and thus can be altered by policy and does not follow any law of social mobility that passes status as a biological trait.

Conclusion

In this work, we have examined the multi-generational persistence

of the socioeconomic status and have found out that measured mobility decreases as more generations are considered, contradicting the notion that regression to the mean occurs after only three generations and that the advantages and disadvantages of ancestors become irrelevant (Becker and Tomes, 1986, p. S1). Our results support Clark's postulation of the magnitude of socioeconomic status persistence, but his "law" of social mobility, which suggests that social status is inherited like a biological trait (Clark et al., 2014, p. 9), cannot be confirmed. Spatial and temporal differences in social mobility levels suggest that various factors affect social mobility, and policy measures can be used to influence it.

The results are based on a newly introduced methodology, addressing Mare's call for new methods to analyze intergenerational family structures (2011, p. 2). Scraped genealogical databases from the Internet are used to be able to look at as many successive generations as possible. That also demonstrates the value of privately created genealogical data sets for scientific research and how they can be used. Further research in social and economic history and demographic studies can benefit greatly from analyzing additional genealogical databases. The GEDBAS data utilized in this study presents opportunities for investigating additional research questions. In future studies, it is recommended to incorporate additional indicators of socioeconomic status beyond occupation. While occupation is a primary focus of this study, it is acknowledged that it is not the only indicator of socioeconomic status and that classification systems used are not sufficient. Therefore, additional data is needed with other indicators of socioeconomic status, such as tax payments, assets, land ownership, and wills. It is worth highlighting that this is an ongoing line of research (e.g. Barone and Mocetti, 2021). These data sets can be generated for smaller geographic units, such as individual cities or counties, as well as for different time periods. Temporal and spatial differentiation would enable the study of interesting issues, such as the evolution of socioeconomic mobility over time. Genealogical databases hold considerable potential to provide the necessary data for such research.

6. Appendix

6.1 Indexing GEDBAS data

On November 1st, 2024, a total of 3,228 GEDCOM files were scraped from the GEDBAS database. The available data underwent a preprocessing step, which included cleansing and reduction of the data, particularly for location, occupation, and date information. The occupational and location data was processed according to the procedure described by Goldberg and Moeller (2022), while the date information was converted from GEDCOM-specific notation to an intuitively readable date format.

However, not all the cleaned data sets were useful for the investigation. Files with missing occupational data were disregarded. Additionally, files with a ratio of occupational information to persons of less than 10 percent or that were identical to other files were eliminated. Two files were considered identical if the number of rows, persons, families, places, and occupations matched. To identify and remove duplicate uploaded GEDCOM files, they were examined using this method. In cases of partial matches between files, further reduction occurred after the transformation of the data, which is described below.[12]

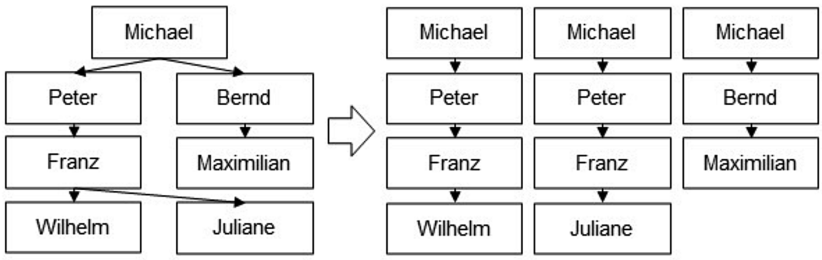
The basic structure of a panel data set, also known as a "data set," results from a patrilinear linkage across the generations. The data's graph structure in a tree format is unsuitable for classical modelling and must be transformed into a more suitable format. In this study, a patrilinear approach was used to structure the data. The patrilineal structure is justified in the section "Structure of the genealogical data" (also in the appendix). To structure the data in this way, individual trees were split up, resulting in one panel data set

[12] The importance of this aspect can be illustrated by an example: Johann Henrich Klingelhöffer (born 1670 in Marburg) appears in the three files. As a pastor, he also has a high socioeconomic status. In the further processing of this data, this duplication exponentiates itself so that it clearly distorts the analysis of the region of Westphalia (a pastor has HISCAM value of 98, moreover, he appears 38 times in the data due to 38 patrilineal descendants each).

per respondent in the lowest generation sequence (i.e., without their own children, see Figure 15). After cleaning the data, deleting irrelevant files, and structuring the data in this way, approximately 58,000 data sets remained.

The number of resulting data sets in this study increases linearly with the number of offspring without further children. Consequently, individuals with many children appear more frequently in the data than those with fewer descendants. This may lead to a disproportionate influence of patterns occurring in families with a particularly large number of offspring during the early period of observation. However, Clark and Cummins argue that the number of offspring in England (1780-1879) was random and only minimally dependent on the "quality" of the parents (Clark and Cummins, 2019, p. 2). On the other hand, Boberg-Fazlic, Sharp and Weisdorf found that socioeconomic status does influence family size in pre-industrial England (2011, pp. 384-385). Despite this, it is assumed that there is no difference. It is worth noting that most records in the data are ascendance-oriented, which has little impact on descendance-oriented records. Therefore, this effect is deemed negligible. A more detailed discussion of the impact of this assumption on the bias of the data can be found later in the appendix ("Effects of dividing genealogical structures into panel data sets").[13]

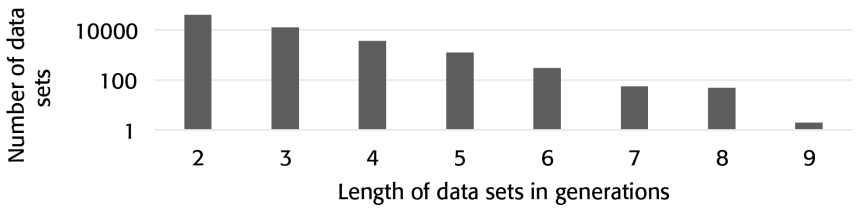
Figure 15
Transformation of the graph into individual time series



[13] Reference should also be made to Blanc's discussion on the preparation of such family tree data and possible restrictions (2024a, p. 5).

Adopting a patrilineal view in this study results in data sets of varying lengths. There is a regressive relationship between the length of the data sets and their frequency (see Figure 16). Some data sets consist of only one generation and are thus deleted, as neither an inter- nor a multigenerational analysis can be conducted with them.

Figure 16
Number of data sets by the length of data sets, logarithmic scaling



The period from 1600 to 1929 is divided into cohorts of 30 years, assuming that the average age of parents at the birth of a child is 30 years. Each cohort represents a generation. The oldest record is assigned to a cohort based on birth records, and if not available, the year of birth of the first child is used for estimation. Subsequent generations are assigned to other cohorts regardless of actual lifetime, resulting in a panel data set with a chronological sequence of data points.[14] However, the available data are incomplete and unbalanced. For instance, some occupational details may be missing, and there are no variables invariant in time as the data sets represent families with different individuals.

[14] ID: Number of the record; source: File name of the underlying GEDCOM file from GEDBAS; INDI: GEDBAS-internal identification number of a person; INDI_father: Identification number of the father of the INDI person; cohort: Clustering of birth years in ten cohorts of 30 years (1600 to 1899); birth_year: Year of birth given or calculated; death_year: Date of death given; age: Age at death; hisco: HISCO code; hisclass: HISCLASS category; hiscam: HISCAM value based on HISCO classification of occupation (1 to 99); kldb: KldB code; region_1820: Allocation of the place of birth (or another place in the life course, if the place of birth is not specified) to the administrative unit which the place had in 1820 (other years are also possible or different ones in the comparison).

6.2 Structure of the Genealogical Data

The GEDBAS data comprises interconnected networks of individuals (nodes) linked by their kinship relationships (edges). These resulting graphs can be analyzed from various perspectives, which must consider the structure of the underlying networks. Two primary approaches are the ascendance-oriented approach, which focuses on the ancestors of a test person, and the descendance-oriented approach, which examines the descendants. There can also be mixed variants, such as when kinship networks are analyzed.

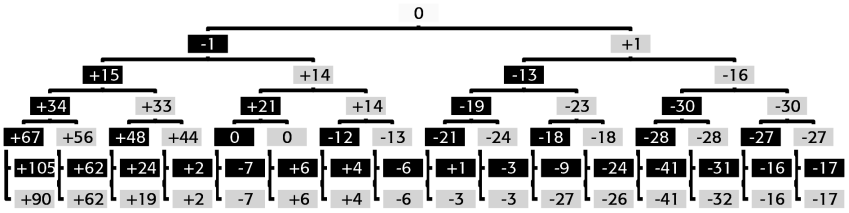
Although most GEDCOM files follow an ascendance-based approach in principle, this is often not strictly followed due to including information on siblings and other relatives. As a result, there are typically about four times more people without parents than without children in the files. For purely descent-oriented files, the ratio is theoretically less than 1 and approaches 0 as the volume increases. This poses challenges for descendance-oriented investigations, as distant cousins are less frequently included in a file. To analyze the ascendancy approach, it's important to specify which part of it is being considered. There are two main views: a comprehensive view in which all ancestors are included, and a patrilineal view that follows only the father's side of the family.

In the following section, we will adopt a patrilineal approach, as it is essential to reconstruct Clark's surname methodology's structure. The transmission of surnames during the period under consideration was predominantly patrilineal. This approach prompts us to ask which patrilineal correlations exist in the data and over how many generations. Another advantage of this approach is that it includes all those people who had no children or no known descendants today. The number of childless people was not insignificant, with perpetually single women accounting for 10 to 15 percent of the population according to the "European Marriage Pattern" (Dennison and Ogilvie, 2014, pp. 651-652). Additionally, some individuals were married but childless, and it is crucial to consider these individuals because childlessness could have social reasons such as mar-

riage restrictions. On the other hand, a dynastic approach based on the comparison of (distant) cousins (Adermon, Lindahl and Palme, 2021) does not consider this aspect.

However, it should be noted that the data often does not include a complete list of ancestral generations. As a result, information on patrilineal ancestors is included more frequently, resulting in an imbalance in favor of patrilineal links (as shown in Figure 17). This imbalance may be due to several reasons, such as the difficulty in obtaining information on women and their origins, or the fact that genealogists may be more interested in researching their surname origins, which is often associated with the patrilineal line. Nonetheless, it is important to acknowledge this bias and consider other sources of information to gain a more comprehensive understanding of the ancestral relationships in the data.

Figure 17
Distribution of ancestors in the ascendancy of persons without children in percent, the ascending person in light grey (above), males in black, females in grey, data in the appendix (see Table 10)

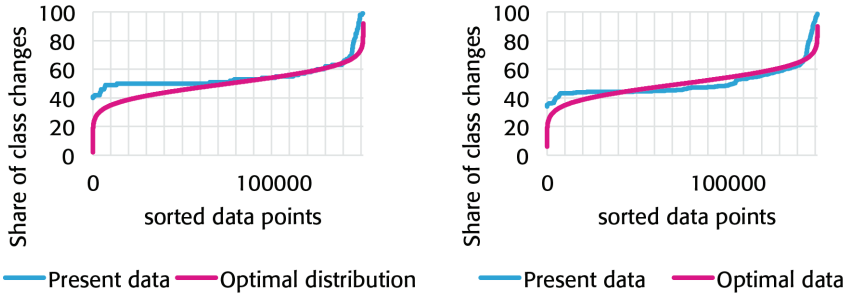


6.3 Classification of Occupational Data

Operationalization of occupational data is necessary for the study. The first thing to note is that there is a total of 379,956 occupations in the data (GEDCOM tag: OCCU) – duplicate designations not eliminated. Nevertheless, before a statement can be made about the socioeconomic status of an occupation, it is necessary to classify these occupations. Various methods exist for classifying historical occupations. In this paper, two methods are considered and applied.

One is HISCO, a system that claims to be internationally applicable (van Leeuwen and Maas, 2010, p. 432). In addition to the classification of the occupation, HISCO offers the possibility of a status dimension, which can contain information on "ownership, level of artisan career, the rank of an employee, level of tertiary education and status titles" (Zijdeman and Lambert, 2010, p. 113). The HISCAM prestige scale based on HISCO is also used (van Leeuwen and Maas, 2010, p. 433). The underlying assumption in creating the scale is that individuals interact more with each other the closer their social position is (Zijdeman and Lambert, 2010, p. 115). The result is a scale from 1 (lowest prestige) to 99 (highest prestige) (Lambert et al., 2013, p. 6). The digits below 40, however, are practically not used. Higher digits, on the other hand, occur disproportionately often. This circumstance is also evident in the data (see Figure 18, left). Approximately, HISCAM is normally distributed with a standard deviation of 10 and an expected value of 50 (Lambert et al., 2013, p. 81). After mean and variance correction, these values are also achieved (see Figure 18, right).

Figure 18
HISCAM distribution with available and optimal data



By utilizing the approximated distribution as the benchmark for the HISCAM distribution, a bias is introduced. In a completely mobile society, the likelihood of observing a HISCAM value exceeding 90 would be 0.001 percent. However, such high HISCAM values

occur far more frequently in the data. As a result, the increasing HISCAM values are unrealistic in a mobile society, suggesting the presence of a large deterministic component. This could potentially affect the estimation that the actual mobility is higher than the calculated mobility.

The HISCAM scales are divided into country-specific and universal scales, which are further categorized based on time and gender. In the case of Germany, due to the limited population data available, it is advised to use the universal scale in developing the HISCAM scale for the country (van Leeuwen et al., 2013). As per this recommendation, the universal scale is utilized in this study.

Each HISCO code is also associated with a HISCLASS (van Leeuwen and Maas, 2011), which is a nominal scale with twelve categories including higher managers, higher professionals, lower managers, lower professionals, clerical and sales personnel, lower clerical and sales personnel, foremen, medium-skilled workers, farmers, lower-skilled workers, lower-skilled farm workers, unskilled workers, and unskilled farm workers. However, many German-language job titles are not covered by HISCO, with only 1,306 German occupations being recorded as of April 22, 2020. For instance, occupations such as "Bürstenbinder" (brush binder), "Erbrichter" (hereditary judge), or "Büchsenmacher" (gunsmith) are not recognized by HISCO (International Institute of Social History, 2021). An extended list with 2,030 entries is used here, but it still does not include all the occupations that occur. Therefore, not all GEDBAS occupations can be classified using HISCO, and the data could be distorted if only certain occupational groups were classified. Moreover, there is no recognizable structure to the available occupations, and the recorded occupations appear to be random.

In contrast to HISCO, which is designed for international use, KldB 2010 is specifically tailored to the occupational structure of German-speaking countries (Paulus and Matthes, 2013, p. 5). Like HISCO, KldB uses a five-digit coding system. The first digit represents the occupational area, categorizing jobs into one of ten major sectors. The second digit indicates the occupational group within

this area, refining the classification into more specific fields. The third digit identifies the main occupational group within the occupational group, detailing the job functions further. The fourth digit designates the occupational subgroup, providing the most detailed level of classification within the first four digits. However, the last digit in KldB represents the required level of the occupation, ranging from helper/apprentice activities to highly complex activities (xxxx0 to xxxx4). Additionally, when a 9 is present in the fourth digit in combination with the five-digit code, it indicates the supervisory or managerial function of the occupation (xxx93 or xxx94). Although KldB 2010 is primarily designed for existing occupations in Germany (e.g. for labor market statistics, job placement and occupational research), historical occupational titles have been added by the Historical Data Centre Saxony-Anhalt. In this study, we use the database status from 27 May 2020, which includes 108,254 coded historical occupational variants. We employ the KldB classification as a supplementary variable to HISCO since it is specifically tailored to the German language area and provides a more comprehensive classification. However, neither HISCO nor KldB is perfectly suited to our purposes, as both have limitations in terms of job title, time, and place. Going forward, it will be an important task for social history to develop more comprehensive solutions in this area.

To classify the occupational data, we employed the algorithm developed by Goldberg and Moeller (2022). This algorithm first cleans the data and then utilizes various string-matching techniques to correct spelling errors, among other things. Our results indicate that ~50 percent of the occupational data can be classified according to HISCO and ~50 percent according to KldB. Additionally, ~30 percent of the data can be assigned to both classifications. The seemingly low recognition rate is due to the inclusion of many foreign-language occupational titles in GEDBAS, as well as many entries in the occupation field that deviate from the standard definition of an occupation. Despite these challenges, the Goldberg and Moeller algorithm has proven to be effective in efficiently classifying most of the occupational data. Nevertheless, if more occupational

details were recognized, even more data sets could be created over many generations.

6.4 Data from Deutsche Reichsstatistik

The Reichsstatistik distinguishes between the Grand Duchy of Hesse and the Province of Hesse-Nassau. However, these divisions are not precisely reflected in the table. The program assigns the region "B11 Hessen" to the GOV[15] object "object_218147," which corresponds to the Grand Duchy of Hesse. On the other hand, the state "B22 Electorate of Hesse" has no equivalent in the Reichsstatistik because it was absorbed into the Prussian province of Hesse-Nassau in 1868. Despite this, a significant portion of Hesse-Nassau province was previously the Electorate. Therefore, the Electorate is used as a counterpart to the Prussian province of Hesse-Nassau.

Table 6
Data from the Deutsche Reichsstatistik on population distribution (Kaiserliches Statistisches Amt, 1873c, pp. 136, 149, 154f., 158f., 1873a, p. 16)

Land	Resident population in 1871	N. of inhabitants per km²	N. of inhabitants in residential areas ≥ 2,000 inhabitants	Share of urban population	N. of inhabitants in residential areas < 2,000 inhabitants	Share of rural population
B18	1,818,533	93.20	497,858	0.27	1,320,681	0.73
B24	1,067,441	89.75	317,998	0.30	727,443	0.68
B10	1,461,562	97.00	483,029	0.33	978,522	0.67
B22	1,400,370	88.10	402,630	0.30	997,740	0.71
B11	852,894	111.10	303,290	0.36	549,604	0.64
B19	4,852,026	64.10	1,112,211	0.23	3,739,815	0.77
B21	2,556,244	170.50	1,263,553	0.49	1,292,691	0.51
A13	3,579,347	132.70	1,303,078	0.36	2,276,269	0.64
A11	1,775,175	87.90	480,961	0.27	1,294,214	0.73
A09	2,103,174	83.30	796,774	0.38	1,306,400	0.62
A20	1,961,437	51.00	438,088	0.22	1,523,348	0.78
Correlation	–	**-0.08**	–	**0.09**	–	**-0.04**

The number of residents in 1871 is used as a reference for calculating the proportions, and the correlation pertains to the ranking in Table 4 (bolded column).

[15] The GOV is the "Geschichtliches Ortsverzeichnis", a database of places mainly in Germany, which is used by the algorithm according to G87oldberg and Moeller.

TABLE 7
Data from the Deutsche Reichsstatistik on religious denomination
(Kaiserliches Statistisches Amt, 1873c, pp. 188c-188d)

Land	N. of Catholic inhabitants	Share of Catholic inhabitants	N. of Protestant inhabitants	Share of Protestant inhabitants
B18	553,542	0.30	1,248,860	0.69
B24	13,041	0.01	1,047,941	0.98
B10	942,560	0.64	491,008	0.36
B22	371,736	0.27	988,041	0.71
B11	238,080	0.28	585,399	0.69
B19	3,455,329	0.71	1,340,218	0.28
B21	53,642	0.02	2,493,556	0.98
A13	2,628,173	0.73	906,867	0.25
A11	949,118	0.53	806,464	0.45
A09	126,735	0.06	1,966,696	0.94
A20	233,631	0.12	1,711,728	0.87
Correlation	**–**	**0.10**	**–**	**-0.10**

The number of residents in 1871 is used as a reference for calculating the proportions, and the correlation pertains to the ranking in Table 4 (bolded column).

TABLE 8
Data from the Deutsche Reichsstatistik on industrial structure
(Kaiserliches Statistisches Amt, 1879, pp. 548-551)

Land	N. of enterprises	Share of enterprises with ≤ five employees	N. of small enterprises	Share of enterprises with > five employees	N. of bigger enterprises	N. of industrial employed	Share of industrial employed individuals	N. of employees per enterprise
B18	148,702	217,420	0.12	70,565	0.04	287,985	0.16	1.94
B24	87,138	129,399	0.12	61,372	0.06	190,771	0.18	2.19
B10	105,237	158,246	0.11	80,163	0.05	238,409	0.16	2.27
B22	107,278	145,927	0.10	73,537	0.05	219,464	0.16	2.04
B11	60,388	96,235	0.11	38,148	0.04	134,383	0.16	2.23
B19	547,925	547,925	0.11	159,526	0.03	707,451	0.15	1.29
B21	369,459	369,459	0.14	262,885	0.10	632,344	0.25	1.71
A13	401,721	401,721	0.11	321,258	0.09	722,979	0.20	1.80
A11	170,499	170,499	0.10	181,974	0.10	352,473	0.20	2.07
A09	208,499	208,499	0.10	139,140	0.07	347,639	0.17	1.67
A20	128,030	177,838	0.09	84,306	0.04	262,144	0.13	2.05
Correlation	**–**	**–**	**-0.48**	**–**	**0.32**	**–**	**0.03**	**0.36**

The number of residents in 1871 is used as a reference for calculating the proportions, and the correlation pertains to the ranking in Table 4 (bolded column).

Table 9
Data from the Deutsche Reichsstatistik on economic indicators (Kaiserliches Statistisches Amt, 1873b, pp. 78-93) and on the education of recruits (Kaiserliches Statistisches Amt, 1878, p. 95)

Land	Import and export duties	Customs revenue per inhabitant	Share of soldiers in the replacement year 1875/76 without formal education
B18	961,213	0.53	0.02
B24	276,709	0.26	0.00
B10	1,971,669	1.35	0.22
B22	1,890,477	1.35	0.53
B11	994,912	1.17	0.35
B19	2,176,354	0.45	1.79
B21	3,730,185	1.46	0.23
A13	7,498,648	2.10	0.74
A11	1,331,580	0.75	1.05
A09	1,524,641	0.72	0.32
A20	2,544,271	1.30	0.84
Correlation	**–**	**0.47**	**0.51**

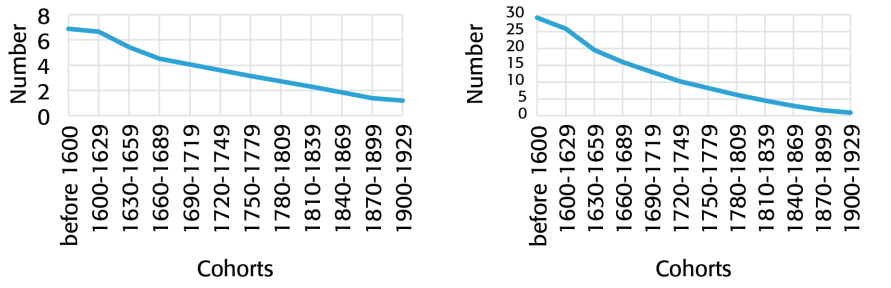
The number of residents in 1871 is used as a reference for calculating the proportions, and the correlation pertains to the ranking in Table 4 (bolded column).

6.5 Effects of Dividing Genealogical Structures into Panel Data Sets

Pedigrees are divided into individual patrilineal ancestral lines, which often results in fathers with many patrilineal descendants appearing disproportionately frequently. For instance, if a person appears on average 1.86 times in the cohort from 1840 to 1869, their frequency increases linearly to an average five-fold occurrence in the cohort from 1630 to 1659 (refer to Figure 19, left). However, the standard deviation of the frequency of individuals in the earlier cohorts increases to a level of approximately 30 (refer to Figure 19, right). This suggests a highly skewed distribution with outliers. One such example is John Andrews, a farmer born around 1619 in Essex County, England, who appears a staggering 3,182 times. This unusually high number is attributed to a detailed descent-oriented ge-

nealogical file. Despite the skewed distribution of frequencies, the median across all cohorts is only 1, underscoring the fact that most individuals appear only once in the data.[16]

Figure 19
Average occurrence of individuals by cohort (left) and standard deviation of this frequency (right)

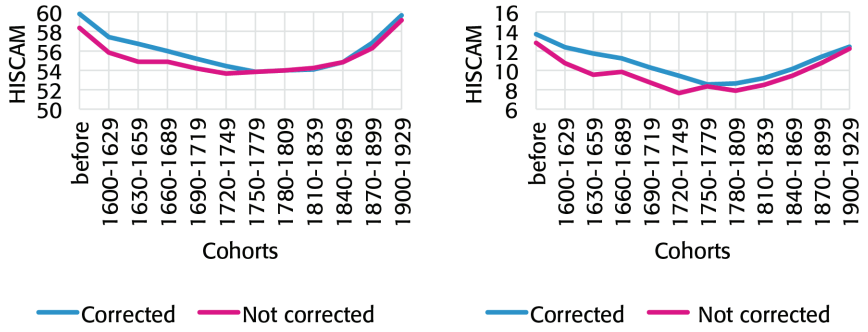


Considering the impact of the division on the HISCAM values across different cohorts, it is observed that a significant effect on the average can only be detected in the cohorts prior to 1750 (see Figure 20, left). In these cohorts, the average value of the shared data is approximately two points lower on the HISCAM scale, indicating a disproportionate presence of individuals with low-rated occupations.[17] It is worth noting that the standard deviation on the HISCAM scale is lower in almost all cohorts with bias, as expected (see Figure 20, right).

[16] At this point, it is worth taking a moment to think about who does not appear in the data at all. In addition to all women, these are primarily people without fathers, for example illegitimate sons. If they do not even have sons of their own, they are completely excluded from this view.

[17] The classification of certain occupations as "low-rated" is based on a comparison of the data within each temporal cohort, rather than a comparison to the broader population. It should be noted that this classification may not accurately reflect the occupational status of the general population during the period in question. It is possible that the data overrepresents socially advantaged individuals, particularly in the early cohorts, which may skew the average. Genealogical research tends to focus on higher-ranking individuals who are more likely to leave behind occupational details and other historical traces, leading to a potential bias in the data. Furthermore, the absence of data for the largest part of the population prior to 1600, coupled with a significant increase in average HISCAM values during that period, lends further support to this hypothesis.

Figure 20
Average of HISCAM average per cohort (left) and HISCAM standard deviation per cohort (right), with bias (without correction) in light, with bias (correction) removed in dark



The effect of bias seems to be within an acceptable range, so the temporal division of the data can still be used. However, it is crucial to take into account the bias when interpreting the data, especially in cohorts prior to 1810. It is important to note that in a region-specific analysis with limited data, chance may lead to a more significant bias. Therefore, it is crucial to exercise caution and consider the limitations of the data when drawing conclusions.

6.6 Derivation of β

Social mobility is traditionally modeled using an AR(1) process, as exemplary shown in studies by Clark et al. (2015) and Adermon, Lindahl, and Palme (2021, p. 1528).[18] In this model, the socioeconomic status y^t of generation t is expressed as a linear function of the status y_{t-1} of the parental generation.

[18] However, the AR(1) model is considered as too simplistic to fully capture the complex determinants of social mobility. For instance, it fails to account for a direct influence of the grandparent generation. Nevertheless, Clark et al. use this model to establish a "law of social mobility", assuming that an AR(1) process suffices if complete information about the parent generation is available (2014, pp. 293-294).

$$y_t = \alpha + \beta y_{t-1} + u_t \tag{5}$$

Here, β represents the intergenerational transmission coefficient, and u_t captures the stochastic component (i.e., the role of chance). Assuming that the variance of y_t and y_{t-1} is equal (σ_y^2), the residual standard deviation σ_u can be derived as:

$$\sigma_u = \sqrt{\sigma_y^2 \bullet \left(1 - \beta^2\right)} \tag{6}$$

To estimate β, a custom numerical approach is used that deviates from classical Ordinary Least Squares (OLS) or Maximum Likelihood Estimation (MLE). The core idea is to find the value of β for which the standardized residuals $u_t = y_t - y_{t-1}$ are as symmetrically centered around the expected value μ as possible, under the assumption of normality. For each candidate value of β in the interval [0, 1], the residuals are computed and standardized as:

$$z_t = \frac{u_t - \mu}{\sigma} \tag{7}$$

Then, the value z_t is evaluated using the cumulative distribution function (CDF) $\Phi(z_t)$ of the standard normal distribution. To account for both tails of the distribution symmetrically, the CDF is transformed as:

$$v_t = 2\Phi\left(z_t\right) - 1 \tag{8}$$

This transformation maps the distribution onto the interval [–1, 1], with zero at the center (i.e., when $u_t = \mu$). The absolute deviation from the center is then computed as:

$$\left|v_t\right| = \sqrt{\left(2\Phi\left(z_t\right) - 1\right)^2} \tag{9}$$

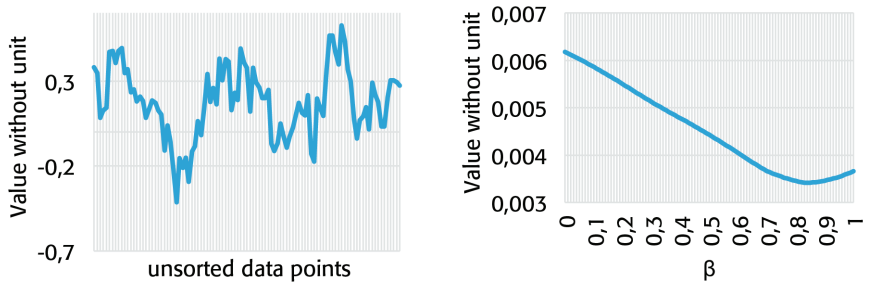
Finally, an average deviation function $S(\beta)$ is constructed by aggregating over all generations:

$$S\left(\beta\right) = \frac{1}{n-1} \sum_{t=2}^{n} \sqrt{\left(2\Phi\left(z_t\right) - 1\right)^2} \tag{10}$$

This function is minimized numerically by evaluating β on a fine grid (e.g., in steps of 0.01). The β that minimizes $S(\beta)$ is selected as

the best-fitting estimate $\hat{\beta}$. To do so, the function is evaluated on a fine grid of candidate values between 0.00 and 1.00. Figure 21 illustrates this process using a simulated time series of 100 values with a true $\beta = 0.80$. The right panel shows the resulting values of $S(\beta)$. In this illustrative example, the minimum of the residual function is found at $\beta = 0.85$, which is close to the used value.

Figure 21
Simulation of an AR(1) process with 101 random values (left) and estimation of β (right)



This method does not rely on closed-form likelihood expressions but instead operationalizes the assumption of normal residuals by leveraging the symmetry of the standard normal CDF. It thereby offers a robust, interpretable approach to estimating β in intergenerational mobility studies, even in the presence of non-classical error structures or when seeking robustness to extreme values.

To obtain a general estimate for $\hat{\beta}$, which can be applied to multiple data sets, $\hat{\beta}$ must be estimated for each individual data set. Once $\hat{\beta}$ has been estimated for all data sets, aggregation is necessary to draw a conclusion about all the data sets combined. The arithmetic mean is often used for this purpose, which is calculated by adding up all of the $\hat{\beta}$ values and dividing by the number of data sets (n). The resulting average value provides an indication of the typical value of β across all data sets.

6.7 Distribution of the ascendancy

TABLE 10
Distribution of the ascendancy in the GEDCOM files

Kekule	Number	Actual share	Proportion with equal distribution	Deviation in percent	Kekule	Number	Actual share	Proportion with equal distribution	Deviation in percent
1	985	1.000	1	0	65	90	0.046	0.03125	46
2	565	0.495	0.5	-1	66	76	0.039	0.03125	23
3	576	0.505	0.5	1	67	74	0.038	0.03125	20
4	386	0.286	0.25	15	68	64	0.032	0.03125	4
5	385	0.286	0.25	14	69	63	0.032	0.03125	2
6	293	0.217	0.25	-13	70	42	0.021	0.03125	-32
7	284	0.211	0.25	-16	71	42	0.021	0.03125	-32
8	253	0.168	0.125	34	72	29	0.015	0.03125	-53
9	251	0.166	0.125	33	73	28	0.014	0.03125	-55
10	228	0.151	0.125	21	74	30	0.015	0.03125	-51
11	216	0.143	0.125	14	75	30	0.015	0.03125	-51
12	152	0.101	0.125	-19	76	39	0.020	0.03125	-37
13	146	0.097	0.125	-23	77	39	0.020	0.03125	-37
14	132	0.087	0.125	-30	78	31	0.016	0.03125	-50
15	132	0.087	0.125	-30	79	31	0.016	0.03125	-50
16	184	0.104	0.0625	67	80	32	0.016	0.03125	-48
17	172	0.097	0.0625	56	81	30	0.015	0.03125	-51
18	163	0.092	0.0625	48	82	23	0.012	0.03125	-63
19	159	0.090	0.0625	44	83	23	0.012	0.03125	-63
20	110	0.062	0.0625	0	84	30	0.015	0.03125	-51
21	110	0.062	0.0625	0	85	25	0.013	0.03125	-59
22	97	0.055	0.0625	-12	86	23	0.012	0.03125	-63
23	96	0.054	0.0625	-13	87	22	0.011	0.03125	-64
24	87	0.049	0.0625	-21	88	28	0.014	0.03125	-55
25	84	0.048	0.0625	-24	89	28	0.014	0.03125	-55
26	91	0.052	0.0625	-18	90	27	0.014	0.03125	-56
27	91	0.052	0.0625	-18	91	27	0.014	0.03125	-56
28	80	0.045	0.0625	-28	92	22	0.011	0.03125	-64
29	80	0.045	0.0625	-28	93	19	0.010	0.03125	-69
30	81	0.046	0.0625	-27	94	30	0.015	0.03125	-51
31	81	0.046	0.0625	-27	95	24	0.012	0.03125	-61

(continued)

(continued)

Kekule	Number	Actual share	Proportion with equal distribution	Deviation in percent	Kekule	Number	Actual share	Proportion with equal distribution	Deviation in percent
32	124	0.062	0.03125	99	96	38	0.019	0.03125	-38
33	115	0.058	0.03125	84	97	38	0.019	0.03125	-38
34	98	0.049	0.03125	57	98	28	0.014	0.03125	-55
35	98	0.049	0.03125	57	99	28	0.014	0.03125	-55
36	75	0.038	0.03125	20	100	37	0.019	0.03125	-40
37	72	0.036	0.03125	15	101	37	0.019	0.03125	-40
38	62	0.031	0.03125	-1	102	30	0.015	0.03125	-51
39	62	0.031	0.03125	-1	103	30	0.015	0.03125	-51
40	56	0.028	0.03125	-10	104	15	0.008	0.03125	-76
41	56	0.028	0.03125	-10	105	15	0.008	0.03125	-76
42	64	0.032	0.03125	3	106	24	0.012	0.03125	-61
43	64	0.032	0.03125	3	107	22	0.011	0.03125	-64
44	63	0.032	0.03125	1	108	23	0.012	0.03125	-63
45	63	0.032	0.03125	1	109	23	0.012	0.03125	-63
46	57	0.029	0.03125	-9	110	16	0.008	0.03125	-74
47	57	0.029	0.03125	-9	111	16	0.008	0.03125	-74
48	61	0.031	0.03125	-2	112	24	0.012	0.03125	-61
49	59	0.030	0.03125	-5	113	24	0.012	0.03125	-61
50	59	0.030	0.03125	-5	114	20	0.010	0.03125	-68
51	59	0.030	0.03125	-5	115	20	0.010	0.03125	-68
52	55	0.028	0.03125	-12	116	19	0.010	0.03125	-69
53	44	0.022	0.03125	-29	117	19	0.010	0.03125	-69
54	46	0.023	0.03125	-26	118	15	0.008	0.03125	-76
55	45	0.023	0.03125	-28	119	15	0.008	0.03125	-76
56	36	0.018	0.03125	-42	120	40	0.020	0.03125	-35
57	36	0.018	0.03125	-42	121	39	0.020	0.03125	-37
58	42	0.021	0.03125	-33	122	20	0.010	0.03125	-68
59	41	0.021	0.03125	-34	123	20	0.010	0.03125	-68
60	51	0.026	0.03125	-18	124	26	0.013	0.03125	-58
61	51	0.026	0.03125	-18	125	20	0.010	0.03125	-68
62	50	0.025	0.03125	-20	126	23	0.012	0.03125	-63
63	50	0.025	0.03125	-20	127	20	0.010	0.03125	-68
64	90	0.046	0.03125	46					

6.8 HISCAM values

The mean and variance correction involves two steps: First, the mean is corrected to 50 per cohort. Second, the variance is corrected to 100 per cohort. In the process, HISCAM values below 50 after the mean correction are corrected with a different amount than those above 50. This leads to compression or stretching of the distribution. It is important to note that the correction values are determined in the program code and may vary for different data.

TABLE 11
Correction parameters to produce a constant mean (50) and a constant variance (100)

Cohort	Mean value correction	Variance correction for HISCAM values ≥ 50	Variance correction for HISCAM values < 50
1600-1629	-4.683480205	3.14756	-1.94060
1630-1659	-4.529587488	2.88940	-1.44248
1660-1669	-5.314463159	-0.42336	0.17212
1690-1719	-4.450852312	2.89248	-1.25364
1720-1759	-4.197589589	3.99668	-1.61524
1750-1779	-4.008747109	3.69656	-1.32912
1780-1809	-4.259381843	3.49948	-1.49096
1810-1839	-4.489012208	2.41036	-1.26740
1840-1869	-5.044529332	0.96688	-0.42576
1870-1899	-6.572772599	-1.21336	0.77880

The corrected HISCAM values are subject to unit root tests, both with and without temporal differentiation by cohort. Table 12 shows the results of the unit root test with temporal differentiation. A panel data set is a set of data that represents a successive patriline over time, and only balanced panel data sets without missing values are considered. For example, there are 6,575 panel data sets from the 1600-1629 cohort to the 1630-1659 cohort, which corresponds to two cohorts and usually two generations. The unit root test is applied to all data sets of a selected period simultaneously. If the test fails to

reject the null hypothesis of a unit root, the time series is non-stationary. The Breitung test is used because it is well-suited for balanced data (Breitung, 2001; Breitung and Das, 2005). Since the Breitung test does not provide p-values for panels with two periods in STATA, the values for these panels are formatted in italics. The null hypothesis of a unit root is unlikely to be rejected for legged tests (see Table 12), which would render subsequent calculations over these periods invalid. As a result, time-differentiated mobility calculations are of limited use in these cases.

Table 12
Unit root test with a time component

Cohort	1630-1659	1660-1689	1690-1719	1720-1749	1750-1779	1780-1809	1810-1839	1840-1869	1870-1899
1600-1629	6,575	2,259	647***	217***	37	1	0	0	0
1630-1659		*7,697*	2,772***	975***	299***	103***	45***	36***	2
1660-1689			*9,190*	4,162	1,620***	671***	200***	90***	16***
1690-1719				*10,971*	4,045***	1,262***	392***	137***	22*
1720-1749					*11,355*	3,933***	1,379***	442***	69***
1750-1779						*11,240*	3,384***	1,015***	163***
1780-1809							*10,254*	2,881	540***
1810-1839								*9,431*	1,648***
1840-1869									*5,073*

Number of balanced panel data sets, respective panel data set: from a cohort in a row to cohort in the column, only completely balanced data sets (variable: hiscam), unit root test according to Breitung (Breitung, 2001; Breitung & Das, 2005), null hypothesis: unit root present, *** $p < 0.005$, ** $p < 0.01$, * $p < 0.10$, test is not possible for numbers in italics.

In Table 13, the panel data sets are presented according to the number of generations, disregarding the cohort as irrelevant. This consideration is more important for the calculations than the previous one since the panel data sets in this study are not separated in time, but always according to the number of related generations. In contrast to the temporal differentiation in Table 12, all data up to and including the eight contiguous generation do not exhibit a unit root. Nevertheless, this study analyzes up to six contiguous generations because less absolute data in over seven and eight generations.

TABLE 13
Unit root test without time component

N. of generations	2	3	4	5	6	7	8	9	10
N. of panels	39,056	12,392***	3,624***	1,214***	300***	56***	48***	2	0

Number of balanced data sets as a function of the number of generations (variable: hiscam), unit root test according to Breitung (Breitung, 2001; Breitung & Das, 2005), null hypothesis: unit root present, *** $p < 0.005$, ** $p < 0.01$, * $p < 0.10$, test is not possible for numbers in italics.

6.9 Spatial Differences in the Stability Indicators

TABLE 14
Differences of mean values per persistence indicators between provinces

	B18	B24	B10	B22	B11	B19	B21	A13	A11	A09	B20
	–	0.01	-0.01	0.00	0.03**	-0.03	-0.03	-0.04***	-0.01	-0.02	0.06***
		0.84***	1.32***	0.45***	-0.25	1.08***	1.19***	0.79***	0.85***	0.91***	0.42**
B18		0.37*	1.97***	-0.27	-0.09	0.64***	1.03***	-1.20***	0.31	1.21***	0.96***
		0.00	-0.01	-0.00	-0.08***	-0.01	-0.02***	0.06***	-0.03***	-0.03**	-0.05***
		0.05***	0.00	0.05***	-0.01	0.05***	0.05***	0.05***	0.05***	-0.04***	-0.04***
	-0.01	–	-0.02*	-0.01	0.02	-0.04***	-0.04***	-0.06***	-0.03**	-0.03**	0.04***
	-0.84***		0.48***	-0.39***	-1.10***	0.24**	0.34***	-0.05	0.00	0.06	-0.42***
B24	-0.37*		1.60***	-0.63***	-0.46**	-0.27*	0.66***	-1.57***	-0.05	0.84***	0.59**
	-0.00		-0.01*	-0.01	-0.08***	-0.01*	-0.03***	0.05***	-0.04***	-0.03***	-0.06***
	-0.05***		-0.05***	-0.00	-0.06***	0.00	-0.00	0.00	0.02**	-0.00	-0.09***
	0.01	0.02*	–	0.01	0.04***	-0.02**	-0.02***	-0.04***	-0.01	-0.01	0.06***
	-1.32***	-0.48***		-0.87***	-1.57***	-0.24***	-0.13**	-0.53***	-0.48***	-0.41***	-0.90***
B10	-1.97***	-1.6***		-2.23***	-2.05***	-1.33***	-0.94***	-3.17***	-1.65***	-0.76***	-1.01***
	0.01	0.01*		0.00	-0.07***	-0.00	-0.02***	0.06***	-0.03***	-0.02**	-0.05***
	-0.00	0.05***		0.05***	-0.01*	0.05***	0.05***	0.05***	0.07***	0.05***	-0.04***
	-0.00	0.01	-0.01	–	0.03**	-0.03***	-0.03***	-0.05***	-0.02*	-0.02*	0.05***
	-0.45***	0.39***	0.87***		-0.70***	0.63***	0.74***	0.34**	0.40***	0.46***	-0.03
B22	0.27	0.63***	2.23***		0.18	0.90***	1.3***	-0.94***	0.58***	1.48***	1.22***
	0.00	0.01	-0.00		-0.07***	-0.00	-0.02***	0.06***	-0.03***	-0.02**	-0.05***
	-0.05***	0.00	-0.05***		-0.06***	0.00	0.00	0.00	0.02***	0.00	-0.08***
	-0.03**	-0.02	-0.04***	-0.03**	–	-0.06***	-0.06***	-0.07***	-0.04***	-0.05***	0.03*
	0.25	1.10***	1.57***	0.70***		1.34***	1.44***	1.04***	1.10***	1.16***	0.67***
B11	0.09	0.46**	2.05***	-0.18		0.73***	1.12***	-1.11***	0.40*	1.30***	1.05***
	0.08***	0.08***	0.07***	0.07***		0.07***	0.06***	0.14***	0.04***	0.05***	0.02**
	0.01	0.06***	0.01*	0.06***		0.06***	0.06***	0.06***	0.08***	0.06***	-0.03**

(continued)

(continued)

	B18	B24	B10	B22	B11	B19	B21	A13	A11	A09	B20
	0.03	0.04***	0.02**	0.03***	0.06***	–	0.00	-0.02	0.01	0.01	0.08***
	-1.08***	-0.24**	0.24***	-0.63***	-1.34***		0.10	-0.29**	-0.24*	-0.18	-0.66***
B19	-0.64***	-0.27*	1.33***	-0.90***	-0.73***		0.39***	-1.84***	-0.32*	0.57***	0.32
	0.01	0.01*	0.00	0.00	-0.07***		-0.01*	0.07***	-0.02***	-0.02*	-0.05***
	-0.05***	-0.00	-0.05***	-0.00	-0.06***		-0.00	0.00	0.02**	-0.00	-0.09***
	0.03	0.04***	0.02***	0.03***	0.06***	-0.00	–	-0.02	0.01	0.01	0.08***
	-1.19***	-0.34***	0.13**	-0.74***	-1.44***	-0.10		-0.40***	-0.34***	-0.28**	-0.77***
B21	-1.03***	-0.66***	0.94***	-1.30***	-1.12***	-0.39***		-2.23***	-0.71***	0.18	-0.07
	0.02***	0.03***	0.02***	0.02***	-0.06***	0.01*		0.08***	-0.01*	-0.00	-0.03***
	-0.05***	0.00	-0.05***	-0.00	-0.06***	0.00		0.00	0.02***	0.00	-0.09***
	0.04***	0.06***	0.04***	0.05***	0.07***	0.02	0.02	–	0.03**	0.02	0.10***
	-0.79***	0.05	0.53***	-0.34**	-1.04***	0.29**	0.4***		0.05	0.12	-0.37**
A13	1.20***	1.57***	3.17***	0.94***	1.11***	1.84***	2.23***		1.52***	2.41***	2.16***
	-0.06***	-0.05***	-0.06***	-0.06***	-0.14***	-0.07***	-0.08***		-0.09***	-0.08***	-0.11***
	-0.05***	-0.00	-0.05***	-0.00	-0.06***	-0.00	-0.00		0.02**	-0.00	-0.09***
	0.01	0.03**	0.01	0.02*	0.04***	-0.01	-0.01	-0.03**	–	-0.01	0.07***
	-0.85***	-0.00	0.48***	-0.40***	-1.10***	0.24*	0.34***	-0.05		0.06	-0.42**
A11	-0.31	0.05	1.65***	-0.58***	-0.40*	0.32*	0.71***	-1.52***		0.89***	0.64**
	0.03***	0.04***	0.03***	0.03***	-0.04***	0.02***	0.01*	0.09***		0.01	-0.02*
	-0.05***	-0.02**	-0.07***	-0.02***	-0.08***	-0.02**	-0.02***	-0.02**		-0.02*	-0.11***
	0.02	0.03**	0.01	0.02*	0.05***	-0.01	-0.01	-0.02	0.01	–	0.08***
	-0.91***	-0.06	0.41***	-0.46***	-1.16***	0.18	0.28**	-0.12	-0.06		-0.49**
A09	-1.21***	-0.84***	0.76***	-1.48***	-1.30***	-0.57***	-0.18	-2.41***	-0.89***		-0.25
	0.03**	0.03***	0.02**	0.02**	-0.05***	0.02*	0.00	0.08***	-0.01		-0.03**
	0.04***	0.00	-0.05***	-0.00	-0.06***	0.00	-0.00	0.00	0.02*		-0.09***
	-0.06***	-0.04***	-0.06***	-0.05***	-0.03*	-0.08***	-0.08***	-0.1***	-0.07***	-0.08***	–
	-0.42**	0.42***	0.90***	0.03	-0.67***	0.66***	0.77***	0.37**	0.42**	0.49**	
B20	-0.96***	-0.59**	1.01***	-1.22***	-1.05***	-0.32	0.07	-2.16***	-0.64**	0.25	
	0.05***	0.06***	0.05***	0.05***	-0.02**	0.05***	0.03***	0.11***	0.02*	0.03**	
	0.04***	0.09***	0.04***	0.08***	0.03**	0.09***	0.09***	0.09***	0.11***	0.09***	

Order: $\bar{\beta}_{province}$, $\sigma_{y_{province}}$, $|\bar{y}_{all} - \bar{y}_{province}$, KldB stability, HISCLASS stability), significance level by t-test, * $p \leq 0{,}05$, ** $p \leq 0{,}01$, *** $p \leq 0{,}001$.

References

Adermon A., Lindahl M., Palme M. (2021), "Dynastic Human Capital, Inequality, and Intergenerational Mobility", in *American Economic Review*, 111(5), pp. 1523-1548, https://doi.org/10.1257/aer.20190553.

Baranyi G. et al. (2023), "Early Life PM2.5 Exposure, Childhood Cognitive Ability and Mortality between Age 11 and 86: A Record-linkage Life-course Study from Scotland", in *Environmental Research*, 238, https://doi.org/10.1016/j.envres.2023.117021.

Barone G., Mocetti S. (2021), "Intergenerational Mobility in the Very Long Run: Florence 1427-2011", in *The Review of Economic Studies*, 88(4), pp. 1863–1891, https://doi.org/10.1093/restud/rdaa075.

Becker G.S., Tomes N. (1986), "Human Capital and the Rise and Fall of Families", in *Journal of Labor Economics*, 4(3), pp. S1-S39.

Becker S.O., Woessmann L. (2009), "Was Weber Wrong? A Human Capital Theory of Protestant Economic History", in *The Quarterly Journal of Economics*, 124(2), pp. 531-596, https://doi.org/10.1162/qjec.2009.124.2.531.

Blanc G. (2024a), *Demographic Transitions, Rural Flight, and Intergenerational Persistence: Evidence from Crowdsourced Genealogies*, Working paper, preprint.

– (2024b), *The Cultural Origins of the Demographic Transition in France*, Working paper, preprint, https://doi.org/10.2139/ssrn.3702670.

Boberg-Fazlic N., Sharp P., Weisdorf J. (2011), "Survival of the Richest? Social Status, Fertility and Social Mobility in England 1541-1824", in *European Review of Economic History*, 15(3), pp. 365-392, https://doi.org/10.1017/S136149161100013X.

Braun S.T., Stuhler J. (2018), "The Transmission of Inequality Across Multiple Generations: Testing Recent Theories with Evidence from Germany", in *The Economic Journal*, 128(609), pp. 576-611, https://doi.org/10.1111/ecoj.12453.

Breitung J. (2001), "The Local Power of Some Unit Root Tests for Panel Data", in B.H. Baltagi, T.B. Fomby, R. Carter Hill (eds.), *Nonstationary Panels, Panel Cointegration, and Dynamic Panels*,

Bringley: Emerald (Advances in Econometrics), pp. 161-177, https://doi.org/10.1016/S0731-9053(00)15006-6.

Breitung J., Das S. (2005), "Panel Unit Root Tests under Cross-sectional Dependence", in *Statistica Neerlandica*, 59(4), pp. 414-433, https://doi.org/10.1111/j.1467-9574.2005.00299.x.

Clark G. et al. (2014), *The Son Also Rises. Surnames and the History of Social History*, Princeton University Press, Princeton, Oxford.

– (2015), "Surnames: A New Source For The History of Social Mobility", in *Explorations in Economic History*, 55, pp. 3-24, https://doi.org/10.1016/j.eeh.2014.12.002.

Clark G., Cummins N. (2015), "Intergenerational Wealth Mobility in England, 1858–2012: Surnames and Social Mobility", in *The Economic Journal*, 125(582), pp. 61-85, https://doi.org/10.1111/ecoj.12165.

– (2019), *The Child Quality-Quantity Tradeo?, England, 1780-1879: A Fundamental Component of the Economic Theory of Growth is Missing*, Working Paper, http://neilcummins.com/qq_paper_2019_CEPR.pdf (Accessed: 1 December 2021).

Clark G., Cummins N., Curtis M. (2022), *The Mismeasure of Man: Why Intergenerational Occupational Mobility is Much Lower than Conventionally Measured, England, 1800-2021*, Rochester, NY, https://papers.ssrn.com/abstract=4144664 (Accessed: 18 April 2023).

– (2024), "Three New Occupational Status Indices for England and Wales, 1800-1939", in *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 57(1), pp. 41-66, https://doi.org/10.1080/01615440.2024.2368458.

Dalman E. (2025), "The Legacy of Names. Persistence in Social Status in Sweden 1865–2015", in *Research in Social Stratification and Mobility*, 97, https://doi.org/10.1016/j.rssm.2025.101033.

Dennison T., Ogilvie S. (2014), "Does the European Marriage Pattern Explain Economic Growth?", in *The Journal of Economic History*, 74(3), pp. 651-693, https://doi.org/10.1017/S0022050714000564.

Dribe M., Helgertz J. (2016), "The Lasting Impact of Grandfathers: Class, Occupational Status, and Earnings over Three Generations

in Sweden 1815-2011", in *The Journal of Economic History*, 76(4), pp. 969-1000, https://doi.org/10.1017/S0022050716000991.

Eggen B., Rupp M. (2007), *Kinderreichtum - Eine Ausnahme in der neueren Geschichte?*, Statistisches Monatsheft Baden-Württemberg(3), pp. 6-14.

Engel C.L.E. (1869), *Die Ergebnisse der Volkszählung und Volksbeschreibung vom 3. December 1867, I. Theil*, Verlag des königlichen statistischen Bureaus, Berlin.

– (1871), *Die Ergebnisse der Volkszählung und Volksbeschreibung vom 3. December 1867, II. Theil*, Verlag des königlichen statistischen Bureaus, Berlin.

Fertig G. et al. (2018), "Das postmalthusianische Zeitalter: Die Bevölkerungsentwicklung in Deutschland, 1815-1871", in *Vierteljahrschrift für Sozial- und Wirtschaftsgeschichte*, 105(1), pp. 6-33.

Gellatly C. (2015), "Reconstructing Historical Populations from Genealogical Data Files", in G. Bloothooft et al. (eds.), *Population Reconstruction*, Cham, Heidelberg, New York, Dordrecht, London, Springer, pp. 111-128, https://doi.org/10.1007/978-3-319-19884-2_6.

Goldberg J.M., Moeller K. (2022), "Automatisierte Identifikation und Lemmatisierung historischer Berufsbezeichnungen in deutschsprachigen Datenbeständen", in *Zeitschrift für digitale Geisteswissenschaften*, Heft 7, https://doi.org/10.17175/2022_002.

Güell M. et al. (2018), "Correlating Social Mobility and Economic Outcomes", in *The Economic Journal*, 128(612), pp. F353-F403, https://doi.org/10.1111/ecoj.12599.

Güell M., Rodríguez Mora J.V., Solon G. (2018), "New Directions in Measuring Intergenerational Mobility: Introduction", in *The Economic Journal*, 128(612), pp. F335-F339, https://doi.org/10.1111/ecoj.12607.

Hällsten M., Kolk M. (2021), *The Shadow of Peasant Past: Seven Generations of Inequality Persistence in Northern Sweden*, Working Paper, https://doi.org/10.31235/osf.io/yjksz.

Häner M., Schaltegger C.A. (2021), "Fällt der Apfel weit vom Stamm? Ein Überblick über den Forschungsstand zur intergen-

erationellen sozialen Mobilität", in *Perspektiven der Wirtschaftspolitik*, 22(2), pp. 103-120, https://doi.org/10.1515/pwp-2020-0052.
– (2022), "The Name Says It All. Multigenerational Social Mobility in Basel (Switzerland), 1550–2019", in *Journal of Human Resources*, preprint, https://doi.org/10.3368/jhr.0621-11749R2.
Harviainen J.T., Björk B.-C. (2018), "Genealogy, GEDCOM, and Popularity Implications", in *Informaatiotutkimus*, 37, pp. 4-14, https://doi.org/10.23978/inf.76066.
Hochstadt S. (1983), "Migration in Preindustrial Germany", in *Central European History*, 16(3), pp. 195-224.
Hoffmann F. (2012), *Ein den thatsächlichen Verhältnissen entsprechendes Bild nicht zu gewinnen*, Quellenkritische Untersuchungen zur preußischen Gewerbestatistik zwischen Wiener Kongress und Reichsgründung, Steiner, Stuttgart.
Hradil S. (2005), *Soziale Ungleichheit in Deutschland*, 8th ed., Verlag für Sozialwissenschaften, Wiesbaden.
Imhof A.E. (1981), *Die gewonnenen Jahre: von der Zunahme unserer Lebensspanne seit dreihundert Jahren, oder von der Notwendigkeit einer neuen Einstellung zu Leben und Sterben: ein historischer Essay*, C.H. Beck, München.
Kaiserliches Statistisches Amt (1873a), *Vierteljahrshefte zur Statistik des Deutschen Reichs für das Jahr 1873. Erstes Heft*, Verlag des Königlich Preussischen Statistischen Bureaus, Berlin.
– (1873b), *Vierteljahrshefte zur Statistik des Deutschen Reichs für das Jahr 1873. Viertes Heft*, Verlag des Königlich Preussischen Statistischen Bureaus, Berlin.
– (1873c), *Vierteljahrshefte zur Statistik des Deutschen Reichs für das Jahr 1873. Zweites Heft*, Verlag des Königlich Preussischen Statistischen Bureaus, Berlin.
– (1878), *Monatshefte zur Statistik des Deutschen Reichs für das Jahr 1878. Oktober-Heft*, Puttkammer & Mühlbrecht, Berlin.
– (1879), *Statistik des Deutschen Reichs. Die Ergebnisse der Deutschen Gewerbezählung vom 1. Dezember 1875*, Puttkammer & Mühlbrecht, Berlin.

Lambert P.S. et al. (2013), "The Construction of HISCAM: A Stratification Scale Based on Social Interactions for Historical Comparative Research", in *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 46(2), pp. 77-89, https://doi.org/10.1080/01615440.2012.715569.

Lindahl M. et al. (2015), "Long-term Intergenerational Persistence of Human Capital: An Empirical Analysis of Four Generations", in *Journal of Human Resources*, 50(1), pp. 1-33.

Long J., Ferrie J. (2013), "Intergenerational Occupational Mobility in Great Britain and the United States since 1850", in *American Economic Review*, 103(4), pp. 1109-1137, https://doi.org/10.1257/aer.103.4.1109.

– (2018), "Grandfathers Matter(ed): Occupational Mobility Across Three Generations in the US and Britain, 1850-1911", in *The Economic Journal*, 128(612), pp. F422-F445, https://doi.org/10.1111/ecoj.12590.

Mare R.D. (2011), "A Multigenerational View of Inequality", in *Demography*, 48(1), pp. 1-23.

Modalsli J., Vosters K. (2024), "Spillover Bias in Multigenerational Income Regressions", in *Journal of Human Resources*, 59(3), pp. 743-776, https://doi.org/10.3368/jhr.0520-10919R2.

Paulus W., Matthes B. (2013), "Klassifikation der Berufe 2010 – Struktur, Codierung und Umsteigeschlüssel", in *FDZ-Methodenreport*, 8, https://doku.iab.de/fdz/reporte/2013/MR_08-13.pdf (Accessed: 1 December 2021).

Peri-Rotem N. (2016), "Religion and Fertility in Western Europe: Trends across Cohorts in Britain, France and the Netherlands", in *European Journal of Population*, 32, pp. 231-265, https://doi.org/10.1007/s10680-015-9371-z.

Pfister C. (2010), *Bevölkerungsgeschichte und historische Demographie 1500-1800, Bevölkerungsgeschichte und historische Demographie 1500-1800*, Oldenbourg Wissenschaftsverlag, München.

Pfister U., Fertig G. (2020), "From Malthusian Disequilibrium to the Post-Malthusian Era: The Evolution of the Preventive and Positive Checks in Germany, 1730-1870", in *Demography*, 57, pp. 1145-1170, https://doi.org/10.1007/s13524-020-00872-w.

Price J. et al. (2021), "Combining Family History and Machine Learning to Link Historical Records: The Census Tree Data Set", in *Explorations in Economic History*, 80, https://doi.org/10.1016/j.eeh.2021.101391.

Rosenbaum-Feldbrügge M. (2019), "The Impact of Parental Death in Childhood on Sons' and Daughters' Status Attainment in Young Adulthood in the Netherlands, 1850-1952", in *Demography*, 56(5), pp. 1827-1854, https://doi.org/10.1007/s13524-019-00808-z.

Schüren R. (1989), *Soziale Mobilität. Muster, Veränderungen und Bedingungen im 19. und 20. Jahrhundert*, Scripta Mercaturae, St. Katharinen.

Solon G. (2018), "What Do We Know So Far about Multigenerational Mobility?", in *The Economic Journal*, 128(612), pp. F340-F352, https://doi.org/10.1111/ecoj.12495.

Song X. et al. (2020), "Long-term Decline in Intergenerational Mobility in the United States since the 1850s", in *Proceedings of the National Academy of Sciences*, 117(1), pp. 251-258, https://doi.org/10.1073/pnas.1905094116.

Stelter R., Alburez-Gutierrez D. (2022), *Representativeness Is Crucial for Inferring Demographic Processes from Online Genealogies: Evidence from Lifespan Dynamics*, Proceedings of the National Academy of Sciences of the United States of America, 119(10), p. e2120455119, https://doi.org/10.1073/pnas.2120455119.

Torche F., Corvalan A. (2018), "Estimating Intergenerational Mobility with Grouped Data: A Critique of Clark's the Son Also Rises", in *Sociological Methods & Research*, 47(4), pp. 787-811, https://doi.org/10.1177/0049124116661579.

Van Bavel J. et al. (2011), "Family Size and Intergenerational Social Mobility during the Fertility Transition", in *Demographic Research*, 24, pp. 313-344, https://doi.org/10.4054/DemRes.2011.24.14.

van Leeuwen M.H.D. et al. (2013), *HISCAM: Estimating Social Interaction and Stratification Scales for the 19th and 20th Century, CAMSIS: Social Interaction and Stratification Scales*, https://www.camsis.stir.ac.uk/hiscam, Accessed: 1 December 2021.

– (2016), "Social Mobility in France 1720-1986: Effects of Wars,

Revolution and Economic Change", in *Journal of Social History*, 49(3), pp. 585-616.

van Leeuwen M.H.D., Maas I. (2010), "Historical Studies of Social Mobility and Stratification", in *Annual Review of Sociology*, 36, pp. 429-451, https://doi.org/10.1146/annurev.soc.012809.102635.

– (2011), *Hisclass: A Historical International Social Class Scheme*, Universitaire Pers Leuven, Leuven.

Verein für Computergenealogie (2020), *GEDBAS, GEDBAS*, https://gedbas.genealogy.net/ (Accessed: 17 April 2023).

– (2021), *GEDBAS/FAQ – GenWiki*, http://wiki-de.genealogy.net/GEDBAS/FAQ#Was_passiert_mit_den_Daten_lebender_Personen.3F (Accessed: 14 April 2021).

Vosters K. (2018), "Is the Simple Law of Mobility Really a Law? Testing Clark's Hypothesis", in *The Economic Journal*, 128(612), pp. F404-F421, https://doi.org/10.1111/ecoj.12516.

Ward Z. (2023), "Intergenerational Mobility in American History: Accounting for Race and Measurement Error", in *American Economic Review*, 113(12), pp. 3213-3248, https://doi.org/10.1257/aer.20200292.

Westoff C.F., Jones E.F. (1979), "The End of 'Catholic' Fertility", in *Demography*, 16(2), pp. 209-217, https://doi.org/10.2307/2061139.

Xie Y., Killewald A. (2013), "Intergenerational Occupational Mobility in Great Britain and the United States since 1850: Comment", in *American Economic Review*, 103(5), pp. 2003-2020, https://doi.org/10.1257/aer.103.5.2003.

Zijdeman R.L., Lambert P.S. (2010), "Measuring Social Structure in the Past: A Comparison of Historical Class Schemes and Occupational Stratification Scales on Dutch 19th and Early 20th Century Data", in *Belgisch Tijdschrift voor Nieuwste Geschiedenis*, XL(1-2), pp. 111-141.