The South African Families Database

By Jeanne Cilliers

To cite this article: Cilliers, J. (2021). The South African Families Database. *Historical Life Course Studies*, *11*, 97–111. https://doi.org/10.51964/hlcs11095

HISTORICAL LIFE COURSE STUDIES

Content, Design and Structure of Major Databases with Historical Longitudinal Population Data

VOLUME 11, SPECIAL ISSUE 5, 2020

GUEST EDITORS
George Alter
Kees Mandemakers
Hélène Vézina

MISSION STATEMENT

HISTORICAL LIFE COURSE STUDIES

Historical Life Course Studies is the electronic journal of the *European Historical Population Samples Network* (EHPS-Net). The journal is the primary publishing outlet for research involved in the conversion of existing European and non-European large historical demographic databases into a common format, the Intermediate Data Structure, and for studies based on these databases. The journal publishes both methodological and substantive research articles.

Methodological Articles
This section includes methodological articles that describe all forms of data handling involving large historical databases, including extensive descriptions of new or existing databases, syntax, algorithms and extraction programs. Authors are encouraged to share their syntaxes, applications and other forms of software presented in their article, if pertinent, on the EHPS-Net website.

Research articles
This section includes substantive articles reporting the results of comparative longitudinal studies that are demographic and historical in nature, and that are based on micro-data from large historical databases.

Historical Life Course Studies is a no-fee double-blind, peer-reviewed open-access journal supported by the European Science Foundation (ESF, http://www.esf.org), the Scientific Research Network of Historical Demography (FWO Flanders, http://www.historicaldemography.be) and the International Institute of Social History Amsterdam (IISH, http://socialhistory.org/). Manuscripts are reviewed by the editors, members of the editorial and scientific boards, and by external reviewers. All journal content is freely available on the internet at hlcs.nl.

Co-Editors-In-Chief:
Paul Puschmann (Radboud University) & Luciana Quaranta (Lund University)
hislives@kuleuven.be

The European Science Foundation (ESF) provides a platform for its Member Organisations to advance science and explore new directions for research at the European level. Established in 1974 as an independent non-governmental organisation, the ESF currently serves 78 Member Organisations across 30 countries. EHPS-Net is an ESF Research Networking Programme.

The European Historical Population Samples Network (EHPS-net) brings together scholars to create a common format for databases containing non-aggregated information on persons, families and households. The aim is to form an integrated and joint interface between many European and non-European databases to stimulate comparative research on the micro-level.
Visit: http://www.ehps-net.eu.

HISTORICAL LIFE COURSE STUDIES
VOLUME 11 (2021), published 05-11-2021

The South African Families Database

Jeanne Cilliers

Lund University

ABSTRACT

Very little is known about what family life looked like for settlers in colonial South Africa during the 18th or 19th century, nor how events over these centuries might have affected demographic change. The primary reason for this lacuna is a shortage of adequate data. Historians and genealogists have, over the last century, worked to combine the rich administrative records that are available in the Cape Archives in Cape Town and beyond, into a single genealogical volume of all settlers living in the 18th, 19th and early 20th century. Until recently, this valuable resource was not in a format that would enable its use for the type of event-history analyses that have come to dominate the field of contemporary historical demography. This is now changing with the introduction of the South African Families database (SAF). SAF is one of very few databases known to document a full population of immigrants and their families over several generations. This article provides a brief background to, and technical overview of, the construction of the SAF. It discusses both the merits and limitations of its use in longitudinal demographic studies and offers a look into the types of studies it can enable.

Keywords: Historical demography, Genealogies, Longitudinal data, Life courses, Intermediate Data Structure, South Africa

e-ISSN: 2352-6343
DOI article: https://doi.org/10.51964/hlcs11095

1 INTRODUCTION

Assembling archival materials and historical registries to reconstruct family lineages of the European settlers to South Africa from the 17th to 20th centuries allows for an investigation into long-term economic and demographic trends across more than just two or three generations. Questions relating to the inter-generational transmission of socioeconomic status or about demographic processes such as fertility, migration, and marriage, that have previously gone unanswered, re-emerge.

Thanks to the wealth of documents kept by the Dutch East India Company (VOC) and the British colonial government when they ruled South Africa, much is already known about the establishment of the South African colonial society (Fourie, 2014). Less is known about what family life looked like for settlers in the 18th and 19th century nor how events over this period might have affected the way in which decisions around household formation were made. This is exacerbated to some extent by the fact that South Africa does not have a research hub for historical demography to encourage researchers to collect and transcribe data from the archives. As a result, South African historical demography remains in its infancy.

The South African Families Database (hereafter SAF) is a genealogical registry of settler families. It is one of very few in the world that is known to document a full population of immigrants and their families over several generations spanning nearly three centuries. The registers were painstakingly compiled by historians and genealogists using baptism and marriage registers, death notices, and individual family genealogies. The time-intensive nature of manual data transcription and a lack of computing power has meant that up until fairly recently, researchers opted to draw only small samples from these data, and as a result they had never been used in their entirety. Over the last decade these records have been turned into a functional database. The SAF database now includes information on all families known to have settled in South Africa and their descendants, complete until 1910, containing over half a million individuals.

This article provides a brief background to and technical overview of the construction of the South African Families database. It discusses both the strengths and limitations of its use in longitudinal demographic studies and offers a look into research currently being undertaken with these data at their core.

2 WHO WERE THE CAPE SETTLERS?

The Dutch, while not the first Europeans to ever traverse the southern parts of Africa — the Portuguese having done so a century prior — were the first to settle at the Cape of Good Hope, landing in 1652. In that year three ships of the VOC, under the Commander Jan van Riebeek, arrived in Table Bay with the first company men. The VOC, with its base in Batavia, was a powerful monopolistic chartered company and the Cape was to serve the Company's ships as a rest stop on their passage to India. Of course, the Cape was not previously uninhabited. VOC company men settled on lands wrested from the indigenous Khoesan populations and their movements further inland were characterized by tension and violence between the groups.

Notions of family life amongst early European settlers at the Cape likely derived from the diverse cultural and religious practices of VOC employees' homelands. The end of Thirty Years War in 1648 saw European soldiers and refugees widely dispersed across the continent. Immigrants from Germany, Scandinavia, and Switzerland journeyed to Holland in the hope of finding employment and were amongst those who would make the six-month journey to settle the southern tip of Africa. Beyond this, the company filled its ranks with farm labourers, artisans, and unskilled workers from both rural and urban areas who spoke variations of French, Dutch, and German.

A consequential event of the 17th century at the Cape, was the arrival of about 170 French Huguenots in 1688 and 1689, by which time the free settler population had reached about six hundred. Cultural adaptation took place rapidly since new identities had to be shaped in a settler environment. De Kiewiet (1941, p. 6) described the arrival of the Huguenots as giving the Cape "more truly than before the contours and substance of a colony". He notes that although the Huguenots differed from the Dutch

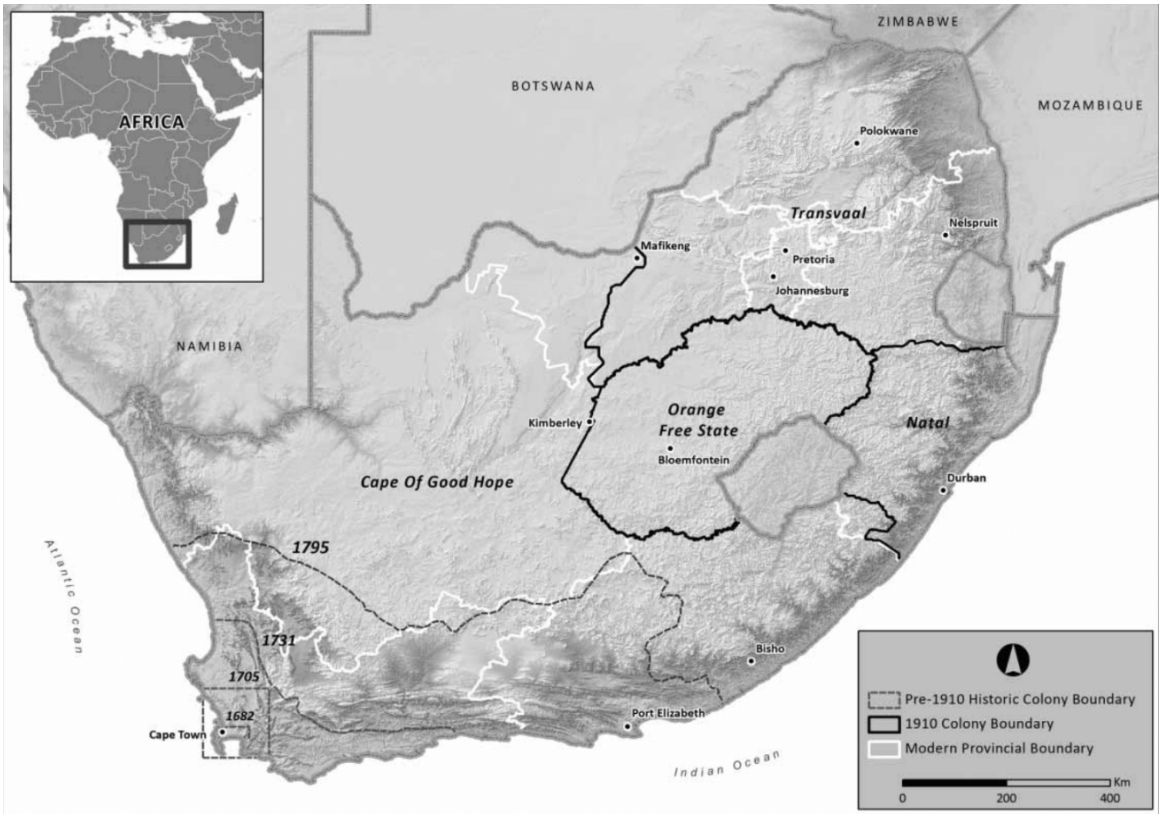
settlers in language, they were united by equal devoutness and tradition and "in two generations or less the groups had grown together and become one" (de Kiewiet, 1941, p. 6).

By the beginning of the 18th century free settlers had increased in number and influence and become more and more independent of the authority of Company officials. With the exception of the smallpox epidemics of 1713 and 1755, which resulted in slight declines in the population growth rate, the 18th century experienced a gross population growth rate of around 2.6% per annum (van Duin & Ross, 1987, p. 12). A steady flow of immigration of European settlers would continue so that by the end of the VOC's governance in 1795 the Colony was home to nearly 15,000 settlers (van Duin & Ross, 1987).

The British annexation of the Cape in 1795, and again in 1806 after a brief interlude of Batavian rule (1803–1806), brought immigrants from Britain to the Colony. Most notably some 4,000 settlers arrived in the Eastern Cape in 1820, as beneficiaries of a major scheme of assisted migration. The result of the arrival of an overtly British pressure group; the abolition of slavery; economic motives relating to insecure tenure of land and the abundance of fertile land beyond the frontier; and inadequate protection from native depredations were among the reasons cited for the mass exodus, beginning in 1835 of settlers further into the interior, known as the *Great Trek* (Neumark, 1957, p. 20).

These newly settled regions later formed the two independent Boer republics of the Orange Free State (1848) and the Transvaal (1852) and the colony of Natal (1843), which, together with the Cape Colony, became the four provinces of the Union of South Africa in 1910 (see Figure 1). The discovery of diamonds (1866) and gold (1886) in the two Boer republics boosted the population and income of settler South Africa. Migration to the diamond and gold fields increased rapidly, both from within the region and from outside its borders. Kimberley in the Orange Free State was the hub of the diamond industry, but its wealth was minor in comparison to the immense wealth generated by the discovery of gold in the Witwatersrand region of the Transvaal.

Figure 1 *Map showing the settler expansion from the south-western Cape*



Note: The expansion until 1795 is shown by way of the grey dashed lines; the four provinces that later constituted the Union of South Africa in 1910 by way of the black lines; and the modern-day provincial boundaries of South Africa, by way of white lines.

Source: Cilliers & Fourie, 2012.

While much is known about the political events of the pre-Union period, less is known about changes in living standards. The 17th- and 18th-century Cape Colony is generally considered to have been poor, almost entirely dependent on agriculture, although pockets of wealth could be found close to the market in Cape Town (Guelke & Shell, 1983). Recent scholarship has raised doubts about this view of the Cape Colony: Fourie (2013) uses probate inventories to show that 18th-century Cape settlers owned, on average, greater quantities of luxury goods than many of their European counterparts. Fourie and van Zanden (2013) find that Cape settlers' per capita income was in line with the most prosperous countries of the time, Holland and England.

How living standards, societal inequalities, and economic developments might have factored into individuals' choices about when and whom to marry, or what the ideal family size might be in this context, has previously been based on qualitative research. Anecdotal evidence seems to suggest exceptionally large family sizes. For example, Penn (2014) describes a woman in 1727, in her early 30s already the mother of seven children, who would go on to bear 11 children in total. Ross (1975) tells of a woman dying at the age of 49 at the birth of her twelfth child, whose husband would incidentally go on to father another 12 children with his second wife. While these cases appear to be outliers, they serve to highlight the difficulties of drawing conclusions based on a limited number of observations. Fertility rates for the early Cape Colony come from a study by Guelke (1988) in which the average number of children per woman are calculated from a sample size of fewer than 300, for just two years, 1705 and 1730. Simkins and van Heyningen (1989) offer similar snap-shot crude birth rate calculations using aggregated census data from 1891 and 1904 respectively. These aggregate censuses, available roughly decennially from the second half of the 19th century, while sufficiently broad in scope, do not allow for the possibility to follow individual households or family lineages over time. Evidently this body of literature requires an update. The SAF database is a springboard for a new generation of research that can address this lacuna.

3 THE SOURCE MATERIAL

The lineages that form the basis of SAF were compiled from thousands of source documents. According to the Genealogical Institute of South Africa (GISA) these sources include but are not limited to, baptism and marriage records of the Dutch Reformed Church archives in Cape Town; marriage documents of the courts of Cape Town, Graaff-Reinet, Tulbagh, Colesberg, collected from a card index in the Cape Archives Depot; death notices in the estate files of Cape Town and Bloemfontein; registers of the Reverends Archbell and Lindley; voortrekker baptismal register in the Dutch Reformed Church archive in Cape Town; marriage registers of the magistrate of Potchefstroom; other notable genealogical publications including *Geslachtregister der Oude Kaapsche Familien* [Genealogies of Old Cape Families] (De Villiers, 1894); *Die Herkoms van die Afrikaner* [The Origins of the Afrikaner], 1657–1867 (Heese, 1971); *The Family Register of the South African Nation* (Malherbe, 1966); *Some Frontier Families* (Mitford-Baberton & White, 1968), and various individual families genealogical publications.

Varying degrees of measurement error may have been introduced during the process of data compilation and digitization. The first is the possibility of errors in the original source documents. Misspelling of names and misreporting of dates are likely, given the differential precision applied by the members of the clergy and colonial administration responsible for the maintenance of the respective records. Next, mistakes will have inevitably cropped up in the process of compiling the genealogies. Many of the source documents were copies of originals that had been lost, in some the writing was faded, indistinct or illegible, or had already been transposed a number of times. The degree to which genealogists made discretionary choices in such instances can never be fully known. These issues will be handled systematically in section 5.

The resulting volumes, *South African Genealogies* (2008) and *South African Families* (2012) represent over a century of effort by South African genealogists, many of whom devoted their careers to creating and expanding these registers. In doing so they have, perhaps unintentionally, provided a rich source for exploring South African settler demographic history. An excerpt from the Cilliers lineage (Figure 2) shows the format of a typical register. A short text biography of the progenitor is provided, often containing some details about his region of origin or journey to the Cape. In this example, we are told

that Josué Cellier was born in 1667 in Orleans, France, and arrived at the Cape, aboard a vessel named the "Reygersdaal" in 1700, together with his wife, Elisabeth Couvert whom he had married that same year. They settled on "Het Kruyspad" farm in the district of Brackenfell and later moved to "Orleans" in Daljosaphat. She would go on to marry Paul Roux in 1722 after Josue's death in 1721. Their (Josue and Elisabeth's) children are listed below.

Figure 2 *Excerpt from 'South African Families' (2012)*

CELLIERS / CILLIERS / CILLIE

Josué Cellier * Orleans, Frankryk c. 1667
a. aan Kaap 1700 aan boord Reygersdaal met sy vrou, vestig aanvanklik te "Het Kruyspad", dist Brackenfell en Later "Orleans", Daljosaphat. Volgens Boucher is Josue Celliers moontlik die seun v Josue Celliers en sy vrou Judith Rouilly. Hierdie egpaar het 'n seun Nicolaas in die kerk te Bazoches-en-Dunois laat doop. † " De Orleans", dist Drakenstein Okt. 1721 x Frankryk c. 1700, Elisabeth COUVERT * Orleans, Frankryk c. 1676 † c. 1743 (sy xx c. 1722 Paul Roux † Drakenstein 7.2.1723)

b1 Josué ≈ Drakenstein 2.1.1701 † dist. Drakenstein 19.4.1770, ongetroud
b2 Jan ≈ c. 1702 † c. 1755, burger v Drakenstein x Paarl 5.12.1728 Anna MARAIS * ≈ c. 1707 (wed. v. Gabriel Rossouw) † dist. Drakenstein 11.1.1765 d.v. Charles Marais en Anna de Ruelle
c1 Jan ≈ Paarl 9.10.1729 † dist. Drakenstein 6.6.1766 x Tulbagh 8.10.1751 Susanna MALHERBE ≈ Drakenstein 15.2.1733 † c.1754 d.v. Pierre Malherbe en Elisabeth Cellier xx Paarl 11.7.1756 Sara Margaretha ROSSOUW ≈ Drakenstein 5.8.1736 † Drakenstein 18.7.1821 d.v. Daniel Rossouw en Sara Hanekom
d1 Johannes ≈ Paarl 30.7.1752 † dist. Drakenstein 5.6.1816 x Paarl 12.10.1783 Anna Maria NAUDE ≈ Paarl 12.10.1760 † dist. Drakenstein 22.7.1809 d.v. Jacob Naude en Susanna du Toit
e1 Johannes Francois ≈ Paarl 19.9.1784 † dist. Paarl 10.9.1843 x Paarl 13.6.1806 Anna Magdalena ROSSOUW * 2.3.1788 ≈ Paarl 9.3.1788 † dist. Drakenstein 7.7.1822 d.v. Pieter Rossouw en Anna Cilliers xx Cradock 5.12.1824 Maria Magdalena BREED * 7.2.1807 ≈ Graaff-Reinet 26.4.1807 d.v. Johannes Augustus Breed en Johanna Venter
f1 Anna Magdalena * 12.7.1807 ≈ Paarl 9.8.1807 † Prince Alfred Hamlet 6.6.1873 x Paarl 5.4.1829 (Johannes) Cornelis Jeremias GOOSEN ≈ 10.9.1809 † 6.10.1892 s.v. Gideon Jacobus Goosen en Hester Catharina Malan
f2 Johannes Francois * 13.3.1810 ≈ Paarl 8.4.1810 † Paarl 31.10.1879 x Paarl 8.9.1835 Maria Johanna DU TOIT * c. 1815 † Paarl 9.6.1874 d.v. Daniel du Toit en Maria Elizabeth Marais

4 DATA CAPTURING AND CODING

At the outset, transforming these registers into a functional format fit for analysis proved an enormous task, which spanned the better part of 2011. The first step in the data capturing process was to create a custom-designed data-transposing software that was able to convert what was essentially long text strings demarcated with proprietary symbology into a format compatible with conventional statistical software packages which also captured only the relevant information. This was a cumbersome process as the programme, while innovative, was not able to distinguish between successive families and meant that data had to be read in on a family by family basis. Resulting from typesetting inconsistencies in the SAF volumes, many records still required substantial post-transcription cleaning. Some family lineages, the Cilliers for example, were compiled by the Genealogical Institute of South Africa (GISA) in Afrikaans while others were in English. For consistency, all output was translated to English.

The first resulting dataset captured only the following individual-level information: names, surnames, birth, baptism, marriage and death dates and position in the genealogical tree. Soon after this initial phase of transcription, however, GISA undertook to revise and republish the registers, with the aim of correcting errors where possible, and extending the series to contain complete family registers of all

settler families up to 1930. A new edition of the genealogical registers was published by the GISA in 2014 and contained complete family registers of all settler families from 1652 to approximately 1830 as well as those of new progenitors of settler families up to 1867 for families with surnames starting with the letters A–Z, and up to 1930 for families with surnames starting with the letters A–K. In 2016, GISA completed their revisions of L surname families before bequeathing the ownership and copyright of the series to the Genealogical Society of South Africa (GSSA). Although several registers M–Z have subsequently been updated, and continue to be revised, the latest release of the SAF database only includes up to the L revision.

To transcribe these new versions of the registers a more sophisticated data transcription programme was designed to extract more information. This process was completed in 2013 and a new dataset containing both the original set of variables, as well as new information on occupation (where available), locations of vital events, and spousal information including birth, baptism and death dates and places as well as maiden names (where applicable) and parents' names. The inclusion of the new information was limited to surnames starting with A–K information but provided a significant sample size increase. The inclusion of the revised and expanded A–K data into the original dataset was permitted since having a surname A–K was not found to make an individual systematically different from those with surname starting with L–Z on observable characteristics, including age at first marriage and net fertility. Moreover, no systematic differences between the two versions of the data, other than the increased sample size, indicate that any errors that might remain in the data can be safely attributed to the underlying data, rather than the transcription process. The SAF database is freely available for academic use. An anonymized version of the database will be made publicly available while use of the full version (including personal identifiers) can be made available upon request.

As will become apparent, additional variables were critical to enable the broader usability of the dataset for the purposes of longitudinal or event-history analysis. This is because the original structure of the genealogies is patrilineal. That is, children appear under their father's lineage and are not directly linked to their mothers. In the first example in Figure 2, this would not be problematic because a list of all the offspring of Josué and Elisabeth is given. If, however, Josué had remarried and continued to have children with a second wife, the listed offspring would have to be assigned to their respective mothers. To do so, information about death and/or marriage dates of the spouses is needed. However, this is not always available which limits research into for example female fertility.

In addition to the information captured directly from the source, a number of new variables were generated during this stage of transcription to facilitate the linking of individuals to both of their parents, and the tracing of familial relationships with relative ease over multiple generations. To generate unique individual identity codes, genealogical codes were concatenated to surnames to indicate the relative position of individuals on their family tree. The genealogical codes follow the de Villiers-Pama numbering system. The de Villiers-Pama System is similar to the Henry Numbering System more commonly used in the United States, except that each digit (or group of two digits for numbers larger than 9) is preceded by a generation letter. The progenitor of a particular family, or the first ancestor of that family entering the country is assigned the letter "a". This designates him or her as the "a" generation. The "a" is followed by a number showing which child he/she was. "a3" would mean that the person was the third child in the "a" generation. The children of a3 will be the "b" generation. They will be numbered according to how they were born — the eldest or first born being b1; the second b2; b3 etc. Children descending from the "b" generation will be the "c" generation and so on. An illustration is provided in Figure 3.

On top of the unique individual ID codes assigned, individuals were also assigned a sibship ID, which equates to their individual ID with the last digit (or group of two digits for birth orders higher than 9) removed. The final two entries from the excerpt in Figure 2, Anna Magdelena and Johannes Francois, would therefore have individual IDs: CILLIERS_a1b2c1e1f1 and CILLIERS_a1b2c1e1f2, respectively and would share the sibling ID: CILLIERS_a1b2c1d1e1f. Their father is identifiable by removing the last character from their sibling ID: CILLIERS_a1b2c1d1e1.

Sex is never stated in the records nor is it immediately discernible from the de Villiers-Pama genealogical coding. A sex variable was generated for every individual post-transcription, through a semi-automated process whereby sex was attributed based on the likelihood that first and second names of individuals and their spouses matched a predetermined list of common South African names. All ambiguous cases were dealt with manually. Individuals for whom a sex was indeterminable constitute around 1% of the database.

Figure 3 *Example of the de Villiers-Pama structure*

```
a  Progenitor
   b1  Child
      c1  Grandchild
          d1  Great-grandchild
          d2  Great-grandchild
       c2  Grandchild
       c3  Grandchild
    b2  Child
       c1  Grandchild
           d1  Great-grandchild
           d2  Great-grandchild
           d3  Great-grandchild
       c2  Grandchild
       c3  Grandchild
```

Since women appear as wives in their husband's households but are not directly linked to their own children through the transcription process, a mother's ID variable was generated and assigned to each individual. In cases where a man was only married once in his lifetime (94.5% of the fathers in the sample), matching mothers to their children was a relatively straightforward process using the individual and sibling identification codes. In these cases individuals who share a sibling identifier all are assigned the same mother's ID (the ID of their father's only wife). Cases where men married more than once require more careful distinction of children belonging to the first wife from children belonging to the second, third, or in some rare cases, fourth wife. An algorithm using the previous wife's death date, subsequent marriage date, and the birth dates of all of the children, allows for the linking of children to the correct mother. In the event that there was more than one wife and a birth or death date was missing, a successful match cannot be made. As a result, 18% of non-progenitor individuals in the database have a missing mother's ID in the database.

With all familial relationships clearly established in the database, conversion to longitudinal format to allow for event-history analysis, was fairly straightforward. The only familial relationship that remains untraced is that of children to their mother's ancestors. This is because females appear as children in their father's genealogy i.e., under their maiden names, and then as wives in their husband's genealogy, i.e. under their married names. The possibility to make this linkage does exist using these maiden names, and record-linkage strategies to do so are currently being explored.

5 REPRESENTATIVENESS

Specific concerns related to data appropriateness or representativeness for a given research question are already covered extensively in the various publications which make use of the SAF database. These will be discussed briefly below but a few general points regarding data quality are worth making here.

Family lineages have long been used by demographers in their studies on past demographic behaviour. The common problems associated with the use of genealogical data in historical demography research are already well documented (Hollingsworth, 1969; Willigan & Lynch, 1982; Zhoa, 2001) and they are obviously biased towards the fertile and the marriageable. By definition, a genealogy is the written record of a family descended from a common ancestor or ancestors, and as a result, most genealogies are the records of members of *surviving* patrilineages. These families would most likely have experienced favourable demographic conditions which resulted in their survival. The use of these genealogies may, therefore, not be representative of the history of the whole population in question (Zhoa, 2001, p. 181). As Willigan and Lynch (1982, p. 112) argue: "Genealogies were often designed to emphasize not only the glorious aspects of a lineages past but also its durability through time. Consequently, members who contributed little to the group's duration were likely to be missing or underrepresented. This category might include individuals who did not reach maturity and those who

survived but had no children, or who had children who themselves died at a young age or failed to reproduce. This creates a bias towards long generations (late marriage, remarriage, late child-bearing, high fertility) and long life." In general, the greater the number of generations recorded, the smaller the impact of the selective bias, as long as the genealogy does not suffer severely from other types of under-registration. If the genealogy is shallow in generational depth or the members of the first few generations consist of a large part of the population being investigated, the selective biases are more likely to affect the outcome. Otherwise, their influences can be negligible. The SAF database benefits from great generational depth (see Table 1). However, as a result of small population sizes (the entire free burgher population consisting of less than 1000 individuals before 1700) and very small sample sizes for the period 1652–1699, using SAF to study the period prior to 1700 is not advised.

Table 1 *Distribution of individuals across generations, by birth cohort*

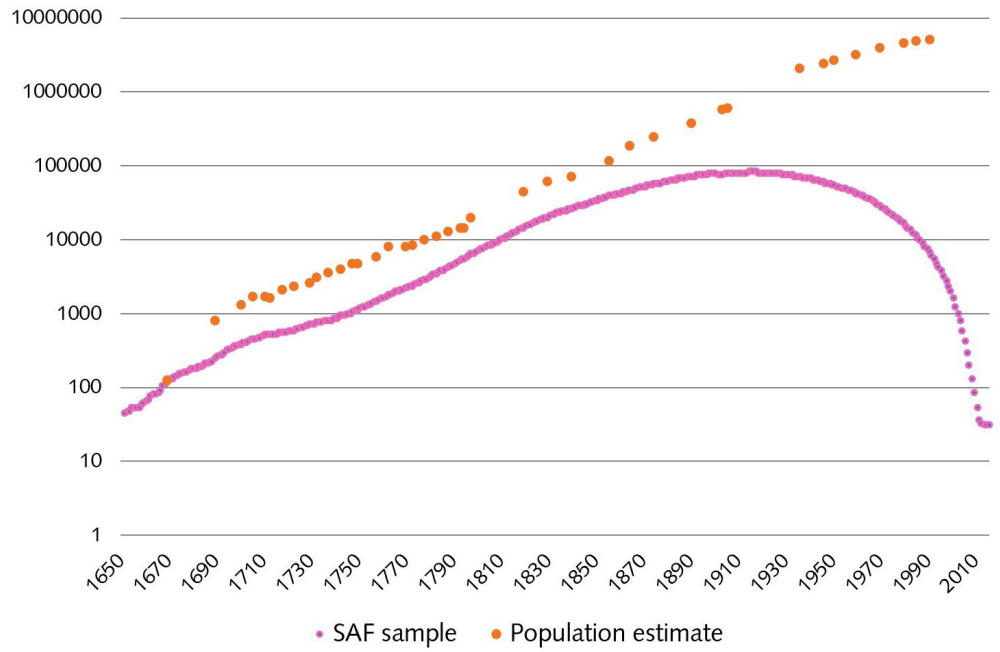
Generation	1650–1749	1750–1849	1850–1949	Total	% of sample
1	204	1,624	4,343	6,171	1.40
2	2,599	15,641	22,516	40,756	9.27
3	1,988	14,569	24,024	40,581	9.23
4	627	19,316	28,388	48,331	10.99
5	25	26,913	34,913	61,851	14.07
6	1	27,391	52,739	80,131	18.23
7	0	9,862	77,971	87,833	19.98
8	0	767	52,701	53,468	12.16
9	0	18	17,277	17,295	3.93
10	0	1	3,019	3,020	0.69
11	0	0	154	154	0.04
12	0	0	17	17	0.00
Total	5,444	116,102	318,062	439,608	100.00

It is also necessary to address the representativeness of the SAF data in terms of the size of the documented historical population. While GISA asserts that the registers are complete up until 1869 for all families and complete to 1930 for families with surnames starting with letters A–L, the registers also contain information on individuals up to the present. This information only exists, however, where families have taken it upon themselves to keep information on their family trees publicly up to date. This calls into question the representativeness of the registers after 1930, since it is unclear what kind of a bias this self-selection into the registers would introduce.

Moreover, as illustrated by Figure 4 which plots the sample size against the actual population over the whole period, the sample closely correlates with estimates of the total settler population for the 18th century and 19th century. By the early 20th century absolute SAF sample size slows considerably relative to the total settler population, and by roughly 1912, the sample size reaches a turning point and begins to decrease in size.

The year 1910 which marks the political unification of the two British colonies, the Cape Colony and Natal, and the two Boer republics, the Orange Free State and the South African Republic, seems an appropriate year up to which this sample could be used as a representative source of information on European settlers and their descendants in South Africa. A further limitation is that SAF do not follow individuals who emigrated from South Africa, nor is there any clear way of discerning outmigrants from those whose lineages ended for other, unrelated reasons. The year 1910 as a cut-off point is additionally useful, since it precludes users from possible violations of privacy regulations which protect the data for a period of one hundred years.

Figure 4 *Sample size versus population estimate*



Note: Log scale. Population size provided for years for which a population estimate is available.

Sources: Census of the colony of the Cape of Good Hope, 1856, 1865, 1891, 1904, 1910; Elphick & Giliomee, 1989; Ross, 1975; Sadie, 2000; and own calculations.

Beyond mirroring the general trend in population growth, Cilliers and Mariotti (2019) provide further comparisons of the age, sex, and regional distributions of SAF to census data, where possible, confirming that the database does not suffer from systematic compositional bias.

Of additional concern is partial or incomplete data on individuals. While the size and scope of the SAF data are its greatest advantage, it must be noted that not all entries contain complete information. Of the full dataset, which contains 671,385 observations, many entries are empty save for a name and surname. Close to two thirds of these entries contain a birth or a baptism date, while only one quarter contains a death date, and less than one fifth contains a marriage year. These statistics can be found in Table 2.

When individuals whose data are partial or incomplete are removed from the study in question, the sample size is substantially reduced. If we consider the SAF sample for which there are complete birth and death dates, it effectively captures approximately 30% of the total estimated population over time, reducing from around 1865 to about 10% around 1910 (see Figure 5). In addition, if there is a systematic relationship between the demographic event under investigation and the likelihood that information is incomplete, this will introduce additional bias to the study.

Table 2 *Frequency of observations in the dataset for selected variables*

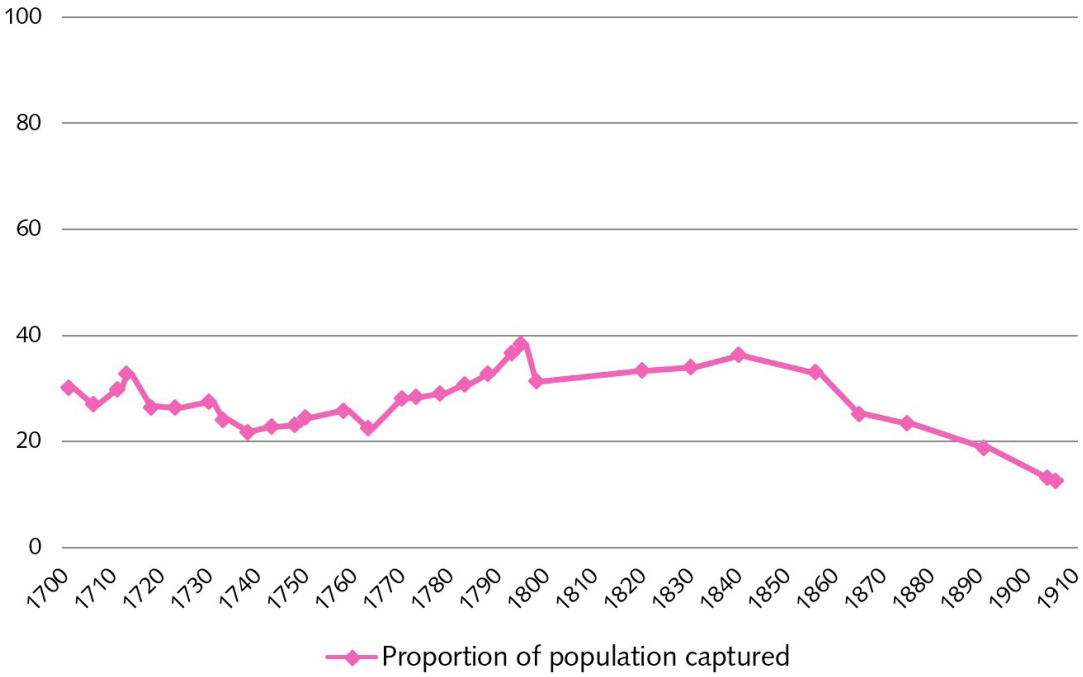
Variable name	Females	Males	Sex unknown	Total
Individuals	305,260	360,936	5,189	671,385
Individuals with known fathers	300,625	333,214	4,251	638,090
Individuals with known mothers	238,757	276,724	2,399	517,880
Individuals with known year of birth/baptism	219,089	255,902	2,029	477,020
Individuals with known year of first marriage	67,602	94,115	119	161,836
Individuals with known year of death	40,978	98,894	459	140,331
Individuals with all data known	13,771	32,977	12	46,760

6 SOCIO-ECONOMIC STATUS

The only measure of socio-economic status provided in SAF are occupations. Occupations in the database have been coded according to the Historical International Standard Classification of Occupations, hereafter HISCO (van Leeuwen, Maas, & Miles, 2002), and then classified according to the Historical International Social Class Scheme, hereafter HISCLASS (van Leeuwen & Maas, 2011). Bias arising from the incomplete or inconsistent reporting of occupations is of particular concern. However, comparisons with available census data reveal that the reporting of occupations do not appear to be systematically related to the relative ranking of certain occupations in society.

Comparing the white working age male population from Cape Colony censuses to estimates for equivalent time-periods from SAF in table 3, shows that although discrepancies exist between the SAF and the true occupational structure of the population, the general levels and trends are correlated. Still, occupations are typically not reported for women, and roughly only 10% of men in SAF have one or more occupations listed chronologically (not associated with a specific date or individual's age), providing a less than ideal measure of socio-economic status.

Figure 5 *SAF sample with complete birth and death dates as a proportion of the total settler population, 1700–1908*



Sources: Census of the colony of the Cape of Good Hope, 1856, 1865, 1891, 1904, 1910; Elphick & Giliomee, 1989; Ross, 1975; Sadie, 2000; and own calculations.

Table 3 *Share of the European/white working age male population with specified occupations, from available Cape of Good Hope censuses, compared to SAF by skill group*

Skill group	1850 SAF	1865 Census	1900 SAF	1911 Census
White collar	20.5	29.7	33.4	29.3
Farmer	64.4	55.3	49.8	47.8
Skilled/semi-skilled	8.4	7.5	13.9	19.0
Unskilled	6.8	7.5	3.0	3.8
N	1,602	48,485	5,327	493,562

7 LINKING TO EXTERNAL SOURCES

7.1 MANUAL RECORD LINKAGE: PROBATE INVENTORIES

Given the limitations of SAF in terms of refined socio-economic variables, additional sources can help to complete the database (see Figure 6 for the time-coverage of supplementary datasets). These required the development of suitable record linkage strategies. The database has already been manually supplemented with information from probate inventories compiled by the Master of the Orphan Chambers (MOOC). The Orphan Chamber was set up in 1673 and operated until 1834 and the inventories of the Orphan Chamber (MOOC 8-series) are an invaluable source for researchers interested in the lives of people at the early Cape. The inventories list all the possessions in a deceased estate, including livestock and slaves, and were a relatively complete and undisturbed reflection of households at the time of appraisal, which usually took place within days of death. In the rural districts, possessions were inventoried by neighbours, relatives or friends and sent to Cape Town. A clerk then copied the appraisal into a standard format, though the original details were retained (TANAP, 2010). The MOOC 8-series was manually linked to SAF based on individuals' unique first name(s) and surname strings and their birth and death dates (where available) resulting in the linkage of 2,117 of the 4,160 probate inventories, representing just over 50% (Fourie & Swanepoel, 2018).[1] This has proven to be a valuable addition to the database enabling the study of intergenerational transmission of wealth, crucially, for both males and females (Cilliers, Fourie, & Swanepoel, 2019).

Beyond their material wealth, Cape settlers held substantial shares of wealth in slaveholdings. The slave valuation and compensation records from 1834, the year slave abolition was enacted by the British Empire, can be found in the Cape Town Archives. They contain information on slaves (names, sex, age, place of birth, and value) and the slaveholders. Martins, Cilliers and Fourie (2020) manually linked slaveholder data to SAF for the Stellenbosch district. Linkage for the remaining districts of the colony is currently underway.

7.2 AUTOMATED RECORD LINKAGE: TAX CENSUSES

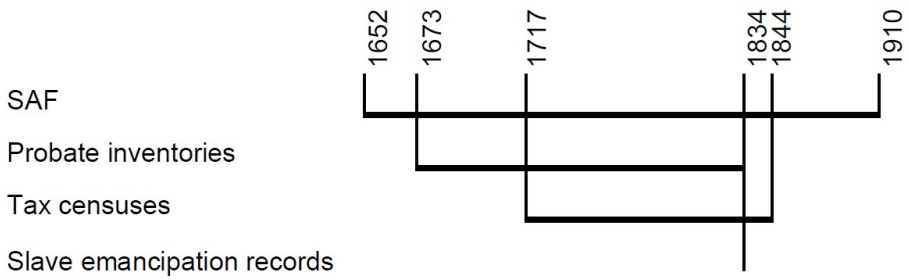
Further supplementary data comes from the *opgaafrollen*, annual tax censuses collected between 1663 and 1844, first by the Dutch East India administration and after 1795 by the British colonial administration, of all free households of the Colony. Household-level information includes name and surname of the head of the household and spouse, the number of children present in the household, the number of slaves and indigenous Khoisan employed, and several agricultural inputs and outputs, including cattle, sheep, horses, wheat sown, wheat reaped, vines, and wine produced. The series of *opgaafrollen* housed in the Cape Archives (1717–1844) are in the process of being transcribed under the umbrella of the Cape of Good Hope Panel project, a joint venture between Lund University and Stellenbosch University (Fourie & Green, 2018).

The *opgaafrollen* returns are, themselves, in the process of being linked across years to create an annual panel of household production. The data from one district served as the pilot study for the development of an automated probabilistic record linkage strategy that will soon be rolled out to the remaining districts (Rijpma, Cilliers, & Fourie, 2020). At the time of writing, individuals in SAF have been linked to household heads in the *opgaafrollen* for only the Graaff-Reinet district of the Cape of Good Hope Panel. Once complete, the Cape of Good Hope Panel will be the longest dataset of its kind in existence, spanning a period that stretches beyond any one lifetime. The inclusion of inputs such as household size and labor employed and outputs such as grain, wine and stock, allows for the testing of theories of economic growth and development, labor markets, industrial organization, political economy, migration and institutional economics, but also to develop and apply new econometric techniques.

The combination of the Cape of Good Hope Panel with SAF allows for a number of novel multigenerational studies. Firstly, it contains a heterogeneous group of individuals or households, whose behavior may be vastly different even within the same region. Secondly, following households over time allows for the study of reactions to changes in economic and social circumstances or (exogenous) institutional changes. With an intergenerational panel even more could be done: how these exogenous shocks affect families over multiple generations, and whether these processes are time-persistent and dependent on the initial conditions from which those families started out could be ascertained.

1 Non-unique name and surname combinations make linkage impossible since the correct individual cannot be selected from a list of possible candidates. Still, a linkage rate of 50% is generally considered to be high for historical data.

Figure 6 *Coverage of supplementary datasets*



Note: The line for the tax censuses represents the potential for linkage (only one pilot district has, at the time of writing, been successfully linked: Graaff Reinet district, 1786–1834), while all the other lines represent completed linkage between sources.

8 LOCATIONS

Certain dimensions of SAF remain unexplored. Locations are just one. Just over one third of all reported birth/baptism dates are accompanied by location information (residence of parents for the births and place of registration for the baptisms). For the most part these data are not yet cleaned or geocoded and therefore not yet available for public use. Where previous studies have made use of the location information from SAF, it has been made possible by drawing sub-samples from the database, for which location information was processed and assigned a relevant broader district categorization in lieu of exact co-ordinates. Geocoding all location information available in SAF is one of the priority steps to be taken in the further development of the database.

9 SELECTED STUDIES USING SAF

Genealogical data are particularly useful for the study of individuals, families, or communities across multiple generations. They are additionally well-suited for cohort analysis (Hollingsworth, 1969) since individuals belonging to the same cohort will have typically experienced the same vital event, birth or marriage for example, during the same period. With these advantages in mind and since data constraints define the limitation of studies, it is useful to ask which types of questions these data are best-suited to answer. Given the relative completeness of birth recording and the capacity to link individuals across generations, the obvious topics are fertility and intergenerational mobility.

Cilliers and Mariotti (2019) provide a complete series of female fertility estimates in the Cape Colony from 1700 to 1909. While previous research used portions of these data (Cilliers & Fourie 2012; Gouws, 1987) this was the first paper to use the full SAF database to date the onset of the South African fertility transition, which was found to have begun in the late 1870s to women born in the 1850s. Cilliers and Mariotti (2021) take further advantage of the longitudinal nature of the database to revisit the discussion on family limitation through stopping and spacing behavior prior to and during the fertility transition. Using split population estimation (cure models), the study finds that physiology and fecundity were the main determinants of both stopping and spacing behaviour prior to the fertility transition. The paper does not find evidence of explicit parity-dependent control for either stopping or spacing although some evidence of variation in birth interval lengths driven by postponement is found. During the transition, an increase in both stopping behavior and variation in birth interval length driven by postponement is found, followed by an increase in spacing after the transition.

Exploiting the intergenerational character of the data, Piraino, Mullier, Cilliers and Fourie (2014) investigate the intergenerational transmission of longevity between parents and offspring and find a positive and significant association between parents' and offspring's life duration, as well as between siblings. While these correlations persist over time, the magnitude of the effect is relatively small.

The effect of grandparents' longevity on that of grandchildren is insignificant, but cousin correlations suggest that inequality in longevity might persist across more than two generations. It was suggested that family and environmental factors shared by cousins could explain these results.

10 INTERMEDIATE DATA STRUCTURE

The SAF userbase has, for the most part, been limited to a handful of scholars interested in revising the existing historiography of the Cape Colony. While this is by no means an undeserving objective, this article serves to demonstrate that the value added by SAF could extend far beyond this. Most striking is the dearth of comparative studies emerging from this database. To attract a broader userbase of international historical demographers, and to facilitate comparative demographic studies, a standardization of sorts was warranted. It was for this reason that the decision was made to transfer SAF into the Intermediate Data Structure (IDS) (Alter & Mandemakers, 2014).

While IDS is not the only way to store and extract data, it is favoured by a growing number of longitudinal historical databases, not least because of its simple format that can suit many different types of data but also because it solves many of the problems related to the time-dependent nature of most historical demographic data. The principles of IDS involves two layers of data: 1) data about individuals and relations between individuals, and 2) data about contexts and relations between individuals and contexts. This design yields the five principal IDS tables. The INDIVIDUAL table, containing individual attributes; the INDIV_INDIV table, containing individual relations; the CONTEXT table, containing attributes of a geographical space where individuals reside together; the CONTEXT_CONTEXT table, containing how contexts are related to one another; and finally, the INDIV_CONTEXT table, relating the two layers of data.

Following the step-by-step guide provided by Klancher Merchant and Alter (2017), ENTITY and RELATIONSHIP tables were created for SAF. These are the necessary files to enable use of the IDS Transposer — an online tool which automatically transforms prepared data into the IDS standard. This produced two of the five IDS tables mentioned above: INDIV and INDIV_INDIV, since location information in SAF is not yet ready for wider use. These two "pilot" tables are now available for public use and following further development of the database it is hoped that full use of the IDS standard is on the horizon.

11 CONCLUSION

By shedding new light on the demographic characteristics of European settlers in 18th-, 19th- and early 20th-century Cape Colony, a severely under-researched topic in South African economic history, we can begin to move beyond a mere restatement of the history, producing results which not only challenge the existing understanding of South African historiography, but which add to the international debate around the nature and causes of demographic transitions. The limitations of genealogical data in terms of national representativeness and under-enumeration bias cannot be overlooked, but the research possibilities which capitalise on its highly valuable longitudinal and individual-level properties are boundless.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my various co-authors, Johan Fourie, Erik Green, Martine Mariotti, Igor Martins, Sean Millier, Patrizio Piraino, Auke Rijpma, and Christie Swanepoel who have contributed their time, effort, and expertise towards making the SAF database a functional source for future research.

I am grateful to George Alter and Luciana Quaranta for their advice and guidance with IDS conversion.

REFERENCES

Alter, G., & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for longitudinal historical microdata, version 4. *Historical Life Course Studies*, *1*, 1–26. doi: 10.51964/hlcs9298

Cape of Good Hope (South Africa). (1856). *Census of the colony of the Cape of Good Hope. 1856*. Cape Town: Solomon.

Cape of Good Hope (South Africa). (1866). *Census of the colony of the Cape of Good Hope. 1865.* Cape Town: Solomon.

Cape of Good Hope (South Africa). (1891). *Census of the colony of the Cape of Good Hope. 1891.* Cape Town: Solomon.

Cape of Good Hope (South Africa). (1905). *Census of the colony of the Cape of Good Hope. 1904.* Cape Town: Government Printer.

Cape of Good Hope (South Africa). (1911). *Census of the colony of the Cape of Good Hope. 1910.* Cape Town: Government Printer.

Cilliers, J., & Fourie, J. (2012). New estimates of settler life span and other demographic trends in South Africa, 1652–1948. *Economic History of Developing Regions*, *27*(2), 61–86. doi: 10.1080/20780389.2012.745663

Cilliers, J., & Fourie, J. (2014). Die huwelikspatrone van Europese setlaars aan die Kaap, 1652–1910. *New Contree*, *69*, 45–70. Retrieved from http://dspace.nwu.ac.za/handle/10394/10906

Cilliers, J., & Fourie, J. (2018). Occupational mobility during South Africa's industrial take-off. *South African Journal of Economics*, *86*(1), 3–22. doi: 10.1111/saje.12177

Cilliers, J., Fourie, J., & Swanepoel, C. (2019). 'Unobtrusively into the ranks of colonial society': Intergenerational wealth mobility in the Cape Colony over the eighteenth century. *Economic History of Developing Regions*, *34*(1), 48–71. doi: 10.1080/20780389.2019.1574565

Cilliers, J., & Mariotti, M. (2019). The shaping of a settler fertility transition: Eighteenth- and nineteenth-century South African demographic history reconsidered. *European Review of Economic History*, *23*(4), 421–445. doi: 10.1093/ereh/hey019

Cilliers, J., & Mariotti, M. (2021). Stop! Go! What can we learn about family planning from birth timing in settler South Africa, 1835–1950? *Demography, 58*(3), 901–925. doi: 10.1215/00703370-9164749

de Kiewiet, C. W. (1941). *A history of South Africa, social & economic*. Oxford: Clarendon Press; New York: Oxford University Press.

De Villiers, C. C. (1893). *Geslacht-register der oude Kaapse familien. Kaapstad: Van de Sandt de Villiers & Co.*

Elphick, R., & Giliomee, H. (Eds.). (1989). *The shaping of South African society, 1652–1840* (1st Wesleyan ed.). Middletown, CT: Wesleyan University Press.

Fourie, J. (2013). The remarkable wealth of the Dutch Cape Colony: Measurements from eighteenth-century probate inventories. *The Economic History Review*, *66*(2), 419–448. doi: 10.1111/j.1468-0289.2012.00662.x

Fourie, J. (2014). The quantitative Cape: A review of the new historiography of the Dutch Cape Colony. *South African Historical Journal*, *66*(1), 142–168. doi: 10.1080/02582473.2014.891646

Fourie, J., & Green, E. (2018). Building the Cape of Good Hope Panel. *History of the Family*, *23*(3), 493–502. doi: 10.1080/1081602X.2018.1509367

Fourie, J., & Swanepoel, C. (2018). 'Impending ruin' or 'remarkable wealth'? The role of private credit markets in the 18th-century Cape Colony. *Journal of Southern African Studies*, *44*(1), 7–25. doi: 10.1080/03057070.2018.1403218

Fourie, J., & van Zanden, J. L. (2013). GDP in the Dutch Cape Colony: The national accounts of a slave-based society. *South African Journal of Economics*, *81*(4), 467–490. doi: 10.1111/SAJE.12010

Genealogical Institute of South Africa. (2008). *South African Genealogies*. Stellenbosch: Genealogical Institute of South Africa.

Genealogical Institute of South Africa. (2012). *South African Families*. Stellenbosch: Genealogical Institute of South Africa.

Gouws, N. B. (1987). The demography of whites in South Africa prior to 1820. *Southern African Journal of Demography*, *1*(1), 7–15. Retrieved from https://hdl.handle.net/10520/AJA16824482_8

Guelke, L., & Shell, R. (1983). An early colonial landed gentry: Land and wealth in the Cape Colony, 1682–1731. *Journal of Historical Geography, 9*(3), 265–286. doi: 10.1016/0305-7488(83)90183-4

Guelke, L. (1988). The anatomy of a colonial settler population: Cape Colony 1657–1750. *The International Journal of African Historical Studies*, *21*(3), 453–473. doi: 10.2307/219451

Heese, J. A. (1971). *Die herkoms van die Afrikaner, 1657–1867.* Kaapstad: A. A. Balkema.

Hollingsworth, T. H. (1968). The importance of the quality of the data in historical demography. *Daedalus*, *97*(2), 415–432. Retrieved from http://www.jstor.org/stable/20023820

Klancher Merchant, E., & Alter, G. (2017). IDS Transposer: A users guide. *Historical Life Course Studies, 4*, 59–96. doi: 10.51964/hlcs9339

Malherbe, D. F. du T. (1966). *Family register of the South African nation.* Stellenbosch: Tegniek.

Martins, I., Cilliers, J., & Fourie, J. (2019). Legacies of loss: The intergenerational outcomes of slaveholder compensation in the British Cape Colony. *Lund Papers in Economic History. Development Economics* (No. 197). Lund: Lund University, Department of Economic History. Retrieved from https://portal.research.lu.se/portal/files/61357218/LUPEH_197.pdf

Mitford-Barberton, I., & White, V. (1969). *Some frontier families: Biographical sketches of 100 Eastern Province families before 1840*. Cape Town: Human and Rousseau.

Neumark, S. D. (1957). *Economic influences on the South African frontier, 1652–1836*. Stanford, CA: Stanford University Press.

Penn, N. (2014). Casper, Crebis and the knegt: Rape, homicide and violence in the eighteenth-century rural Western Cape. *South African Historical Journal*, *66*(4), 611–634. doi: 10.1080/02582473.2014.925961

Piraino, P., Muller, S., Cilliers, J., & Fourie, J. (2014). The transmission of longevity across generations: The case of the settler Cape Colony. *Research in Social Stratification and Mobility*, *35*, 105–119. doi: 10.1016/j.rssm.2013.08.005

Rijpma, A., Cilliers, J., & Fourie, J. (2020). Record linkage in the Cape of Good Hope Panel. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *53*(2), 112–129. doi: 10.1080/01615440.2018.1517030

Ross, R. (1975). The 'White' population of South Africa in the eighteenth century. *Population Studies*, *29*(2), 217–230. doi: 10.1080/00324728.1975.10410200

Sadie, J. (2000). *The economic demography of South Africa* (Doctoral dissertation). Stellenbosch: Stellenbosch University. Retrieved from http://hdl.handle.net/10019.1/51963

Simkins, C., & van Heyningen, E. (1989). Fertility, mortality, and migration in the Cape Colony, 1891–1904. *The International Journal of African Historical Studies*, *22*(1), 79–111. doi: 10.2307/219225

Swanepoel, C. (2017). *The private credit market of the Cape Colony, 1673-1834: An investigation into the role of wealth, property rights, and social networks* (Doctoral dissertation). Stellenbosch: Stellenbosch University. Retrieved from http://hdl.handle.net/10019.1/100828

TANAP. (2010). *Towards a New Age of Partnership*. www.tanap.net

TEPC project. (2008). *Transcription of Estate Papers at the Cape of Good Hope Project.* /z-wcorg/.

van Duin, P., & Ross, R. (1987). *The economy of the Cape Colony in the 18th century*. Leiden: Centre for the History of European Expansion.

van Leeuwen, M. H. D., & Maas, I. (2011). *HISCLASS: A Historical International Social Class Scheme*. Leuven: Leuven University Press.

van Leeuwen, M. H. D., Maas, I., & Miles, A. (2002). *HISCO: Historical International Standard Classification Of Occupations*. Leuven: Leuven University Press.

Willigan, J. D., & Lynch, K. A. (1982). *Sources and methods of historical demography*. Cambridge, MA: Academic Press.

Zhoa, Z. (2001). Chinese genealogies as a source for demographic research: A further assessment of their reliability and biases. *Population Studies*, *55*(2), 181–193. doi: 10.1080/00324720127690