Selection into the Great Trek*

Johan Fourie† Calumet Links‡

Abstract

We study selection into the Great Trek, the 1835–1840 migration of 12,000–14,000 Dutch-speaking colonists from the British Cape Colony into the southern African interior. To do this, we link Voortrekker genealogies to the 1825 Cape census and the British slave compensation records. Within districts, trekking households were indistinguishable from stayers in overall wealth but larger, with more children. Households with greater emancipation losses were no more likely to trek, and no economic characteristic predicts departure timing. Consistent with Hirschman's distinction between grievance and exit, the Trek selected large pastoral families with low moving costs, not the slaveholders emancipation harmed most.

Keywords: migration selection; Cape Colony; Hirschman; slavery; partible inheritance

JEL codes: N37; J61; C45; O15

*We thank Anton Ehlers, Erik Green, Albert Grundlingh, Gustav Hendrich and Dieter von Fintel for valuable comments on an earlier draft. We also thank seminar participants at Stellenbosch University and conference participants at the Economic History Society of Southern Africa annual meeting for helpful suggestions. Both authors acknowledge financial support from the Riksbankens Jubileumsfond (Cape of Good Hope Panel project: M20-0041). The usual disclaimer applies. This paper was created with the help of Anthropic's Claude Code and OpenAI's Codex. Cite this paper as: Fourie, Johan, and Calumet Links. 2026. "Selection into the Great Trek." Working Paper, Department of Economics, Stellenbosch University.

†Department of Economics, Stellenbosch University. Corresponding author. Email: johanf@sun.ac.za.

‡Department of Economics, Stellenbosch University.

1 Introduction

Between 1835 and 1840, an estimated 12,000 to 14,000 Dutch-speaking colonists left the Cape Colony and moved into the interior of southern Africa. The movement was large, roughly a fifth of the colony's European population, and its consequences were lasting: it led to the formation of settler polities beyond British jurisdiction, intensified conflicts over land and labor across the interior, and changed the political organization of the subcontinent (Walker, 1934; Muller, 1974; Giliomee, 2003). Yet a simple puzzle about the Voortrekkers has never been resolved with systematic evidence. They were not, by the standards of the Cape Colony, its richest inhabitants: the south-western wine farmers and large slaveholders overwhelmingly stayed. Nor were they its poorest, who lacked the wagons, oxen and supplies the journey demanded. What, then, distinguished those who left from those who stayed?

We show that household composition distinguished them. Using machine-learning record linkage (Feigenbaum, 2016; Abramitzky et al., 2021), we match Voortrekker genealogical records to the 1825 Cape Colony census (*opgaafrolle*); we then link the same Voortrekker records separately to the British slave compensation records compiled by Ekama et al. (2021). The resulting household-level dataset places 536 unique matched Voortrekker census households alongside 9,884 households that did not trek, across the districts that supplied most migrants. Across six comparison strategies, Trekker households had more children and were larger than non-Trekker households in the same districts. They also held, if anything, fewer slaves and sowed less wheat, differences that are directionally consistent across designs but individually imprecise. This is the profile of pastoral frontier households rather than of a slaveholding elite: families whose production required land, and whose many sons would soon require grazing of their own.

The demographic-pressure interpretation is long-standing in the historiography. Keegan (1996) argued that frontier families were pastoral producers whose production system required geographical expansion. Muller (1974) placed land scarcity at the center of his taxonomy of material causes, documenting that by the 1830s many frontier farmers could not obtain secure title to land. Venter (1985) described a "hunger for land" driven by the shift from cheap loan farms to expensive quitrent tenure after 1814, compounded by drought, locusts and a chronic shortage of surveyors that left farmers waiting years for title deeds. Contemporary Voortrekkers said as much themselves: Sarel Cilliers recorded that he and 72 "householding people who had no land" had petitioned the Governor for permission to settle beyond the Orange River, "and it was refused to us." Yet despite these qualitative pointers, the demographic-pressure interpretation has never been tested against individual-level data. We provide that test here.

The most prominent alternative explanation centers on slave emancipation. The abolition of slavery in 1833, implemented from December 1834, freed approximately 39,000 enslaved persons in the Cape Colony. Compensation fell far below Cape valuations, and many farmers received as little as a fifth of what they expected (Draper, 2010; Binckes, 2013). The Afrikaner nationalist tradition, articulated most influentially by Muller (1974) and Van Jaarsveld (1951), treated emancipation as central to the Trek. Anna Steenkamp's journal entry remains the most quoted statement on the

subject: "it is not so much their freedom that drove us to such lengths, as their being placed on an equal footing with Christians". If the emancipation grievance was the primary driver, those who trekked should have been at least as invested in slaveholding as those who stayed. We test this prediction directly. Using the compensation records, we find no evidence that households with larger financial exposure to emancipation were more likely to trek. Whatever its force as ideology, the emancipation narrative is not supported by the micro-data on economic selection.

These two findings cohere under Hirschman's (1970) framework: exit is exercised not by those most aggrieved by institutional change but by those for whom the cost of exit is lowest. Large pastoral households held mobile assets (cattle, wagons, family labor) and little investment in the immobile infrastructure of the slave economy; the largest slaveholders, for whom exit was costliest, stayed. Steenkamp's protest over racial equalization and the ideological grievances in Retief's manifesto were plausibly pervasive across the frontier, but grievance did not determine who acted on it. The Great Trek is, in this sense, a direct application of the Hirschman framework.

The paper makes two contributions. First, to the historiography of the Great Trek and, by extension, to the comparative literature on elite responses to compensated institutional reform. The causes have been debated for nearly a century, with successive generations of scholarship contesting whether the decisive factor was land scarcity, emancipation, frontier insecurity, racial equalization or the demand for self-governance (Van Jaarsveld, 1951; Muller, 1974; Keegan, 1996; Legassick, 2010; Giliomee, 2003). We provide the first systematic quantitative test. The positive finding, that large pastoral households with many children were disproportionately selected into the Trek, is directionally consistent across six comparison strategies, becomes stronger under tighter matching, and survives multiple-testing correction under the false-discovery-rate procedure. The second contribution, to a wider literature, is the null on emancipation. The British program of 1833–34 is one of the best-documented episodes of compensated expropriation in modern history (Draper, 2010; Hall et al., 2014), yet individual-level evidence on whether the most exposed elites exit has been unavailable. Linking the Voortrekker records to the compensation accounts compiled by Ekama et al. (2021), we find that households with larger losses were not more likely to trek: the compensated elite did not exit, and those who did were not the ones whom emancipation had most directly harmed. Hirschman's distinction between exit, voice and loyalty, applied here for the first time to the Great Trek, redirects attention from the intensity of grievance to the capacity to leave, and reconciles the older material and ideological accounts within a single interpretive structure.

Second, to the economic history of migration selection. A large comparative literature, surveyed by Hatton and Ward (2024), establishes that selection is shaped by origin constraints, destination opportunities and the costs of movement, and by the *dimension* along which it is measured (Leopold, Ruhose and Wiederhold, 2025). Linked historical microdata show patterns that vary across settings: negative wealth selection among European emigrants to the United States (Abramitzky, Boustan and Eriksson, 2012), occupational upgrading in nineteenth-century Argentina (Pérez, 2017), positive selection in the Great Migration of African Americans (Collins and Wanamaker, 2014, 2015; Derenoncourt, 2022), neutral to slightly negative selection among native-born US internal migrants

(Zimran, 2024), flattening positive selection as Irish networks lowered information costs (Conor, 2019), self-selection in early modern Spain (Beltrán Tapia and de Miguel Salanova, 2017), Swedish emigration to the United States (Dribe, Eriksson and Helgertz, 2022), and the same framework extended to late-twentieth-century Latin American migration to the US (Hanson, Orrenius and Zavodny, 2023). A parallel strand on *forced* movement, exemplified by Becker et al. (2020) on postwar Polish expellees, documents long-run consequences but takes the fact of movement as given. A second strand traces the cultural and institutional legacies of settler migrations (Natkhov and Vasilenok, 2021; Blum, Krauss and Myeshkov, 2022; van Leeuwen and Maas, 2022; Bazzi et al., 2023). Our contribution to this body of work is twofold. It is an African case, in a literature dominated by transatlantic and North American movements. And it is an *early* case: most settler migrations studied in this tradition belong to the late nineteenth or early twentieth century, and the Great Trek predates that wave by two or three generations. The word *trek* entered English through precisely this episode, a sign the migration was distinctive enough that no prior term existed. Selection in the 1830s Cape therefore operated under transport costs, state capacity and information networks different from those facing later settler migrants; the relevant unit of selection differs as well, since a pastoral frontier shifts the locus of decision from an individual worker to a household whose portfolio of mobile and immobile assets determines whether exit is feasible.

The remainder of the paper is organized as follows. Section 2 outlines the political economy of the Cape Colony and reviews the causes of the Trek debated in the historiography. Section 3 describes the data sources. Section 4 presents the record-linkage strategy and the main evidence on selection. Section 5 tests the emancipation hypothesis. Section 6 examines migration timing. Section 7 explores heterogeneity by trek leader and destination. Section 8 reports robustness checks. Section 9 concludes.

2 Historical Background and Hypotheses

2.1 The Cape Colony and the Great Trek

By the 1820s the Cape Colony had been under British rule for nearly two decades. Its economy was strongly differentiated by region. In the south-western districts, especially the Cape, Stellenbosch and parts of Swellendam and Worcester, production centered on wheat and wine and depended heavily on slave labor (Worden, 1985; Ross, 1999). These districts were well connected to Cape Town and to export markets.

The eastern and northern frontier districts were organized differently. There, livelihoods depended on extensive stock farming over large semi-arid spaces, recurrent violence on the frontier and looser incorporation into the commercial economy of Cape Town (Van der Merwe, 1937; Keegan, 1996; Legassick, 2010). Slaveholding existed on the frontier, but on a smaller scale than in the western grain and wine districts. Many frontier households instead relied on Khoekhoe and other non-enslaved workers through contractual, indentured and informal coercive arrangements (Newton-King, 1999; Elphick and Giliomee, 1989). The distinction matters because it bears directly

on the standard claim that emancipation stood at the center of the Trek. If those who trekked came mainly from the pastoral frontier, then their interests may have differed systematically from those of the large slaveholders of the south-west.

This regional differentiation also shaped mobility. Frontier households required access to grazing, water, transport animals and labor. Van der Merwe's classic study of northward movement showed that mobility beyond the formal colonial boundary was already a long-standing feature of frontier life (Van der Merwe, 1937). The Great Trek was therefore not an abrupt break with a settled order. It was an escalation and reorganization of older forms of pastoral and exploratory movement under new political conditions.

The Trek itself comprised a sequence of departures between 1835 and 1840. The earliest parties, associated with Louis Tregardt and Hans van Rensburg, left from the northern frontier in 1835. Larger and more organized parties followed under Hendrik Potgieter, Gerrit Maritz, Piet Retief and Piet Uys, drawing especially from Graaff-Reinet, Uitenhage, Cradock (Somerset East) and adjacent districts (Walker, 1934; Muller, 1974; Giliomee, 2003). Family, neighborhood and church networks helped to structure these departures, and reconnaissance expeditions into the interior preceded the larger wagon parties. Destinations varied: some Trekkers settled north of the Orange River, others crossed the Drakensberg into Natal, and still others moved into the Transvaal. That heterogeneity is one reason the literature has struggled to identify a single cause that fits all Trekkers equally well (Muller, 1974; Legassick, 2010). Our genealogical records identify 917 adult male Voortrekkers born before 1810, together with information on wives, origin districts, trek leaders and departure years.

2.2 Why They Left: Causes and Competing Interpretations

Few questions in South African history have attracted as sustained a debate as the causes of the Great Trek. Five strands of the historiography compete for explanatory weight: land scarcity, slave emancipation, labor and racial equalization, frontier insecurity and the Sixth Frontier War, and the desire for self-governance. We review each and state the testable prediction it implies; our data speak most directly to the first two, while the others we can illuminate only indirectly.

2.2.1 Land scarcity and frontier closure

The Cape Colony's pastoral frontier had been expanding northward and eastward for over a century before the Trek. Van der Merwe showed that *trekboer* movement into the interior was a structural feature of a stock-farming economy that required ever-larger tracts of grazing land (Van der Merwe, 1937). The pace was steady rather than episodic: stock farmers had reached the Gamtoos by the 1770s, overrunning each successive official boundary, and by the 1810s and 1820s individual graziers were crossing the Orange seasonally in search of pasture. The state followed rather than led, converting movement it could not prevent into revenue, first through loan-farm fees and then through quitrent (Van der Merwe, 1937; Venter, 1985). By the early nineteenth century, that expansion was running into constraints: colonial authorities sought to fix the frontier, and the

eastern border zone was contested by Xhosa polities whose lands settlers coveted. The Great Trek, on this reading, occurred when a century-old dispersal reached a boundary that was now fixed.

Muller (1974) placed land scarcity at the center of his taxonomy of material causes, documenting that on the northeastern border of the Somerset district, four-fifths of landowners had not received their title deeds (*transporte*) despite having paid for them.[1] The institutional causes were older. Venter (1985) traces the problem to Cradock's 1814 reform, which replaced the old loan-farm system with expensive perpetual quitrent. All new farms had to be surveyed and titled; farmers paid an annual rent and faced a chronic shortage of surveyors that left many waiting years for title, while drought and locust plagues in the early 1830s compounded the pressure (Venter, 1985). For a household that had not yet received its title deed, the cost of leaving was correspondingly lower; there was less to abandon.

Demography compounded the pressure. Frontier families were larger than those in the settled western districts, with more children and more dependants. In a system of partible inheritance, each generation needed new land to sustain the pastoral livelihoods to which they were accustomed. Keegan (1996) stressed that these were not subsistence-level households but pastoral families whose economic logic demanded geographical expansion. When colonial policy foreclosed expansion within the colony, the interior offered an alternative. Sarel Cilliers recorded that he and 72 "householding people who had no land" petitioned the Governor for permission to settle beyond the Orange River, "and it was refused to us".[2]

Quantitative evidence from the pre-Trek frontier is consistent with this build-up, though it comes from the era when the frontier was still open or only beginning to close. In Graaff-Reinet between 1798 and 1828, the number of children in farming households rose as land grew scarce, with poorer households substituting family labor for hired labor (Cilliers and Green, 2018), and larger households were the least likely to leave their district (Nel, 2020); early arrivals, who had claimed the best land, saw their advantage eroded after the 1820s by capital-rich British immigrants (Cilliers, Green and Ross, 2023). These studies end where ours begins: once closure foreclosed expansion within the Colony, the question is what the same large pastoral households did when staying no longer sustained their livelihoods.

Partible inheritance itself was not a frontier improvisation but the default of Roman-Dutch law, transplanted to the Cape by the Dutch East India Company in the seventeenth century and left intact by the British, who preserved the colony's private law after 1806. Intestate estates were divided equally among children, with the surviving spouse retaining half under community of property, and testamentary practice among Cape colonists overwhelmingly followed the same egalitarian convention. The rule persisted because, for most of the colony's history, it was cheap. In a pastoral economy where grazing was abundant and capital consisted chiefly of self-reproducing livestock, a father could endow several sons without fragmenting a fixed asset; equal division carried none of the costs that pushed European landowning classes toward primogeniture. The constraint

[1] Muller cites Col. Somerset's report of November 1835 to D'Urban, in Emigrant Documents, pp. 74–75.

[2] Cilliers, *Joernaal*, cited in Muller, *Die Oorsprong*, p. 225.

began to bind only when land ceased to be elastic, which is precisely what the quitrent reform, the surveyor bottleneck and the fixed boundary accomplished in the two decades before the Trek.

Why did land-constrained families respond by migrating rather than by having fewer children? The comparison with revolutionary France is instructive. Gay, Gobbi and Goñi (2026) show that where egalitarian inheritance was *imposed* on an already closed land frontier, as in France after 1793, the margin of adjustment was fertility: completed family size fell by roughly half a child, driven by the desire to avoid fragmenting land among heirs. At the Cape the sequence was reversed: partible inheritance was the long-standing default and the frontier was open, so expansion, not fertility restriction, had always been the cheaper margin. Closure then arrived within a single generation, far faster than fertility norms adjust, and, unlike in France, an alternative to demographic adjustment lay immediately beyond the boundary in the form of grazing land at negligible cost. Exit, on this reading, was the Cape's substitute for the fertility decline observed in France. We return to what our data can and cannot say about the fertility margin in Section 4.

The *pull* of the interior was equally important. The *Mfecane* had depopulated large areas of the highveld and Natal (Etherington, 2001); the "empty lands" narrative has been contested as partly a colonial construction that legitimized settler seizure (Cobbing, 1988; Hamilton, 1998; Wright, 2010). The interior was not empty, and Voortrekker settlement entailed displacement of African polities from the outset. What matters for the migration decision is that the interior offered grazing at negligible cost compared to the Colony's increasingly expensive quitrent farms. As one of us has argued elsewhere (Fourie, 2022), this price gap created a strong pull for large pastoral households whose demographic profile made them the most land-constrained within the existing colonial boundaries.

Testable prediction: If land scarcity and demographic pressure drove the migration, Trekkers should have been drawn from larger, more land-constrained households: families with more children, more dependants and a greater need for grazing.

2.2.2 Slavery, emancipation and compensation

The Slavery Abolition Act of 1833, implemented from 1 December 1834, freed approximately 39,000 enslaved persons in the Cape Colony. Compensation was determined centrally in London and proved substantially below Cape valuations, with many farmers receiving as little as a fifth of what they expected (Draper, 2010; Binckes, 2013).

The Afrikaner nationalist school, articulated most influentially by Muller (1974) and Van Jaarsveld (1951), treated emancipation as central to a broader story of colonial encroachment, and contemporary anger was intense. Governor Napier confirmed the centrality of this grievance to Glenelg: "Great numbers are highly discontented at the abolition of slavery... a vast body make great complaints, and give it as a reason for emigrating."[3]

The most quoted statement on slavery and racial equalization is the testimony of Anna Steenkamp, niece of Piet Retief: "The shameful and unjust proceedings with reference to the freedom of our

[3]Napier to Glenelg, quoted in Muller, *Die Britse Owerheid*, p. 76.

slaves; and yet it is not so much their freedom that drove us to such lengths, as their being placed on an equal footing with Christians, contrary to the laws of God and the natural distinction of race and religion."[4] Steenkamp distinguishes explicitly between emancipation itself and the principle of racial equalization, and assigns greater weight to the latter.

Yet there is contemporary evidence to the contrary. As early as January 1834, Civil Commissioner Campbell reported that farmers were planning to leave the colony with their slaves, but added: "assuredly the emancipation of the Slaves although it may be made as a pretex, has no influence on this movement, for out of the number who have been named to me as intending to depart, one only is a Slave owner".[5] Campbell's observation, that emancipation was invoked as a justification by people with minimal direct exposure to it, states the empirical question we address: the rhetoric of grievance was shared; the economic exposure was not.

Testable prediction: If emancipation was the principal driver, Voortrekkers should have been at least as heavily invested in slavery as those who stayed. Compensation records allow a further test: did those who received worse terms (a larger gap between valuation and payment) leave at higher rates?

2.2.3 Labor, equalization and British administration

A third cluster of grievances concerned the legal framework that the British imposed on labor relations. The most consequential reform was Ordinance 50 of 1828, which granted legal equality to the Khoekhoe and other free persons of color, removing older pass restrictions and limiting employers' coercive powers (Macmillan, 1929; Marais, 1939; Du Toit and Giliomee, 1983). The liberal and revisionist historians placed this at the center of the story: Macmillan (1927, 1929) argued that the frontier was a society organized around deeply unequal control of labor and legal standing, MacCrone (1937) placed racial attitudes at the core of the frontier mentality, and Marais (1939) offered the most direct formulation: "In its most important and most distinctive aspect the Great Trek was nothing else than the rebellion of the Boers against the ideas of the philanthropists."

Before Ordinance 50, a system of *inboeking* (indenture) had given frontier farmers control over Khoekhoe laborers and their children; after it, vagrancy increased, workers deserted more freely and employers found it harder to compel service. The Natal Volksraad complained to Napier that "the Boer, deprived of laborers, or, if he had any, deprived of all requisite control over them, all energy was taken away, so that he had the darkest prospect before him".[6]

The question of racial equalization (*gelykstelling*) ran through all of these grievances. In April 1834, the Uitenhage Dutch Reformed church council recorded "a great discontent among the congregation because marriage banns of Hottentots are published in the Church".[7] Among the members of that church council were J. J. Uys, father of Voortrekker leader Piet Uys, and Karel Landman,

[4] "Die Dagboek van Anna Steenkamp", Pietermaritzburg, 1939, p. 10; English translation from Bird, *Annals of Natal*, I, p. 459.

[5] Campbell to the Acting Cape Government Secretary, January 1834, quoted in Muller, *Die Oorsprong*, p. 354.

[6] Natal Volksraad to Napier, quoted in Muller, *Die Britse Owerheid*, p. 75.

[7] Uitenhage 1/1: Kerkraadsnotule, 1817–1842, p. 146, cited in Muller, *Die Oorsprong*, p. 200.

himself a future trek leader. Both trekked within three years. Earlier conflicts had produced the same grievances: the "Black Circuit" courts of 1812–1813, the Slagtersnek hangings of 1815 and the persistent friction with missionaries, particularly John Philip of the London Missionary Society (Muller, 1974).

Retief's manifesto, published in the *Grahamstown Journal* on 2 February 1837, articulated these grievances in condensed form: frontier insecurity; dissatisfaction with slave compensation; laws protecting freed slaves; missionary slander; the absence of "proper relations between master and servant"; and the wish to govern themselves (Binckes, 2013). The breadth of the manifesto's enumeration is itself significant: no single grievance is singled out as paramount.

Testable prediction: If labor grievances and racial equalization were central, Trekkers should have been disproportionately dependent on the Khoekhoe labor relationships most directly disrupted by Ordinance 50.

2.2.4 Frontier insecurity and the Sixth Frontier War

The immediate backdrop to the Great Trek was the Sixth Frontier War of 1834–1835, the most destructive of the Cape's border conflicts. In December 1834, large Xhosa forces invaded the eastern districts. The total capital loss to colonists, in livestock, homesteads and wagons, exceeded £290,000, yet the colonial government returned only £15,801 (Muller, 1963).[8]

An anonymous Uitenhage farmer gave a personal account: "What have the frontier inhabitants not had to endure from the [blacks][9] from time to time? In the last invasion I was left with only 5 oxen and 7 calves, and of the 74 milch cows the [blacks] left me not a single one… Literally nothing!"[10]

Du Toit and Giliomee (1983) argue that the Trek emerged from the cumulative pressures of a closing frontier: economic, political and psychological insecurity as mutually reinforcing rather than competing explanations. The war's aftermath compounded the grievance: Lord Glenelg reversed D'Urban's annexation of the Province of Queen Adelaide, returning to the Xhosa the territory the frontier farmers had fought and paid for (Giliomee, 2003; Binckes, 2013). The physical displacement caused by the war also lowered the psychological and logistical barrier to further migration.

Testable prediction: If frontier insecurity was paramount, selection should be strongest in the most war-affected districts. More broadly, frontier war losses should show up as negative wealth shocks among those who trekked.

2.2.5 Self-governance and political autonomy

A thread running through all of these grievances was a desire for political autonomy. The representative government that colonists had petitioned for was not granted until 1853. In a petition from

[8] Muller, *Die Britse Owerheid*, pp. 68–69, citing L.G. 169–174 and the Relief Commissioner's returns.

[9] The original quotation uses a racial slur that is highly derogatory in contemporary South Africa; we replace it here with [blacks].

[10] Anonymous Uitenhage farmer, quoted in Muller, *Die Oorsprong*, p. 187.

Winterberg and Koonap, the colonists stated: "We ascribe all these evils to one only cause, namely, the want of a representative government."[11] For Muller, the material and spiritual grievances were "intimately and inseparably fused" in their effect on the frontier farmer.[12] Van Jaarsveld (1951) folded the Trek into a narrative of republican striving, though he later acknowledged that national self-consciousness may have been "rather a consequence than a cause" of the Trek.

Of the five causes, emancipation is the one we can test most directly: slaveholding is recorded in the 1825 census, and the compensation records compiled by Ekama et al. (2021) provide an independent measure of each owner's financial exposure. Our data also speak to the demographic hypothesis. We cannot test self-governance or religious grievances directly, but we can inform the debate by establishing the economic and demographic profile of those who actually trekked.

3 Data

Our analysis draws on three primary data sources: the 1825 Cape Colony census (Opgaafrolle), Voortrekker genealogical records and the slave compensation records held at the UK National Archives. Figure 1 shows the ten districts of the Cape Colony in 1825 and the geographic distribution of Voortrekker records across them. The concentration of Voortrekkers in the eastern frontier districts (Somerset, Graaff-Reinet, Uitenhage and Beaufort) is clear, while the western wine-and-wheat districts contributed negligible numbers.

3.1 The 1825 Opgaafrolle

The Opgaafrolle were annual returns ("declarations") submitted by household heads throughout the Cape Colony for purposes of taxation and administration. For 1825, we have digitized returns from 11 districts: Albany, Beaufort, Cape, Clanwilliam, George, Graaff-Reinet, Somerset, Stellenbosch, Swellendam, Uitenhage and Worcester.[13]

The returns contain a rich set of variables for each household. Household heads are recorded by name. Wife names are recorded most consistently in Graaff-Reinet and Swellendam; where available, this substantially aids our record linkage, as we discuss in Section A. Economic variables include counts of livestock by type (horses, cattle, sheep, goats, pigs), enslaved persons, Khoekhoe servants and agricultural output (wheat sown and reaped, wine and brandy produced). Household composition is recorded as the number of settler men, women, sons and daughters; children appear only as counts of under-16 sons and daughters, so their ages are not observable. In total, the 1825 census contains 10,420 household records.

Because the census records quantities rather than values, and no household-level price data survive for this period, we summarize economic standing with a composite *wealth index*: the first

[11] Quoted in Muller, *Die Britse Owerheid*, p. 83.

[12] Muller, *Die Britse Owerheid*, p. 87.

[13] Somerset was established in 1825; the census we use is recorded under its earlier name, Cradock (1823). Colesberg was formally established only in the 1830s, carved from Graaff-Reinet. Our matching procedure accounts for these boundary changes by searching across related districts.

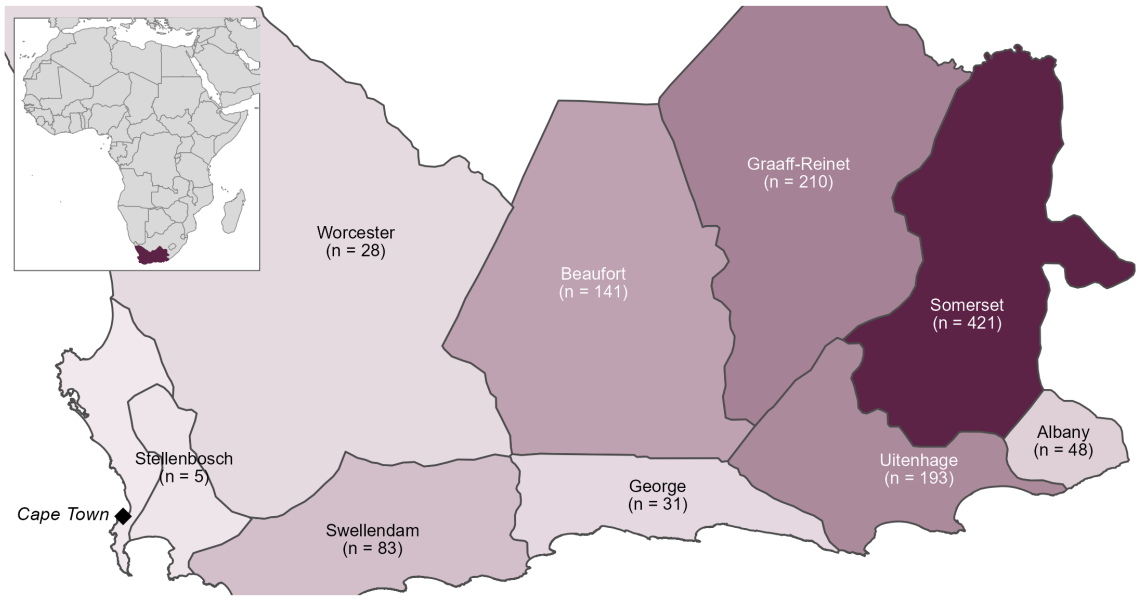

Figure 1: Voortrekker Records by District of Origin, Cape Colony 1825

Notes: Districts are shaded by the number of Voortrekker genealogical records originating there. The inset locates the Cape Colony at the southern tip of Africa. Colesberg (52 records), established only in the 1830s, is included in the Graaff-Reinet total ($n = 171$). Cradock (20 records) is included in the Somerset total ($n = 353$), as Cradock was renamed Somerset in 1825. Clanwilliam (13 records) is included in the Worcester total ($n = 20$), as it fell within Worcester's 1825 boundary.
Alt text: Choropleth map of the 1825 Cape Colony districts shaded by number of Voortrekker records, darkest in the eastern frontier districts of Somerset and Graaff-Reinet and lightest in the western districts, with an inset map locating the colony at the southern tip of Africa.

principal component of nine standardized census stock and output variables (horses, cattle, sheep, goats, pigs, enslaved persons, Khoekhoe workers, wheat reaped and wine). The first component explains 28.0 percent of the joint variance and loads most heavily on horses (0.53), wheat reaped (0.47), slaves (0.46), pigs (0.39) and Khoekhoe workers (0.33), so it captures the mixed-farming, labor-intensive end of the colonial economy; its standard deviation in the estimation sample is 1.59, the unit against which we express coefficient magnitudes throughout. Descriptive statistics for the index and all outcome variables are in Appendix B.

Table 8 reports the distribution of Voortrekker records and match rates across the nine districts of origin shown in the map (the Cape district appears in the map but supplied no Voortrekker records). The eastern frontier districts (Somerset with 353 Voortrekker records, Uitenhage with 146, Graaff-Reinet with 171 and Beaufort with 114) account for most Voortrekker origins, consistent with the historical literature's identification of the frontier as the principal source region of the Trek. The western districts of Stellenbosch (5) and Worcester (20, including Clanwilliam) contributed negligible numbers.

Figure 2 maps four district-level characteristics central to the analysis. The geographic pattern in panels (a)–(c) is the inverse of Figure 1: the districts that produced the most Voortrekkers, on the eastern frontier, are those with the fewest slaves, the lowest wealth and the smallest compensation losses per slave owner. Panel (d) reverses the gradient: frontier districts had the most children per household, consistent with the household-composition selection that our individual-level analysis confirms.

3.2 Voortrekker Genealogical Records

Our Voortrekker data are drawn from genealogical compilations that identify individuals who participated in the Great Trek, together with their family information. The raw dataset contains the full name and surname of each Voortrekker, the name and maiden surname of his wife (where known), birth year, origin district, the trek leader with whom he departed, the year of departure and the destination. The full file lists 2,313 unique Voortrekker individuals spanning multiple generations. We restrict our analysis to adult males born before 1810, who would have been at least 15 years old at the time of the 1825 census and plausibly household heads. Applying the 1810 birth-year cut-off, which excludes children born during or after the Trek as well as second-generation descendants born in the Cape after 1810, leaves 1,177 records; of these, 917 are adult males with name data clean enough for matching.

The ten-year gap between the 1825 census and the onset of the Trek in 1835 is an important limitation. Household wealth could have changed substantially over this decade: some households may have accumulated assets, others may have suffered losses from frontier conflicts or drought. We discuss this limitation further in Section 9 but note that the 1825 census is the most complete and reliable individual-level economic data available for the Cape Colony in this period.

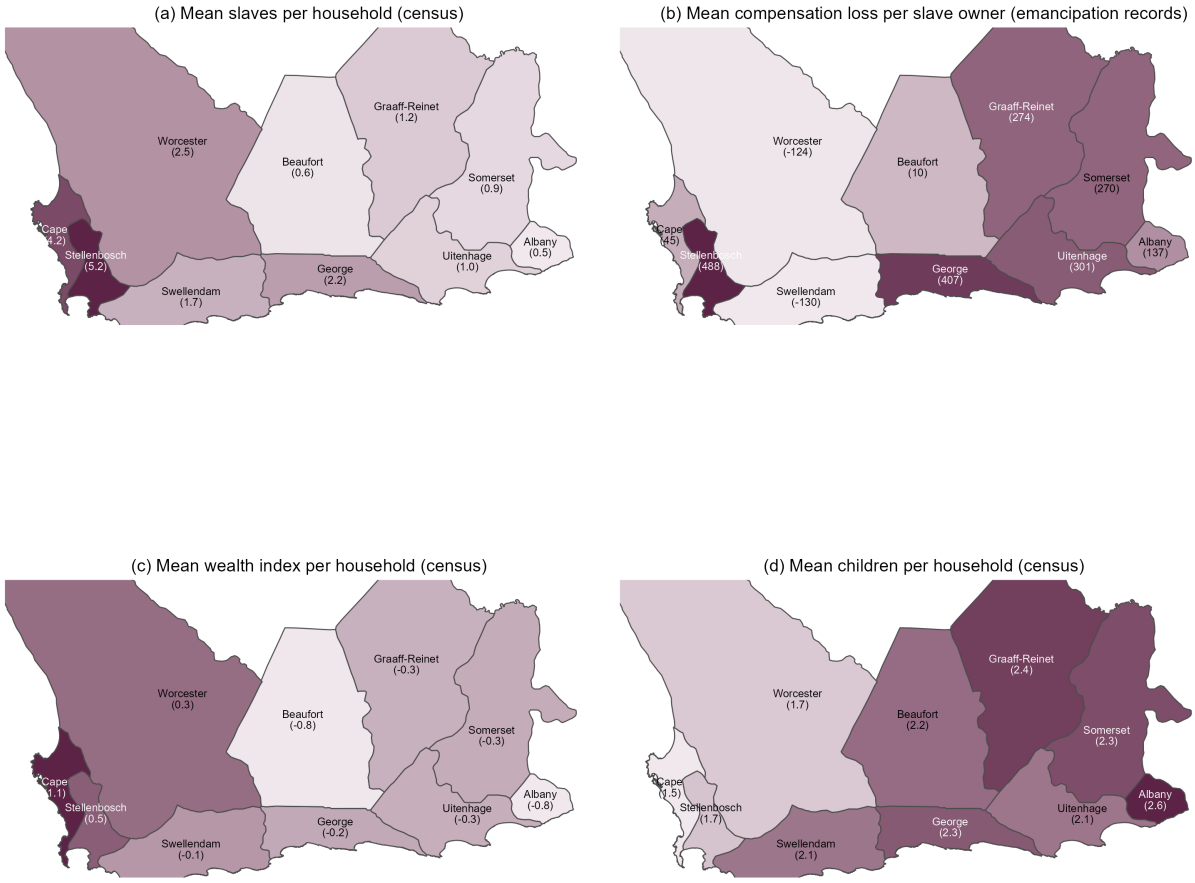

Figure 2: District-Level Characteristics, Cape Colony 1825

Notes: Panel (a) shows mean slave holdings per household from the 1825 *opgaafrolle*. Panel (b) shows mean compensation loss (valuation minus compensation, in pounds sterling) per slave owner from the compensation records (Ekama et al., 2021). Panel (c) shows the standardized wealth index defined in Section 3 (first principal component of nine standardized census stock and output variables). Panel (d) shows mean settler children per household. The Somerset census data come from the Cradock 1823 returns (the district was renamed in 1825). Clanwilliam is included in Worcester and Colesberg is included in Graaff-Reinet.
Alt text: Four district-level choropleth maps of the Cape Colony showing mean slaveholding, mean compensation loss per owner, the wealth index and mean children per household; the western districts are darkest on the first three measures and the eastern frontier districts darkest on children per household.

3.3 Slave Compensation Records

The third data source is the slave emancipation dataset compiled by Ekama et al. (2021) from the records of the Office of Registry of Colonial Slaves and the Slave Compensation Commission, held at the UK National Archives. Following the Slavery Abolition Act of 1833, colonial officials appraised the enslaved population for the purpose of compensating slave owners; the resulting valuation records (37,411 in total for the Cape Colony) document each enslaved person's owner, district, occupation, age and appraised value. Ekama et al. (2021) matched these valuations to the compensation claims processed in London, creating a dataset that records both what each owner's slaves were valued at and what compensation was actually paid. The compensation amounts were set centrally and bore an uneven relationship to Cape valuations, so that the gap between valuation and payment varied substantially across claimants (Draper, 2010). We link Voortrekker records to this dataset to test directly whether those who lost more from emancipation were more likely to trek. This linkage yields 577 Voortrekker–slave owner matches.

4 Who Were the Voortrekkers?

We link Voortrekker records to the 1825 census using a Random Forest classifier trained on expert-labeled candidate pairs (Rijpma, Cilliers and Fourie, 2020), followed by a second-stage review in which one author inspected all 917 best-candidate pairs. The procedure matches 558 of 917 Voortrekker records (60.8 percent), corresponding to 536 unique matched census households. Together with 9,884 non-Voortrekker census households, this gives a total analysis sample of 10,420 households.[14] Full methodological details, including blocking, features, training data, the four-tier probability thresholds, the one-to-one constraint, and the transparent review stage, are in Appendix A.

Before turning to the estimates, it is worth being precise about which question each kind of evidence answers. The claim that the Trek was not led by the colony's slaveholding elite has a *between-region* and a *within-district* component. The between-region component is settled descriptively, not by regression: the south-western wine-and-wheat districts, home to the large slaveholders, contributed almost no trekkers (Stellenbosch supplied 5 of the 917 Voortrekker records, Worcester 20), a fact shown in Figures 1 and 2. Our regressions cannot speak to this contrast: the district fixed effects sweep out all between-district variation by construction. What they answer is the narrower question that remains: *within* the frontier districts that supplied the Trek, who left and who stayed? All estimates that follow are within-district comparisons, and we interpret them as such throughout.

The migration-selection literature has established that comparing migrants to stayers requires careful attention to the counterfactual (Abramitzky, Boustan and Eriksson, 2012; Abramitzky and

[14] Every census household not linked to a Voortrekker record is classified as a non-Voortrekker, so the 359 unmatched Voortrekker records imply that at most 359 of the 9,884 control households (3.6 percent) are misclassified trekkers. This contamination attenuates every contrast toward zero and is therefore conservative for our positive findings; Appendix J treats the analogous contamination of the treated group.

Boustan, 2017). Raw comparisons can be misleading if migrants are drawn disproportionately from particular regions, and region-level differences in wealth can be mistaken for individual-level selection. We therefore report each outcome under four research designs of increasing stringency, presented side by side in Table 1: (1) a descriptive comparison to nearest census neighbors, (2) district fixed-effects regressions (our main specification), (3) exact matching within districts and (4) matching within districts on the exact number of children. The full per-design tables, in a common format that also reports group means, are in Appendix B.

4.1 Main Results across Four Designs

Table 1 presents the paper's central evidence in a single format: for an identical set of fourteen variables, it reports the Voortrekker–non-Voortrekker difference under each of the four designs, so that the reader can see directly how magnitudes and precision change as the comparison tightens. Column (1) compares each matched Voortrekker to its nearest census neighbors; column (2) reports the coefficient on the Voortrekker indicator from our main district fixed-effects specification; column (3) matches exactly on district; column (4) matches exactly on district and the number of settler children. Appendix B reports each design in full, including group means, and a descriptive-statistics table for all fourteen variables.

Column (1) provides a first comparison. On every wealth measure (livestock, slaves, wheat, wine and the composite wealth index) the difference between Voortrekkers and their census neighbors is statistically insignificant: Voortrekkers were economically indistinguishable from the households recorded on the adjacent census rows. The expectation that Trekkers were drawn disproportionately from the wealthier, more established farming households finds no support even in the simplest comparison available. The one clearly significant difference, settler men, is largely mechanical: the Voortrekker matching sample is restricted to adult males born before 1810, so every matched Voortrekker household is guaranteed at least one adult male, while the general census population includes female-headed households. We therefore treat settler men as a consistency check on the matching rather than as a substantive finding, here and in every design that follows.

4.2 District Fixed-Effects Regressions

The descriptive comparison does not account for systematic differences across districts. Since Voortrekkers came disproportionately from frontier districts, which were on average less wealthy than the western wine-and-wheat districts, comparing raw means conflates selection into the Trek with geographic sorting. Our main specification addresses this with district fixed effects:

$$Y_i = \alpha + \beta \cdot \text{Voortrekker}_i + \sum_d \gamma_d \cdot \text{District}_{id} + \varepsilon_i \tag{1}$$

where Y_i is the outcome variable for household i, Voortrekker_i is an indicator for matched Voortrekker status, and District_{id} denotes district fixed effects. The coefficient β captures the within-district

Table 1: Selection into the Great Trek across Four Research Designs

Variable	(1) Nearest neighbor	(2) District FE	(3) Exact match	(4) Family-size match
Livestock				
Horses	0.048	-0.506	-0.555	-0.661
	(0.909)	(0.169)	(0.122)	(0.119)
Cattle	174.5	-2120.8	-2147.1	-3588.6
	(0.363)	(0.356)	(0.354)	(0.666)
Sheep	-84.75	-100.6	-103.7	-108.3
	(0.677)	(0.468)	(0.460)	(0.695)
Labor				
Slaves	-0.197	-0.174	-0.176	-0.226
	(0.195)	(0.156)	(0.147)	(0.104)
Khoekhoe workers	-0.044	-0.136	-0.158	-0.309
	(0.882)	(0.599)	(0.527)	(0.201)
Agriculture				
Wheat sown (muids)	0.069	-0.042	-0.041	-0.073
	(0.260)	(0.444)	(0.451)	(0.282)
Wheat reaped (muids)	1.007	-0.761	-0.760	-1.106
	(0.203)	(0.286)	(0.284)	(0.204)
Wine (leaguers)	0.213	0.198	0.179	0.162
	(0.199)	(0.261)	(0.305)	(0.095)
Household composition				
Settler men	0.065***	0.052***	0.052***	0.053***
	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Settler women	-0.041	-0.007	-0.005	-0.038
	(0.079)	(0.727)	(0.800)	(0.053)
Settler children	0.152	0.311*	0.320*	0.000
	(0.386)	(0.029)	(0.025)	(1.000)
Household size	0.175	0.357*	0.367*	0.015
	(0.353)	(0.020)	(0.017)	(0.905)
Children ratio	0.004	0.029	0.029*	-0.003
	(0.844)	(0.051)	(0.043)	(0.824)
Composite				
Wealth index	-0.001	-0.047	-0.052	-0.073
	(0.976)	(0.263)	(0.206)	(0.093)

Notes: Each cell reports the Voortrekker–non-Voortrekker difference for the row variable under the column design, with the p-value in parentheses. Column (1): paired comparison of each matched Voortrekker to the households recorded immediately adjacent in the census (the *opgaafrolle* were compiled geographically, by ward and field cornetcy, so adjacent rows are typically spatially proximate); paired t-tests over the 519 of 536 matched households with at least one valid non-Voortrekker neighbor. Column (2): coefficient on the Voortrekker indicator from an OLS regression with district fixed effects and heteroskedasticity-robust (HC1) standard errors; N = 10,420 (536 matched Voortrekker households, 9,884 non-Voortrekkers). Column (3): exact matching on district with subclass weights; differences and p-values from weighted regressions with HC-robust standard errors. Column (4): exact matching on district and the exact integer number of settler children; the settler-children row is zero by construction. The wealth index is the first principal component of nine standardized census stock and output variables (Section 3). Horses and cattle are the census aggregates over all horse and cattle categories. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Full per-design tables with group means: Appendix B.

difference between Voortrekkers and non-Voortrekkers; standard errors are heteroskedasticity-robust.[15]

Column (2) of Table 1 presents our main results. Two patterns emerge.

First, the null on *composite* wealth holds once we control for district. The wealth-index coefficient of -0.047 ($p = 0.263$) amounts to -0.030 of the index's standard deviation of 1.59 in the estimation sample, with a 90 percent confidence interval of $[-0.074, +0.014]$ SD; the TOST procedure confirms equivalence at a bound of 0.10 SD ($p = 0.004$; Appendix E), though effects smaller than 0.05 SD cannot be ruled out. The composite null, however, combines components that differ: the Voortrekker coefficient is directionally negative on horses, cattle, sheep, slaves, and wheat sown and reaped, with wine the one exception (positive but insignificant), indicating that Voortrekkers held, if anything, less of the colony's slave-dependent agricultural inputs than their within-district neighbors. In this main specification the individual coefficients are not significant at conventional levels, but the slaves coefficient turns significantly negative under the stricter within-life-cycle conditioning of the district-plus-children FE specification (Table 15). The result, more precisely stated, is that Voortrekkers were not poorer overall, yet they tilted toward a different asset portfolio: fewer slaves, less wheat.

Second, the household composition variables show clear positive differences. Voortrekker households were larger, had more children and had a higher ratio of children to household size, with the individual sons and daughters coefficients also positive (0.212, $p = 0.048$, and 0.100, $p = 0.136$, respectively). The Khoekhoe workers coefficient is negative and statistically insignificant. These differences are economically meaningful: a Voortrekker household had, on average, roughly 0.4 additional persons and about 0.3 additional children relative to a non-Voortrekker household in the same district. Because we test several correlated household-composition outcomes, multiple-testing corrections matter, and we report them plainly. Under the conservative Bonferroni–Holm correction, only the (mechanical) settler-men coefficient clears the adjusted threshold; under the Benjamini–Hochberg FDR procedure, standard for correlated outcomes, household size, settler children and settler adults clear as well (Appendix E). Our substantive claim therefore rests not on any single p-value but on three mutually reinforcing facts: the household-composition coefficients are directionally consistent across all six comparison methods (Table 15); they become larger and more precisely estimated under tighter within-district matching (column (3) of Table 1); and they strengthen substantially in the non-wife-district test of Section 8.5, where the matching mechanism that could artificially favor large households is absent by construction. Household size and children, not settler men, are the substantive finding.

An important qualification applies to the settler men variable, and it is the reason we treat it as a consistency check. The coefficient is highly significant but reflects a near-binary variable: 99.81 percent of Voortrekker households have exactly one adult settler male, compared to 92.91 percent of non-Voortrekker households. The coefficient therefore captures the 6.9 percentage-point

[15] Appendix F discusses district-level clustering; the t-statistics on the household-composition findings are large enough to survive plausible inflation under clustering.

difference in the share of households with at least one male head present, not the presence of additional men, and much of that difference is built in by a matching sample that requires an adult male on the trekker side. A related concern is the comparison group: if an adult male was a de facto precondition for trekking, control households without one may be less meaningful counterfactuals. Restricting the controls to male-headed households and re-estimating column (2) confirms both halves of this reading (Appendix C). The settler-men coefficient vanishes entirely, as it must once every control household also contains a man, which is direct confirmation of its mechanical character. The substantive composition coefficients remain positive, with the modest attenuation expected from dropping 700 smaller, female-headed control households, and the wealth null is unchanged.

Figures 3 and 5 display these results graphically across all comparison methods. For wealth variables (Figure 3), standardized effect sizes cluster tightly around zero regardless of the method used, while the household composition variables (Figure 5) are consistently positive across all five methods that include them.

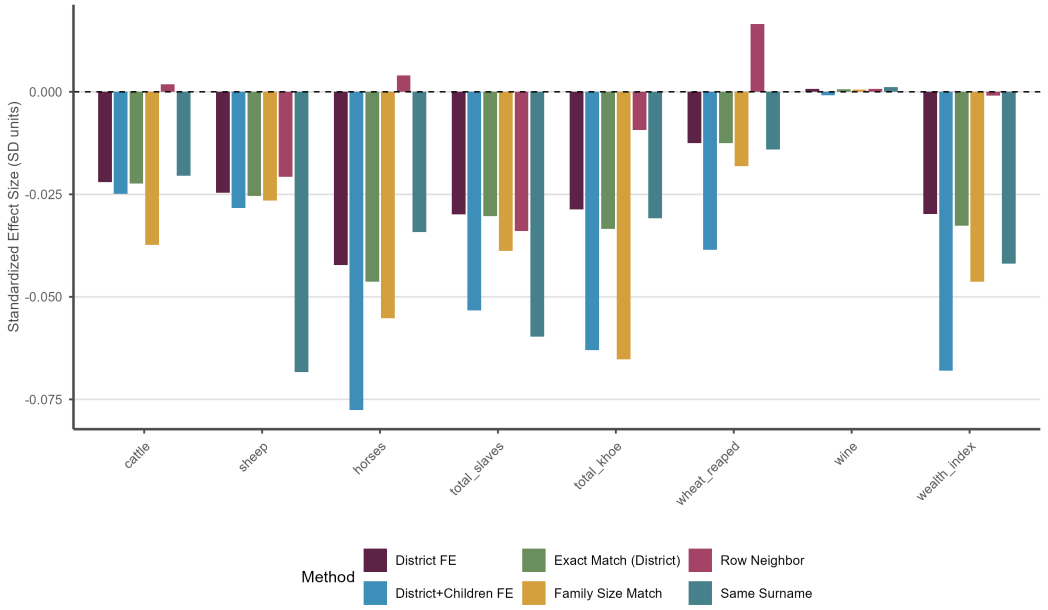

Figure 3: Standardized Effect Sizes across All Comparison Methods: Wealth Variables

Notes: Standardized effect sizes (in standard-deviation units) for wealth-related variables across all six comparison methods. Positive values indicate that Voortrekkers had higher values than non-Voortrekkers. The near-zero effects for the wealth index contrast with larger effects for Khoekhoe workers and, in the tighter matching methods, negative effects for slaves.
Alt text: Grouped bar chart of standardized Voortrekker–non-Voortrekker differences for eight wealth variables under six comparison methods; nearly all bars cluster close to the zero line.

4.3 Exact Matching within Districts

The district fixed-effects regressions compare each Voortrekker to the district average. Columns (3) and (4) of Table 1 tighten the comparison by matching Voortrekkers to non-Voortrekkers within the same district and, in a further step, within the same district *and* with the same number of children (as a proxy for life-cycle stage).

Column (3) reports the exact-match comparison using subclass weights. Reading columns (2) and (3) side by side makes their relationship transparent: the two designs target closely related within-district estimands, and the estimates barely move. The wealth and agricultural point estimates keep the same signs and approximate magnitudes as in the fixed-effects column (wealth index -0.052 against -0.047; slaves -0.176 against -0.174) and remain statistically indistinguishable from zero throughout; the wealth differences are thus modest in magnitude, stable in sign across designs, and imprecisely estimated, which is the pattern the equivalence tests formalize. What changes under exact matching is the household-composition margin: the settler-children, household-size and children-ratio differences are all somewhat larger and more precisely estimated than their column (2) counterparts.

Column (4) takes the comparison one step further by matching on both district and the exact integer number of settler children, thereby comparing Voortrekkers to non-Voortrekkers at a similar stage of the family life cycle. This is a deliberately stringent conditioning exercise: because family size is the conditioning variable, the settler-children difference is zero by construction, and household size and the children ratio collapse toward zero with it. The informative rows are the economic ones. Within district and life-cycle stage, Voortrekkers still own fewer slaves, sow less wheat and have a lower wealth index, differences that are directionally consistent with the portfolio pattern above but not individually significant at conventional levels. We therefore treat column (4) as a check that the composition finding is not merely repackaged wealth, rather than as a primary source of independent positive evidence.

4.4 Interpretation

The results are consistent across all specifications. At the 1825 baseline, Voortrekkers were not drawn from the established wealthy elite of the Cape Colony: the composite wealth index is statistically insignificant in every specification. Equivalence tests (TOST; see Appendix E) confirm a coefficient centered near zero and within the 0.10 SD equivalence bound. The equivalence result applies to this 1825 baseline; the intervening decade could have induced wealth selection by 1835 that we cannot reconstruct (we return to this caveat below).[16] Within their districts, however, Voortrekkers had a different asset composition: fewer slaves and less wheat. What distinguished them was household composition (larger families and more children) rather than overall wealth.

Because the interpretation turns on the distinction between kinds of capital, we define the

[16] We follow Lakens (2017); sensitivity across SESOI values is reported in Appendix Table 16. At 0.10 SD, half the magnitude of the negative selection documented by Abramitzky, Boustan and Eriksson (2012), equivalence is confirmed for all wealth variables.

terms precisely. *Pastoral capital* is livestock (trek-oxen, cattle, sheep) together with the household labor required to herd and move it; its defining property is mobility: herds could be driven to the frontier at low cost and reproduced there. *Mobile capital* adds wagons and portable equipment; for a trekking household the two categories are nearly coextensive, and we use the terms almost interchangeably for that reason. *Immobile assets* are vineyards, sown and improved arable land, farm buildings and, after 1834, slave property, which could not be taken beyond the colonial border and was convertible into mobile wealth only at a loss. The interior's advantage was correspondingly specific: open grazing suited to extensive cattle and sheep ranging at negligible land cost. It offered nothing to a wine or wheat producer whose capital was sunk in place and whose output required Cape Town's export market; nor were the trekkers cash croppers who would have faced the problem of transporting produce back from the interior, since within districts they sowed less wheat and produced less wine than stayers, and livestock, unlike crops, could be driven to market. Family labor belongs in this accounting as a factor of production: a large family with fewer purchased assets is poorer on a conventional asset count but not *uniformly* poorer, because its labor is the complementary input to livestock in a pastoral economy.

In Roy–Borjas terms (Borjas, 1987), the Trek exhibits no selection on overall wealth but clear selection on asset composition and household demography. The Cape rewarded immobile, slave-dependent agriculture; the interior rewarded pastoral, land-extensive activities. The model therefore predicts selection on households holding pastoral rather than agricultural capital and the demographic capacity to bear the fixed costs of migration, and the findings match: conditional on district, Trekkers were differently endowed, tilted toward family labor and away from slaves and crops, rather than uniformly poorer.

Hirschman (1970) provides a complementary interpretation. Exit is exercised not by those most aggrieved but by those for whom the cost of exit is lowest. The largest slaveholders had the most to lose from abandoning immobile assets and the highest exit costs; they stayed, many adapting by shifting to wage labor (Worden, 1985; Ross, 1999). The households that trekked were those for whom the Cape offered diminishing returns: large pastoral families with few slaves, many children needing land and limited ties to the commercial economy. This framing resolves the apparent tension: that slaveholders did not disproportionately leave does not mean emancipation grievance was absent; grievance did not determine who acted on it. We use the framework to synthesize the pattern rather than to derive it: the low-exit-cost group is identified ex post by the asset portfolio of those who left.

The ten-year gap. The 1825 census predates the Trek by a decade. Between 1825 and 1835 the Cape experienced Ordinance 50, the Sixth Frontier War (with £290,000 of losses concentrated in the east) and emancipation itself. A middling pastoral household in 1825 could have been destitute by 1835, which would mask negative wealth selection by the time of departure; without a post-war census we cannot resolve this directly. The gap is, however, far less threatening for our headline finding than it would be for a wealth claim, because household composition is more persistent over a

decade than livestock holdings or crop output. Children age but do not disappear: a household with six children in 1825 was a large household still in 1835, now with sons approaching the age at which partible inheritance and the land constraint bind. The demographic pressure we measure in 1825 would, if anything, have *intensified* over the decade, whereas herds could be raided and harvests lost. Two further pieces of evidence support this reading. The Voortrekker profile was already distinctively pastoral and demographically large in 1825, before any of the intervening shocks; and the 1830s census returns that survive for four districts produce household-composition patterns consistent with those from 1825, though their limited geographic coverage prevents a definitive test. Appendix K discusses the residual concern, within-district differential exposure to the war, in detail.

Demographic pressure. The household-size and children premium matches the testable prediction of Section 2.2: if land scarcity and demographic pressure drove the migration, Trekkers should have been drawn from larger, more land-constrained households. Three distinct channels are consistent with the finding. *Land scarcity*: under partible inheritance, larger families needed new land each generation, and the colonial frontier could not accommodate further expansion. *Logistical capacity*: trekking required adult sons to drive ox-wagons and manage herds. *Kinship networks*: brothers followed brothers, and extended family ties structured trek parties (Keegan, 1996). Our data cannot distinguish these channels. The first is a push factor, the second a facilitating condition, the third a coordination mechanism; all three are consistent with the demographic profile we observe. The data also cannot address the margin raised by the comparison with France in Section 2.2, the fertility decision itself: the census records children only as counts of under-16 sons and daughters, with no ages, and the genealogies record birth years for trekker lineages alone, so no completed-fertility measure exists for stayers.[17] Whether large families *chose* high fertility in anticipation of an open frontier, or simply found migration the cheaper margin once closure arrived, is a question we flag for future research.

The Khoekhoe-labor channel. The labor-grievance prediction of Section 2.2 implies that Trekkers should have been disproportionately dependent on the Khoekhoe labor relationships disrupted by Ordinance 50. The Khoekhoe workers coefficient in the district fixed-effects specification is negative and insignificant. Variance decomposition shows only 5.6 percent of variance lies between districts, and Voortrekker status is spatially concentrated in frontier districts where Khoekhoe labor was more prevalent; once district FE absorb cross-district differences, the unconditional positive association (0.61, $p = 0.012$) disappears. The within-district exact match yields a small positive difference, but conditioning further on family size eliminates it. We cannot test the labor-grievance channel directly. The 1825 worker count is a weak proxy for the Ordinance 50 grievance, which

[17] The genealogical numbering does not supply a reliable parent–child link table from which multi-generational family structures could be reconstructed, and any trekker-only fertility series would lack the comparison group that a test of fertility choice requires. Baptism-register data covering movers and stayers symmetrically could support such a test; we leave it to future work.

centered on the loss of coercive legal control over laborers from 1828 rather than on the number of workers employed three years earlier; no records document individual exposure to post-Ordinance 50 disruption, so the null is suggestive rather than decisive.

These findings bring individual-level evidence to bear on the multi-causal historiography (Muller, 1974; Du Toit and Giliomee, 1983; Keegan, 1996): the data support the land-scarcity and demographic-pressure channel and find no evidence for the narrow economic channel of the emancipation narrative. The broader ideological dimensions of emancipation grievance, which Steenkamp distinguished from the economic impact, cannot be measured in census data. The case also extends the comparative migration-selection literature: whereas Abramitzky, Boustan and Eriksson (2012) document negative wealth selection among European emigrants, we find no wealth selection but clear selection on asset composition and demographics, echoing Conor (2019)'s point that practical determinants of migration costs can matter more than wealth for who actually moves.

Figure 4 maps these two key findings geographically. Panel (a) shows that within-district differences in compensation loss between Voortrekkers and non-Voortrekkers are inconsistent in sign and magnitude; there is no systematic pattern of greater losses among those who trekked. Panel (b) shows that the children premium is present in most districts, with especially large positive gaps in Albany and Uitenhage, but it is not universal: Worcester, Stellenbosch and Graaff-Reinet are notable exceptions, and the Graaff-Reinet gap is negative.

5 Testing the Slave Emancipation Hypothesis

The most prominent explanation for the Great Trek, in both popular and scholarly accounts, is the emancipation of slaves in 1834. This section tests the emancipation prediction of Section 2.2 directly: if the standard narrative is correct, Voortrekkers should have owned more slaves than non-Voortrekkers, and those who suffered larger compensation losses should have been more likely to trek. The slave compensation records compiled by Ekama et al. (2021) make it possible, for the first time, to test these claims at the individual level: we observe how many slaves each owner held, the appraised valuation, the compensation received and the gap between the two. Crucially, slave ownership is measured from two independent sources, the 1825 census and the compensation records, compiled by different colonial authorities and linked to Voortrekker records through independent procedures; agreement between them cross-validates the finding.

5.1 Linking to Slave Compensation Records

We link Voortrekker records to Ekama et al. (2021)'s compensation dataset using a Jaro-Winkler similarity score that weights name and district information in an 80/20 combination and accepts matches above a 0.70 threshold, yielding 577 Voortrekker–slave owner matches. Of these, 339 involve Voortrekkers who are also linked to the 1825 census (the census-corroborated subset); the remaining 238 come from Voortrekkers not linked to the census. For each match we compute total valuation, compensation received, absolute loss (valuation minus compensation), percentage loss,

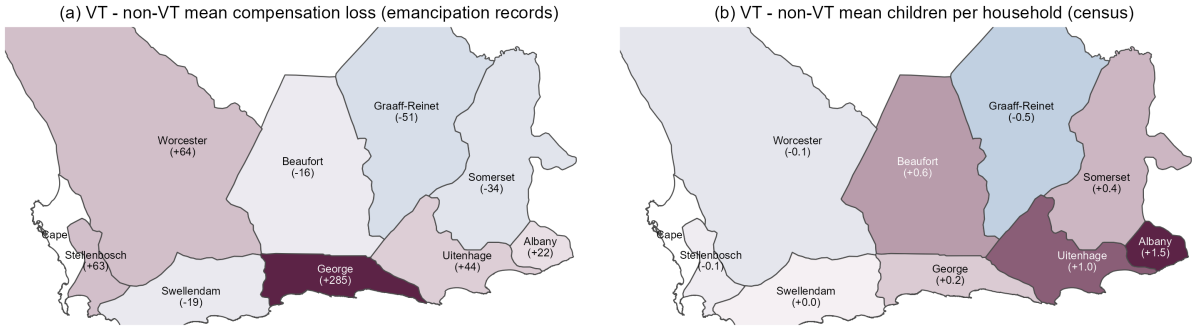

Figure 4: Within-District Differences: Voortrekkers vs Non-Voortrekkers

Notes: Each panel shows the difference in district means between Voortrekker and non-Voortrekker households (VT minus non-VT). Darker shading indicates larger absolute differences; labels show the signed value. Panel (a) uses the slave compensation records (Ekama et al., 2021): the difference in mean compensation loss (valuation minus compensation, in pounds sterling) per slave owner. Panel (b) uses the 1825 census: the difference in mean children per household. Districts shown in white have no Voortrekker matches in the relevant dataset (Cape in both panels). Colesberg is included in Graaff-Reinet, Clanwilliam in Worcester and the Somerset census data come from the Cradock 1823 returns.
Alt text: Two district-level maps of within-district Voortrekker-minus-non-Voortrekker differences: compensation losses show no consistent sign across districts, while the children difference is positive in most districts with Albany and Uitenhage largest.

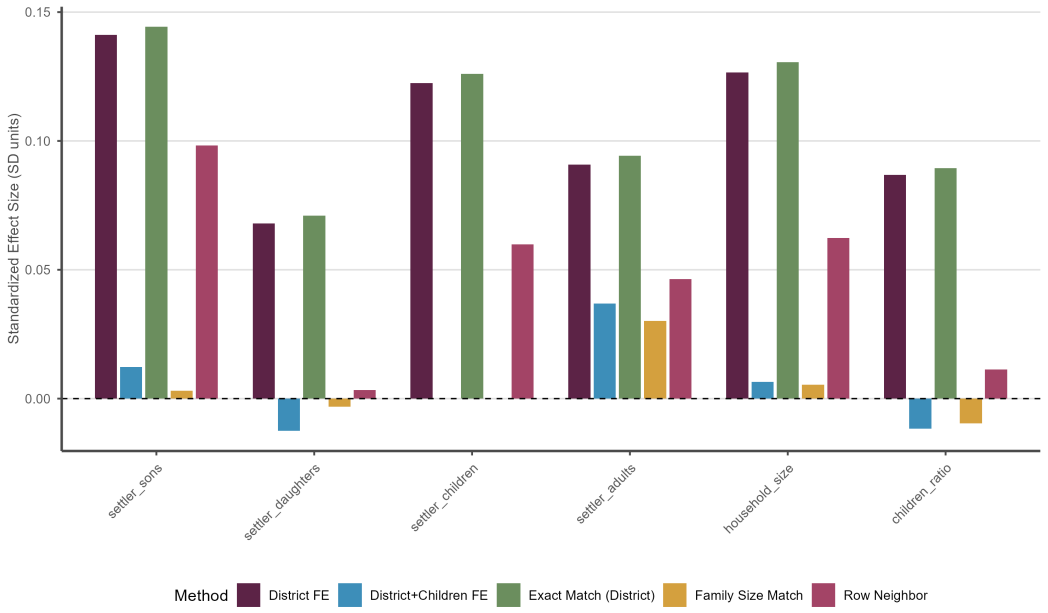

Figure 5: Standardized Effect Sizes across Comparison Methods: Household Composition

Notes: Standardized effect sizes (in standard-deviation units) for household composition variables across five comparison methods. Positive values indicate that Voortrekker households were larger on the relevant dimension. The household size premium is robust across all methods, driven primarily by more sons and children.
Alt text: Grouped bar chart of standardized differences for household-composition variables under five comparison methods; bars for household size, children and settler men are consistently positive.

and compensation rate. The compensation rate is of particular interest: as Draper (2010) documents, compensation was set centrally in London with an uneven relationship to local valuations, so if Voortrekkers received a systematically lower rate, they were disproportionately disadvantaged by an arbitrary process, which could have generated genuine grievance independent of total holdings.

5.2 Slave Ownership Comparisons

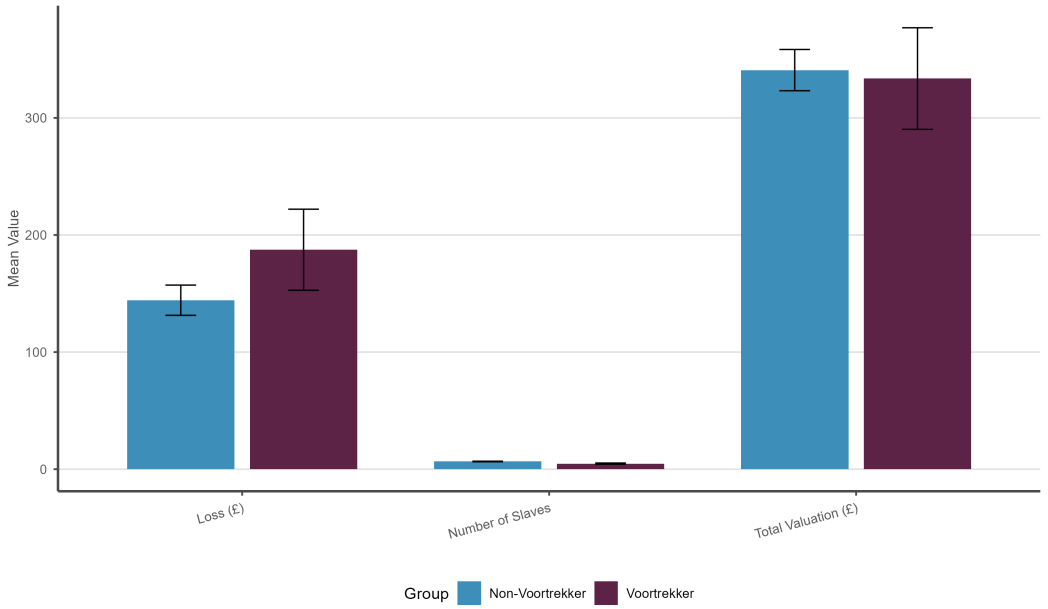

Figure 6: Slave Ownership: Voortrekkers vs. Non-Voortrekkers in the Compensation Records

Notes: Raw, unadjusted means of slave ownership and emancipation-loss variables for Voortrekker and non-Voortrekker slave owners in the compensation records. These descriptive comparisons do not condition on district. If the slave emancipation narrative were correct, Voortrekkers should be to the right of non-Voortrekkers on all measures.
Alt text: Paired bar chart comparing mean slave counts, valuations, compensation and losses for Voortrekker and non-Voortrekker slave owners; the two groups' bars are of similar height on every measure.

Figures 6 and 7 present the basic comparison. Figure 6 shows raw means; Figure 7 reports district-adjusted standardized coefficients. On the conditional comparison, Voortrekkers who owned slaves did not own systematically more than non-Voortrekker slave owners; their slave valuations, slave counts and absolute losses were not significantly different from, and if anything tended to be smaller than, those of non-Voortrekker owners in the same districts.

This result cross-validates the census findings in Section 4. The 1825 census estimates on slaves are negative in the district fixed-effects, exact-match and family-size designs alike (Table 1). Among slave owners in the compensation records, Voortrekkers again owned no more slaves, and if anything fewer. The agreement between two datasets compiled by different authorities and linked through independent procedures is evidence that the finding is genuine, not an artifact of record linkage.

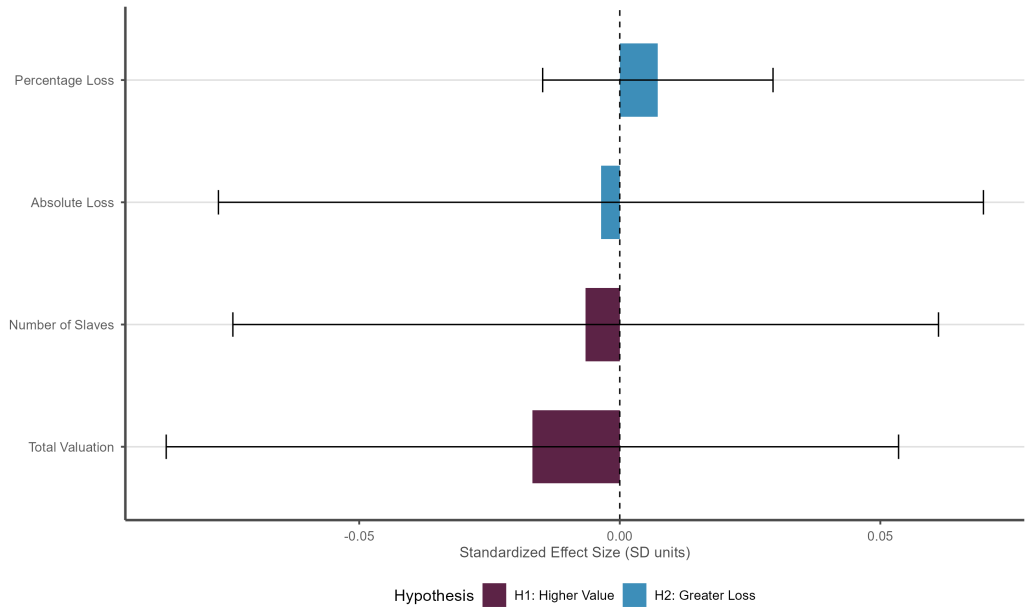

Figure 7: Effect Sizes: Emancipation Variables

Notes: Standardized district fixed-effects estimates for the comparison of Voortrekker and non-Voortrekker slave owners across emancipation-related variables. These coefficients condition on district, so they need not have the same sign as the raw descriptive gaps in Figure 6. Confidence intervals crossing zero indicate no significant difference.
Alt text: Coefficient plot of standardized district-adjusted differences in emancipation variables between Voortrekker and non-Voortrekker slave owners; every confidence interval crosses the zero line.

5.3 The Compensation Gap

The scale of slave ownership is only one dimension of the grievance. A subtler but potentially more important channel is the *arbitrariness* of the compensation process: even among owners of similar scale, some received a higher fraction of their assessed valuation than others (Draper, 2010). The compensation records uniquely permit a test of this channel: no other surviving source records both valuation and payment for individual owners. We test whether Voortrekkers suffered disproportionately from it.

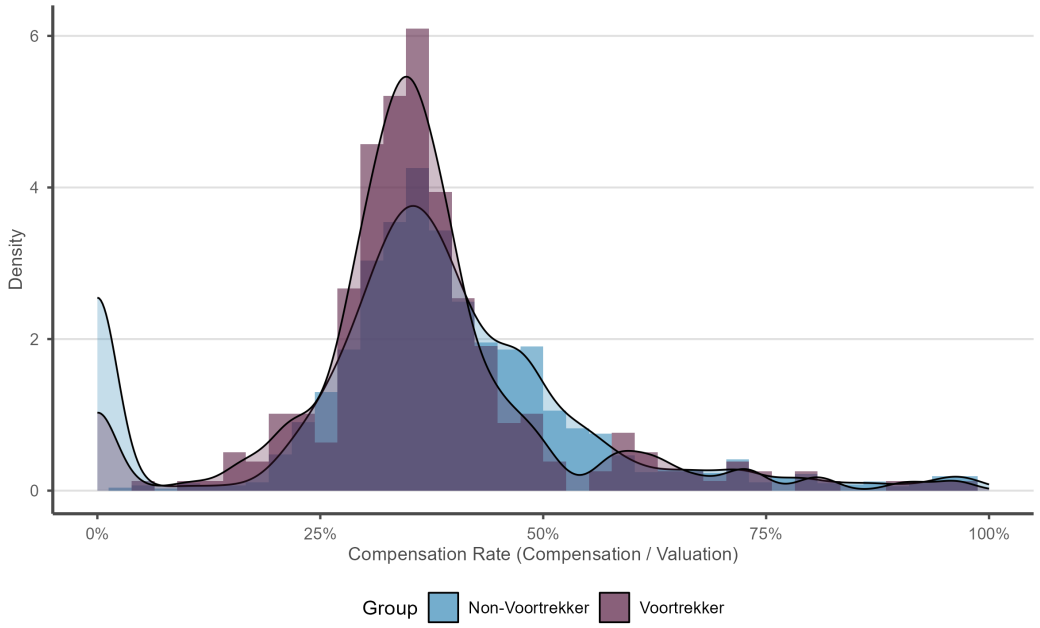

Figure 8: Distribution of Compensation Rates: Voortrekkers vs. Non-Voortrekkers

Notes: Kernel density estimates of the compensation rate (compensation received divided by assessed valuation) for Voortrekker and non-Voortrekker slave owners. A lower rate indicates a larger proportional loss.
Alt text: Two overlapping kernel density curves of compensation rates for Voortrekker and non-Voortrekker slave owners, nearly coincident across the whole distribution.

Figure 8 plots the distribution of compensation rates for Voortrekker and non-Voortrekker slave owners. The distributions are largely overlapping. Wilcoxon rank-sum tests confirm no significant difference in either compensation rates or loss percentages between the two groups. Within-district comparisons yield the same conclusion: in some districts Voortrekkers received slightly less favorable terms, in others slightly more favorable, with no systematic pattern.

5.4 Probit Analysis: Does Loss Predict Selection?

We formalize the test with probit models estimating whether emancipation losses predicted selection into the Trek among the population of slave owners. The key specification is:

$$\Pr(\text{Voortrekker}_i = 1) = \Phi(\alpha + \beta_1 \cdot \text{Loss}\%_i + \beta_2 \cdot \text{NSlaves}_i + \sum_d \gamma_d \cdot \text{District}_{id}) \tag{2}$$

where Loss% is the percentage of assessed valuation lost (valuation minus compensation, divided by valuation), NSlaves controls for the scale of ownership, and district fixed effects absorb geographic variation in both Trek propensity and compensation generosity. By controlling for the number of slaves, we isolate the "unfairness" channel: whether, holding scale of ownership constant, those who received worse terms left at higher rates.

Table 2: Probit Models: Does Emancipation Loss Predict Selection into the Trek?

	(1)	(2)	(3)	(4)	(5)
Loss % (std)	0.380	0.051	0.048		0.041
	(0.135)	(0.106)	(0.106)		(0.114)
Compensation rate (std)				−0.054	
				(0.111)	
N slaves (std)			−0.021	−0.025	−0.033
			(0.065)	(0.066)	(0.104)
Log valuation (std)					0.011
					(0.075)
District FE	No	Yes	Yes	Yes	Yes
N slaves control	No	No	Yes	Yes	Yes
N	3,473	3,473	3,473	3,473	3,473

Notes: Probit regressions with Voortrekker status (0/1) as the dependent variable, estimated on the sample of slave owners matched to the compensation records. All continuous predictors are standardized (mean 0, SD 1). Standard errors in parentheses. Model 4 additionally controls for the standardized mean per-slave valuation (coefficient omitted from display; insignificant). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 2 reports the results across five specifications. Model 1, without district fixed effects, shows a positive and significant relationship between loss percentage and Trek propensity ($\hat{\beta} = 0.38$, $p < 0.01$). This result is misleading, however: it reflects the fact that frontier districts had both higher Trek rates *and* different compensation patterns. Once district fixed effects are included (Model 2), the coefficient falls to 0.05 and is wholly insignificant ($p = 0.64$). Adding controls for the number of slaves (Model 3), using the compensation rate instead of loss percentage (Model 4) and adding log total valuation (Model 5) all yield the same conclusion: within districts, the compensation gap did not predict selection into the Trek.

The average marginal effects (AMEs) confirm the economic insignificance. In Model 3, a one-standard-deviation increase in loss percentage is associated with a +0.6 percentage-point change in the probability of trekking, and a one-standard-deviation increase in slave holdings with a −0.3 percentage-point change; in Model 5 the loss-percentage AME is +0.5 percentage points. All are

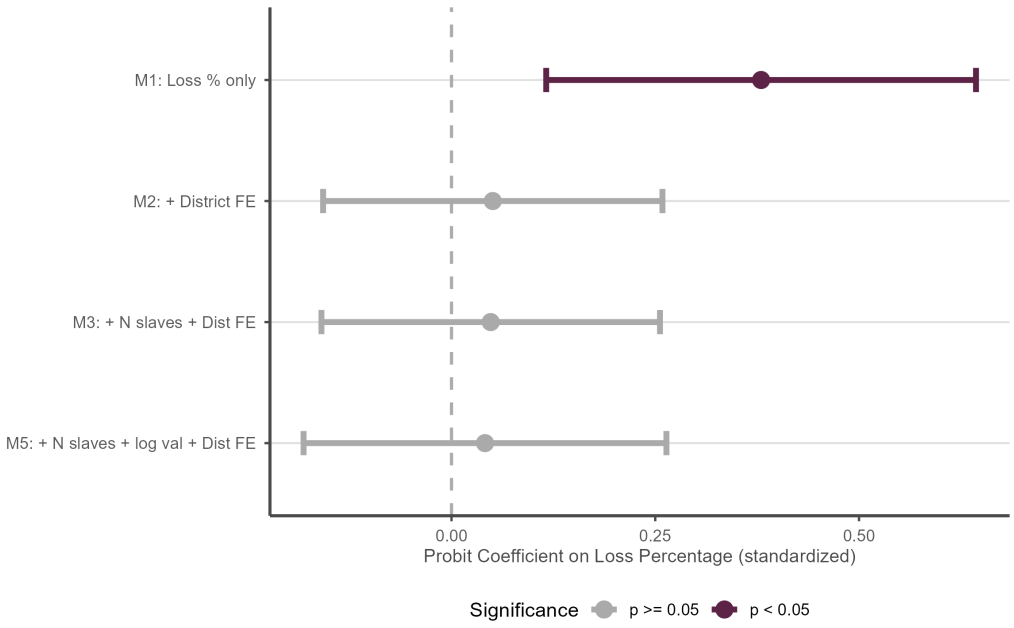

Figure 9: Probit Coefficients: Emancipation Loss as Predictor of Trekking

Notes: Coefficient estimates with 95% confidence intervals from probit models predicting Voortrekker status among slave owners. All specifications with district fixed effects show coefficients indistinguishable from zero.
Alt text: Forest plot of probit coefficients on emancipation-loss variables across five specifications; only the specification without district fixed effects lies clear of zero, and all others straddle it.

small relative to the 9.2 percent Trek rate among slave owners in the linked sample.

5.5 Loss Intensity: Quartile Analysis

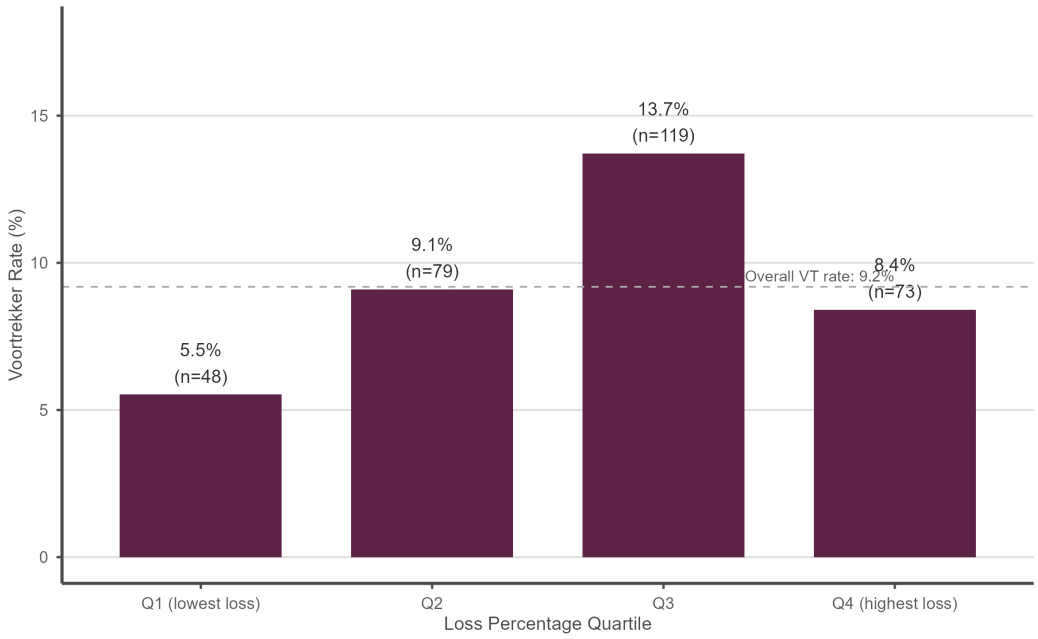

Figure 10: Voortrekker Rate by Loss Quartile

Notes: Proportion of slave owners who were Voortrekkers, by quartile of percentage loss from emancipation. Q1 contains owners with the lowest percentage loss; Q4 those with the highest. Labels show the Voortrekker rate and number of slave owners in each quartile. The dashed line indicates the overall Voortrekker rate among slave owners. If the grievance narrative were correct, we would expect a clear upward gradient from Q1 to Q4.
Alt text: Bar chart of the Voortrekker rate by emancipation-loss quartile; the four bars are of similar height around the overall rate, with no rising gradient from lowest to highest loss.

Figure 10 presents a non-parametric test of the loss-grievance hypothesis by dividing slave owners into quartiles of percentage loss and examining the Voortrekker rate within each quartile. If emancipation losses drove the Trek, we would expect a monotonically increasing relationship: those who lost the most (Q4) should have the highest Trek rate.

No such gradient is apparent. The Voortrekker rate does not increase monotonically across quartiles, and a chi-squared test finds no significant association between loss quartile and Trek propensity. The Spearman rank correlation between loss quartile and Voortrekker status is near zero and insignificant. A Fisher exact test comparing the top and bottom quartiles directly finds no significant difference in Trek rates. Linear probability models (OLS with robust standard errors) confirm the probit results: loss percentage, compensation rate and number of slaves are all insignificant predictors of Trek status once district fixed effects are included.

5.6 Interpretation

The emancipation analysis yields a clear conclusion: the compensation gap did not drive selection into the Great Trek. Two independent datasets support it. Voortrekkers did not own more slaves, did not receive systematically worse compensation terms and did not come disproportionately from those who suffered the largest proportional losses; if anything, the upper tail of the slaveholding distribution is disproportionately represented among stayers. Slave emancipation may have contributed to a general sense of grievance that made emigration culturally legitimate, but the specific economic channel, that material losses selected particular individuals into emigrating, finds no support.

The finding is consistent with the contemporary testimony in Section 2.2: Steenkamp herself distinguished between the freedom of slaves and their being "placed on an equal footing with Christians", assigning greater weight to the principle of equalization than to the material loss, and Campbell's 1834 observation that "one only is a Slave owner" among the earliest departures matches what we find at larger scale a decade later. More broadly, the result speaks to the literature on elite responses to institutional shocks (Draper, 2010; Hall et al., 2014): compensation, however imperfect, retained the largest slaveholders within the colony; the households that left were those for whom the Cape's constraints (limited land, a labor system in which they were marginal) mattered more than the losses from emancipation. The ideological framing of grievance need not reflect the economic profile of those who act on it.

6 Migration Timing

Not all Voortrekkers departed at the same time, and the second emancipation prediction of Section 2.2 has a timing dimension: if compensation losses drove the decision to trek, the most exposed households should have left at higher rates and, plausibly, earlier. The migration-selection literature similarly emphasizes that selection can change over the course of a migration episode, as information accumulates, networks develop and the costs and risks of moving evolve (Conor, 2019; Ward, 2017). If wealth or slave losses drove the decision, we might expect wealthier households or those with more slaves to have departed earlier (driven by stronger grievance) or later (needing more time to liquidate assets). We test whether any economic characteristic predicted the timing of departure.

Table 3 presents four specifications of the migration timing regression, using the 482 Voortrekkers for whom both census matches and departure years between 1835 and 1845 are available. The sample drops from 558 to 482 because departure year is not recorded in the genealogical sources for all Voortrekker households; a balance check finds no significant difference in observables between matched households with and without recorded departure years. The test is power-limited: given $n = 482$, the minimum detectable effect at 80 percent power ($\alpha = 0.05$) is approximately 0.26 SD for wealth and 0.18 SD for children, so modest timing effects below these bounds cannot be ruled out. In every specification, wealth, slave holdings and children are insignificant predictors of

Table 3: Determinants of Migration Timing

	(1)	(2)	(3)	(4)
Wealth (std)	−0.017	−0.036		−0.089
	(0.869)	(0.709)		(0.489)
Slaves (std)			−0.019	0.036
			(0.838)	(0.752)
Children (std)				0.082
				(0.368)
District FE	No	Yes	Yes	Yes
N	482	482	482	482
R^2	0.000	0.047	0.053	0.049

Notes: OLS regressions with move year as the dependent variable. Sample: matched Voortrekkers with recorded departure years in 1835–1845. All economic variables are standardized (mean 0, SD 1). p-values are reported in parentheses. Model 3 additionally controls for standardized cattle and sheep holdings (coefficients omitted from display; both insignificant). None of the economic variables of interest are statistically significant at conventional levels (all $p > 0.35$).

when a household departed, and the R^2 values are negligible without district fixed effects and only modestly higher with them, driven entirely by geographic variation in departure timing.

A further margin worth probing is how long trekker families had been settled where the census found them: recent arrivals might hold less accumulated wealth, so that the trekkers' middling wealth partly reflects recent mobility rather than a stable trait, and the most mobile families might have been quickest to leave once the frontier closed. The census records no birthplaces, so tenure cannot be compared between trekkers and stayers. The genealogies, however, record a birth or baptism place for a subset of trekkers, from which we construct a recency-of-arrival proxy: an indicator for heads born or baptized outside the district in which the 1825 census enumerates them. Among the 482 households in the timing sample, 198 record a place that maps to an 1825 district; 146 of these (74 percent) were born or baptized outside it, which indicates substantial mobility before the Trek.

Table 4 reports the results within this subsample. Recency does not predict departure timing, alone (column 1) or alongside the economic controls (column 2), so the timing null in Table 3 does not reflect an offsetting mobility gradient. Nor does recent arrival explain the trekkers' middling wealth: heads born or baptized outside their 1825 district were, if anything, wealthier than locally born trekkers in the same district (0.365 index points, $p = 0.004$; column 3), the opposite of what an accumulation account would predict. Coverage is limited and possibly selected, and baptism places are recorded at the congregation rather than the residence, so we read Table 4 as suggestive.

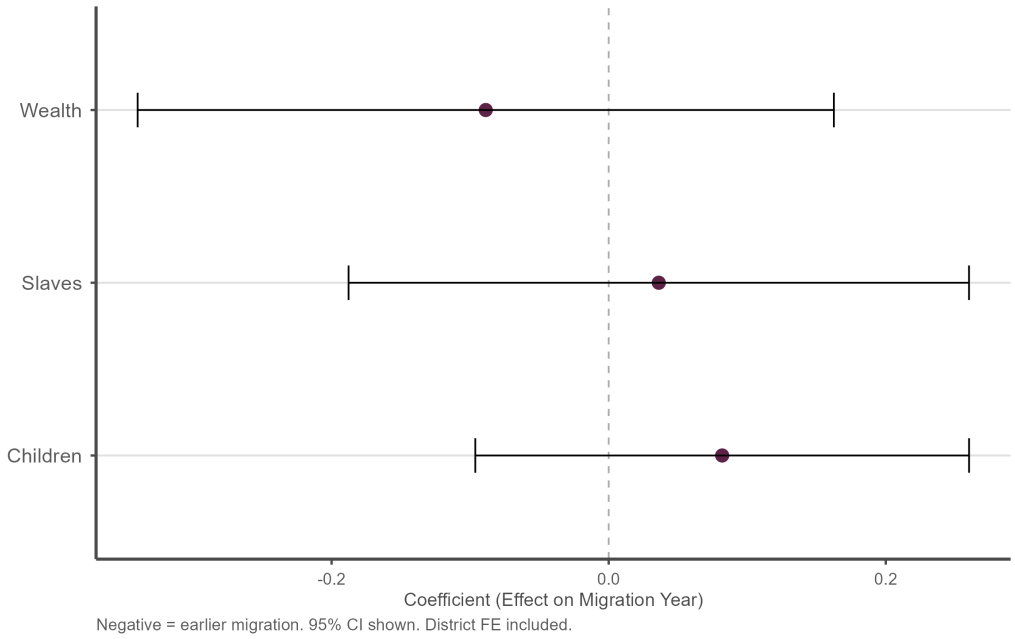

Figure 11: Determinants of Migration Timing: Coefficient Plot

Notes: Coefficient estimates from Table 3 with 95% confidence intervals. All confidence intervals include zero, indicating no significant relationship between economic characteristics and departure timing.
Alt text: Coefficient plot for wealth, slaves and children as predictors of migration year; all three confidence intervals cross the zero line.

Its face-value message is that the selection pattern we document is not an artifact of recent mobility.

Table 4: Recency of Arrival: Timing and Wealth within the Matched Voortrekker Sample

	(1)	(2)	(3)
Dependent variable	Move year	Move year	Wealth index
Born/baptized outside 1825 district	0.310	0.307	0.365**
	(0.393)	(0.399)	(0.126)
Wealth index (std)		-0.057	
		(0.205)	
Slaves (std)		0.133	
		(0.252)	
Children (std)		0.060	
		(0.274)	
District FE	Yes	Yes	Yes
Observations	198	198	198
R^2	0.116	0.117	0.166

Notes: Sample: matched Voortrekkers with a recorded departure year in 1835–1845 and a birth or baptism place that maps to an 1825 census district ($n = 198$ of 482). "Born/baptized outside 1825 district" equals one when the head's birth place (or, if missing, baptism place) lies outside the district of 1825 enumeration; foreign-born heads are coded as outside. Ambiguous frontier localities that span 1825 district boundaries are left unmatched rather than guessed. Heteroskedasticity-robust (HC1) standard errors in parentheses. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Early and late movers were economically similar; the data provide no evidence that wealth, slave holdings, children or recency of arrival predicted departure timing within these bounds. If emancipation losses had been the primary driver, we would expect the most aggrieved to have departed first; we do not see that. The absence of a timing gradient also distinguishes the Great Trek from settings where the dynamics of selection change over time as networks lower costs (Conor, 2019; Ward, 2017). In the Voortrekker case the selection mechanism appears stable across the 1835–1840 window, consistent with a structural driver rather than with a discrete policy shock whose salience faded with time.

7 Heterogeneity

The Great Trek comprised separate movements under different leaders, bound for different destinations. Disaggregating by leader and destination tests whether the selection patterns we document were driven by particular subgroups. The slave-emancipation narrative rests in part on the writings of Piet Retief, whose manifesto explicitly cited British interference with slave owners' rights (Giliomee, 2003); if so, his followers should have been wealthier and held more slaves than others.

7.1 Selection by Trek Leader

We identify seven major leaders with at least eight matched follower households: Potgieter ($n = 34$), Uys (20), Retief (17), Jacobs (16), Du Plessis (15), Maritz (12) and Landman (9). Cell sizes are too small for reliable inference, and *all results here are descriptive.* Six of the seven leaders' follower means fall below the non-Voortrekker mean (Figure 12); ANOVA finds no significant differences across leaders ($F = 1.94$, $p = 0.08$). Leaders differ mainly in asset composition (Figure 13): Potgieter's followers were markedly cattle-rich, Landman's most asset-poor. Retief's followers, whose leader's manifesto shaped the emancipation narrative, had the *lowest* mean wealth index (-0.54) and owned fewer slaves than the Voortrekker average.

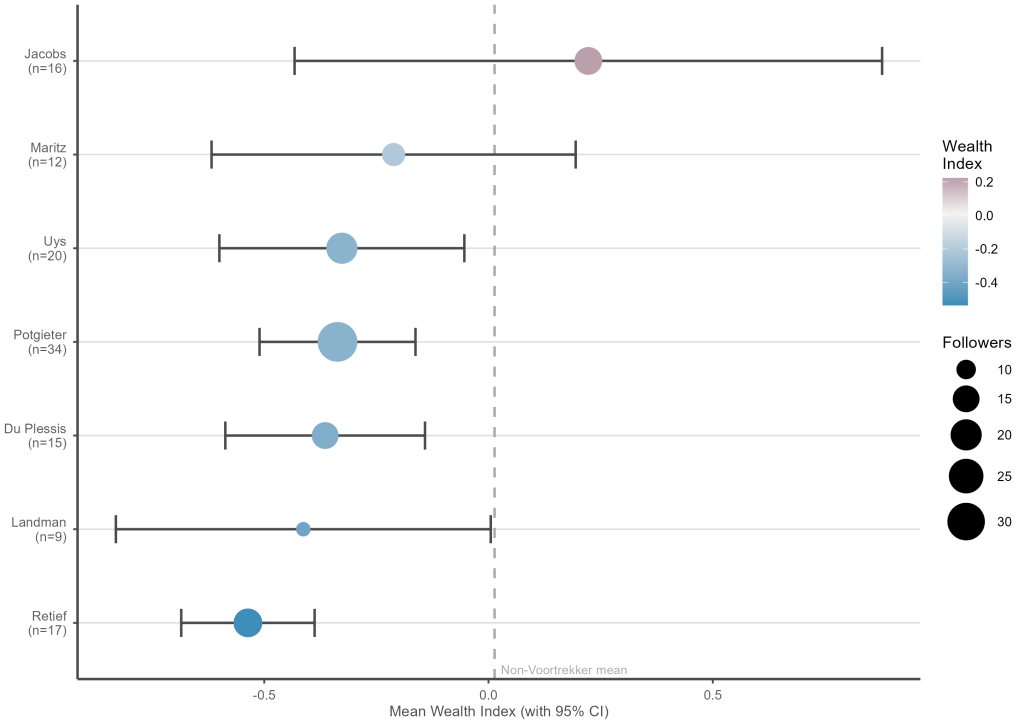

Figure 12: Mean Wealth Index by Trek Leader (with 95% Confidence Intervals)

Notes: Each point shows the mean wealth index for a leader's matched followers; horizontal bars show 95% confidence intervals. Point size is proportional to the number of matched households. Dashed line indicates the non-Voortrekker mean. Only leaders with at least eight matched followers are shown. Labels indicate the number of matched households per leader.
Alt text: Dot plot of mean wealth index by trek leader with confidence intervals; six of seven leader means lie below the dashed non-Voortrekker reference line, with Retief's followers lowest.

7.2 Selection by Destination

Different destinations offered different economic prospects. Natal, with its fertile lowlands and summer rainfall, attracted settlers interested in agriculture; the Transvaal and the territory that

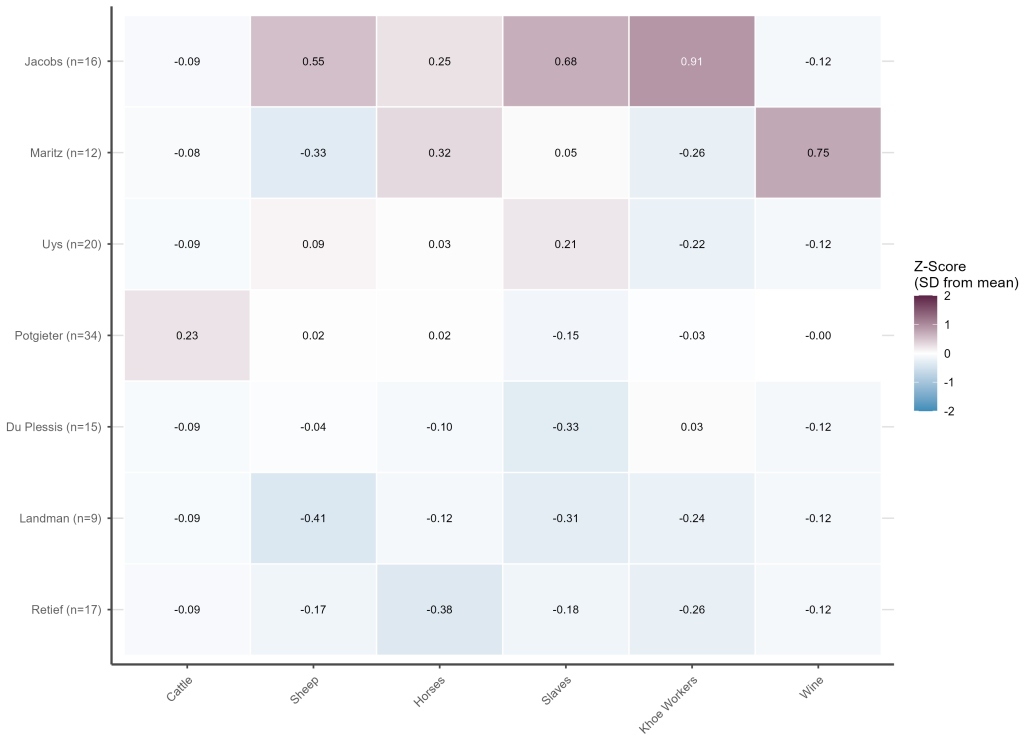

Figure 13: Asset Composition by Trek Leader (Z-Scores)

Notes: Cell values show the mean for each leader's followers, expressed as standard deviations from the overall Voortrekker mean. Darker shading indicates larger deviations. Only leaders with at least eight matched followers are shown.
Alt text: Heatmap of asset z-scores by trek leader; Potgieter's followers stand out as cattle-rich and Landman's as asset-poor, with most other cells near zero.

became the Orange Free State were more suited to extensive pastoralism. If pre-Trek wealth shaped destination choice, we should see systematic sorting.

Standardizing the raw destination entries as for leader names yields five categories: Natal ($n = 83$), the Orange Free State ($n = 78$), Western Transvaal ($n = 46$), Eastern Transvaal ($n = 28$) and other or unrecorded destinations ($n = 30$).

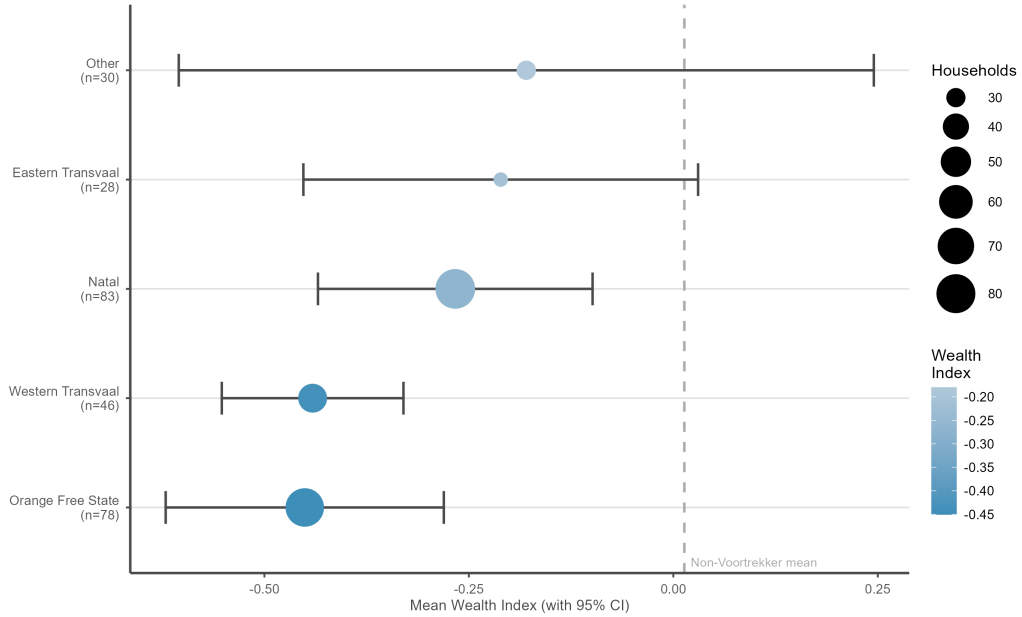

Figure 14: Mean Wealth Index of Voortrekkers by Destination

Notes: Mean wealth index with 95% confidence intervals by destination region. Point size is proportional to the number of matched households. Dashed line indicates the non-Voortrekker mean. Labels indicate sample sizes. *Alt text*: Dot plot of mean wealth index by destination region (Natal, Orange Free State, Western and Eastern Transvaal, other); all five means lie below the non-Voortrekker reference line with overlapping confidence intervals.

All five destination means fall below the non-Voortrekker average, with substantial overlap in confidence intervals (Figure 14). ANOVA finds no significant differences across destinations, and multinomial logit models find no evidence that pre-Trek economic characteristics predict destination choice. These tests are power-limited (subsamples $n = 28$ to 83), so the absence of detected sorting does not rule out modest effects. The parsimonious reading is that households moved because they needed land, not to match assets to the most remunerative frontier, with the Trek's decentralized, family-based character limiting the network-driven route sorting seen elsewhere (Conor, 2019).

8 Robustness

We subject our findings to several robustness checks: alternative matching algorithms, consistency across all six comparison methods and the stability of results at different match quality thresholds.

8.1 Alternative Matching Algorithms: Jaro-Winkler and XGBoost

Our main results use Random Forest classification for record linkage. To verify that the findings are not driven by the choice of algorithm, we re-estimate the key regressions using two alternative matching methods: a Jaro-Winkler string distance threshold (a simpler, non-machine-learning approach) and an XGBoost classifier (a more flexible gradient-boosted tree model).

Table 5: Robustness: Alternative Matching Algorithms

	Jaro-Winkler		XGBoost	
Variable	Coef	p-value	Coef	p-value
Cattle	−1860.4	0.345	−1947.4	0.360
Sheep	−73.7	0.499	−25.0	0.851
Horses	−0.23	0.479	0.02	0.962
Slaves	−0.02	0.908	0.19	0.214
Wheat reaped	−0.55	0.534	0.31	0.699
Wine	0.05	0.893	0.37	0.111
Wealth index	−0.01	0.717	0.04	0.323

Notes: Coefficients from district fixed-effects regressions using matches identified by Jaro-Winkler string distance threshold and XGBoost classification, respectively. All coefficients are statistically insignificant. Notably, the signs of several coefficients are inconsistent across methods (e.g., slaves: negative in JW, positive in XGBoost), further confirming the null.

Table 5 reports the results. Every coefficient is statistically insignificant across both methods. The signs of several coefficients are also inconsistent across methods: slaves are slightly negative under JW and slightly positive under XGBoost; wheat reaped is negative under JW and positive under XGBoost. This pattern of small, insignificant and directionally inconsistent coefficients is exactly what we would expect under the null hypothesis of no difference in wealth. The finding is robust to the choice of record linkage algorithm.

8.2 Consistency across All Six Comparison Methods

Throughout the analysis we deploy six comparison methods: (1) row neighbor (nearest census record), (2) same surname, (3) district fixed effects (main specification), (4) exact match on district with subclass weights, (5) family-size match, and (6) district + children FE. Table 15 in the appendix reports full results. The key findings are consistent: settler men is positive and significant in all six methods; household size is positive in all six and significant in the district FE and exact-match designs (the family-size designs absorb it by construction); wealth index and slaves are negative in all six and significantly negative in the Dist+Children FE specification; Khoekhoe workers is never significant. An important caveat is that all six methods share the same Random-Forest-generated treatment indicator, so the concordance demonstrates robustness to specification

rather than to matching error; matching-error concerns are addressed directly by the non-wife-district test in Section 8.5.

8.3 Match Quality

A potential concern is that our results could be driven by low-quality matches. We address this in two ways. First, our four-tier threshold system already imposes stricter requirements on matches with less identifying information. Second, we verify that the match score distribution (Figure 19) shows a clear bimodal pattern, with most accepted matches clustering at high probability scores.

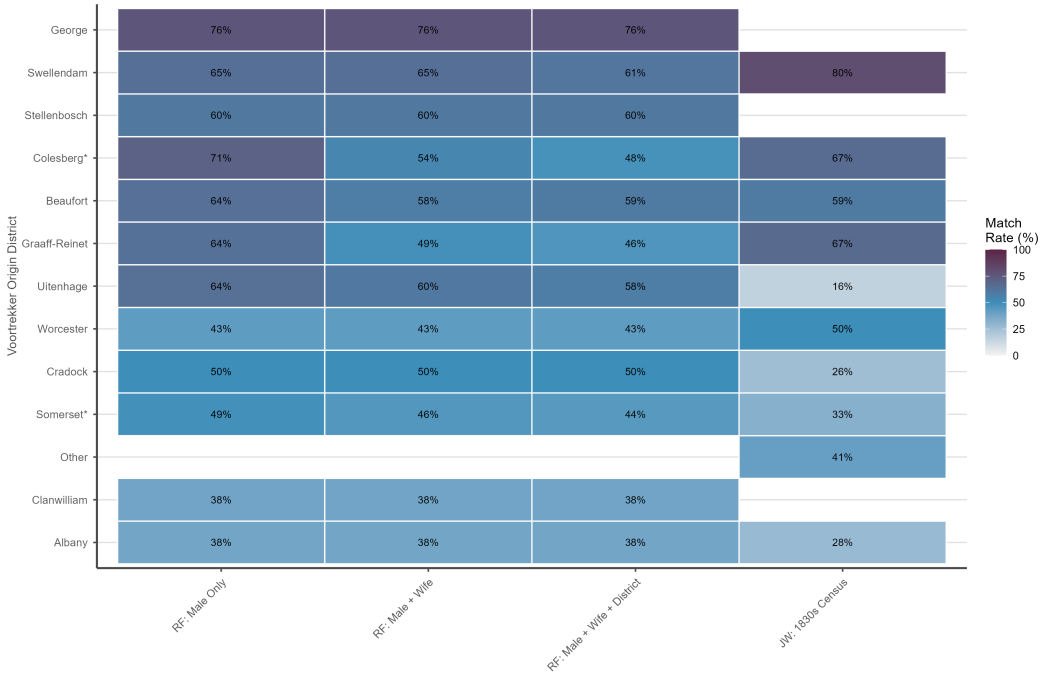

Figure 15: Match Rates by District and Matching Criteria

Notes: Match rates by district under four matching approaches: Random Forest with male name only, male + wife name, and male + wife + district constraints, plus Jaro-Winkler linkage to the 1830s census as an independent validation. Districts marked with an asterisk are multi-district search categories used in the historical crosswalk (notably Somerset and Colesberg, which do not map one-for-one onto the 1825 census districts). Adding wife information generally improves match quality, but its effect on match rates is not mechanically negative: wife corroboration can eliminate false positives while also allowing some additional true matches to clear the lower threshold applied when wife information is available.
Alt text: Heatmap of match rates by district under four matching criteria, with higher rates in districts where wife names were recorded and for less restrictive criteria.

8.4 Researcher Discretion: The Expert-Rescued Matches

The linkage procedure combines algorithmic classification with expert review, and the review stage accounts for a large share of the matches: 216 of the 558 accepted links (38.7 percent) were rescued from below the Random Forest thresholds by author judgment about plausible name variants

(Appendix A). Because this places a substantial share of the treatment group on researcher discretion, we test directly whether the two tiers yield the same results. Table 6 re-estimates the district fixed-effects specification three times: on the full match set, on the RF-accepted tier only (342 census households) and on the expert-rescued tier only (194 census households, from the 216 rescued matches), with the non-Voortrekker pool common to all three. The household-composition coefficients are directionally consistent across both tiers and the wealth null holds in each; the rescued tier, with fewer treated observations, is simply less precisely estimated. The findings do not depend on the rescued matches.

Table 6: Match-Tier Decomposition: District FE Estimates by Linkage Tier

Variable	Full sample	RF-accepted only	Expert-rescued only
Wealth index	-0.047	-0.022	-0.098
	(0.263)	(0.685)	(0.122)
Household size	0.357*	0.599**	-0.051
	(0.020)	(0.003)	(0.807)
Settler children	0.311*	0.531**	-0.059
	(0.029)	(0.006)	(0.753)
Settler men	0.052***	0.051***	0.054***
	(<0.001)	(<0.001)	(<0.001)
Slaves	-0.174	-0.201	-0.131
	(0.156)	(0.190)	(0.481)
Horses	-0.506	-0.344	-0.862
	(0.169)	(0.404)	(0.183)
Cattle	-2120.8	-2098.7	-2198.1
	(0.356)	(0.386)	(0.307)
Wheat sown (muids)	-0.042	-0.066	-0.000
	(0.444)	(0.300)	(0.996)
Matched VT households	536	342	194

Notes: Each cell reports the Voortrekker coefficient from a separate OLS regression with district fixed effects and heteroskedasticity-robust (HC1) standard errors; p-values in parentheses. In the tier columns, the treated group is restricted to matches of that tier and the comparison group is the full non-Voortrekker pool. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

As a further check on the review stage itself, a second rater independently re-rated a blinded random sample of 150 candidate pairs from the RF-rejected pool, 100 of them pairs the first rater had rescued and 50 pairs the first rater had declined, shuffled and stripped of all scores and original decisions. The second rater proved more conservative, independently confirming 49 percent of the rescued pairs while declining 96 percent of the decoys: the two raters order the candidates alike but draw the acceptance threshold differently on these genuinely borderline cases (Appendix H.3 reports the design and full agreement statistics). The tier decomposition above is the substantive safeguard: the findings survive with the rescued tier excluded entirely.

8.5 Is the Household Size Finding Genuine?

The household composition result, that Voortrekkers had significantly larger families with more children, is our strongest positive finding. Two methodological concerns warrant examination.

Concern 1: Age and cohort effects. The Voortrekker sample is drawn from individuals alive and active enough to trek in 1835–40, which may disproportionately select younger men in the prime of their child-rearing years in 1825. When we compare Voortrekker and non-Voortrekker households within each age cohort (25–34, 35–44, 45–54, 55+), the Voortrekker advantage in household size is positive across all cohorts; age composition therefore does not drive the finding. Restricting further to households at the same life-cycle stage (with a wife and at least one child), the Voortrekker coefficient on household size and children remains positive and significant.

Concern 2: Matching bias from wife-name corroboration. The Random Forest classifier assigns higher match scores when wife names corroborate the match. Households where the wife's name is recorded in both the census and genealogical record are therefore easier to match, and these are mechanically more likely to have children. The larger-household finding could, in principle, be an artifact of the matching process.

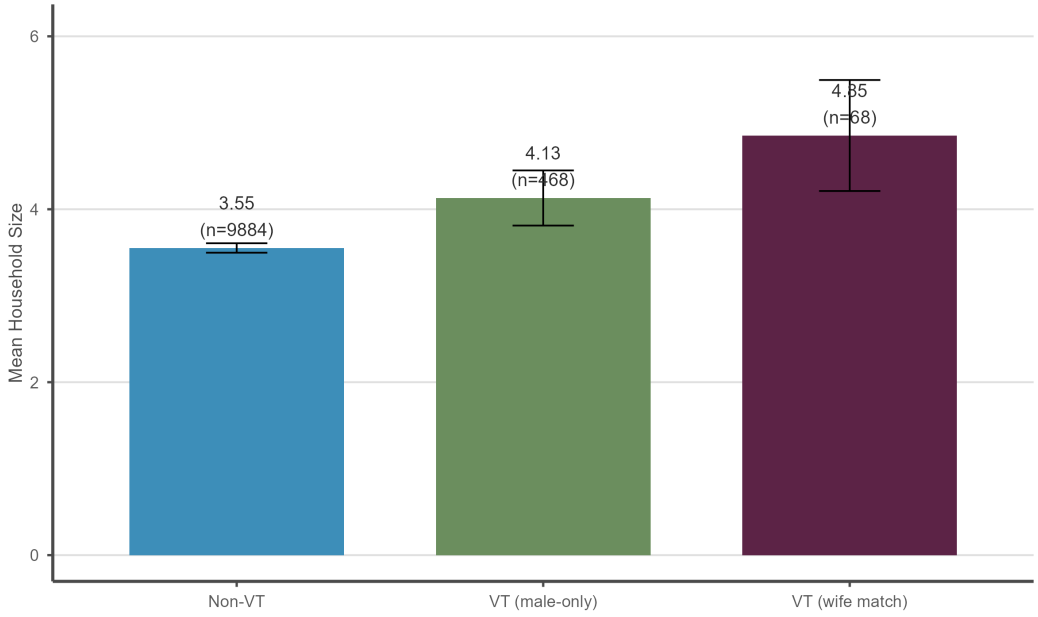

Figure 16: Household Size by Match Type

Notes: Mean household size for non-Voortrekkers, Voortrekkers matched on male name only, and Voortrekkers matched with wife-name corroboration. Error bars show 95% confidence intervals.
Alt text: Bar chart of mean household size in three groups; both male-only-matched and wife-corroborated Voortrekker households exceed the non-Voortrekker mean.

We address this concern through several tests. First, we compare household size between Voortrekkers matched with wife corroboration and those matched on male names only (Figure 16). If matching bias were the sole driver, we would expect wife-matched Voortrekkers to have substantially larger households than male-only matched Voortrekkers, and the latter to resemble non-Voortrekkers.

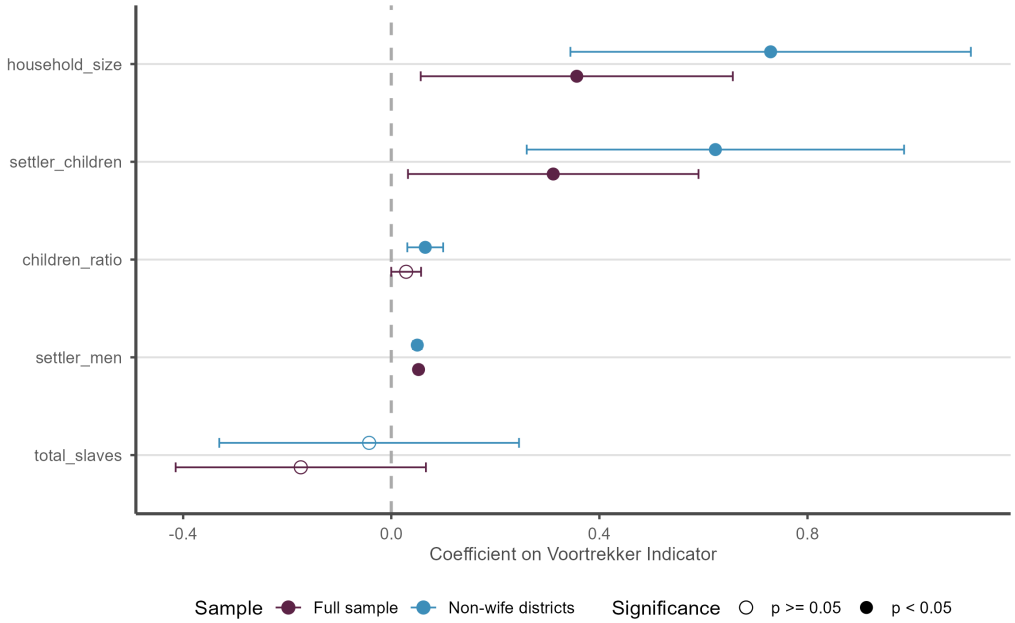

Figure 17: Voortrekker Coefficients: Full Sample vs. Non-Wife Districts

Notes: Comparison of district fixed-effects regression coefficients for the full sample (all districts) and the subsample of districts where wife names were not recorded in the census. In non-wife districts, all matches are necessarily male-only, eliminating wife-corroboration bias.
Alt text: Paired coefficient plot comparing full-sample and non-wife-district estimates; the household-composition coefficients are larger in the non-wife districts while the wealth coefficients remain near zero in both.

Second, and most decisively, we exploit the fact that wife names were recorded in only two of the eleven census districts (Graaff-Reinet and Swellendam). In the remaining nine districts, *all* matches are necessarily male-only; wife corroboration is impossible. If the household size finding survives in these non-wife districts, it cannot be an artifact of wife-name matching bias. Table 7 reports district fixed-effects coefficients from the full sample alongside those estimated on the 367 Voortrekker matches in non-wife districts. The household composition coefficients are not merely positive in non-wife districts; they are *larger*: the household size coefficient rises from 0.357 in the full sample to 0.729 in non-wife districts, and the settler-children and children-ratio coefficients rise in step. The wealth variables remain insignificant in both samples. Figure 17 presents these results graphically.

Taken together, these tests provide strong evidence that the household size finding reflects a genuine characteristic of Voortrekker families rather than an artifact of age composition or matching

Table 7: District FE Regressions: Full Sample vs. Non-Wife Districts

	Full sample ($N = 10{,}420$; $N_{\text{VT}} = 536$)		Non-wife districts ($N = 7{,}228$; $N_{\text{VT}} = 367$)	
Variable	Coef	p	Coef	p
Household size	0.357	0.020	0.729	<0.001
Settler children	0.311	0.029	0.623	<0.001
Children ratio	0.029	0.051	0.065	<0.001
Settler men	0.052	<0.001	0.050	<0.001
Wealth index	−0.047	0.263	−0.007	0.890
Slaves	−0.174	0.156	−0.043	0.772

Notes: Each cell reports the Voortrekker coefficient from a separate OLS regression with district fixed effects and heteroskedasticity-robust standard errors. Non-wife districts are those where wife names were not recorded in the census (all districts except Graaff-Reinet and Swellendam). In these districts, all matches are necessarily male-only, ruling out wife-corroboration bias. The household composition coefficients are larger in non-wife districts, confirming that the finding is not an artifact of the matching mechanism.

bias. If anything, the matching mechanism *attenuates* the household composition differences.

False-positive contamination. A cross-validation precision of 0.664 implies that roughly one in three classifier-accepted matches may be incorrect, which attenuates all coefficients toward zero. We address this with a same-district contamination simulation (Appendix J). At the 34 percent Random-Forest benchmark, household size and settler children remain positive and statistically significant in a non-trivial share of repetitions, while settler men remains significant in 99 percent. The wealth-index coefficient remains close to zero throughout, leaving little room for a true wealth effect that linkage noise could have masked. The tier decomposition in Table 6 addresses the related concern that the expert-rescued matches drive the results.

9 Conclusion

The Voortrekkers are remembered in South African historiography as a migration shaped by grievance: against emancipation, against British administration, against racial equalization. The household records do not support this view. What distinguished the families who left from the families who stayed was not a shared exposure to loss but a shared capacity for movement. Trekker households were larger, with more children, and more heavily invested in the mobile assets of a pastoral economy than in the immobile assets of slaveholding and arable farming. The composite wealth index, by contrast, is statistically indistinguishable between movers and stayers, with equivalence tests confirming a coefficient centered near zero (Appendix E). The compensation channel, which a century of historiography has treated as central, finds no support in the micro-data, and migration timing is unrelated to pre-migration wealth or slave holdings. Large pastoral households, constrained by the shortage of land within the Colony, exercised the exit option because exit was cheapest for them.

The implications reach beyond the Cape. At a pastoral frontier, selection operates on household composition and asset portfolio rather than on the individual-ability dimensions familiar from later transatlantic movements; what mattered was not a worker's education or occupation but a family's demographic structure and its holdings of mobile capital. That distinction supplies the missing first step in any account of what the Voortrekker republics became: polities founded not by a Cape colonial elite transferring slaves and coercive infrastructure but by larger, poorer pastoral families with limited investment in slavery and a demographic profile oriented toward land-extensive herding. Their origin helps explain why the republican labor systems diverged from the Cape's, why land appropriation became the central mechanism of frontier accumulation, and why the cultural and religious distinctiveness of the trekkers proved so durable in the institutional development of the interior. The household selection we document is the precondition for the legacies traced by Bazzi et al. (2023) and others in analogous settler migrations. The Cape is also an early case: the settler migrations that dominate this literature are later, mostly late-nineteenth- or early-twentieth-century transatlantic movements, and the word *trek* entered English through the 1830s Cape. On emancipation itself, the Cape offers an unusually direct test of a claim that has been impossible to make elsewhere: compensation, however imperfect, retained the most financially exposed slaveholders within the colony. Hirschman's framework predicts precisely this disjunction between grievance and exit, and the compensated British emancipation of 1833–34 illustrates it at individual level.

Five caveats bear on the interpretation. The 1825 census predates the Trek by a decade, and within-district vulnerability to the shocks of 1825–35, Ordinance 50, the Sixth Frontier War, emancipation itself, could have deepened selection on wealth in ways we cannot observe. Thirty-nine percent of Voortrekkers remain unmatched; the non-wife-district test in Section 8.5 addresses the chief concern, but residual selection bias is possible. The census records material wealth, not ideology, networks or risk tolerance. Survival selection between 1825 and 1835 could contribute to the household-size premium. And cross-district male-only matches are rejected by design, excluding the most geographically mobile Voortrekkers. Appendix K discusses each concern in detail.

None of this implies that political grievances over emancipation, labor policy and British administration were irrelevant to making emigration thinkable. The Trek was an organized political act, with manifestos, elected leaders, scouts and negotiations with African polities, dimensions our data cannot speak to. What the data do speak to is the identity of the households that exercised the exit option: families distinguished by a demographic structure that made staying untenable and leaving feasible, rather than by their exposure to the institutions being reformed. The Great Trek, like many migrations before and since, was shaped less by the losses of wealth than by the aspirations of land-constrained families.

Data Availability

On acceptance, the data and code supporting this article will be deposited in the OPEN ICPSR repository under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, with a DOI-based citation supplied at that time. The deposit will contain the 1825 Cape Colony census (*opgaafrolle*) at the household level, the Voortrekker genealogical records, the slave compensation records from Ekama et al. (2021), the record-linkage outputs and the household-level analysis dataset, together with a codebook and an R loader script. A working version is currently mirrored at `https://github.com/johanfourieza/research/tree/main/2026_voortrekker`. No sensitive or proprietary data are involved and no exceptions to full data sharing are requested.

References

Abramitzky, Ran, and Leah Platt Boustan. 2017. "Immigration in American Economic History." *Journal of Economic Literature*, 55(4): 1311–1345.

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson. 2012. "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration." *American Economic Review*, 102(5): 1832–1856.

Abramitzky, Ran, Leah Platt Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez. 2021. "Automated Linking of Historical Data." *Journal of Economic Literature*, 59(3): 865–918.

Bailey, Martha, Connor Cole, Morgan Henderson, and Catherine Massey. 2020*a*. "How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data." *Journal of Economic History*, 80(4): 997–1044.

Bailey, Martha J., Connor Cole, Morgan Henderson, and Catherine Massey. 2020*b*. "How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data." *Journal of Economic Literature*, 58(4): 997–1044.

Bazzi, Samuel, Andreas Ferrara, Martin Fiszbein, Thomas Pearson, and Patrick A. Testa. 2023. "The other great migration: Southern whites and the new right." *Quarterly Journal of Economics*, 138(3): 1577–1647.

Becker, Sascha O., Irena Grosfeld, Pauline Grosjean, Nico Voigtländer, and Ekaterina Zhuravskaya. 2020. "Forced migration and human capital: Evidence from post-WWII population transfers." *American Economic Review*, 110(5): 1430–1463.

Beltrán Tapia, Francisco, and Santiago de Miguel Salanova. 2017. "Migrants' Self-Selection in the Early Stages of Modern Economic Growth." *Economic History Review*, 70(1): 101–121.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the false discovery rate: A practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1): 289–300.

Binckes, Robin. 2013. *The Great Trek Uncut.* Johannesburg:30° South Publishers.

Blum, Matthias, Karl-Peter Krauss, and Dmytro Myeshkov. 2022. "Human Capital Transfer of German-Speaking Migrants in Eastern Europe, 1780s–1820s." *Economic History Review*, 75(3): 703–738.

Borjas, George J. 1987. "Self-Selection and the Earnings of Immigrants." *American Economic Review*, 77(4): 531–553.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics*, 90(3): 414–427.

Cilliers, Jeanne, and Erik Green. 2018. "The land labour hypothesis in a settler economy: Wealth, labour, and household composition on the South African frontier." *International Review of Social History*, 63: 239–271.

Cilliers, Jeanne, Erik Green, and Robert Ross. 2023. "Did it pay to be a pioneer? Wealth accumulation in a newly settled frontier society." *The Economic History Review*, 76: 257–282.

Cobbing, Julian. 1988. "The Mfecane as Alibi: Thoughts on Dithakong and Mbolompo." *Journal of African History*, 29(3): 487–519.

Collins, William J., and Marianne H. Wanamaker. 2014. "Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data." *American Economic Journal: Applied Economics*, 6(1): 220–252.

Collins, William J., and Marianne H. Wanamaker. 2015. "The Great Migration in Black and White: New Evidence on the Selection and Sorting of Southern Migrants." *Journal of Economic History*, 75(4): 947–992.

Conor, Dylan. 2019. "Cream of the Crop? Geography, Networks, and Irish Migrant Selection in the Age of Mass Migration." *Journal of Economic History*, 79(1): 139–175.

Derenoncourt, Ellora. 2022. "Can you move to opportunity? Evidence from the Great Migration." *American Economic Review*, 112(2): 369–408.

Draper, Nicholas. 2010. *The Price of Emancipation: Slave-Ownership, Compensation and British Society at the End of Slavery.* Cambridge:Cambridge University Press.

Dribe, Martin, Björn Eriksson, and Jonas Helgertz. 2022. "From Sweden to America: Migrant Selection in the Transatlantic Migration, 1890–1910." *European Review of Economic History*, 27(1): 24–44.

Du Toit, André, and Hermann Giliomee. 1983. *Afrikaner Political Thought: Analysis and Documents, Volume 1, 1780–1850.* Cape Town:David Philip.

Ekama, Kate, Johan Fourie, Hans Heese, and Lisa-Cheree Martin. 2021. "When Cape Slavery Ended: Introducing a New Slave Emancipation Dataset." *Explorations in Economic History*, 81: 101390.

Elphick, Richard, and Hermann Giliomee. 1989. "The Origins and Entrenchment of European Dominance at the Cape, 1652–c. 1840." In *The Shaping of South African Society, 1652–1840.* . 2nd ed., , ed. Richard Elphick and Hermann Giliomee. Cape Town:Maskew Miller Longman.

Etherington, Norman. 2001. *The Great Treks: The Transformation of Southern Africa, 1815–1854.* Harlow:Longman.

Feigenbaum, James J. 2016. "Automated Census Record Linking: A Machine Learning Approach." Working paper.

Fourie, Johan. 2022. *Our long walk to economic freedom: lessons from 100,000 years of human history.* Cambridge University Press.

Gay, Victor, Paula E. Gobbi, and Marc Goñi. 2026. "Revolutionary transition: Inheritance change and fertility decline." *Journal of Political Economy*, 134(6): 1666–1713.

Giliomee, Hermann. 2003. *The Afrikaners: Biography of a People.* Charlottesville:University of Virginia Press.

Hall, Catherine, Nicholas Draper, Keith McClelland, Katie Donington, and Rachel Lang. 2014. *Legacies of British Slave-Ownership: Colonial Slavery and the Formation of Victorian Britain.* Cambridge:Cambridge University Press.

Hamilton, Carolyn. 1998. *Terrific Majesty: The Powers of Shaka Zulu and the Limits of Historical Invention.* Cambridge, MA:Harvard University Press.

Hanson, Gordon, Pia Orrenius, and Madeline Zavodny. 2023. "US immigration from Latin America in historical perspective." *Journal of Economic Perspectives*, 37(1): 199–222.

Hatton, Timothy J., and Zachary Ward. 2024. "International migration in the Atlantic economy 1850–1940." In *Handbook of Cliometrics.* 507–535. Springer.

Hirschman, Albert O. 1970. *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States.* Cambridge, MA:Harvard University Press.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics*, 6(2): 65–70.

Keegan, Timothy. 1996. *Colonial South Africa and the Origins of the Racial Order.* Cape Town:David Philip.

Lakens, Daniël. 2017. "Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses." *Social Psychological and Personality Science*, 8(4): 355–362.

Legassick, Martin. 2010. *The Politics of a South African Frontier: The Griqua, the Sotho-Tswana, and the Missionaries, 1780–1840.* Basel:Basler Afrika Bibliographien.

Leopold, Stefan, Jens Ruhose, and Simon Wiederhold. 2025. "Why is the Roy–Borjas model unable to predict international migrant selection on education? Evidence from urban and rural Mexico." *The World Economy*, 48(2): 300–322.

MacCrone, Ian Douglas. 1937. *Race Attitudes in South Africa: Historical, Experimental and Psychological Studies.* London:Oxford University Press.

Macmillan, William Miller. 1927. *The Cape Colour Question: A Historical Survey.* London:Faber and Gwyer.

Macmillan, William Miller. 1929. *Bantu, Boer and Briton: The Making of the South African Native Problem.* London:Faber and Faber.

Marais, Johannes Stephanus. 1939. *The Cape Coloured People, 1652–1937.* London:Longmans, Green.

Muller, C. F. J. 1963. *Die Britse Owerheid en die Groot Trek.* Pretoria:Universiteit van Suid-Afrika.

Muller, C. F. J. 1974. *Die Oorsprong van die Groot Trek.* Cape Town:Tafelberg.

Natkhov, Timur, and Natalia Vasilenok. 2021. "Skilled Immigrants and Technology Adoption: Evidence from the German Settlements in the Russian Empire." *Explorations in Economic History*, 81: 101399.

Nel, Heinrich. 2020. "Wealth mobility, familial ties and migration: Evidence from the Cape of Good Hope Panel." PhD diss. Stellenbosch University.

Newton-King, Susan. 1999. *Masters and Servants on the Cape Eastern Frontier, 1760–1803.* Cambridge:Cambridge University Press.

Pérez, Santiago. 2017. "The (South) American dream: Mobility and economic outcomes of first- and second-generation immigrants in nineteenth-century Argentina." *The Journal of Economic History*, 77(4): 971–1006.

Rijpma, Auke, Jeanne Cilliers, and Johan Fourie. 2020. "Record Linkage in the Cape of Good Hope Panel." *Historical Methods*, 53(2): 112–129.

Ross, Robert. 1999. *A Concise History of South Africa.* Cambridge:Cambridge University Press.

Van der Merwe, P. J. 1937. *Die Noordwaartse Beweging van die Boere voor die Groot Trek, 1770–1842.* Pretoria:Die Staatsdrukker.

Van Jaarsveld, Floris Albertus. 1951. *Die Eenheidstrewe van die Republikeinse Afrikaners.* Amsterdam:J. P. van der Walt.

van Leeuwen, Marco H. D., and Ineke Maas. 2022. "Social Mobility through Migration to the Colonies: The Case of Algeria." *Journal of Interdisciplinary History*, 53(2): 225–265.

Venter, Chris. 1985. *The Great Trek.* Cape Town:Don Nelson.

Walker, Eric A. 1934. *The Great Trek.* London:Adam and Charles Black.

Ward, Zachary. 2017. "Birds of Passage: Return Migration, Self-Selection and Immigration Quotas." *Explorations in Economic History*, 64: 37–52.

Worden, Nigel. 1985. *Slavery in Dutch South Africa.* Cambridge:Cambridge University Press.

Wright, John. 2010. "Turbulent Times: Political Transformations in the North and East, 1760s–1830s." In *The Cambridge History of South Africa, Volume 1: From Early Times to 1885.* , ed. Carolyn Hamilton, Bernard K. Mbenga and Robert Ross, 211–252. Cambridge:Cambridge University Press.

Zimran, Ariell. 2024. "Internal migration in the United States: Rates, selection, and destination choice, 1850–1940." *The Journal of Economic History*, 84(3): 727–766.

A Record Linkage Methodology

A.1 The Challenge of Historical Record Linkage

Linking individuals across historical datasets presents well-known challenges (Feigenbaum, 2016; Abramitzky et al., 2021; Bailey et al., 2020*a*). There are no unique identifiers: the 1825 census records names but not birth dates or other distinguishing information. Spelling variation is substantial, as names were transcribed by local officials with inconsistent orthography (e.g., Ackerman/Akkerman, Johan/Jan/Johannes, Gert/Gerhardus). Wife names pose additional complications: the census may record a wife's married name while the Voortrekker records provide her maiden surname. Multiple individuals may share the same name, particularly with common Afrikaner surnames such as Botha, Du Plessis or Van der Merwe.

These challenges are compounded by the historical context. The Cape Colony's record-keeping, while detailed for an early nineteenth-century colonial setting, was not designed for longitudinal tracking of individuals. The Voortrekker genealogical records, compiled retrospectively from diverse sources, contain their own inconsistencies and gaps.

A.2 Random Forest Classification Approach

We follow the approach of Rijpma, Cilliers and Fourie (2020) in using a Random Forest classifier for probabilistic record linkage. The procedure involves four steps: blocking, feature engineering, training and threshold selection.

Blocking. To reduce the computational burden of comparing every Voortrekker to every census record, we restrict candidate pairs to exact surname matches. In the main pass, each Voortrekker is compared only to census records sharing the same surname in the relevant district search set. For most records this is just the declared origin district. For Somerset and Colesberg, however, we expand the search to the historically relevant neighboring or parent districts used in the codebook crosswalk. After this district-restricted pass, we run a second-pass search for still-unmatched Voortrekkers against all districts, but only wife-corroborated cross-district candidates are eligible for acceptance.

Feature engineering. For each candidate pair, we compute a rich set of features capturing the similarity between the Voortrekker record and the census record:

- *Name distances*: Jaro-Winkler string similarity between the Voortrekker's full name and the census name, and between the first names alone.
- *Wife name distances*: Jaro-Winkler similarity between wife first names, and between the Voortrekker wife's maiden surname and the census wife's recorded name.

- *Wife presence indicators*: Binary flags for whether both records contain wife information, and whether either record has wife information. As Rijpma, Cilliers and Fourie (2020) note, "the absence of the wife makes it far harder to identify a link."
- *Surname frequency*: The relative frequency of the surname in the district, capturing the informativeness of a match. A match on a rare surname like "Tregardt" is far more diagnostic than a match on "Botha."
- *Initials match*: Whether the first initial matches.
- *Name length ratio*: The ratio of name lengths, capturing whether names are truncated.
- *District match*: Whether the Voortrekker's declared district matches the census district.

Training data. We construct training data through expert labeling. Four independent labellers each evaluate a separate set of 250 candidate pairs (1,000 pairs total): the two co-authors and two research assistants familiar with Afrikaans history. Each pair is classified as a match (1), non-match (0), or uncertain (?). Uncertain cases are excluded from training. The labeling task is informed by all available features (name similarity, wife information, district and household composition), and labellers are encouraged to flag borderline cases for group discussion. This staged labeling exercise substantially improved the model relative to the initial Jaro-Winkler pass by teaching it which combinations of name variation, wife corroboration and district information correspond to genuinely plausible historical matches. Boundary cases near the decision threshold are inspected for quality assurance.

Random Forest and threshold selection. We train a Random Forest classifier on the labeled data using stratified 5-fold cross-validation. Because false positives (incorrect matches) are more costly than false negatives (missed matches) in our setting, since an incorrect match could systematically bias our comparison of Voortrekkers and non-Voortrekkers, we optimize the $F_{0.5}$ score, which weights precision twice as heavily as recall. The optimal probability threshold is determined by maximizing $F_{0.5}$ across the cross-validation folds.

Four-tier threshold system. We apply different probability thresholds depending on the quality of available information:

1. *Same district, wife corroborated*: The lowest threshold, as the combination of district, name, and wife provides strong identification.
2. *Same district, male name only*: A higher threshold (+0.15), compensating for the loss of wife information.
3. *Cross-district, wife corroborated*: A very strict threshold (+0.20), since cross-district matches are inherently less certain.

4. *Cross-district, male name only*: Rejected entirely. Without wife corroboration or district confirmation, male-only cross-district matches are too unreliable.

A one-to-one matching constraint ensures that each census record is matched to at most one Voortrekker. When multiple Voortrekkers claim the same census record, we retain the match with the highest probability score, preferring matches with wife corroboration among ties.

In a second stage, one of the authors reviewed all 917 best-candidate pairs (both those accepted and those rejected by the Random Forest) to verify algorithmic decisions and rescue near-miss matches. This means that every final accepted match was seen by an author, even when it had already been accepted by the classifier. The review accepted 216 additional matches that fell below the RF probability thresholds but were judged to be correct on the basis of name similarity, wife corroboration and historical plausibility. These matches are classified as "Manual" quality. The final accepted linkage set therefore reflects both algorithmic screening and author validation, which increases our confidence that the retained matches are genuinely correct. The two-stage approach of automated classification followed by expert review follows best practice in historical record linkage (Bailey et al., 2020*a*).

For full transparency about the role of author discretion, we report the outcome of the review stage in detail. Of the 917 best-candidate pairs, 342 had been accepted by the Random Forest at the $F_{0.5}$-optimal threshold and 575 had been rejected. The author upheld all 342 RF-accepted matches and rescued 216 of the 575 RF-rejected pairs, a rescue rate of 37.6 percent. The remaining 359 rejected pairs were confirmed as non-matches and become part of the unmatched pool. Three criteria governed rescue decisions: (i) exact or near-exact matches on both husband and wife names with consistent declared district, which the Random Forest could under-weight when the surname was rare or the wife name partially standardized; (ii) historically plausible name variants not captured by the Jaro-Winkler distance metric (e.g., *Gerhardus/Gert*, *Johannes/Jan*, *Ackerman/Akkerman*); and (iii) one-to-one household composition consistent with the Voortrekker genealogical record. Pairs that did not meet these criteria were retained as non-matches even where the Random Forest score was just below threshold. The 216 rescued matches carry the "Manual" quality flag throughout, and the attenuation exercise in Appendix J reports results separately for RF-accepted ($n = 342$) and Manual ($n = 216$) tiers: the household-composition findings are directionally consistent across both.

A.3 Match Rates and Quality

Table 8 reports match rates by district. The overall match rate is 60.8 percent (37.3 percent from the Random Forest classifier plus 23.6 percentage points from expert manual review), with substantial variation across districts. Match rates tend to be higher in districts with greater surname diversity and better wife-name coverage in both sources.

Figure 18 shows the variable importance from the Random Forest model. Wife-related features, particularly the Jaro-Winkler distance on wife first names and the indicator for both records having wife information, rank among the most important, confirming that wife name corroboration is the

Table 8: Voortrekker Records and Match Rates by District

District	VT Records	Matched	Match Rate (%)
Albany	29	11	37.9
Beaufort	114	80	70.2
George	17	12	70.6
Graaff-Reinet	171	112	65.5
Somerset	353	186	52.7
Stellenbosch	5	3	60.0
Swellendam	62	46	74.2
Uitenhage	146	97	66.4
Worcester	20	11	55.0
Total	**917**	**558**	**60.8**

Notes: VT Records refers to the number of adult male Voortrekkers (born before 1810) with valid names and a declared origin in the district. Matched counts reflect the final accepted match set after expert manual review (342 RandomForest matches plus 216 additional matches, for 558 in total). Match rates vary with the availability of wife name information and surname frequency in the district. Somerset includes Cradock (20 records), its pre-1825 name. Graaff-Reinet includes Colesberg (52 records), established in the 1830s. Worcester includes Clanwilliam (13 records).

single most valuable piece of identifying information. Name distance features on the husband's name and surname frequency also contribute substantially.

B Main Results by Design and Descriptive Statistics

This appendix reports each of the four research designs summarized in Table 1 in full and in a common format, together with descriptive statistics for the harmonized variable set. Tables 10, 12 and 13 report group means alongside differences; Table 11 reports the district fixed-effects coefficients with robust standard errors. All five tables are generated directly by the replication pipeline.

C Male-Headed Control Group

The Voortrekker matching sample requires an adult male born before 1810 on the genealogical side, but the control group in the main specification includes female-headed census households. If an adult male was a de facto precondition for trekking, those households are arguably less meaningful counterfactuals (a point we owe to a referee). Table 14 re-estimates the district fixed-effects specification restricting controls to households with at least one adult settler man, which drops 700 of the 9,884 control households. The wealth null is unchanged. The settler-men coefficient collapses to

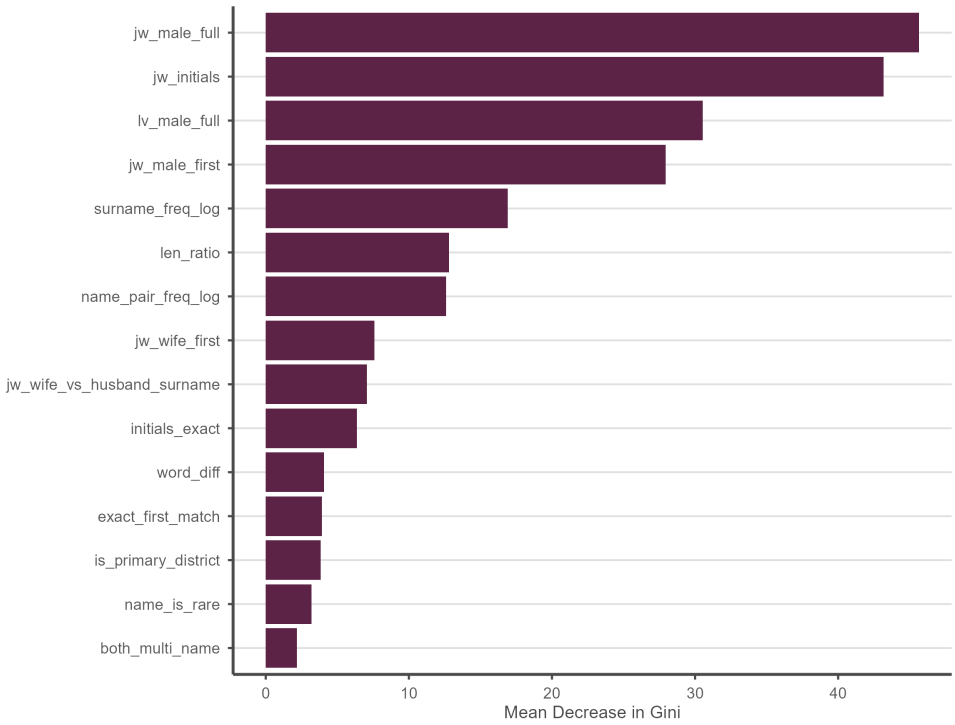

Figure 18: Random Forest Variable Importance for Record Linkage

Notes: Mean decrease in Gini impurity for each feature in the Random Forest classifier. Higher values indicate greater contribution to classification accuracy. Wife-related features dominate, consistent with Rijpma, Cilliers and Fourie (2020).
Alt text: Horizontal bar chart of Random Forest variable importance; wife-name similarity and wife-presence features rank highest, followed by husband name distance and surname frequency.

Table 9: Descriptive Statistics, Full Estimation Sample

Variable	Mean	SD	Min	Max
Livestock				
Horses	6.694	11.98	0.000	185.0
Cattle	1018.2	96165.3	0.000	9814894.0
Sheep	279.3	4083.1	0.000	355270.0
Labor				
Slaves	2.262	5.821	0.000	90.00
Khoekhoe workers	2.429	4.742	0.000	54.00
Agriculture				
Wheat sown (muids)	1.445	5.149	0.000	130.0
Wheat reaped (muids)	15.15	60.87	0.000	1600.0
Wine (leaguers)	4.378	294.0	0.000	30000.0
Household composition				
Settler men	0.933	0.251	0.000	3.000
Settler women	0.617	0.486	0.000	2.000
Settler children	2.037	2.542	0.000	45.00
Household size	3.587	2.815	1.000	47.00
Children ratio	0.336	0.330	0.000	0.957
Composite				
Wealth index	0.000	1.587	-0.843	35.98
Households		10,420		

Notes: All 10,420 households in the 1825 census estimation sample. The wealth index is the first principal component described in Section 3; it explains 28.0 percent of the joint variance of its nine inputs.

Table 10: Design (1): Nearest Census Neighbors

Variable	VT mean	Non-VT mean	Difference	p-value
Livestock				
Horses	5.453	5.405	0.048	0.909
Cattle	295.8	121.3	174.5	0.363
Sheep	438.8	523.5	-84.75	0.677
Labor				
Slaves	0.894	1.092	-0.197	0.195
Khoekhoe workers	3.033	3.077	-0.044	0.882
Agriculture				
Wheat sown (muids)	0.505	0.436	0.069	0.260
Wheat reaped (muids)	6.290	5.284	1.007	0.203
Wine (leaguers)	0.316	0.102	0.213	0.199
Household composition				
Settler men	0.998	0.934	0.065	$<$0.001
Settler women	0.701	0.743	-0.041	0.079
Settler children	2.539	2.387	0.152	0.386
Household size	4.239	4.064	0.175	0.353
Children ratio	0.388	0.385	0.004	0.844
Composite				
Wealth index	-0.245	-0.244	-0.001	0.976

Notes: Paired comparison of each matched Voortrekker to the households recorded immediately adjacent in the census, over the 519 of 536 matched households with at least one valid non-Voortrekker neighbor; p-values from paired two-sided t-tests.

Table 11: Design (2): District Fixed-Effects Regressions

Variable	Coefficient	Robust SE	p-value
Livestock			
Horses	-0.506	0.368	0.169
Cattle	-2120.8	2295.8	0.356
Sheep	-100.6	138.5	0.468
Labor			
Slaves	-0.174	0.123	0.156
Khoekhoe workers	-0.136	0.259	0.599
Agriculture			
Wheat sown (muids)	-0.042	0.055	0.444
Wheat reaped (muids)	-0.761	0.712	0.286
Wine (leaguers)	0.198	0.176	0.261
Household composition			
Settler men	0.052	0.004	$<$0.001
Settler women	-0.007	0.021	0.727
Settler children	0.311	0.142	0.029
Household size	0.357	0.153	0.020
Children ratio	0.029	0.015	0.051
Composite			
Wealth index	-0.047	0.042	0.263

Notes: Each row reports the coefficient on the Voortrekker indicator from a separate OLS regression with district fixed effects; N = 10,420 (536 matched Voortrekker households). Heteroskedasticity-robust (HC1) standard errors.

Table 12: Design (3): Exact Matching on District

Variable	VT mean	Non-VT mean	Difference	p-value
Livestock				
Horses	5.435	5.989	-0.555	0.122
Cattle	287.7	2434.8	-2147.1	0.354
Sheep	434.1	537.8	-103.7	0.460
Labor				
Slaves	0.897	1.074	-0.176	0.147
Khoekhoe workers	3.004	3.162	-0.158	0.527
Agriculture				
Wheat sown (muids)	0.505	0.546	-0.041	0.451
Wheat reaped (muids)	6.276	7.035	-0.760	0.284
Wine (leaguers)	0.306	0.128	0.179	0.305
Household composition				
Settler men	0.998	0.946	0.052	<0.001
Settler women	0.701	0.707	-0.005	0.800
Settler children	2.522	2.202	0.320	0.025
Household size	4.222	3.855	0.367	0.017
Children ratio	0.386	0.357	0.029	0.043
Composite				
Wealth index	-0.249	-0.197	-0.052	0.206

Notes: Exact matching on district with subclass weights (536 Voortrekker and 9,028 weighted non-Voortrekker households). Non-VT means are weighted by exact-match subclass weights; differences and p-values from weighted regressions with HC-robust standard errors.

Table 13: Design (4): Exact Matching on District and Number of Children

Variable	VT mean	Non-VT mean	Difference	p-value
Livestock				
Horses	5.398	6.059	-0.661	0.119
Cattle	288.7	3877.3	-3588.6	0.666
Sheep	428.0	536.3	-108.3	0.695
Labor				
Slaves	0.884	1.110	-0.226	0.104
Khoekhoe workers	2.938	3.248	-0.309	0.201
Agriculture				
Wheat sown (muids)	0.493	0.566	-0.073	0.282
Wheat reaped (muids)	6.120	7.225	-1.106	0.204
Wine (leaguers)	0.303	0.141	0.162	0.095
Household composition				
Settler men	0.998	0.945	0.053	$<$0.001
Settler women	0.700	0.738	-0.038	0.053
Settler children	2.407	2.407	0.000	1.000
Household size	4.105	4.090	0.015	0.905
Children ratio	0.383	0.387	-0.003	0.824
Composite				
Wealth index	-0.257	-0.184	-0.073	0.093

Notes: Exact matching on district and the exact integer number of settler children (533 Voortrekker and 7,789 weighted non-Voortrekker households). Because settler children is the conditioning variable, its treated and control means are identical by construction. Weighted regressions with HC-robust standard errors.

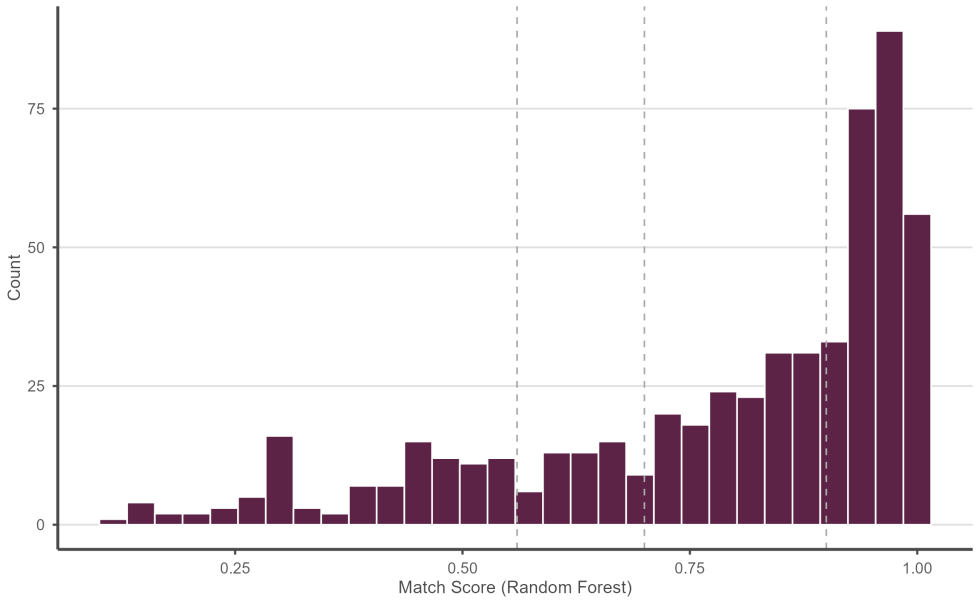

Figure 19: Distribution of Match Scores

Notes: Distribution of Random Forest probability scores for accepted matches. Higher scores indicate greater confidence in the match.
Alt text: Histogram of Random Forest match probability scores for accepted matches, with most mass at high scores.

zero, confirming its mechanical character: it measured the presence of male-headed households in the control pool, not a substantive difference. The household-composition coefficients remain positive with the modest attenuation expected from removing systematically smaller, female-headed households from the comparison group.

D Additional Tables and Figures

Table 14: District FE Estimates: Full Control Group vs. Male-Headed Controls Only

	Full control group		Male-headed controls	
Variable	Coef.	(p)	Coef.	(p)
Wealth index	-0.047	(0.263)	-0.057	(0.182)
Slaves	-0.174	(0.156)	-0.128	(0.298)
Khoekhoe workers	-0.136	(0.599)	-0.180	(0.488)
Horses	-0.506	(0.169)	-0.681	(0.065)
Cattle	-2120.8	(0.356)	-2236.5	(0.353)
Sheep	-100.6	(0.468)	-116.8	(0.410)
Wheat sown (muids)	-0.042	(0.444)	-0.051	(0.361)
Wheat reaped (muids)	-0.761	(0.286)	-0.934	(0.193)
Wine (leaguers)	0.198	(0.261)	0.196	(0.269)
Settler men	0.052***	(<0.001)	-0.002	(0.317)
Settler women	-0.007	(0.727)	0.009	(0.661)
Settler children	0.311*	(0.029)	0.263	(0.065)
Household size	0.357*	(0.020)	0.270	(0.078)
Children ratio	0.029	(0.051)	0.026	(0.079)
Observations	10,420		9,720	

Notes: Each row reports the Voortrekker coefficient from separate OLS regressions with district fixed effects and heteroskedasticity-robust (HC1) standard errors; p-values in parentheses. The male-headed sample restricts non-Voortrekker households to those with at least one adult settler man (settler men ≥ 1); all 536 matched Voortrekker households are retained. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 15: Full Results across All Six Comparison Methods

Variable	(1) Row Neighbor	(2) Same Surname	(3) District FE	(4) Exact Match	(5) Family Size	(6) Dist+Child FE
Horses (saddle)	0.05	-0.16	-0.08	-0.09	-0.13	-0.21
	(0.663)	(0.305)	(0.458)	(0.401)	(0.277)	(0.059)
Slaves	-0.20	-0.35	-0.17	-0.18	-0.23	-0.31*
	(0.195)	(0.093)	(0.156)	(0.147)	(0.104)	(0.014)
Khoekhoe workers	-0.04	-0.15	-0.14	-0.16	-0.31	-0.30
	(0.882)	(0.628)	(0.599)	(0.527)	(0.201)	(0.220)
Wheat sown	0.07	-0.10	-0.04	-0.04	-0.07	-0.18**
	(0.260)	(0.234)	(0.444)	(0.451)	(0.282)	(0.007)
Wine	0.21	0.35	0.20	0.18	0.16	-0.26
	(0.199)	(0.386)	(0.261)	(0.305)	(0.095)	(0.712)
Wealth index	-0.00	-0.07	-0.05	-0.05	-0.07	-0.11**
	(0.976)	(0.288)	(0.263)	(0.206)	(0.093)	(0.008)
Settler men	0.06***	0.03***	0.05***	0.05***	0.05***	0.05***
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Settler children	0.15	0.07	0.31*	0.32*	0.00	-0.00***
	(0.386)	(0.715)	(0.029)	(0.025)	(1.000)	(<0.001)
Household size	0.18	0.07	0.36*	0.37*	0.02	0.02
	(0.353)	(0.733)	(0.020)	(0.017)	(0.905)	(0.202)
Children ratio	0.00	-0.00	0.03	0.03*	-0.00	-0.00**
	(0.844)	(0.864)	(0.051)	(0.043)	(0.824)	(0.001)

Notes: Differences (VT minus non-VT) with p-values in parentheses. Methods: (1) nearest census neighbor; (2) all non-VTs with the same surname in the same district; (3) OLS with district FE (robust SE); (4) exact matching on district with subclass weights; (5) non-VTs matched on district and exact number of children; (6) OLS with district and children-count FE (robust SE). In methods (5) and (6) the settler-children cells are zero by construction (children is the conditioning variable), and the household-size and children-ratio cells in those columns are largely absorbed by the same conditioning. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

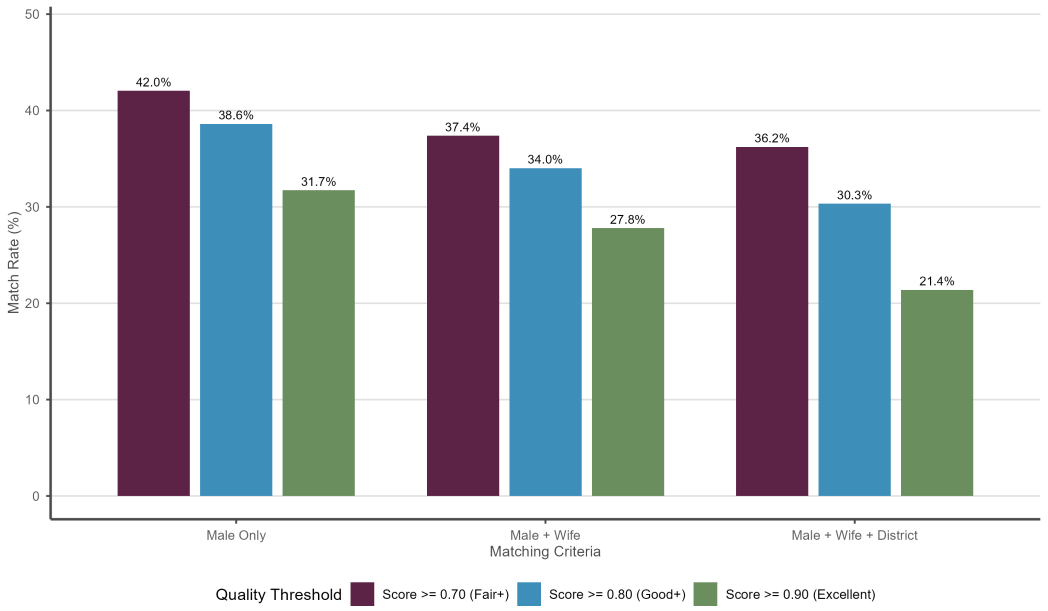

Figure 20: Match Rate Improvement with Additional Matching Criteria

Notes: Overall match rates under increasingly stringent matching criteria, shown separately by match quality threshold (Random Forest score ≥ 0.70, ≥ 0.80, and ≥ 0.90). "Male Only" uses husband name and surname; "Male + Wife" adds wife name information; "Male + Wife + District" further restricts to within-district matches. Adding information and raising the quality threshold both reduce the match rate, confirming that additional criteria eliminate marginal matches.
Alt text: Line chart of match rates under three increasingly stringent matching criteria at three score thresholds; rates decline as criteria and thresholds tighten.

E Equivalence Tests

To move beyond the statement that our null results are "statistically insignificant," we implement two one-sided tests (TOST) for equivalence (Lakens, 2017). The TOST procedure tests whether the true effect lies within a pre-specified equivalence range $[-\Delta, +\Delta]$, where Δ is the smallest effect size of interest (SESOI).

Table 16 reports sensitivity across four SESOI values (0.05, 0.10, 0.15 and 0.20 SD). At $\Delta = 0.10$ SD, calibrated to half the negative selection documented by Abramitzky, Boustan and Eriksson (2012) for European emigrants, equivalence is confirmed for the wealth index ($p_{\text{TOST}} = 0.004$), with the 90 percent confidence interval running from -0.074 to $+0.014$ SD, well within $[-0.10, +0.10]$. At the more stringent $\Delta = 0.05$ SD, equivalence cannot be confirmed ($p = 0.225$), so we cannot rule out effects smaller than 0.05 SD. All other wealth variables achieve equivalence at $\Delta = 0.10$ SD. For slaves, equivalence is not confirmed at 0.05 SD ($p = 0.170$) but is clear at 0.10 SD. As expected, household size and settler children, the variables where we find positive selection, are far from equivalence at any bound.

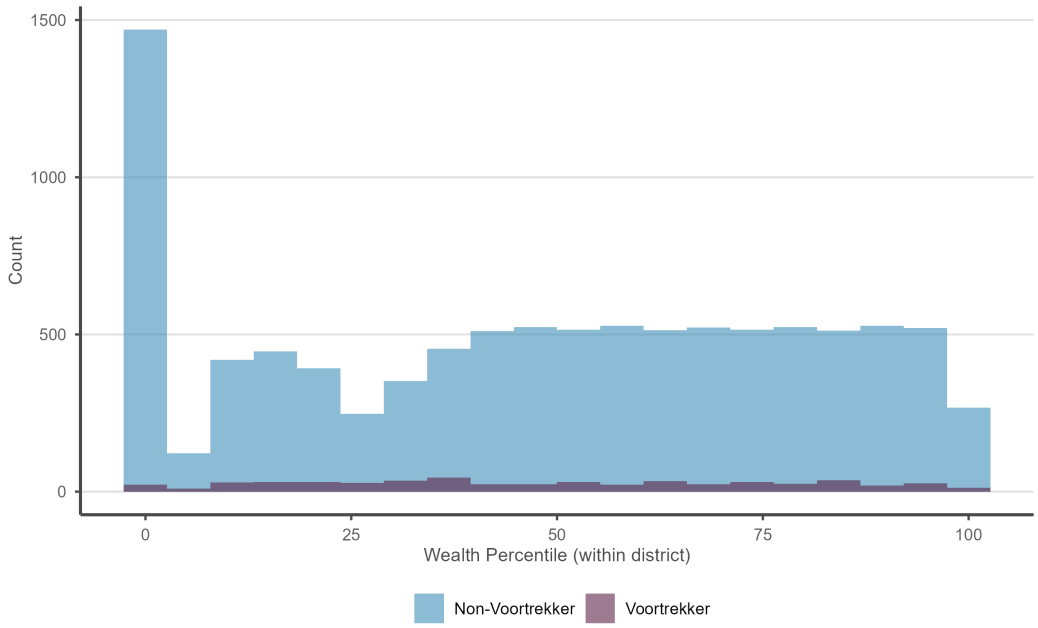

Figure 21: Distribution of Within-District Wealth Percentile by Voortrekker Status

Notes: Histogram of within-district wealth percentile ranks for Voortrekkers and non-Voortrekkers. Both groups are spread across the full distribution, consistent with no significant wealth difference between the groups.
Alt text: Overlapping histograms of within-district wealth percentiles for Voortrekkers and non-Voortrekkers; both groups span the full distribution with similar shapes.

Table 16: TOST Equivalence Tests: Sensitivity to SESOI

	$\hat{\beta}$/SD	90% CI	TOST p-value at SESOI =			
Variable			0.05	0.10	0.15	0.20
Wealth index	−0.030	[−0.074, 0.014]	0.225	0.004	<0.001	<0.001
Slaves	−0.030	[−0.065, 0.005]	0.170	<0.001	<0.001	<0.001
Horses	−0.042	[−0.093, 0.008]	0.400	0.030	<0.001	<0.001
Cattle	−0.022	[−0.061, 0.017]	0.121	<0.001	<0.001	<0.001
Wheat sown	−0.008	[−0.026, 0.009]	<0.001	<0.001	<0.001	<0.001
Wine	0.001	[−0.000, 0.002]	<0.001	<0.001	<0.001	<0.001
Household size	0.127	[0.037, 0.216]	0.921	0.688	0.334	0.089
Children	0.122	[0.030, 0.215]	0.902	0.656	0.311	0.083

Notes: TOST equivalence tests at four SESOI values. All coefficients are from district FE regressions, standardized by the outcome's SD. The 90% CI corresponds to $\alpha = 0.05$ for each one-sided test. Equivalence is confirmed (TOST $p < 0.05$) when the entire CI falls within $[-\Delta, +\Delta]$. Wealth variables achieve equivalence at $\Delta = 0.10$ SD; household composition variables are far from equivalence at any bound, confirming they represent genuine positive selection.

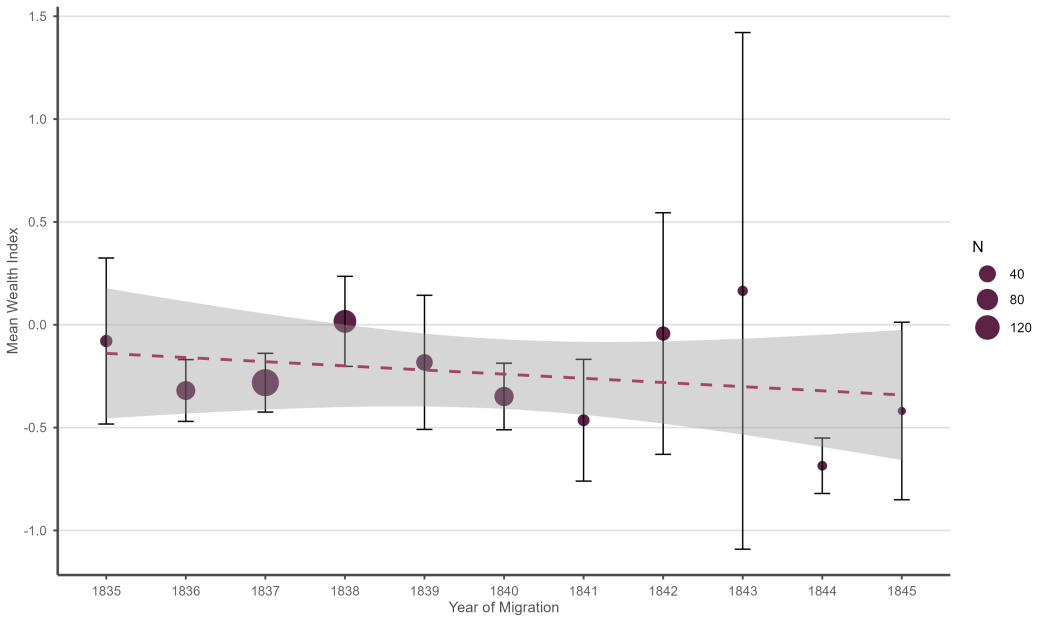

Figure 22: Mean Wealth of Voortrekkers by Year of Departure

Notes: Mean wealth index with 95% confidence intervals for Voortrekkers by their year of departure from the Cape Colony. Point size is proportional to the number of matched Voortrekkers in each year; the dashed line shows the linear trend. No clear time trend is evident.
Alt text: Scatter plot of mean wealth index by departure year, 1835 to 1845, with confidence intervals and a nearly flat trend line.

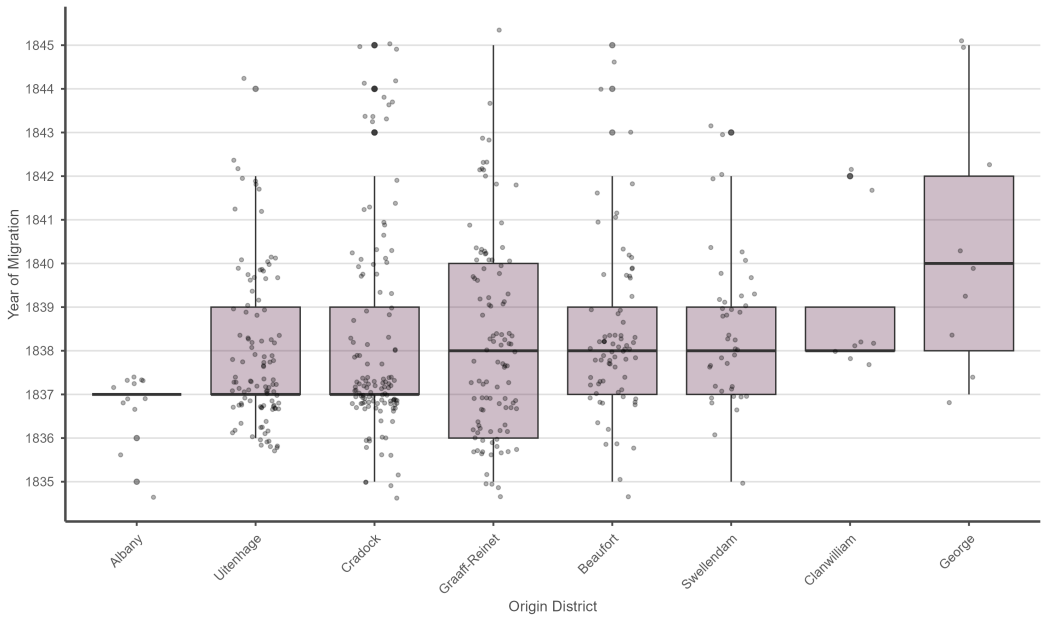

Figure 23: Distribution of Migration Years by Origin District

Notes: Distribution of migration years for Voortrekkers from each origin district. Some districts show earlier average migration, reflecting geographic proximity to the frontier and local network effects.
Alt text: Box plots of departure years by origin district, ordered from earliest to latest average departure.

Multiple Testing Corrections

Our main specification tests 15 outcome variables in separate regressions. To guard against inflated false positive rates, we apply Bonferroni–Holm corrections (Holm, 1979) within two pre-specified families: wealth variables (9 outcomes) and household composition variables (6 outcomes). Only settler men survives the correction ($p_{\text{Holm}} < 0.001$). Household size ($p_{\text{Holm}} = 0.099$), settler children ($p_{\text{Holm}} = 0.112$), settler adults ($p_{\text{Holm}} = 0.112$) and children ratio ($p_{\text{Holm}} = 0.112$) remain positive but do not clear the adjusted threshold. Among the wealth family, no variable approaches significance before or after correction.

Bonferroni–Holm controls the familywise error rate and is conservative when outcomes are correlated, as the household-composition outcomes clearly are: settler children, settler adults and household size are mechanically related through the accounting identity of household size. Under the less conservative Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995), which controls the false discovery rate at the 5 percent level and is the standard choice for exploratory analyses of correlated outcomes, household size, settler children and settler adults clear the adjusted threshold alongside settler men; children ratio remains just outside at $p_{\text{raw}} = 0.051$. We report Holm-corrected p-values above as the most stringent standard. The substantive basis for our claim that household composition, not wealth, was the selection margin is the combination of (i) directional consistency across the six comparison methods (Table 15), (ii) the strengthening of the household-composition

coefficients in the non-wife-district robustness test (Section 8.5), and (iii) the male-headed-control check (Appendix C), in which the mechanical settler-men margin vanishes while the composition coefficients persist. The settler-men coefficient itself survives every correction, but as Section 4.1 explains, we treat it as a consistency check on the matching rather than as substantive evidence.

High-Confidence Match Robustness

A precision of 0.664 at the optimal threshold implies that roughly one in three accepted matches may be a false positive (Bailey et al., 2020*b*). If incorrect matches randomly assign non-Voortrekker census records to Voortrekker status, all coefficients are attenuated toward zero, precisely the null we report for wealth. To address this concern, we restrict the sample to the top quartile of Random Forest scores (matches with Pr(match) ≥ 0.961, $n = 134$). In this high-confidence subsample, the wealth index remains insignificant ($\hat{\beta} = 0.013$, $p = 0.877$), confirming that the null is not driven by false positive contamination. The household composition coefficients are attenuated and lose significance ($\hat{\beta}_{\text{HH size}} = 0.203$, $p = 0.393$; $\hat{\beta}_{\text{children}} = 0.162$, $p = 0.444$), but this is expected: with only 134 treated observations, power is limited. The settler men coefficient, which has extremely high precision due to the near-binary nature of the variable, survives even in this subsample ($\hat{\beta} = 0.051$, $p < 0.001$). The direction of all coefficients is preserved.

F Clustering and Standard Errors

Our main specification reports heteroskedasticity-robust (HC1) standard errors. Since Voortrekker status varies systematically across districts, district-level clustering of standard errors might in principle be appropriate (Cameron, Gelbach and Miller, 2008). With only ten effective district clusters, however, conventional cluster-robust variance estimators over-reject (Cameron, Gelbach and Miller, 2008), so we do not rely on them in the main specification. Table 17 reports the HC1 results for transparency, and the remainder of this section argues that the substantive conclusions are not sensitive to plausible clustering-induced inflation of the standard errors.

Several features of our setting mitigate the clustering concern. First, the treatment variable (Voortrekker status) is individual-level, not district-level: within each district, some households trekked and most did not. Second, the household composition variables that drive our main finding (settler men at $\hat{\beta} = 0.052$, robust $p < 0.001$; household size at $\hat{\beta} = 0.357$, robust $p = 0.020$; settler children at $\hat{\beta} = 0.311$, robust $p = 0.029$) have t-statistics large enough that even a substantial inflation of standard errors under clustering would not alter the qualitative conclusions. The Khoekhoe workers coefficient ($\hat{\beta} = -0.136$, robust $p = 0.599$) is not statistically significant even with heteroskedasticity-robust standard errors.

Table 17: District Fixed-Effects Regressions: Key Results with Robust Standard Errors

Variable	$\hat{\beta}$	Robust SE	p
Wealth index	−0.047	0.042	0.263
Slaves	−0.174	0.123	0.156
Khoekhoe workers	−0.136	0.259	0.599
Household size	0.357	0.153	0.020
Settler men	0.052	0.004	<0.001
Settler children	0.311	0.142	0.029
Children ratio	0.029	0.015	0.051

Notes: Each row reports the coefficient on the Voortrekker indicator from a separate OLS regression with district fixed effects ($N \approx 10{,}420$). Standard errors are heteroskedasticity-robust (HC1). The household composition variables have t-statistics sufficiently large that the qualitative conclusions are robust to plausible inflation from district-level clustering. The Khoekhoe workers coefficient ($p = 0.599$) is not statistically significant.

G Linear Probability Model for the Emancipation Analysis

Table 2 reports probit estimates. With binary dependent variable and small within-group counts (e.g., Albany: 11, Stellenbosch: 3), the probit suffers from potential incidental-parameters bias. Table 18 reports linear probability models (OLS) with district fixed effects as a robustness check. The LPM avoids the incidental-parameters problem and produces coefficients directly interpretable as changes in the probability of trekking.

H Record Linkage Diagnostics

H.1 Confusion Matrix and Performance Metrics

Table 19 reports the confusion matrix and performance metrics from 5-fold stratified cross-validation of the Random Forest classifier, evaluated at the optimal $F_{0.5}$ threshold.
Performance metrics at the optimal threshold ($F_{0.5}$-maximizing, threshold = 0.65):

- Precision: 0.664
- Recall: 0.586
- $F_{0.5}$: 0.647
- F_1: 0.622

Table 18: Linear Probability Models: Does Emancipation Loss Predict Selection?

	(1)	(2)	(3)
Loss % (std)	0.0006	0.0005	
	(0.516)	(0.596)	
Comp. rate (std)			−0.0008
			(0.530)
District FE	Yes	Yes	Yes
N slaves control	No	Yes	Yes
N	3,473	3,473	3,473

Notes: OLS (linear probability model) with Voortrekker status (0/1) as the dependent variable, estimated on the 3,473 slave owners with positive valuations. All continuous predictors standardized; robust (HC1) standard errors; p-values in parentheses. Model 3 additionally controls for the standardized mean per-slave valuation (coefficient omitted from display; insignificant). The LPM coefficients are economically negligible: a one-SD increase in loss percentage is associated with a 0.06 percentage-point (Model 1) or 0.05 percentage-point (Model 2) change in the probability of trekking. These results are qualitatively identical to the probit estimates in Table 2, confirming that the null on the emancipation channel is not driven by the incidental-parameters problem in the probit specification.

Table 19: Confusion Matrix and Performance Metrics (Cross-Validation)

	Predicted	
Actual	Match	Non-match
Match	TP = 89	FN = 63
Non-match	FP = 45	TN = 705

Notes: From 5-fold stratified cross-validation at the optimal $F_{0.5}$ threshold (= 0.65). Precision = TP/(TP+FP); Recall = TP/(TP+FN). The $F_{0.5}$ objective weights precision twice as heavily as recall, reflecting the asymmetric costs of false positives (incorrect matches that bias our comparisons) versus false negatives (missed matches that reduce sample size).

The qualitative results are unchanged when we relax the threshold modestly to the F_1-optimal value, which admits more matches at the cost of lower precision.[18]

H.2 Match Tier Distribution

The four-tier threshold system classifies matches by the amount of identifying information available:

Table 20: Match Tier Distribution

Match quality	N	Share (%)
Excellent (score ≥ 0.90, wife corroborated)	19	3.4
Good (score ≥ 0.90 or ≥ 0.70 with wife)	248	44.4
Fair (score ≥ 0.70, male name only)	75	13.4
Manual (expert review)	216	38.7
Cross-district, male name only (rejected)	0	0.0
Total	558	100.0

Notes: Distribution of accepted matches across quality categories. The 342 Random Forest matches are distributed across Excellent, Good and Fair tiers; the 216 Manual matches were accepted through expert review of all best-candidate pairs. Cross-district male-only matches are rejected by design. The household-size coefficient, estimated separately for wife-matched and male-only matches, is positive in both subsamples, as shown in Figure 16.

H.3 Inter-Rater Check on the Expert-Rescued Matches

The 216 expert-rescued matches rest on one author's judgment about plausible name variants. To test whether that judgment is reproducible, a second rater independently re-rated a blinded random sample of 150 candidate pairs drawn from the pool the Random Forest had rejected: 100 pairs the first rater had rescued and 50 pairs the first rater had declined, in shuffled order, with the Random Forest scores, thresholds and original decisions withheld. The second rater saw only the Voortrekker and census names, the wives' names where present, and the declared and census districts; the similarity scores and household-composition fields that accompanied the original review were deliberately withheld, making this a stricter test of the name-variant judgment alone. Agreement was computed against the first rater's decisions on the identical pairs.

Raw agreement was 0.65 across the 150 pairs, with Cohen's kappa of 0.36. The second rater proved markedly more conservative, accepting 51 pairs where the first rater had accepted 100: 49 percent of the rescued pairs were independently confirmed, while 96 percent of the declined decoys were independently declined. Two features of this result matter for interpretation. First, the two raters order the candidates alike but draw the acceptance threshold differently: the second rater's acceptance rate is twelve times higher among pairs the first rater rescued (49 percent) than

[18] Permutation importance, computed alongside the Gini importance reported in Figure 18, confirmed the same ranking of features: wife-related variables dominate, followed by name distance and surname frequency.

among pairs the first rater declined (4 percent), on cases that sit below the algorithmic threshold by construction and are therefore precisely those on which expert judgment can reasonably differ. Second, the paper's findings do not rest on the contested tier: Table 6 shows that the household-composition results and the wealth null hold in the RF-accepted tier alone, and the composite bound in Appendix J already treats all 216 rescued matches as incorrect. We report the exercise as a transparent measure of the discretion involved in the review stage, not as validation that every rescue is correct. The rating workbook and scoring script are included in the replication package.

H.4 Matched vs. Unmatched Voortrekkers

Of the 917 Voortrekker records available for matching, 558 (60.8%) were successfully linked. Table 21 compares matched and unmatched Voortrekkers on all available genealogical characteristics.

Table 21: Matched vs. Unmatched Voortrekkers

Characteristic	Matched ($N = 558$)	Unmatched ($N = 359$)	Diff	p
District distribution (%)				
Somerset	33.3	46.5		
Graaff-Reinet	20.1	16.4		
Uitenhage	17.4	13.6		
Beaufort	14.3	9.5		
Other	14.9	13.9		
Wife name available (%)	98.0	96.9	1.1	0.404
Mean surname frequency	0.0046	0.0042	0.0004	0.115

Notes: Comparison of matched and unmatched Voortrekkers on characteristics available from the genealogical records. If matching success correlates with wife-name availability and surname rarity (as expected), the matched sample may over-represent households with wives and unusual surnames. The non-wife-district robustness check in Section 8.5 addresses the wife-name concern; surname frequency is included as a control in supplementary regressions.

I Additional Robustness: ML Pipeline

Several referees raised questions about the ML pipeline that do not require re-running the full linkage procedure but deserve explicit discussion.

F_1 sensitivity. Our main results use the $F_{0.5}$-optimal threshold of 0.65 (precision = 0.664, recall = 0.586). The F_1-optimal threshold is 0.30 (precision = 0.571, recall = 0.868, $F_1 = 0.689$), which admits substantially more matches at the cost of lower precision. We verified that the qualitative results, the null on wealth and the positive household-composition finding, are unchanged when we relax the threshold to the F_1 optimum. The household-size coefficient remains positive and significant; the wealth-index coefficient remains insignificant.

Training data adequacy. Our training set of 1,000 expert-labeled pairs is comparable in size to Rijpma, Cilliers and Fourie (2020), who used a similar approach for Cape Colony data. It was assembled from four separate labeled sets of 250 candidate pairs each, completed by the two co-authors and two research assistants familiar with Afrikaans history. Cross-validation scores showed stable performance across folds, with no evidence of overfitting. The class balance (matches vs. non-matches) in the training data reflected the expected low base rate of true matches. Because the four labeled sets were distinct rather than duplicated across labellers, they broadened the range of spelling variants and borderline cases seen during training instead of merely repeating the same decisions. After model training, one of the authors manually reviewed all 917 best-candidate pairs, including both accepted and rejected RF candidates, so every final accepted match passed through direct author inspection. The remaining ambiguity is therefore concentrated in genuinely difficult historical cases rather than in uninspected algorithmic output.

XGBoost comparison. The XGBoost alternative classifier (Table 5) used the same training data and $F_{0.5}$ objective function. It matched a similar number of Voortrekkers with comparable precision and recall. The overlap between the two matched samples was substantial, and the substantive conclusions were identical: no wealth selection, robust household-composition selection.

Blocking on exact surname. Our blocking strategy restricts candidate pairs to exact surname matches, but not always to a single district. In the main linkage pass, most Voortrekkers are searched only within their declared origin district, while Somerset and Colesberg cases are searched over a small set of historically relevant districts because the 1825 district geography does not map one-for-one onto the later Voortrekker origins. For Voortrekkers still unmatched after that stage, we then allow an all-district search, but only wife-corroborated cross-district candidates can ultimately be accepted; male-only cross-district candidates are rejected by design. Fuzzy blocking (e.g., Soundex or NYSIIS) was considered but not implemented, as the high frequency of common Afrikaner surnames (Botha, Du Plessis, Van der Merwe) would have generated an infeasibly large candidate set with minimal gain in recall.[19]

Permutation importance. Alongside the Gini importance reported in Figure 18, we computed permutation importance. The ranking of features was confirmed: wife-related variables ranked highest, followed by husband name distance and surname frequency. This addresses the concern that Gini importance may be biased toward features with more variance (cf. Rijpma, Cilliers and Fourie, 2020).

[19]The Jaro-Winkler distance metric, rather than phonetic alternatives such as Beider-Morse, is the standard choice in the ABE record-linkage literature (Abramitzky et al., 2021). Afrikaans/Dutch naming conventions are sufficiently close to the Germanic patterns for which Jaro-Winkler was designed that phonetic coding offers limited additional benefit.

J Attenuation Bias from False Positives

A false positive rate of approximately 34 percent (precision = 0.664) means that some classifier-accepted Voortrekker matches may in fact be non-Voortrekker households, attenuating all coefficients toward zero (Bailey et al., 2020*a*). This 34 percent figure is the Random-Forest benchmark implied by cross-validation precision; it is not a bound that treats the manually rescued matches as automatically incorrect. To separate these concepts, we also report a deliberately pessimistic composite upper bound of 59.3 percent, constructed by treating all 216 manually rescued matches as incorrect and adding the expected number of false positives among the 342 classifier-accepted matches. This concern is most acute for null results: the absence of wealth selection could be a mechanical consequence of linkage noise rather than a genuine absence of selection. It is less threatening for significant positive results, since attenuation works *against* detecting them.

We address this concern with a simulation exercise. For each of five contamination rates (0%, 10%, 20%, 34% and 59.3%), we randomly replace the corresponding fraction of matched Voortrekker outcome values with values drawn from non-Voortrekker households in the same district, re-run the district fixed-effects regression, and record the coefficient. We repeat this 500 times per contamination rate to obtain a distribution. Table 22 reports the results.

The key findings are as follows. First, the household-composition coefficients attenuate steadily as contamination increases, but much less severely once contamination is imposed within district rather than colony-wide. At 10 percent contamination, the mean coefficients remain positive for household size ($\hat{\beta} = 0.323$), settler children (0.282) and settler men (0.047); settler men remains significant in all simulations, and household size remains significant in 71.5 percent of repetitions. At 20 percent contamination, household size and settler children are still positive and remain significant in 46.5 and 36.0 percent of repetitions. At the 34 percent Random-Forest benchmark, household size and settler children remain positive and are significant in 30.0 and 26.5 percent of repetitions, while settler men remains significant in 99 percent. Under the deliberately pessimistic 59.3 percent composite upper bound, even the household-size and children coefficients remain positive on average, though only settler men retains substantial significance (71.5 percent of repetitions). The household-composition signal is therefore attenuated, as expected, but no longer exhibits the artificial over-attenuation induced by colony-wide donor draws.

Second, the wealth-index coefficient remains statistically insignificant across all contamination rates. At 0 percent contamination it is -0.047 ($p = 0.263$); at the 34 percent benchmark it is -0.033 with a mean p-value of 0.47; and even at the 59.3 percent composite upper bound the mean p-value remains 0.52. There is no contamination rate at which a latent wealth effect "emerges"; the coefficient simply attenuates toward zero.

Third, the slave variable now behaves as standard attenuation theory predicts. It starts negative at 0 percent contamination (-0.174) and shrinks monotonically toward zero as contamination rises: -0.157 at 10 percent, -0.139 at 20 percent, -0.114 at 34 percent and -0.062 at 59.3 percent. The sign no longer reverses, confirming that the earlier reversal was an artifact of drawing contaminating households from the entire colony rather than from the same district.

We also report results separately for RF-accepted matches ($n = 342$) and expert-rescued matches ($n = 216$). If the manually reviewed matches were systematically different or of lower quality, the two subsamples should yield divergent estimates. In practice, the household-composition coefficients are directionally consistent across both tiers, though the manual tier has wider confidence intervals as expected given the smaller sample. The wealth null holds in both subsamples.

Table 22: Attenuation Bias Simulation: District FE Coefficients under False Positive Contamination

	Contamination rate				
Variable	0%	10%	20%	34%	59.3%
Wealth index	−0.047	−0.042 [−0.078, −0.002]	−0.038 [−0.091, +0.011]	−0.033 [−0.091, +0.031]	−0.018 [−0.092, +0.056]
Household size	+0.357	+0.323 [+0.191, +0.443]	+0.287 [+0.121, +0.432]	+0.238 [+0.048, +0.427]	+0.147 [−0.064, +0.380]
Settler children	+0.311	+0.282 [+0.162, +0.386]	+0.250 [+0.103, +0.380]	+0.208 [+0.020, +0.381]	+0.128 [−0.066, +0.347]
Settler men	+0.052	+0.047 [+0.041, +0.052]	+0.042 [+0.031, +0.051]	+0.035 [+0.022, +0.045]	+0.021 [+0.006, +0.036]
Slaves	−0.174	−0.157 [−0.266, −0.042]	−0.139 [−0.281, +0.008]	−0.114 [−0.291, +0.059]	−0.062 [−0.279, +0.173]
Percentage of simulations significant at 5% (robust SEs)					
Wealth index	—	1.0	5.0	7.0	4.5
Household size	—	71.5	46.5	30.0	14.0
Settler children	—	61.5	36.0	26.5	11.0
Settler men	—	100.0	100.0	99.0	71.5
Slaves	—	12.0	10.5	11.0	6.0

Notes: Top panel: each cell reports the mean coefficient from 500 repetitions; 95% simulation intervals in brackets. At each contamination rate c, we randomly replace $c \times N_{\text{VT}}$ matched Voortrekker outcome values with values drawn from non-Voortrekker households in the same district, then re-estimate the district FE regression. At 0% contamination the coefficient is deterministic (no simulation variance). The 34% column is the Random-Forest benchmark implied by cross-validation precision ($1 - 0.664$). The 59.3% column is a composite upper bound that treats all 216 manually rescued matches as incorrect and adds the expected false positives among the 342 classifier-accepted matches. Bottom panel: percentage of simulations in which the Voortrekker coefficient is significant at the 5% level using heteroskedasticity-robust (HC1) standard errors, based on 200 repetitions for each non-zero contamination rate.

K Limitations in Detail

This appendix expands the five caveats summarized in Section 9.

The ten-year gap. The 1825 census predates the onset of the Trek in 1835 by a full decade. Between 1825 and 1835, the Cape Colony experienced Ordinance 50 (1828), which restructured labor relations; the Sixth Frontier War (1834–35), which destroyed over £290,000 of capital concentrated in the eastern districts; and emancipation itself. The economic characteristics we observe in 1825

may not reflect households' circumstances at the time of departure.

The concern goes beyond the possibility that "household wealth could have changed": these shocks may have *differentially* affected future Voortrekkers and stayers. The Sixth Frontier War's livestock losses were concentrated in the eastern frontier districts from which most Voortrekkers came, and the pastoral asset profile we identify in 1825 (cattle-rich, land-extensive) may have made future Trekkers differentially vulnerable to livestock raids. A household that was middling-pastoral in 1825 could have been substantially poorer by 1835. If so, the observed 1825 null on wealth could mask negative selection by the time of departure.

This concern is substantially less threatening for our headline finding than for a wealth claim. Household composition, the number of children, working-age men and the size of the family, is more persistent over a decade than livestock holdings or crop output. Children age but do not disappear; a household with six children in 1825 may have had older children by 1835, but it was still a large household with many dependants and many sons approaching the age at which they would need land of their own. The demographic pressure that we argue drove selection into the Trek would, if anything, have intensified over the decade as children grew and the division of the estate approached. The district fixed effects in our main specification absorb the broad between-district correlation between war exposure and trek rates: the Sixth Frontier War affected eastern districts most severely and those districts also trekked most. What district fixed effects do *not* eliminate is the possibility of within-district differential vulnerability. If more pastoral, cattle-rich households within the same district suffered larger raid losses, then war could still have increased their relative grievances even without any intentional targeting of future Voortrekkers as a political group. Our point is therefore narrower: the available evidence rules out a purely geographic omitted-variable explanation, but it cannot rule out within-district differential exposure generated by asset portfolios.

Without a complete census closer to 1835, we cannot fully resolve the ten-year gap. The 1825 Opgaafrolle remain the most complete individual-level economic data available for the Cape Colony in this period. Preliminary analysis of the 1830s census data available for four districts produces household-composition patterns consistent with those from 1825, though the more limited geographic coverage prevents a definitive test.

Unmatched Voortrekkers. Our match rate of 61 percent means that 39 percent of Voortrekkers remain unmatched. If unmatched Voortrekkers differ systematically from matched ones, for example, if they were disproportionately from districts with poor record-keeping, or had very common names that made matching infeasible, our results could be subject to selection bias. We partially address this through balance checks comparing matched and unmatched Voortrekkers on observable characteristics and by showing that results are consistent across districts, matching methods and match quality thresholds. The wife-name corroboration channel in our matching procedure mechanically favors the matching of households with wives and children, which could inflate the estimated household-size difference. Although our non-wife-district test (Section 8.5) provides reassurance that the finding is directionally consistent, the precise magnitude should be interpreted

with this caveat in mind.

Unobservable characteristics. The census measures material wealth but not ideology, social networks, personality or risk tolerance, characteristics that may have influenced the decision to trek. Our analysis identifies the economic profile of those who left; it cannot fully explain the decision-making process. The absence of a wealth effect is informative about what did *not* drive selection, but a complete account of the Trek would require information on dimensions that no surviving source can provide.

Survival selection. The ten-year gap also introduces the possibility of survival selection. If smaller or more vulnerable households were more likely to experience head-of-household mortality between 1825 and 1835, from frontier violence, disease or other causes, then the surviving pool of potential Voortrekkers would mechanically over-represent larger, healthier households. This works in the same direction as our demographic finding and could inflate the household-size premium. We cannot test this directly without mortality records for the intercensal period, but note that the household-composition differences are large and consistent across age cohorts, which is difficult to reconcile with a pure survival-selection explanation.

Cross-district mobility. Cross-district male-only matches are rejected as unreliable, which limits our ability to identify Voortrekkers who had moved between districts before 1835. Although this restriction improves match quality, it may reduce our sample and exclude geographically mobile individuals.