Biplots for historical household data: Evidence from Cape tax records*

Johan Fourie† Sugnet Lubbe‡ Johané Nienkemper-Swanepoel§ Dieter von Fintel¶

Abstract

Historical sources like tax lists, inventories or censuses record many quantities for each household, but historians often summarise them one quantity at a time. We show how a principal component analysis (PCA) biplot maps combinations of holdings and measures what the map leaves out. Our worked example uses 20,913 household-year returns from the Cape Colony's tax censuses, from eligible years within 1804–1825. We explain how source decisions and transformations shape the display, how to read calibrated axes and test the result. Two dimensions retain 70.1% of the standardised, log-transformed variation. The first separates households with small recorded holdings from those with large ones; the second separates viticulture from grain and livestock. This picture changes little when households are resampled. A classification into household types is less stable: the number of groups changes, as does the share of each asset's inequality that they account for. A common coordinate system allows comparison across periods, but cannot distinguish economic change from recording changes. Household records reveal what aggregates hide: most of the variation in a colony-wide cross-section lies within districts. The paper provides a practical guide to biplots and the checks needed to judge when the picture can be trusted.

Keywords: biplots; principal component analysis; historical microdata; household production; inequality; Cape Colony

JEL codes: N17; C38; D63

*We thank the Riksbankens Jubileumsfond (Cape of Good Hope Panel grant M20-0041) for financial support. This paper was created with the help of Anthropic's Claude Code (Opus and Fable) and OpenAI's Codex (Sol and Astra), with an additional AI-assisted review by Refine.ink. Cite this paper as: Fourie, Johan, Sugnet Lubbe, Johané Nienkemper-Swanepoel, and Dieter von Fintel. 2026. "Biplots for historical household data: Evidence from Cape tax records." Working Paper, Department of Economics, Stellenbosch University.

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

‡Department of Statistics and Actuarial Science, Stellenbosch University. Email: slubbe@sun.ac.za.

§Department of Statistics and Actuarial Science, Stellenbosch University. Email: nienkemperj@sun.ac.za.

¶Department of Economics, Stellenbosch University. Email: dieter2@sun.ac.za.

1 Introduction

Every year, officials at the Cape Colony asked settler households to declare what they owned and produced. A single return from Stellenbosch in the early nineteenth century lists the enslaved men and women attached to the household, its Khoekhoe labourers, its cattle, horses, sheep and goats, its wagons, the number of vines on the farm and the wine, brandy, wheat, barley and rye it produced. More than a hundred thousand of these returns survive for Stellenbosch and Drakenstein alone. Together they form a table with one row for each household in each year and a column for each quantity. Such a table contains a remarkable amount of information about how colonial farms worked. It is also almost impossible to read.

Historians usually summarise such tables one column at a time: the average number of cattle, the share of households owning enslaved people, a Gini coefficient for vines. These summaries discard the combinations within each household. Two populations can have identical average holdings of every asset yet differ entirely in how those assets are combined: in one, the households with the most vines might also hold the most livestock; in the other, vines and livestock might belong to different households. Households with similar combinations may also differ greatly in scale.

This paper explains how to use the principal component analysis (PCA) biplot for historical household records. A biplot maps a data table: households appear as points and recorded quantities as calibrated axes. A reader can project a household's point onto an axis to read its approximate holding of cattle or wine. Compressing many columns into two dimensions loses information. The method reports how much it loses, overall, for each quantity and for each household.

Our worked example uses the Cape tax censuses, the *opgaafrollen*. The main display contains 20,913 household-year returns for Stellenbosch and Drakenstein from eligible years within 1804–1825. A pooled comparison adds eligible years from 1738, and a separate cross-section covers district returns across the colony dated 1823–1825. The application shows how decisions about source coverage, blanks and transformations shape a biplot, and how to check whether the resulting map supports a historical claim. Table 1 sets out these decisions.

We choose PCA because our question concerns the overall structure of the table. Its linear approximation provides axes that can be read in source units throughout the display. Newer methods such as t-SNE and UMAP better preserve near neighbours in our data, but lack calibrated axes; distances in their displays do not measure economic differences. For differences between predefined groups, an appendix illustrates canonical variate analysis.

Two dimensions retain 70.1% of the variation in the standardised, log-transformed Cape table. The first describes scale: households on the right hold more of nearly everything. The second separates vines, wine and brandy from grain and several kinds of livestock. This picture barely moves when we resample households or replace the log transformation with the inverse hyperbolic sine. Without transformation, a few very large holdings dominate the map and fit falls to 49.7%.

In the colony-wide cross-section, differences within districts account for 86.5% of the standardised, transformed variation. District averages would discard most of that variation before any biplot was drawn. Individual returns allow us to ask whether each asset's inequality lay mainly between

broad types of household or within them.

The classification is less stable. A clustering procedure selects two groups, but usually chooses three when households are resampled. Excluding the asset being measured from the grouping reverses the balance of within- and between-group inequality in vines and wine. A stable map of holdings therefore need not imply a stable typology of households. Neither the biplot nor the decomposition identifies what caused these patterns.

Inventories and related records have long been used to reconstruct household economies. English probate inventories describe combinations of production and consumption (Overton et al., 2004); marriage and death inventories in Württemberg reveal the asset and debt positions behind participation in credit markets (Ogilvie, Küpker and Maegraith, 2012). At the Cape, probate inventories have been used to measure settler wealth (Fourie, 2013), and tax censuses to estimate wheat productivity (Fourie and Greyling, 2023), the relationship between enslaved labour and wine output (Fourie and Greyling, 2024) and the productivity of indigenous labour (Fourie and Green, 2015). Fourie et al. (2024) describe the Stellenbosch–Drakenstein transcription used here. We show how to display holdings jointly and measure the variation omitted. Applying the procedure elsewhere requires attention to each source's population and accounting categories.

The paper also complements research on preindustrial inequality (Milanović, Lindert and Williamson, 2011; Alfani, 2021), including at the Cape (Fourie and Von Fintel, 2011, 2014; Fourie, 2025). Single-asset indices cannot show how different holdings move together or how much concentration occurs among households with similar production profiles. We use an asset-specific Theil decomposition across profiles estimated from the same records (Theil, 1967; Bourguignon, 1979), testing how much the answer depends on the grouping. These are distributions of recorded physical quantities; measuring total wealth would also require prices.

Biplot geometry is well established (Gabriel, 1971; Gower and Hand, 1996; Gower, Lubbe and Le Roux, 2011; Gower, Le Roux and Gardner-Lubbe, 2015), with accessible introductions (Greenacre, 2010; Greenacre et al., 2022). Ciccarelli and Proietti (2013) provide an economic-history precedent: dynamic specialisation biplots trace the industrial configuration of Italian provinces across four benchmark years. Software also supports biplot comparisons across time (Ganey and Nienkemper-Swanepoel, 2026). We address the measurement problems of historical household records, including unrecorded fields, transformation units and repeated observations. We also compare nonlinear alternatives by the local relationships they preserve (van der Maaten and Hinton, 2008; McInnes, Healy and Melville, 2018; Wang et al., 2021; Triana-Martinez et al., 2024). Observing households rather than provinces allows us to examine the distributions behind aggregate profiles, although a common temporal frame cannot separate economic change from changes in recording.

2 How a PCA biplot works

2.1 The idea in plain terms

Imagine each household as a point in space. With two recorded quantities, say cattle and vines, we can plot every household on an ordinary scatter diagram. With three, we could build a three-dimensional model. With the 13 quantities in our main analysis, each household is a point in 13-dimensional space, which no one can visualise. PCA solves this problem much as a photographer does when photographing a sculpture: it looks for the viewpoint from which the flat image preserves as much of the object's shape as possible. Formally, it finds the flat plane through the cloud of household points on which their shadows are most spread out. Equivalently, it is the plane that keeps the points' distances from the plane as small as possible.

The two directions that define this plane are the first two *principal components*. The first is the single direction along which households differ most. The second is the direction, at right angles to the first, along which they differ most once the first has been accounted for. Each component is a weighted combination of the original quantities, and its weights tell us what it measures. If every quantity has a positive weight on the first component, households far along it hold more of everything, and the component describes scale. If the second component gives positive weights to vines and wine and negative weights to wheat and sheep, it contrasts two kinds of farm.

A plain PCA plot shows only the households. A *biplot* adds the variables, each drawn as a line marked with tick values like a ruler. To read a household's approximate holding of wine, drop a perpendicular from its point to the wine axis and read the nearest tick. The position preserves only the part of its record in the two retained dimensions. We can measure the share of variation retained for the whole table, each quantity and each household.

Two decisions shape the map before any of this happens. The first concerns units. Vines are counted in thousands and wagons in ones, so without adjustment the vines column would dominate every distance. The standard remedy is to *standardise* each column, rescaling it to have mean zero and the same spread, so that each quantity starts with equal weight. The second concerns skewness. Most households held a few cattle and a few held hundreds. A logarithmic transformation compresses the large holdings so that the map is not organised around a handful of the largest farms. Both decisions change what "similar" means, and Section 2.4 discusses them in detail.

2.2 The data matrix and its approximation

The rows of our data table are household-years and its columns are recorded quantities. Let x_{ij} denote quantity j for observation i, and $y_{ij} = g_j(x_{ij})$ its transformed value (for example, $\log(1 + x_{ij})$). For correlation PCA, we standardise each transformed column: $z_{ij} = (y_{ij} - \bar{y}_j)/s_j$, where $s_j^2 = n^{-1} \sum_i (y_{ij} - \bar{y}_j)^2$. We use this population-divisor convention throughout; using $n - 1$ would change the units of the scores but not the principal directions or variance shares. Covariance PCA instead centres the transformed variables without rescaling them, so each keeps its original

Table 1: Decisions in a historical biplot, in the order a researcher faces them

Decision	Question to ask of the source and the research question	Section
Observed population	Which returns count as observed? Can a wholly blank return be distinguished from a household with nothing?	3.1, 3.2
Blanks and totals	Does a blank mean zero, an omitted entry or a field absent from the form? Do totals have the same components in every year compared?	3.2
Variable set	Does the list give some activities several columns, and therefore more weight in distances?	2.2, 4.3
Transformation	Which differences should count most: differences among large holdings, or between owning nothing and owning a little?	2.4, 4.3
Standardisation	Should each quantity receive equal initial weight (correlation PCA), or should measured scales be kept (covariance PCA)?	2.4
Weighting of records	Should a household observed for ten years count ten times? Should crowded years count more?	3.2, 4.4
Factorisation and calibration	Which scaling convention is used, and are the axes calibrated in source units?	2.2, 2.3
Fit	How well are the quantities and households relevant to the claim represented, not only the table as a whole?	4.2
Sensitivity	Does the picture survive other transformations, variable sets, weights and household resampling?	4.3
Interpretation	What can the display show, and what requires other evidence?	4.1, 5

variance. PCA decomposes the resulting matrix Z by the singular value decomposition:

$$Z = UDV^{\mathsf{T}}, \qquad \widehat{Z}_r = U_r D_r V_r^{\mathsf{T}}. \tag{1}$$

Here D contains the singular values in decreasing order; each measures how much variation one dimension carries. The subscript r denotes the first r dimensions. Among all matrices of rank at most r, $\widehat{Z}_r$ has the smallest sum of squared differences from Z. This guarantee applies to the table as constructed; it cannot establish whether the chosen observations, variables and weights adequately describe household economic life.

In the row-principal factorisation used for our displays, household coordinates are $A = U_r D_r$ and variable coordinates are $B = V_r$. The approximate standardised value of quantity j for household i is the inner product $\widehat{z}_{ij} = a_i^{\mathsf{T}} b_j$. With all dimensions retained, distances between household points reproduce the Euclidean distances between the rows of Z. With two dimensions, differences in omitted directions disappear. Households close on the map can therefore still differ substantially on poorly represented variables.

Distance measures similarity in recorded characteristics. Households with similar agricultural returns may differ greatly in unrecorded land or buildings, so proximity cannot establish comparable welfare, status or production strategies. The variable list also allocates influence: several measures of the same activity give it several contributions to every distance, even after standardisation.

Other factorisations divide the singular values differently between households and variables:

$A_\alpha = U_r D_r^\alpha$ and $B_\alpha = V_r D_r^{1-\alpha}$. The product, and hence the approximation, is the same, but the geometry of each set of points changes. Setting $\alpha = 1$ gives the row-principal convention; $\alpha = 0$ assigns the singular values to the variables; $\alpha = 1/2$ shares them symmetrically. No single choice supports every interpretation of distances and angles at once. A caption should therefore state the convention, and the plotted coordinates should be saved with the fitted object. Lengthening the arrows for visibility without adjusting their calibration breaks the link between a household's position and its approximate values.

2.3 Reading household values and variable relationships

A calibrated axis turns the inner product into something the eye can read. If b_j is a nonzero variable vector, the point $tb_j/\|b_j\|^2$ on its line receives the tick label t in standardised units. Projecting household point a_i perpendicularly onto that line then gives its approximate value $\widehat{z}_{ij}$. For a variable transformed by $\log(1 + x)$ and standardised, the ticks can be relabelled in source units, such as head of cattle or leaguers of wine, using $x = \exp(\bar{y}_j + s_j t) - 1$. These readings are reconstructions from the plotted dimensions, not the recorded values. They coincide only when the discarded part of that entry is zero (Gower, Le Roux and Gardner-Lubbe, 2015).

Suppose a variable has direction $b_j = (0.4, 0.2)$ and a household sits at $a_i = (2, 1)$. Its approximate standardised value is $2(0.4) + 1(0.2) = 1.0$. The household lies on the variable axis; any point with the same inner product receives the same predicted value. If the transformed mean were $\log(3)$ and the standard deviation 0.5, the implied quantity would be $3\exp(0.5) - 1$, or about 3.95. Appendix B reports recorded, transformed and reconstructed values for selected observations from the analysis.

The origin represents the mean of each transformed variable, rather than zero holdings. A household with nothing may lie far from the centre after standardisation. Converting the origin back to source units does not recover arithmetic means, because the transformation is nonlinear. A point near the origin is close to the transformed average on the displayed dimensions.

Angles between axes approximate correlations only under particular conditions. In a correlation representation, the coordinates $L = V_r D_r/\sqrt{n}$ are the correlations of the standardised variables with the retained components, and their inner products approximate the original correlations. The cosine of the displayed angle approximates the correlation well only when both variables are well represented in the plane. The row-principal directions V_r we plot do not generally have this interpretation. Our statements about associations therefore use the actual correlations between transformed variables (Appendix Table 10).

Zero correlation need not mean quantities are unrelated: the relationship may be nonlinear. Poor representation can also produce an apparently right angle. A positive association between enslaved-person holdings and output is not a marginal product of labour: PCA treats columns symmetrically and does not hold other inputs constant. Its geometry alone cannot explain how production was organised.

2.4 Choices that shape the map

The transformation decides which differences matter most. In levels, ten additional cattle add the same squared difference whether a household starts with one or with a hundred. A log-like transformation compresses differences among large holdings and gives distinctions near zero more relative weight. Variation among smaller holdings becomes easier to see, but important differences in scale may be hidden. PCA's least-squares approximation does not require normally distributed variables. Choose a transformation for its effect on the comparison, rather than its ability to produce a symmetric distribution or raise the explained-variance share.

Many households record no vines or no enslaved people, and $\log(x)$ is undefined at zero. Using $\log(x + c)$ with $c > 0$ separates zero and positive holdings according to the choice of c. A small offset pushes nonowners far from owners; a large one draws them together. Judge the offset against the units and distribution of positive entries. The inverse hyperbolic sine is also defined at zero, but its compression depends on the input units. Chen and Roth (2024) establish related unit-dependence problems for treatment effects with outcomes that include zeros. Their theorem does not concern PCA, but motivates examining how a transformation weights ownership differences in an unsupervised description.

Replacing x by ax (say, converting muids to bushels) while keeping c fixed changes $\log(x+c)$ by more than a constant. Converting both, to ax and ac, gives $\log(ax+ac) = \log(a)+\log(x+c)$. Centring removes $\log(a)$, so the representations coincide. Sensitivity checks can compare inconsistent unit changes with correctly converted offsets. Adding a constant without a nonlinear transformation does not alter a correlation.

Correlation PCA gives every nonconstant transformed column a variance of one and hence equal initial weight. Covariance PCA may be more appropriate when variables share commensurate units and variance differences carry substantive meaning. Reporting both shows whether the historical description depends on equalising the spread of each variable.

2.5 Alternatives to PCA and when to use them

Robust PCA reduces the influence of contaminated or unusually influential observations (Hubert, Rousseeuw and Vanden Branden, 2005), although a historically genuine large holding should not be discarded as error. Recent work distinguishes unusual entire records from unusual cells and allows for missing entries (Centofanti, Hubert and Rousseeuw, 2024); archival blanks still require explanation. Sparse PCA builds components from fewer variables, at the cost of a penalty choice and a different approximation (Zou, Hastie and Tibshirani, 2006). Kernel PCA captures nonlinear relationships through a chosen similarity function, but recovering interpretable values of the original variables requires additional work (Schölkopf, Smola and Müller, 1998). Greenacre et al. (2022) survey these extensions.

Other methods start from different kinds of data. Metric multidimensional scaling begins from dissimilarities between observations and suits cases in which a justified distance exists but a conventional quantitative table does not. Classical scaling with Euclidean distances is closely related to

PCA, but its map of scores does not automatically provide calibrated axes for the original measurements. Correspondence analysis is designed for contingency tables and their relative profiles, and multiple correspondence analysis extends related ideas to categorical data. These methods suit occupations or cross-classified counts. Combining physical quantities with nominal categories without first specifying a suitable distance, however, would leave the comparison undefined in substantive terms.

The newer alternatives include t-SNE, which matches neighbourhood similarities, and UMAP, which represents a neighbourhood graph (van der Maaten and Hinton, 2008; McInnes, Healy and Melville, 2018). They can reveal nonlinear organisation but depend on neighbourhood-size and optimisation settings, which should be reported along with random seeds. Gaps in their displays neither measure economic distance nor demonstrate discrete classes. Assessments of local and global structure can yield different verdicts on the same display (Wang et al., 2021). We use neighbourhood comparisons as a diagnostic while retaining PCA for our question about global structure.

Nonlinear methods can also be combined with variable axes. Triana-Martinez et al. (2024) construct local linear representations around a UMAP embedding of agricultural data, so a variable can point in different directions in different parts of the display. Reading those directions requires the additional local model, whereas a calibrated PCA axis applies across the whole display. Log-ratio methods address yet another object: compositions of comparable parts, such as shares of a budget (Aitchison, 1986). How they treat zeros depends on whether a zero reflects rounding, sampling or genuine absence (Lubbe, Filzmoser and Templ, 2021). A method that requires positive entries does not justify treating a household that owned nothing as one that owned a little.

3 Preparing historical household data

3.1 The Cape records and their observed population

The *opgaafrollen* record settler households' demographic characteristics and economic quantities for tax purposes. The Stellenbosch–Drakenstein transcription runs from 1685 to 1844 (Fourie et al., 2024), and related records have supported research on the colonial economy (Fourie, 2013; Fourie and Von Fintel, 2011, 2014; Fourie and Greyling, 2023, 2024; Fourie, 2025). Repeated recording of the same activities allows us to study their joint structure, although fiscal purposes and changing categories constrain comparisons. The series was never designed to measure a constant household production function.

Farming and exchange operated within the institutions of the colonial economy. Viticulture was especially important around Stellenbosch and Drakenstein, while livestock and grain entered households in varying combinations. Enslaved labour formed part of this production system, alongside settler family labour and the labour of Khoekhoe people (Worden, 1985; Fourie and Green, 2015). These relationships were shaped by coercion and by unequal access to land and resources, and the quantities in the returns cannot capture them fully. A count of labourers says nothing about effort,

skill, tasks or working conditions, and an output entry does not show that the output was sold.

The unit of observation is the settler household return. Enslaved people and Khoekhoe labourers appear as entries on those returns, not as households whose own resources can be reconstructed. Inequality in recorded enslaved-person holdings therefore describes how ownership was concentrated among settler households. It cannot measure inequality among enslaved people or describe the welfare of the Cape population as a whole. Other sources may support those questions, but their coverage cannot be inferred from this one.

What the returns record also changed over time. Figure 1 shows in which years each analysis quantity can be constructed under our component and coverage rules. It is not an inventory of every original field: individual livestock entries can exist in years in which the specified total is unavailable. Early returns disaggregate some livestock categories; later forms combine or replace them. A richer form introduced in 1804, during Batavian rule, includes more detailed labour and livestock information (Fourie et al., 2024), and subsequent years change fiscal coverage and the status of enslaved labour. A change of administration alone, however, cannot establish where a field came from or show that two similarly named totals measure the same thing.

The geographical comparison combines district returns dated 1823–1825, which we describe as circa 1825. The preparation records retain the actual year, district, sheet and source row, and the associated agricultural study describes the transcription (Fourie and Greyling, 2023). Because the returns were recorded in different years, the comparison is not a simultaneous colonial census. Differences from the Stellenbosch series can reflect geography, dates, variable coverage and selection as well as production, and these must be specified before any agreement between the two displays can be interpreted.

3.2 Blanks, totals, selection and repeated observations

A blank may mean the household held none, the clerk omitted an entry, or the form never asked for it. Filling all blanks with zeros would erase these distinctions. We record coverage by variable and year in the panel, and by variable, district and year in the cross-section. A field with no numeric entry anywhere in that cell is treated as unavailable. At least one numeric entry shows the field was in use, although enumerators may have filled it in differently. The rule prevents unavailable fields from being filled with zeros; the meaning of other blanks remains conditional.

A return enters the main population only if at least one relevant economic quantity is explicitly recorded. A return blank in every active quantity cannot be verified as a household with nothing. Within otherwise observed returns, we treat remaining blanks in covered fields as zeros and keep indicators that distinguish these inferred zeros from explicit entries. Sensitivity analyses include the wholly blank returns under the stronger zero assumption and examine populations with more complete recording. Restricting the analysis to complete cases is no solution, because it can select on nonownership and on how conscientious an enumerator was. These comparisons show how much results depend on the blank rule; they cannot establish that any one selected population is unbiased.

A total is comparable across years only with the same components and coverage. An unavailable

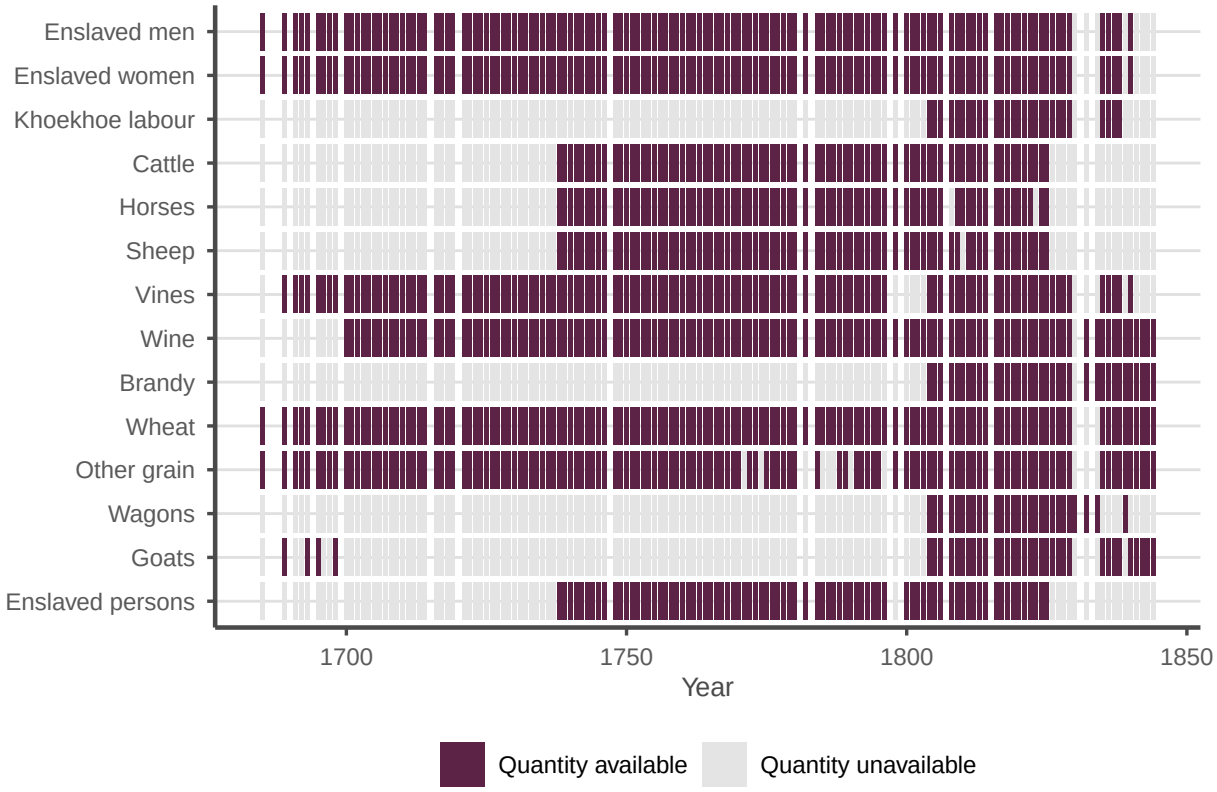


Figure 1: Availability of constructed analysis quantities in the Stellenbosch–Drakenstein transcription, 1685–1844. A dark tile indicates at least one nonmissing constructed value after the stated component and blank rules. A light tile indicates an unavailable quantity; years absent from the source have no tile. This empirical coverage measure does not establish that remaining archival blanks represent zero.

component leaves the sum missing; a blank in a covered component enters as zero only under the stated reporting assumption. We define other grain as barley plus rye throughout, without adding oats in later years. The source dictionary (Appendix Table 4) records components and units, and temporal comparisons retain only years supporting these definitions.

Riding-horse entries are wholly absent in 1808 and 1823, wool-bearing sheep in 1810, and wether-sheep entries after 1825. Total horses and total sheep therefore differ in coverage within the nominal 1804–1829 regime. The main analysis uses fully covered years within 1804–1825. The temporal comparison starts from component definitions available from 1738 and excludes years lacking a required common field. Appendix A lists eligible years and variables.

Requiring at least three positive quantities would select households partly on the combinations the biplot describes, dropping specialised producers and households with few holdings. The main analysis admits a return with a single recorded activity when coverage and recording conditions are met. An explicit zero counts as a recorded entry, distinguishing a household that declared nothing from a wholly blank return. Appendix Table 5 reports sample flow; the three-positive-variable restriction is a sensitivity comparison.

The circa-1825 preparation register identifies observations that a ceiling rule would remove, preserving their values and source locations. We retain these unresolved entries in the baseline and compare estimates without them. Statistical influence cannot distinguish transcription errors from genuine large holdings, and a transformation cannot establish that an entry is correct.

The main analysis contains 20,913 household-year observations and 8,072 distinct household identifiers. Pooling returns gives a household recorded for ten years ten times the weight of one recorded once, and gives a year with many returns more influence than a year with few. The baseline PCA therefore describes observed household-years. Equal-household and equal-year weighting show how these patterns affect the map. Weighting cannot recover missing households or repair mistaken links, and the returns remain an observed fiscal population, not a probability sample of the colony.

Only four eligible returns have explicit values in every underlying active component. These support a PCA of rank at most three, but neither a full-rank 13-variable benchmark nor an informative population comparison. Treating wholly blank returns as zero holdings expands the sample to 28,869 returns and raises two-dimensional fit to 74.0%; it also changes the selected grouping. Requiring three positive variables retains 13,804 returns with 63.2% fit. Higher fit does not make a rule preferable: the rules define different tables to approximate.

4 Reading and testing the Cape biplots

4.1 What the main biplot shows

Figure 2 displays the eligible 1804–1825 returns after applying $\log(1+x)$ to each quantity in source units and standardising it. Each grey dot is a household-year. Coloured lines carry ticks in source units: head of livestock, number of vines, leaguers of wine and brandy, muids of grain. Quantities increase towards the labelled end.

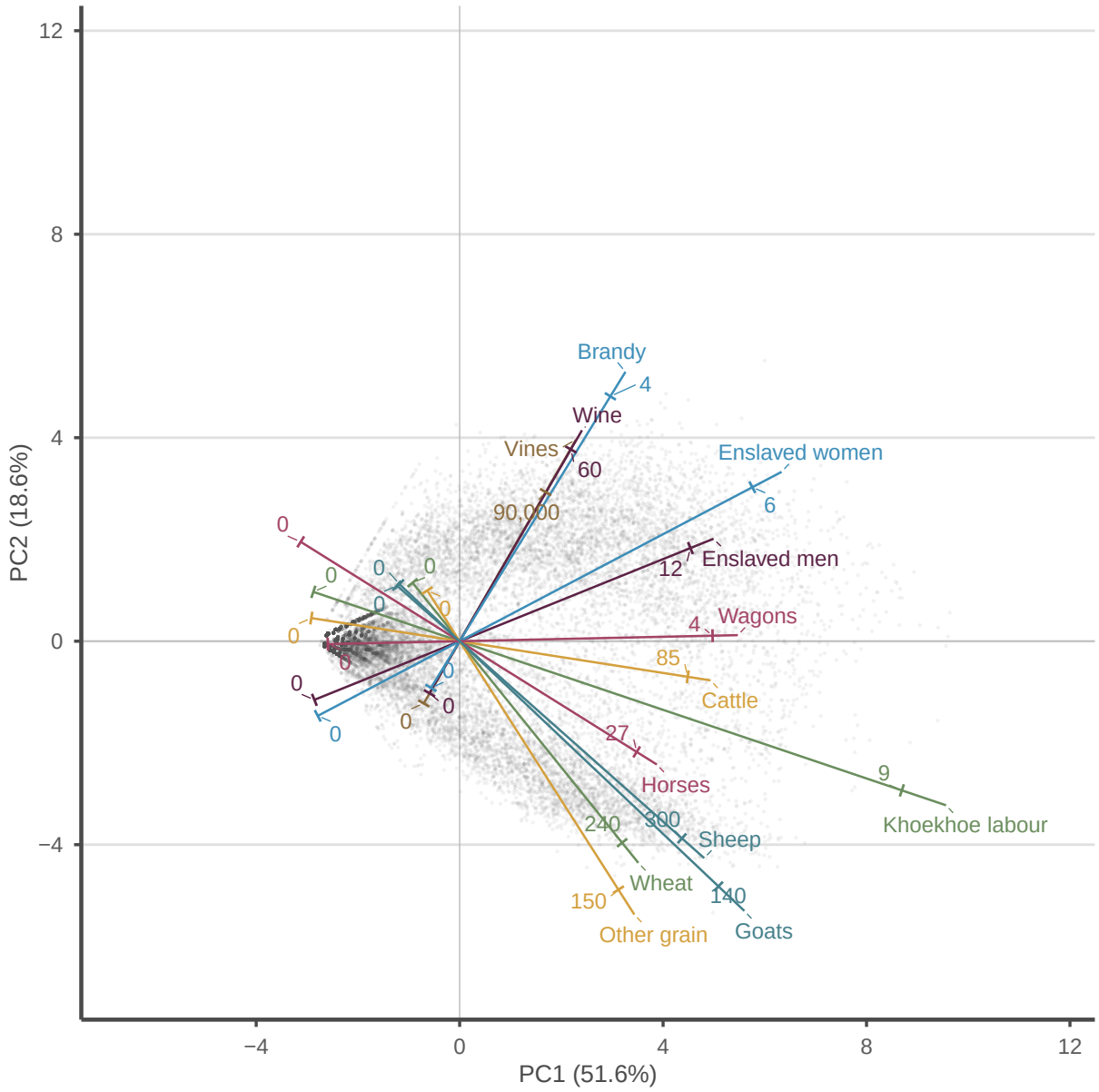


Figure 2: PCA biplot of eligible household-year records within 1804–1825. Row-principal coordinates represent the standardised $\log(1+x)$ data. Variable-axis ticks are quantities in source units; perpendicular projection gives their two-dimensional approximations. The equal aspect ratio preserves displayed angles and distances. The plane accounts for 70.1% of total variation.

Every axis points to the right. The first dimension, accounting for 51.6% of variation, describes scale: the dense mass of points on the left represents households with small or zero entries in most columns; the scattered points on the right are larger farms. Vertically, vines, wine and brandy point upwards, while wheat, other grain, sheep, goats and horses point downwards. The second dimension, accounting for a further 18.6%, describes the production mix. The cloud is continuous, with no obvious gap separating kinds of household (Section 5).

A household whose perpendicular projection onto the wine axis falls at the tick labelled 60 has an approximate output of 60 leaguers. Projecting the same point onto the wheat axis gives its approximate harvest.

The scale interpretation reflects recorded quantities rising together, rather than a priced measure of wealth. The second dimension describes their relative mix after transformation and standardisation. Households could combine vines, grain and livestock; the map shows the balance without assigning them to mutually exclusive sectors.

The common direction of agricultural and enslaved-person holdings is consistent with earlier Cape research on farm scale and coerced labour; the viticultural direction matches accounts of regional wine farming. These associations do not establish that a larger enslaved workforce raised productivity, which requires an output-input relationship and attention to confounding factors. Historical accounts likewise cannot validate an apparent gap or bend in the cloud.

4.2 How much the map leaves out

The plane retains 70.1% of the standardised table's variation. No universal threshold makes this sufficient: high overall fit can coexist with poor representation of the quantity a historical claim depends on.

Three diagnostics answer that question. *Overall fit* is the share of the table's total squared variation that the two-dimensional approximation retains. *Variable predictivity* is the same share for a single column: how much of the variation in, say, wine output the map reproduces. *Household predictivity* is the corresponding share for a single row. These are measures of reconstruction, not forecasts of future or unobserved outcomes. A household very close to the transformed centre has little variation to reproduce, so its predictivity ratio can be unstable; where the denominator is exactly zero, the ratio is undefined and reported as such. Residuals in source units complement these ratios, because a small error on the transformed scale can represent a large absolute difference for a large holding.

Wine and vines are well represented, with two-dimensional predictivities of 92.8% and 83.7% (Table 2 and Appendix Table 9). The plane reproduces only 34.8% of Khoekhoe labour's variation, despite its long, prominent axis in Figure 2. Individual readings also require attention to recorded values and residuals: a dense cloud does not mean every return is precisely reconstructed.

Table 2: Two-dimensional fit of the fitted representations. D denotes eligible returns within 1804–1825; B denotes eligible returns within 1738–1795; bridge pools comparable quantities across these periods; colony denotes district returns dated 1823–1825. Appendix B gives variable diagnostics and reconstruction checks.

Sample	N	PC1 (%)	PC2 (%)	Two PCs (%)	Median row fit (%)
D	20,913	51.6	18.6	70.1	71.7
B	39,850	48.2	18.5	66.7	70.5
bridge	60,522	49.4	19.7	69.1	75.6
colony	8,405	42.0	19.4	61.4	52.1

4.3 How sensitive the map is to our choices

We hold households and variables fixed while comparing levels, $\log(1+x)$, the inverse hyperbolic sine and $\log(x+c)$ for $c = 0.01, 0.1, 1$, recalculating column means and scales for each fit. Appendix Table 12 reports zero shares and positive-quantity distributions so offsets can be judged against actual entries. Unit-conversion checks distinguish holding the offset fixed from converting it with the units.

The transformation matters, but mainly at the extremes (Figure 3 and Appendix C). Two-dimensional fit is 70.1% in the baseline, 69.7% under the inverse hyperbolic sine and 65.3% with a log offset of 0.01, but only 49.7% in standardised levels. To measure whether the plane itself moves, we use principal angles, which compare two planes while ignoring rotations or sign changes within them. The largest principal angle between the baseline plane and the alternative is 0.67 degrees under the inverse hyperbolic sine, so the two planes are almost identical. In levels, it is 26.22 degrees (Appendix Table 11). Compressing large holdings, in other words, substantially changes what the map shows; the precise form of compression does not. Covariance PCA, alternative variable sets and the positive-entry filter are reported separately, because they also change the weighting or the population.

Vines, wine and brandy carry distinct information but give viticulture three contributions to every distance. Dividing labour by sex or status similarly multiplies its influence. The variable-selection comparison assesses these choices: aggregation simplifies, while disaggregation may reveal differences a total conceals. A higher fit alone cannot justify dropping a variable.

We draw 500 resamples of household identifiers with replacement, keeping all returns for each selected household together. After recalculating means and scales and refitting PCA, we compare each plane with the baseline. Variable-direction angles are adjusted for orientation and angular wrapping; components with similar variation can rotate within a stable plane. The largest principal angle from the baseline has a median of 1.28 degrees and a 97.5th percentile of 2.27 degrees (Appendix Table 17). The plane is very stable.

These resamples measure dependence on the observed household histories. They say nothing about missing pages, omitted households, common year shocks or the meaning of blanks. Period and weighting comparisons assess uneven recording over time, but cannot support inference about

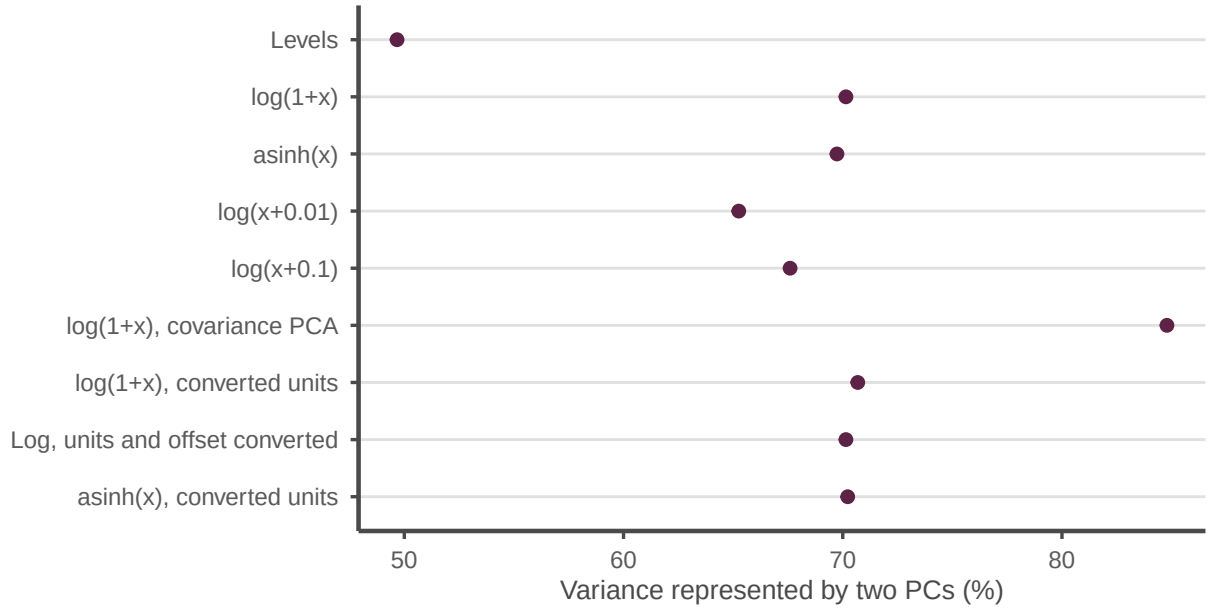


Figure 3: Two-dimensional fit under alternative transformations and variable weighting on the same eligible sample. The covariance specification retains transformed-variable variances; the other specifications standardise them. Table 11 also reports changes in the retained subspace.

all Cape households.

On a fixed sample of 2,000 records, we fit t-SNE with perplexities 15 and 40 and UMAP with neighbourhood sizes 15 and 40 and minimum distance 0.1, each under two random seeds. All eight fits preserve more of each household's 15 nearest neighbours than the two-dimensional PCA projection. They also better approximate transformed quantities using means of neighbours in the display (Appendix Table 26). These diagnostics concern the fitted sample, not prediction for new households. This consistent improvement prompted a kernel PCA comparison with three bandwidths; none improves both diagnostics over ordinary PCA. Nonlinear displays better preserve local resemblance, while PCA provides a global description in calibrated units. Separation in either does not establish household classes.

4.4 Comparing periods in a common frame

Historians often want to know how households changed over time. Fitting a separate biplot for each period would not answer that question, because each fit produces its own axes, and differences between the maps would mix real change with changes in the axes themselves. We therefore fit one transformation, centre, scale and set of loadings to a pooled sample of common variables and project each period's records into that same frame. The common quantities are total enslaved-person holdings, cattle, horses, sheep, vines, wine, wheat, and barley plus rye. Coverage supports observed years within 1738–1795 and 1804–1825, with gaps; it does not support a continuous comparison from 1685 to 1829. The pooled sample contains 60,522 records and 18,787 household

identifiers, and its two-dimensional fit is 69.1%. The two panels of Figure 4 share identical axes, limits and aspect ratios, so a given location means the same combination of holdings in both periods. The approach follows an established dynamic-biplot logic (Ciccarelli and Proietti, 2013; Ganey and Nienkemper-Swanepoel, 2026).

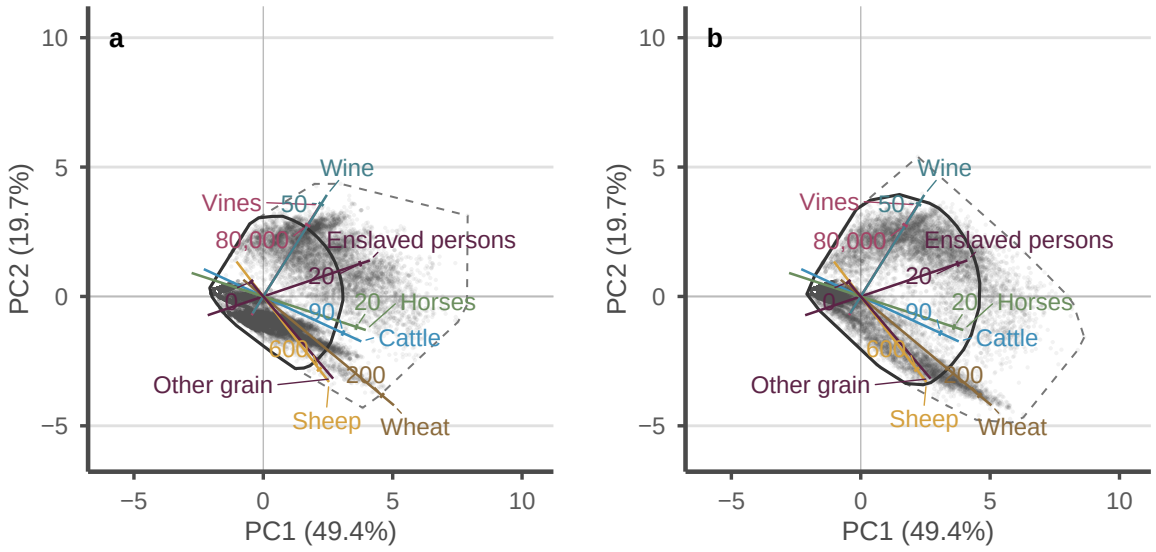


Figure 4: Period distributions in a pooled PCA frame: eligible observations within 1738–1795 in panel (a), left, and 1804–1825 in panel (b), right. The solid boundary encloses the 90% of records closest to the period centroid in the two-dimensional frame; the dashed boundary is the full convex hull. These are descriptive regions, not confidence intervals. Common variables and eligible years are documented in Appendix A. Transformation, centre, scale, loadings, limits and aspect ratio are fixed.

The later distribution extends further along the second dimension, towards both viticulture and grain and livestock. A fixed frame removes movement caused by refitting axes, but cannot establish comparable measurement. Changing livestock components or enumeration practices, entry into the source and household links can all shift the distribution. Household fixed effects cannot distinguish a recording change affecting everyone from an economic change affecting everyone; nor can persistence after an administrative change. One way to investigate would combine households linked across the new form with independent evidence spanning both periods on the enumeration of several active quantities. Any separation would still depend on explicit assumptions about measurement and entry into the source.

A pooled frame may represent larger periods better than smaller ones; period-specific fit (Appendix Table 19) shows how much of each distribution it captures. Equal-year and equal-household weighting test the influence of crowded years and long histories. The choice should follow the

population the researcher intends to describe.

The full convex hull, the smallest polygon enclosing every point, is sensitive to extremes and tends to grow as observations are added. A central region encloses a specified share of the distribution. We report both and compare equal-size subsamples across periods to check for sample-size effects (Appendix Table 20), while estimating on the full eligible data. Subsampling tests whether envelope changes could arise from adding points to the same distribution; it cannot balance historical coverage.

The panels compare period distributions rather than following individual households. Studying change within households would require an explicitly defined linked panel.

5 Comparing household profiles and inequality

5.1 What household records add to an aggregate profile

Most historical biplots, like most historical statistics, are drawn from aggregates such as provinces or districts. Household records make a different kind of analysis possible. District means can describe an aggregate combination of output and holdings while concealing how it is distributed across households. Suppose two districts each record ten cattle per household on average. In one, every household might own ten; in the other, half might own none and half twenty. The two districts look identical in a table of means but differ in inequality. With two assets, identical distributions of each asset can also conceal different combinations within households. Only the household records can resolve either difference.

The circa-1825 cross-section contains 8,405 eligible household returns, with two-dimensional fit of 61.4% (Appendix Figure 6). In the full standardised, transformed space, 86.5% of variation lies within districts (Appendix Table 27). District means would remove most of this variation. The historical importance of differences between districts nevertheless depends on the question, rather than their share of variation alone.

District averages and household distributions also answer different graphical questions. Averaging household scores within a fixed map differs from fitting a new PCA to a small table of district means. Because of the transformation, the average of transformed quantities is not the transformation of the arithmetic average. The variance accounting in Appendix D uses the full transformed space and splits squared deviations into within-district and between-district parts. The household biplot in Appendix A shows the same records, but its two-dimensional plane should not be mistaken for a complete display of all within-district differences.

Ciccarelli and Proietti (2013) use provincial observations to reveal differences below the regional level. Our household observations reveal differences within local aggregates. Each level suits different questions: provincial data can describe industrial specialisation; household data can show how holdings were combined and distributed.

5.2 A descriptive grouping and asset-specific inequality

We group households to ask how each asset's inequality divides between and within groups. The k-means procedure seeks to minimise squared distances from households to their group centres. We apply it to the full standardised, transformed table, choosing from $k = 2, \ldots, 8$. Sensitivity checks repeat the procedure on the first two principal components, then three, up to full rank. All components together reproduce the standardised distances exactly; we do not assume discarded variation is noise.

The number of groups is chosen by the average silhouette width (Rousseeuw, 1987), which compares how close each household is to members of its own group with how close it is to the nearest other group. The silhouette is a diagnostic within a given distance representation. A larger silhouette in a reduced space does not show that the reduced space is economically more meaningful. Selection uses 1,500 sampled records, with silhouettes evaluated on 750 and ten random starts; the selected centres then initialise a refit on every eligible observation. Where two values of k tie within a declared numerical tolerance, we choose the smaller. Comparisons across retained dimensions examine whether households stay together and how the groups differ in size and asset profile. Household resampling repeats the entire procedure. Even stable groups need not correspond to natural classes: k-means will partition a continuous distribution whenever it is asked to.

The grouping mainly separates households with lower recorded holdings from those with higher ones. The higher-holdings group, about a third of the records, holds several times, and often more than ten times, as much of each recorded quantity (Appendix Table 23). The number of groups, however, is not stable. The silhouette advantage of two groups over three is small (Appendix Table 21). Using two to five components selects three groups, while using six or more selects two. The adjusted Rand index, which measures agreement between two groupings on a scale where one means identical, is 0.762 between the two-component grouping and the full-space reference. More decisively, 199 of 200 household resamples select three groups, whereas the original full-space selection gives two; mean agreement with the reference is 0.764. The high-holdings group is especially prone to being divided further. We therefore retain the reference grouping as a declared accounting device and do not interpret it as a settled typology of Cape households.

DBSCAN groups densely connected points and can leave isolated observations unassigned; HDBSCAN considers a hierarchy of density levels (Hahsler, Piekenbrock and Doran, 2019; Campello, Moulavi and Sander, 2013). Their density and minimum-size choices require historical judgement. Soft k-means allows graded membership through, for example, distance-based weights with entropy regularisation; mixture models supply membership probabilities under distributional assumptions (MacKay, 2003; Fraley and Raftery, 2002). We use a compact partition for a transparent decomposition, while testing its limitations.

To measure inequality, we use the Theil T index, calculated separately for each asset on its untransformed quantity. The index is zero when every household holds the same amount and rises as holdings become more concentrated. Its useful property is that it splits exactly into two parts: inequality within groups and inequality between group averages. For this calculation, enslaved-person

holdings include men, women, sons and daughters; the main PCA instead uses separate columns for men and women. Households with zero holdings remain in the population and contribute zero through the limit $0 \log 0 = 0$. We do not add one before calculating the index, because that would change the distribution being measured. If $\mu > 0$ is the overall mean, the decomposition is

$$T = \frac{1}{n} \sum_i \frac{x_i}{\mu} \log \frac{x_i}{\mu} = \sum_g s_g T_g + \sum_g s_g \log \frac{\mu_g}{\mu}, \qquad s_g = \frac{n_g \mu_g}{n \mu}. \tag{2}$$

The first sum on the right is within-group inequality: each group's own Theil index T_g, weighted by the group's share s_g of the asset. The second sum is between-group inequality, which would remain if every household held its group's average (Theil, 1967; Bourguignon, 1979). A group with a zero mean holds none of the asset and contributes zero to both sums. If no household holds the asset, there is nothing to normalise, and the index is reported as undefined. Appendix D gives the implementation and additivity checks.

Table 3: Inequality in recorded quantities and its within- and between-profile decomposition. Theil T retains zero holdings and uses unshifted quantities. Profiles are estimated from the same multivariate records; the decomposition is descriptive.

Asset	N	T	Within	Between	Within (%)
Enslaved persons	20,913	0.797	0.481	0.316	60.4
Cattle	20,913	1.041	0.476	0.565	45.7
Horses	20,913	0.865	0.456	0.409	52.7
Vines	20,913	1.431	0.845	0.585	59.1
Wine	20,913	1.679	1.074	0.605	64.0

The index describes observed household-years, with inferred zeros in covered fields. Assets being decomposed also helped form the groups, so the between-group component partly reflects the partitioning criterion. It cannot measure an effect of group membership or independently test whether the groups explain concentration.

Table 3 and Figure 5 show the results. Under the reference grouping, the within-group shares are 60.4% for enslaved-person holdings, 45.7% for cattle, 52.7% for horses, 59.1% for vines and 64.0% for wine. For vines and wine, most inequality lies among households within the same broad group. When the asset being measured is left out and the grouping is re-estimated, those shares fall to 35.8% and 43.7%, and most inequality lies between groups. Excluding one asset does not make the grouping independent of that activity, because related quantities remain in the clustering. Grouping by year or by whether a household recorded any enslaved people produces yet other allocations (Appendix Table 25). On a fixed sample, each asset's total inequality stays the same; only its allocation between and within groups changes with the partition.

The allocation of concentration is conditional on the grouping. Explaining it through coercion, entry costs in viticulture or the household life cycle requires evidence beyond a partition estimated from the same quantities and shown to be unstable.

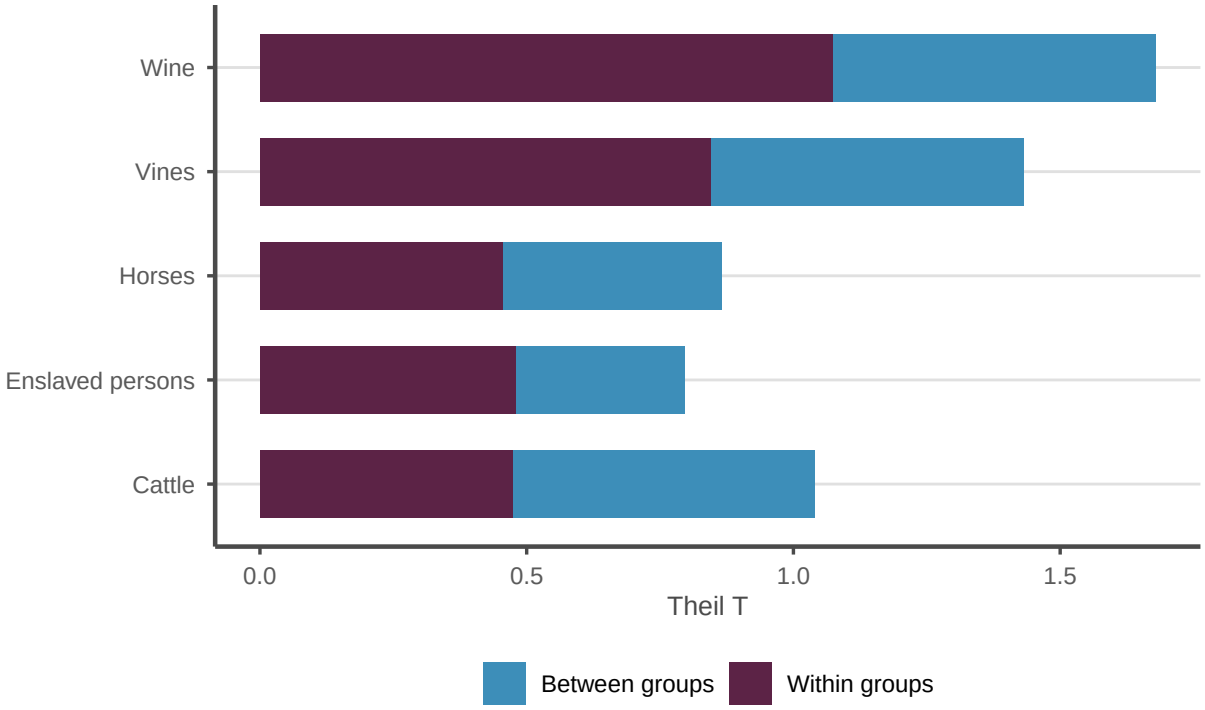


Figure 5: Within- and between-profile components of Theil inequality in unshifted recorded quantities. Zero holdings remain in the population. Components sum to total inequality for each asset; profiles are estimated from the same multivariate data, so this is descriptive accounting conditional on the partition.

6 Conclusion

We began with a table that is almost impossible to read: more than a hundred thousand Cape tax returns, each listing a household's labourers, livestock, vines and harvests. A PCA biplot turns it into a map. Two dimensions retain 70.1% of the standardised, transformed variation, describing scale and the balance between viticulture and grain and livestock. The map barely changes under household resampling or between the log and inverse-hyperbolic-sine transformations. The selected number of household groups changes under resampling. Stability of the visual approximation and of a classification require separate tests.

Historians have long reconstructed household economies from inventories and tax returns (Overton et al., 2004; Ogilvie, Küpker and Maegraith, 2012). At the Cape, the same returns have measured wheat productivity, the relationship between enslaved labour and wine output, and the productivity of indigenous labour (Fourie and Greyling, 2023, 2024; Fourie and Green, 2015). Each study examines a few quantities at a time. The biplot shows every recorded holding jointly and reports how much the display leaves out. In the Cape returns, it describes the settler economy as a continuum of scale crossed by a contrast between viticulture and grain and livestock. The instability of the household groups cautions against reading that continuum as a set of discrete farm types. The procedure depends on source criticism before calculation and checks afterwards. Coverage and accounting rules define the table; transformation and weighting determine which differences matter; fit must be assessed for the variables and households relevant to a claim, with attention to repeated records. Table 1 sets out these steps for other sources.

Research on preindustrial inequality usually measures concentration one asset at a time (Milanović, Lindert and Williamson, 2011; Alfani, 2021; Fourie and Von Fintel, 2014). Household records show how holdings of different assets were combined and where their concentration lay. In the colony-wide cross-section, 86.5% of the variation lies within districts; district averages would discard it. The asset-specific Theil decomposition then asks whether each asset's inequality lay between broad household profiles or within them. The answer depends on the partition: excluding vines or wine from the grouping reverses that asset's within- and between-group ordering. That sensitivity is itself a result. It shows how much a decomposition of inequality depends on the typology imposed on the data. For the biplot literature, the paper moves the method from aggregates to households. Ciccarelli and Proietti (2013) trace the specialisation of Italian provinces. Household observations reveal the distributions behind such aggregate profiles, but they bring measurement problems that provincial tables conceal: unrecorded fields, zero holdings, transformation units and repeated records of the same household.

These results concern settler returns under the stated blank rule. They cannot describe inequality among enslaved people or across the whole colonial population, and measuring wealth would require valuations. The biplot and the decomposition describe; neither identifies what caused the patterns they display. How much Cape farming changed across the 1804 change in the form remains unresolved: changes in enumeration or in the recorded population could produce similar movement, and a fixed frame cannot separate the two. Independent evidence spanning both peri-

ods on how several quantities were enumerated could help, under stated assumptions about entry into the source (Section 4.4). Other sources, from probate inventories to agricultural censuses, require their own assessment of coverage, accounting categories and dependence between records. A source without repeated identifiers may need a different resampling unit.

These limits concern the sources more than the method, and the sources are changing. Transcription projects are turning tax lists, inventories and censuses into household panels that cover whole colonies and link households across generations. A common frame can already place several periods on one map. Animated biplots can follow the distribution year by year, showing when viticulture spread and which households moved with it (Ganey and Nienkemper-Swanepoel, 2026). With linked households, the same frame can trace individual trajectories through the map across the life cycle and across generations. Combined with the checks set out here, the biplot allows historians to see how whole economies were organised, household by household. Summaries built one column at a time could not show this.

Data, code and computational assistance

Tables and numerical statements are generated from saved analysis objects. Numerical checks document consistency between the analysis and its reported results, and the build records the software environment and random seeds. Access to the underlying transcriptions requires permission from their custodians.

The authors used generative AI tools at all stages of this research, including data management, coding, hypothesis testing, and drafting and reviewing the manuscript. These tools were Claude Code (Anthropic; Claude Opus and Fable models) and OpenAI Codex (Sol and Astra models). Refine.ink provided an additional AI-assisted review of the manuscript. The authors reviewed and verified all AI-generated output and take full responsibility for the content of the final version.

Data availability

The code, aggregate results and documentation that support the findings of this study are openly available at `https://github.com/johanfourieza/research/tree/main/2026/biplots`. Access to the underlying transcriptions requires permission from their custodians.

References

Aitchison, John. 1986. *The Statistical Analysis of Compositional Data.* London:Chapman and Hall.

Alfani, Guido. 2021. "Economic Inequality in Preindustrial Times: Europe and Beyond." *Journal of Economic Literature*, 59(1): 3–44.

Bourguignon, François. 1979. "Decomposable Income Inequality Measures." *Econometrica*, 47(4): 901–920.

Campello, Ricardo J. G. B., Davoud Moulavi, and Jörg Sander. 2013. "Density-Based Clustering Based on Hierarchical Density Estimates." 160–172. Springer.

Centofanti, Fabio, Mia Hubert, and Peter J. Rousseeuw. 2024. "Robust Principal Components by Casewise and Cellwise Weighting."

Chen, Jiafeng, and Jonathan Roth. 2024. "Logs with Zeros? Some Problems and Solutions." *Quarterly Journal of Economics*, 139(2): 891–936.

Ciccarelli, Carlo, and Tommaso Proietti. 2013. "Patterns of industrial specialisation in post-Unification Italy." *Scandinavian Economic History Review*, 61(3): 259–286.

Fourie, Johan. 2013. "The Remarkable Wealth of the Dutch Cape Colony: Measurements from Eighteenth-Century Probate Inventories." *Economic History Review*, 66(2): 419–448.

Fourie, Johan. 2025. "Inequality in the Cape Colony, 1685–1844." *South African Journal of Science*, 121(11/12): Art. 20532.

Fourie, Johan, and Dieter Von Fintel. 2011. "A History with Evidence: Income Inequality in the Dutch Cape Colony." *Economic History of Developing Regions*, 26(1): 16–48.

Fourie, Johan, and Dieter Von Fintel. 2014. "The Dynamics of Inequality in a Newly Settled, Pre-Industrial Society: The Case of the Cape Colony." *Cliometrica*, 8(3): 289–318.

Fourie, Johan, and Erik Green. 2015. "The Missing People: Accounting for the Productivity of Indigenous Populations in Cape Colonial History." *Journal of African History*, 56(2): 195–215.

Fourie, Johan, and Jan Greyling. 2023. "Wheat Productivity in the Cape Colony in 1825: Evidence from Newly Transcribed Tax Censuses." *Agrekon*, 62(1): 98–115.

Fourie, Johan, and Jan Greyling. 2024. "Slave Labor Productivity and Wine Output: Stellenbosch, 1680–1828." *Journal of Wine Economics*, 18(3): 191–204.

Fourie, Johan, Erik Green, Christiaan Burger, Chris De Wit, Kate Ekama, Jan Greyling, Hans Heese, Jan-Hendrik Pretorius, Robert Ross, and Dieter Von Fintel.

2024. "'Waar is beter dorp in Zuid Africa dan Stellenbos?': What the Stellenbosch–Drakenstein Tax Censuses Reveal." *South African Historical Journal*, 76(4): 420–446.

Fraley, Chris, and Adrian E. Raftery. 2002. "Model-Based Clustering, Discriminant Analysis, and Density Estimation." *Journal of the American Statistical Association*, 97(458): 611–631.

Gabriel, K. Ruben. 1971. "The Biplot Graphic Display of Matrices with Application to Principal Component Analysis." *Biometrika*, 58(3): 453–467.

Ganey, Raeesa, and Johané Nienkemper-Swanepoel. 2026. "moveEZ: An R Package for Animated Biplots."

Gower, John C., and David J. Hand. 1996. *Biplots.* London:Chapman and Hall.

Gower, John C., Niël J. Le Roux, and Sugnet Gardner-Lubbe. 2015. "Biplots: quantitative data." *WIREs Computational Statistics*, 7(1): 42–62.

Gower, John C., Sugnet G. Lubbe, and Niël J. Le Roux. 2011. *Understanding Biplots.* Chichester:Wiley.

Greenacre, Michael. 2010. *Biplots in Practice.* Bilbao:Fundación BBVA.

Greenacre, Michael, Patrick J. F. Groenen, Trevor Hastie, Alfonso Iodice D'Enza, Angelos Markos, and Elena Tuzhilina. 2022. "Principal component analysis." *Nature Reviews Methods Primers*, 2.

Hahsler, Michael, Matthew Piekenbrock, and Derek Doran. 2019. "dbscan: Fast Density-Based Clustering with R." *Journal of Statistical Software*, 91(1): 1–30.

Hubert, Mia, Peter J. Rousseeuw, and Karlien Vanden Branden. 2005. "ROBPCA: A New Approach to Robust Principal Component Analysis." *Technometrics*, 47(1): 64–79.

Le Roux, Niël J., and Sugnet Gardner-Lubbe. 2025. "A two-group canonical variate analysis biplot for an optimal display of both means and cases." *Advances in Data Analysis and Classification*, 19: 721–748.

Lubbe, Sugnet, Peter Filzmoser, and Matthias Templ. 2021. "Comparison of zero replacement strategies for compositional data with large numbers of zeros." *Chemometrics and Intelligent Laboratory Systems*, 210.

MacKay, David J. C. 2003. *Information Theory, Inference, and Learning Algorithms.* Cambridge:Cambridge University Press.

McInnes, Leland, John Healy, and James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction."

Milanović, Branko, Peter H. Lindert, and Jeffrey G. Williamson. 2011. "Pre-Industrial Inequality." *Economic Journal*, 121(551): 255–272.

Ogilvie, Sheilagh, Markus Küpker, and Janine Maegraith. 2012. "Household Debt in Early Modern Germany: Evidence from Personal Inventories." *Journal of Economic History*, 72(1): 134–167.

Overton, Mark, Jane Whittle, Darron Dean, and Andrew Hann. 2004. *Production and Consumption in English Households, 1600–1750.* London:Routledge.

Rousseeuw, Peter J. 1987. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics*, 20: 53–65.

Schölkopf, Bernhard, Alexander J. Smola, and Klaus-Robert Müller. 1998. "Nonlinear Component Analysis as a Kernel Eigenvalue Problem." *Neural Computation*, 10(5): 1299–1319.

Theil, Henri. 1967. *Economics and Information Theory.* Amsterdam:North-Holland.

Triana-Martinez, Jenniffer Carolina, Andrés Marino Álvarez-Meza, Julian Gil-González, Tom De Swaef, and Jose A. Fernandez-Gallego. 2024. "Crop Water Status Analysis from Complex Agricultural Data Using UMAP-Based Local Biplot." *Remote Sensing*, 16(15).

van der Maaten, Laurens, and Geoffrey Hinton. 2008. "Visualizing Data using t-SNE." *Journal of Machine Learning Research*, 9: 2579–2605.

Wang, Yingfan, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. 2021. "Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization." *Journal of Machine Learning Research*, 22(201): 1–73.

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

Zou, Hui, Trevor Hastie, and Robert Tibshirani. 2006. "Sparse Principal Component Analysis." *Journal of Computational and Graphical Statistics*, 15(2): 265–286.

A Source definitions and analysis populations

The source register distinguishes explicit values, inferred zeros in fields the form requested, and unavailable fields. A coverage cell is a year in the panel and a district-year in the circa-1825 cross-section; it need not correspond to one archival form. Remaining blanks are therefore subject to the stated interpretation. A derived total is unavailable when a required component field is unavailable in that cell. Blanks within covered component fields enter the sum as zero only under the conditional reporting rule. Our preparation records link each quantity to its source and record its units and component definitions, so that new archival evidence can inform these decisions.

Eligible years in the earlier comparison are 1738–1795 except 1747, 1771, 1774, 1781–1783, 1785–1787 and 1790. The main later-period analysis uses 1804–1825 except 1807, 1808, 1810, 1815 and 1823. The pooled common-variable comparison uses these covered periods with its own record-level eligibility rule. The preparation records document the variable-by-year coverage decisions. An exclusion under the numerical coverage rule does not independently establish that an archival form omitted the field: a field that was entirely zero and always left blank would also fail this conservative criterion.

Table 4: Definitions of active and derived quantities in this analysis. Units refer to source entries; source-specific assumptions are documented in the preparation records.

Variable	Definition	Unit
Enslaved men	Recorded adult male enslaved persons	count
Enslaved women	Recorded adult female enslaved persons	count
Khoekhoe labour	Recorded Khoekhoe labourers	count
Cattle	Cattle: work and breeding	count
Horses	Horses: riding and breeding	count
Sheep	Sheep: breeding, wethers and wool-bearing	count
Vines	Vines	count
Wine	Wine output	leaguers
Brandy	Brandy output	leaguers
Wheat	Wheat harvest	muids
Other grain	Barley and rye harvest	muids
Wagons	Wagons	count
Goats	Goats	count
Enslaved persons	Enslaved persons: men, women, sons and daughters	count

The main population includes eligible low-asset and specialised records. A requirement of three positive active variables defines a separate sensitivity population. The strict observed-entry population is also reported separately, because removing records with inferred zeros can select on nonownership and recording practice. Neither exercise recovers a representative census of all residents. The circa-1825 comparison preserves actual dates and identifies unresolved extreme records; estimates with and without those observations show the consequences of including them.

Table 5: Flow from source records to analysis populations, with household-year observations and distinct identifiers distinguished.

sample	stage	n	households
Panel source	all source rows	142,293	48,380
Regime D	nominal period 1804-1829	41,969	15,644
Regime D	comparable columns before wholly blank returns excluded	28,869	11,747
Regime D	comparable complete variables; no asset-count filter	20,913	8,072
Regime D	old three-positive-variable rule within revised sample	13,804	5,125
Regime B	comparable complete variables	39,850	11,350
Bridge	comparable complete variables 1738-1825	60,522	18,787
Colony	accepted head markers	9,570	9,570
Colony	comparable columns before wholly blank returns excluded	9,563	9,563
Colony	complete nonnegative active quantities; no ceiling exclusions	8,405	8,405

Table 6: Completeness of explicit recording across the raw economic components underlying the main variables.

Primary records	Source components	Complete records	Active quantities
20,913	21	4	13

Table 7: Coverage and eligible temporal comparisons under the common-variable construction.

Sample	Variables	Records	First	Last	Observed years
D	13	20,913	1,804	1,825	17
B	8	39,850	1,738	1,795	48
bridge	8	60,522	1,738	1,825	65
colony	10	8,405	1,823	1,825	3

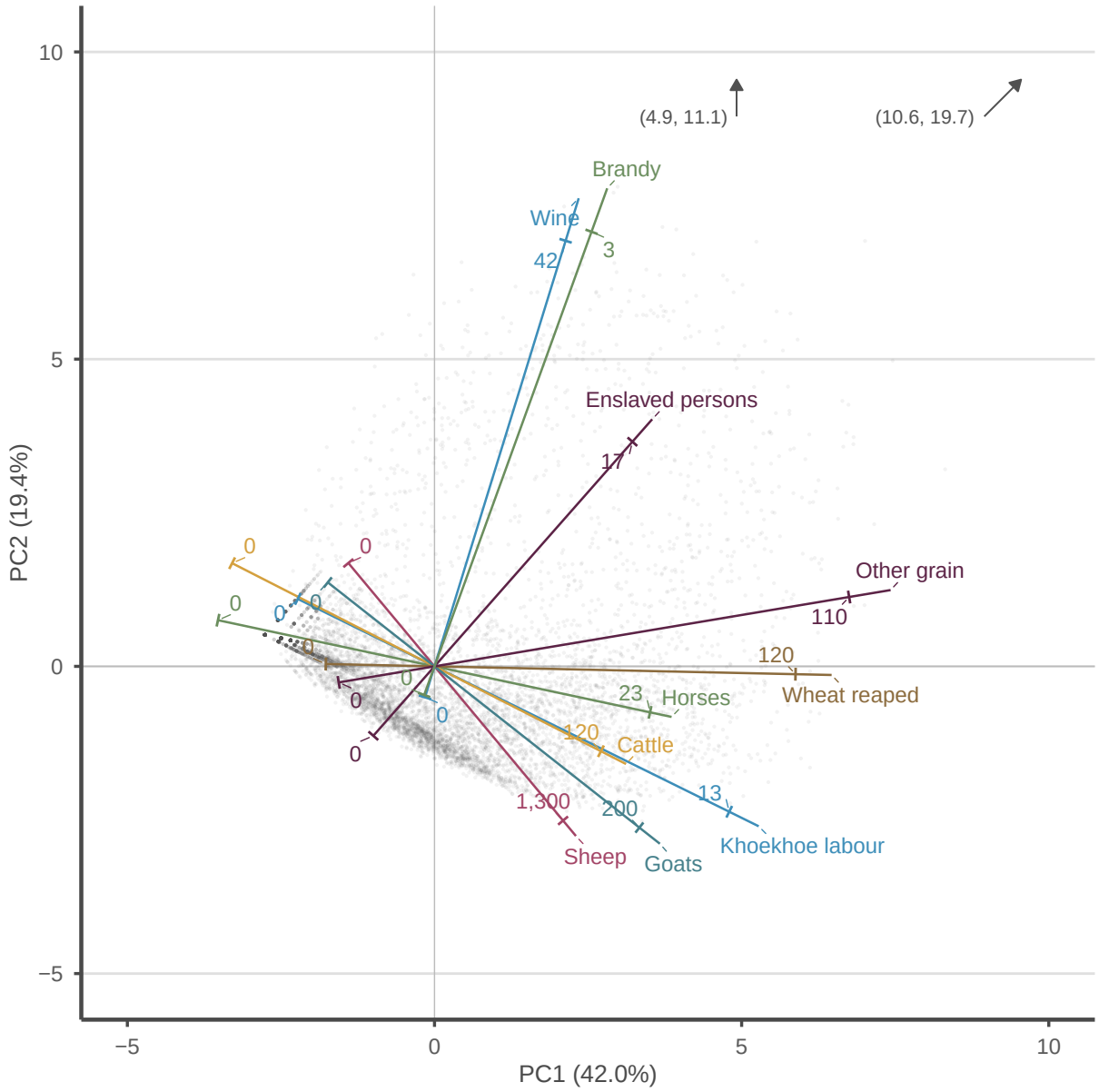


Figure 6: PCA representation of eligible district returns dated 1823–1825. The cross-section uses its own documented variable set and standardisation. Its 61.4% two-dimensional fit does not make its coordinates directly comparable with the main within-period PCA. Two returns lie above the displayed frame; arrows at its edge mark their direction and give their coordinates.

B Biplot geometry and reconstruction

For an orthogonal rank-r projection, the residual $E = Z - \widehat{Z}_r$ is orthogonal to the fitted matrix in the relevant row and column decompositions. Overall fit is $\sum_{h \le r} d_h^2 / \sum_h d_h^2$. Variable predictivity is $q_j = \sum_i \widehat{z}_{ij}^2 / \sum_i z_{ij}^2$, and observation predictivity is $q_i = \sum_j \widehat{z}_{ij}^2 / \sum_j z_{ij}^2$. Both ratios are undefined at a zero denominator. The code verifies full-rank reconstruction and the inner-product calibration, and confirms that the squared norm of the rank-r reconstruction equals the sum of the retained squared singular values. These identities are checked independently of the plotting software.

The calibration exercise selects anonymous example rows at declared points of the observed score distribution. It reports the source quantity, transformed value and two-dimensional reconstruction. The rows are chosen for instruction and are not claimed to be historically representative. Inverse transformation can yield an approximate value outside the admissible range, including a negative approximation for a quantity that cannot be negative. Such a result signals projection error; it should not be silently truncated before fit is assessed, nor presented as a recorded holding.

Table 8: Recorded quantities and reconstructions for the worked calibration example.

Variable	Recorded	z	Predicted z	Predicted quantity
Enslaved men	0	-1.042	-0.741	0.331
Enslaved women	0	-0.915	-0.615	0.232
Khoekhoe labour	3	1.017	-0.490	0.198
Cattle	0	-0.992	-0.742	0.556
Horses	1	-0.759	-0.663	1.236
Sheep	0	-0.620	-0.581	0.081
Vines	0	-0.691	-0.654	0.200
Wine	0	-0.610	-0.690	-0.109
Brandy	0	-0.525	-0.675	-0.071
Wheat	0	-0.631	-0.533	0.211
Other grain	0	-0.527	-0.459	0.117
Wagons	0	-0.885	-0.777	0.070
Goats	0	-0.554	-0.506	0.083

C Sensitivity checks for the household and temporal displays

The transformation grid fixes the eligible rows and columns. Each specification uses the means and scales of its own transformed data to estimate the decomposition. Subspace comparisons use principal angles or equivalent rotation-invariant summaries. The matched-offset unit conversion is an exact centring identity and is checked numerically. Specifications that also change variable weighting or the eligible population, including covariance PCA and changes in the active variable set, are labelled separately.

The common temporal construction uses the same transformations and variable definitions in both panels. Fixed-frame period fit divides reconstructed squared distances by squared distances from the pooled transformed centre, using pooled scales. A separately fitted period PCA instead

Table 9: Variable predictivity in the main two-dimensional representation.

Variable	PC1 (%)	PC2 (%)	Two PCs (%)	RMSE
Enslaved men	66.0	3.9	69.8	0.549
Enslaved women	44.3	4.4	48.7	0.716
Khoekhoe labour	33.4	1.4	34.8	0.807
Cattle	74.0	0.6	74.6	0.504
Horses	65.1	9.2	74.3	0.507
Sheep	53.1	15.0	68.1	0.565
Vines	40.2	43.5	83.7	0.404
Wine	44.8	47.9	92.8	0.269
Brandy	43.7	41.5	85.2	0.385
Wheat	48.4	26.8	75.1	0.499
Other grain	38.6	34.0	72.6	0.523
Wagons	78.2	0.0	78.2	0.467
Goats	40.8	13.2	54.0	0.678

Table 10: Selected correlations among the active transformed quantities. These descriptive correlations are calculated from the analysis matrix and do not use independent-row significance tests.

First quantity	Second quantity	Correlation
Enslaved men	Enslaved women	0.754
Enslaved men	Khoekhoe labour	0.390
Enslaved men	Cattle	0.641
Cattle	Horses	0.719
Cattle	Sheep	0.629
Enslaved men	Wine	0.598
Cattle	Wine	0.503
Vines	Wine	0.915
Vines	Brandy	0.804
Wine	Brandy	0.932
Wheat	Other grain	0.789
Sheep	Goats	0.632

Table 11: Transformation and specification sensitivity. Fixed-sample transformation contrasts are distinguished from changes in weighting, variables or population.

Specification	N	Two PCs (%)	Maximum angle
Levels	20,913	49.7	26.217
log(1+x)	20,913	70.1	0
asinh(x)	20,913	69.7	0.674
log(x+0.01)	20,913	65.3	4.887
log(x+0.1)	20,913	67.6	3.185
log(1+x), covariance PCA	20,913	84.8	44.784
log(1+x), converted units	20,913	70.7	0.700
Log, units and offset converted	20,913	70.1	0
asinh(x), converted units	20,913	70.2	0.812

Table 12: Zero shares and positive source quantities in the main analysis population. Minimum, median and 90th percentile refer to positive entries; zero shares include the stated conditional treatment of covered partial blanks.

Variable	Zero (%)	Minimum	Median	P90	Maximum
Enslaved men	37.4	1	4	12	86
Enslaved women	46.8	1	2	6	56
Khoekhoe labour	59.2	1	3	9	970
Cattle	46.7	1	26	85	983
Horses	15.9	1	5	27	520
Sheep	68.8	1	60	300	30,400
Vines	67.1	1	40,000	90,000	700,000
Wine	69.3	0.125	20	60	3,803
Brandy	71.6	0.062	1.500	4	87
Wheat	69.0	0.250	55	240	1,766
Other grain	75.9	0.750	33	153	50,000
Wagons	50.2	1	2	4	63
Goats	74.3	1	42	138.400	18,024

Table 13: Sensitivity to the eligible population and influential observations.

Specification	N	Two PCs (%)	Maximum angle
At least three positive quantities	13,804	63.2	3.717
Exclude largest 1% of distances	20,703	69.3	0.547
at least one observed component	20,913	70.1	0
reporting intensity upper half	10,470	60.4	10.397
reporting intensity upper quartile	5,596	58.1	21.141
Exclude flagged source entries	20,899	70.2	0.086

Table 14: Sensitivity to treating wholly blank economic returns as zero holdings.

Sample	Primary N	Broad N	Primary fit (%)	Broad fit (%)
D	20,913	28,869	70.1	74.0
B	39,850	46,321	66.7	68.3
bridge	60,522	75,190	69.1	71.7
colony	8,405	9,563	61.4	64.3

Table 15: Changes in the descriptive partition when wholly blank economic returns are included under a zero assumption.

Primary N	Broad N	Primary groups	Broad groups	Common-record ARI
20,913	28,869	2	3	0.717

Table 16: Sensitivity to equal observation, household and period weights.

Specification	N	Two PCs (%)	Maximum angle
Equal household weights	20,913	70.6	5.169
Equal year weights	20,913	70.0	0.762

Table 17: Household-bootstrap stability of the retained principal subspace. Principal angles are in degrees; eigenvalue ratios are dimensionless.

Statistic	Median	2.5 percentile	97.5 percentile
max angle	1.283	0.701	2.275
rms angle	0.946	0.563	1.636
eigen ratio 12	2.782	2.639	2.928
eigen ratio 23	3.118	2.974	3.250

Table 18: Orientation-adjusted variable-direction angles in degrees under household resampling. The intervals summarise the 2.5th and 97.5th percentiles and are not tests of independence.

Variable	Reference	Aligned lower	Aligned upper
Enslaved men	21.971	19.478	24.396
Enslaved women	27.739	24.361	31.103
Khoekhoe labour	-18.637	-23.954	-13.374
Cattle	-8.866	-10.576	-6.940
Horses	-32.013	-33.144	-30.821
Sheep	-41.560	-43.177	-39.936
Vines	60.010	59.239	60.650
Wine	59.881	59.343	60.398
Brandy	58.380	57.661	59.040
Wheat	-51.120	-52.317	-49.795
Other grain	-57.414	-58.505	-56.339
Wagons	1.249	-0.515	3.294
Goats	-43.441	-45.873	-41.121

Table 19: Period sample sizes and fit within the fixed pooled representation.

Decade	Records	Households	Fixed fit (%)	Separate fit (%)	Maximum angle
1730	945	648	62.0	70.9	49.815
1740	4,905	1,917	66.5	72.7	43.570
1750	5,950	1,929	66.5	69.3	34.340
1760	7,616	2,532	64.7	65.5	29.564
1770	8,105	2,962	62.9	65.3	26.623
1780	5,336	2,954	66.0	64.4	22.622
1790	6,993	2,908	69.9	65.2	14.589
1800	5,034	3,524	72.3	75.8	9.225
1810	9,148	3,678	73.8	78.7	16.789
1820	6,490	2,774	72.8	77.3	17.948

Table 20: Full-data and equal-size-sample hull areas in squared pooled-PC units. One saved sample is used per decade. Central regions contain the 90% closest to each cloud's own centroid, including boundary ties. This is a diagnostic comparison, not a resampling distribution or confidence region.

Decade	N	Sample N	Full hull	Sample hull	Full central	Sample central
1,730	945	500	29.515	28.099	17.549	16.118
1,740	4,905	500	45.137	32.944	21.802	19.841
1,750	5,950	500	39.672	28.562	20.067	15.434
1,760	7,616	500	52.267	34.161	17.502	14.490
1,770	8,105	500	47.137	31.532	19.363	15.800
1,780	5,336	500	44.164	30.054	16.543	13.352
1,790	6,993	500	38.417	26.514	16.755	14.221
1,800	5,034	500	56.052	41.892	34.122	29.304
1,810	9,148	500	56.720	43.516	32.687	30.309
1,820	6,490	500	61.261	50.945	30.326	29.354

re-estimates its own centre and scales. The two fit shares therefore have different denominators and should not be compared as improvements in a single approximation problem. Full hulls describe the observed extremes; central regions describe the stated proportion under the declared construction. Equal-size comparisons use the same sampling rule in each period, and both panels use coordinates from the pooled model.

D Household partitions and distributional diagnostics

The reference k-means partition uses the full standardised data. The dimension search repeats estimation on the first r unwhitened scores for every r from two to full rank, comparing $k = 2, \ldots, 8$. For each representation, the selected solution follows the declared silhouette and tie rule, and the comparison with the reference reports assignments and substantive profiles as well as the silhouette. A high silhouette obtained by discarding dimensions does not show that the discarded information was irrelevant. Random starts reduce dependence on a particular local solution but do not guarantee the global optimum.

Table 21: Selection of the number of groups in the full-rank representation.

Groups	Silhouette	Smallest group share	Selected
2	0.432	0.354	TRUE
3	0.426	0.169	FALSE
4	0.384	0.118	FALSE
5	0.316	0.115	FALSE
6	0.256	0.105	FALSE
7	0.253	0.070	FALSE
8	0.249	0.051	FALSE

The nonlinear investigation uses a fixed sample for the t-SNE and UMAP grids. The diagnostic records retain the neighbourhood comparisons together with the settings needed to reproduce each

Table 22: Selected partitions across retained component dimensions, compared with the full-space reference.

Dimensions	Groups	Variance (%)	ARI to full
2	3	70.1	0.762
3	3	76.1	0.761
4	3	81.3	0.760
5	3	85.8	0.759
6	2	89.2	0.998
7	2	91.8	0.999
8	2	93.9	0.999
9	2	95.5	0.999
10	2	97.0	0.999
11	2	98.5	1.000
12	2	99.7	1.000
13	2	100	1

Table 23: Recorded quantities by the selected descriptive profile. Group labels describe the fitted partition.

Quantity	Profile 1	Profile 2
Records	13,307	7,606
Share (%)	63.630	36.370
Enslaved men	1.089	7.431
Enslaved women	0.688	2.860
Khoekhoe labour	0.623	3.912
Cattle	4.078	49.973
Horses	2.716	19.628
Sheep	4.666	117.134
Vines	2679.695	35419.058
Wine	1.441	20.523
Brandy	0.069	1.338
Wheat	3.287	79.168
Other grain	1.463	51.623
Wagons	0.273	2.687
Goats	2.112	42.780

Table 24: Stability of the descriptive partition under household-aware refitting.

Statistic	Median	2.5 percentile	97.5 percentile
k	3	3	3
r	13	13	13
ari	0.762	0.746	0.782
pair jaccard	0.782	0.768	0.799
mean best group jaccard	0.743	0.730	0.757

Table 25: Within-group inequality under the reference partition, year, recorded slaveholding and clustering with the focal asset excluded. Slaveholding includes men, women, sons and daughters.

Asset	Grouping	T	Within (%)
Enslaved persons	clusters	0.797	60.4
Cattle	clusters	1.041	45.7
Horses	clusters	0.865	52.7
Vines	clusters	1.431	59.1
Wine	clusters	1.679	64.0
Enslaved persons	year	0.797	99.4
Cattle	year	1.041	99.1
Horses	year	0.865	96.5
Vines	year	1.431	98.6
Wine	year	1.679	97.4
Enslaved persons	slaveholding	0.797	55.2
Cattle	slaveholding	1.041	83.5
Horses	slaveholding	0.865	85.5
Vines	slaveholding	1.431	82.6
Wine	slaveholding	1.679	84.5
Enslaved persons	focal asset excluded	0.797	65.5
Cattle	focal asset excluded	1.041	50.2
Horses	focal asset excluded	0.865	51.5
Vines	focal asset excluded	1.431	35.8
Wine	focal asset excluded	1.679	43.7

Table 26: Neighbourhood overlap and error in approximating standardised transformed quantities by embedding-neighbour means. Diagnostics use 15 neighbours on the fitted sample; they do not measure prediction on new observations.

Method	Parameter	Seed	Neighbour overlap	Profile RMSE
PCA	2	20,260,908	0.246	0.493
t-SNE	15	20,260,914	0.500	0.362
t-SNE	40	20,260,914	0.506	0.365
UMAP	15	20,260,914	0.456	0.383
UMAP	40	20,260,914	0.413	0.398
t-SNE	15	20,260,915	0.500	0.362
t-SNE	40	20,260,915	0.509	0.362
UMAP	15	20,260,915	0.454	0.386
UMAP	40	20,260,915	0.422	0.397
Kernel PCA (RBF)	0.500	20,260,908	0.211	0.550
Kernel PCA (RBF)	1	20,260,908	0.229	0.507
Kernel PCA (RBF)	2	20,260,908	0.243	0.494

Table 27: Variation within and between districts in the full standardised circa-1825 data.

Records	Within district (%)	Between districts (%)
8,405	86.5	13.5

fit. All eight fitted alternatives improve both neighbourhood overlap and original-variable prediction error over the two-component PCA projection, which prompted a bounded kernel comparison. For pairwise squared distances d_{ij}^2, the radial-basis kernel is $K_{ij} = \exp[-d_{ij}^2/(2\sigma^2)]$, with bandwidths of 0.5, 1 and 2 times the median positive pairwise Euclidean distance. The first two principal coordinates of the centred kernel do not improve those diagnostics over ordinary PCA. This result concerns this grid and sample, not all nonlinear models; the investigation is a bounded comparison rather than an exhaustive search.

The inequality code implements equation (2) directly, retaining all observed zeros. Missing quantities are excluded only under an explicit sample definition and are never replaced inside the inequality function. Zero-mean groups contribute zero; an entirely zero asset produces an undefined overall index. Numerical checks include $T(0, 2) = \log 2$, equality for equal holdings, invariance to multiplication by a positive constant, and equality of total inequality to independently calculated within and between terms. Inequality calculated on shifted quantities $x + 1$ is a different estimand and is not reported.

E A two-group canonical variate display

PCA chooses the directions along which households vary most overall. Canonical variate analysis (CVA) instead chooses directions that best separate predefined groups, relative to the variation within them. For transformed observations z_i, group means $\bar{z}_g$ and overall mean $\bar{z}$, define the within-group and between-group scatter matrices

$$W = \sum_g \sum_{i \in g} (z_i - \bar{z}_g)(z_i - \bar{z}_g)^\mathsf{T}, \qquad B = \sum_g n_g (\bar{z}_g - \bar{z})(\bar{z}_g - \bar{z})^\mathsf{T}. \tag{3}$$

The first canonical direction maximises $v^\mathsf{T} B v / (v^\mathsf{T} W v)$, or equivalently solves $Bv = \lambda W v$ with suitable normalisation. With two groups, B has rank at most one, so there is only one discriminating direction. A second plotted direction must therefore follow an additional criterion; it is not a second independent dimension of group discrimination.

The example compares returns recording enslaved-person holdings with returns recording none, excluding the defining ownership quantities from the active variables. The fitted input is $\log(1 + x)$ for the remaining quantities. We use the weighted two-group construction in `biplotEZ`, with `weightedCVA="weighted"` and `low.dim="sample.opt"` (Le Roux and Gardner-Lubbe, 2025). If C denotes a full canonical basis, C_2 its two display columns, and $C^{(2)}$ the first two rows of its inverse, the reconstruction is $\widehat{Z} = ZC_2C^{(2)}$. The additional direction minimises $\|Z - \widehat{Z}\|_F^2$ over the admissible bases that reconstruct the group means exactly. This total sample residual criterion is not an extra discriminant eigenvector; for two groups, no second nonzero discriminant direction exists.

The appropriate method depends on the comparison. More than two groups permit up to $G - 1$ discriminating directions. Unequal group sizes require an explicit weighting convention,

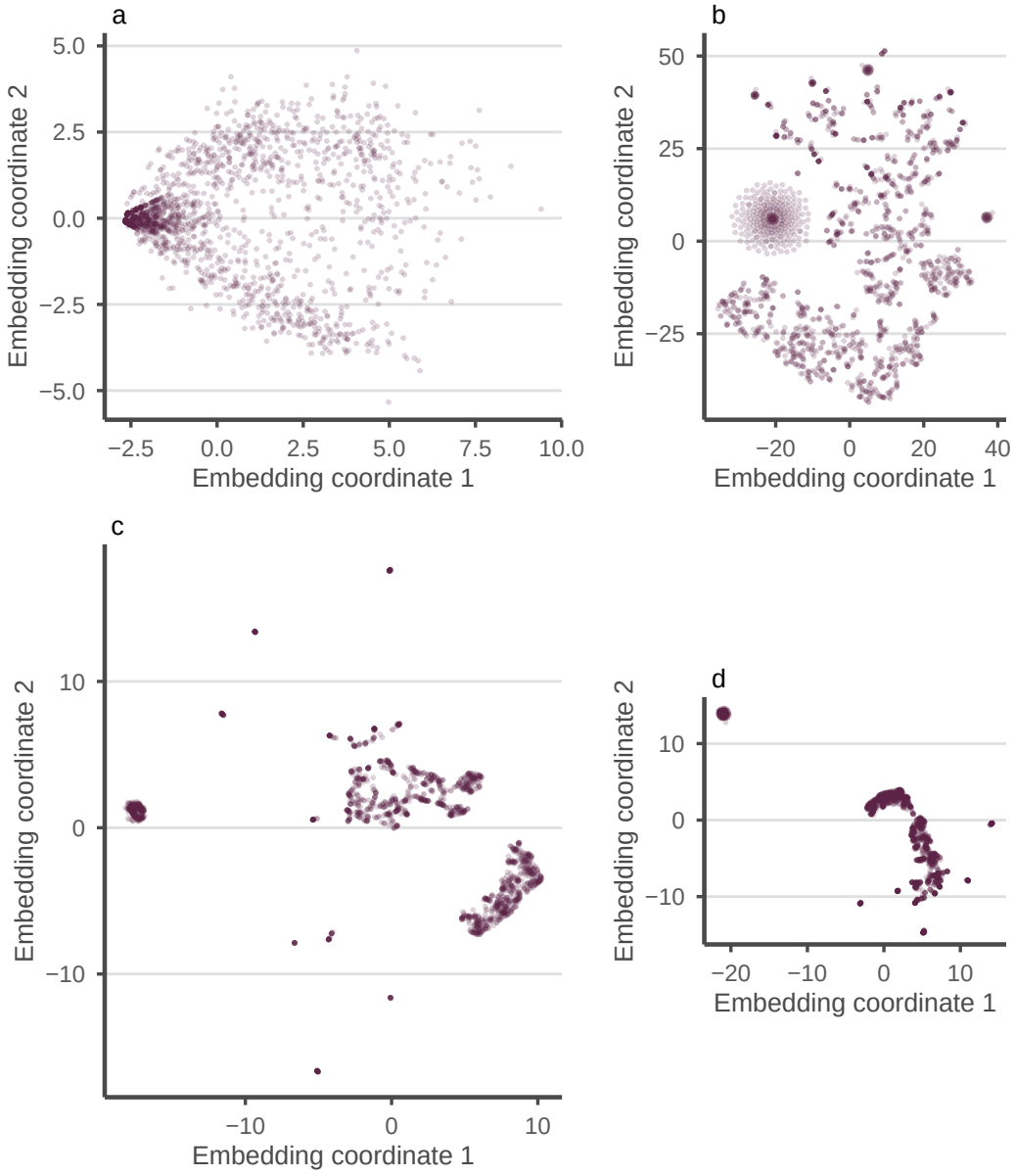


Figure 7: Display sensitivity on a fixed subsample. Panels in reading order show two-PC scores, t-SNE with perplexity 15, and UMAP with neighbourhood sizes 15 and 40 (minimum distance 0.1); nonlinear displays use seed 20260914. Apparent separation is not independent validation of household classes.

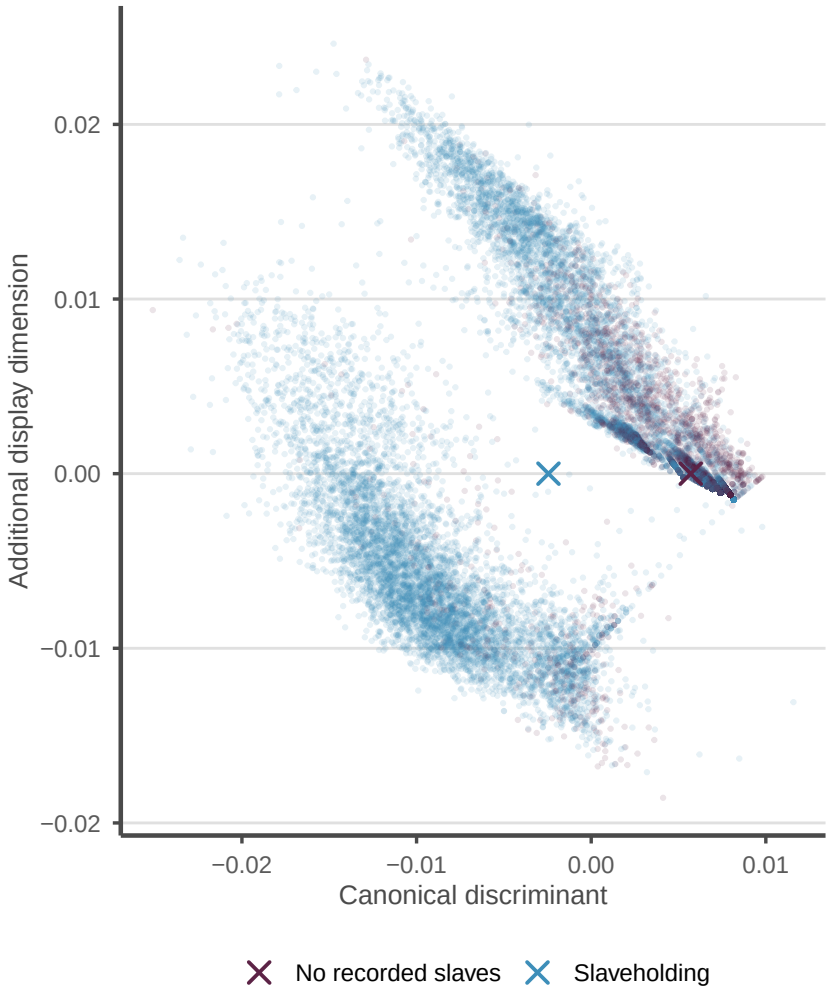


Figure 8: Two-group canonical variate display comparing returns with and without recorded enslaved-person holdings. Defining ownership variables are excluded from the active set. The first direction concerns separation relative to within-group scatter; the second follows the sample-optimal two-group criterion. Group differences are descriptive and do not identify a causal effect of ownership.

and singular within-group scatter requires a justified regularisation or a restricted variable space. PCA with group labels shown as supplementary information remains preferable when the question concerns total variation rather than maximised separation. A regression or classification model serves yet other objectives and needs its own assumptions and validation. In particular, if a group is defined by ownership of an asset, including that defining quantity among the discriminating variables risks a circular demonstration; the example therefore excludes it from the active variables.