

Online Appendix to: Uprooted: Migration, Coercion, and the Roots of Social Connectedness*

Johan Fourie[†]

This appendix provides supplementary tables and figures for “Uprooted: Migration, Coercion, and the Roots of Social Connectedness.” All country-level specifications cluster standard errors on the two country identifiers that define each undirected dyad, while US county regressions use state-level clustering and district-level regressions use heteroskedasticity-robust standard errors.

A. PPML Specification Diagnostics

Table A1 reports formal specification tests comparing the log-OLS and PPML estimators. The purpose is not to declare a single winner, but to show why the country-level ancestry coefficient is sensitive to estimator choice.

*This is the supplementary material for Fourie, Johan. 2026. “Uprooted: Migration, Coercion, and the Roots of Social Connectedness.” Working Paper, Department of Economics, Stellenbosch University. That paper was produced with substantial assistance from large language models, principally Anthropic’s Claude, with OpenAI’s Codex used for independent code and reference checking; all code, results and claims have been reviewed and verified by the author. Cite the paper, not this appendix.

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

Table A1: PPML specification diagnostics

Test	Statistic	Interpretation
Park-type test (slope)	1.926	Gamma-like variance ($\text{Var} \propto E[y]^2$) inconsistent with a strict Poisson variance benchmark
RESET test (OLS-log)	$p < 0.001$	Rejects log-linear form
RESET test (PPML)	$p = 1.0$	Does not reject PPML form
Gamma PML (anc_log)	0.186* (0.091)	Significant under Gamma
PPML with FE (anc_log)	0.046 (0.046)	Insignificant

Notes: Park-type test regresses $\log(\text{resid}^2)$ on $\log(\text{fitted values})$; a slope near 2 indicates Gamma-type heteroskedasticity. RESET tests add squared and cubed fitted values; joint significance indicates misspecification. Gamma PML estimated via GLM with Gamma family and log link on positive-SCI subsample ($N = 14,028$). PPML with FE estimated via `fepois` with country fixed effects. Taken together, the diagnostics indicate estimator sensitivity rather than a clean validation of one specification.

Variance inflation factors. A natural concern about the kitchen-sink horse race in Table 2 of the main paper is multicollinearity among the bilateral channels. Table A2 reports variance inflation factors for the regressors in the kitchen-sink specification (excluding the country fixed effects). All values lie comfortably below the conventional threshold of 10, indicating that the precision pattern in the main paper is not driven by inflated standard errors from collinear regressors.

Table A2: Variance inflation factors, kitchen-sink horse race

Regressor	VIF
<code>anc_log</code>	1.22
F_{ST} genetic distance	1.89
Linguistic distance	1.79
Religious distance	1.25
Folklore similarity	1.80
Pre-1870 trade (log)	1.34
Colonial-era trade (log)	1.69
Post-war trade (log)	1.28
Log distance	1.81
Contiguity	1.27
Colonial tie	1.12
Maximum VIF	1.89

Notes: VIFs computed via `car::vif` on the kitchen-sink specification of Table 2 of the main paper (excluding country fixed effects). All VIFs are below 2, well under the conventional cut-off of 10. Multicollinearity is therefore not a plausible explanation for the precision pattern in the kitchen-sink horse race.

B. Expanded US County Analysis

Table A3 extends the county-level analysis by adding Nigerian, Ethiopian, Korean and Japanese origin measures. The county-origin variables combine ACS Table B04006 for European and African ancestries, Table B02015 for Asian-origin groups and Table B03001 for Mexican origin.

Table A3: Expanded county-level ancestry coefficients

Origin	β	SE	N	p -value
Ireland	0.371***	(0.038)	3,140	< 0.001
Italy	0.188***	(0.020)	3,121	< 0.001
Poland	0.169***	(0.024)	3,007	< 0.001
Japan	0.121***	(0.012)	1,611	< 0.001
Korea	0.176***	(0.017)	1,893	< 0.001
India	0.218***	(0.019)	1,893	< 0.001
France	0.115***	(0.025)	3,127	< 0.001
Ethiopia	0.299***	(0.036)	825	< 0.001
Philippines	0.172***	(0.019)	2,367	< 0.001
UK	0.250***	(0.034)	3,145	< 0.001
Germany	0.223***	(0.029)	3,150	< 0.001
Mexico	0.577***	(0.042)	3,133	< 0.001
Nigeria	0.175***	(0.018)	1,261	< 0.001

Notes: Dependent variable: $\log(\text{SCI}_{c,k})$. Each row is a separate regression of county-level SCI on log county-origin share, controlling for $\log(\text{pop})$ and state fixed effects. County-origin shares use ACS 2018–2022 five-year extracts: B04006 for European and African ancestries, B02015 for Asian-origin groups and B03001 for Mexican origin, all divided by county total population (B01003). Standard errors clustered at the state level. China excluded because the county SCI extract does not contain a China partner series. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

C. Herfindahl Concentration Test

Hypothesis 1 predicts that the ancestry-SCI elasticity should be larger for more geographically concentrated ancestral groups. Table A4 reports the Herfindahl-Hirschman Index of geographic concentration for each origin group alongside its estimated coefficient.

Table A4: Geographic concentration and ancestry-SCI elasticity

Origin	HHI	β
Philippines	0.0053	0.172
India	0.0030	0.218
Mexico	0.0013	0.577
Poland	0.0009	0.169
France	0.0007	0.115
Italy	0.0007	0.188
Germany	0.0005	0.223
UK	0.0004	0.250
Ireland	0.0004	0.371
Correlation(HHI, β)		-0.166

Notes: HHI computed as $\sum_c (s_c)^2$ where s_c is county c 's share of the national total for each origin group. The cross-group correlation between concentration and elasticity is weakly negative (-0.166), so this measure does not support the directional concentration prediction of Hypothesis 1.

D. “American” Ancestry and the Identity Channel

Many Americans of British descent report “American” rather than a specific European origin on the Census, particularly in the South and Appalachia. If the ancestry-SCI relationship operates through active ethnic identification, “American” ancestry should not predict SCI with any specific country. If it operates through deeper demographic roots, it should.

Counties reporting higher “American” ancestry show a significant positive association with SCI with the United Kingdom ($\beta = 0.061$, $p < 0.001$). In a horse race with British ancestry, both variables are significant: British $\beta = 0.054$ ($p = 0.002$), American $\beta = 0.071$ ($p < 0.001$). This suggests that the ancestry-SCI relationship reflects deep demographic roots that persist even when subjective ethnic identification has dissolved.

E. South Africa: Internet Access and Urbanisation Controls

Table A5 reports the SA district-level regressions with Census 2022 internet access as an additional control.

Table A5: South Africa: robustness to internet access and urbanisation controls

	S1 EU baseline	S1c EU + internet	S6 Af. placebo	S6c Plac. + internet
% White	0.578*** (0.014)	0.596*** (0.016)	0.230*** (0.030)	0.243*** (0.036)
Internet access		-0.430*** (0.126)		-0.258 (0.273)
Partner FE	Yes	Yes	Yes	Yes
Within R^2	0.705	0.709	0.332	0.334
N	728	728	312	312

Notes: Dependent variable: $\log(\text{SCI})$. All specifications include $\log(\text{pop})$. Internet access = share of households with any internet (Census 2022). The placebo coefficient does not shrink when internet is controlled; the correlation between %White and internet access is $r = 0.19$. Heteroskedasticity-robust standard errors. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A6 applies the same logic to the slave-origin proxy. The positive %Coloured coefficient does not collapse when internet access, formal dwelling share, urban share, or all three are added. This is why the main text treats the modern racial category as an imprecise demographic correlate rather than as the decisive test of Hypothesis 3.

Table A6: South Africa: modern %Coloured proxy with urbanisation controls

	S6 Baseline	S6c + internet	S6d + formal	S6e + urban	S6f + all three
% Coloured	0.158*** (0.016)	0.152*** (0.016)	0.140*** (0.016)	0.123*** (0.020)	0.155*** (0.021)
Internet access		0.958*** (0.188)			0.797*** (0.208)
Formal dwellings			1.036*** (0.194)		0.828*** (0.201)
Urban share				0.101* (0.044)	-0.047 (0.049)
Partner FE	Yes	Yes	Yes	Yes	Yes
Within R^2	0.241	0.290	0.294	0.252	0.319
N	364	364	364	364	364

Notes: Dependent variable: $\log(\text{SCI})$. All specifications include $\log(\text{pop})$ and $\log(\text{dist})$ and are estimated on the seven slave-origin partner countries used in column S6 of Table 5: MZ, MG, ID, IN, MY, LK and MU. Internet access, formal dwellings and urban share are district-level controls from the 2022 Census. Heteroskedasticity-robust standard errors. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A7 reports the distribution of matched slave-origin shares by partner country. The

identifying variation is visibly dominated by Mozambique, which appears in all 12 matched districts and accounts for 730 of the 914 matched individuals. Madagascar, Indonesia and India contribute some additional support, but Mauritius, Sri Lanka and Malaysia are effectively near-zero cells. The pooled slave-origin coefficient in the main text should therefore be read as Mozambique-weighted evidence rather than as a sharply estimated average across seven well-measured origins.

Table A7: Cape slave-origin variation by partner country

Partner	Total slaves	Nonzero districts	Mean share	SD share	Max share
Mozambique	730	12	0.789	0.103	1.000
India	63	8	0.059	0.058	0.182
Madagascar	53	7	0.059	0.075	0.250
Indonesia	43	7	0.051	0.065	0.200
Sri Lanka	4	2	0.001	0.003	0.008
Mauritius	3	2	0.001	0.003	0.008
Malaysia	1	1	0.001	0.002	0.008

Notes: Based on the 914 matched slave-emancipation records used in the Cape Colony regressions. Shares are district-level origin shares among foreign-origin enslaved people, computed across the 12 districts with any matched slave-origin data; districts with zero share for a given partner are included when calculating the mean and standard deviation. Algeria and Timor-Leste are excluded because they are not part of the seven-partner slave-origin set used in the main regressions.

Table A8: Validation of the farm-name classifier using Huguenot-linked farms

Classifier label	Farms	Share
Dutch	27	0.659
French	14	0.341

Notes: Based on 41 uniquely matched farms in the Huguenot validation file, constructed by linking historically Huguenot-associated owner surnames to named farms. The key point is not the exact split but the direction of the mismatch: even within a Huguenot-linked subset, most farms are classified as Dutch by the toponymic algorithm. This indicates that farm-name language reflects colonial naming conventions as well as settler ethnicity and should therefore be interpreted as a noisy historical settlement marker rather than a literal ethnicity measure.

Atlantic and Intra-American Slave-Trade Corridor Specifications

This subsection supports the external-validity exercise in Section 8 of the main paper. The historical corridor variables are constructed from the trans-Atlantic and intra-American Slave Voyages databases using the number of enslaved people *disembarked* on each voyage, aggregated to modern country pairs after mapping historical embarkation places or regions and disembarkation ports or regions to present-day sovereign states. Exact place matches are used where possible; region-level matches are used when the historical geography is clear; and broad unresolved labels such as “Africa, port unspecified” are left unmapped rather than assigned mechanically. The final trans-Atlantic analysis sample contains 265 Africa-Americas dyads, of which 178 have strictly positive mapped exposure. The intra-American sample contains 153 Americas-Americas dyads, of which 45 have positive mapped exposure. A handful of dyads lack complete gravity controls, so the regressions in Table A9 use 263, 170 and 151 observations respectively (full Africa-Americas, positive-exposure trans-Atlantic, and intra-American).

All specifications mirror the country-level gravity framework in the main paper. The dependent variable is $\log(\text{SCI}_{ij})$. The slave-trade regressor is $\log(1 + \text{slave corridor}_{ij})$, where the corridor measure is symmetrised at the country-pair level. The horse-race columns add the paper’s preferred ancestry measure, anc_log_{ij} . Every specification includes log distance, contiguity, common official language, colonial ties and country fixed effects, with standard errors clustered on the two country

identifiers defining each undirected dyad.

Table A9: Atlantic and intra-American slave-trade corridor regressions

Dependent Variable:	log(SCI)				
	Trans-Atlantic			Intra-American	
Model:	(1)	(2)	(3)	(4)	(5)
<i>Variables</i>					
log(1 + slave corridor)	0.025 (0.015)	0.025 (0.015)	1.09×10^{-5} (0.044)	-0.047 (0.078)	-0.041 (0.080)
log shared ancestry		0.085 (0.065)	0.158 (0.087)		0.165 (0.184)
<i>Fixed-effects</i>					
Country FE	Yes	Yes	Yes	Yes	Yes
<i>Fit statistics</i>					
Observations	263	263	170	151	151
Within R ²	0.48956	0.49415	0.51382	0.66234	0.66375

Clustered (iso3_i & iso3_j) standard-errors in parentheses

*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05, .: 0.1*

Dependent variable: log(SCI).

Columns 1-3 use the Africa-Americas trans-Atlantic sample; columns 4-5 use the Americas-only intra-American sample.

Column 3 restricts the trans-Atlantic sample to dyads with strictly positive mapped slave-corridor exposure.

All specifications include gravity controls and country fixed effects, with standard errors clustered on the two country identifiers.

F. South Africa: Genetics-Based Coloured Ancestry Robustness

As an exploratory robustness exercise, I use recently published population-genetic evidence on the South African Coloured population. Lankheet et al. (2025) report ancestry shares for 22 sampled Coloured communities, together with site coordinates and autosomal sample sizes. I map those sampled communities into modern GADM2 districts and use the site-level admixture shares to split the census %Coloured category into genetics-based Asian-linked, African-linked and Khoesan-linked components.

I do not use these variables in the main analysis. The reason is simple: the published genetics sample maps into only 10 of the 52 districts used in the South Africa regressions. A sampled-district

robustness test using the Asian-linked component is directionally positive but highly imprecise ($\beta = 0.038$, $SE = 0.071$, $N = 40$ district-partner observations), and the analogous African-linked split does not have enough observations to support stable estimation. The exercise is therefore better viewed as proof of concept than as evidence capable of adjudicating the main hypothesis.

The value of the genetics material is prospective. It shows that novel source material can, in principle, be used to decompose the broad census %Coloured category into historically meaningful ancestry components. With wider geographic coverage, future genetics-based work could help distinguish more sharply between Khoesan, African, South Asian and Southeast Asian contributions to the social connectedness patterns documented in the South Africa section.

G. Per-Country Region-Level Results: New Zealand and Australia

Table A10 reports the per-country, region-level regressions of $\log(\text{SCI})$ on birthplace shares that underlie the pooled Australasian analysis in Section 7 of the main paper, documenting the individual-country coefficients behind the free-versus-penal comparison.

Table A10: Per-country region-level results: New Zealand (1881) and Australia (1901)

	New Zealand only ($N = 13$)			Australia only ($N = 5$)	
	β	Asym. p	Perm. p	β	Asym. p
% UK-born $\rightarrow$ GB	0.280	< 0.001	0.137	0.268	0.134
% English-born $\rightarrow$ GB	0.279	0.003	0.081	0.317	0.171
% Scottish-born $\rightarrow$ GB	0.124	0.390	—	—	—
% Irish-born $\rightarrow$ IE	0.102	0.590	—	—	—

Notes: Each row is a separate regression of $\log(\text{SCI})$ on $\log$ birthplace share, controlling for $\log(\text{pop})$. No fixed effects (single country-partner pair). Perm. p : two-sided permutation p -value from 5,000 permutations. Heteroskedasticity-robust standard errors. Australian regressions have only 5 observations (2 degrees of freedom) and should be interpreted as descriptive.

H. GADM2 District-Level Analysis: New Zealand and Australia

Table A11 reports regressions pooling NZ and AU at the GADM2 district level (29 NZ districts with borough-level birthplace data and 384 AU districts with state-level birthplace data). The Australian observations dominate the pooled sample (384 of 413 observations), and all AU districts within the same state share identical birthplace shares. Standard errors are clustered at the GADM1 region/state level (18 clusters). None of the pooled GADM2-level birthplace coefficients are significant, reflecting the dominance of Australian districts whose state-level birthplace shares provide no within-state variation. The main text therefore relies on the NZ-only district evidence and stacked NZ partner-specific tests in Table 4 rather than these pooled regressions.

Table A11: GADM2 district-level analysis: historical birthplace and SCI

	(P1) %UK→GB	(P2) %Eng→GB	(P3) %Scot→GB	(P4) %Irish→IE	(P5) Pooled
Birthplace share	−0.170 (0.154)	−0.239 (0.175)	−0.042 (0.127)	0.193 (0.125)	0.200 (0.176)
log(pop)	0.118*** (0.029)	0.105*** (0.024)	0.115** (0.038)	0.351*** (0.066)	0.232*** (0.055)
Country FE	Yes	Yes	Yes	Yes	Yes
Partner FE	No	No	No	No	Yes
Within R^2	0.026	0.028	0.022	0.113	0.069
N	413	413	413	413	826
Clusters	18	18	18	18	18

Notes: Dependent variable: log(SCI). Pooled sample: 29 NZ GADM2 districts (1881 Census, 65 boroughs aggregated via crosswalk) and 384 AU GADM2 districts (1901 Census, 5 states). Standard errors clustered at the GADM1 region/state level (18 clusters). Birthplace shares are constant within GADM1 regions for most NZ districts and within states for all AU districts, so the GADM2-level variation comes primarily from the dependent variable. None of the birthplace coefficients are statistically significant at the GADM2 level, indicating that the ancestry-SCI relationship operates at the regional (GADM1) rather than the district (GADM2) level. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

The insignificance of the GADM2-level results is informative. Within a region such as Otago, districts with more Scottish-borough heritage (Dunedin: 20.8% Scottish-born) are not measurably more connected to the UK than districts with less (Central Otago: 13.0% Scottish-born). The ancestry-SCI relationship operates at the regional level – between Canterbury (English) and Otago (Scottish) – consistent with the institutional-persistence mechanism: settlement schemes created regional identities and institutions that sustain homeland connections at the provincial rather than the neighbourhood level.

I. US County Placebo Matrix

The US county-level analysis in Section 6 of the main text reports the diagonal of a 9×9 placebo matrix: each ancestry against its matched partner country (Irish → Ireland, Italian → Italy, etc.). This appendix reports the full matrix, running each of the nine ancestry groups against *every* partner country. Each cell is a separate fixed-effects regression of log(SCI) on log(%ancestry) and log(population), with state fixed effects and state-clustered standard errors. The diagonal cells are the matched-ancestry results that appear in Table 3 of the main text; the off-diagonal cells test whether the ancestry → SCI relationship is partner-specific or generic.

If the diagonal cells are large and the off-diagonal cells are small, then the ancestry effect is sharply concentrated on the matched homeland – ruling out the alternative interpretation that some county-level feature (urbanisation, education, cosmopolitan orientation) drives ancestry →

partner connectedness for any partner country.

Table A12: Placebo coefficient matrix: each ancestry against every partner country

Ancestry (row)	Partner country (column)								
	UK	Germany	Ireland	Italy	Mexico	Poland	France	India	Philipp.
UK	0.250***	0.216***	0.293***	0.188***	-0.217*	0.177***	0.175***	-0.001	-0.093
Germany	0.209***	0.223***	0.288***	0.175***	-0.062	0.174***	0.144**	-0.038	-0.028
Ireland	0.276***	0.239***	0.371***	0.225***	-0.291**	0.255***	0.193***	0.013	-0.109
Italy	0.184***	0.159***	0.245***	0.188***	-0.111*	0.204***	0.163***	0.049	0.038
Mexico	-0.022	-0.025	-0.013	0.001	0.577***	0.007	0.008	0.039*	0.061**
Poland	0.133***	0.128***	0.180***	0.130***	-0.068	0.169***	0.115***	0.027	0.015
France	0.142***	0.125***	0.169***	0.116***	-0.153**	0.129***	0.115***	-0.004	0.029
India	0.078***	0.062***	0.081***	0.060***	0.012	0.056***	0.085***	0.218***	0.074***
Philippines	0.070***	0.083***	0.082***	0.082***	0.032	0.089***	0.079***	0.096***	0.172***

Notes: Each cell reports the coefficient on $\log(\% \text{ancestry})$ from a separate regression of $\log(\text{SCI}_{c,k})$ on $\log(\% \text{ancestry}_{c,a})$ and $\log(\text{pop}_c)$, with state fixed effects and standard errors clustered at the state level. Diagonal cells (bold) reproduce Table 3 in the main text. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A13: Diagonal vs off-diagonal contrast in the US county placebo matrix

Statistic	Value
Mean diagonal coefficient (matched ancestry $\rightarrow$ matched partner)	0.254
Mean off-diagonal coefficient (mismatched)	0.079
Median diagonal	0.218
Median off-diagonal	0.082
Paired t -test (each diagonal vs its row's off-diagonal mean)	$t = 3.22$
p-value	0.012
Diagonal cells significant at $p < 0.05$	9 / 9
Off-diagonal cells significant at $p < 0.05$	50 / 72
Diagonal is row maximum (Ireland, Mexico, India, Philippines)	4 / 9

Notes: Summary statistics for the full 9×9 US county placebo matrix in Table A12. The paired t -test compares each diagonal coefficient to the average of the off-diagonal cells in the same row, i.e. the same ancestry against its mismatched partners.

The corrected placebo matrix shows a clearer diagonal premium than the earlier, miscoded version. The mean diagonal coefficient is 0.254 versus 0.079 off diagonal, and the paired test of each diagonal cell against its row's off-diagonal mean is significant ($t = 3.22$, $p = 0.012$). All 9 diagonal cells are significant at the 5% level, although 69% of off-diagonal cells remain significant as well. The reason is collinearity in county-level ancestry composition: counties with many Irish-Americans also have many Italian-, German- and Polish-Americans, so a regression of $\log(\text{SCI})$ to country k on a single ancestry share inevitably loads on whatever broader settlement pattern that ancestry tracks. A clean partner-specific test would require orthogonalising the ancestry shares

across the nine groups, which the available ACS sample sizes do not support.

What the placebo *does* demonstrate cleanly is stronger matched-partner specificity on average, not a universal row-by-row dominance result. The diagonal is the row maximum for Ireland, Mexico, India and the Philippines, while the remaining rows still load on correlated settlement patterns. The placebo therefore strengthens the case that matched homeland ties matter, but it also shows that the county ancestry vectors are too collinear for a sharp partner-specific test in every row.

J. South Africa: Asia and Latin America Placebos

Section 8 of the main text reports a stacked Europe vs Africa placebo specification with a White-times-Europe interaction term. The interaction is large and significant, indicating that the differential effect of %White on European versus African connectedness is not driven by generic cosmopolitan orientation. To close off the residual concern that this differential is unique to the Europe-Africa comparison, this appendix extends the placebo logic to two additional partner regions with no historical South African tie:

- *Asia placebo*: CN, JP, KR, TH, VN. India, Indonesia, Malaysia, Sri Lanka, Mauritius and the Philippines are explicitly excluded because of their historical SA links (indentured migration, slave origins, post-war labour movements).
- *Latin America placebo*: BR, AR, CL, CO, PE, UY. No country in this set has a meaningful historical migration link to South Africa.

If the cosmopolitanism alternative is correct, the $\log(\%White) \times \text{Europe}$ interaction should shrink towards zero when Europe is compared against *any* non-African partner region. If the matched-ancestry interpretation is correct, the Europe interaction should remain large.

Table A14: South Africa: stacked specifications with Asia and Latin America placebos

	(1) EU vs Africa	(2) EU vs Asia	(3) EU vs LatAm	(4) Four-region pool
$\log(\%White) \times \text{Europe}$	0.348***	0.307***	0.249***	0.308***
$\log(\%White)$ (placebo base)	0.230***	0.265***	0.313***	–
Partner FE	Yes	Yes	Yes	Yes
Region dummies	–	–	–	Yes
N	1,040	936	1,040	1,560

Notes: Dependent variable: $\log(SCI_{dk})$ where d indexes 52 GADM2 districts and k indexes partner countries. All specifications include $\log(\text{pop})$, $\log(\text{distance})$ and partner fixed effects. Asia partners: CN, JP, KR, TH, VN. LatAm partners: BR, AR, CL, CO, PE, UY. Africa partners: MZ, LS, SZ, ZM, TZ, MW. Standard errors are heteroskedasticity-robust. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

The bare $\log(\%White)$ coefficients on the three placebo regions (Africa: 0.230, Asia: 0.265, LatAm: 0.313) are similar in magnitude, indicating that whiter districts are indeed more internationally connected on a generic cosmopolitan axis – a feature the paper should not deny. But the

White $\times$ Europe interaction is large and consistent across all three placebo choices (0.249–0.348) and remains 0.308 in the four-region pool. Europe sits roughly 0.25–0.35 above the cosmopolitanism baseline that the other three placebo regions all share, and this differential is the matched-ancestry channel: it is what voluntary European settlement adds to whatever generic international orientation is associated with the White population share. The robustness of this differential to the choice of placebo region rules out the alternative interpretation that Europe is unusual on a generic “rich-country” axis.

K. Additional Log-OLS Defense

Section 9 of the main text reports mixed evidence on functional form: PPML with country fixed effects produces an imprecise ancestry coefficient near zero, whereas log-OLS and Gamma PML yield positive coefficients. This appendix adds three further diagnostics: a Manning-Mullahy GLM family-selection criterion, a graphical mean-variance diagnostic, and an out-of-sample 5-fold cross-validation comparison across log-OLS, PPML and Gamma PML.

K.1 Manning-Mullahy GLM family-selection criterion

Manning and Mullahy (2001) propose using the slope of a Park-type regression — the log of squared residuals on the log of fitted values — as a guide to the appropriate GLM family. The slope estimates the variance-mean exponent k in $\text{Var}(y) \propto E[y]^k$. The mapping is:

k	Appropriate family
≈ 0	Gaussian / OLS
≈ 1	Poisson / PPML
≈ 2	Gamma PML
≈ 3	Inverse Gaussian

The empirical Park slope from PPML residuals on the SCI sample is 1.93, placing the SCI’s variance structure closer to the Gamma family than to the Poisson family. This is evidence against treating the Poisson variance benchmark as exact for SCI, but it is not by itself sufficient to validate log-OLS or to rule out PPML on all margins. log-OLS and Gamma PML nevertheless deliver comparable positive ancestry coefficients ($\beta = 0.144$ and $\beta = 0.186$ respectively), whereas PPML with fixed effects is much smaller and imprecise.

K.2 Variance-vs-mean diagnostic

To make the variance-mean exponent visible, I bin the analysis sample into deciles of fitted SCI from the country-FE log-OLS model, compute the empirical variance of SCI within each bin, and plot $\log(\text{var})$ against $\log(\text{mean})$ across deciles. The slope of this scatter is a non-parametric estimate

of the empirical mean-variance exponent. Reference lines for slope = 1 (Poisson) and slope = 2 (Gamma) anchor the interpretation.

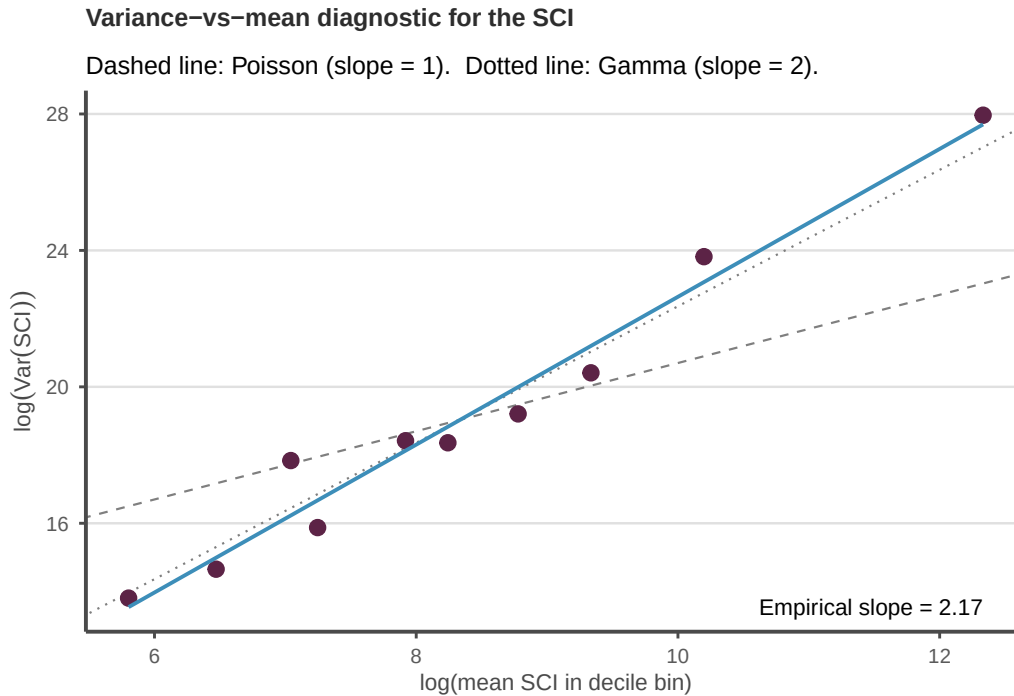


Figure A1: Variance-vs-mean diagnostic for the SCI. Each point is one decile of fitted SCI from the country-FE log-OLS model. Dashed line: Poisson reference (slope = 1). Dotted line: Gamma reference (slope = 2). The empirical slope across deciles is reported in the figure annotation.

The empirical relationship is closer to slope = 2 than to slope = 1, confirming the Park-test diagnostic in K.1.

K.3 Out-of-sample predictive comparison

One might still argue that the diagnostics in K.1 and K.2 are theoretical rather than substantive, and that PPML's robustness properties make it the preferred default regardless of the variance structure. As a substantive check, I conduct a 5-fold cross-validation in which each fold trains log-OLS, PPML and Gamma PML on 80% of country pairs and predicts the held-out 20%. RMSE is reported on both the log scale (where log-OLS is fit) and the level scale (where PPML is fit).

Table A15: 5-fold cross-validation: out-of-sample prediction error

Estimator	log-RMSE	level-RMSE
log-OLS	1.066	815,079
PPML	1.653	655,817
Gamma PML	1.303	3,624,577

Notes: Mean RMSE across 5 folds; standard deviations across folds in the run log. Each fold trains on 80% of country pairs and predicts the held-out 20%. log-RMSE is computed on $\log(\text{SCI})$; level-RMSE on the back-transformed prediction.

The cross-validation result is split: each estimator dominates the metric on which it is fit. log-OLS achieves the lowest log-scale RMSE (1.066 versus 1.653 for PPML), whereas PPML achieves the lowest level-scale RMSE (655,817 versus 815,079 for log-OLS). Gamma PML underperforms both estimators on the level scale because back-transformed Gamma predictions are sensitive to the long right tail of SCI. This split reinforces the main text’s conclusion that no single estimator resolves the issue cleanly.

Taken together, K.1 through K.3 show why the estimator choice is genuinely contested. The Park slope and the empirical mean-variance relationship both point away from a strict Poisson variance structure, whereas the RESET tests and level-scale prediction performance are more congenial to PPML. log-OLS performs best on the log scale, PPML performs best on the level scale, and Gamma PML lies between them on coefficient magnitude. The appendix therefore motivates caution about treating the benchmark log-OLS coefficient as a uniquely privileged parameter, even though the broader pattern of evidence remains consistent with a positive ancestry-connectedness relationship.

L. Bilateral Air Traffic and Facebook Penetration Robustness

Section L reports two further robustness exercises: a bilateral-air-traffic control and Facebook-penetration subsamples.

Bilateral direct flights. Table A16 re-estimates the kitchen-sink horse race of Table 2 of the main paper, adding $\log(1+n_{\text{routes}})$ as a country-pair control. The route-count variable is constructed from the OpenFlights database (OpenFlights, 2017): for each pair of countries I count the number of distinct airport-to-airport routes served by at least one direct scheduled flight, symmetrise across the two directions and take the log. The ancestry coefficient remains positive and statistically significant once bilateral aviation is controlled for.

Table A16: Horse race with bilateral direct-flight routes (OpenFlights)

	(1)	(2)
	Kitchen sink	+ log(routes)
<i>Dep. var.: $\log(SCI_{ij})$</i>		
anc_log	0.084**	0.075**
	(0.030)	(0.029)
$\log(1 + n_{\text{routes}})$		0.147**
		(0.049)
F_{ST} , language, religion, folklore, trade	Yes	Yes
Country FE	Yes	Yes
Within R^2	0.606	0.608
N	11,323	11,323

Notes: Both columns include the standard gravity controls (log distance, contiguity, colonial tie) and country fixed effects. Standard errors clustered on the two country identifiers defining each undirected pair. The route count is constructed from the OpenFlights routes file (2017 snapshot) by counting distinct airport-pair routes between the two countries, symmetrised across directions and entered in logs. The ancestry coefficient is slightly attenuated by the addition of bilateral routes (0.084 $\rightarrow$ 0.075) but remains significant at the 1% level; bilateral aviation is itself a significant correlate of social connectedness.

Facebook penetration thresholds. Table A17 restricts the country-level baseline of Table 1 of the main paper to country pairs in which both countries exceed two estimable penetration cutoffs: 25% and 50% of the population on Facebook in 2021. The 75% cutoff leaves fewer than 100 country pairs and is therefore omitted. The ancestry coefficient remains positive and statistically significant in both restricted subsamples.

Table A17: Baseline estimated on Facebook-penetration subsamples

	(1)	(2)	(3)
	Full sample	FB pen. $\geq 25\%$	FB pen. $\geq 50\%$
<i>Dep. var.: $\log(SCI_{ij})$</i>			
anc_log	0.144*** (0.031)	0.189*** (0.036)	0.158*** (0.032)
Gravity controls	Yes	Yes	Yes
Country FE	Yes	Yes	Yes
<i>N</i>	14,026	859	593

Notes: Each column re-estimates the gravity-plus-fixed-effects baseline of Table 1 of the main paper after restricting the sample to country pairs in which both countries exceed the indicated 2021 Facebook penetration threshold. Penetration is measured as monthly active Facebook users divided by total population, using DataReportal’s 2021 country-level snapshots for 54 large economies. The 75% threshold is omitted because the implied subsample has fewer than 100 country pairs and so does not support a clustered country-FE regression. The ancestry coefficient is, if anything, slightly larger in the high-penetration subsamples than in the full sample, indicating that the country-level result is not driven by very-low-penetration country pairs. The exercise narrows but does not eliminate the “Western network” concern: bilateral connectedness on closed networks (VK, WeChat) remains unobservable.

M. Data sources

This appendix collects variable-by-variable definitions and construction details for every dataset used in the paper, beyond the Social Connectedness Index and the Putterman-Weil World Migration Matrix described in Section 3 of the main text.

M.1 Bilateral distance measures

The horse race in the main paper requires a comprehensive set of competing bilateral channels.

Gravity variables. The CEPII Gravity Database (Head, Mayer and Ries, 2010) provides geodesic distance (population-weighted), contiguity, common official language and colonial relationship for the 2021 cross-section. These are the standard controls in any gravity model of bilateral outcomes.

Slave-trade corridors. For the external-validity test of Hypothesis 3, I use the trans-Atlantic and intra-American modules of the Slave Voyages database (SlaveVoyages Consortium, 2024). In both files, the historical flow variable is the recorded number of enslaved people *disembarked* rather than embarked. I aggregate those flows to modern country-pair corridors by mapping historical embarkation places and regions in Africa or the Americas, and historical disembarkation ports and regions in the Americas, to present-day sovereign states. The mapping proceeds conservatively: exact places and ports are used where possible, administrative or region-level matches are used where

the historical geography is clear, and broad unresolved labels such as “Africa, port unspecified” are left unmapped rather than forced into arbitrary modern-country assignments. After symmetrising the resulting country-pair corridor measures, the usable trans-Atlantic sample consists of 265 Africa-Americas dyads, of which 178 have strictly positive mapped exposure; the corresponding intra-American sample contains 153 Americas-Americas dyads, of which 45 have positive mapped exposure. The mapping covers 70.9 percent of total recorded trans-Atlantic disembarkations and 83.4 percent of total recorded intra-American disembarkations. These measures should be read as historical flow-based proxies for forced-migration exposure, not as descendant stocks analogous to the Putterman-Weil ancestry matrix.

Genetic distance. F_{ST} genetic distance from Spolaore and Wacziarg (2009), with the updated bilateral measures of Spolaore and Wacziarg (2018), captures the share of total genetic variation attributable to between-group differences. I use both the modern (post-migration) weighted measure and the pre-1500 measure. Coverage: 12,246 of 14,028 analysis pairs (87 percent).

Out-of-Africa migratory distance. As a coarser proxy for deep relatedness, I use the migratory-distance data of Ashraf and Galor (2013): each country’s migratory distance from Addis Ababa, intended to capture the serial-founder process in human expansion. I convert this to a bilateral proximity measure by taking the negative absolute difference between the two countries’ migratory distances, so larger values indicate more similar deep migratory histories. I also use the accompanying predicted-diversity measure as an auxiliary robustness control. Coverage: 12,880 of 14,028 pairs for migratory distance (92 percent).

Linguistic distance. Tree-weighted linguistic distance from Spolaore and Wacziarg (2016) is a continuous measure based on the structure of language family trees. It subsumes the binary common-language indicator and provides a finer-grained measure of communicative proximity. Coverage: all 14,028 pairs.

Religious distance. Bilateral religious distance from Spolaore and Wacziarg (2016), computed from the distribution of adherents across religious groups in the year 2000. Coverage: all 14,028 pairs.

Cultural distance. Cultural distance from Desmet and Wacziarg (2021), based on the World Values Survey. Coverage is limited to the 5,050 pairs (36 percent) for which both countries have WVS data.

Folklore similarity. I compute bilateral cosine similarity from the Michalopoulos and Xue (2021) motif catalogue, which records the intensity of 2,564 mythological motifs across 199 countries. For each pair, folklore similarity is the cosine of the angle between the two countries’ motif intensity vectors. This captures deep cultural transmission through oral traditions – creation myths, trickster tales, cosmological narratives – that predate modern language families, religious boundaries and institutional structures. Michalopoulos and Xue (2021) show that folklore similarity predicts trust, trade and a range of economic outcomes. Coverage: 12,561 pairs (90 percent).

Historical trade. TRADHIST version 4 (Fouquin and Hugot, 2016) provides bilateral trade flows from 1827 to 2014. I aggregate flows into three periods: pre-1870 (capturing the first era

of globalisation and the twilight of the slave trade), colonial era (1870–1945, spanning the height of European imperialism) and post-war (1946–2014, covering decolonisation and modern globalisation). For each pair-period, I compute the log of average bilateral trade. Historical trade is an important control because countries with shared ancestry may also trade more, and trade itself generates social connections through business relationships, supply chain interactions and associated migration. Coverage: 11,966 pairs with data in at least one period (85 percent).

Contemporary migrant stocks. The United Nations bilateral migrant stock matrix (United Nations Department of Economic and Social Affairs, 2024) records the number of foreign-born residents from each origin country in each destination country as of mid-year 2020. Because the country-pair regressions are undirected, I symmetrise the bilateral matrix by taking the maximum of the two directed stocks, $\max\{\text{stock}_{i \rightarrow j}, \text{stock}_{j \rightarrow i}\}$, and then construct $\log(1 + \text{migrant stock}_{ij})$. This distinguishes the ancestry-SCI association from the mechanical effect of contemporary diasporas maintaining Facebook friendships with homeland contacts.

M.2 American Community Survey ancestry data

For the county-level analysis, I combine county-level ACS 2018–2022 five-year extracts for European and African ancestries from Table B04006, detailed Asian-origin groups from Table B02015, and Mexican origin from Table B03001 (US Census Bureau, 2022). I convert each count to a county share of total population across 3,222 US counties. English and Scottish are combined as “British” for the UK regression. The extraction also includes Chinese ancestry, but China is omitted from the reported regressions because the county SCI extract does not contain a China partner series, leaving nine estimable origin groups.

Two limitations deserve mention. First, many Americans of British descent report “American” rather than a specific European origin, particularly in the South and Appalachia where British settlement was earliest and assimilation most complete. This downward bias in British ancestry shares works against finding a significant UK result and makes the significant coefficient ($\beta = 0.040$) a conservative estimate. Second, ACS ancestry is self-reported and reflects subjective ethnic identification rather than verified genealogy. People may report an ancestry they identify with culturally rather than one that reflects their actual demographic descent. To the extent that contemporary cultural identification drives both ancestry reporting and Facebook friendships, the estimated coefficient may overstate the role of demographic history and instead partly reflect current ethnic identity. In that case, the concern is confounding rather than corroboration.

M.3 South Africa district-level data

For the South Africa analysis, I combine four data sources at the level of 52 GADM2 administrative districts.

Population composition. The 2022 South African Census (Statistics South Africa, 2024) reports population by group (Black African, Coloured, Indian/Asian, White) at the district level. I compute the share of each group in total district population. The White share ranges from 0.2% (rural

Eastern Cape) to 20.4% (parts of Western Cape/Gauteng), with a mean of 6.7%. The Coloured population – descended substantially from Cape slaves, Khoesan and mixed-origin populations – averages 13.5% nationally but is concentrated in the Western and Northern Cape. I therefore use it only as a rough demographic proxy, not as a direct measure of slave-origin ancestry. The Indian/Asian population (mean 1.6%) is concentrated in KwaZulu-Natal, reflecting indentured labour migration from India between 1860 and 1911, and provides the baseline proxy for the indentured-migration margin.

Indian indentured records. I additionally draw on the digitised Natal indentured records of Waetjen and Vahed (2014). The linked extract contains 152,273 individual records across nine workbooks, of which 89,027 can be assigned to South African districts using employer-location strings. The matched records are concentrated in six KwaZulu-Natal districts. The records include the migrants’ origins within India and the South African district to which they were assigned or first observed. This source allows the South Africa section to move beyond the coarse modern %Indian proxy toward a district-by-origin measure of semi-voluntary migration that is more comparable to the direct slave-origin data used for the Cape Colony. But the district assignment remains incomplete: 63,246 records are not yet placed, and the available linkage does not establish whether those unmatched records differ systematically by origin region or arrival cohort. I therefore treat the historical indentured totals as a useful but still provisional robustness check rather than as a fully representative district census of indentured migration. Throughout, I use the modern population share as the baseline specification and report the historical indentured totals as a robustness check; the historical coefficient is positive and marginally significant, but still less precisely estimated than the modern district-level proxy.

Home language. From the Census 2022 individual-level data (10% sample, 4.2 million records; Statistics South Africa, 2024), I aggregate the share of each district’s population speaking each of the 11 official languages plus Shona, Chichewa and Portuguese. I construct a binary *common language* indicator that equals one when more than 5% of a district’s population speaks a language historically linked to the partner country: Afrikaans to the Netherlands, English to the United Kingdom and Ireland, Portuguese to Portugal and Mozambique, Sesotho to Lesotho, Chichewa to Malawi and Siswati to Eswatini. The underlying language-country crosswalk also maps Shona to Zimbabwe, but Zimbabwe is not part of the preferred African-placebo set reported in Table 5.

Urbanisation controls. To check whether the White population share is merely a proxy for district-level wealth and international connectivity rather than ancestry per se, I construct urbanisation controls from the 2022 Census household file (10% sample, 1.34 million records; Statistics South Africa, 2024). For each of the 52 GADM2 districts, I compute the weighted share of households with internet access (mean 60.6%, range 41.3–73.2%), the share in formal dwellings (mean 84.6%) and the share classified as urban (mean 58.3%). Internet access is the main control discussed in the text; formal dwellings and urban share are used in the appendix robustness checks, including the slave-proxy sensitivity analysis. The correlation between the White population share and district-level internet access is low ($r = 0.19$), indicating limited collinearity between the two

variables.

M.4 Cape Colony historical data

For the Cape Colony deep dive I exploit two historical datasets covering 14 modern districts that fall within the boundaries of the colonial Cape.

Slave Emancipation Dataset. The slave emancipation dataset of Ekama et al. (2021) contains 36,417 individual records from the 1834 emancipation. Each record reports the enslaved person’s recorded origin, the owner’s surname and – for 77% of records – a place name that typically corresponds to a farm. I use this dataset for two distinct purposes. First, I match the place names to the farm boundaries first published by Fourie (2022) to geolocate enslaved individuals within modern GADM2 districts, using historical district boundaries to disambiguate farms with common names. This yields 914 matched records across 12 districts, with recorded origins mapped to modern countries: Mozambique (730 individuals, 80%), India (63), Madagascar (53), Indonesia (43) and smaller numbers from Sri Lanka, Mauritius, Algeria and Malaysia. The resulting district-level origin shares provide a direct historical measure of the geographic composition of forced migration into each Cape Colony district, but the usable variation is heavily concentrated in Mozambique; partner-specific shares for the smaller origins are sparse and often zero in most districts. Second, I extract owner surnames from the same dataset and link them to a Huguenot surname dictionary and to a parsed 1820-settler surname list, building district-level shares of slaveholders carrying Huguenot or 1820-settler British surnames. These surname measures are the paper’s preferred historical settlement variables and are used in Table 6 of the main text.

Farm boundaries. The boundary file from Fourie (2022) contains 5,503 farm-level polygons from the Cape Colony circa 1850. I use these polygons primarily as a spatial backbone for matching the slave emancipation records to modern districts. A robustness check in Appendix N additionally classifies each farm name by linguistic origin – Dutch, English, French (Huguenot) or German – using a rule-based classifier built on toponymic suffixes and place-name markers. Dutch names dominate (96%), with English names concentrated in the 1820 settler districts of the Eastern Cape and French names in the Cape Winelands. The validation table in that appendix shows that the classifier captures naming conventions imperfectly; farm names are therefore reported only as a robustness check, with the surname measures preferred in the main text.

M.5 New Zealand and Australia data

The Australasian evidence in Section 7 of the main paper draws on three sources. The 1881 New Zealand Census records the birthplace of every non-Maori resident at borough level. I aggregate 65 borough-level birthplace tabulations to 29 modern GADM2 districts using a manual borough-to-district crosswalk, and to 13 GADM1 regions for the descriptive figures. The crosswalk relies on the historical boundaries of provinces and counties at the time of the 1881 enumeration.

The 1901 Australian Census tabulates colony-level birthplace for New South Wales, Victoria, Queensland, South Australia and Tasmania, drawn from the Historical and Colonial Census Data

Archive (Australian Bureau of Statistics et al., 2019). Western Australia is excluded because comparable colony-level birthplace data are not available in the working sample. Because birthplace is recorded at the colony level, the 503 GADM2 districts within each Australian colony share identical birthplace shares; the Australian observations therefore enter the pooled NZ + AU regression aggregated to five colony-level cells.

Convict-transportation totals are taken from Anderson and Maxwell-Stewart (2014), Cowley et al. (2023) and Nicholas (1988). For each colony I record total convicts transported between 1788 and 1868: approximately 83,000 to New South Wales (1788–1850, roughly 500 of them transported from other British colonies), 73,500 to Tasmania (1803–1853, from the Tasmania-specific reconstruction of Cowley et al., 2023, which differs from the Anderson–Maxwell-Stewart figure by about 1,000), 9,700 to Western Australia (1850–1868, outside the working sample), 3,000 “Pentonvillians” to Port Phillip (modern Victoria, 1846–1850), and zero to South Australia, founded in the 1830s explicitly as a free colony. Queensland is coded with zero primary transportation: its only penal episode was the secondary settlement at Moreton Bay (1824–1839), which received transportees reconvicted elsewhere in Australia rather than a direct transportation flow. I divide each colony’s convict total by its 1901 population to obtain a convict-intensity measure, used both as a continuous covariate and to assign each colony to a regime category (*penal* for NSW and Tasmania; *free* for South Australia, Victoria and Queensland). New Zealand regions are coded with `convict_share` = 0.

N. Cape Colony: farm-name robustness

The main paper leads the Cape Colony historical evidence with owner-surname measures from the slave emancipation records of Ekama et al. (2021), which can be linked directly to historically Huguenot or 1820-settler British surnames. As an alternative proxy for settler origin, I classify the Cape Colony farm boundaries first published by Fourie (2022) as Dutch, English, French (Huguenot) or German on the basis of toponymic suffixes and known place-name markers. Farm names are a noisier proxy for settler origin than owner surnames because Dutch toponymic conventions dominated the colony for nearly two centuries, so even British and French settlers often farmed land carrying Dutch names. The validation table in Section E (Table A8) makes this explicit: among uniquely matched Huguenot-associated farms, two-thirds are still classified as Dutch by the rule-based algorithm.

For transparency, Table A18 reports the farm-name specifications. The English farm-name share is significantly associated with European SCI; the French farm-name share is also positive but smaller. Both measures point in the same direction as the surname results in Table 6 of the main paper, but the surname measures attach attribution more cleanly to specific settler waves. The farm-name regressions are reported here for readers who wish to verify that the surname-led conclusion is not an artefact of the new measurement strategy.

Table A18: Cape Colony: farm-name robustness specifications

	(F1) Baseline	(F2) +Eng	(F3) +Fr	(F4) +Both	(F5) Kitchen
<i>Dep. var.: log(SCI)</i>					
% White	0.785*** (0.027)	0.818*** (0.028)	0.775*** (0.026)	0.806*** (0.028)	0.654*** (0.031)
English farms		0.045*** (0.011)		0.040*** (0.012)	0.023. (0.012)
French farms			0.021** (0.006)	0.017* (0.006)	0.012* (0.006)
Slave origins					−0.027**† (0.010)
Partner FE	Yes	Yes	Yes	Yes	Yes
Within R^2	0.789	0.806	0.800	0.813	0.692
N	196	196	196	196	294

Notes: Dependent variable: log(SCI). All specifications include log(pop) and common language. F5 pools EU and slave-origin partners (294 obs). †As in Table 6 of the main paper, the F5 slave-origin coefficient reflects a single pooled log(%White) slope across both partner groups and is a pooling artifact; the corrected slave-origin-only estimate is −0.012 (SE 0.007, $p = 0.11$; see main-text Section 8.3). English/French farm-name shares classified from the 5,503 farm boundaries first published by Fourie (2022). Slave origins from the slave emancipation dataset of Ekama et al. (2021). Heteroskedasticity-robust standard errors. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$.

O. US Counties: Homeland Specificity

The placebo matrix in Appendix I shows that county ancestry shares are collinear, so a single-ancestry regression could in principle load on a broad immigrant or gateway-county factor rather than a focal homeland tie. Table A19 addresses this directly. For each origin I re-estimate the homeland-SCI regression (i) jointly controlling for every other tracked ancestry share, and (ii) on a residualised “excess ancestry” measure – the focal log-ancestry orthogonalised against all the others. In both specifications the focal coefficient remains positive and significant for all nine origins. The point estimates are close to the single-ancestry baselines (e.g. Ireland 0.267 joint, 0.189 excess; Mexico 0.575/0.517; the UK 0.179/0.137), confirming that each ancestry predicts its *own* homeland’s connectedness over and above the correlated ancestry bundle.

Table A19: US counties: focal-ancestry coefficient controlling for correlated ancestry bundles

Origin	N	Joint multi-ancestry	Residualised excess
United Kingdom	3,145	0.179***	0.137*
Germany	3,150	0.171***	0.115**
Ireland	3,140	0.267***	0.189*
Italy	3,121	0.166***	0.130***
Mexico	3,133	0.575***	0.517***
Poland	3,007	0.149***	0.117***
France	3,127	0.081***	0.077**
India	1,893	0.219***	0.190***
Philippines	2,367	0.172***	0.155***

Notes: Dependent variable: log(SCI) with the matched homeland. “Joint multi-ancestry” adds log shares of all other tracked ancestries; “residualised excess” regresses homeland SCI on the focal log-ancestry orthogonalised against the others. Both include log(pop) and state fixed effects, with standard errors clustered at the state level. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

P. South Africa: European-Distance Robustness

With partner fixed effects, the European %White specifications identify distance only from the within-South-Africa gradient to each European capital. Although that gradient is small relative to the common 9,000–10,500 km baseline, Table A20 shows the %White result is essentially unchanged when district-to-partner distance is nonetheless included: the baseline coefficient moves from 0.578 to 0.570, and the kitchen-sink coefficient from 0.404 to 0.398. The European result does not depend on the omission of the distance control.

Table A20: South Africa: %White coefficient with and without district-to-Europe distance

	Baseline (S1)	Kitchen sink (S3)
log(%White), no distance	0.578***	0.404***
log(%White), + log(dist)	0.570***	0.398***

Notes: European partner sample, 14 partners. All specifications include log(pop) and partner fixed effects; the kitchen sink adds log(%Coloured), log(%Indian) and common language. Heteroskedasticity-robust standard errors. *** $p < 0.001$.

Q. Power of the Slavery Null

The slave-origin baseline is a null, so its interpretation depends on the power of the design. The estimated standard error on the slave-origin coefficient is 0.008, implying a minimum detectable effect of about 0.023 at 80% power and 5% size. The design could therefore have detected a

positive slave-origin elasticity an order of magnitude smaller than the voluntary-settler benchmark (%White $\rightarrow$ Europe, 0.578) and well below the indenture benchmark (%Indian $\rightarrow$ India, 0.078). The null is thus informative: it rules out homeland ties of the magnitude that free and semi-free migration generate, even though the matched slave-origin signal is concentrated in 39 non-zero district $\times$ partner cells, dominated by Mozambique.

R. South Africa: Partner-Specific Within-District Z-Scoring

A district that is simply more internationally connected overall would show a high SCI with every partner. To separate the ancestry-specific signal from this generic-connectivity profile, I z-score $\log(\text{SCI})$ across partner countries *within* each district – removing each district’s own level of international connectedness – and re-run the three regime regressions on the z-scored outcome. The European premium survives strongly (%White $\rightarrow$ Europe = 0.152, $p < 0.001$); the African-neighbour placebo shrinks to marginal significance (%White $\rightarrow$ Africa = 0.028, $p = 0.09$); and the slave-origin coefficient is a clean null (%Coloured $\rightarrow$ slave origins = -0.008 , $p = 0.33$). Removing the district’s general connectivity level therefore sharpens rather than erases the regime contrast: the European-ancestry tie is partner-specific, not a by-product of a broadly high-connectivity demographic profile.

S. Additional Robustness Checks

This section reports the full versions of the robustness checks summarised in Section 9 of the main paper. The PPML, Park-test, Gamma and variance-inflation diagnostics are in Appendices A and K; the bilateral air-traffic and Facebook-penetration checks are in Appendix L.

Permutation test. A naive permutation that shuffles the ancestry variable independently across country pairs 1,000 times places the actual coefficient (0.144) 11.1 standard deviations above the mean of the permutation distribution (mean = 0.0004, SD = 0.013). Because dyadic data exhibit node-level dependence, this edge-level scheme could in principle understate the null variance. I therefore also run a network-preserving (node-label) permutation in the spirit of the quadratic-assignment procedure: rather than shuffling individual dyads, I relabel the countries and reassign the rows and columns of the ancestry matrix together, preserving each country’s full vector of ancestry ties. The resulting null is essentially the same width (SD = 0.013), and the actual coefficient remains 10.8 standard deviations above the node-permutation mean ($p < 0.001$; none of the 1,000 relabelings reaches the actual value). The two schemes agree because the two-way country fixed effects already absorb the node-level dependence that the node permutation is designed to preserve. The permutation test and the two-way clustered standard error (0.031) answer different questions – whether the observed network alignment could arise by chance, versus sampling uncertainty – and the paper relies on the clustered standard error for inference.

Spatial trends and the critical persistence literature. A growing body of work questions the statistical foundations of persistence studies. Conley and Kelly (2025) show that smooth spatial trends and autocorrelation can generate spuriously large t -statistics in cross-country regressions, and that 29 of 30 prominent persistence studies lose significance under their spatial-basis diagnostic. This paper’s identification strategy is largely insulated from that concern. The county-level analysis identifies from within-state variation in ancestry composition, not from cross-country spatial gradients; state fixed effects absorb the broad regional trends that drive Conley and Kelly’s critique. The New Zealand and South Africa exercises similarly exploit within-country variation, holding the national institutional and geographic environment broadly constant. The country-level horse race is more exposed to spatial confounding, which is one reason I treat it as descriptive background rather than the paper’s sharpest evidence. Arroyo Abad and Maurer (2021) identify five common pitfalls in the persistence literature: failure to recognise institutional change, vague mechanisms, insufficient use of historical sources, compression of history, and failure to account for geography. The mechanism here – social infrastructure transmitted through diaspora networks – is specific rather than residual, and the subnational analyses use dated historical sources (the 1881 New Zealand Census, the 1834 Cape Colony emancipation records, 1850 farm polygons) rather than compressing centuries of history into a single treatment indicator. Voth (2021) draws a useful distinction between “apples-and-oranges” persistence studies, which link dissimilar historical and modern variables, and “apples-on-apples” studies that track the same class of outcome over time. The county and New Zealand exercises are closer to the latter: historical demographic composition predicts a modern social outcome that is conceptually related to the migration event itself.

Alternative dependent variables. The ancestry coefficient is stable across alternative transformations of the SCI. Using $\text{asinh}(\text{SCI})$ yields $\beta = 0.144$; using $\log(1 + \text{SCI})$ yields $\beta = 0.144$; using the percentile rank of SCI yields $\beta = 0.016$ ($p < 0.001$). Together with the winsorisation and influential-observation checks, these results suggest that the association is not an artefact of a particular outcome transformation or a handful of extreme pairs.

Winsorisation. Winsorising the ancestry variable at the 99th percentile slightly increases the coefficient (from 0.144 to 0.163), indicating that the result is not driven by extreme outlier pairs.

Influential observations. I compute Cook’s distance for all observations in the baseline specification. The most influential pair is Portugal-São Tomé and Príncipe (Cook’s $d = 0.018$), a pair with zero shared ancestry but very high SCI, probably reflecting ongoing postcolonial ties not captured in the Putterman matrix. Dropping the top 20 influential observations has negligible effects on the ancestry coefficient.

Out-of-Africa migratory distance. As an additional robustness check, I replace the detailed F_{ST} measure with the coarser deep-history proxy of Ashraf and Galor (2013): bilateral proximity in migratory distance from Addis Ababa. Adding this variable to the country fixed-effects specification

leaves the ancestry coefficient virtually unchanged ($\beta = 0.149$, $SE = 0.031$), while the migratory-distance proxy itself is small and statistically insignificant ($\beta = -0.003$, $SE = 0.006$). Adding the Ashraf-Galor predicted-diversity controls does not alter this conclusion: the ancestry coefficient remains positive and significant ($\beta = 0.164$, $SE = 0.031$), and the migratory-distance term remains imprecisely estimated. The coarse Out-of-Africa proxy therefore does not subsume the ancestry channel in the way that the richer F_{ST} measure partially does.

Inference. All country-level specifications use standard errors clustered on the two country identifiers comprising each undirected pair, following Egger and Tarlea (2015) and Pfaffermayr (2023). In the data, each dyad is stored once with alphabetically ordered country indices, so these are not directional origin and destination roles. This approximately doubles standard errors relative to pair-level clustering – for example, the baseline ancestry standard error increases from 0.017 to 0.031 – but does not affect the significance of any key result. County-level specifications use standard errors clustered at the state level. For the small-sample New Zealand regressions, I report permutation p -values alongside asymptotic inference; the district-level UK-born coefficient remains significant asymptotically but has a weaker permutation p -value of 0.143, and the stricter stacked district-and-partner fixed-effects test yields a permutation p -value of 0.944. For the Cape Colony slave-origin coefficient, the strongly negative pooled kitchen-sink estimate is a pooling artifact (Section 8.3); the corrected slave-origin-only coefficient (-0.012) is small and only marginally distinguishable from its randomisation null (two-sided permutation $p = 0.034$).

Alternative ancestry constructions. The results are robust in sign, though not equally precise, across alternative constructions of the ancestry variable. The raw maximum (`anc_max`) yields a coefficient of 1.546 with country fixed effects, and the bilateral sum (`anc_sum`) performs almost identically, as expected given that 99.7 percent of pairs have non-zero ancestry in at most one direction. But both raw constructions are only marginally estimated with the reported clustered standard errors, whereas the log transformation remains precise and strongly significant. The product construction, which is zero whenever either directional share is zero, fails entirely – only 39 pairs have reciprocal non-zero ancestry. These results confirm that the log transformation, by converting the extreme right skew of the raw variable into a more symmetric distribution, is the appropriate functional form for this analysis.

Table A21: Alternative ancestry variable constructions

	(1) Max	(2) Sum	(3) Product	(4) Log
anc_max	1.546 (0.940)			
anc_sum		1.549 (0.939)		
anc_prod			194.7 (282.0)	
anc_log				0.144*** (0.031)
Country FE	Yes	Yes	Yes	Yes
Within R^2	0.558	0.558	0.557	0.560
N	14,026	14,026	14,026	14,026

Notes: Dependent variable: $\log(\text{SCI}_{ij})$. All specifications include gravity controls and country fixed effects. Standard errors clustered on the two country identifiers defining each undirected pair. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

T. Limitations

Several caveats deserve emphasis. First, the analysis is descriptive. Although the identification toolkit builds a strong circumstantial case against spuriousness, the results should be interpreted as robust associations, not causal effects.

Second, the Putterman matrix is an estimate, not a census. It relies on historical demographic sources that vary in quality across countries and may not fully capture certain migration flows, particularly within Africa and Asia. This concern is especially relevant for forced-migration settings, where origin labels are coarser, archival survival is uneven, and descendant populations were deliberately separated from origin identities. The Cape slave-origin exercise should therefore be read together with the broader Atlantic corridor results rather than as a complete origin-by-origin test.

Third, ACS ancestry data are self-reported. Self-reported ancestry is influenced by social desirability, ethnic salience and generational distance from the migration event. The tendency of many British-descent Americans to report “American” ancestry biases the UK coefficient downward but does not threaten the validity of the overall pattern.

Fourth, Facebook penetration varies across countries, and several major populations – notably China and Russia – maintain their primary social ties on closed platforms (WeChat, VK) for which no equivalent bilateral connectedness index has been published. Country fixed effects absorb cross-country variation in penetration levels, and the penetration-restricted subsamples in Section 9

confirm that the result is not driven by very-low-penetration country pairs. But it would be misleading to claim that the SCI is a global measure of social connectedness: it is a measure of connectedness on the world's largest open social network, and the closed-platform analogues remain unobserved.

Fifth, the SCI measures the intensity of Facebook friendship links, not the content of those links. It cannot distinguish family ties from tourism, study, work, recent migration or other forms of contact. More direct measures of cultural affinity – shared interests, group membership or page likes – would be useful, but are not part of the public SCI release used here.

Sixth, the SCI is not a time series. Meta advises against comparing SCI across different data releases because methodology changes over time. The analysis is cross-sectional and cannot speak to trends in the ancestry-connectedness relationship.

U. Supplementary Figures

This section collects supplementary figures referenced in the main text. Country-level descriptive and robustness figures appear first, followed by the expanded US county figures, the New Zealand birthplace composition, and the South African per-partner figures.

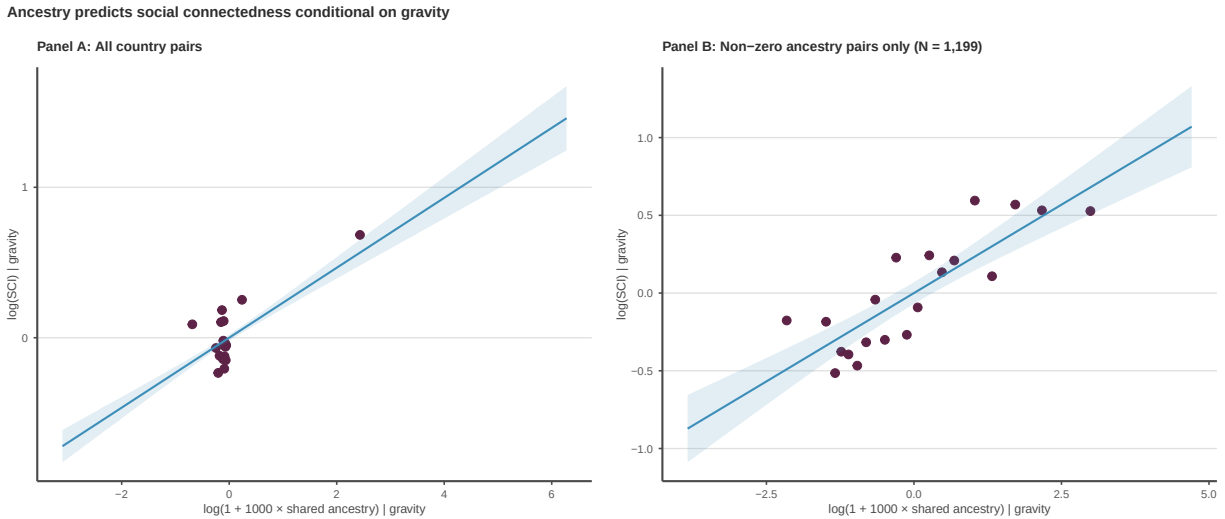


Figure A2: Shared ancestry and social connectedness: binned scatter plot. Each dot represents the mean residual $\log(\text{SCI})$ for a vigintile of residual anc_log after partialling out the gravity controls used in the baseline specification (log distance, contiguity, common language and colonial ties), but not country fixed effects.

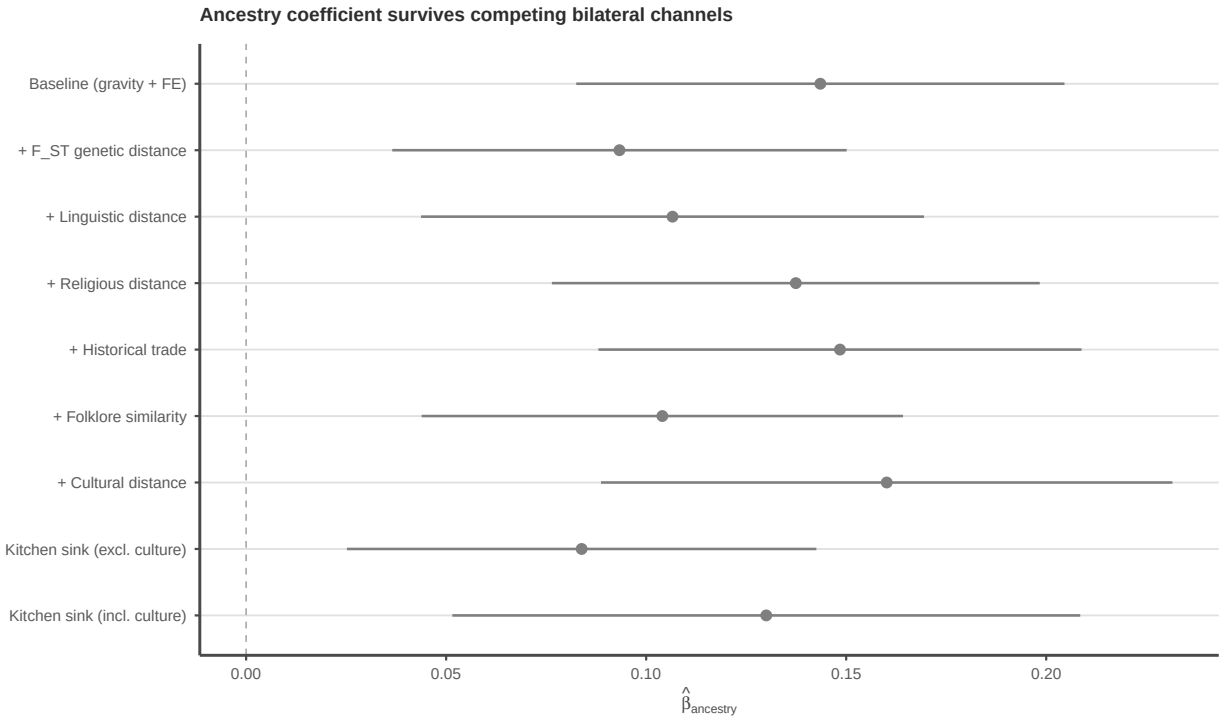


Figure A3: Horse race: ancestry versus competing bilateral channels. Forest plot of ancestry coefficients when each channel is added individually (left) and in kitchen-sink specifications (right).

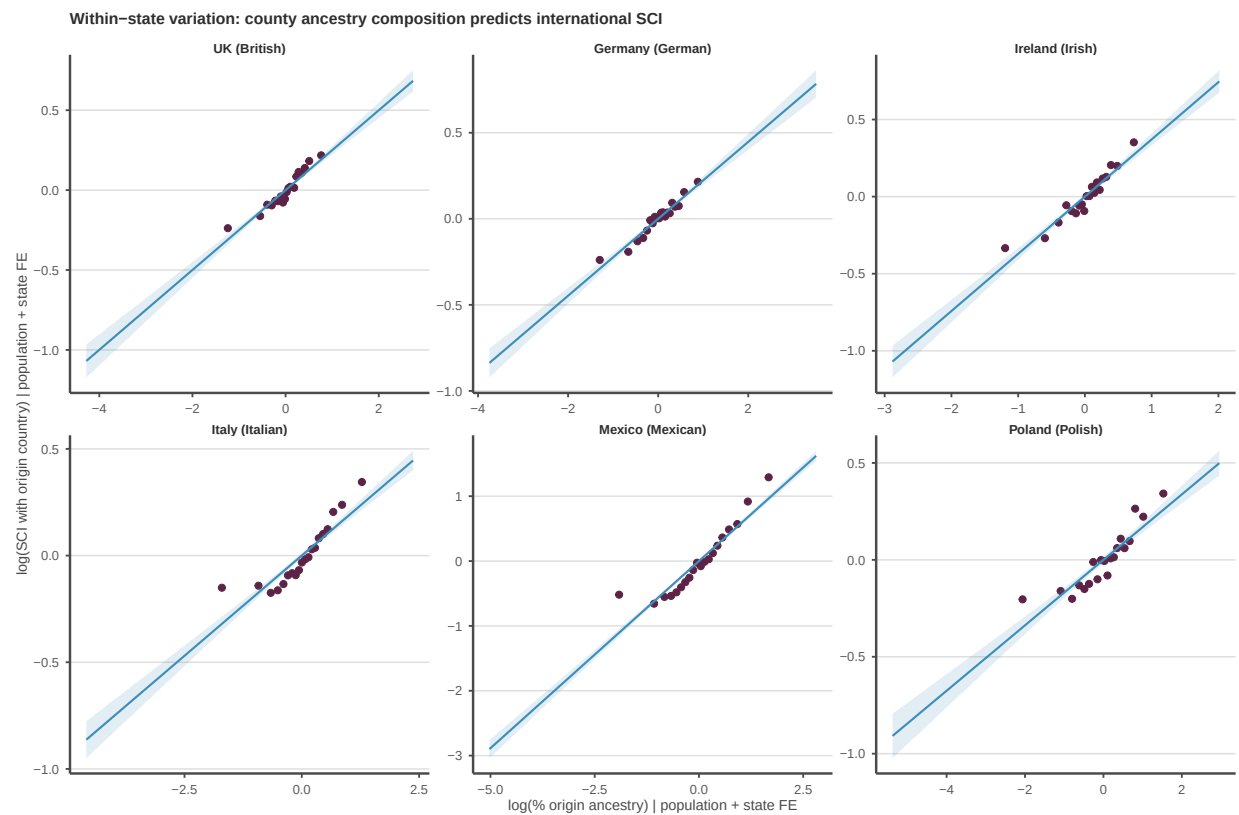


Figure A4: US county-level: ancestry share and SCI with corresponding homeland, six origin groups.

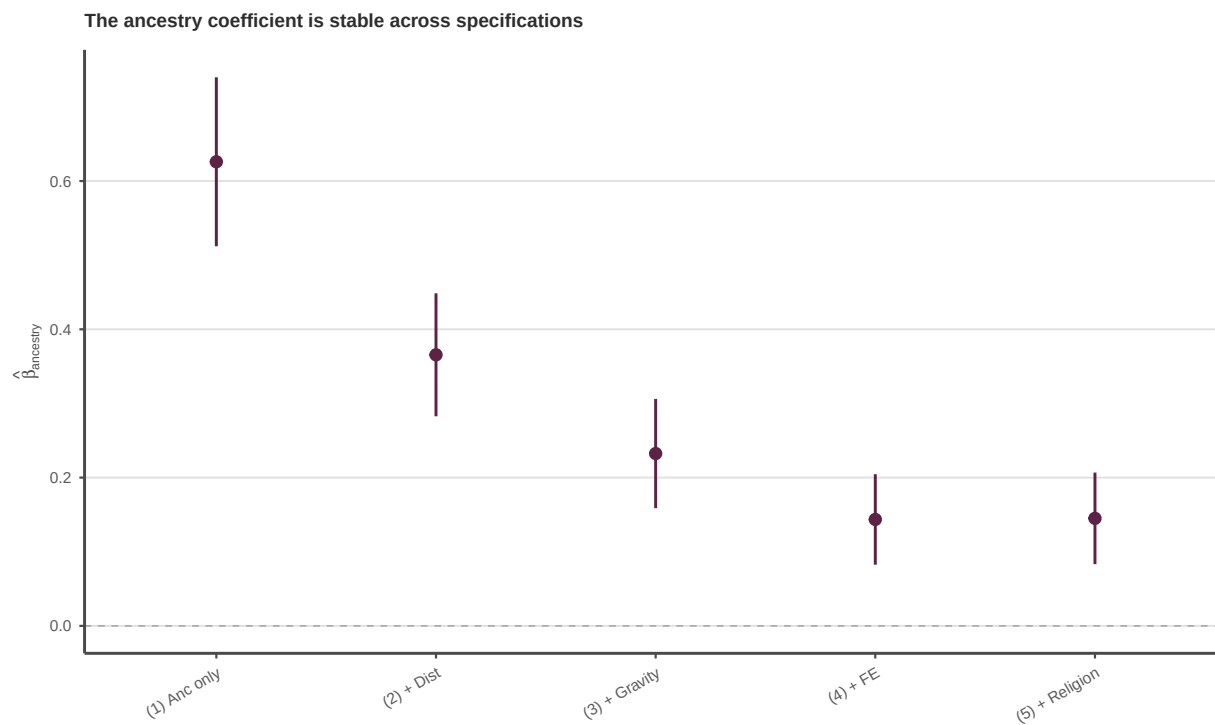


Figure A5: Coefficient stability: ancestry coefficient under sequential addition of controls.

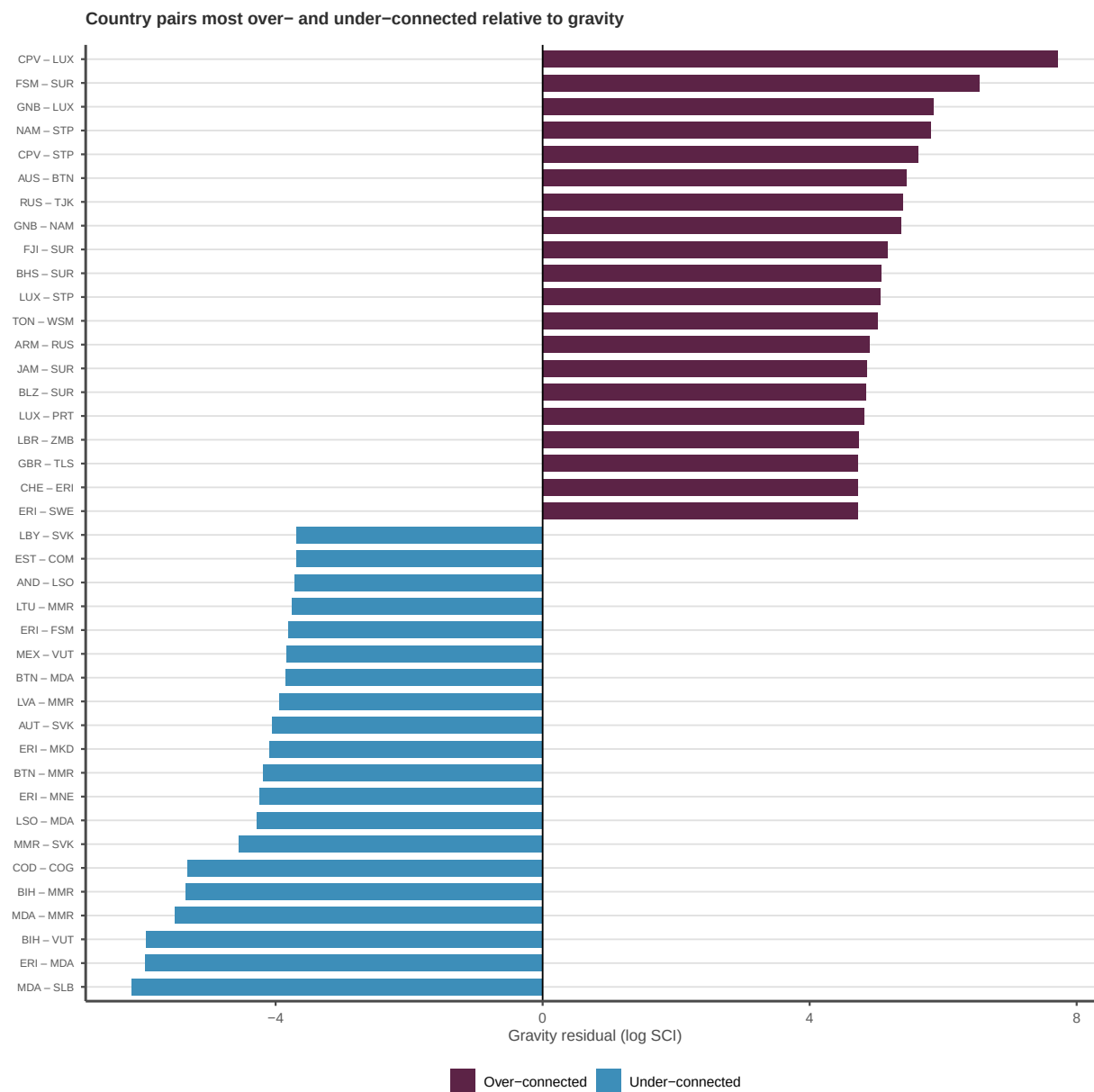


Figure A6: Most over- and under-connected country pairs relative to the gravity model.

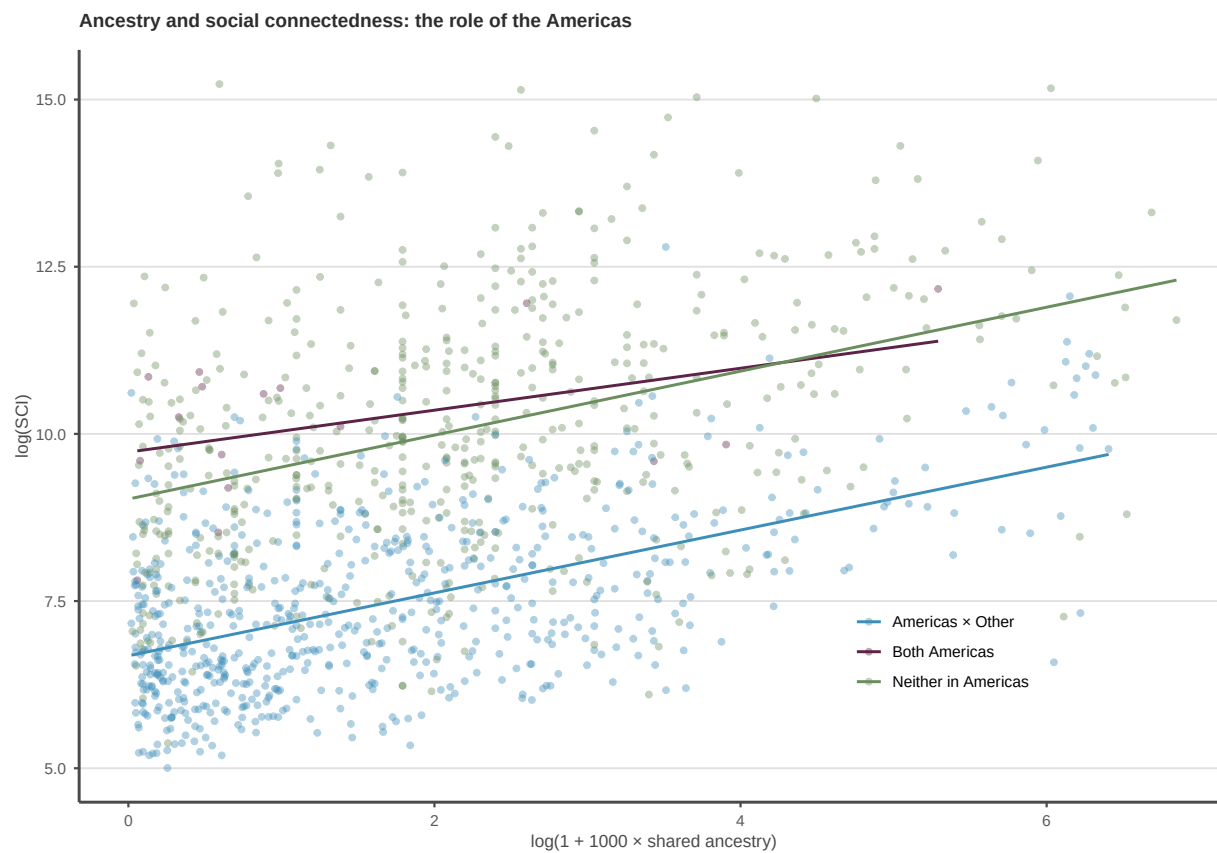


Figure A7: Ancestry and SCI in the Americas subsample.

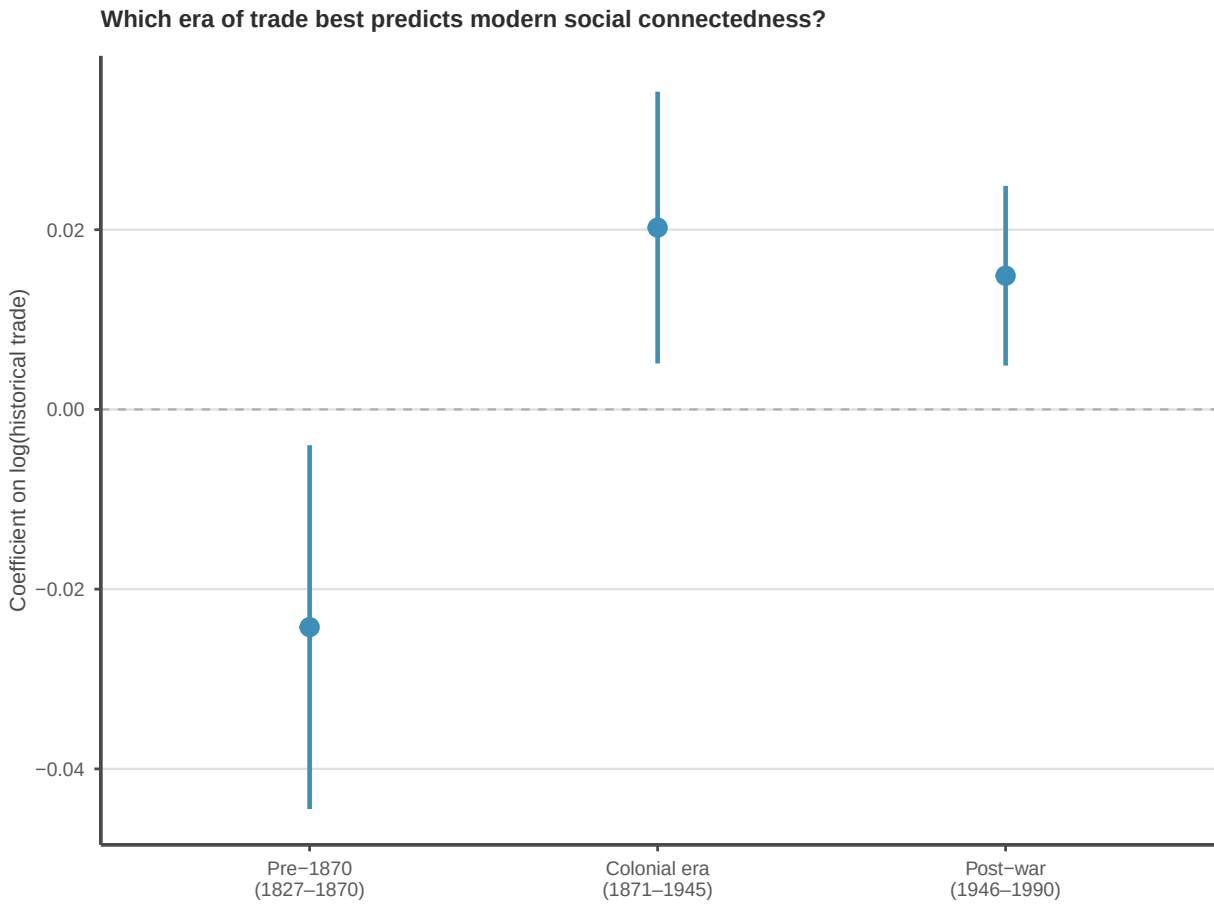


Figure A8: Historical trade as a competing channel: trade coefficients by period.

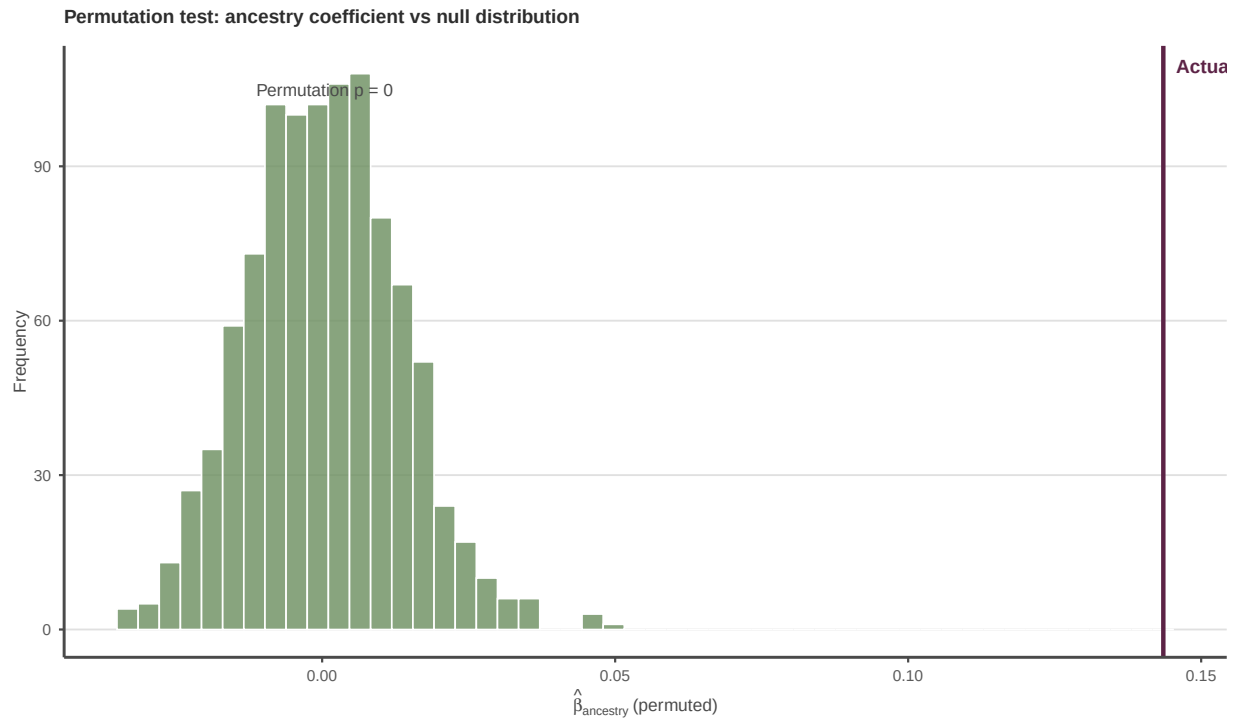


Figure A9: Permutation test: distribution of ancestry coefficients under 1,000 random permutations. The vertical line marks the actual coefficient.

Ancestry coefficients across all specifications

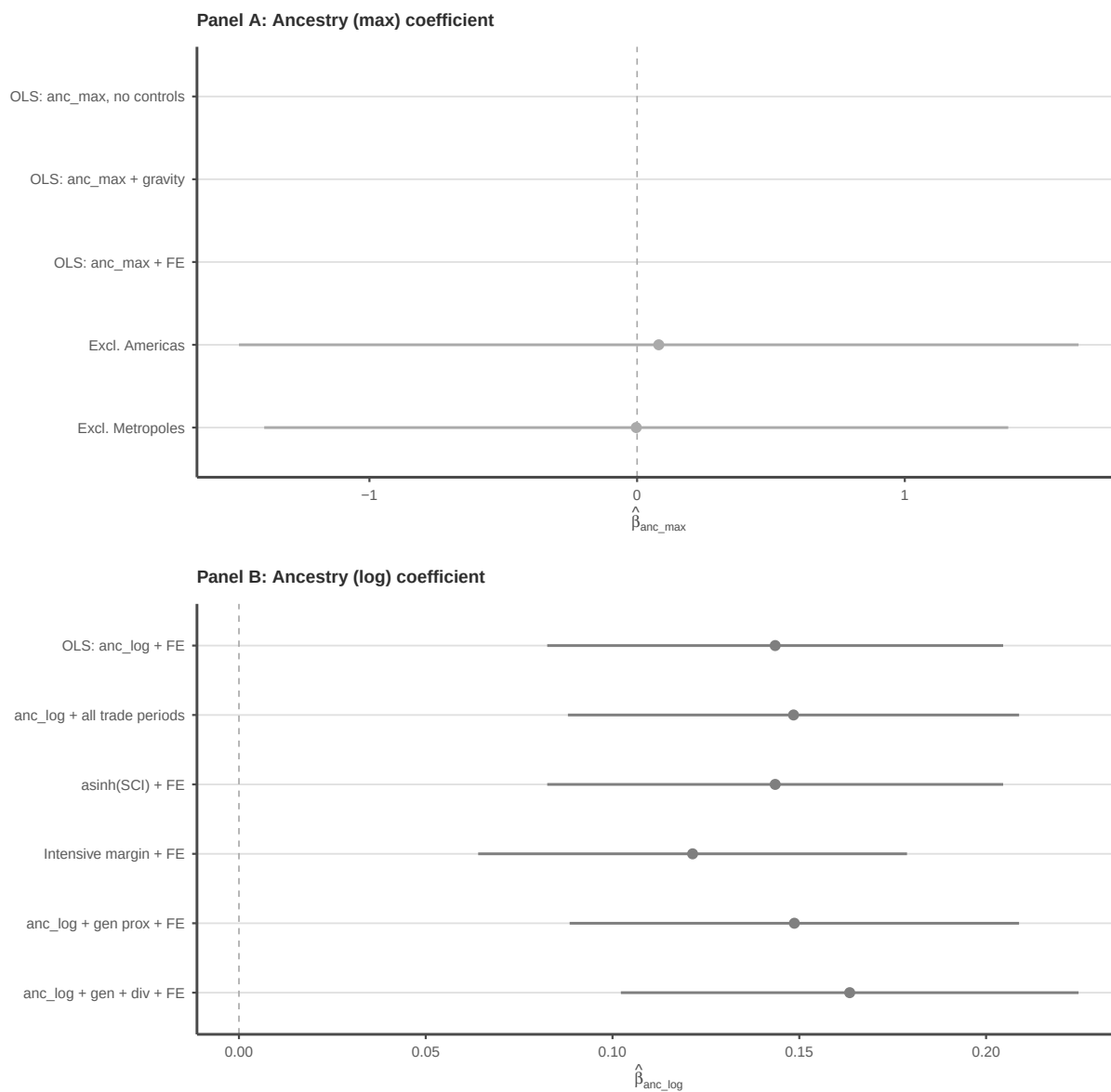


Figure A10: Summary forest plot of ancestry coefficients across specifications.

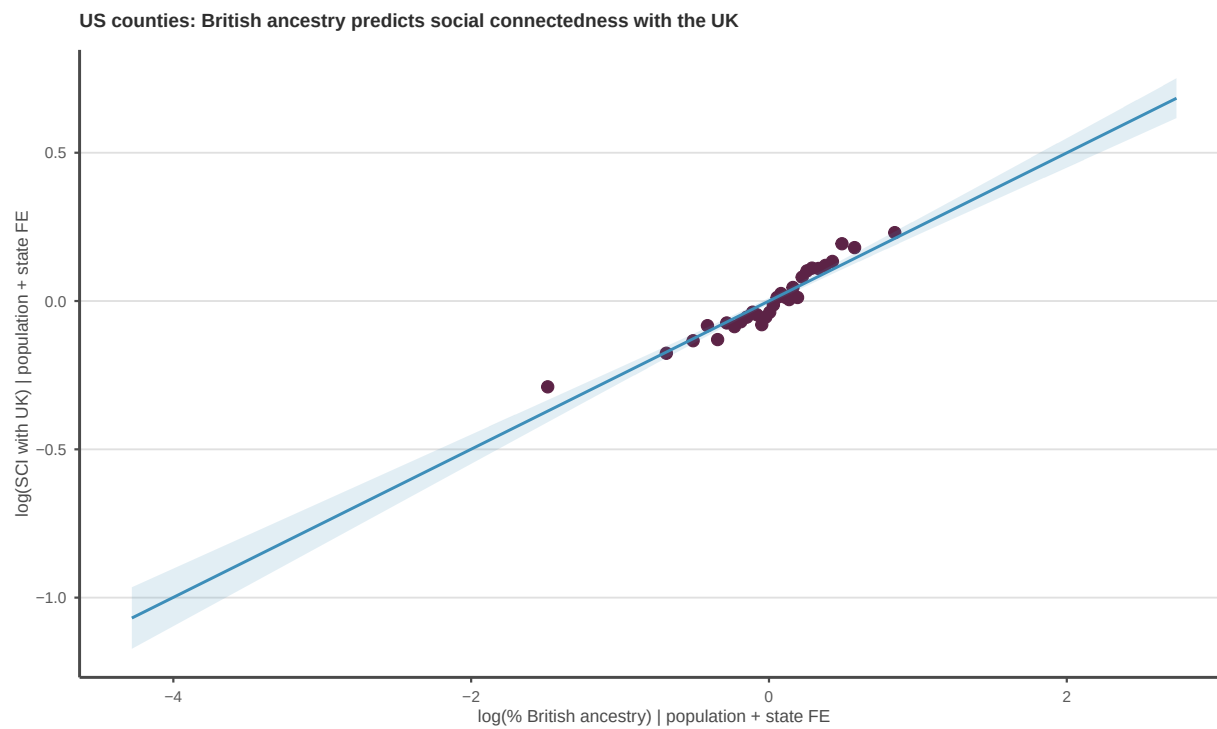


Figure A11: US county-level: British ancestry share and SCI with the United Kingdom.

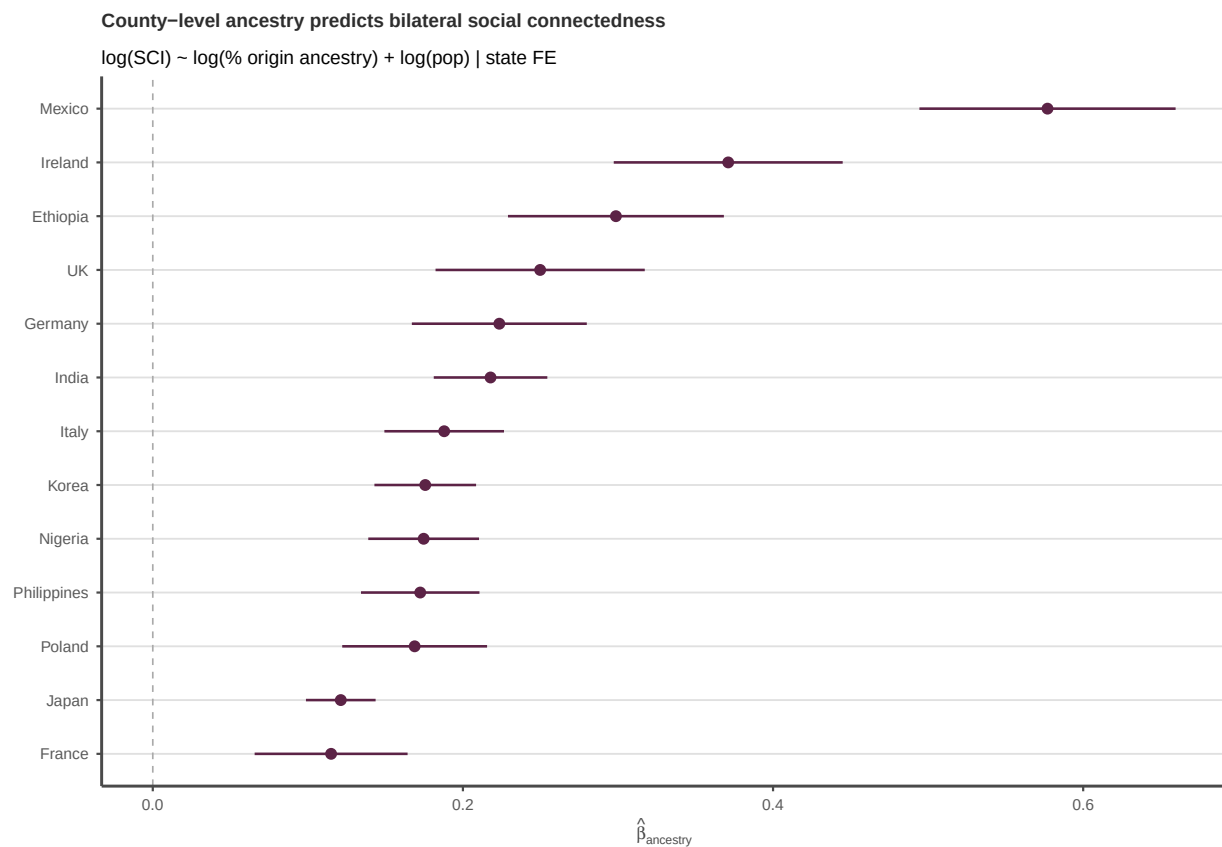


Figure A12: US county-level: forest plot of ancestry coefficients for 13 estimated origin groups in the expanded sample.

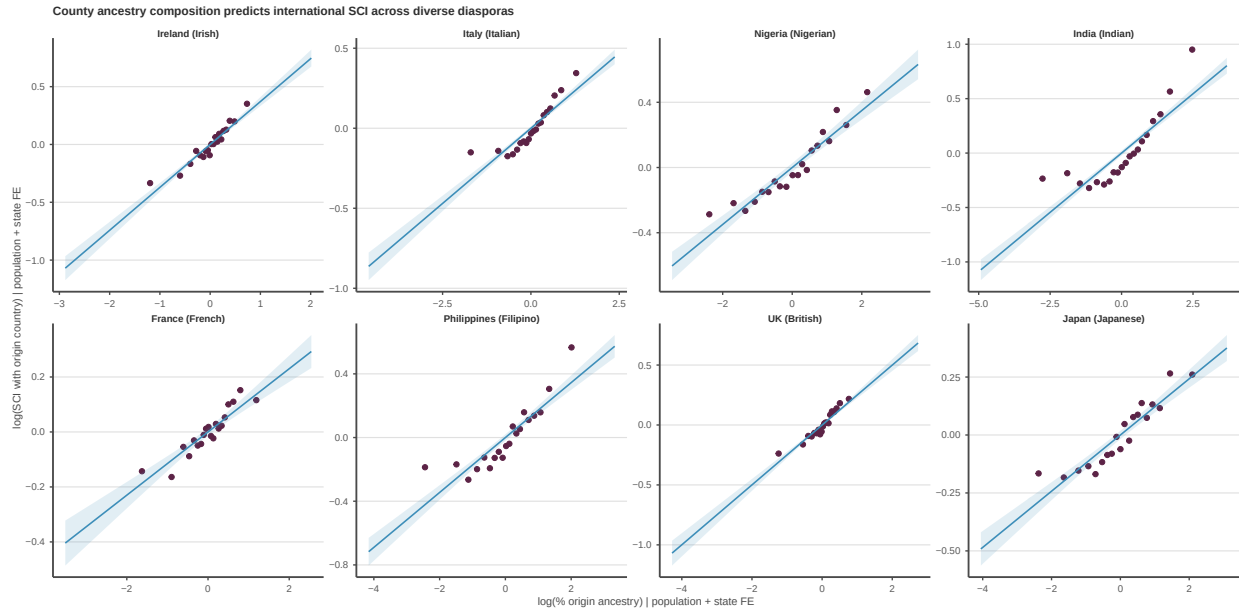


Figure A13: US county-level: ancestry share and SCI for eight selected origin groups from the expanded sample.

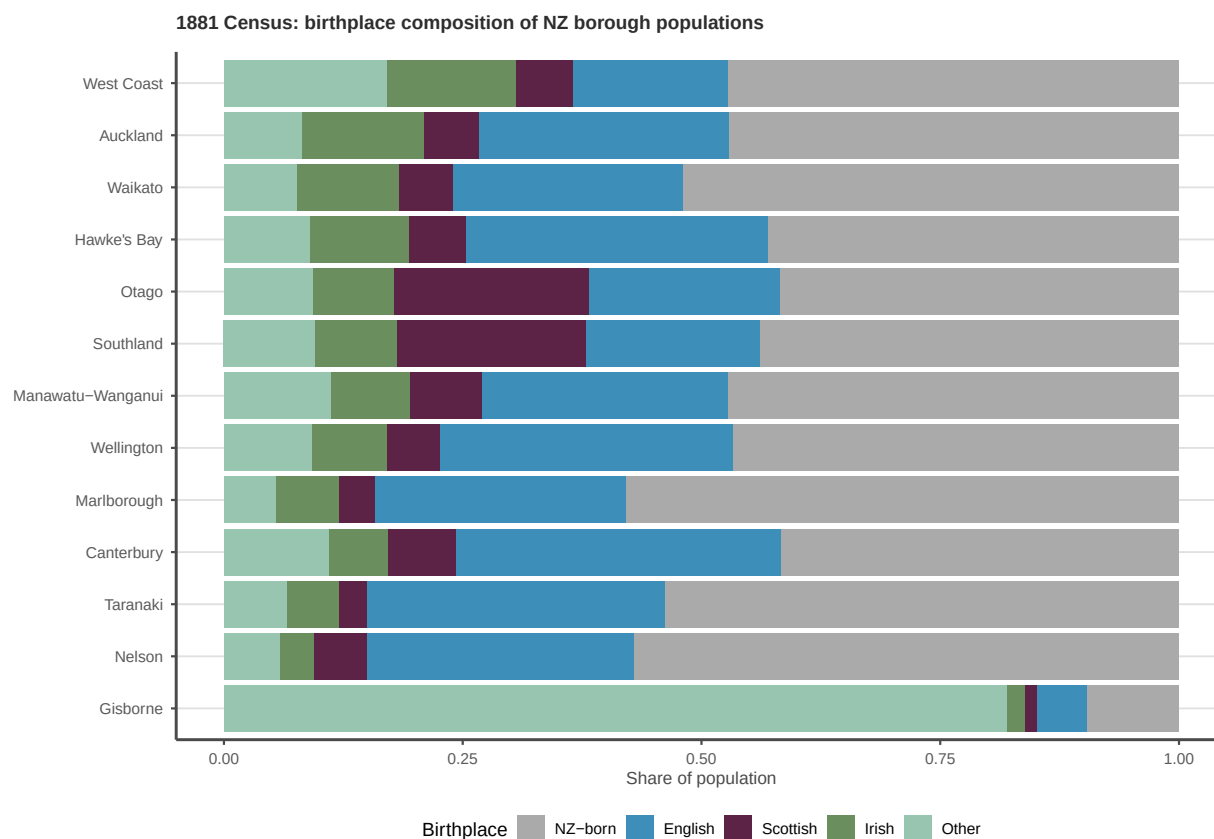


Figure A14: New Zealand: 1881 Census birthplace composition of borough populations by modern GADM1 region.

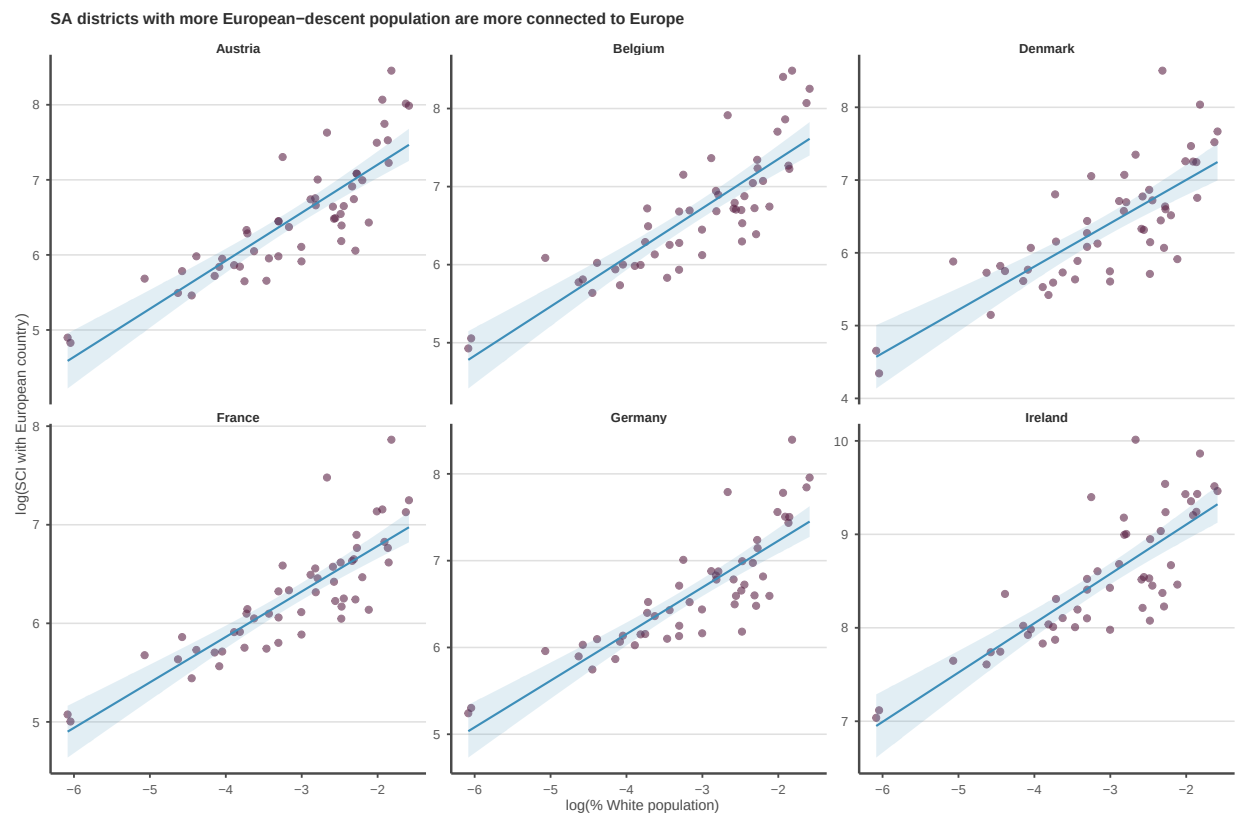


Figure A15: South Africa: $\log(\% \text{ White population})$ and $\log(\text{SCI})$ with six European partner countries.

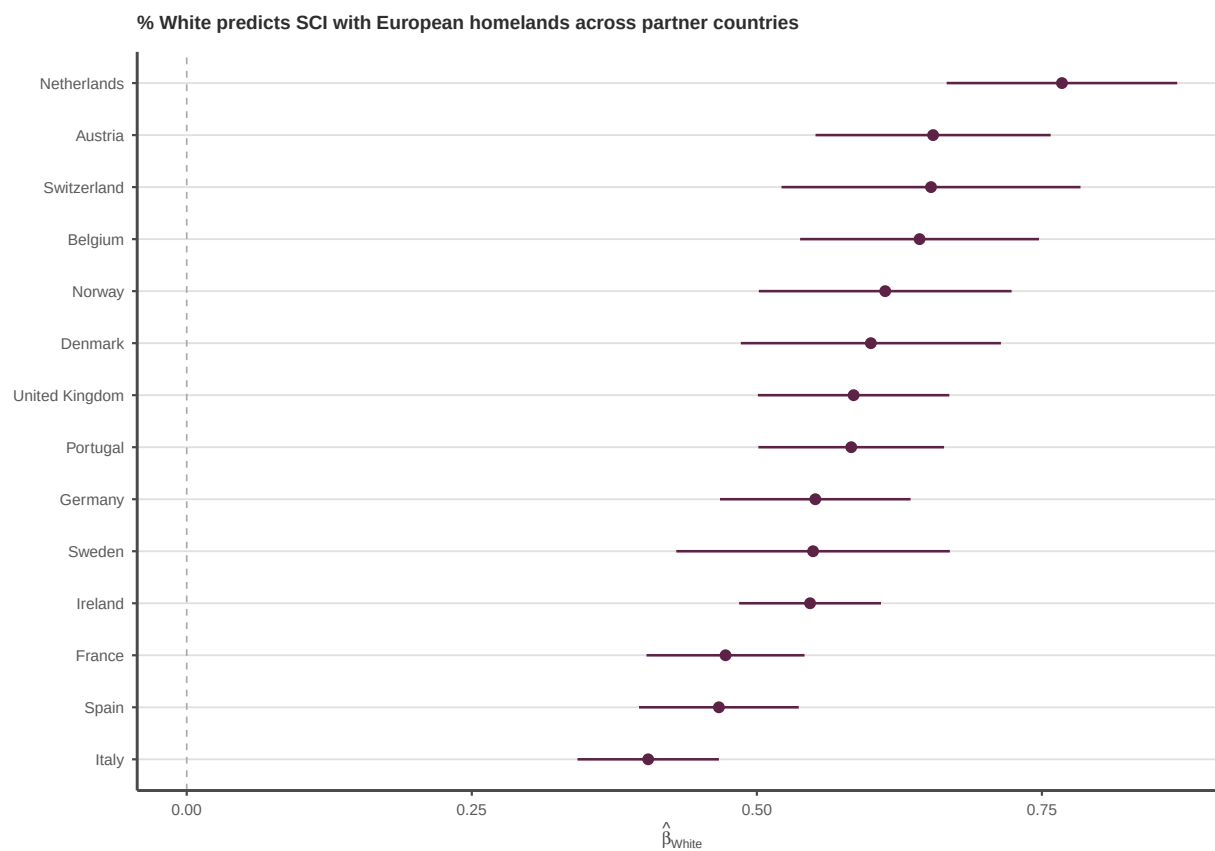


Figure A16: South Africa: % White coefficient estimated separately for each European partner country.

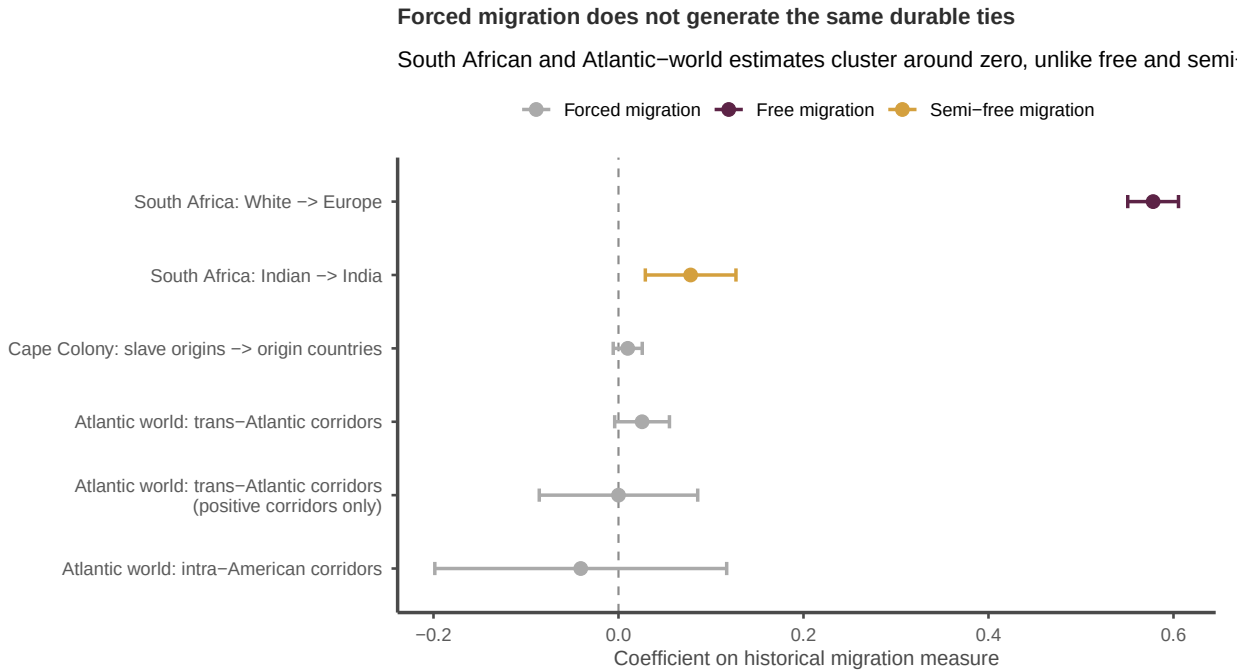


Figure A17: External validity of Hypothesis 3. The point estimates and 95% confidence intervals compare the positive homeland-oriented connectedness effects of free and semi-free migration in South Africa with the forced-migration estimates from the Cape Colony, the trans-Atlantic slave trade, and the intra-American slave trade. The forced-migration estimates cluster around zero.

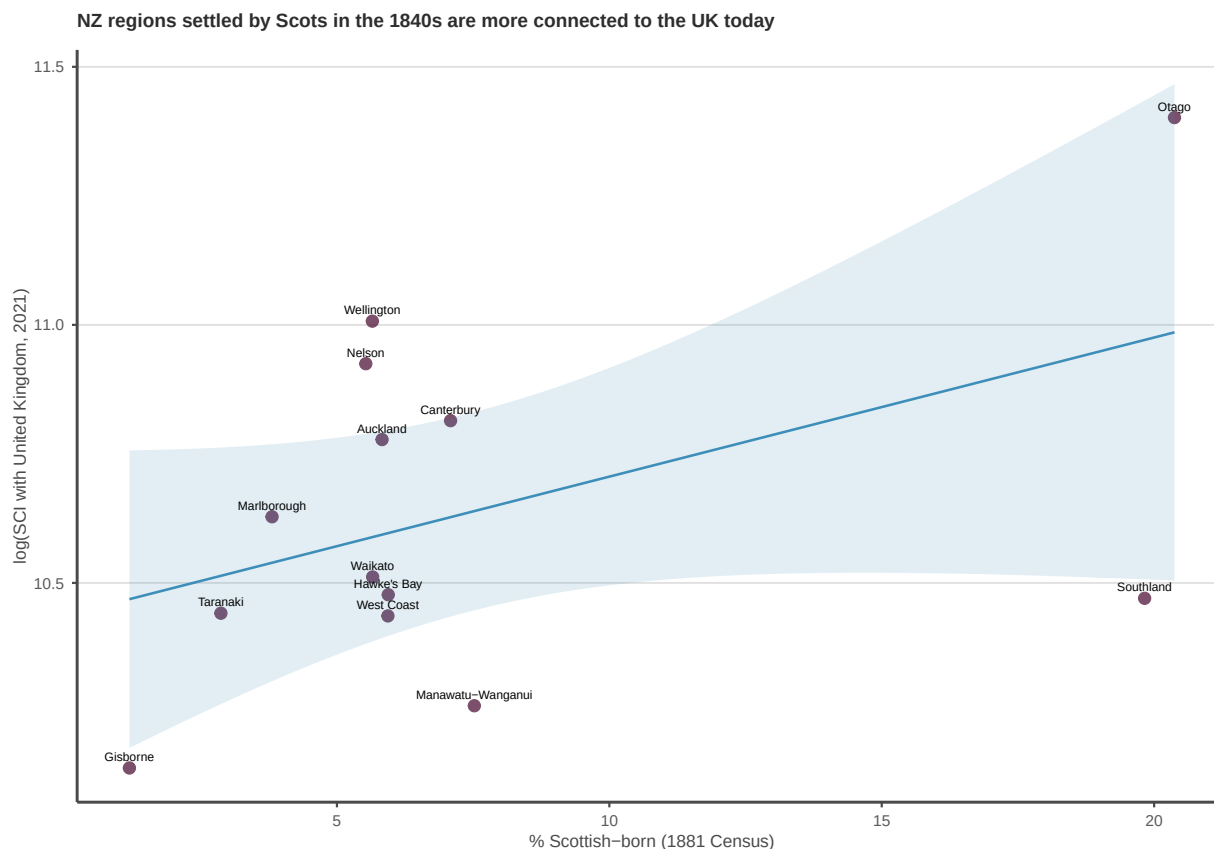


Figure A18: New Zealand: 1881 Scottish-born share and SCI with the United Kingdom (2021), suggestive rather than conclusive. Each point is a modern GADM1 region; the horizontal axis shows the share of the region's 1881 borough population born in Scotland. The Scottish-born coefficient in Table 4 is positive but imprecise, and one high-Scottish-share region records relatively low UK SCI, so the pattern should be read as indicative of the broad regional contrast rather than as a precise Scottish-specific relationship.

V. Supplementary Tables

Table A22: Regional subsamples: much of the global variation comes from settler-colony pairs, but the estimated elasticity is also large within Europe

	Americas	Americas $\times$ Rest	Europe	Africa	Asia
anc_log	0.224 (0.152)	0.124** (0.046)	0.243** (0.085)	0.018 (0.063)	0.004 (0.092)
Country FE	Yes	Yes	Yes	Yes	Yes
Within R^2	0.646	0.222	0.411	0.572	0.374
N	376	3,914	701	1,273	701

Notes: Dependent variable: $\log(\text{SCI}_{ij})$. All specifications include gravity controls and country fixed effects. Americas = both countries in the Americas; Americas $\times$ Rest = one country in the Americas, one elsewhere; Europe = both countries in Europe; Africa = both in Africa; Asia = both in Asia. Standard errors clustered on the two country identifiers defining each undirected pair. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A23: Australasia pooled: directional pattern is consistent with weaker UK-SCI elasticity in convict-heavy colonies, but the formal tests do not reject the null

	(A1)	(A2)	(A3)
	Pooled	+ Convict interaction	Free vs Penal split
<i>Dep. var.: log(SCI to GB)</i>			
% UK-born	0.217. (0.108)	0.250** (0.080)	
% UK-born $\times$ convict_share		-2.55 (5.57)	
convict_share (level)		-4.64 (12.03)	
% UK-born (free regime)			0.250** (0.080)
% UK-born (penal regime)			-0.865 (0.513)
penal regime (level)			-2.046. (1.003)
log(pop)	0.201* (0.072)	0.230* (0.083)	0.230* (0.082)
Country FE (NZ vs AU)	Yes	Yes	Yes
Within R^2	0.534	0.596	0.596
N	18	18	18
Perm. p (interaction)		0.826	

Notes: Pooled sample: 13 New Zealand GADM1 regions (1881 Census borough-level birthplace aggregated) and 5 Australian colonies (1901 Census birthplace: NSW, VIC, QLD, SA, TAS), $N = 18$. The Australian observations are aggregated from 503 GADM2 districts back to 5 colonies because birthplace and convict-transportation are recorded at colony level. Convict-share intensity is colony-level convicts transported between 1788 and 1868 (NSW $\sim 83,000$; TAS $\sim 73,500$; QLD 0; VIC $\sim 3,000$; SA 0) divided by 1901 colony population; New Zealand regions are coded with convict_share = 0. Queensland's only penal episode was the secondary Moreton Bay settlement, which received transportees reconvicted elsewhere in Australia, so it is coded as free with zero primary transportation. Regime: *penal* = NSW, TAS; *free* = QLD, VIC, SA, NZ. Specifications include country fixed effects (NZ vs AU). Perm. p : two-sided permutation p -value from 5,000 permutations of the convict-share vector across the five Australian colonies. The interaction in A2 is correctly signed (negative) but is statistically indistinguishable from zero on both inferential bases (asymptotic $p = 0.66$; permutation $p = 0.83$). The penal-regime cell in A3 is identified from two colony observations (NSW and Tasmania), so the -0.865 coefficient should be read as a small-sample point estimate rather than a precisely estimated effect. Tasmania's convict-to-population ratio of 0.43 is roughly seven times NSW's 0.06, so the interaction is mechanically dominated by Tasmania. Heteroskedasticity-robust standard errors. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$.

References

- Anderson, Clare, and Hamish Maxwell-Stewart.** 2014. "Convict Labour and the Western Empires, 1415–1954." In *The Routledge History of Western Empires.* , ed. Robert Aldrich and Kirsten McKenzie, 102–117. London:Routledge.
- Arroyo Abad, Leticia, and Noel Maurer.** 2021. "History Never Really Says Goodbye: A Critical Review of the Persistence Literature." *Journal of Historical Political Economy*, 1: 31–68.
- Ashraf, Quamrul, and Oded Galor.** 2013. "The "Out of Africa" Hypothesis, Human Genetic Diversity, and Comparative Economic Development." *American Economic Review*, 103(1): 1–46.
- Australian Bureau of Statistics, Australian Data Archive, Len Smith, Tim Rowse, and Stuart Hungerford.** 2019. "Historical and Colonial Census Data Archive (HCCDA)." Digitised census tables from Australian colonies, 1833–1901.
- Conley, Timothy G., and Morgan Kelly.** 2025. "The Standard Errors of Persistence." *Journal of International Economics*, 153: 104027.
- Cowley, Trudy, Lucy Frost, Kris Inwood, Rebecca Kippen, Hamish Maxwell-Stewart, Monika Schwarz, John Shepherd, Richard Tuffin, Mark Williams, John Wilson, and Paul Wilson.** 2023. "Reconstructing a Longitudinal Dataset for Tasmania." In *Sowing: The Construction of Historical Longitudinal Population Databases.* , ed. Kees Mandemakers, George Alter, Hélène Vézina and Paul Puschmann, 455–482. Nijmegen:Radboud University Press.
- Desmet, Klaus, and Romain Wacziarg.** 2021. "The Cultural Divide." *Economic Journal*, 131(637): 2058–2088.
- Egger, Peter H., and Filip Tarlea.** 2015. "Multi-way clustering estimation of standard errors in gravity models." *Economics Letters*, 134: 144–147.
- 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.
- Fouquin, Michel, and Jules Hugot.** 2016. "Two Centuries of Bilateral Trade and Gravity Data: 1827–2014." CEPII CEPII Working Paper No. 2016-14.
- Fourie, Johan.** 2022. "The Settlers of South Africa: Economic Forces of the Expanding Frontier." In *Migration in Africa: Shifting Patterns of Mobility from the 19th to the 21st Century.* , ed. Michiel de Haas and Ewout Frankema, 135–154. Routledge.
- Head, Keith, Thierry Mayer, and John Ries.** 2010. "The Erosion of Colonial Trade Linkages after Independence." *Journal of International Economics*, 81(1): 1–14.

- Lankheet, Imke, Rickard Hammarén, Lucía Ximena Alva Caballero, Maximilian Larena, Helena Malmström, Cecile Jolly, Himla Soodyall, Michael de Jongh, and Carina M. Schlebusch.** 2025. “Wide-scale geographical analysis of genetic ancestry in the South African Coloured population.” *BMC Biology*, 23(1): 219.
- Manning, Willard G., and John Mullahy.** 2001. “Estimating log models: to transform or not to transform?” *Journal of Health Economics*, 20(4): 461–494.
- Michalopoulos, Stelios, and Melanie Meng Xue.** 2021. “Folklore.” *Quarterly Journal of Economics*, 136(4): 1993–2046.
- Nicholas, Stephen,** ed. 1988. *Convict Workers: Reinterpreting Australia’s Past*. Cambridge:Cambridge University Press.
- OpenFlights.** 2017. “OpenFlights Airports, Airlines and Routes Database.” <https://openflights.org/data.html>, Accessed 2026; routes file dated 2017.
- Pfaffermayr, Michael.** 2023. “Cross-sectional Gravity Models, PPML Estimation, and the Bias Correction of the Two-Way Cluster-Robust Standard Errors.” *Oxford Bulletin of Economics and Statistics*, 85(5): 1111–1134.
- SlaveVoyages Consortium.** 2024. “Slave Voyages.” <https://www.slavevoyages.org/>, Accessed data from the trans-Atlantic and intra-American slave trade databases.
- Spolaore, Enrico, and Romain Wacziarg.** 2009. “The Diffusion of Development.” *Quarterly Journal of Economics*, 124(2): 469–529.
- Spolaore, Enrico, and Romain Wacziarg.** 2016. “Ancestry, Language and Culture.” In *The Palgrave Handbook of Economics and Language*, ed. Victor Ginsburgh and Shlomo Weber, 174–211. Hampshire:Palgrave Macmillan.
- Spolaore, Enrico, and Romain Wacziarg.** 2018. “Ancestry and Development: New Evidence.” *Journal of Applied Econometrics*, 33(5): 748–762.
- Statistics South Africa.** 2024. “South African Census 2022, 10% sample [dataset]. Version 1.” *Pretoria: Statistics South Africa [producer]; Cape Town: DataFirst [distributor]*.
- United Nations Department of Economic and Social Affairs.** 2024. “International Migrant Stock 2024.” Bilateral migrant stock matrix, mid-year 2020 estimates.
- US Census Bureau.** 2022. “American Community Survey 2018–2022 5-Year Estimates, Table B04006: People Reporting Ancestry.” <https://data.census.gov>.
- Voth, Hans-Joachim.** 2021. “Persistence – Myth and Mystery.” In *The Handbook of Historical Economics*, ed. Alberto Bisin and Giovanni Federico, 243–267. London:Elsevier.

Waetjen, Thembisa, and Goolam Vahed. 2014. "Passages of Ink: Decoding the Natal Indentured Records into the Digital Age." *Kronos*, 40(1): 45–73.