Online Appendices

Soil Heterogeneity, Social Learning, and the Formation of Close-knit Communities

Itzchak Tzachi Raz
The Hebrew University of Jerusalem

Email: iraz@mail.huji.ac.il
Website: www.tzachiraz.com.

November 2024

Contents

A The Soil Heterogeneity Index **4**

B Historical Indicators of Close-Knit Communities **7**

B.1 The Local Name Index, Examples 7

B.2 Kinship Tightness 10

B.3 Validation 12

B.3.1 Relationship with Other Historical Measures of Close-Communities 13

B.3.2 Relationship with Contemporary Close-Knit Social Networks 27

B.3.3 Relationship with Contemporary Cultural and Psychological Characteristics . 34

B.4 The Spatial Distribution of Close-Knit Communities 41

C Further Analysis and Results **42**

C.1 Visualization of County-Level Results 42

C.2 Kinship Tightness 45

C.3 A Decaying Impact 47

C.3.1 Soil Heterogeneity and Close-Knit Communities, Decade-by-Decade 47

C.3.2 The Long-Run Impact 50

C.4 DID Results 53

C.5 Selective Out-Migration 59

C.6 Confounding Factors, Competing Mechanisms, and Mediators 61

C.6.1 Other geoclimatic features of the environment 61

C.6.2 Climatic risk 65

C.6.3 Trade 68

C.6.4 Immigration and birthplace diversity 71

C.6.5 Race and the legacy of slavery 74

C.6.6 Agricultural inequality 76

C.6.7 Modernization and frontier experience 77

D Robustness Checks **83**

D.1 Robustness of County-Level Results 83

D.2 Robustness of Selective In-Migration Results 93

D.3 Robustness of DID and DDD Results 94

D.4 Robustness of Migrants' Children Results 98

D.5 Robustness of Social Learning Results 102

D.6 Robustness of Selective Out-Migration Results 109

E Data Sources and Variables Construction **112**

E.1 Soil Heterogeneity Index 112

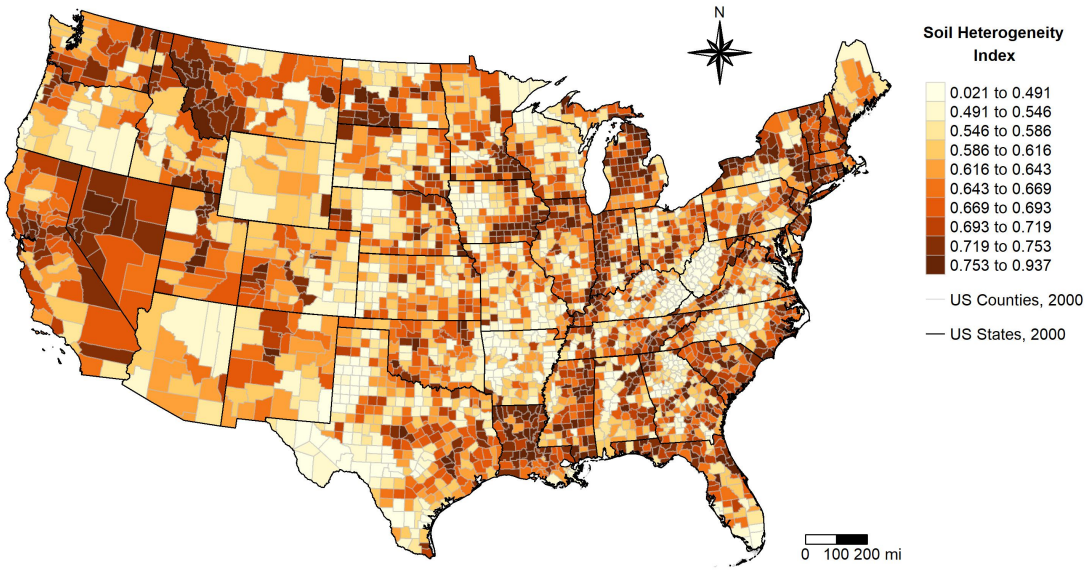
E.2 Historical Measure of Close-Knit Communities 115

E.3 Other Variables 118

References **125**

A The Soil Heterogeneity Index

FIGURE A.1: County-Level Soil Heterogeneity Index, 2000



Note: This figure plots the *Soil Heterogeneity Index* (SHI) for counties in the contiguous U.S. in 2000. Darker colors indicate higher soil heterogeneity. See Appendix E.1 for a description of the SHI construction.

TABLE A.1: Soil Heterogeneity and Geoclimatic Covariates

	Dependent variable: Soil Heterogeneity Index				
	(1)	(2)	(3)	(4)	(5)
Temperature	-0.0057	0.009	0.00036	-0.0063	-0.016
	(0.0036)	(0.0084)	(0.0073)	(0.0042)	(0.023)
Precipitation	0.0021	0.022***	0.01*	-0.019***	-0.018**
	(0.0041)	(0.0062)	(0.0058)	(0.0065)	(0.008)
Slope	0.0089***	0.0096**	0.0037	0.019***	0.029***
	(0.003)	(0.0045)	(0.0047)	(0.0043)	(0.0059)
Elevation	-0.001	-0.017***	-0.0048	-0.0016	-0.033***
	(0.003)	(0.0061)	(0.0039)	(0.0051)	(0.014)
Productivity	0.018***	0.026***	0.033***	0.027***	0.026***
	(0.0029)	(0.0056)	(0.0047)	(0.0043)	(0.0055)
Flow accumulation	0.0058**	0.0049	0.0075***	0.0043	0.0027
	(0.0026)	(0.0037)	(0.0027)	(0.0035)	(0.0038)
River density	0.0046	0.0067**	0.0057*	0.0072***	0.0078***
	(0.0029)	(0.003)	(0.0029)	(0.003)	(0.0031)
Total area	-0.0024	-0.016***	-0.0083***	-0.0058**	-0.011***
	(0.0024)	(0.0028)	(0.0024)	(0.0025)	(0.0027)
Distance to navigated rivers	-0.0095***	-0.024***	-0.012***	-0.011***	-0.018***
	(0.0038)	(0.0037)	(0.0038)	(0.0042)	(0.0035)
Distance to lakes	-0.0039	-0.00093	-0.0044	-0.011***	-0.018***
	(0.0035)	(0.0026)	(0.0034)	(0.0042)	(0.0035)
Distance to shoreline	-0.014***	-0.0066	-0.015***	-0.0045	0.00056
	(0.0038)	(0.0047)	(0.0062)	(0.0037)	(0.0024)
State Fixed Effects		✓			✓
Smooth Location Controls			✓		✓
Geoclimatic Controls				✓	✓

Note: This table reports estimations from cross-sectional county-level regressions in which the SHI is the dependent variable. Each cell in columns 1-3 represents a different regression and the coefficient on the variable listed on the left is reported. In columns 4-5 the coefficients reported are from a single regression that includes all the geoclimatic covariates. The right hand side variables are standardized into z-scores. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE A.2: Soil Heterogeneity Correlates with Measures of Agricultural Heterogeneity

	Dependent variable:					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Variation of Agriculture Suitability (mean = 0, SD = 0.75)					
Soil Heterogeneity	0.770***	0.816***	0.501***	0.494***	0.439***	0.364**
	(0.158)	(0.159)	(0.159)	(0.160)	(0.160)	(0.158)
Observations	3,109	3,109	3,109	3,109	3,109	3,090
R^2	0.013	0.155	0.235	0.237	0.268	0.315
	Panel B: Agricultural Diversity (mean = 0, SD = 1)					
Soil Heterogeneity	0.735**	0.828***	0.769***	0.661**	0.516**	0.556***
	(0.357)	(0.259)	(0.277)	(0.285)	(0.222)	(0.216)
Observations	23,254	23,254	23,254	23,254	23,254	23,215
R^2	0.007	0.342	0.355	0.365	0.392	0.403
State × Year Fixed Effects		✓	✓	✓	✓	✓
Geoclimatic Controls			✓	✓	✓	✓
Smooth Location Controls				✓	✓	✓
Agricultural Suitability Controls					✓	✓
Higher Order Controls						✓

Note: This table reports estimates of multiple regressions in which the dependent variables are other measures of spatial environmental heterogeneity. In Panel A the dependent variable is the mean of standard deviations of agricultural suitability indices for alfalfa, cotton, maize, oat, rye, sugarcane, sweet potato, tobacco, wheat, and white potato. In Panel B the dependent variable is a county-level agricultural diversity index for the years 1880-1935, standardized into z-scores. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. Agricultural suitability controls include average agricultural suitability indices for alfalfa, cotton, maize, oat, rye, sugarcane, sweet potato, tobacco, wheat, and white potato. Higher order controls include the standard deviations of temperature, precipitation, slope, elevation, flow accumulation, and in panel B, also the agricultural suitability indices. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

B Historical Indicators of Close-Knit Communities

B.1 The Local Name Index, Examples

To get some intuition for the variation that is captured by the LNI, Appendix Table B.1 reports a few examples of the LNI in which "local" is defined as the state, focusing on boys names in 1940 in Arkansas and Massachusetts. First, some names are common in one state but rare in the other, leading to large differences in the LNI (Panel A). Consider the following example: in 1940, about 0.49% of boys in the U.S. were named "Billie".[1] However, there was substantial regional variation in the popularity of the name. In Arkansas, about 2.03% of boys were given that name, while in Massachusetts only about 0.0005%. Those striking regional differences in the popularity of the name meant that it caries information regarding the likely state of birth. Put differently, naming their child Billie was a good way for parents in Arkansas to signal their group identity.[2] As a result, a Billie in Arkansas is assigned with a high LNI of 81.40, reflecting a highly local name, but a very low LNI of 0.10 in Massachusetts.[3] A similar pattern exist for "Floyd" and "Jerry", and the opposite pattern for "Francis", "Peter", and "Frederick".

Second, significant differences may occur for names that are common in both locations. Panel B reports five names that are among the top ten most common names in both states. While the gaps in the LNI tend to be lower, they are still significant, and there are more children with these names, implying a significant impact on the average LNI.

Third, very large differences in the LNI may occur even when names are relatively uncommon in both locations (Panel C). For example, only 0.0094% and 0.0005% of boys in Arkansas and Massachusetts, respectively, were named "Walker". But since the name was *relatively* much more common in Arkansas and much less common in Massachusetts, the name's LNI score in Arkansas and Massachusetts was 60.67 and 7.7, respectively. Similarly, while only 0.0055% of boys were named "Wyatt" in Arkansas, the LNI score was 66.27, since it was much more common in the state relative to other locations. In contrast, in Massachusetts no child was given that name.[4]

[1] That is, they appear in census records as "Billie." Their official name may very well be "William." I am agnostic to this question, as communal identification is likely, and arguably, more likely, to present itself also in the name that is used in practice in social contexts rather than only in names that appear in formal records.

[2] More generally, Billie is considered a good "Southern name".

[3] The name presents similar spatial patterns for girls, with an LNI of 77.17 in Arkansas and 2.43 in Massachusetts. A similar pattern also exists using the name "Billy" for boys, with an LNI of 75.74 in Arkansas and 1.09 in Massachusetts.

[4] The name "Wyatt" was generally quite an uncommon name in the U.S. in 1940, given to less than 0.003% of boys in the baseline sample. The name was quite uncommon even in Alabama—the state with the highest share—0.015%. Yet the name was relatively much more common in Alabama than outside of it, leading to a high LNI of 85.26 in Alabama, much higher than the lowest non-zero LNI of 5.26 that was associated with the name in Pennsylvania.
In the empirical analysis, I show that very rare names are not driving the results (Appendix Table D.3).

As these examples demonstrate, when the social and historical context presents multiple competing groups or social identities, the variation captured by the LNI may be quite different from the one captured by first name commonness (e.g., Bazzi et al., 2020; Beck-Knudsen, 2024). Relatively uncommon names, such as "Wyatt," might reflect a desire to signal identification with a particular group rather than a desire to stand out (i.e. and not identify with any group).[5] Similarly, some of the most locally common names, such as "John" or "Robert," are relatively much more common in other locations and may have been chosen out of the desire to identify with some national or universal group rather than the local one. As a result, even when a name commonness measure is defined over the local geography rather than the national level, the LNI and first name commonness might result in different patterns. Therefore, in contexts with multiple social groups and when the focus of interest is not a general desire to fit in or stand out, but rather which particular group individuals identify with, it is essential to consider relative commonness across groups rather than absolute commonness.

[5] As Fryer and Levitt (2004) show, significant shares of Black children in California have unique or almost unique names, and this partly explain the increase in the prevalence of distinctively Black names. Absolutely unique names used by Blacks are "distinctively Black" by definition. For a useful discussion on the difference between unique and disproportionately assigned names see Cook et al. (2014).

TABLE B.1: The Local Name Index, Examples, 1940

	Arkansas		Massachusetts	
	Share	LNI	Share	LNI
	Panel A: Common in one, rare in the other			
Billie	2.0266	81.40	0.0005	0.10
Floyd	0.4415	65.46	0.0332	12.07
Jerry	1.1492	62.21	0.0332	4.39
Francis	0.0746	16.51	1.5644	82.05
Peter	0.0244	8.20	0.9035	78.17
Frederick	0.0401	12.37	0.7546	73.81
	Panel B: Common in both			
James	5.3187	52.83	3.8793	44.80
John	2.7634	39.03	7.1845	63.01
William	2.4476	39.82	5.2227	58.94
Robert	2.4374	29.77	9.2294	62.21
Donald	1.8090	35.59	3.2384	49.91
	Panel C: Rare in both			
Walker	0.0094	60.67	0.0005	7.70
Wyatt	0.0055	66.27	—	—
Carleton	0.0039	35.93	0.0284	81.59
Kevin	—	—	0.0674	88.93

Note: This table reports a few examples of LNI scores attached to children's names in 1940, in which "local" is defined as the state. The sample includes white native-born children between the ages of 0 to 10 with native-born parents. See Appendix E for details on the data and variable construction.

B.2 Kinship Tightness

Three additional outcome variables focus on kinship tightness. It is not trivial that kinship tightness will be affected to the same extent, or even in the same direction, as communal tightness. On the one hand, many conceptual frameworks, such the collectivism-individualism and the communal moral values frameworks, views "communal" or "collectivist" as a "type" that is expected to be more "groupy" in general, and not just with respect to a particular group. Therefore, although the social learning hypothesis is focused on neighbors and communal ties, kinship tightness may also be impacted, although probably to a lower extent. On the other hand, as in the social identity framework, individuals may choose to identify with some groups they belong to more than others, and social reliance on kinship and community can be substitutes (e.g. Buggle and Durante, 2021).

Median Kinship Propinquity (MKP). Two outcomes focus on patrilineal kin propinquity and rely on data and methods from Nelson (2020), which uses the fact that census enumerators generally went from house to house to collect information, and therefore distance on the enumeration form correlates with geographical distance. Nelson's method measures the distance to (i.e., the number of households between) the closest same-race same-surname household, and computes the probability that the surname match was not random and reflects kinship as

$$P(K_{rie}) = \left(1 - \frac{N_{rie} - 1}{N_{re} - 1}\right)^{D_i}$$

where N_{rie} is the number of the same-race (r) same-surname (i) households in the same enumeration district (e), N_{re} is the total number of same-race households in the enumeration district, and D_i is the number of different-surname households that are as close as the nearest same-surname household.

Similar to the national calculation in Nelson (2020), I calculate the county-level median distance between same-surname households, weighted by the probability of a non-random match. I define *Median Kinship Propinquity* (MKP) as one over the median distance.

Kinship Propinquity Rate (KPR). I also follow Nelson (2020) to calculate the county-level *Kinship Propinquity Rate* (KPR), defined as the share of the population with kin residing in the same enumeration distinct, weighted by the probability of a non-random match.

The Strength of Family Ties Index (SFTI). Finally, I use data from the full count censuses between 1860-1940 (Ruggles et al., 2020) to construct a novel county-level *"Strength of Family Ties Index"* (SFTI).[6] I focus on four key variables related to family structure and the choice of living arrangements that are observable in historical censuses and are associated with strong versus weak family ties.

[6] 1850 is excluded because in this year information regarding material status was not recorded.

For each county-year, I calculate (i) the divorce-to-marriage ratio, (ii) the share of elderly people living without a relative, (iii) the share of people living with at least one person who is not their relative, and (iv) the mean size of families. Then, for each year, I conduct a principal component analysis at the county level using these variables as inputs. The first eigenvector, which I refer to as the SFTI, explains between $45-71\%$ of the variance in the four variables, depending on the year and sample. It is also the only component with an eigenvalue that is larger than one in all years and samples ($1.79-2.82$). In all years and samples, the loading on the four variables always has the same sign (positive on family size and negative on the rest). Because there is no natural interpretation for the SFTI units, I standardize it into z-scores within each year.

B.3 Validation

I validate the different indicators of historical close-knit communities in three ways. First, in Appendix Figures B.1-B.7, I show that the historical indicators are positively related to each other. A significant positive association is found between almost all pairs of indicators, both in general (left columns) and within states (right columns).[7] This was expected, as the cultural and psychological attributes that the indicators are designed to capture constitute a correlated bundle of traits that evolved in response to the social demand of a close-knit social structure (Schulz et al., 2019; Henrich, 2020).

Second, I show that the historical indicators of 1940 are positively and significantly associated with contemporary indicators of close-knit social networks and the cultural and psychological traits associated with them. Note that this validation exercise also relies on the persistence of social structure, culture, and psychology. To measure the tightness of contemporary social networks, I use two county-level indicators of local close-knit social networks calculated by Chetty et al. (2022a,b) using active U.S. Facebook users aged 25-44. The first is *Social Clustering*, which measures the average fraction of an individual's friend pairs who are also friends with each other. The second is the *Social Support Ratio*, which measures the proportion of within-county friendships where the pair of friends share a third mutual friend within the same county. I standardize both into z-scores. In Appendix Figures B.8-B.14, I document the positive and significant relationship between the historical indicators and both indicators of contemporary close-knit social networks, both in general (left columns) and within states (right columns).[8] As direct data on historical social structure is unavailable, I view this validation exercise as particularly important.

Finally, I explore the relationship between the historical indicators and three contemporary measures of cultural and psychological traits associated with close-knit social networks. I use a county-level measure of the relative importance of communal moral values from Enke (2020), which utilizes data from Graham et al. (2011), and state-level measures of cultural tightness (Harrington and Gelfand, 2014) and collectivism (Vandello and Cohen, 1999). Note that statistical power is low in the state-level analysis due to the small sample. Appendix Figures B.15-B.21 plots the relationships between the variables, both in general (left columns) and within states or census regions (right columns). I find a positive association in 37 out of 42 cases, out of which 26 are statistically significant.

Overall, these tests suggest that the indicators of historically close-knit communities capture meaningful cultural and psychological variation that relates to social structure.

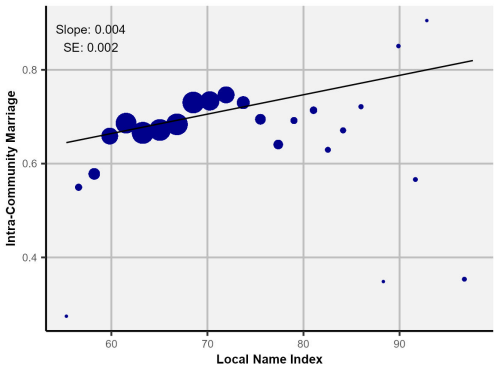
[7] The only exception is the RHI, which is not significantly associated with the ICM, KPR, and SFTI when state fixed-effects are not partialled out.

[8] Again, the only exception is the RHI, which is significantly associated with Support Ratio only within states.

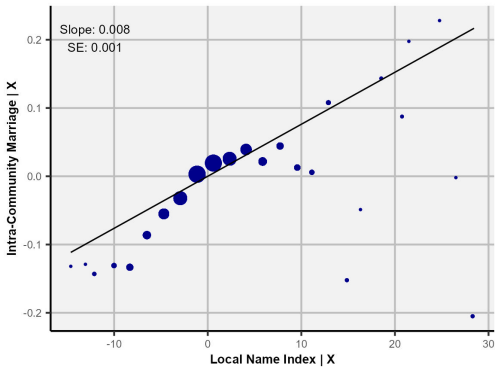
B.3.1 Relationship with Other Historical Measures of Close-Communities

FIGURE B.1: The LNI and Other Historical Indicators of Close-Communities

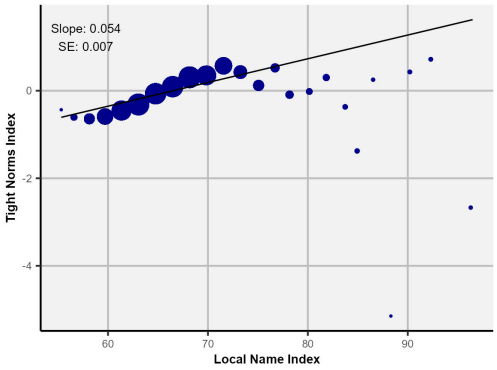
(A) Intra-Community Marriage



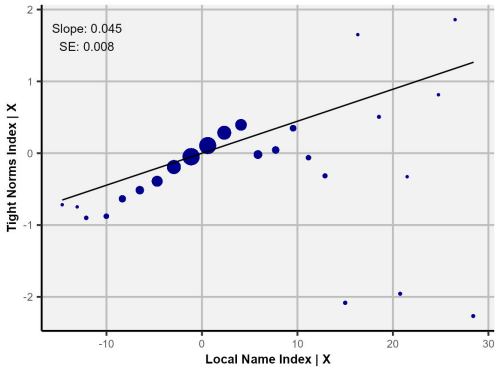
(B) Intra-Community Marriage | X



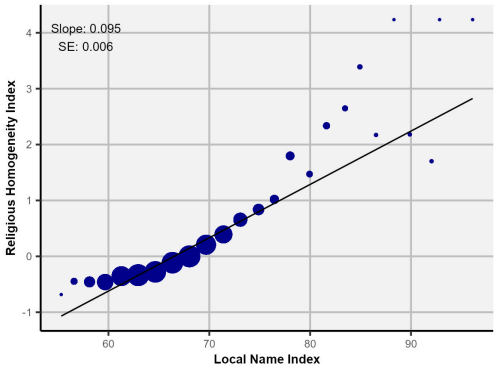
(C) Tight Norms Index



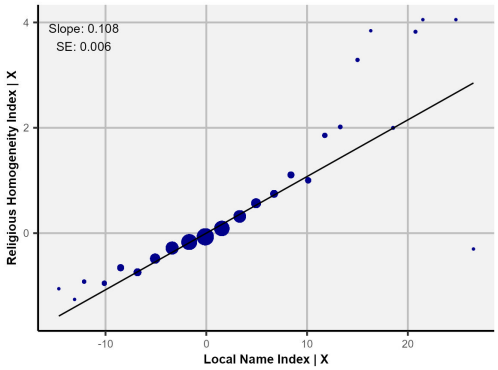
(D) Tight Norms Index | X



(E) Religious Homogeneity Index



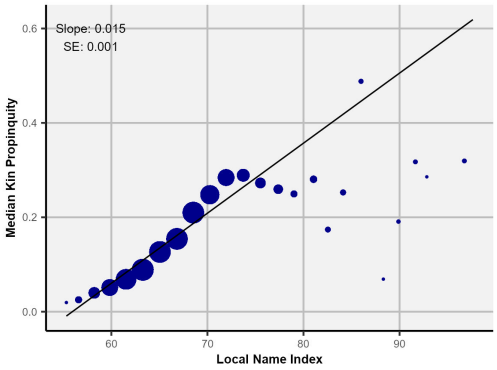
(F) Religious Homogeneity Index | X



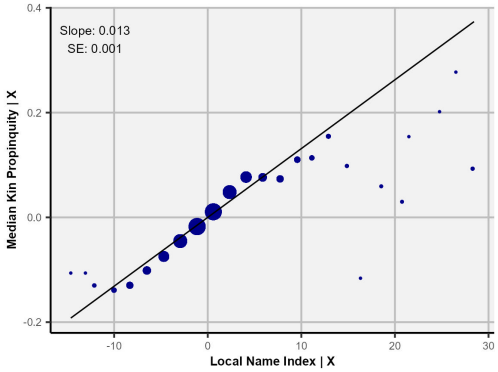
Note: This figure plots the relationship between the LNI and the other historical measures of close-knit communities in 1940: the ICM (subfigures A-B), the TNI (subfigures C-D), and the RHI (subfigures E-F). The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

FIGURE B.1: The LNI and Other Historical Indicators of Close-Communities (cont.)

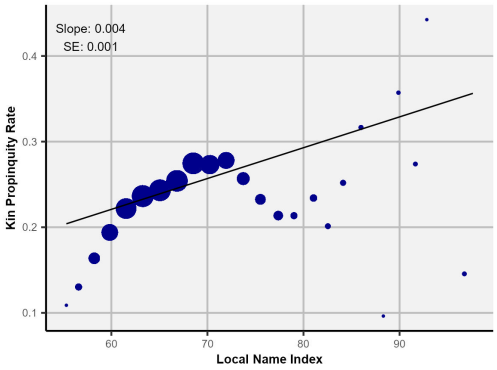
(G) Median Kin Propinquity



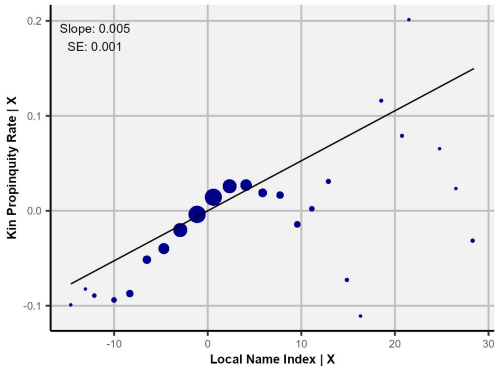
(H) Median Kin Propinquity | X



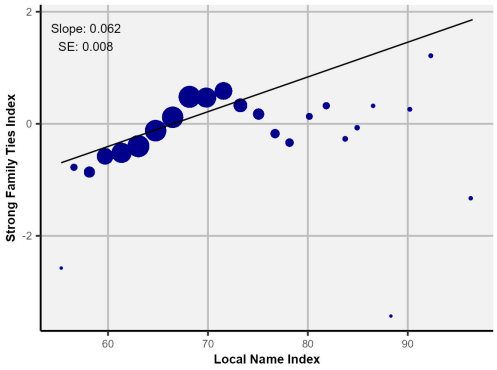
(I) Kin Propinquity Rate



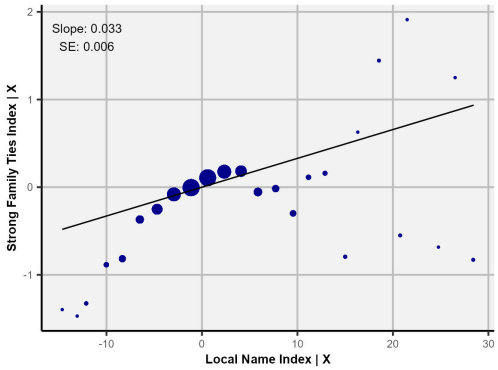
(J) Kin Propinquity Rate | X



(K) Strength of Family Ties Index



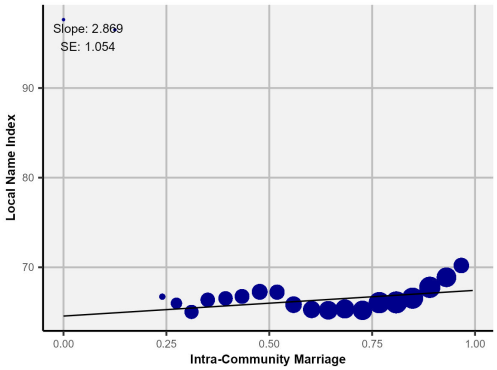
(L) Strength of Family Ties Index | X



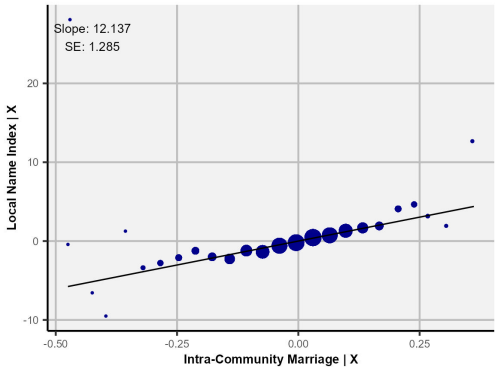
Note: This figure plots the relationship between the LNI and the other historical measures of close-knit communities in 1940: the MKP (subfigures G-H), the KPR (subfigures I-J), and the SFTI (subfigures K-L). The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

FIGURE B.2: The ICM and Other Historical Indicators of Close-Communities

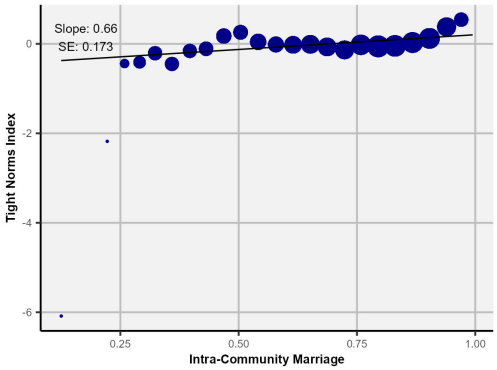
(A) Local Name Index



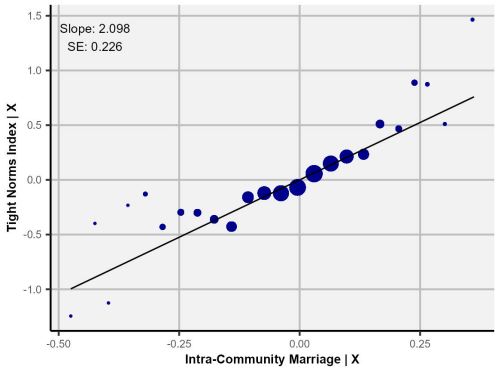
(B) Local Name Index | X



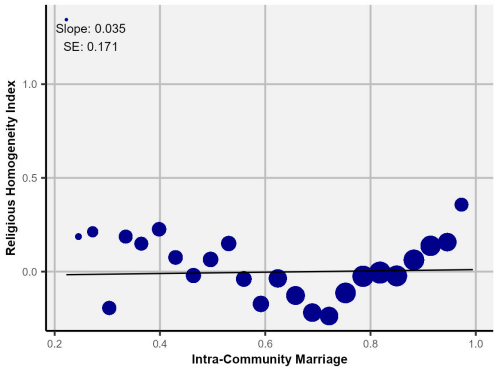
(C) Tight Norms Index



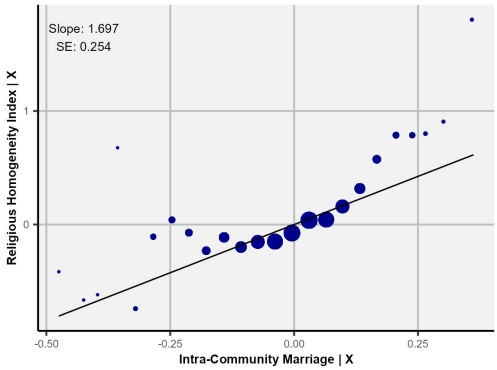
(D) Tight Norms Index | X



(E) Religious Homogeneity Index



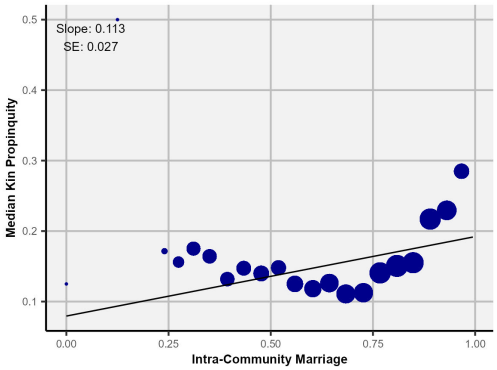
(F) Religious Homogeneity Index | X



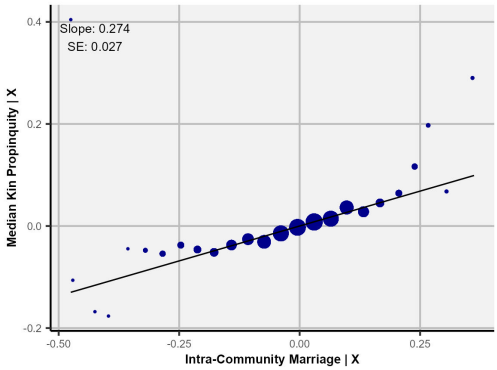
Note: This figure plots the relationship between the ICM and the other historical measures of close-knit communities in 1940: the LNI (subfigures A-B), the TNI (subfigures C-D), and the RHI (subfigures E-F). The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

FIGURE B.2: The ICM and Other Historical Indicators of Close-Communities (cont.)

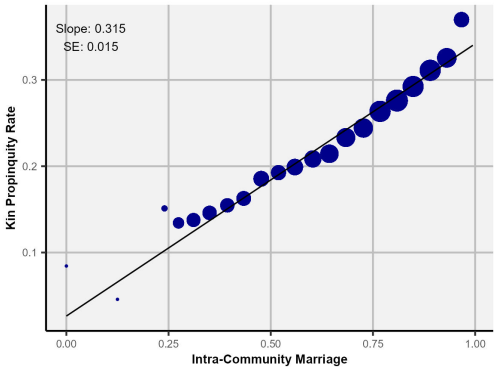
(G) Median Kin Propinquity



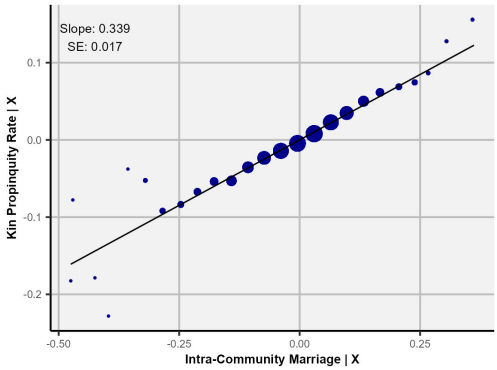
(H) Median Kin Propinquity | X



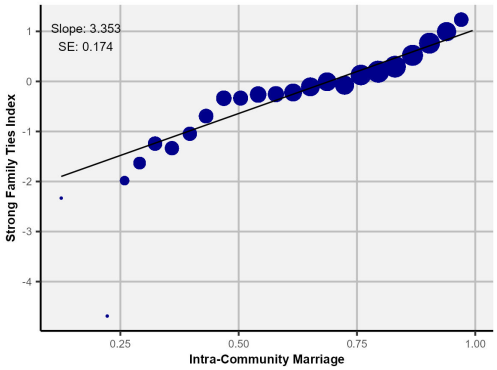
(I) Kin Propinquity Rate



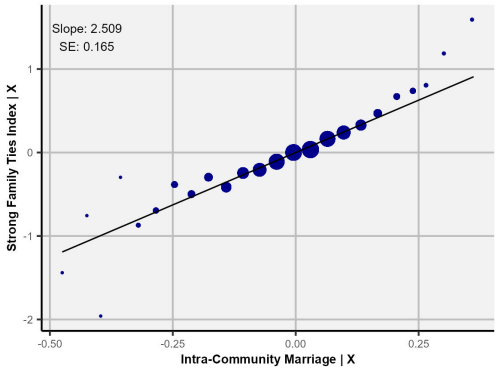
(J) Kin Propinquity Rate | X



(K) Strength of Family Ties Index



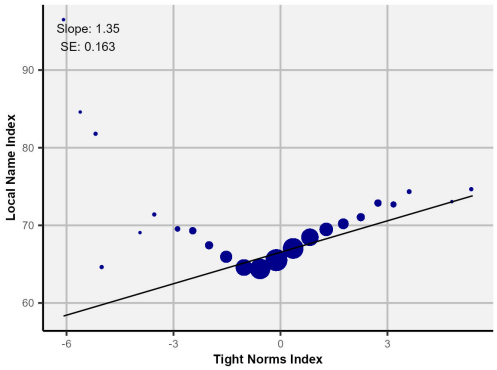
(L) Strength of Family Ties Index | X



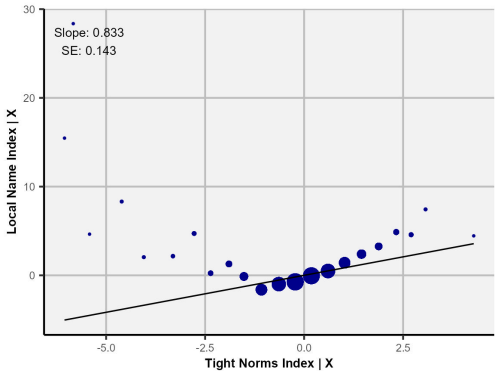
Note: This figure plots the relationship between the ICM and the other historical measures of close-knit communities in 1940: the MKP (subfigures G-H), the KPR (subfigures I-J), and the SFTI (subfigures K-L). The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

FIGURE B.3: The TNI and Other Historical Indicators of Close-Communities

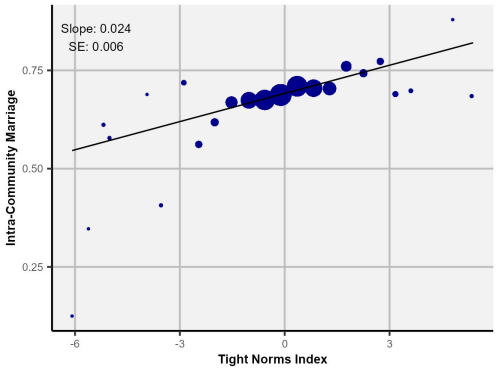
(A) Local Name Index



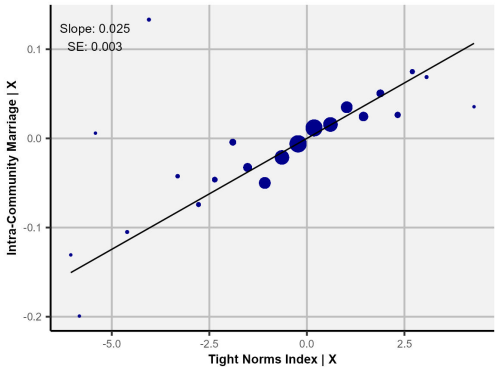
(B) Local Name Index | X



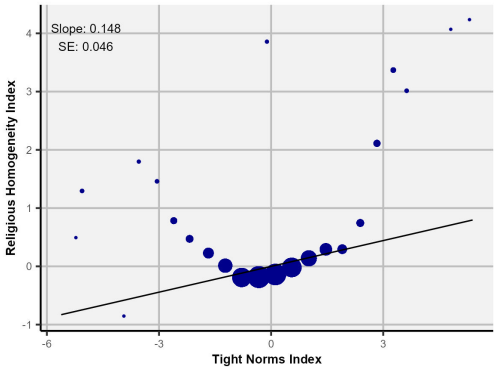
(C) Intra-Community Marriage



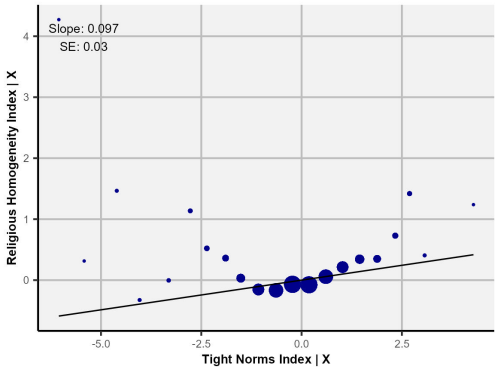
(D) Intra-Community Marriage | X



(E) Religious Homogeneity Index



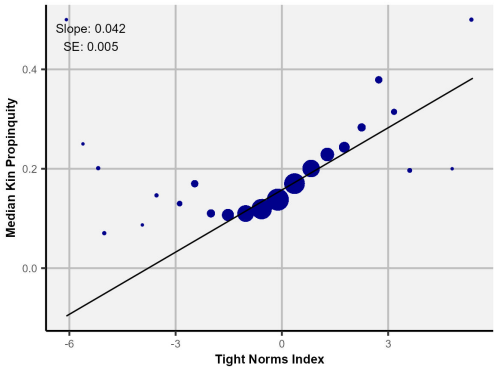
(F) Religious Homogeneity Index | X



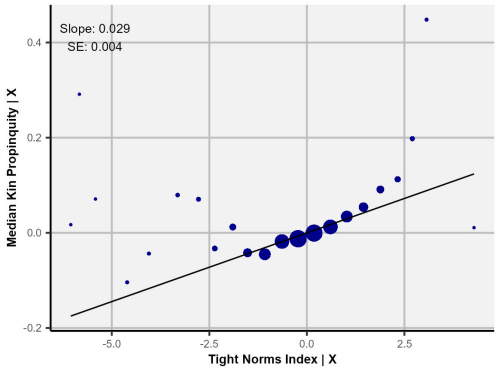
Note: This figure plots the relationship between the TNI and the other historical measures of close-knit communities in 1940: the LNI (subfigures A-B), the ICM (subfigures C-D), and the RHI (subfigures E-F). The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

FIGURE B.3: The TNI and Other Historical Indicators of Close-Communities (cont.)

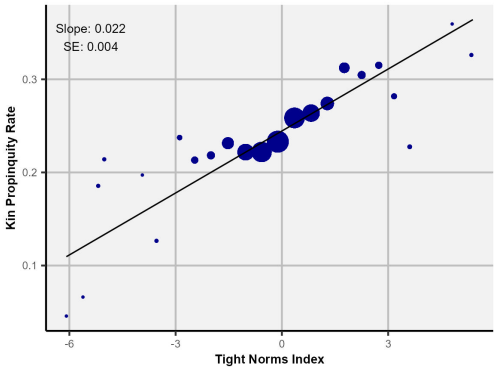
(G) Median Kin Propinquity



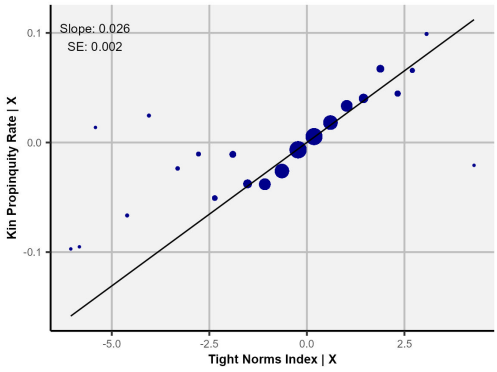
(H) Median Kin Propinquity | X



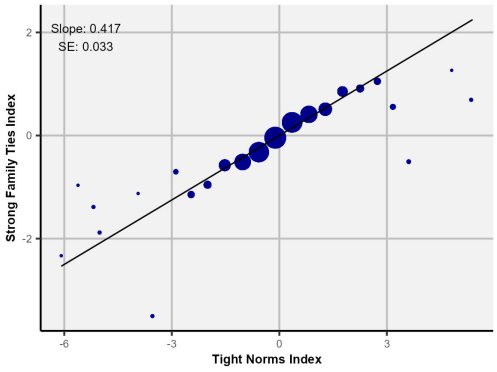
(I) Kin Propinquity Rate



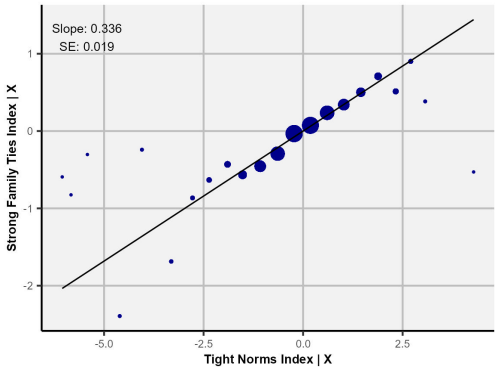
(J) Kin Propinquity Rate | X



(K) Strength of Family Ties Index



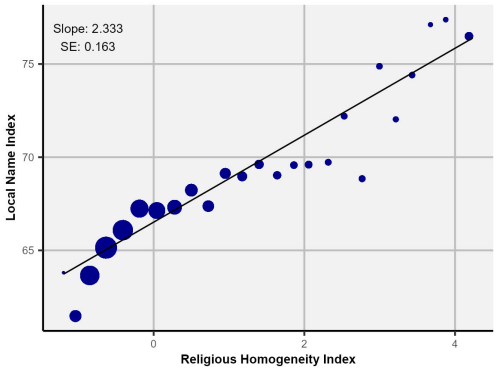
(L) Strength of Family Ties Index | X



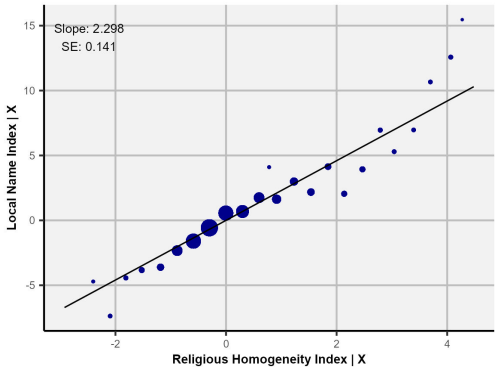
Note: This figure plots the relationship between the TNI and the other historical measures of close-knit communities in 1940: the MKP (subfigures G-H), the KPR (subfigures I-J), and the SFTI (subfigures K-L). The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

FIGURE B.4: The RHI and Other Historical Indicators of Close-Communities

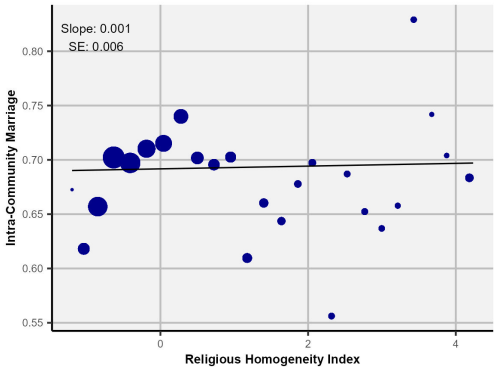
(A) Local Name Index



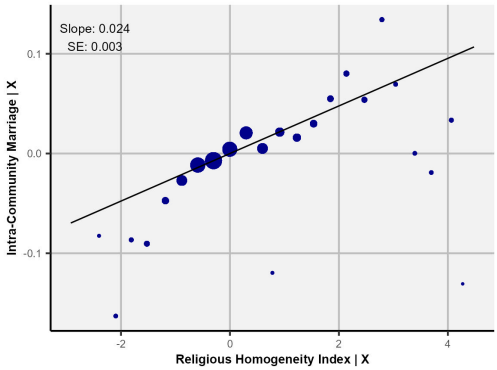
(B) Local Name Index | X



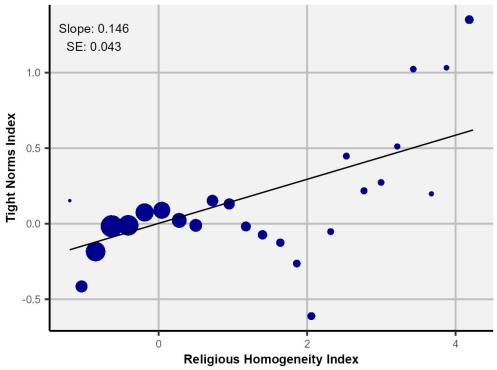
(C) Intra-Community Marriage



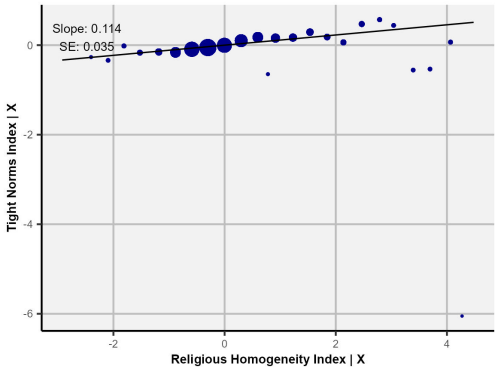
(D) Intra-Community Marriage | X



(E) Tight Norms Index



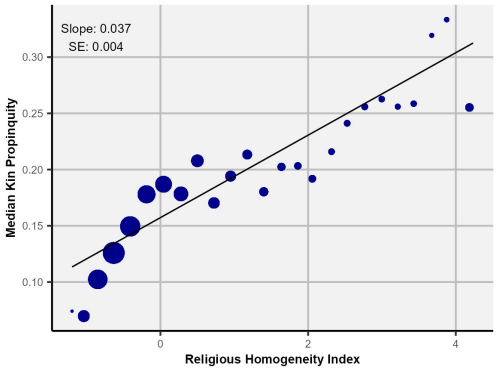
(F) Tight Norms Index | X



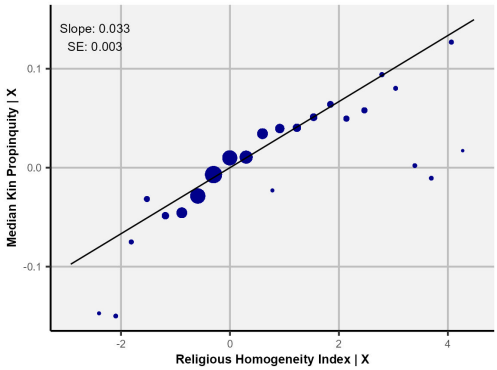
Note: This figure plots the relationship between the RHI and the other historical measures of close-knit communities in 1940: the MKP (subfigures G-H), the KPR (subfigures I-J), and the SFTI (subfigures K-L). The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

FIGURE B.4: The RHI and Other Historical Indicators of Close-Communities (cont.)

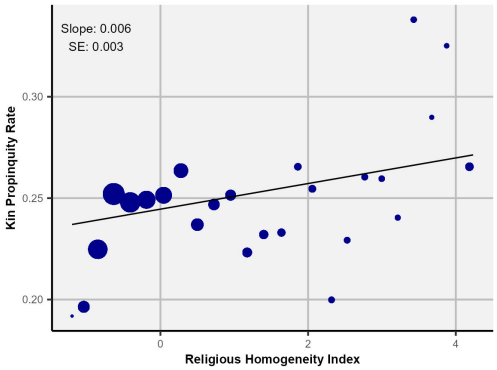
(G) Median Kin Propinquity



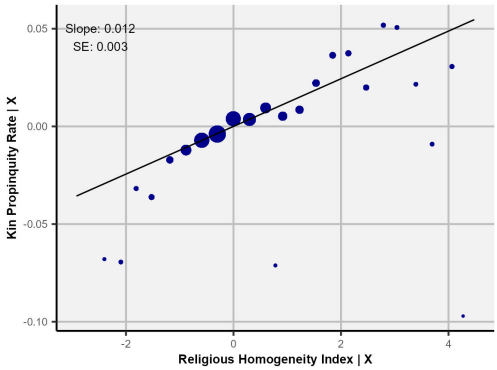
(H) Median Kin Propinquity | X



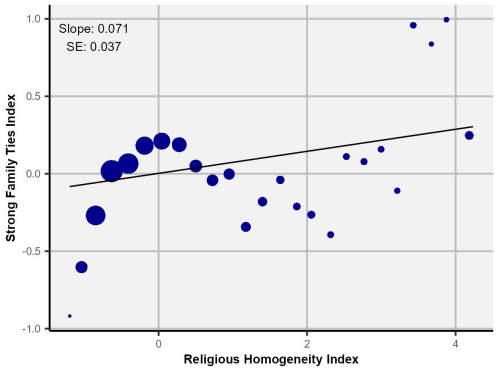
(I) Kin Propinquity Rate



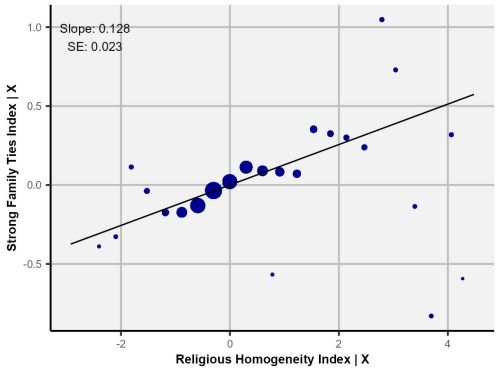
(J) Kin Propinquity Rate | X



(K) Strength of Family Ties Index



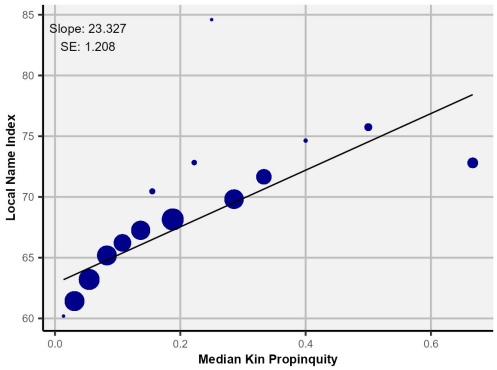
(L) Strength of Family Ties Index | X



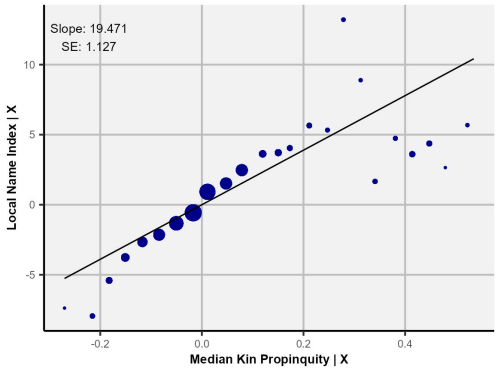
Note: This figure plots the relationship between the RHI and the other historical measures of close-knit communities in 1940: the MKP (subfigures G-H), the KPR (subfigures I-J), and the SFTI (subfigures K-L). The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

FIGURE B.5: The MKP and Other Historical Indicators of Close-Communities

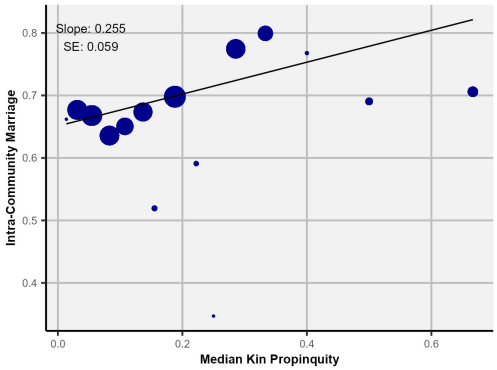
(A) Local Name Index



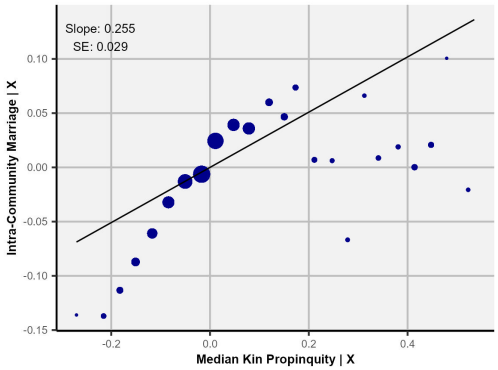
(B) Local Name Index | X



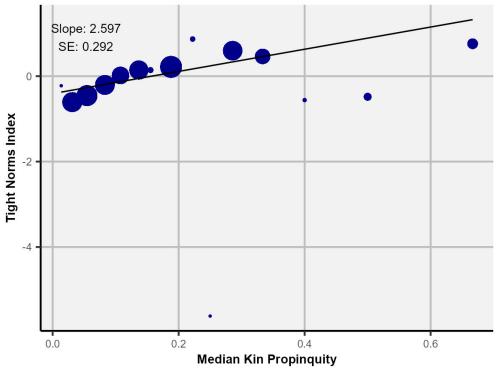
(C) Intra-Community Marriage



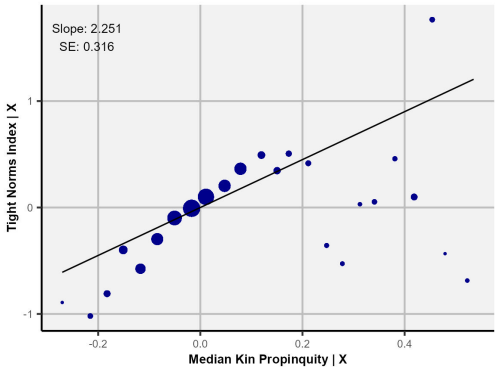
(D) Intra-Community Marriage | X



(E) Tight Norms Index



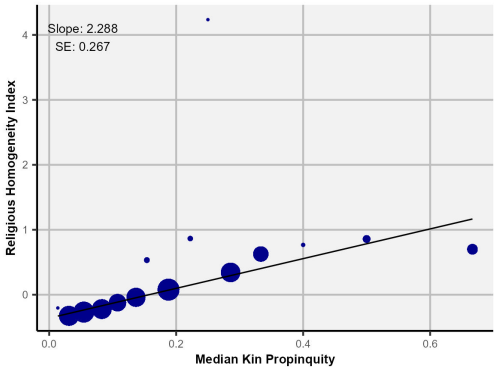
(F) Tight Norms Index | X



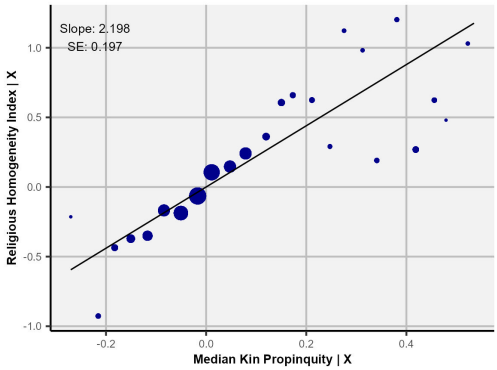
Note: This figure plots the relationship between the MKP and the other historical measures of close-knit communities in 1940: the LNI (subfigures A-B), the ICM (subfigures C-D), and the TNI (subfigures E-F). The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

FIGURE B.5: The MKP and Other Historical Indicators of Close-Communities (cont.)

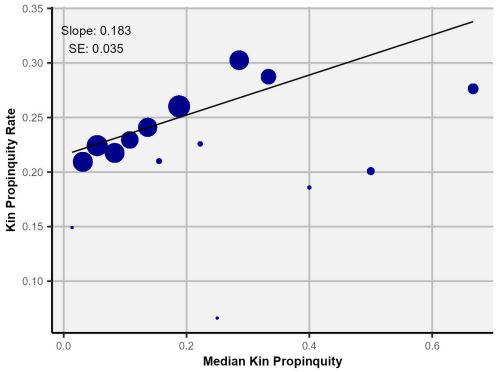
(G) Religious Homogeneity Index



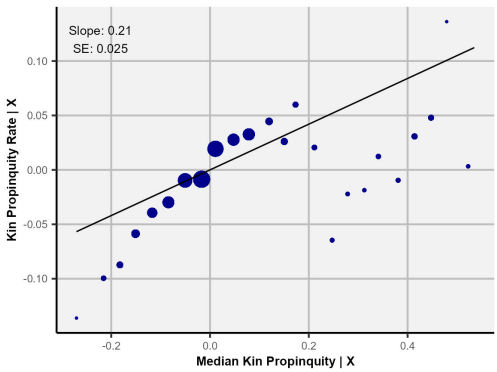
(H) Religious Homogeneity Index | X



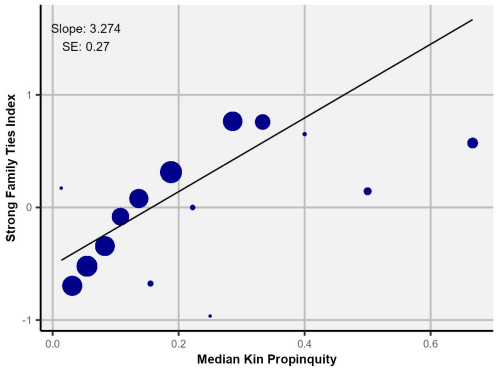
(I) Kin Propinquity Rate



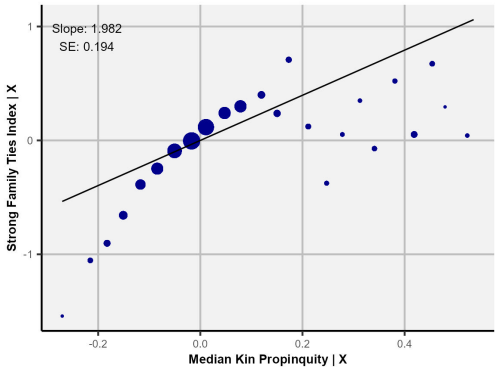
(J) Kin Propinquity Rate | X



(K) Strength of Family Ties Index



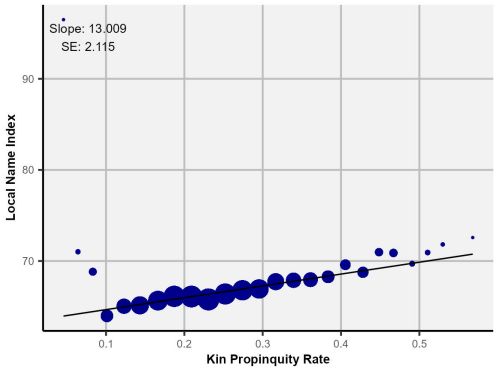
(L) Strength of Family Ties Index | X



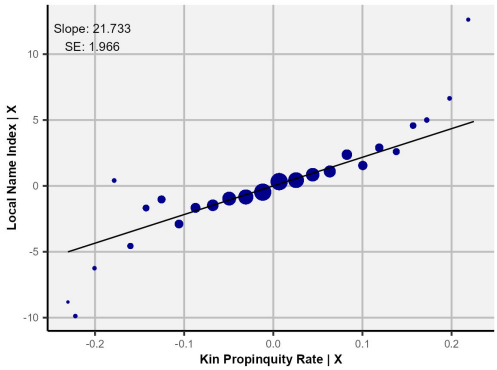
Note: This figure plots the relationship between the MKP and the other historical measures of close-knit communities in 1940: the MKP (subfigures G-H), the KPR (subfigures I-J), and the SFTI (subfigures K-L). The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

FIGURE B.6: The KPR and Other Historical Indicators of Close-Communities

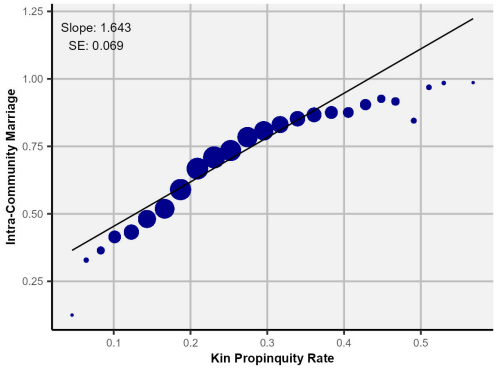
(A) Local Name Index



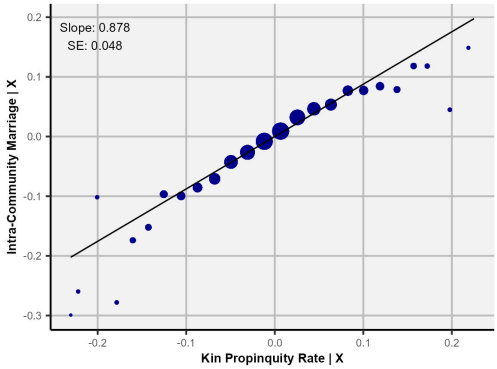
(B) Local Name Index | X



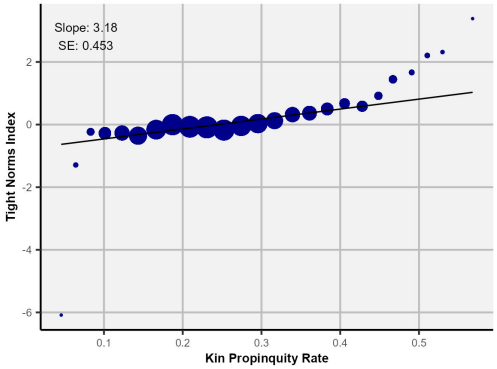
(C) Intra-Community Marriage



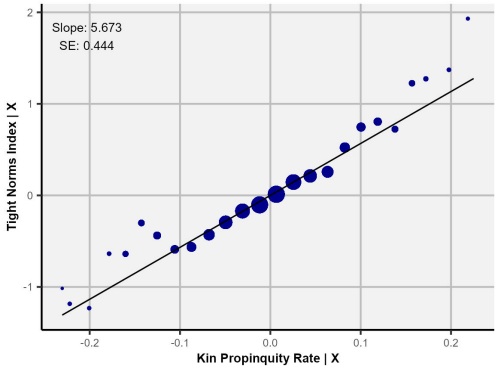
(D) Intra-Community Marriage | X



(E) Tight Norms Index



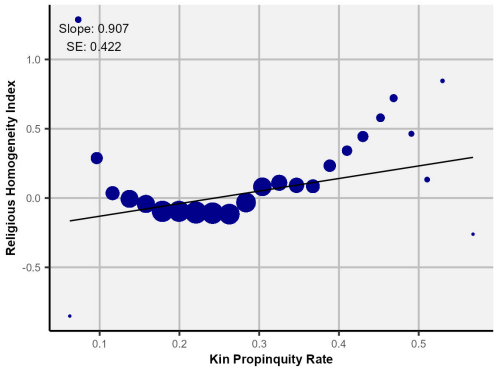
(F) Tight Norms Index | X



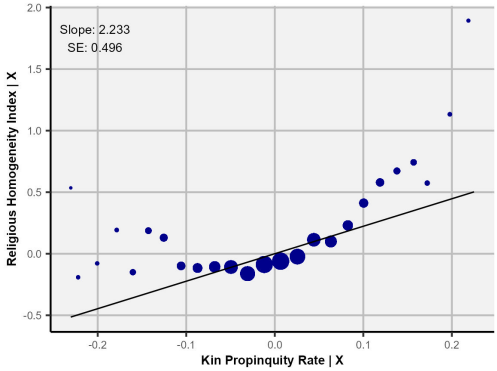
Note: This figure plots the relationship between the KPR and the other historical measures of close-knit communities in 1940: the LNI (subfigures A-B), the ICM (subfigures C-D), and the TNI (subfigures E-F). The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

FIGURE B.6: The KPR and Other Historical Indicators of Close-Communities (cont.)

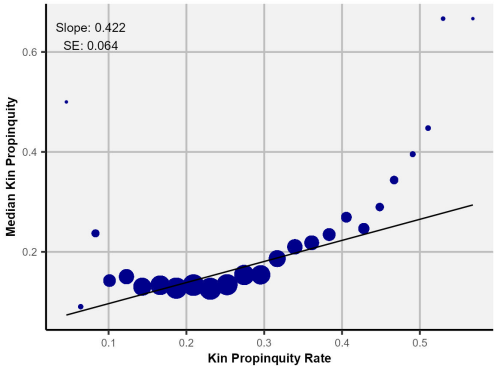
(G) Religious Homogeneity Index



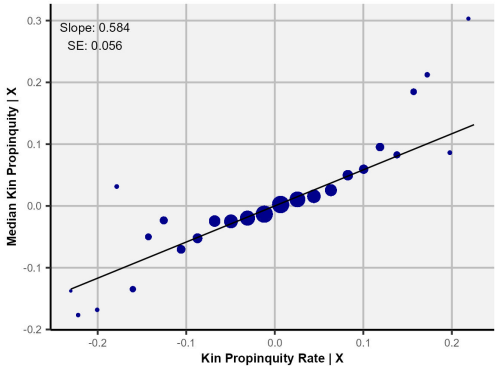
(H) Religious Homogeneity Index | X



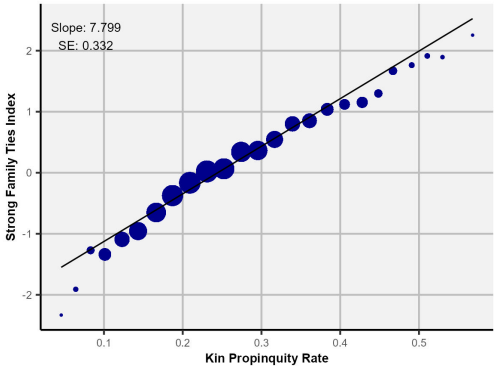
(I) Median Kin Propinquity



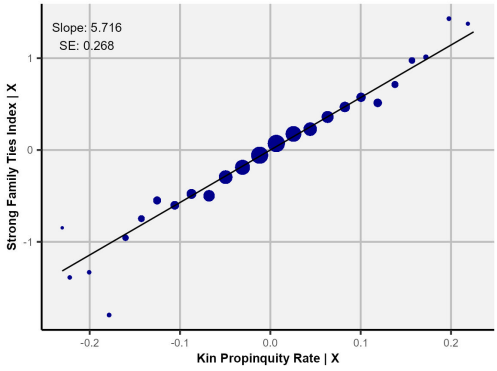
(J) Median Kin Propinquity | X



(K) Strength of Family Ties Index



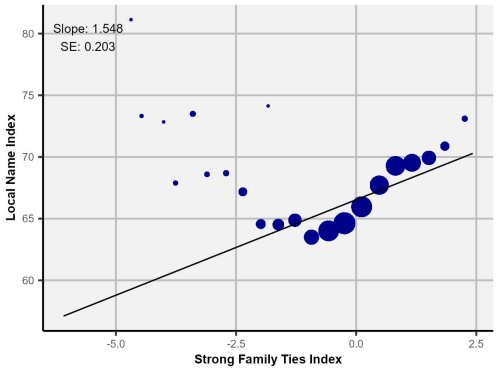
(L) Strength of Family Ties Index | X



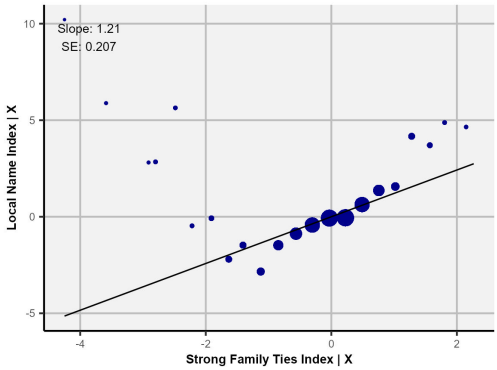
Note: This figure plots the relationship between the KPR and the other historical measures of close-knit communities in 1940: the MKP (subfigures G-H), the KPR (subfigures I-J), and the SFTI (subfigures K-L). The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

FIGURE B.7: The SFTI and Other Historical Indicators of Close-Communities

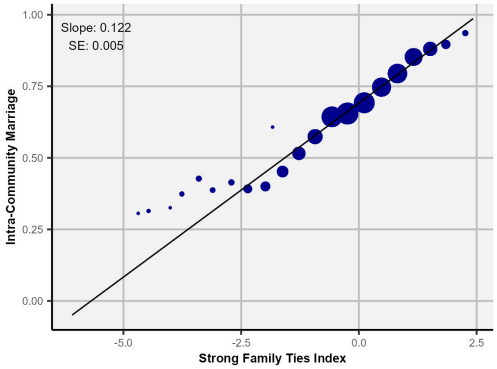
(A) Local Name Index



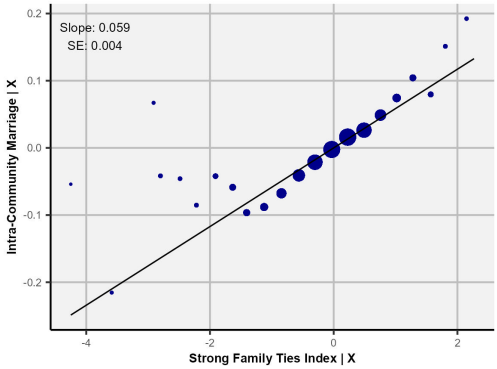
(B) Local Name Index | X



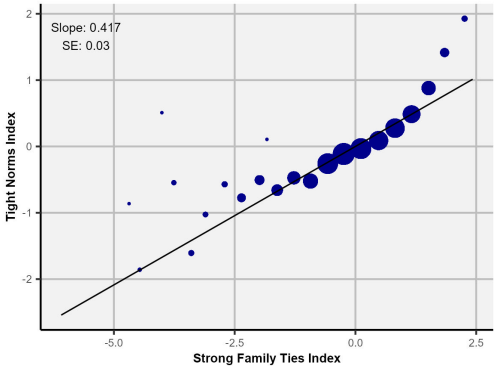
(C) Intra-Community Marriage



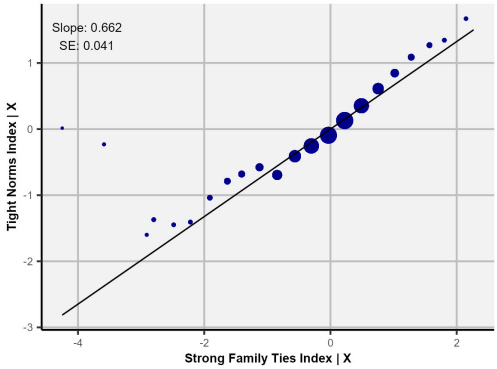
(D) Intra-Community Marriage | X



(E) Tight Norms Index



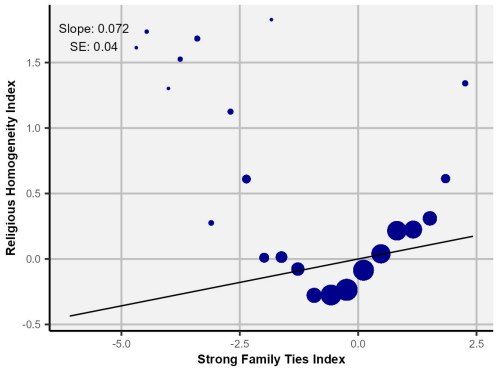
(F) Tight Norms Index | X



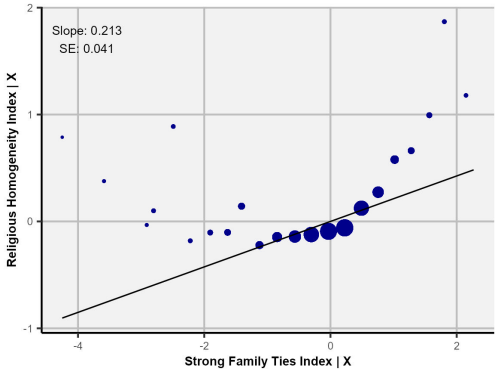
Note: This figure plots the relationship between the SFTI and the other historical measures of close-knit communities in 1940: the LNI (subfigures A-B), the ICM (subfigures C-D), and the TNI (subfigures E-F). The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

FIGURE B.7: The SFTI and Other Historical Indicators of Close-Communities (cont.)

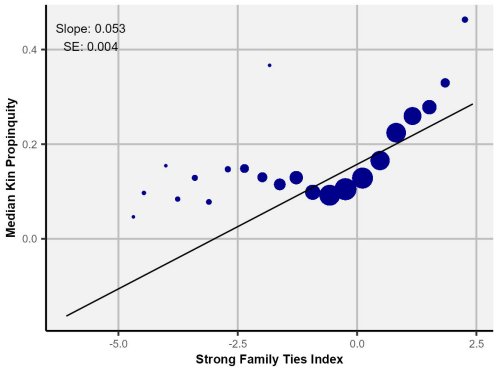
(G) Religious Homogeneity Index



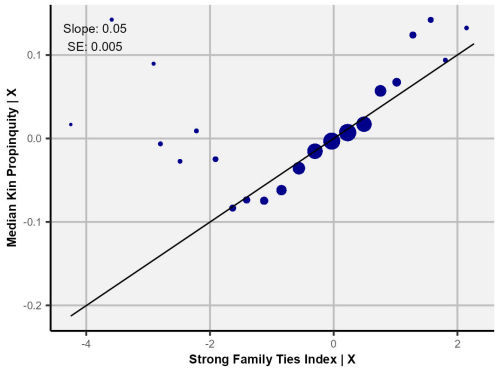
(H) Religious Homogeneity Index | X



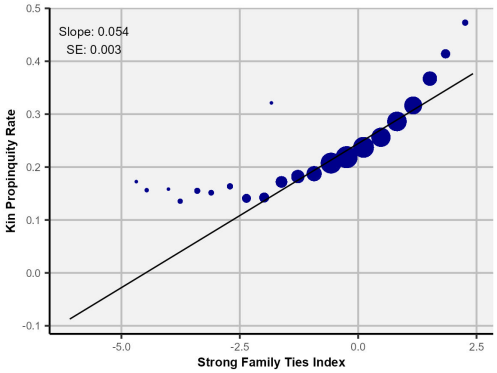
(I) Median Kin Propinquity



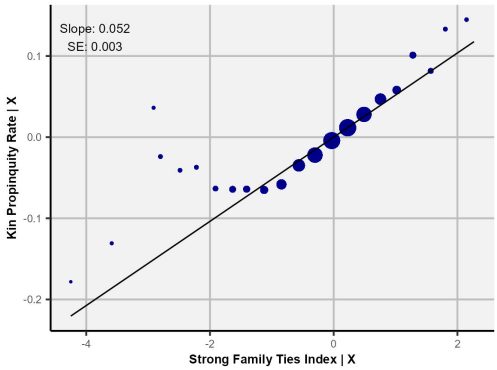
(J) Median Kin Propinquity | X



(K) Kin Propinquity Rate



(L) Kin Propinquity Rate | X

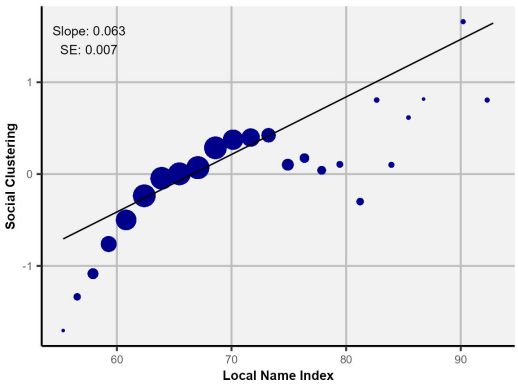


Note: This figure plots the relationship between the SFTI and the other historical measures of close-knit communities in 1940: the MKP (subfigures G-H), the KPR (subfigures I-J), and the SFTI (subfigures K-L). The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

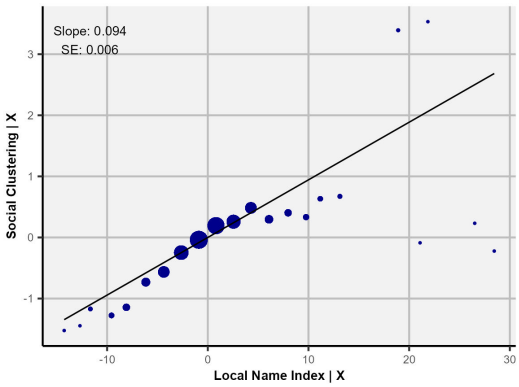
B.3.2 Relationship with Contemporary Close-Knit Social Networks

FIGURE B.8: The LNI and Contemporary Close-knit Social Networks

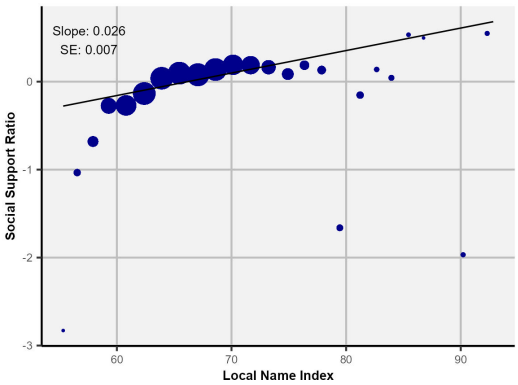
(A) Social Clustering



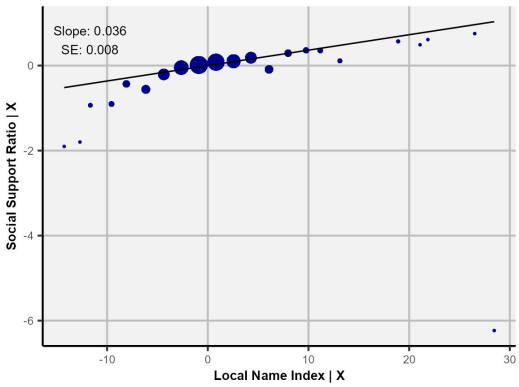
(B) Social Clustering | X



(C) Social Support Ratio



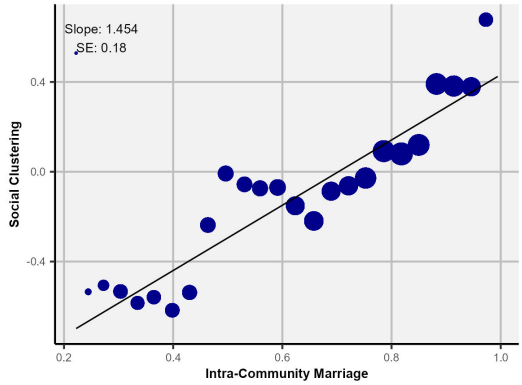
(D) Social Support Ratio | X



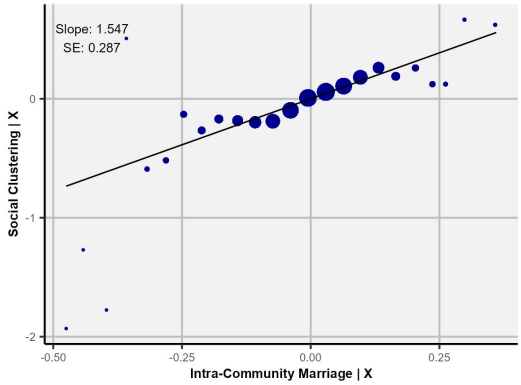
Note: This figure plots the relationship between the LNI in 1940 and contemporary measures of close-knit social networks (Chetty et al., 2022a,b): Social Clustering (subfigures A-B) and Social Support Ratio (subfigures C-D). Observations are counties. The left column presents the raw relationship and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

FIGURE B.9: The ICM and Contemporary Close-knit Social Networks

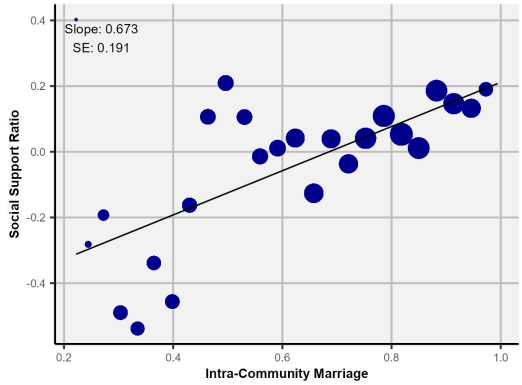
(A) Social Clustering



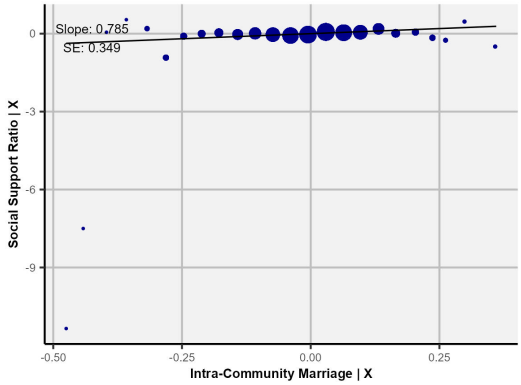
(B) Social Clustering | X



(C) Social Support Ratio

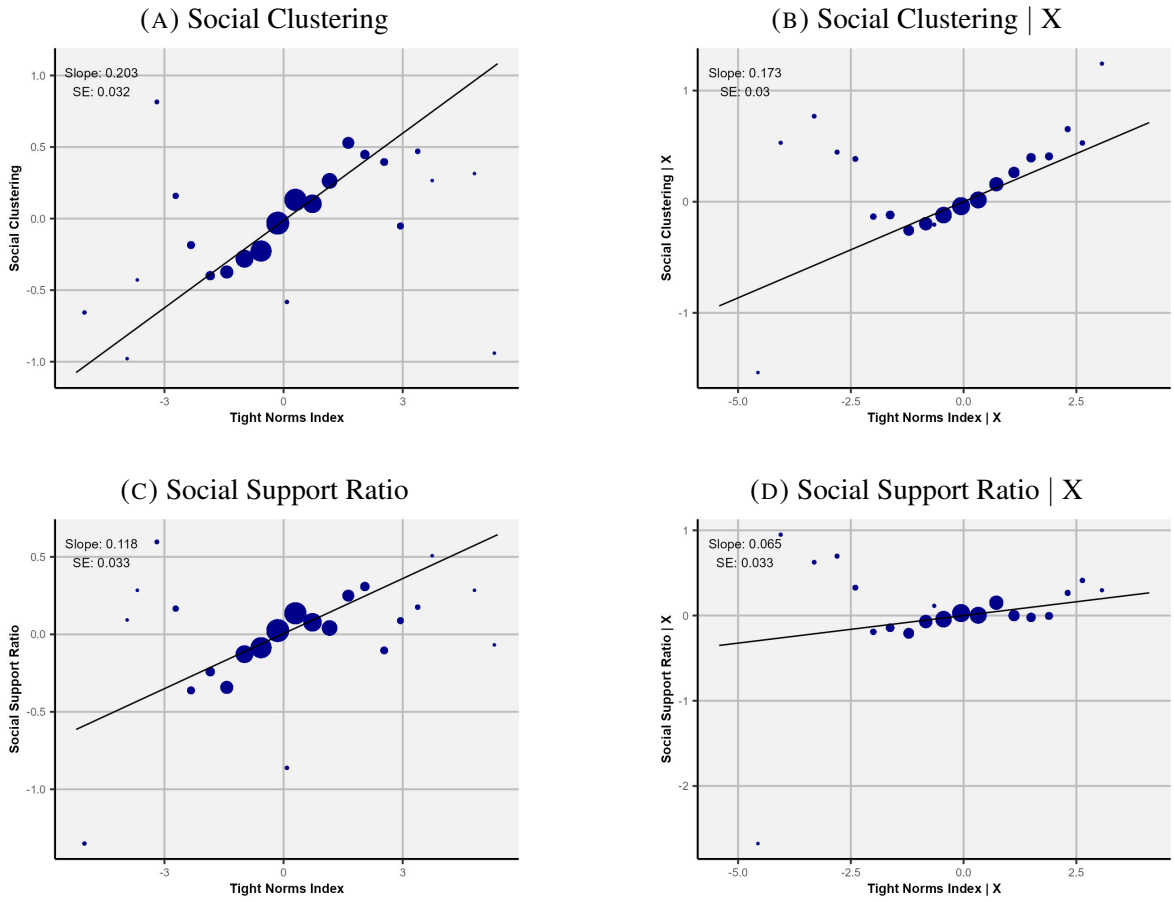


(D) Social Support Ratio | X



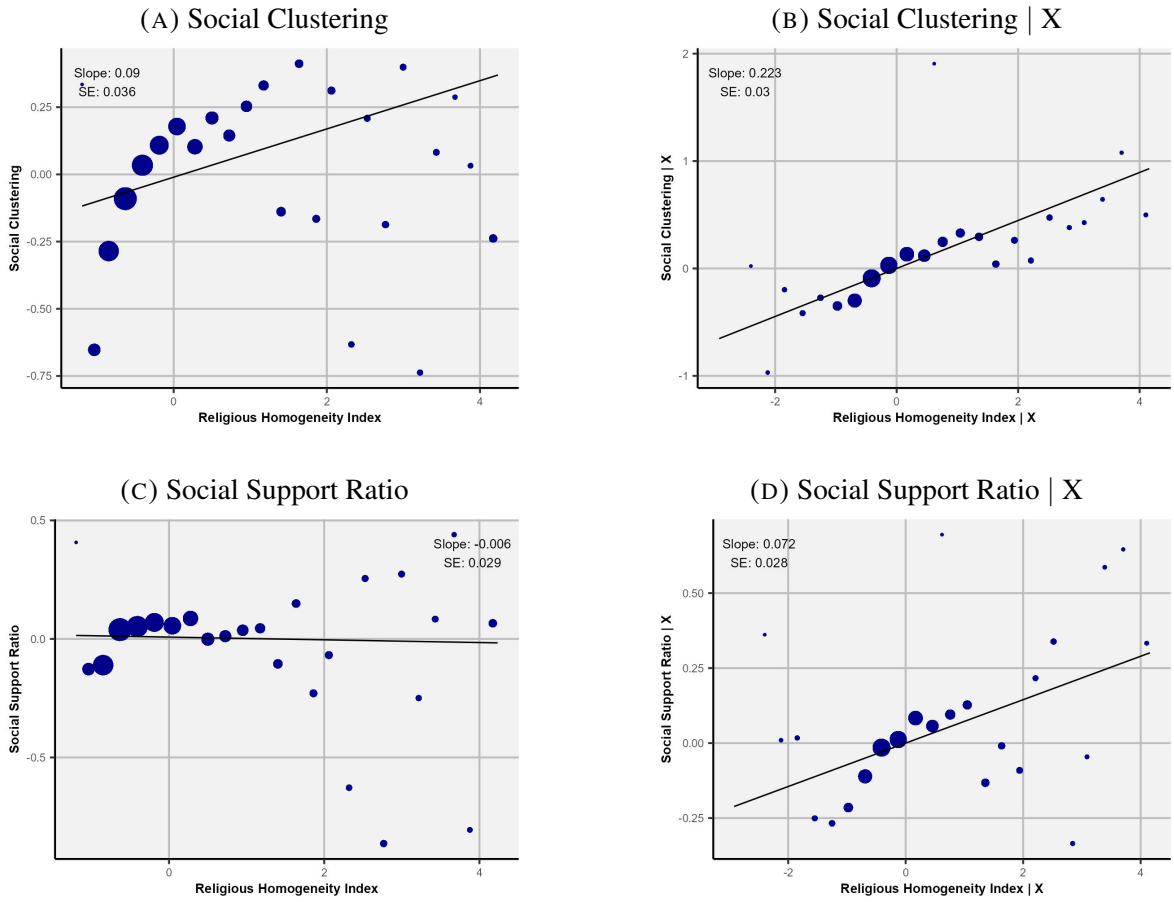
Note: This figure plots the relationship between the ICM in 1940 and contemporary measures of close-knit social networks (Chetty et al., 2022a,b): Social Clustering (subfigures A-B) and Social Support Ratio (subfigures C-D). Observations are counties. The left column presents the raw relationship and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

FIGURE B.10: The TNI and Contemporary Close-knit Social Networks



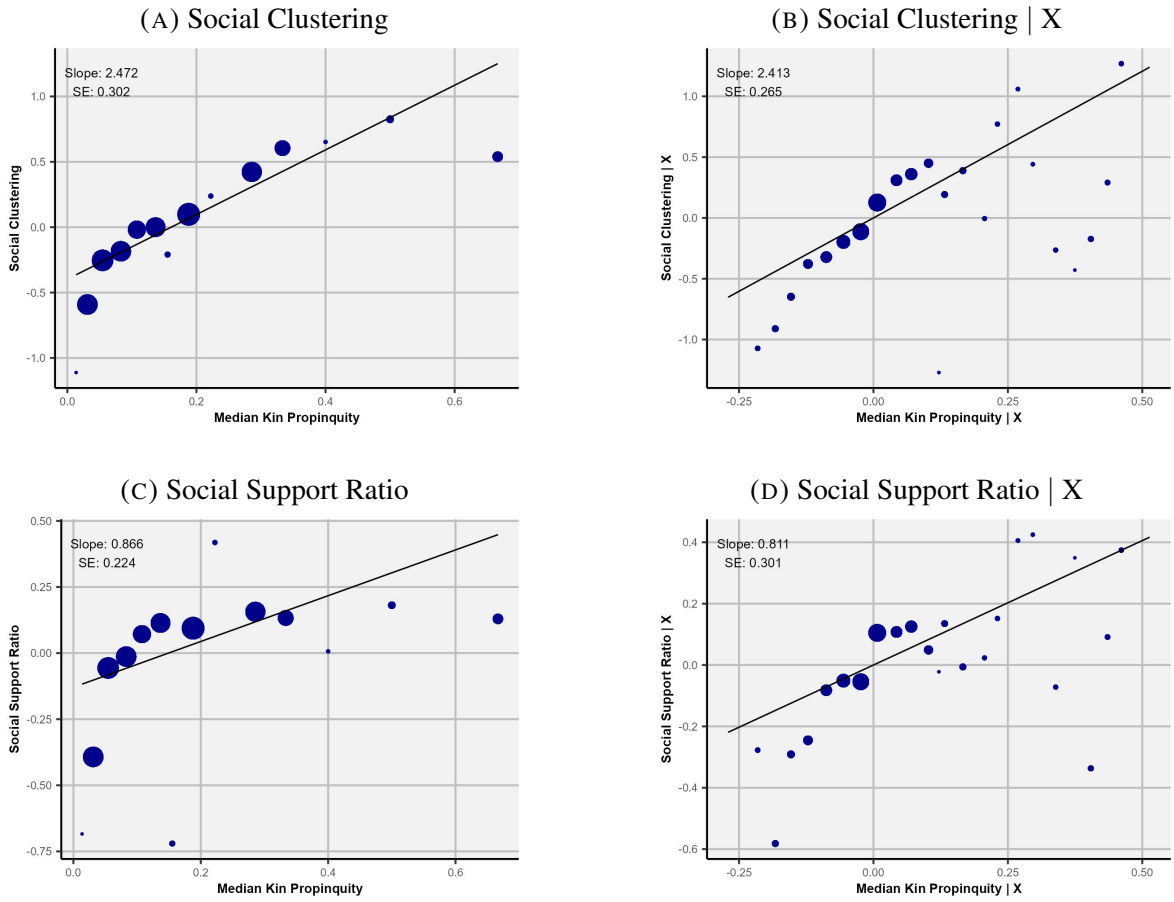
Note: This figure plots the relationship between the TNI in 1940 and contemporary measures of close-knit social networks (Chetty et al., 2022a,b): Social Clustering (subfigures A-B) and Social Support Ratio (subfigures C-D). Observations are counties. The left column presents the raw relationship and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

FIGURE B.11: The RHI and Contemporary Close-knit Social Networks



Note: This figure plots the relationship between the RHI in 1940 and contemporary measures of close-knit social networks (Chetty et al., 2022a,b): Social Clustering (subfigures A-B) and Social Support Ratio (subfigures C-D). Observations are counties. The left column presents the raw relationship and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

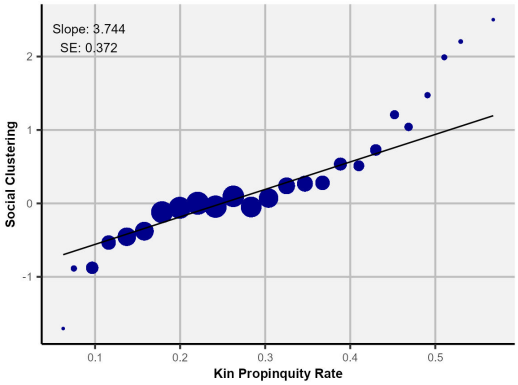
FIGURE B.12: The MKP and Contemporary Close-knit Social Networks



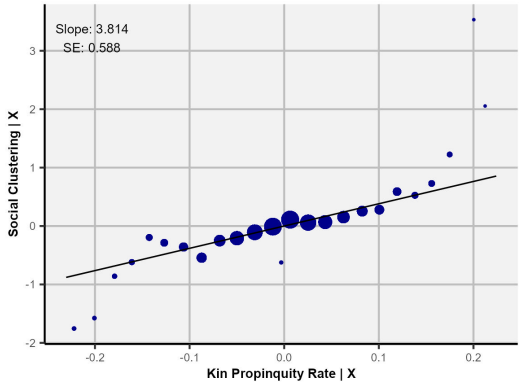
Note: This figure plots the relationship between the MKP in 1940 and contemporary measures of close-knit social networks (Chetty et al., 2022a,b): Social Clustering (subfigures A-B) and Social Support Ratio (subfigures C-D). Observations are counties. The left column presents the raw relationship and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

FIGURE B.13: The KPR and Contemporary Close-knit Social Networks

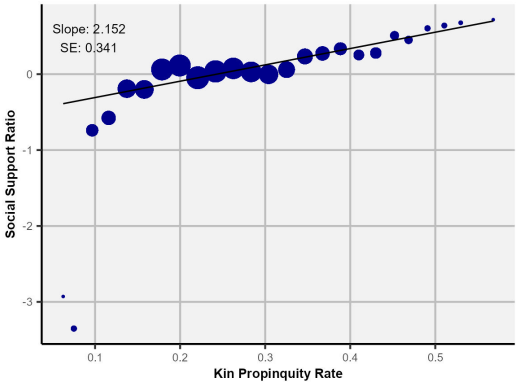
(A) Social Clustering



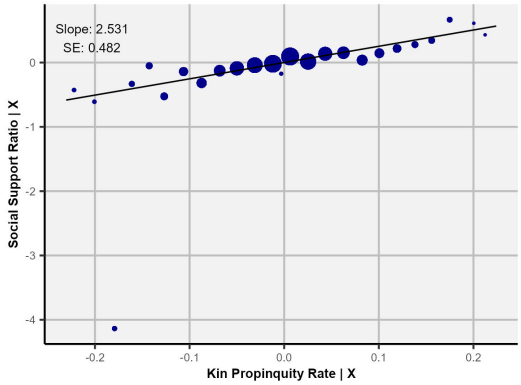
(B) Social Clustering | X



(C) Social Support Ratio

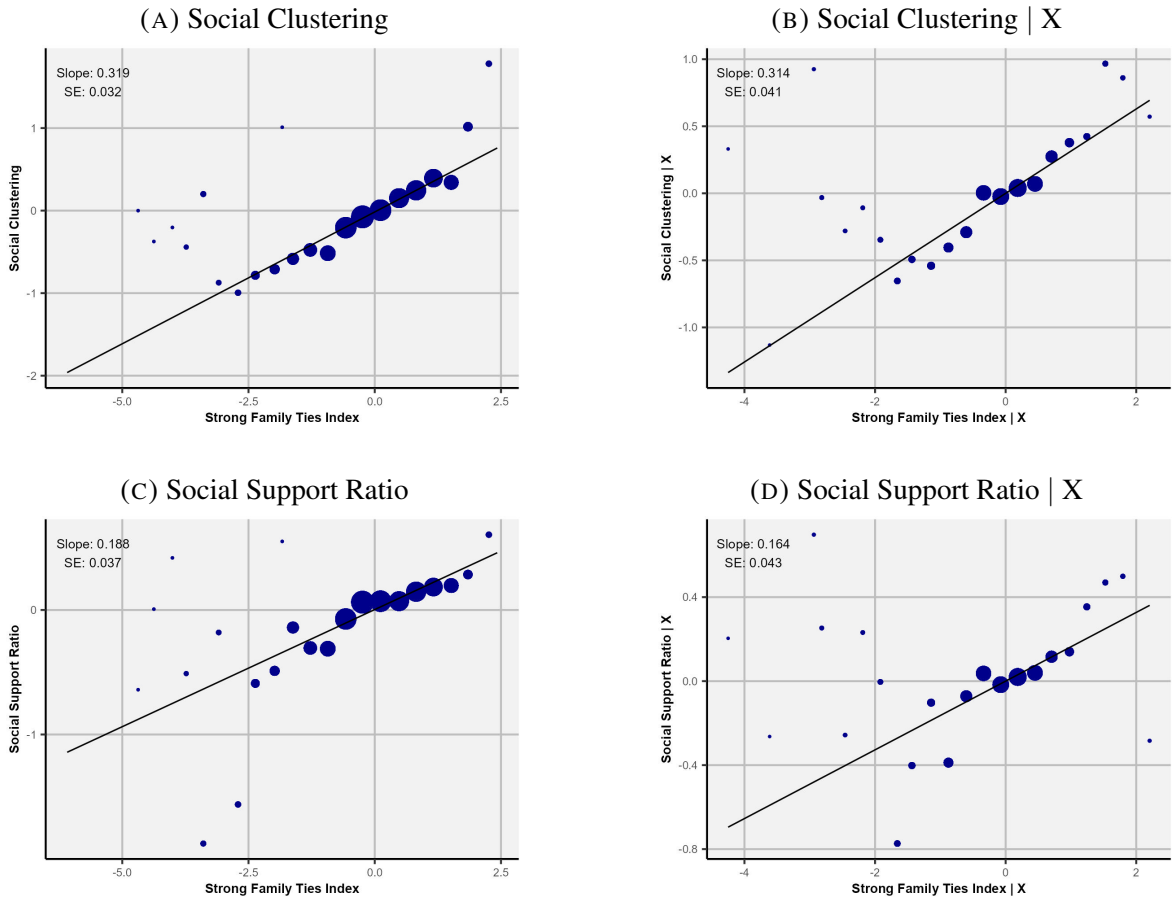


(D) Social Support Ratio | X



Note: This figure plots the relationship between the KPR in 1940 and contemporary measures of close-knit social networks (Chetty et al., 2022a,b): Social Clustering (subfigures A-B) and Social Support Ratio (subfigures C-D). Observations are counties. The left column presents the raw relationship and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

FIGURE B.14: The SFTI and Contemporary Close-knit Social Networks

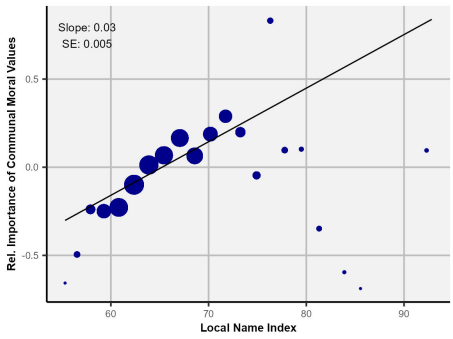


Note: This figure plots the relationship between the SFTI in 1940 and contemporary measures of close-knit social networks (Chetty et al., 2022a,b): Social Clustering (subfigures A-B) and Social Support Ratio (subfigures C-D). Observations are counties. The left column presents the raw relationship and the right column presents the conditional relationship after partialling out state fixed effects. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

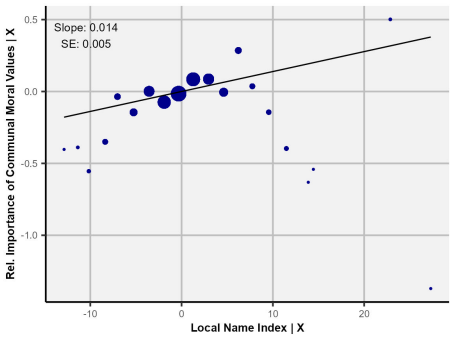
B.3.3 Relationship with Contemporary Cultural and Psychological Characteristics

FIGURE B.15: The LNI and Contemporary Cultural and Psychological Characteristics

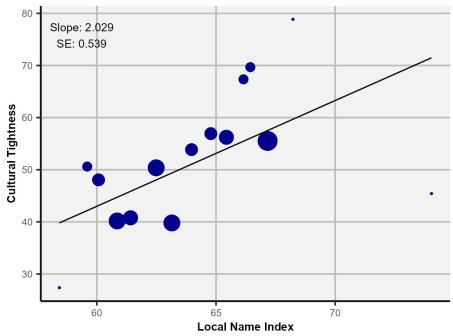
(A) Communal Moral Values



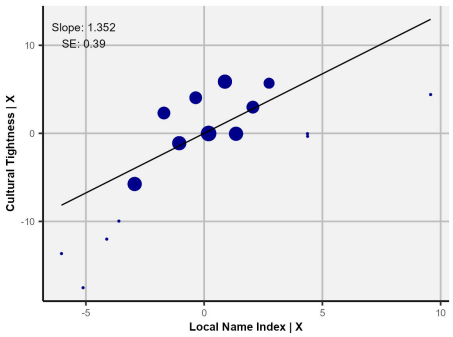
(B) Communal Moral Values | X



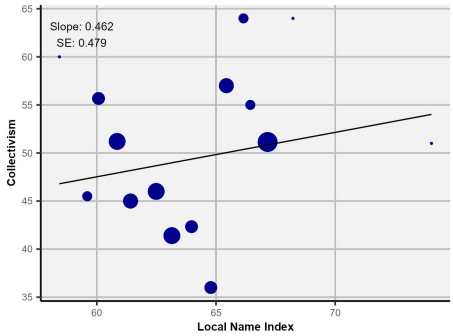
(C) Cultural Tightness



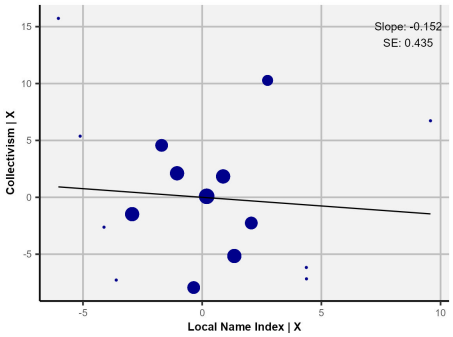
(D) Cultural Tightness | X



(E) Collectivism

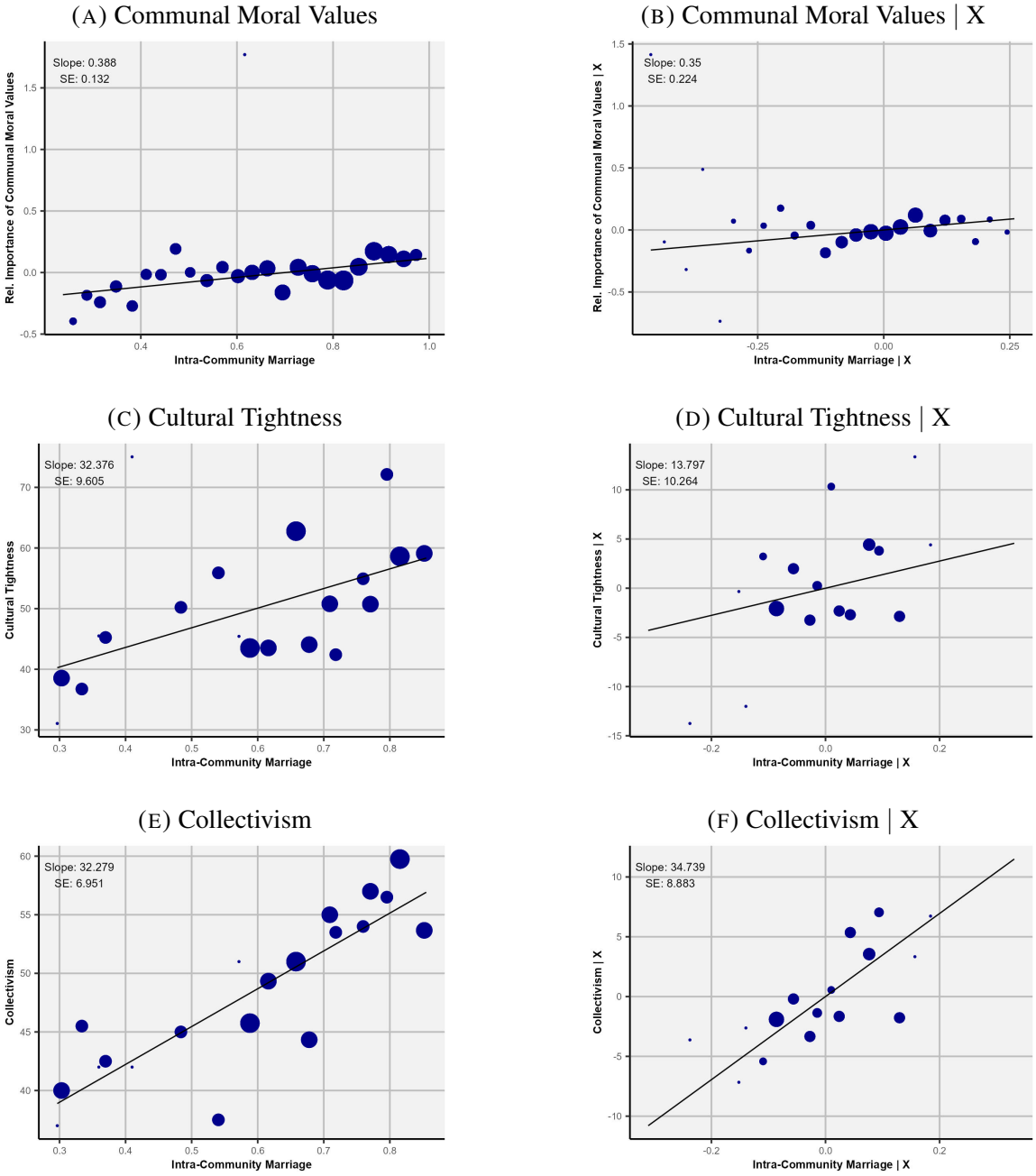


(F) Collectivism | X



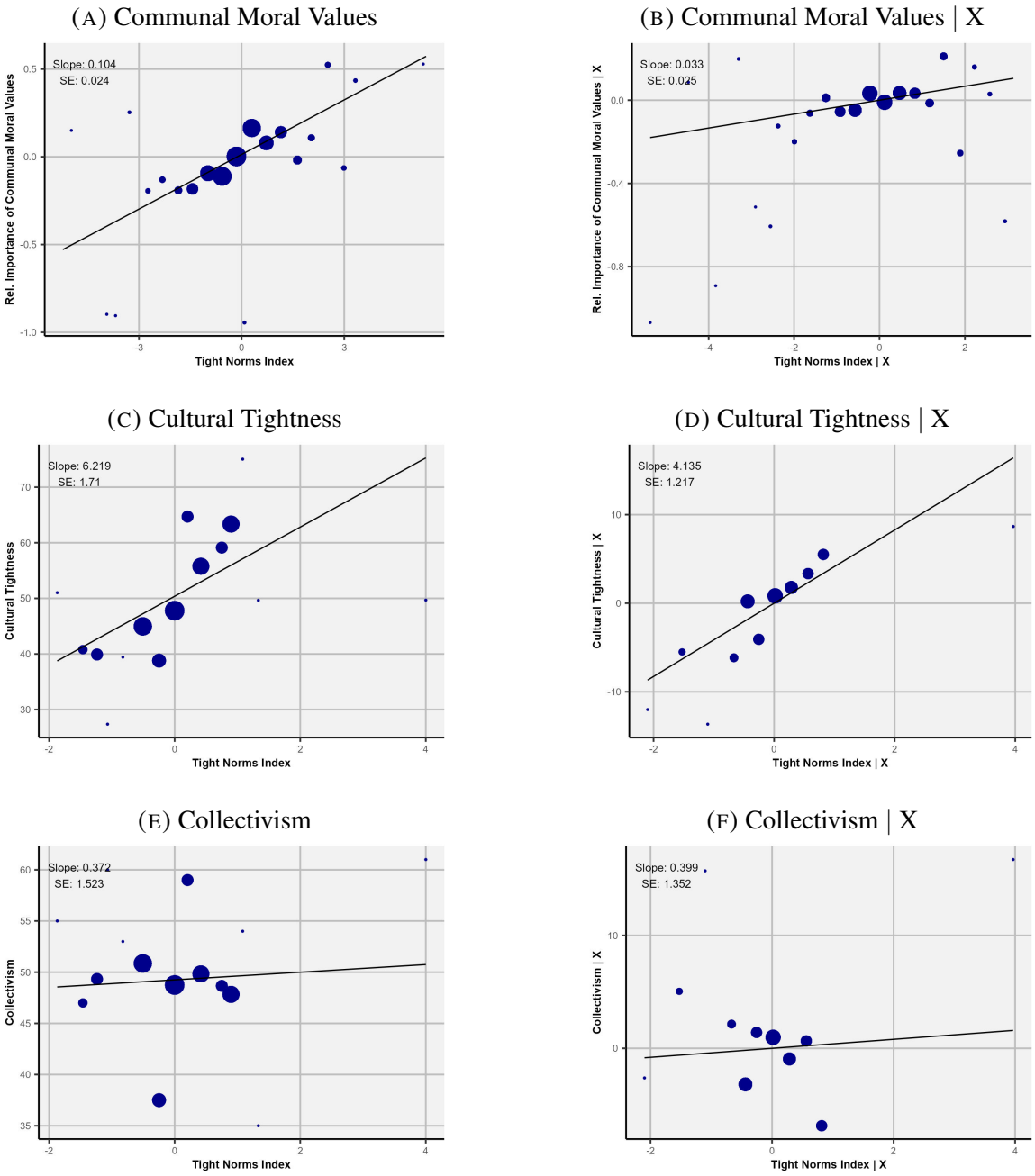
Note: This figure plots the relationship between the LNI in 1940 and contemporary cultural and psychological features of close-knit communities: Communal Moral Values (subfigures A-B) (Enke, 2020), Cultural Tightness (subfigures C-D) (Harrington and Gelfand, 2014), and Collectivism (subfigures E-F) (Vandello and Cohen, 1999). Observations are counties in subfigures A-B and states in subfigures C-F. The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects in subfigures A-B and census region fixed effects in subfigures C-F. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011) in subfigures A-B.

FIGURE B.16: The ICM and Contemporary Cultural and Psychological Characteristics



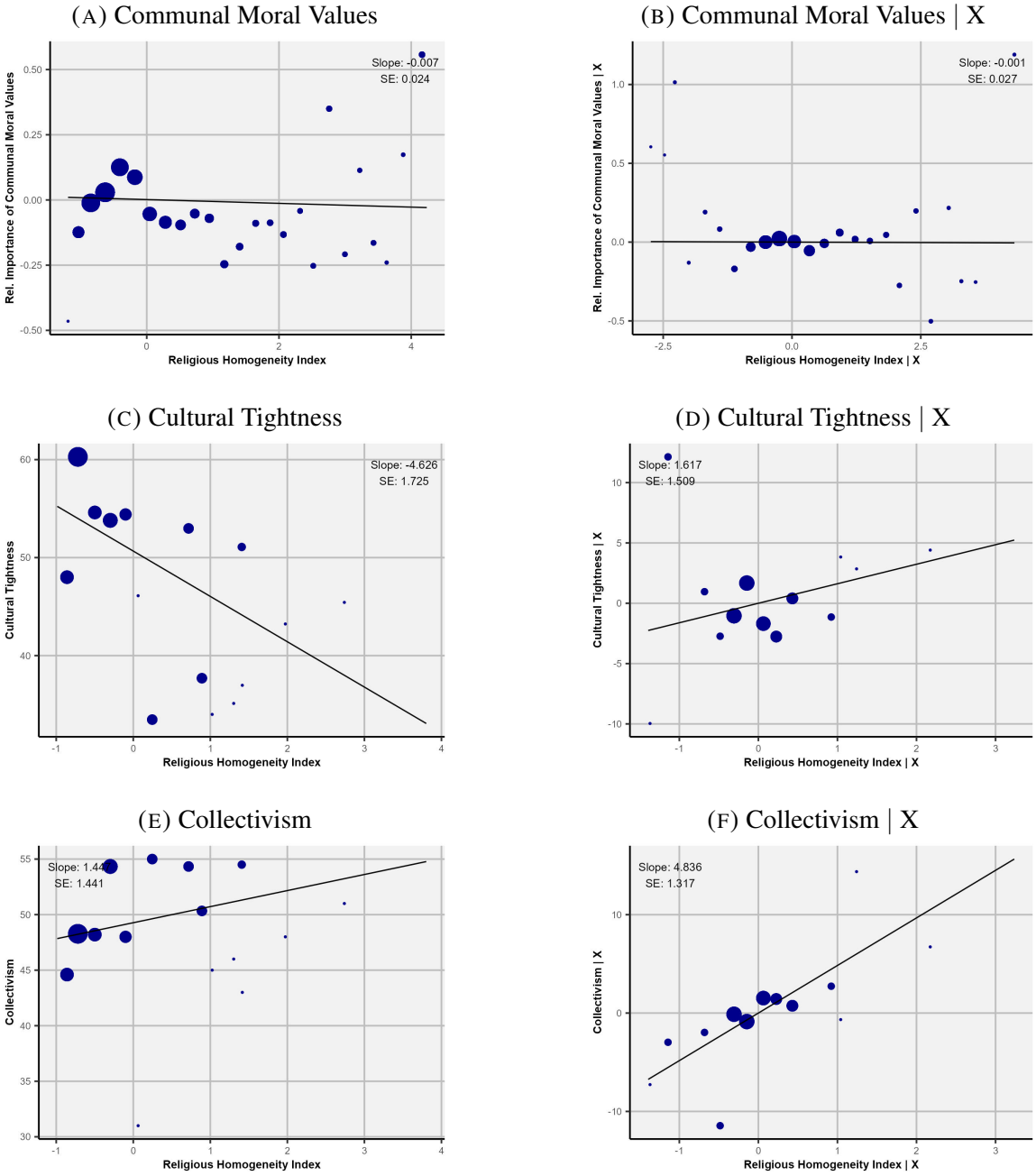
Note: This figure plots the relationship between the ICM in 1940 and contemporary cultural and psychological features of close-knit communities: Communal Moral Values (subfigures A-B) (Enke, 2020), Cultural Tightness (subfigures C-D) (Harrington and Gelfand, 2014), and Collectivism (subfigures E-F) (Vandello and Cohen, 1999). Observations are counties in subfigures A-B and states in subfigures C-F. The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects in subfigures A-B and census region fixed effects in subfigures C-F. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011) in subfigures A-B.

FIGURE B.17: The TNI and Contemporary Cultural and Psychological Characteristics



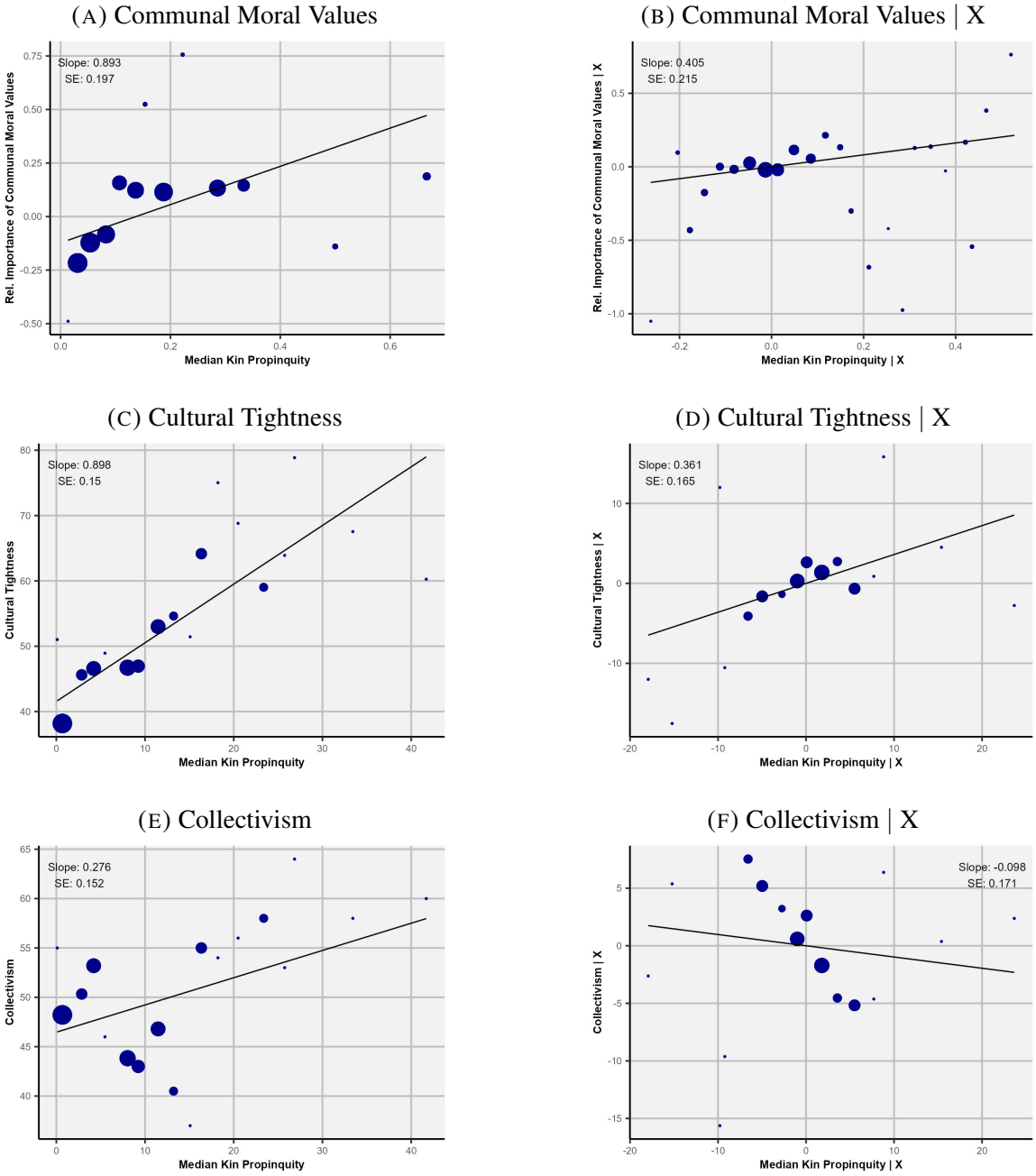
Note: This figure plots the relationship between the TNI in 1940 and contemporary cultural and psychological features of close-knit communities: Communal Moral Values (subfigures A-B) (Enke, 2020), Cultural Tightness (subfigures C-D) (Harrington and Gelfand, 2014), and Collectivism (subfigures E-F) (Vandello and Cohen, 1999). Observations are counties in subfigures A-B and states in subfigures C-F. The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects in subfigures A-B and census region fixed effects in subfigures C-F. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011) in subfigures A-B.

FIGURE B.18: The RHI and Contemporary Cultural and Psychological Characteristics



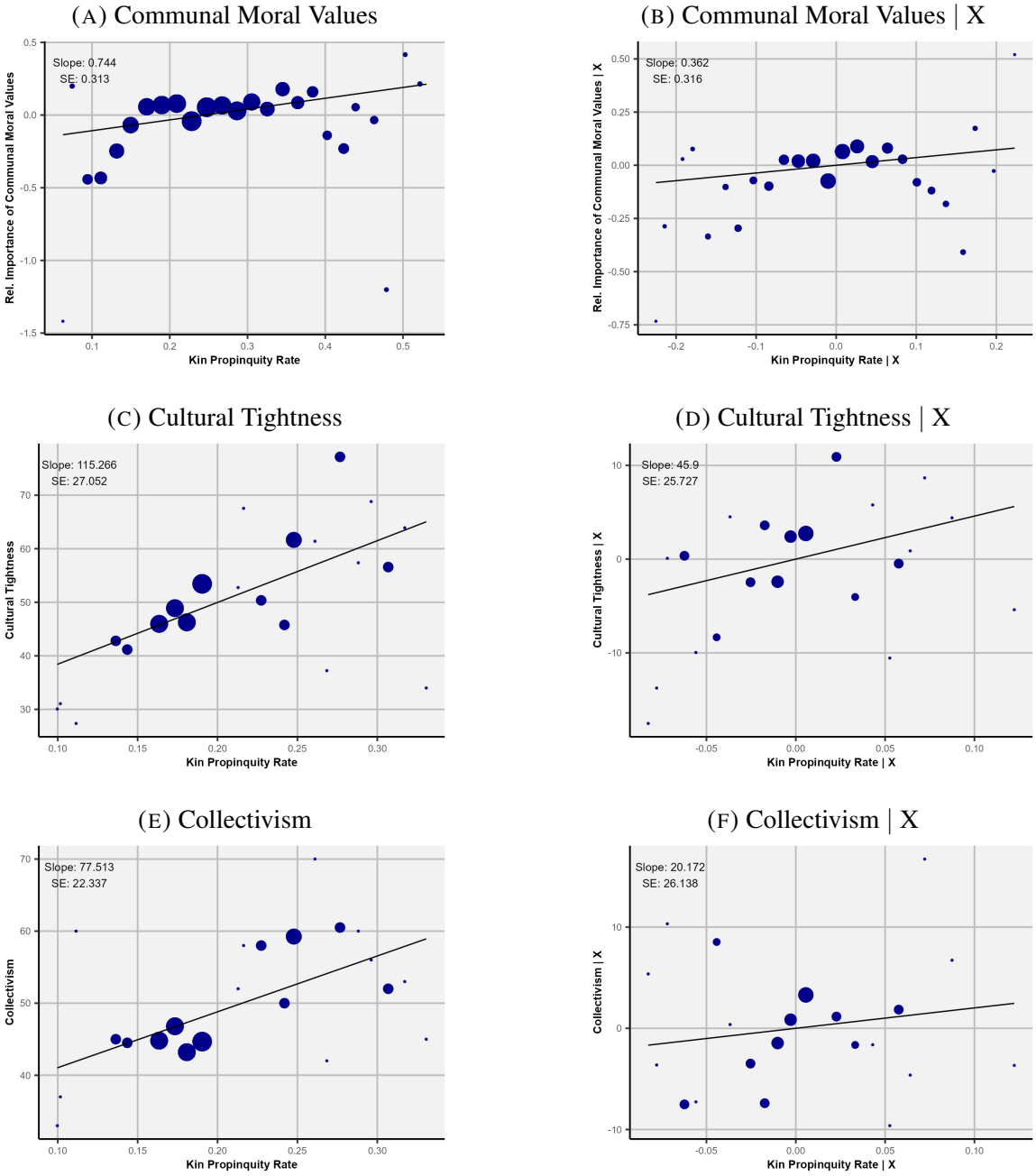
Note: This figure plots the relationship between the RHI in 1940 and contemporary cultural and psychological features of close-knit communities: Communal Moral Values (subfigures A-B) (Enke, 2020), Cultural Tightness (subfigures C-D) (Harrington and Gelfand, 2014), and Collectivism (subfigures E-F) (Vandello and Cohen, 1999). Observations are counties in subfigures A-B and states in subfigures C-F. The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects in subfigures A-B and census region fixed effects in subfigures C-F. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011) in subfigures A-B.

FIGURE B.19: The MKP and Contemporary Cultural and Psychological Characteristics



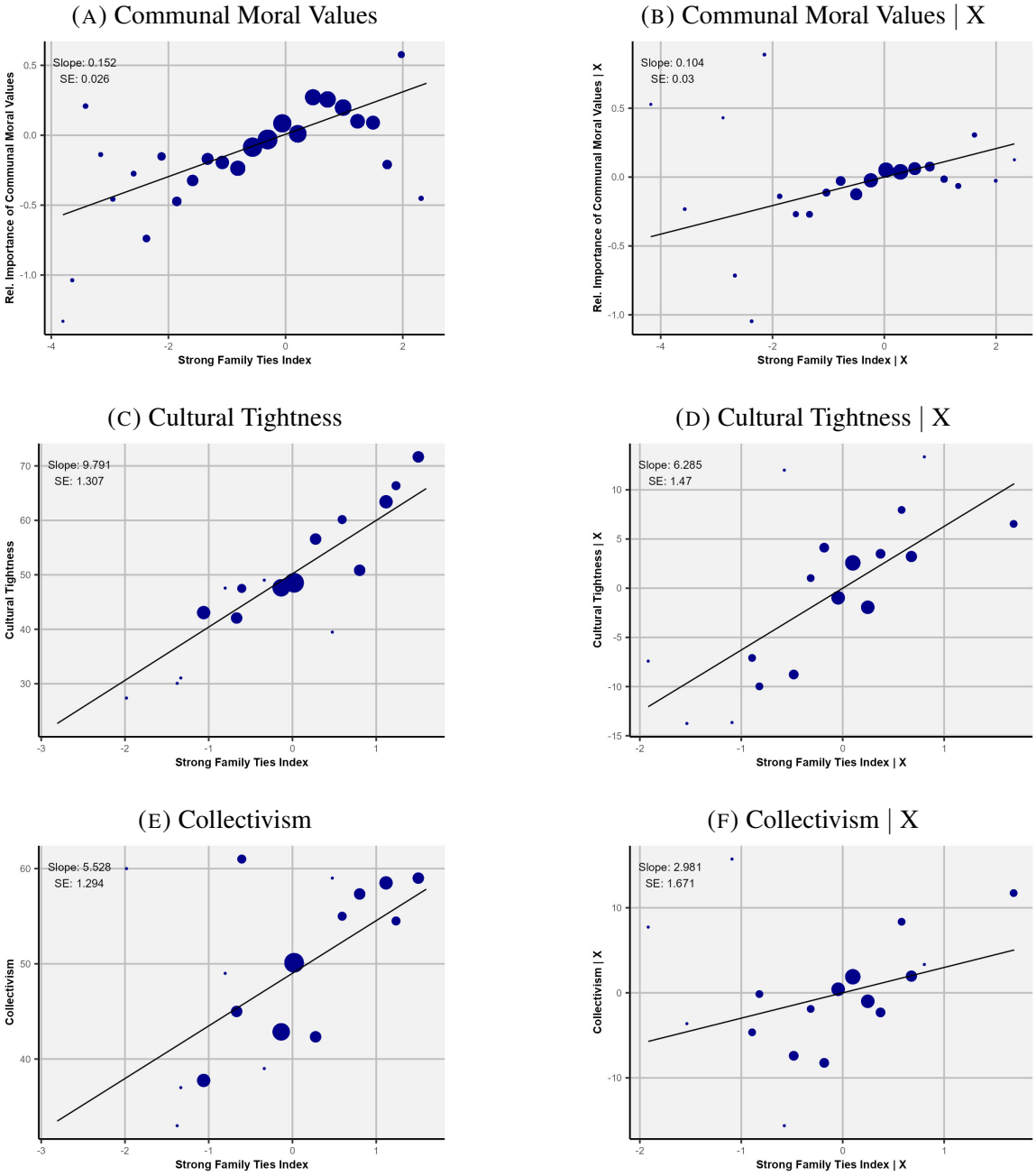
Note: This figure plots the relationship between the MKP in 1940 and contemporary cultural and psychological features of close-knit communities: Communal Moral Values (subfigures A-B) (Enke, 2020), Cultural Tightness (subfigures C-D) (Harrington and Gelfand, 2014), and Collectivism (subfigures E-F) (Vandello and Cohen, 1999). Observations are counties in subfigures A-B and states in subfigures C-F. The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects in subfigures A-B and census region fixed effects in subfigures C-F. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011) in subfigures A-B.

FIGURE B.20: The KPR and Contemporary Cultural and Psychological Characteristics



Note: This figure plots the relationship between the KPR in 1940 and contemporary cultural and psychological features of close-knit communities: Communal Moral Values (subfigures A-B) (Enke, 2020), Cultural Tightness (subfigures C-D) (Harrington and Gelfand, 2014), and Collectivism (subfigures E-F) (Vandello and Cohen, 1999). Observations are counties in subfigures A-B and states in subfigures C-F. The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects in subfigures A-B and census region fixed effects in subfigures C-F. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011) in subfigures A-B.

FIGURE B.21: The SFTI and Contemporary Cultural and Psychological Characteristics

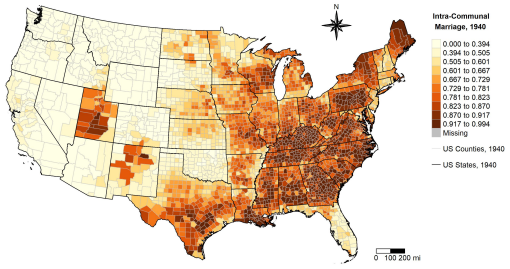


Note: This figure plots the relationship between the SFTI in 1940 and contemporary cultural and psychological features of close-knit communities: Communal Moral Values (subfigures A-B) (Enke, 2020), Cultural Tightness (subfigures C-D) (Harrington and Gelfand, 2014), and Collectivism (subfigures E-F) (Vandello and Cohen, 1999). Observations are counties in subfigures A-B and states in subfigures C-F. The left column presents the raw relationship, and the right column presents the conditional relationship after partialling out state fixed effects in subfigures A-B and census region fixed effects in subfigures C-F. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011) in subfigures A-B.

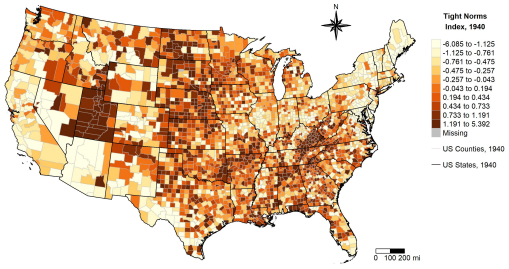
B.4 The Spatial Distribution of Close-Knit Communities

FIGURE B.22: Mapping Close-Knit Communities: County-Level Measures, 1940

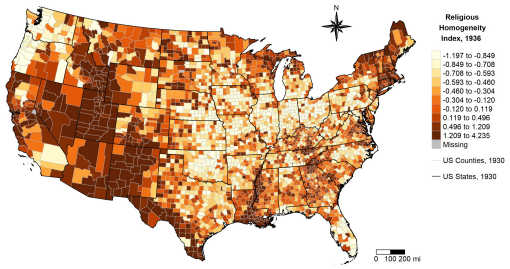
(A) Intra-Community Marriage



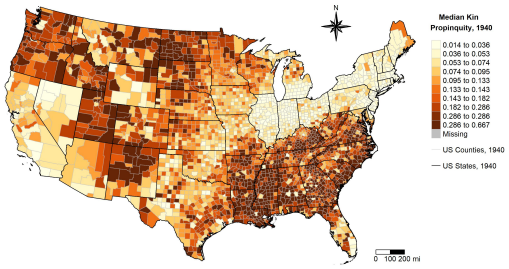
(B) Tight Norms Index



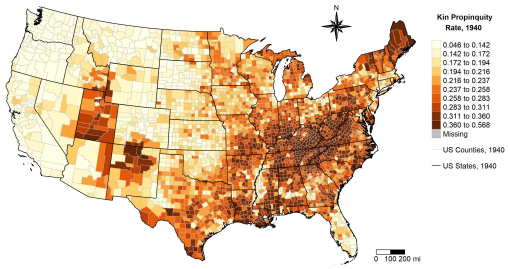
(C) Religious Homogeneity Index



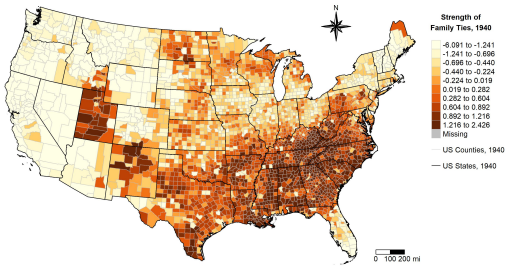
(D) Median Kin Propinquity



(E) Kin Propinquity Rate



(F) Strength of Family Ties Index



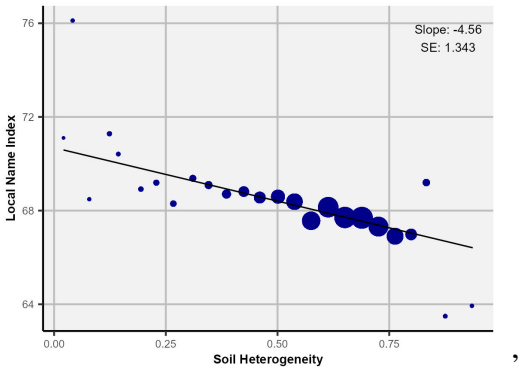
Note: This figure plots the different county-level measures of close-knit communities in 1940: the share of ICM (subfigure A), the TNI (subfigure B), the RHI (subfigure C), the MKP (subfigure D), the KPR (subfigure E), and the SFTI (subfigure F). Subfigures A-B and D-F use data from the 1940 census and limit the sample to native-born whites. Subfigure C uses data from the 1936 census of religious bodies. In all subfigures darker colors indicate a more close-knit community. See Appendix E for details on the data and variable construction.

C Further Analysis and Results

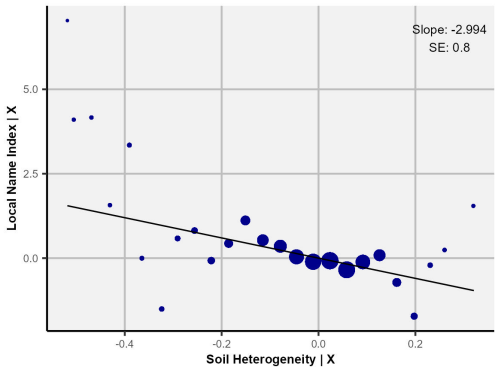
C.1 Visualization of County-Level Results

FIGURE C.1: Bin Scatter Plots

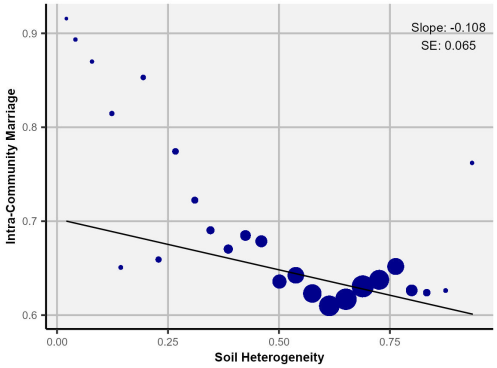
(A) Local Name Index



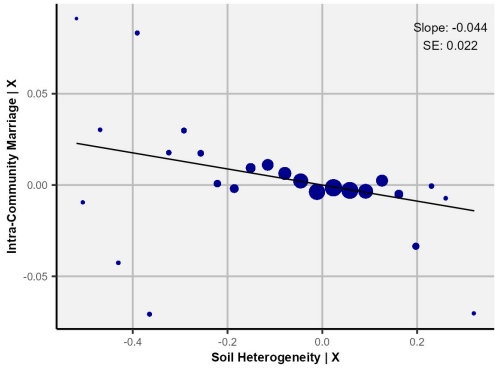
(B) Local Name Index | X



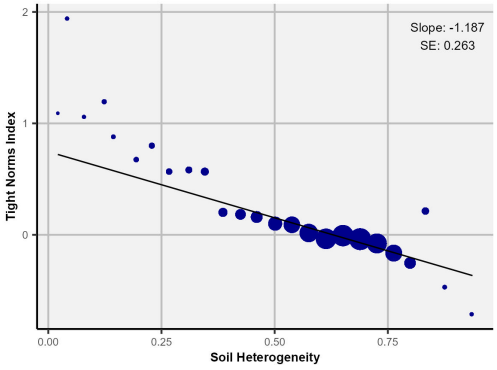
(C) Intra-Community Marriage



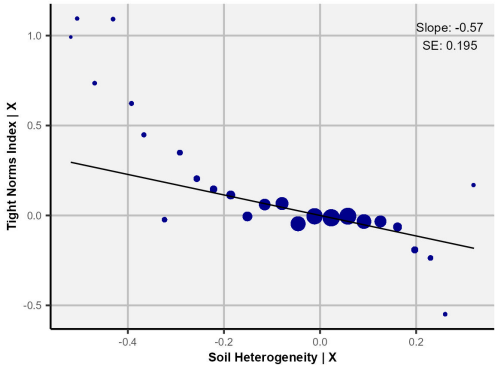
(D) Intra-Community Marriage | X



(E) Tight Norms Index



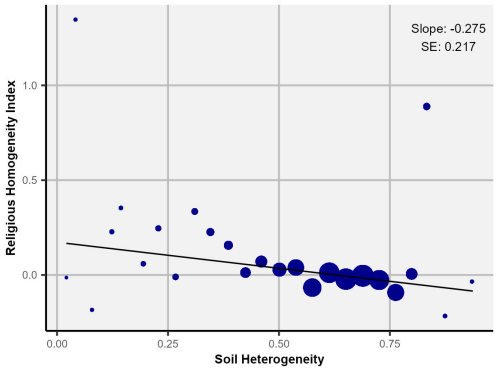
(F) Tight Norms Index | X



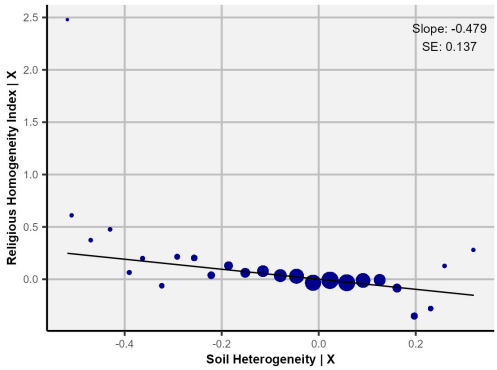
Note: Figure continues to the next page.

FIGURE C.1: Bin Scatter Plots (cont.)

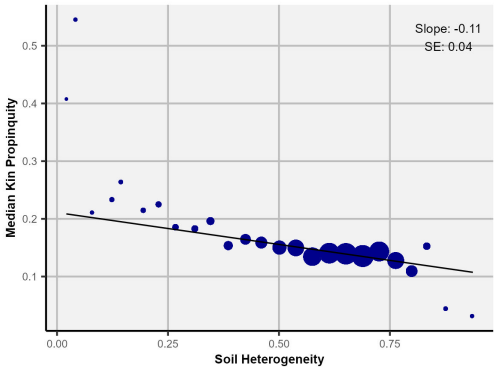
(G) Religious Homogeneity Index



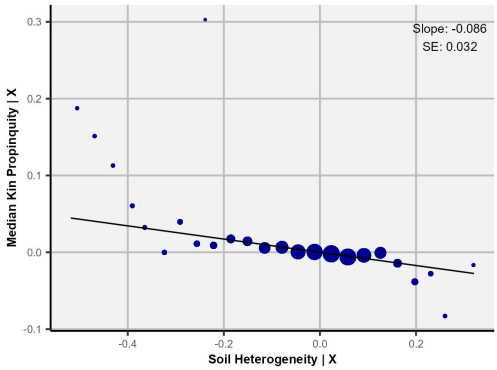
(H) Religious Homogeneity Index | X



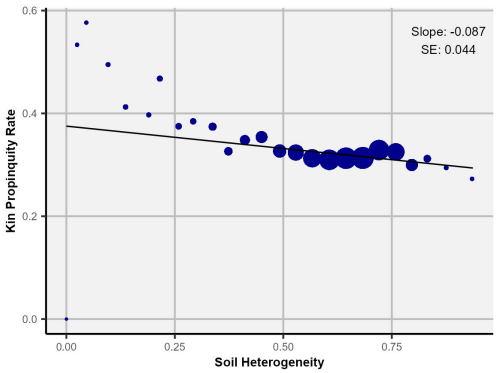
(I) Median Kin Propinquity



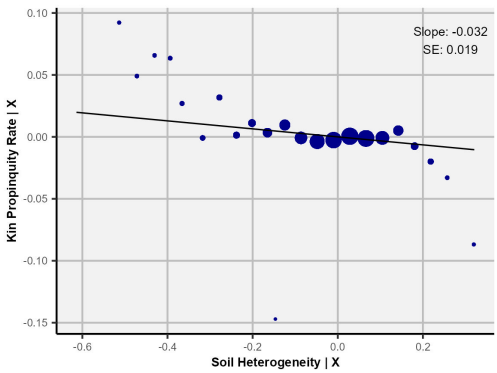
(J) Median Kin Propinquity | X



(K) Kin Propinquity Rate



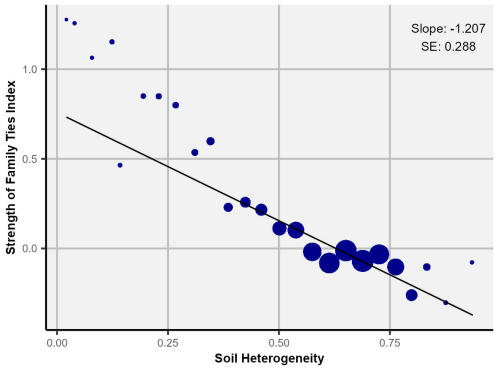
(L) Kin Propinquity Rate | X



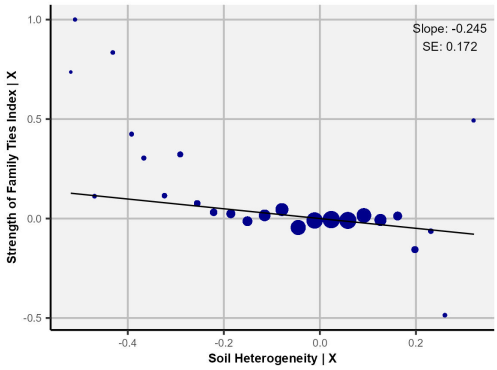
Note: Figure continues to the next page.

FIGURE C.1: Bin Scatter Plots (cont.)

(M) Strength of Family Ties



(N) Strength of Family Ties | X



Note: This figure presents bin scatter plots of the relationship between soil heterogeneity and the different historical measures of close-knit communities: the LNI, the share of ICM, the TNI, the RHI, the MKP, the KPR, and the SFTI. The left column presents the raw relationship, corresponding to column 1 in Table 1, and the right column presents the conditional relationship after partialling out the baseline controls, corresponding to column 4 in Table 1. The size of the points is representative of the number of observations within each bin. See Appendix E for details on the data and variable construction. Standard errors are clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011).

C.2 Kinship Tightness

Appendix Table C.1 reports the estimates of equation 1 when indicators of historical tight kinship are used and for six different specifications. The results are similar to the main finding reported in Table 1, but statistically weaker for two out of the three variables. When the dependent variable is the Median Kin Propinquity (Panel A), the results are highly significant and are highly unlikely to be driven by OVB. In my preferred specification, an increase from SHI = 0 to SHI = 1 is associated with a decrease of 0.086 points ($p-value=0.007$) in the MKP. When the dependent variable is the Kin Propinquity Rate or the Strength of Family Ties Index (Panels B-C), the results are statistically weaker and the data suggest that there is a concern for an OVB. Column 4 suggests that an increase from SHI = 0 to SHI = 1 is associated with a decrease of 0.032 points ($p-value=0.087$) in the KPR and a decrease of 0.245 points ($p-value=0.154$) in the SFTI. As discussed in Online Appendix B.2, there are reasons to expect a weaker impact on kinship ties.

TABLE C.1: Soil Heterogeneity Created Loose-Knit Communities, Kinship Tightness

	Dependent variable:					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Median Kin Propinquity (mean = 0.14, SD = 0.13)					
Soil Heterogeneity	-0.110^{***}	-0.102^{***}	-0.093^{***}	-0.086^{***}	-0.063^{***}	-0.062^{***}
	(0.040)	(0.032)	(0.031)	(0.032)	(0.022)	(0.021)
Oster δ for $\beta = 0$		6.46	4.32	3.47	2.30	2.22
Observations	23,443	23,443	23,443	23,443	23,443	23,381
R^2	0.009	0.359	0.374	0.377	0.405	0.409
	Panel B: Kin Propinquity Rate (mean = 0.32, SD = 0.12)					
Soil Heterogeneity	-0.087^{**}	-0.027	-0.027	-0.032^{*}	-0.020	-0.029^{**}
	(0.044)	(0.023)	(0.021)	(0.019)	(0.015)	(0.015)
Oster δ for $\beta = 0$		0.79	0.87	1.20	0.67	1.05
Observations	23,532	23,532	23,532	23,532	23,532	23,470
R^2	0.007	0.625	0.662	0.677	0.685	0.691
	Panel C: Strength of Family Ties Index (mean = 0, SD = 1)					
Soil Heterogeneity	-1.207^{***}	-0.257	-0.224	-0.245	-0.106	-0.178
	(0.288)	(0.163)	(0.164)	(0.172)	(0.129)	(0.126)
Oster δ for $\beta = 0$		0.47	0.44	0.52	0.22	0.40
Observations	21,671	21,671	21,671	21,671	21,671	21,612
R^2	0.018	0.564	0.594	0.607	0.622	0.631
State × Year Fixed Effects		✓	✓	✓	✓	✓
Geoclimatic Controls			✓	✓	✓	✓
Smooth Location Controls				✓	✓	✓
Agricultural Suitability Controls					✓	✓
Higher Order Controls						✓

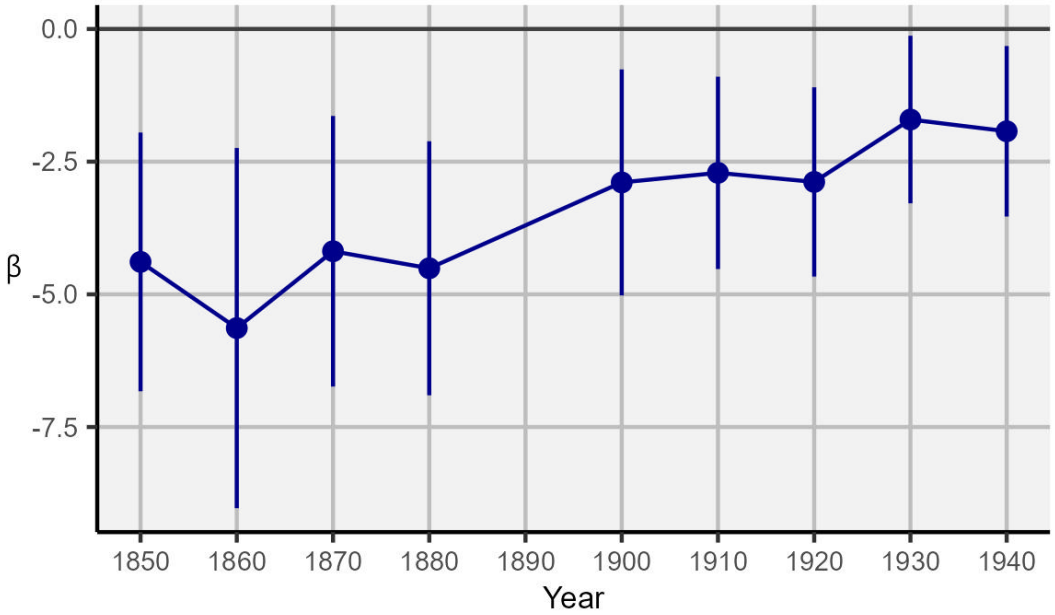
Note: This table reports estimates of equation 1 when the dependent variables are different historical features of close-knit communities, focusing on kinship tightness: the MKP (Panel A), the KPR (Panel B), and the SFT (Panel C). Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. Agricultural suitability controls include average agricultural suitability indices for alfalfa, cotton, maize, oat, rye, sugarcane, sweet potato, tobacco, wheat, and white potato. Higher order controls include the standard deviations of temperature, precipitation, slope, elevation, flow accumulation, and the agricultural suitability indices. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C.3 A Decaying Impact

C.3.1 Soil Heterogeneity and Close-Knit Communities, Decade-by-Decade

In Appendix Figure C.2 and Appendix Table C.2, I study the association between soil heterogeneity and the LNI decade-by-decade rather than pooling together 1850-1940 in a panel specification. I find that the association is always negative and statistically significant, however, it weakens over time. For my baseline specification, the point estimate drops from -4.39 ($p-value < 0.001$) in 1850 to -1.93 ($p-value = 0.02$) in 1940. When I fix the states in the sample to the states that existed in 1850 this pattern is even stronger (not reported for brevity), suggesting that it is not explained by the change in the sample due to the continuing westward expansion. The pattern of a slowly decaying impact is generally shared by the other indicators of close-knit communities (Appendix Figure C.3), although it is weaker and less evident in some of them. The decline in magnitude over 100 years is consistent with soil heterogeneity mattering more in early periods, when farmers were migrating to unfamiliar environments and new communities were forming. This may also explain why the pattern is stronger when the sample excludes states that were settled later.

FIGURE C.2: The Impact Weakens Over Time



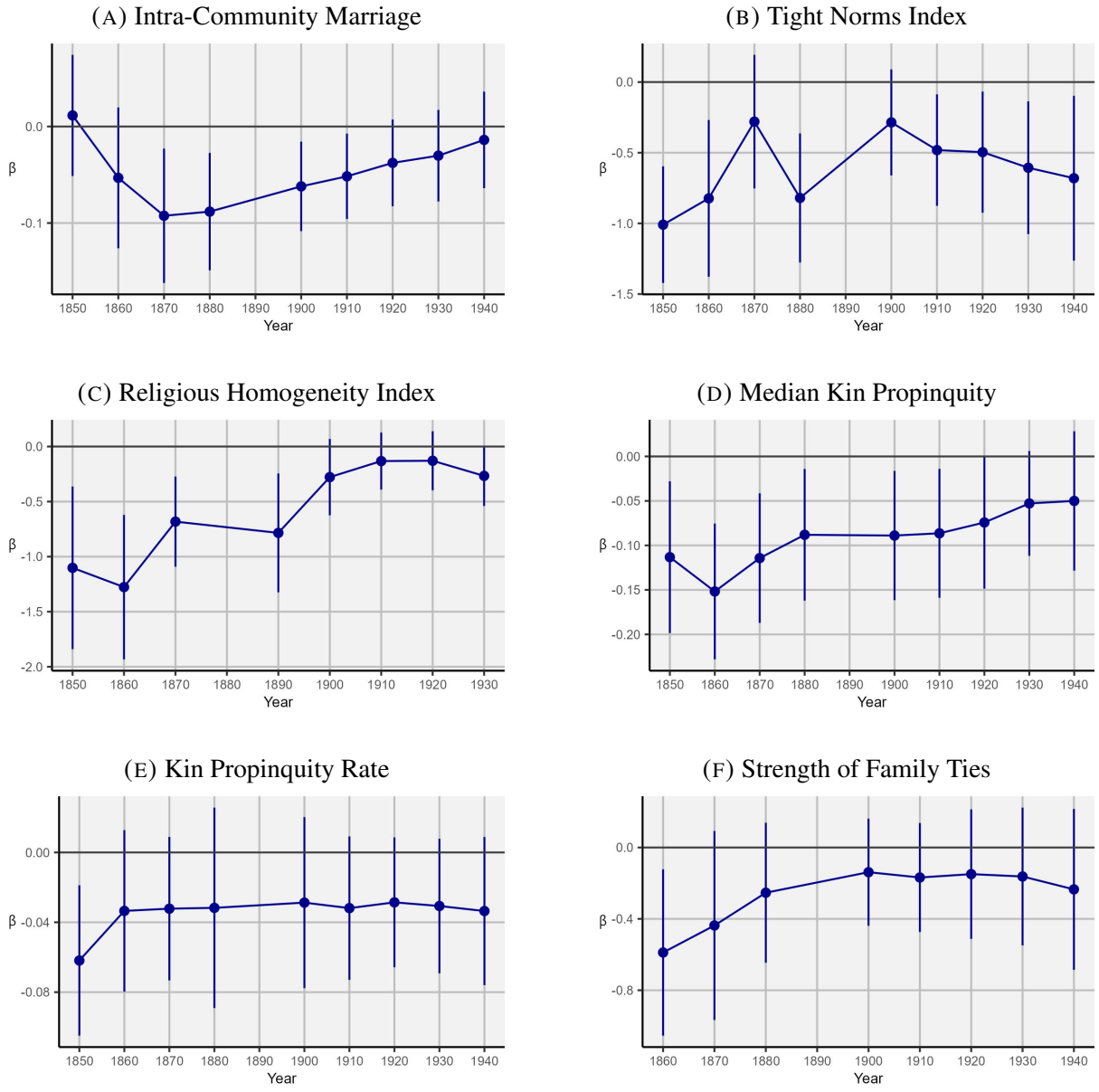
Note: This figure plots the estimates of β and 95% confidence intervals from the baseline specification of equation 1, when the dependent variable is children's LNI in which "local" is defined as the county, estimated separately for each census year. See Appendix E for details on the data and variable construction. Standard errors are clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011)

TABLE C.2: The Association between Soil Heterogeneity and the LNI Weakens Over Time

	Dependent variable: Local Name Index								
	1850	1860	1870	1880	1900	1910	1920	1930	1940
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Soil Heterogeneity	-4.389^{***}	-5.635^{***}	-4.188^{***}	-4.509^{***}	-2.891^{***}	-2.711^{***}	-2.882^{***}	-1.706^{**}	-1.927^{**}
	(1.244)	(1.731)	(1.301)	(1.221)	(1.087)	(0.925)	(0.910)	(0.805)	(0.820)
Dependent Variable Mean	65.95	69.68	69.44	68.79	68.46	67.84	67.32	66.78	66.54
Dependent Variable SD	6.49	8.16	7.84	7.54	5.97	5.14	5.22	4.93	5.02
Observations	1,608	2,031	2,242	2,526	2,819	2,949	3,065	3,098	3,099
R^2	0.665	0.560	0.619	0.606	0.567	0.511	0.535	0.481	0.512
State Fixed Effects	✓	✓	✓	✓	✓	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓

Note: This table reports estimates of equation 1 when the dependent variable is children's LNI in which "local" is defined as the county, estimated separately for each census year. Geoclimatic controls include average temperature, precipitation, slope, elevation, absolute agricultural productivity, flow accumulation, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

FIGURE C.3: The Association between Soil Heterogeneity and Close-Knit Communities Over Time



Note: This figure plots the estimates of β and 95% confidence intervals from the baseline specification of equation 1, when the dependent variables are different historical measures of close-knit communities: the share of ICM (subfigure A), the TNI (subfigure B), the RHI (subfigure C), MKP (subfigure D), the KPR (subfigure E), and the SFTI (subfigure F), estimated separately for each census year. See Appendix E for details on the data and variable construction. Standard errors are clustered at arbitrary grid cells of 100 miles square (Bester et al., 2011)

C.3.2 The Long-Run Impact

I study the following long-run outcome variables:

Close-knit social networks. To study the long-run impact of local soil heterogeneity on the strength of local social ties, I use two county-level measures of social cohesiveness calculated by Chetty et al. (2022a,b) using active U.S. Facebook users aged 25-44. The first is *Social Clustering*, which measures the average fraction of an individual's friend pairs who are also friends with each other. The second is the *Social Support Ratio*, which measures the proportion of within-county friendships where the pair of friends share a third mutual friend within the same county. I standardize both into z-scores.

Communal Morality. I also use contemporary data on individuals' morality collected through an online survey between 2008-2018 using the Moral Foundations Questionnaire (Graham et al., 2011) to construct a measure of *Communal Moral Values* (Haidt, 2008; Haidt and Graham, 2007; Enke, 2020). The Moral Foundation Theory identifies five moral foundations that are divided into two moral approaches: two foundations are universalist—Harm / Care, and Fairness / Reciprocity, and three are communal—In-group / Loyalty, Authority / Respect, and Purity / Sanctity. I define communal moral values as the first eigenvector from a principal component analysis of the five moral foundations. This eigenvector explains about 46% of the variance in the data, and it is the only eigenvector for which the signs of the loadings on the five foundations correspond to the communal versus universalist distinction.

Excess Support for Trump in 2016. Following Enke (2020), which draws a positive link between support for Donald Trump in the 2016 presidential election and voters' "demand" for communal morality, I also use excess support for Trump as another outcome variable, which captures one political consequence of close-knit communities. As in Enke (2020), I define excess support for Trump as the difference between Trump's vote share and that of prior Republican presidential candidates'—Romney and McCain, standardized into z-scores.

Results. I use a framework similar to equation 1 to analyze the long-run data, replacing state-by-year fixed effects with state-fixed effects, and controlling for age, gender, and race fixed-effects in the individual-level communal morality regressions. Table C.3 presents the results. I find evidence that local soil heterogeneity is negatively associated with a close-knit social structure in the long run, however, statistical significance is not high. In my preferred specification (column 4), an increase from SHI = 0 to SHI = 1 is associated with a 0.47 ($p-value = 0.063$) standard deviation drop in long-run social clustering (Panel A), a marginally insignificant -0.303 ($p-value = 0.118$)

standard deviation drop in long-run social support ratio (Panel B), a 0.053 ($p-value = 0.061$) standard deviation drop in communal moral values (Panel C), and a marginally insignificant -0.271 ($p-value = 0.12$) standard deviation drop in excess support for Trump in 2016 (Panel D).

TABLE C.3: Soil Heterogeneity Reduces Long-Run Features of Close-Knit Communities

	Dependent variable:					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Social Clustering (mean = 0, SD = 0.96)					
Soil Heterogeneity	-0.668^{**}	-0.461^{*}	-0.585^{**}	-0.470^{*}	-0.328	-0.249
	(0.319)	(0.245)	(0.240)	(0.252)	(0.200)	(0.199)
Observations	3,055	3,055	3,055	3,055	3,055	3,046
R^2	0.006	0.217	0.247	0.263	0.326	0.338
	Panel B: Social Support Ratio (mean = 0, SD = 0.97)					
Soil Heterogeneity	-0.664^{***}	-0.414^{**}	-0.358^{*}	-0.303	-0.253	-0.128
	(0.228)	(0.206)	(0.194)	(0.194)	(0.187)	(0.153)
Observations	3,055	3,055	3,055	3,055	3,055	3,046
R^2	0.006	0.124	0.157	0.173	0.201	0.206
	Panel C: Communal Moral Values (mean = 0, SD = 1)					
Soil Heterogeneity	0.003	-0.023	-0.037	-0.053^{*}	-0.053^{*}	-0.050^{*}
	(0.033)	(0.033)	(0.030)	(0.029)	(0.030)	(0.030)
Observations	272,695	272,695	272,695	272,695	272,695	266,005
R^2	0.029	0.029	0.029	0.029	0.029	0.030
	Panel D: $\Delta[Trump - Ave.\ GOP]$ (mean = 0, SD = 1)					
Soil Heterogeneity	-0.755^{**}	-0.176	-0.328^{*}	-0.271	-0.142	-0.054
	(0.344)	(0.174)	(0.172)	(0.174)	(0.163)	(0.154)
Observations	3,107	3,107	3,107	3,107	3,107	3,088
R^2	0.007	0.450	0.481	0.489	0.514	0.520
State Fixed Effects		✓	✓	✓	✓	✓
Geoclimatic Controls			✓	✓	✓	✓
Smooth Location Controls				✓	✓	✓
Agricultural Suitability Controls					✓	✓
Higher Order Controls						✓

Note: This table reports estimates of regressions similar to equation 1, when the dependent variables are different long-run measures of close-knit communities: social clustering (Panel A), social support ratio (Panel B), communal moral values (Panel C), and Trump's 2016 gains over Romney and McCain (Panel D). The regressions in Panels A, B, and D are cross-sectional county-level. The regressions in Panel C are individual-level and always control for cohort, race, and gender fixed-effects. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. Agricultural suitability controls include average agricultural suitability indices for alfalfa, cotton, maize, oat, rye, sugarcane, sweet potato, tobacco, wheat, and white potato. Higher order controls include the standard deviations of temperature, precipitation, slope, elevation, flow accumulation, and the agricultural suitability indices. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C.4 DID Results

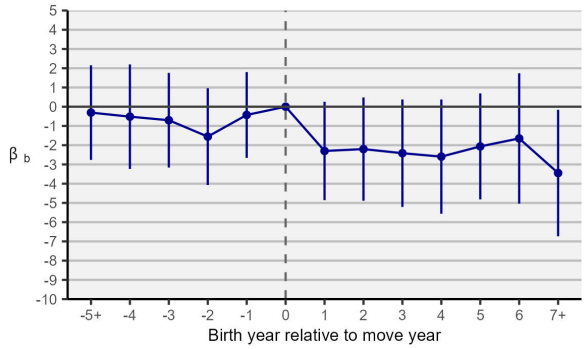
TABLE C.4: The Characteristics of Farmers' and Non-Farmers' Migrant Households

	Household: Farmers' N=216,333	Household: Non-Farmers' N=164,044	Difference p-value
	(1)	(2)	(3)
Father's Age	38.1 (8.12)	36.8 (7.57)	0.000
Number of Children (0-10)	4.70 (2.00)	3.95 (1.74)	0.000
Urban location	0.01 (0.08)	0.27 (0.44)	0.000
Avg. Pre-migration LNI score	52.8 (10.3)	52.6 (11.6)	<0.001
Intra-Community Marriage	0.55 (0.50)	0.53 (0.50)	<0.001

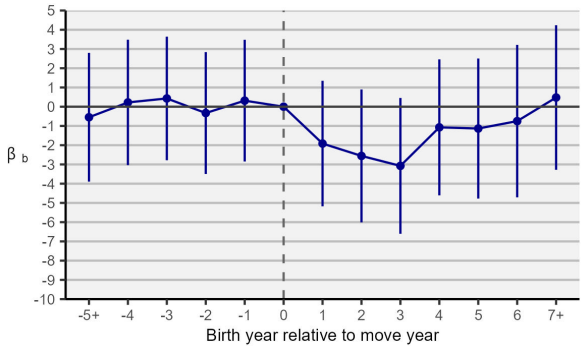
Note: This table reports the mean characteristics or farmers' and non-farmers' domestic out-of-state migrant households from the baseline sample which only includes native-born white households. Standard errors in parenthesis.

FIGURE C.4: The Differential Impact is Not Driven by Age

(A) DID: Father is Young (age $\leq$ median)



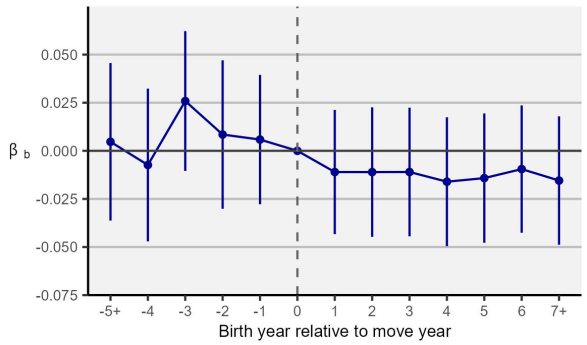
(B) DID: Father is Old (age $>$ median)



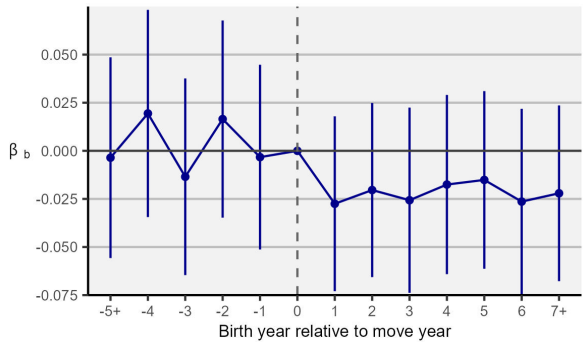
Note: This figure plots the estimates of β_b and 95% confidence intervals from equation 3 when the dependent variable is children's LNI where "local" is defined as the state. The sample includes white native-born children between the ages of 0 to 10 with native-born parents whose families migrated once across states. The sample in subfigure A only includes households in which the father's age is below or equal to the median, and in subfigure B it only includes households in which the father's age is above the median. See Appendix E for details on the data and variable construction. Standard errors are clustered by destination county.

FIGURE C.5: The Differential Impact is Not Driven by a Different Births Profile

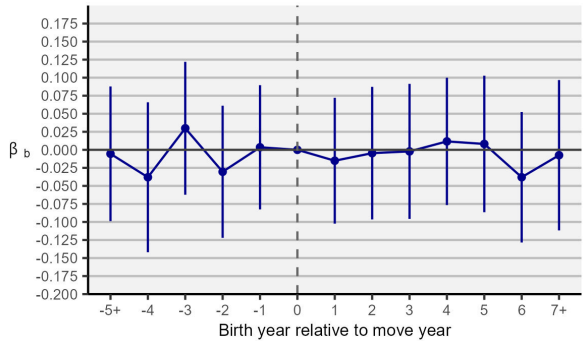
(A) DID: Farmers, birth probability



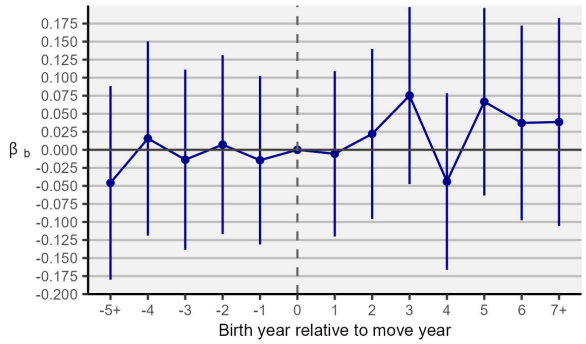
(B) DID: Non-Farmers, birth probability



(C) DID: Farmers, child's gender is male



(D) DID: Non-Farmers, child's gender is male



Note: This figure plots the estimates of β_b and 95% confidence intervals from equation 3. In subfigures A-B the dependent variable is a dummy variable that equals one if a child was born and zero otherwise. In subfigures C-D the dependent variable is a dummy variable that equals one if the gender of a child that was born during this year was male, and zero otherwise. The sample includes white native-born children between the ages of 0 to 10 with native-born parents, whose families migrated once across states. See Appendix E for details on the data and variable construction. Standard errors are clustered at the destination county.

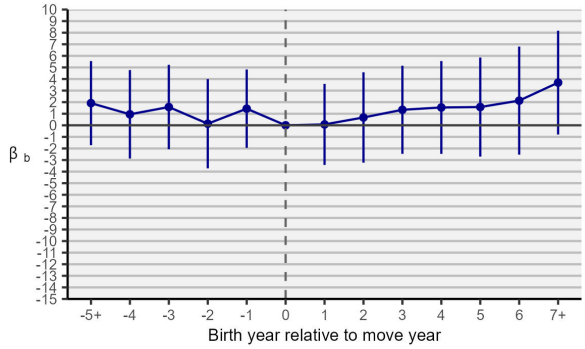
TABLE C.5: The Differential Impact is Not Driven by Rural vs. Urban Locations

	Dependent variable: Local Name Index			
	Difference-in-Differences			Triple-Difference
Sample:	All Households in Rural Locations	Farmers′ Households in Rural Locations	Non-Farmers′ Households in Rural Locations	All Households in Rural Locations
	(1)	(2)	(3)	(4)
Post Migration × Soil Heterogeneity	−2.402***	−3.335***	−0.159	−0.155
	(0.679)	(0.743)	(0.771)	(0.771)
Post Migration × Farmers′ Household × Soil Heterogeneity				−3.180***
				(0.741)
Dependent Variable Mean	54.29	54.40	54.08	54.29
Dependent Variable SD	13.58	13.41	13.91	13.58
Observations	1,075,962	710,260	365,698	1,075,958
R^2	0.359	0.349	0.378	0.359
Family Fixed Effects	✓	✓	✓	✓
Relative YOB Fixed Effects	✓	✓	✓	✓

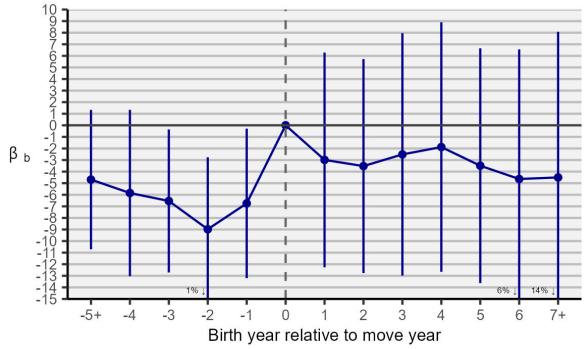
Note: This table presents estimates of the "static" versions of the Difference-in-Differences and Triple-Difference estimation frameworks (equations 3 and 4). The dependent variable is children's LNI where "local" is defined as the state. The sample includes white native-born children between the ages of 0 to 10 with native-born parents, whose families migrated once across states and reside in a rural community. See Appendix E for details on the data and variable construction. Standard errors clustered by destination county in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

FIGURE C.6: The Differential Impact is Not Driven by Rural vs. Urban Locations

(A) DID: Non-Farmers, Rural



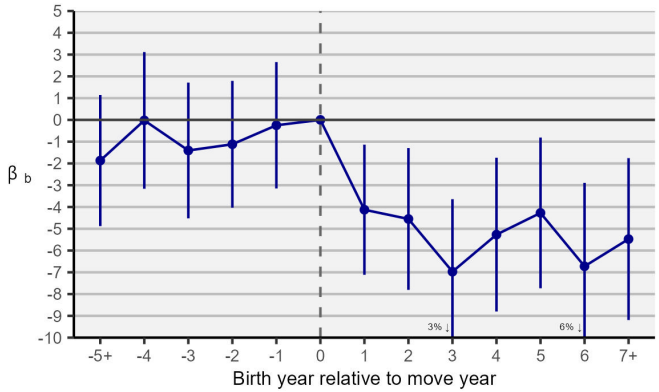
(B) DID: Non-Farmers, Urban



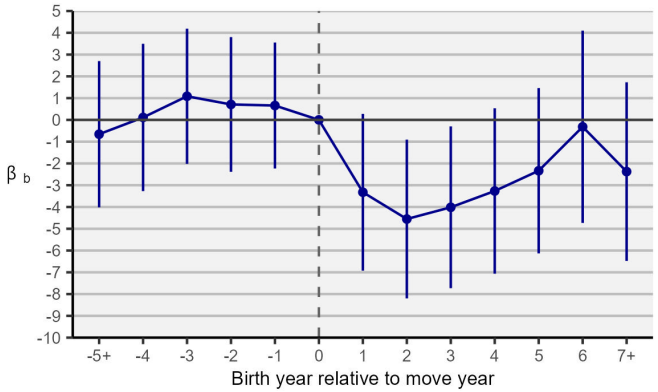
Note: This figure plots the estimates of β_b and 95% confidence intervals from equations 3 when the dependent variable is children's LNI where "local" is defined as the state. The sample includes white native-born children between the ages of 0 to 10 with native-born parents, whose father is not a farmer, and whose families migrated once across states. The sample in subfigure A only includes rural locations and in subfigure B it only includes urban locations. See Appendix E for details on the data and variable construction. Standard errors are clustered by destination county.

FIGURE C.7: Heterogeneous Impact on Farmers by Prior Communalism

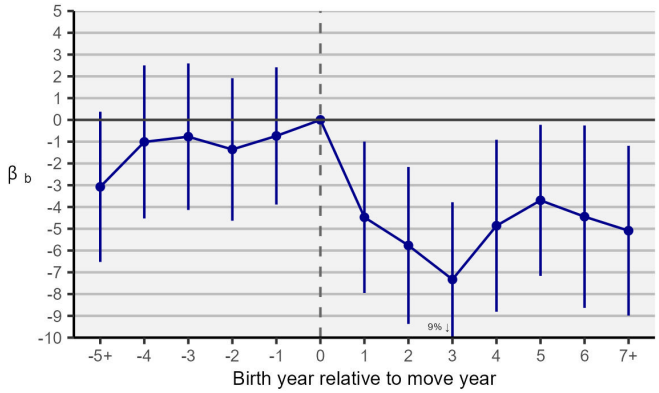
(A) DID: Farmers, High Prior Communal Identification



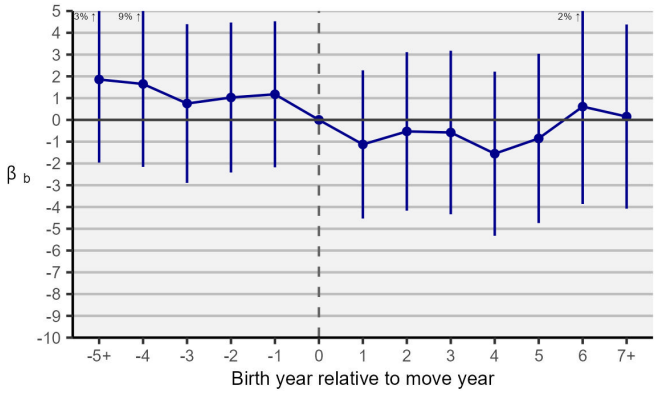
(B) DID: Farmers, Low Prior Communal Identification



(C) DID: Farmers, Intra-Community Marriage



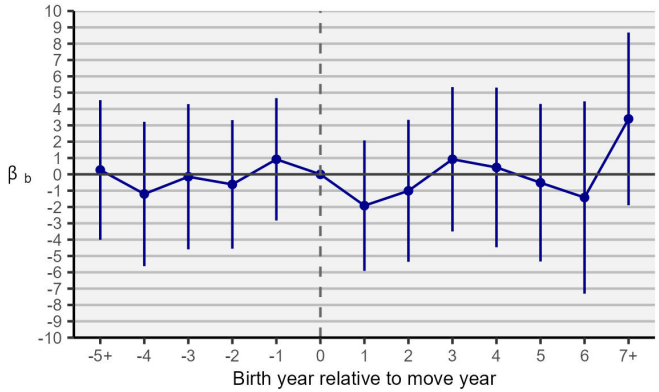
(D) DID: Farmers, Extra-Community Marriage



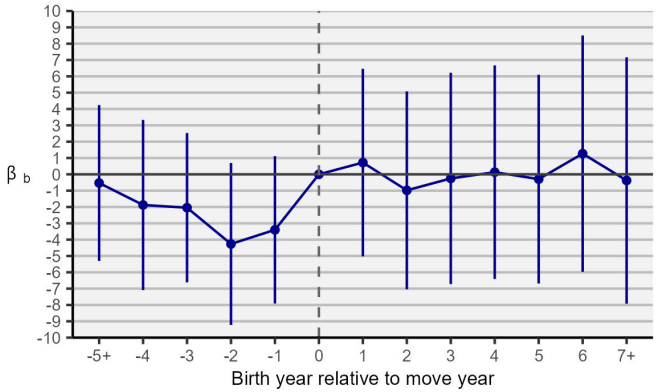
Note: This figure plots the estimates of β_b and 95% confidence intervals from equation 3 when the dependent variable is children's LNI where "local" is defined as the state. The sample includes white native-born children between the ages of 0 to 10 with native-born parents, whose families migrated once across states and whose fathers are farmers. In subfigures A-B, the sample is split by prior levels of communal identification, measured by the average LNI of children born before the migration, and in subfigures C-D, the sample is split by same vs. different birthplace of parents. See Appendix E for details on the data and variable construction. Standard errors are clustered at the destination county.

FIGURE C.8: Heterogeneous Impact on Non-Farmers by Prior Communalism

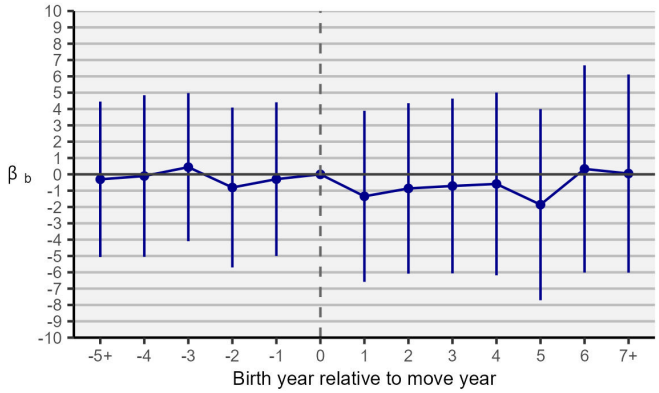
(A) DID: Non-Farmers, High Prior Communal Identification



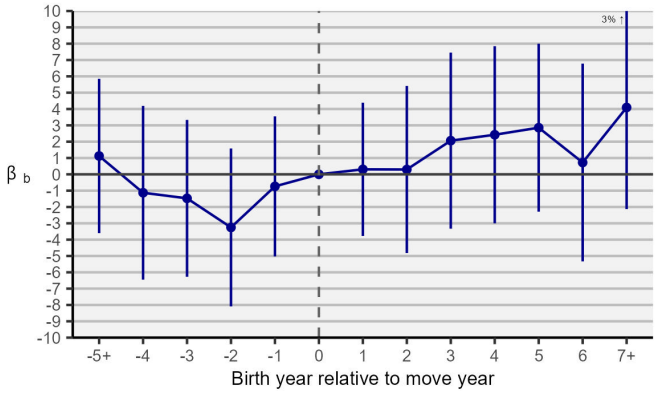
(B) DID: Non-Farmers, Low Prior Communal Identification



(C) DID: Non-Farmers, Intra-Community Marriage



(D) DID: Non-Farmers, Extra-Community Marriage



Note: This figure plots the estimates of β_b and 95% confidence intervals from equation 3 when the dependent variable is children's LNI where "local" is defined as the state. The sample includes white native-born children between the ages of 0 to 10 with native-born parents, whose families migrated once across states and whose fathers are not farmers. In subfigures A-B, the sample is split by prior levels of communal identification, measured by the average LNI of children born before the migration, and in subfigures C-D, the sample is split by same vs. different birthplace of parents. See Appendix E for details on the data and variable construction. Standard errors are clustered at the destination county.

C.5 Selective Out-Migration

TABLE C.6: Selective-Out Migration is Not Concentrated in Older Farmers with Many Children

	Dependent variable: Remaining in the Same County (mean = 0.45, SD = 0.49)				
Sample	All	Age		Number of Children	
		Young	Old	Low	High
	(1)	(2)	(3)	(4)	(5)
Soil Heterogeneity	0.021	0.010	0.029	0.041	−0.013
	(0.037)	(0.048)	(0.050)	(0.045)	(0.056)
Farmer	0.155***	0.149***	0.154***	0.136***	0.150***
	(0.027)	(0.035)	(0.034)	(0.035)	(0.036)
Soil Heterogeneity × Farmer	−0.088**	−0.094*	−0.086	−0.068	−0.089
	(0.041)	(0.054)	(0.053)	(0.053)	(0.055)
Observations	74,922	39,318	35,604	42,053	32,869
R^2	0.033	0.034	0.035	0.036	0.034
State × Year Fixed Effects	✓	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓	✓

Note: This table reports estimates of equation 6 when the dependent variable is the probability of remaining in the same county in the following decade. The sample in column 1 includes all the linked migrants. Columns 2 and 3 restrict the sample to migrants that were younger and older than the median, respectively. Columns 4 and 5 restrict the sample to migrants whose number of children is below and above the median, respectively. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. See Appendix E for details on the data and variable construction. Standard errors clustered by original destination county in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE C.7: Selective-Out Migration is Not Driven by Older Farmers with Many Children

	Dependent variable: Remaining in the Same County (mean = 0.45, SD = 0.49)			
	(1)	(2)	(3)	(4)
Soil Heterogeneity	−0.029	−0.032	−0.007	−0.020
	(0.048)	(0.057)	(0.053)	(0.058)
Farmer	0.115***	0.110***	0.092**	0.095**
	(0.035)	(0.042)	(0.041)	(0.044)
Soil Heterogeneity × Farmer	−0.027	−0.035	−0.002	−0.012
	(0.053)	(0.064)	(0.064)	(0.068)
Soil Heterogeneity × Farmer × High Communal Identification	−0.125*	−0.128*	−0.125*	−0.127*
	(0.068)	(0.068)	(0.067)	(0.067)
Soil Heterogeneity × Farmer × Age > median		0.011		0.035
		(0.068)		(0.075)
Soil Heterogeneity × Farmer × Num. Children > median			−0.030	−0.050
			(0.072)	(0.080)
Observations	74,922	74,922	74,922	74,922
R^2	0.033	0.038	0.036	0.039
State × Year Fixed Effects	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓

Note: This table reports estimates of equation 6 when the dependent variable is the probability of remaining in the same county in the following decade. The sample includes all the linked migrants. Columns 2-4 also include the main effect of being older, having many children, and all the relevant two-term interactions. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. See Appendix E for details on the data and variable construction. Standard errors clustered by original destination county in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C.6 Confounding Factors, Competing Mechanisms, and Mediators

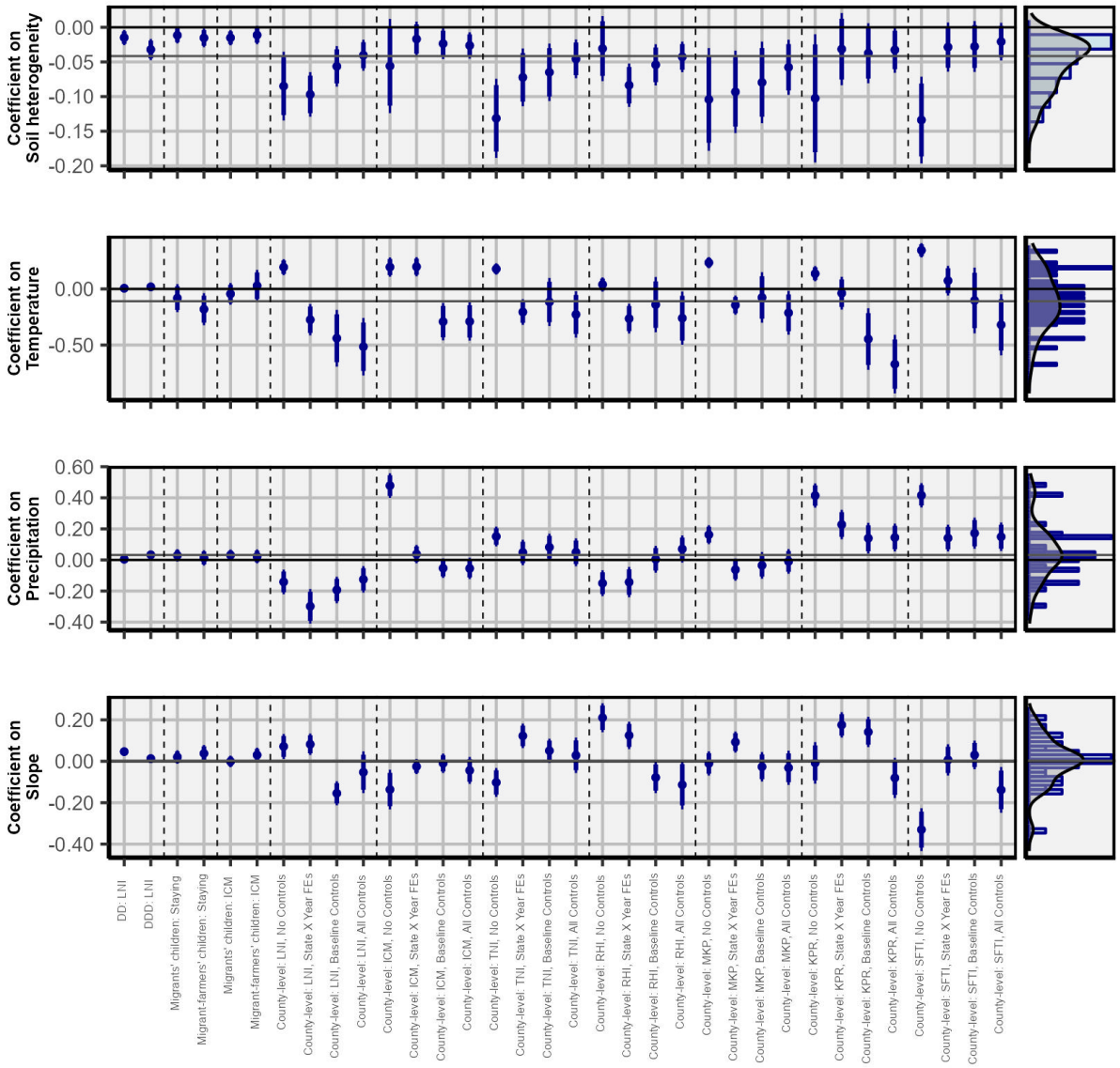
C.6.1 Other geoclimatic features of the environment

One concern is that because soil heterogeneity is correlated with multiple environmental characteristics, the coefficient might be picking up the effect of unobserved geoclimatic features, even conditional on the extensive set of geoclimatic controls. While the data also suggest that such omitted variable bias is unlikely to be driving the results (Oster, 2019), this may still be of some concern.

I provide further evidence that this is not the case by documenting that other geoclimatic features of the environment do not have a stable and robust relationship with the existence of historically close-knit communities that have to do specifically with farmers. Specifically, I estimate the standardized beta coefficients of all the different geoclimatic features of the environment on children's LNI in the DD and DDD frameworks, corresponding to columns 1 and 4 in Table 3, on migrants' children's communal attachment, corresponding to columns 1 and 2 in Table 4, and on the seven different measure of historically close-knit communities in the county-level framework, in four different specifications, corresponding to columns 1, 2, 4, and 6 and all panels in Table 1.

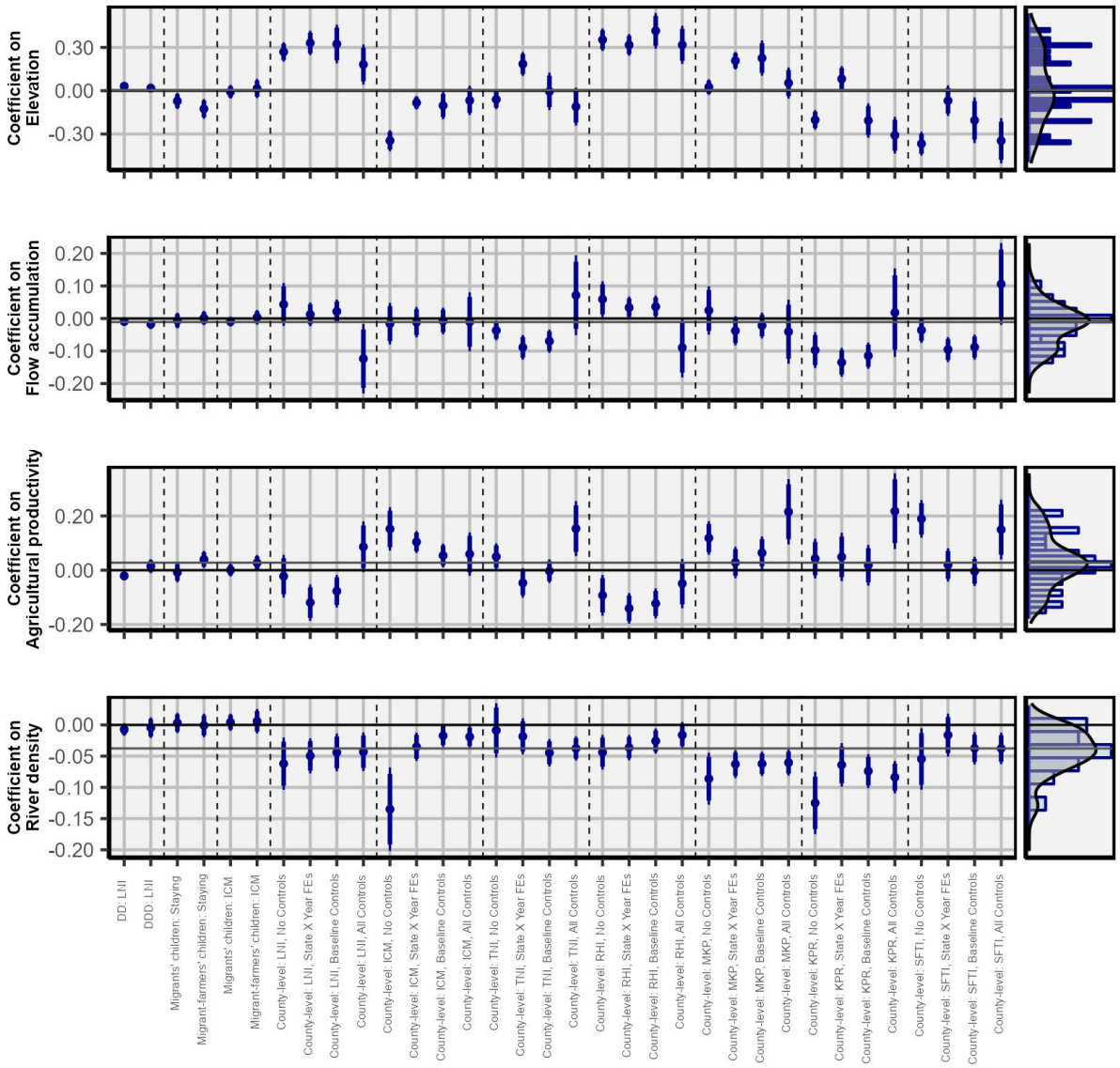
Appendix Figure C.9 presents the results. Only one geoclimatic feature—soil heterogeneity—has a stable and robust negative relationship with close-knit communities and a corresponding treatment effect that is centered on farmers. The estimated coefficients on the other features switch signs across different historical measures of close-knit communities or specifications, and in many cases, they even have significant coefficients operating in different directions. Moreover, the distribution of the coefficients of most geoclimatic features is roughly centered around zero. Since the geoclimatic controls are at least as likely to be correlated with unobserved confounding features of the environment, it seems unlikely that the robust and stable association between soil heterogeneity and close-knit communities is simply picking up some correlated residual geoclimatic variation.

FIGURE C.9: The Impact of Other Geo-climatic Characteristics



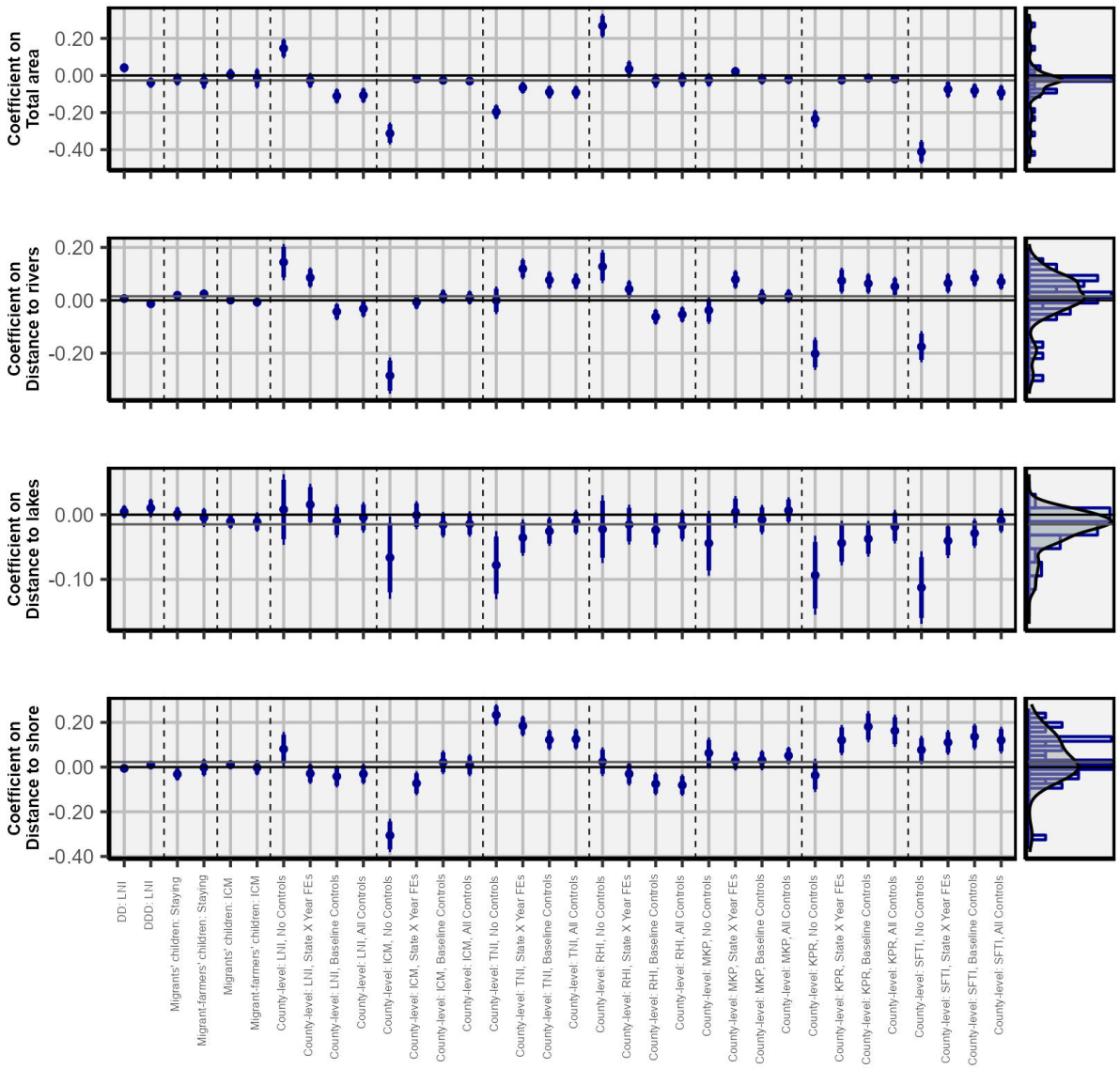
Note: Figure continues to the next page.

FIGURE C.9: The Impact of Other Geo-climatic Characteristics (cont.)



Note: Figure continues to the next page.

FIGURE C.9: The Impact of Other Geo-climatic Characteristics (cont.)



Note: This figure plots standardized beta coefficients of different geoclimatic features of the environment from multiple regressions in which the dependent variables are different measures of historical close-knit communities. The different specifications correspond to columns 1 and 4 in Table 3, columns 1 and 2 in Table 4, and columns 1, 2, 4, and 6 in Table 1. Each row plots the coefficients of a different geoclimatic feature of the environment. The left panel plots the estimated coefficients and their 90% and 95% confidence intervals, and the right panel plots the coefficients' histogram and estimated density. The grey horizontal line in each row plots the median value of the estimated coefficients.

C.6.2 Climatic risk

Another potentially confounding feature of the environment is variation over time, rather than across space. Pre-industrial environmental variation over a long period has been found to decrease traditionalism due to the lower usefulness of knowledge obtained by previous generations (Giuliano and Nunn, 2021), and to increase interpersonal trust and weaken family ties due to extra-familial cooperation that was useful in mitigating climatic risk (Buggle and Durante, 2021). The concern is that the same mechanisms may also be relevant for 1850-1940 American farmers and that environmental spatial heterogeneity and climatic risk are correlated, such that the estimated effect of soil heterogeneity is picking up, at least partly, the impact of the latter.

To explore this possibility, I examine the impact of climatic risk on close-knit communities, and also horse race it with soil heterogeneity. Similar to Buggle and Durante (2021), I define precipitation and temperature risk as the standard deviation of annual precipitation and annual temperature, respectively, using data from GAEZ-FAO (FAO and IIASA, 2020). I aggregate precipitation and temperature risk into a single *Climatic Risk* variable by first standardizing each risk factor into z-scores, and then taking their mean.

Appendix Table C.8 presents the results. Columns 1 and 4 document that there is no significant relationship between climatic risk and all of the outcome variables except the KPR (Panel F) and the SFTI (Panel G). Columns 2 and 5 horse race soil heterogeneity and climatic risk. For all outcome variables, there is little change in the coefficient of soil heterogeneity relative to the baseline estimate (Table 1, column 4). Also, the inclusion of the SHI has little impact on the estimated coefficient of climatic risk, which remains insignificant in all outcomes except the KPR and the SFTI.

It is interesting to note that the finding of a significant negative association between climatic risk and the KPR and STFI is consistent with the findings in Buggle and Durante (2021). At the same time, given the mechanism of substitution suggested in their paper, one could have expected to also find that climatic risk will be associated with *stronger* communal ties. The results in Panels A-D do not support this.

TABLE C.8: Climatic Risk

	Dependent variable:					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Local Name Index			*Panel B*: Intra-Community Marriage		
Soil Heterogeneity		−2.989***	−3.070***		−0.043**	−0.044
		(0.798)	(1.109)		(0.021)	(0.027)
Climatic Risk	0.104	0.068		0.012	0.012	
	(0.397)	(0.390)		(0.009)	(0.009)	
HighClimatic Risk			0.403			0.005
			(0.773)			(0.025)
Soil Heterogeneity ×			0.191			−0.0003
High Climatic Risk			(1.205)			(0.036)
Observations	23,437	23,437	23,437	23,431	23,431	23,431
R^2	0.546	0.548	0.548	0.813	0.813	0.813
	Panel C: Tight Norms Index			*Panel D*: Religious Homogeneity Index		
Soil Heterogeneity		−0.575***	−0.612**		−0.438***	−0.453**
		(0.194)	(0.263)		(0.129)	(0.184)
Climatic Risk	−0.063	−0.069		−0.080	−0.084	
	(0.052)	(0.050)		(0.057)	(0.056)	
HighClimatic Risk			−0.074			−0.057
			(0.146)			(0.136)
Soil Heterogeneity ×			0.095			0.046
High Climatic Risk			(0.220)			(0.211)
Observations	23,322	23,322	23,322	17,216	17,216	17,216
R^2	0.329	0.332	0.332	0.415	0.416	0.416
State × Year Fixed Effects	✓	✓	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓	✓	✓

Note: Table continues to the next page.

TABLE C.8: Climatic Risk (cont.)

	Dependent variable:					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel E: Median Kin Propinquity			*Panel F*: Kin Propinquity Rate		
Soil Heterogeneity		-0.086^{***}	-0.095^{**}		-0.033^{*}	-0.037
		(0.032)	(0.048)		(0.019)	(0.025)
Climatic Risk	-0.0005	-0.002		-0.010^{*}	-0.010^{**}	
	(0.007)	(0.006)		(0.005)	(0.005)	
HighClimatic Risk			-0.013			-0.021
			(0.029)			(0.017)
Soil Heterogeneity ×			0.020			0.011
High Climatic Risk			(0.044)			(0.025)
Observations	23,443	23,443	23,443	23,532	23,532	23,532
R^2	0.373	0.377	0.377	0.676	0.677	0.678
	Panel G: Strength of Family Ties					
Soil Heterogeneity		-0.255	-0.278			
		(0.172)	(0.223)			
Climatic Risk	-0.114^{**}	-0.118^{**}				
	(0.056)	(0.056)				
HighClimatic Risk			-0.112			
			(0.140)			
Soil Heterogeneity ×			0.071			
High Climatic Risk			(0.209)			
Observations	21,671	21,671	21,671			
R^2	0.607	0.608	0.608			
State × Year Fixed Effects	✓	✓	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓	✓	✓

Note: This table reports estimates of multiple regressions in which the dependent variables are different historical features of close-knit communities: the LNI (Panel A), the share of ICM (Panel B), the TNI (Panel C), the RHI (Panel D), MKP (Panel E), the KPR (Panel F), and the SFTI (Panel G). Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. Climatic risk is the average of the standard deviation of annual precipitation, standardized into z-scores, and the standard deviation of mean annual temperature, standardized into z-scores. High climatic risk is an indicator for an above-median climatic risk. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C.6.3 Trade

Two competing mechanisms involve trade, one focuses on trade within the community and the other on trade with outsiders. First, in an environment of high soil heterogeneity, there might be greater scope for farmers to engage in intra-communal trade, either as a way to co-insure against adverse agricultural shocks that affect different soil types in different ways (Buggle and Durante, 2021) or simply due to differences in competitive advantages and specialization (Fenske, 2014).

Two arguments working against this concern. The first is theoretical. If this was the case, then farmers in a soil-heterogeneous location stood to benefit *more* from local cooperation and stronger social ties, relative to farmers in a soil-homogeneous location. This implies that if this mechanism operates in this context, it attenuates the results. The second is historical. Unlike in pre-industrial periods, in nineteenth-century U.S., market integration was high: a significant share of trade was extra-communal, and farmers often produced for far away markets (e.g. cotton production). Also, financial and insurance markets were more developed.

Nevertheless, I perform another empirical exercise in the spirit of Buggle and Durante (2021) to test the possibility that soil heterogeneity facilitated intra-communal trade as a way to insure against climatic risk. In columns 3 and 6 in Appendix Table C.8, I define *High Climatic Risk* as above-median risk and interact it with soil heterogeneity. In all outcomes the interaction term is insignificant, and the coefficient on the main soil heterogeneity effect remains fairly stable, suggesting that this mechanism is unlikely to be a central piece of the story.

The second concern is that soil heterogeneity is correlated with geoclimatic features that made trading more difficult or expensive (e.g., a higher slope and a larger distance to navigated rivers, Appendix Table A.1). While those features are directly controlled for, and the factors that are directly related to long-distance trade cost themselves do not seem to have a robust and stable relationship with close-knit communities (Appendix Figure C.9), there might still be some concern.

Therefore, I test the trade costs-related concern that is most important in this historical context—that soil heterogeneity impeded the development of railroad infrastructure. To test the potential role of railroad development as a competing mechanism, I use 1840-1910 data on railroads (Atack, 2015b) and add to the analysis time-varying indicators for the existence of railroads in the county, as well as within 10, 20, 30, 40, and 50 miles. Appendix Table C.9 presents the results. When interpreting the results of this exercise and similar empirical exercises below, it is important to remember that such variables that are potential outcomes of soil heterogeneity are "bad controls". Therefore, estimates from such exercises should not be given causal interpretations. Instead, they are only meant to provide suggestive evidence regarding mechanisms. I find that controlling for railroad connectivity has little impact on the relationship between soil heterogeneity and close-knit communities, which is evidence

against the competing mechanism of impeding the development of railroads. Overall, both intra-community and extra-community trade do not seem to be an important part of the story.

TABLE C.9: Railroad Connectivity Does Not Account for the Association

	Dependent variable:						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: Local Name Index (mean = 68.4, SD = 6.9)						
Soil Heterogeneity	-3.766^{***}	-2.948^{***}	-2.861^{***}	-2.792^{***}	-2.774^{***}	-2.816^{***}	-2.819^{***}
	(0.927)	(0.812)	(0.793)	(0.771)	(0.767)	(0.769)	(0.779)
Observations	14,175	14,175	14,175	14,175	14,175	14,175	14,175
R^2	0.577	0.595	0.596	0.599	0.600	0.601	0.602
	Panel B: Intra-Community Marriage (mean = 0.61, SD = 0.22)						
Soil Heterogeneity	-0.057^{**}	-0.056^{**}	-0.057^{**}	-0.057^{**}	-0.057^{**}	-0.057^{**}	-0.057^{**}
	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)
Observations	14,169	14,169	14,169	14,169	14,169	14,169	14,169
R^2	0.817	0.817	0.817	0.817	0.817	0.817	0.817
	Panel C: Tight Norms Index (mean = 0, SD = 1)						
Soil Heterogeneity	-0.572^{***}	-0.469^{***}	-0.463^{***}	-0.456^{***}	-0.455^{***}	-0.456^{***}	-0.456^{***}
	(0.183)	(0.164)	(0.162)	(0.159)	(0.159)	(0.159)	(0.160)
Observations	14,063	14,063	14,063	14,063	14,063	14,063	14,063
R^2	0.334	0.347	0.348	0.349	0.349	0.349	0.349
	Panel D: Religious Homogeneity Index (mean = 0, SD = 1)						
Soil Heterogeneity	-0.621^{***}	-0.489^{***}	-0.478^{***}	-0.475^{***}	-0.471^{***}	-0.471^{***}	-0.471^{***}
	(0.163)	(0.138)	(0.135)	(0.133)	(0.132)	(0.131)	(0.132)
Observations	13,726	13,726	13,726	13,726	13,726	13,726	13,726
R^2	0.387	0.408	0.409	0.411	0.412	0.412	0.412
State × Year Fixed Effects	✓	✓	✓	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓	✓	✓	✓
Railroad:							
In county		✓	✓	✓	✓	✓	✓
Within 10 miles			✓	✓	✓	✓	✓
Within 20 miles				✓	✓	✓	✓
Within 30 miles					✓	✓	✓
Within 40 miles						✓	✓
Within 50 miles							✓

Note: Table continues to the next page.

TABLE C.9: Railroad Connectivity Does Not Account for the Association (cont.)

	Dependent variable:						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel E: Median Kin Propinquity (mean = 0.14, SD = 0.12)						
Soil Heterogeneity	−0.104***	−0.085***	−0.083***	−0.082***	−0.082***	−0.082***	−0.082***
	(0.032)	(0.028)	(0.027)	(0.027)	(0.026)	(0.026)	(0.027)
Observations	14,183	14,183	14,183	14,183	14,183	14,183	14,183
R^2	0.347	0.376	0.379	0.382	0.383	0.383	0.383
	Panel F: Kin Propinquity Rate (mean = 0.36, SD = 0.12)						
Soil Heterogeneity	−0.033*	−0.026	−0.027	−0.027	−0.028	−0.028	−0.028
	(0.020)	(0.018)	(0.018)	(0.019)	(0.019)	(0.019)	(0.019)
Observations	14,269	14,269	14,269	14,269	14,269	14,269	14,269
R^2	0.620	0.625	0.626	0.626	0.627	0.627	0.627
	Panel G: Strength of Family Ties Index (mean = 0, SD = 1)						
Soil Heterogeneity	−0.291*	−0.230	−0.230	−0.231	−0.232	−0.231	−0.230
	(0.169)	(0.158)	(0.157)	(0.157)	(0.157)	(0.157)	(0.156)
Observations	12,410	12,410	12,410	12,410	12,410	12,410	12,410
R^2	0.581	0.585	0.585	0.585	0.585	0.585	0.585
State × Year Fixed Effects	✓	✓	✓	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓	✓	✓	✓
Railroad:							
In county		✓	✓	✓	✓	✓	✓
Within 10 miles			✓	✓	✓	✓	✓
Within 20 miles				✓	✓	✓	✓
Within 30 miles					✓	✓	✓
Within 40 miles						✓	✓
Within 50 miles							✓

Note: This table reports estimates of equation 1 when the dependent variables are different historical features of close-knit communities: the LNI (Panel A), the share of ICM (Panel B), the TNI (Panel C), the RHI (Panel D), MKP (Panel E), the KPR (Panel F), and the SFTI (Panel G). Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. In addition to the baseline controls, dummy variables for railroad connectivity are controlled for. The sample only includes 1850-1910 to match railroad infrastructure data available from Atack (2015b). See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C.6.4 Immigration and birthplace diversity

Another set of competing mechanisms concerns the selective sorting of immigrants. First, there are two related concerns: that soil-heterogeneous counties had a higher share of immigrants, and that human capital directed immigration (Michalopoulos, 2012), such that immigrants were selected to soil-heterogeneous counties from more diverse countries of origin because they were trying to find locations that resembled the agricultural conditions they were familiar with (Obolensky et al., 2024). In both cases, soil heterogeneity would have increased population diversity, and diversity, in turn, could have induced loose-knit communities. Moreover, on top of being a competing mechanism, population diversity can be a mediator, as loose-knit communities tend to be more hospitable to foreigners and open to diversity, which can in turn further induce looseness. Second, one might be concerned that immigrants to soil-heterogeneous counties were selected from specific countries of origin that were different in terms of social structure or industrial composition.

The fact that the results are robust to excluding or including immigrants in the sample (Appendix Tables D.4 and D.8) and the evidence on selective in-migration (Section 6.2) suggest that selective sorting of immigrants is unlikely to be a central part of the story. However, due to the centrality of immigration in this historical context, I explore this possibility further.

I address all three concerns in Appendix Table C.10. In column 1, I control for the share of immigrants. In column 2, I control for birthplace diversity, defined as one minus the Herfindahl–Hirschman Index over the share of household heads that were born in different countries. In column 3, I add fixed effects for the most common country of origin among immigrants residing in the county. Finally, in column 4, I control for all three factors together. In all columns the estimated impact of soil heterogeneity remains stable relative to the baseline (Table 1, column 4), suggesting that selective sorting of immigrants is not a central piece of the story.

TABLE C.10: Historical Immigration Does Not Account for the Association

	Dependent variable:			
	(1)	(2)	(3)	(4)
	Panel A: Local Name Index (mean = 67.8, SD = 6.3)			
Soil Heterogeneity	-2.955^{***}	-2.773^{***}	-2.788^{***}	-2.326^{***}
	(0.799)	(0.792)	(0.723)	(0.686)
Observations	23,437	23,437	23,153	23,153
R^2	0.549	0.557	0.563	0.592
	Panel B: Intra-Community Marriage (mean = 0.63, SD = 0.21)			
Soil Heterogeneity	-0.044^{**}	-0.042^{*}	-0.047^{**}	-0.042^{*}
	(0.022)	(0.022)	(0.021)	(0.021)
Observations	23,431	23,431	23,158	23,158
R^2	0.813	0.814	0.818	0.821
	Panel C: Tight Norms Index (mean = 0, SD = 1)			
Soil Heterogeneity	-0.552^{***}	-0.534^{***}	-0.514^{***}	-0.470^{***}
	(0.194)	(0.193)	(0.167)	(0.165)
Observations	23,322	23,322	23,064	23,064
R^2	0.339	0.341	0.344	0.353
	Panel D: Religious Homogeneity Index (mean = 0, SD = 1)			
Soil Heterogeneity	-0.430^{***}	-0.413^{***}	-0.371^{***}	-0.322^{***}
	(0.128)	(0.128)	(0.115)	(0.113)
Observations	17,203	17,203	17,021	17,021
R^2	0.417	0.419	0.431	0.441
State × Year Fixed Effects	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓
Share Immigrants	✓			✓
Birthplace Diversity		✓		✓
Dominate Origin Fixed Effects			✓	✓

Note: Table continues to the next page.

TABLE C.10: Historical Immigration Does Not Account for the Association (cont.)

	Dependent variable:			
	(1)	(2)	(3)	(4)
	Panel E: Median Kin Propinquity (mean = 0.14, SD = 0.13)			
Soil Heterogeneity	-0.084^{***}	-0.081^{**}	-0.077^{***}	-0.070^{***}
	(0.032)	(0.032)	(0.027)	(0.026)
Observations	23,443	23,443	23,172	23,172
R^2	0.385	0.391	0.392	0.407
	Panel F: Kin Propinquity Rate (mean = 0.32, SD = 0.12)			
Soil Heterogeneity	-0.030	-0.027	-0.031^{*}	-0.024
	(0.019)	(0.019)	(0.017)	(0.017)
Observations	23,532	23,532	23,225	23,225
R^2	0.685	0.691	0.688	0.706
	Panel G: Strength of Family Ties Index (mean = 0, SD = 1)			
Soil Heterogeneity	-0.246	-0.235	-0.227	-0.194
	(0.172)	(0.172)	(0.152)	(0.150)
Observations	21,671	21,671	21,440	21,440
R^2	0.607	0.608	0.620	0.627
State × Year Fixed Effects	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓
Share Immigrants	✓			✓
Birthplace Diversity		✓		✓
Dominate Origin Fixed Effects			✓	✓

Note: This table reports estimates of equation 1 when the dependent variables are different historical features of close-knit communities: the LNI (Panel A), the share of ICM (Panel B), the TNI (Panel C), the RHI (Panel D), MKP (Panel E), the KPR (Panel F), and the SFTI (Panel G). Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. In addition to the baseline controls, the patterns of historical immigration are controlled for: column 1 controls for the share of immigrants, column 2 for birthplace diversity, column 3 for dominant country of origin fixed effects, and column 4 for all three. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C.6.5 Race and the legacy of slavery

A related selective sorting concern focuses on racial diversity and historical slavery. For example, soil heterogeneous locations may have been more suitable for slave crops, even conditional on the geoclimatic controls, and the distribution of land ownership and other institutional constraints relating to race may have also created a correlation between racial diversity and soil heterogeneity. But while selective sorting of race is ex-ante possible, its impact on communal ties is unclear. On the one hand, a higher racial diversity can be expected to make social structure looser. On the other, before the great migration, a high black population share was tied to a history of slavery. Slavery and subsequent Jim Crow laws enforced strong race distinctions, and created race-related tension and animosity, which can be expected to strengthen "groupiness". Indeed, the U.S. South tends to be more close-knit (Figure 2 and Appendix Figure B.22).

Empirically, here too it seems unlikely that racial selective sorting was an important part of the story, because the results are robust to excluding or including non-whites in the sample (Appendix Table D.4 and D.8). However, similar to the case of immigration above, due to the centrality of slavery, Jim Crow laws, and the great migration in this historical context, I explore this further.

In columns 1 and 4 of Appendix Table C.11, I directly control for the time-varying black population share. Here too the estimated impact of soil heterogeneity remains remarkably stable relative to the baseline. In columns 2-3 and 5-6, I explore the role of historical slavery. To do so, I first harmonize all the results to 1940 county borders (Hornbeck, 2010). Then, I match the 1850 outcomes to the (harmonized) 1850 slave share, and the 1860-1940 outcomes to the (harmonized) 1860 slave share. Columns 2 and 5 report the baseline estimations when using the harmonized data. Columns 3 and 6, additionally control for historical slavery, which makes almost no difference. I conclude that race and the legacy of slavery are also not important competing mechanisms.

TABLE C.11: Race Does Not Account for the Association

	Dependent variable:					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Local Name Index			*Panel B*: Intra-Community Marriage		
Soil Heterogeneity	−3.009***	−2.940***	−2.974***	−0.044**	−0.052**	−0.051**
	(0.820)	(0.794)	(0.833)	(0.022)	(0.024)	(0.024)
Observations	23,423	23,146	22,949	23,416	23,146	22,954
R^2	0.550	0.646	0.650	0.813	0.789	0.790
	Panel C: Tight Norms Index			*Panel D*: Religious Homogeneity Index		
Soil Heterogeneity	−0.561***	−0.401**	−0.394***	−0.436***	−0.446***	−0.445***
	(0.167)	(0.170)	(0.147)	(0.129)	(0.129)	(0.129)
Observations	23,307	23,027	22,839	17,210	19,715	19,561
R^2	0.365	0.350	0.369	0.417	0.485	0.482
	Panel E: Median Kin Propinquity			*Panel F*: Kin Propinquity Rate		
Soil Heterogeneity	−0.086***	−0.075**	−0.075***	−0.031**	−0.033	−0.032*
	(0.031)	(0.029)	(0.029)	(0.016)	(0.021)	(0.017)
Observations	23,429	23,163	22,968	23,517	23,204	23,006
R^2	0.378	0.390	0.395	0.724	0.652	0.689
	Panel G: Strength of Family Ties					
Soil Heterogeneity	−0.248*	−0.240	−0.244*			
	(0.148)	(0.170)	(0.143)			
Observations	21,657	20,584	20,397			
R^2	0.644	0.645	0.674			
State × Year Fixed Effects	✓	✓	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓	✓	✓
Black Share	✓			✓		
Historical Slave Share			✓			✓

Note: This table reports estimates of multiple regressions in which the dependent variables are different historical features of close-knit communities: the LNI (Panel A), the share of ICM (Panel B), the TNI (Panel C), the RHI (Panel D), MKP (Panel E), the KPR (Panel F), and the SFTI (Panel G). Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. In columns 1 and 4 the data is in contemporaneous county borders, and in columns 2-3 and 5-6 the data is harmonized to 1940 county borders. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C.6.6 Agricultural inequality

Another competing mechanism is that a higher degree of local soil heterogeneity may have contributed to a higher degree of agricultural inequality due to a higher variation in agricultural productivity. A higher degree of agricultural inequality could, in turn, weaken communal ties and result in a loose-knit social structure. To test this, I use county-level data on the distribution of farm sizes for the years 1860-1940 (Manson et al., 2020) to calculate the Gini coefficient of the farms' size distribution and add it as a control. Appendix Table C.12 reports the results. For all seven historical measures of close-knit communities, there is almost no change in the coefficient of soil heterogeneity, suggesting that agricultural inequality is not an important part of the mechanism.

TABLE C.12: Agricultural Inequality Does Not Account for the Association

	Dependent variable:						
	Local Name Index	Intra-Communal Marriage	Tight Norms Index	Religious Homogeneity Index	Median Kin Propinquity	Kin Propinquity Rate	Strength of Family Ties
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Soil Heterogeneity	-2.826^{***}	-0.047^{**}	-0.516^{**}	-0.425^{***}	-0.084^{***}	-0.033^{*}	-0.252
	(0.819)	(0.022)	(0.202)	(0.131)	(0.033)	(0.020)	(0.171)
Observations	21,612	21,605	21,549	18,356	21,616	21,653	21,527
R^2	0.539	0.824	0.347	0.410	0.399	0.669	0.615
State × Year Fixed Effects	✓	✓	✓	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓	✓	✓	✓
Farm Size Gini	✓	✓	✓	✓	✓	✓	✓

Note: This table reports estimates of equation 1 when the dependent variables are different historical features of close-knit communities: the LNI (column 1), the share of ICM (column 2), the TNI (column 3), the RHI (column 4), MKP (column 5), the KPR (column 6), and the SFTI (column 7). Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. Farm size Gini is the Gini coefficient for the distribution of farm size. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C.6.7 Modernization and frontier experience

The selective migration of farmers to and from soil-heterogeneous counties (Appendix Table D.6 and Table 6) suggests that inducing modernization of the local economy is a potential mediator. Also, by limiting farmers' social learning, soil heterogeneity may have reduced agricultural productivity and induced transition out of agriculture. At the same time, environmental heterogeneity also has a direct effect on agricultural diversity (Appendix Table A.2), which has been shown to induce development (Fiszbein, 2022). Therefore, modernization might also be a competing mechanism. While the existing literature documents an association between different aspects of modernization (e.g. urbanization) and a loose-knit social structure (Enke, 2020; Triandis et al., 1990; Vandello and Cohen, 1999), causal links in both directions have been suggested. Moreover, the two can co-evolve (Eruchimovitch et al., 2023). Because economic modernization can be both the cause of a loose-knit social structure (Greenfield, 2009) and its effect (Gorodnichenko and Roland, 2011, 2017), and soil heterogeneity may have a direct impact on modernization, evaluating the role of modernization as a mediator is challenging.

As above, I evaluate the potential role of modernization by controlling for it directly and examining the change in the estimated impact of soil heterogeneity. I use four different proxies of modernization: the share of farmers, the literacy rate among household heads, the share of the population residing in urban locations, and the number of manufacturing establishments per 100 people. Table C.13 presents the results. Column 1 reports for comparison the result from my baseline specification of equation 1 when the sample is restricted to observations for which data on all four modernization proxies is available. In columns 2-5, I control separately for each of the different proxies of modernization, and in column 6, I control for all four proxies in the same regression. The results are mixed. For most measures of close-knit communities, the estimated coefficient of soil heterogeneity remains fairly stable and statistically significant when proxies of modernization are controlled for separately. However, in some cases, there is more significant attenuation and a loss of statistical significance. Moreover, when all proxies are controlled for in the same regression (column 6), the attenuation is stronger, and for four measures significance is lost. Yet even in this case, the coefficients in columns 1 and 6 are not statistically different from each other.

This suggests that modernization can at most account for some, but not all, of the historical association between soil heterogeneity and close-knit communities. It is possible that modernization was partly a competing mechanism alternative to the limited ability to engage in social learning and partly a mediating factor in a longer causal chain in which social learning is the fundamental factor. However, the evidence suggests that the latter is more likely. First, if the causal chain is that soil heterogeneity induced modernization, and modernization in turn induced loose-knit communi-

ties, it is hard to explain why the treatment effect and selective out-migration are concentrated on farmers and particularly those with a high tendency to rely on social networks. Second, any positive relationship between soil heterogeneity and modernization is at least partly the result of selective out-migration, in which farmers with a higher tendency to rely on social networks were more likely to leave soil-heterogeneous locations, presumably due to a limitation on social learning.

Finally, there is also the opposite concern—that soil heterogeneity delayed development, but possibly in a non-linear way, such that soil-heterogeneous counties remained in a frontier status for longer periods, which has been shown to induce "rugged individualism" (Bazzi et al., 2020). I use data from Bazzi et al. (2020) to control for frontier status and total frontier experience to explore the potential role of frontier status as a mediator or confounder. Appendix Table C.14 presents the results. Columns 1-2 focus on the years before the closing of the frontier and explore the potential role of contemporaneous frontier status. Columns 3-4 limit the sample to 1940 and explore the potential role of total frontier experience. The sample is restricted to the main sample of counties in Bazzi et al. (2020). I find that controlling for frontier status and total frontier experience makes little difference, as the estimated impact of soil heterogeneity in columns 1 and 3 are very similar to those in columns 2 and 4, respectively. This suggests that remaining in a frontier status for a longer period is not a central part of the story.

TABLE C.13: Modernization Can At Most Partially Accounts for the Association

	Dependent variable:					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Local Name Index (mean = 67.9, SD = 6.3)					
Soil Heterogeneity	-3.035^{***}	-2.142^{***}	-2.646^{***}	-1.837^{***}	-2.822^{***}	-1.489^{**}
	(0.837)	(0.758)	(0.743)	(0.645)	(0.781)	(0.614)
Observations	14,731	14,731	14,731	14,731	14,731	14,731
R^2	0.551	0.609	0.563	0.597	0.558	0.625
	Panel B: Intra-Community Marriage (mean = 0.63, SD = 0.22)					
Soil Heterogeneity	-0.058^{**}	-0.046^{**}	-0.049^{**}	-0.044^{*}	-0.057^{**}	-0.037
	(0.023)	(0.023)	(0.024)	(0.024)	(0.023)	(0.024)
Observations	14,697	14,697	14,697	14,697	14,697	14,697
R^2	0.823	0.832	0.828	0.828	0.823	0.836
	Panel C: Tight Norms Index (mean = -0.02, SD = 0.96)					
Soil Heterogeneity	-0.491^{***}	-0.365^{**}	-0.417^{***}	-0.278^{*}	-0.469^{***}	-0.232
	(0.186)	(0.177)	(0.161)	(0.149)	(0.180)	(0.142)
Observations	14,636	14,636	14,636	14,636	14,636	14,636
R^2	0.366	0.418	0.385	0.431	0.369	0.450
	Panel D: Religious Homogeneity Index (mean = -0.05, SD = 0.95)					
Soil Heterogeneity	-0.381^{***}	-0.310^{***}	-0.308^{***}	-0.250^{**}	-0.346^{***}	-0.193^{*}
	(0.119)	(0.114)	(0.112)	(0.105)	(0.113)	(0.103)
Observations	11,892	11,892	11,892	11,892	11,892	11,892
R^2	0.419	0.437	0.436	0.445	0.427	0.460
State × Year Fixed Effects	✓	✓	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓	✓	✓
Share Urban		✓				✓
Share Literate			✓			✓
Share Farmers				✓		✓
Manufacturing Est.					✓	✓

Note: Table continues to the next page.

TABLE C.13: Modernization Can At Most Partially Accounts for the Association (cont.)

	Dependent variable:					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel E: Median Kin Propinquity (mean = 0.13, SD = 0.12)					
Soil Heterogeneity	-0.087^{***}	-0.071^{**}	-0.080^{***}	-0.057^{**}	-0.083^{***}	-0.052^{**}
	(0.030)	(0.028)	(0.027)	(0.024)	(0.028)	(0.023)
Observations	14,711	14,711	14,711	14,711	14,711	14,711
R^2	0.395	0.450	0.407	0.481	0.401	0.493
	Panel F: Kin Propinquity Rate (mean = 0.33, SD = 0.11)					
Soil Heterogeneity	-0.037^{*}	-0.023	-0.027^{*}	-0.018	-0.035^{*}	-0.011
	(0.020)	(0.019)	(0.016)	(0.016)	(0.019)	(0.015)
Observations	14,731	14,731	14,731	14,731	14,731	14,731
R^2	0.656	0.699	0.683	0.692	0.657	0.725
	Panel G: Strength of Family Ties Index (mean = 0.02, SD = 0.98)					
Soil Heterogeneity	-0.290^{*}	-0.206	-0.217	-0.120	-0.264	-0.074
	(0.168)	(0.161)	(0.146)	(0.143)	(0.161)	(0.132)
Observations	14,601	14,601	14,601	14,601	14,601	14,601
R^2	0.611	0.634	0.627	0.651	0.615	0.662
State × Year Fixed Effects	✓	✓	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓	✓	✓
Share Urban		✓				✓
Share Literate			✓			✓
Share Farmers				✓		✓
Manufacturing Est.					✓	✓

Note: This table reports estimates of equation 1 when the dependent variables are different historical features of close-knit communities: the LNI (Panel A), the share of ICM (Panel B), the TNI (Panel C), the RHI (Panel D), MKP (Panel E), the KPR (Panel F), and the SFTI (Panel G). Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. In addition to the baseline controls, four modernization indicators are controlled for: share urban, share literate, share farmers and manufacturing establishments per 100 people. The sample is restricted to observations in which data on all four modernization indicators is available. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE C.14: Frontier Experience Does Not Account for the Association

	Dependent variable:			
	Frontier Status 1850-1880		Total Frontier Experience 1940	
	(1)	(2)	(3)	(4)
	Panel A: Local Name Index			
Soil Heterogeneity	-4.508^{***}	-3.966^{***}	-2.542^{***}	-2.405^{***}
	(1.172)	(1.102)	(0.741)	(0.725)
Observations	7,484	7,484	2,038	2,038
R^2	0.644	0.669	0.594	0.598
	Panel B: Intra-Community Marriage			
Soil Heterogeneity	-0.041	-0.043	-0.050^{*}	-0.050^{*}
	(0.032)	(0.032)	(0.026)	(0.026)
Observations	7,489	7,489	2,038	2,038
R^2	0.676	0.677	0.626	0.626
	Panel C: Tight Norms Index			
Soil Heterogeneity	-0.325	-0.275	-0.739^{**}	-0.719^{**}
	(0.217)	(0.219)	(0.322)	(0.319)
Observations	7,370	7,370	2,038	2,038
R^2	0.457	0.478	0.181	0.186
	Panel D: Religious Homogeneity Index			
Soil Heterogeneity	-0.660^{***}	-0.623^{***}	-0.382^{***}	-0.371^{**}
	(0.221)	(0.218)	(0.143)	(0.144)
Observations	4,785	4,785	2,038	2,038
R^2	0.357	0.395	0.368	0.369
State × Year Fixed Effects	✓	✓		
State Fixed Effects			✓	✓
Geoclimatic Controls	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓
Frontier Status		✓		
Total Frontier Experience				✓

Note: Table continues to the next page.

TABLE C.14: Frontier Experience Does Not Account for the Association (cont.)

	Dependent variable:			
	Frontier Status 1850-1880		Total Frontier Experience 1940	
	(1)	(2)	(3)	(4)
	Panel E: Median Kin Propinquity			
Soil Heterogeneity	-0.090^{***}	-0.083^{***}	-0.089^{**}	-0.086^{*}
	(0.022)	(0.022)	(0.044)	(0.044)
Observations	7,495	7,495	2,038	2,038
R^2	0.237	0.252	0.471	0.473
	Panel F: Kin Propinquity Rate			
Soil Heterogeneity	-0.036^{*}	-0.038^{*}	-0.053^{**}	-0.049^{**}
	(0.021)	(0.021)	(0.025)	(0.024)
Observations	7,558	7,558	2,038	2,038
R^2	0.614	0.615	0.527	0.541
	Panel G: Strength of Family Ties Index			
Soil Heterogeneity	-0.421^{**}	-0.430^{**}	-0.497^{**}	-0.463^{*}
	(0.172)	(0.172)	(0.243)	(0.238)
Observations	5,585	5,585	2,038	2,038
R^2	0.443	0.444	0.558	0.566
State × Year Fixed Effects	✓	✓		
State Fixed Effects			✓	✓
Geoclimatic Controls	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓
Frontier Status		✓		
Total Frontier Experience				✓

Note: This table reports estimates of equation 1 when the dependent variables are different historical features of close-knit communities: the LNI (Panel A), the share of ICM (Panel B), the TNI (Panel C), the RHI (Panel D), MKP (Panel E), the KPR (Panel F), and the SFTI (Panel G). Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. The sample is restricted to the baseline sample in Bazzi et al. (2020). In columns 1-2, the sample is 1850-1880 and column 2 additionally controls for contemporary frontier status. In columns 3-4, the sample is 1940 and column 4 additionally controls for total frontier experience. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

D Robustness Checks

D.1 Robustness of County-Level Results

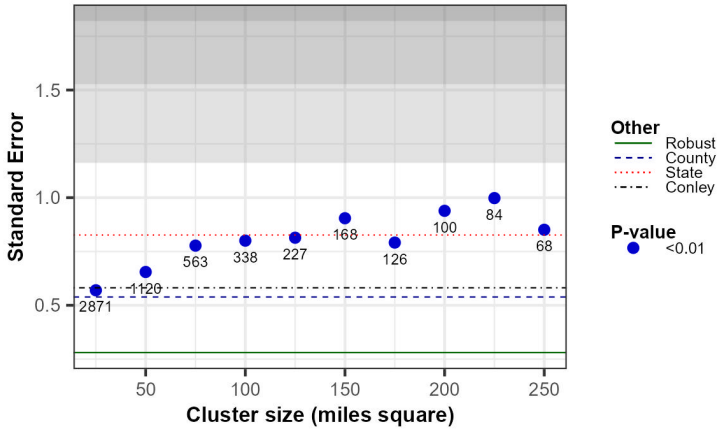
TABLE D.1: Robustness to Dropping Outliers

	Dependent variable:						
	Local Name Index	Intra-Communal Marriage	Tight Norms Index	Religious Homogeneity Index	Median Kin Propinquity	Kin Propinquity Rate	Strength of Family Ties
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: $0.25 < SHI < 0.75$						
Soil Heterogeneity	-2.603^{***}	-0.045^{**}	-0.403^{**}	-0.432^{***}	-0.058^{**}	-0.016	-0.101
	(0.864)	(0.021)	(0.172)	(0.147)	(0.024)	(0.016)	(0.152)
Observations	20,785	20,779	20,678	17,622	20,791	20,872	19,212
R^2	0.553	0.816	0.319	0.400	0.375	0.679	0.599
	Panel B: $P_5 < SHI < P_{95}$						
Soil Heterogeneity	-3.364^{***}	-0.043^{*}	-0.282^{*}	-0.436^{***}	-0.050^{**}	-0.002	0.036
	(0.811)	(0.024)	(0.155)	(0.146)	(0.019)	(0.016)	(0.157)
Observations	21,090	21,083	20,986	17,893	21,100	21,175	19,493
R^2	0.553	0.817	0.326	0.405	0.386	0.674	0.609
State × Year Fixed Effects	✓	✓	✓	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓	✓	✓	✓

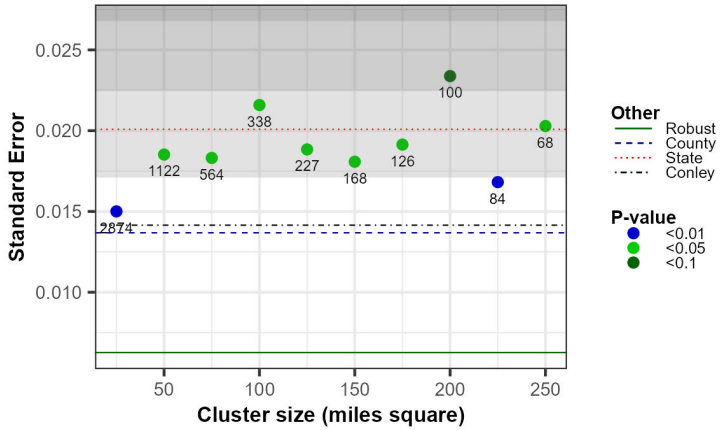
Note: This table reports estimates of equation 1 when the dependent variables are different historical features of close-knit communities: the LNI (column 1), the share of ICM (column 2), the TNI (column 3), the RHI (column 4), the MKP (column 5), the KPR (column 6), and the SFTI (column 7). Outlier counties in terms of soil heterogeneity are dropped. In Panel A counties with absolute low values (below 0.25) and absolute high values (above 0.75) are dropped, and in Panel B the bottom and top 5% of the distribution is dropped. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

FIGURE D.1: Robustness to Different Standard Errors

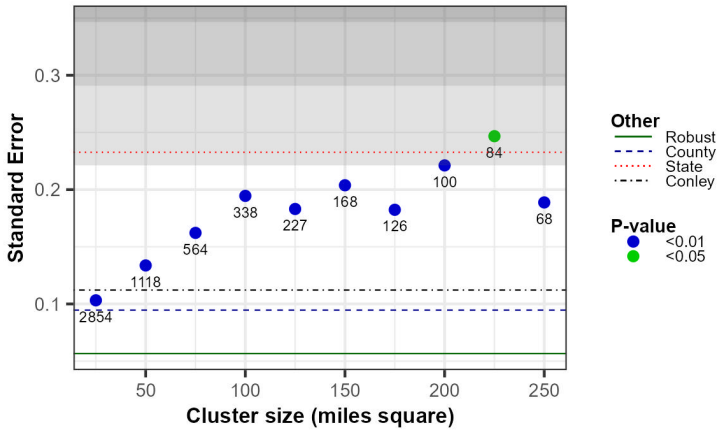
(A) Local Name Index



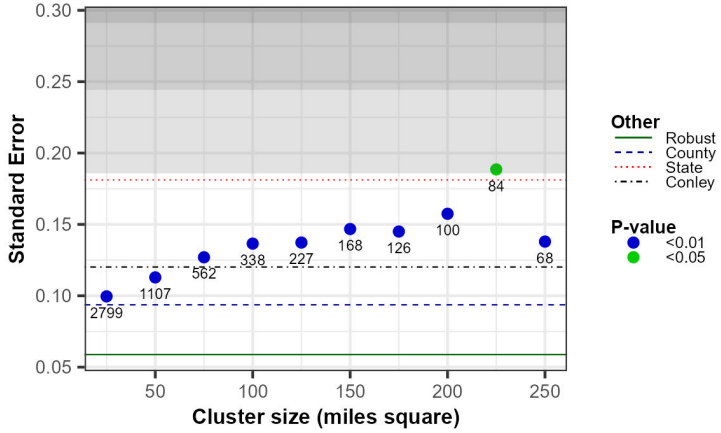
(B) Intra-Community Marriage



(C) Tight Norms Index



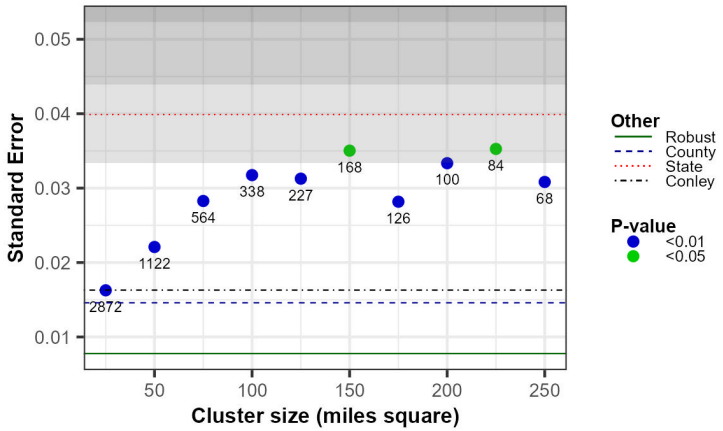
(D) Religious Homogeneity Index



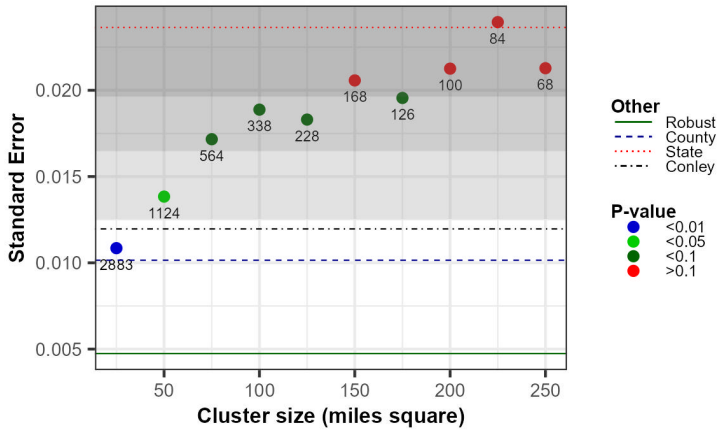
Note: Figure continues to the next page.

FIGURE D.1: Robustness to Different Standard Errors

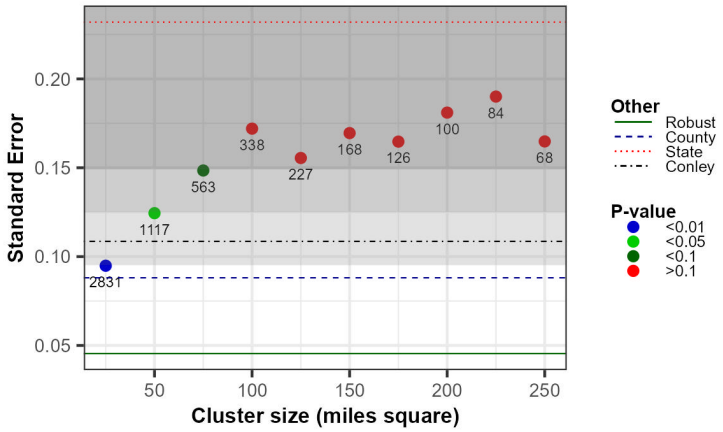
(E) Median Kin Propinquity



(F) Kin Propinquity Rate



(G) Strength of Family Ties Index



Note: This figure plots the standard errors of β from the preferred specification of equation 1 using different approaches for inference. The blue dots represent the standard errors (on the y-axis) using arbitrary grid cells of different sizes (on the x-axis), as proposed by Bester et al. (2011). The numeric label under each dot indicates the number of spatial clusters. The solid black horizontal line plots Conley (1999) standard errors allowing a linearly decaying spatial correlation over 500 miles. The dotted dark green horizontal line plots the HC robust standard errors, the dashed dark blue horizontal line plots the standard errors when clustering at the county level, and the dash-dotted red horizontal line plots the standard errors when clustering at the state level. The background color is indicative of the level of statistical significance. The p-value is < 0.01 in the white area, and < 0.05, < 0.1 and > 0.1 in the light to dark shades of gray. See Appendix E for details on the data and variable construction.

TABLE D.2: Robustness to Dropping States Fixed Effects

	Dependent variable:					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Local Name Index (mean = 67.8, SD = 6.3)					
Soil Heterogeneity	-4.560^{***}	-4.568^{***}	-3.713^{***}	-2.654^{**}	-2.132^{***}	-2.192^{***}
	(1.343)	(1.361)	(1.053)	(1.083)	(0.807)	(0.790)
Observations	23,437	23,437	23,437	23,437	23,437	23,375
R^2	0.006	0.040	0.234	0.344	0.378	0.386
	Panel B: Intra-Community Marriage (mean = 0.63, SD = 0.21)					
Soil Heterogeneity	-0.108^{*}	-0.106	-0.115^{**}	-0.081^{***}	-0.077^{***}	-0.094^{***}
	(0.065)	(0.066)	(0.049)	(0.030)	(0.030)	(0.029)
Observations	23,431	23,431	23,431	23,431	23,431	23,369
R^2	0.003	0.028	0.309	0.577	0.599	0.611
	Panel C: Tight Norms Index (mean = 0, SD = 1)					
Soil Heterogeneity	-1.187^{***}	-1.187^{***}	-0.722^{***}	-0.632^{***}	-0.478^{***}	-0.487^{***}
	(0.263)	(0.263)	(0.230)	(0.229)	(0.168)	(0.162)
Observations	23,322	23,322	23,322	23,322	23,322	23,260
R^2	0.017	0.017	0.155	0.195	0.213	0.215
	Panel D: Religious Homogeneity Index (mean = 0, SD = 1)					
Soil Heterogeneity	-0.275	-0.275	-0.310^{*}	-0.270	-0.151	-0.259^{*}
	(0.217)	(0.217)	(0.179)	(0.189)	(0.155)	(0.147)
Observations	19,881	19,881	19,881	19,881	19,881	19,837
R^2	0.001	0.001	0.207	0.235	0.253	0.268
Year Fixed Effects		✓	✓	✓	✓	✓
Geoclimatic Controls			✓	✓	✓	✓
Smooth Location Controls				✓	✓	✓
Agricultural Suitability Controls					✓	✓
Higher Order Controls						✓

Note: Table continues to the next page.

TABLE D.2: Robustness to Dropping States Fixed Effects (cont.)

	Dependent variable:					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel E: Median Kin Propinquity (mean = 0.14, SD = 0.13)					
Soil Heterogeneity	-0.110^{***}	-0.112^{***}	-0.100^{***}	-0.081^{**}	-0.059^{**}	-0.066^{***}
	(0.040)	(0.040)	(0.036)	(0.036)	(0.026)	(0.025)
Observations	23,443	23,443	23,443	23,443	23,443	23,381
R^2	0.009	0.045	0.143	0.184	0.226	0.230
	Panel F: Kin Propinquity Rate (mean = 0.32, SD = 0.12)					
Soil Heterogeneity	-0.087^{**}	-0.091^{**}	-0.066^{**}	-0.057^{**}	-0.040^{*}	-0.053^{**}
	(0.044)	(0.042)	(0.034)	(0.027)	(0.022)	(0.021)
Observations	23,532	23,532	23,532	23,532	23,532	23,470
R^2	0.007	0.258	0.414	0.519	0.529	0.540
	Panel G: Strength of Family Ties Index (mean = 0, SD = 1)					
Soil Heterogeneity	-1.207^{***}	-1.207^{***}	-0.673^{***}	-0.472^{**}	-0.316^{*}	-0.458^{***}
	(0.288)	(0.288)	(0.231)	(0.205)	(0.161)	(0.154)
Observations	21,671	21,671	21,671	21,671	21,671	21,612
R^2	0.018	0.018	0.353	0.430	0.446	0.462
Year Fixed Effects		✓	✓	✓	✓	✓
Geoclimatic Controls			✓	✓	✓	✓
Smooth Location Controls				✓	✓	✓
Agricultural Suitability Controls					✓	✓
Higher Order Controls						✓

Note: This table reports estimates of equation 1 when the dependent variables are different historical features of close-knit communities: the LNI (Panel A), the share of ICM (Panel B), the TNI (Panel C), the RHI (Panel D), the MKP (Panel E), the KPR (Panel F), and the SFTI (Panel G). Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. Agricultural suitability controls include average agricultural suitability indices for alfalfa, cotton, maize, oat, rye, sugarcane, sweet potato, tobacco, wheat, and white potato. Higher order controls include the standard deviations of temperature, precipitation, slope, elevation, flow accumulation, and the agricultural suitability indices. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE D.3: Alternative Definitions of the LNI

	Dependent variable: Local Name Index					
LNI definition	Baseline	NYSIIS phonetic algorithm	At least 100 name repetitions	At least 1000 name repetitions	At least 100 children in county	State defined as local
	(1)	(2)	(3)	(4)	(5)	(6)
Soil Heterogeneity	-2.994^{***}	-3.098^{***}	-2.952^{***}	-2.947^{***}	-2.453^{***}	-1.788^{**}
	(0.800)	(0.774)	(0.792)	(0.782)	(0.754)	(0.782)
Dependent Variable Mean	67.82	63.52	64.49	61.26	67.12	57.11
Dependent Variable SD	6.30	6.15	6.53	6.56	5.08	4.31
Observations	23,437	23,437	23,434	23,430	22,594	23,437
R^2	0.548	0.514	0.524	0.497	0.559	0.437
State × Year Fixed Effects	✓	✓	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓	✓	✓

Note: This table reports estimates of equation 1 when the dependent variable is children's LNI under different definitions. "Local" is defined as the county in columns 1-5 and as the state in column 6. Column 2 uses the NYSIIS phonetic algorithm to remove variation that results from different spelling of phonetically similar names. In columns 3 and 4 the sample is restricted to include names that are observed at least 100 and 1000 times nationally within the same period, respectively. Column 5 drops from the sample counties with less than 100 children. Geoclimatic controls include average temperature, average precipitation, average slope, average elevation, average absolute agricultural productivity, flow accumulation, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE D.4: Robustness to the Inclusion of Immigrants and Non-Whites

	Dependent variable:			
Sample:	Baseline	Include foreign-born	Include all races	Include all races and foreign-born
	(1)	(2)	(3)	(4)
	Panel A: Local Name Index			
Soil Heterogeneity	-2.994^{***}	-3.050^{***}	-2.809^{***}	-2.927^{***}
	(0.800)	(0.787)	(0.775)	(0.769)
Dependent Variable Mean	67.82	67.36	67.74	67.24
Dependent Variable SD	6.30	6.09	5.98	5.79
Observations	23,437	23,437	23,437	23,437
R^2	0.548	0.543	0.542	0.536
	Panel B: Intra-Community Marriage			
Soil Heterogeneity	-0.044^{**}	-0.046^{**}	-0.043^{**}	-0.048^{**}
	(0.022)	(0.021)	(0.021)	(0.021)
Dependent Variable Mean	0.63	0.63	0.64	0.64
Dependent Variable SD	0.21	0.20	0.21	0.20
Observations	23,431	23,442	23,437	23,445
R^2	0.813	0.796	0.817	0.805
	Panel C: Tight Norms Index			
Soil Heterogeneity	-0.570^{***}	-0.596^{***}	-0.579^{***}	-0.607^{***}
	(0.195)	(0.202)	(0.200)	(0.207)
Dependent Variable Mean	0.00	0.00	0.00	0.00
Dependent Variable SD	1.00	1.00	1.00	1.00
Observations	23,322	23,369	23,342	23,384
R^2	0.332	0.337	0.332	0.337
State × Year Fixed Effects	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓

Note: Table continues to the next page.

TABLE D.4: Robustness to the Inclusion of Immigrants and Non-Whites (cont.)

	Dependent variable:			
Sample:	Baseline	Include foreign-born	Include all races	Include all races and foreign-born
	(1)	(2)	(3)	(4)
	Panel D: Median Kin Propinquity			
Soil Heterogeneity	-0.086^{***}	-0.086^{***}	-0.076^{**}	-0.076^{**}
	(0.032)	(0.032)	(0.032)	(0.032)
Dependent Variable Mean	0.14	0.14	0.14	0.14
Dependent Variable SD	0.13	0.13	0.12	0.12
Observations	23,443	23,446	23,452	23,452
R^2	0.377	0.376	0.366	0.366
	Panel E: Kin Propinquity Rate			
Soil Heterogeneity	-0.032^{*}	-0.034^{*}	-0.049^{***}	-0.049^{***}
	(0.019)	(0.019)	(0.019)	(0.019)
Dependent Variable Mean	0.32	0.32	0.32	0.32
Dependent Variable SD	0.12	0.12	0.11	0.11
Observations	23,532	23,532	23,532	23,532
R^2	0.677	0.671	0.704	0.704
	Panel F: Strength of Family Ties Index			
Soil Heterogeneity	-0.245	-0.238	-0.350^{*}	-0.300
	(0.172)	(0.167)	(0.193)	(0.186)
Dependent Variable Mean	0.00	0.00	0.00	0.00
Dependent Variable SD	1.00	1.00	1.00	1.00
Observations	21,671	21,721	21,686	21,730
R^2	0.607	0.643	0.567	0.602
State × Year Fixed Effects	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓

Note: This table reports estimates of equation 1 when the dependent variables are different historical features of close-knit communities: the LNI (Panel A), the share of ICM (Panel B), the TNI (Panel C), the MKP (Panel D), the KPR (Panel E), and the SFTI (Panel F). The base sample used to calculate the county-level measures in column 1 includes white native-born individuals. In column 2 the sample also includes foreign-born, in column 3 it also includes non-whites, and in column 4 there are no sample restrictions on race or nativity. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE D.5: Robustness to Alternative Distances for SHI Calculation

SHI Cell Distance	Dependent variable: Baseline 25	10	15	20	30	35	40
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: Local Name Index (mean = 67.8, SD = 6.3)						
Soil Heterogeneity	-2.994^{***}	-2.663^{***}	-2.797^{***}	-2.912^{***}	-3.058^{***}	-3.106^{***}	-3.148^{***}
	(0.800)	(0.853)	(0.815)	(0.800)	(0.808)	(0.819)	(0.834)
Observations	23,437	23,437	23,437	23,437	23,437	23,437	23,437
R^2	0.548	0.547	0.547	0.548	0.548	0.548	0.548
	Panel B: Intra-Community Marriage (mean = 0.63, SD = 0.21)						
Soil Heterogeneity	-0.044^{**}	-0.030	-0.036^{*}	-0.041^{*}	-0.046^{**}	-0.048^{**}	-0.050^{**}
	(0.022)	(0.022)	(0.022)	(0.021)	(0.022)	(0.022)	(0.023)
Observations	23,431	23,431	23,431	23,431	23,431	23,431	23,431
R^2	0.813	0.813	0.813	0.813	0.813	0.813	0.813
	Panel C: Tight Norms Index (mean = 0, SD = 1)						
Soil Heterogeneity	-0.570^{***}	-0.442^{**}	-0.499^{***}	-0.541^{***}	-0.593^{***}	-0.611^{***}	-0.627^{***}
	(0.195)	(0.195)	(0.193)	(0.194)	(0.195)	(0.196)	(0.198)
Observations	23,322	23,322	23,322	23,322	23,322	23,322	23,322
R^2	0.332	0.330	0.331	0.331	0.332	0.332	0.332
	Panel D: Religious Homogeneity Index (mean = 0, SD = 1)						
Soil Heterogeneity	-0.479^{***}	-0.445^{***}	-0.462^{***}	-0.474^{***}	-0.479^{***}	-0.475^{***}	-0.469^{***}
	(0.137)	(0.142)	(0.137)	(0.136)	(0.138)	(0.141)	(0.144)
Observations	19,881	19,881	19,881	19,881	19,881	19,881	19,881
R^2	0.407	0.406	0.407	0.407	0.407	0.407	0.406
State × Year Fixed Effects	✓	✓	✓	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓	✓	✓	✓

Note: Table continues to the next page.

TABLE D.5: Robustness to Alternative Distances for SHI Calculation (cont.)

SHI Cell Distance	Baseline 25	10	15	20	30	35	40
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Dependent variable:						
	Panel E: Median Kin Propinquity (mean = 0.14, SD = 0.13)						
Soil Heterogeneity	-0.086^{***}	-0.074^{**}	-0.080^{***}	-0.084^{***}	-0.088^{***}	-0.088^{***}	-0.089^{***}
	(0.032)	(0.030)	(0.031)	(0.031)	(0.032)	(0.033)	(0.034)
Observations	23,443	23,443	23,443	23,443	23,443	23,443	23,443
R^2	0.377	0.376	0.377	0.377	0.377	0.377	0.377
	Panel F: Kin Propinquity Rate (mean = 0.32, SD = 0.12)						
Soil Heterogeneity	-0.032^{*}	-0.023	-0.027	-0.030	-0.034^{*}	-0.035^{*}	-0.035^{*}
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.020)
Observations	23,532	23,532	23,532	23,532	23,532	23,532	23,532
R^2	0.677	0.676	0.676	0.676	0.677	0.677	0.677
	Panel G: Strength of Family Ties Index (mean = 0, SD = 1)						
Soil Heterogeneity	-0.245	-0.155	-0.202	-0.229	-0.255	-0.260	-0.264
	(0.172)	(0.172)	(0.171)	(0.171)	(0.173)	(0.174)	(0.176)
Observations	21,671	21,671	21,671	21,671	21,671	21,671	21,671
R^2	0.607	0.607	0.607	0.607	0.607	0.607	0.607
State × Year Fixed Effects	✓	✓	✓	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓	✓	✓	✓

Note: This table reports estimates of equation 1 when the dependent variables are different historical features of close-knit communities: the LNI (Panel A), the share of ICM (Panel B), the TNI (Panel C), the RHI (Panel D), the MKP (Panel E), the KPR (Panel F), and the SFTI (Panel G). In each column the SHI is calculated over areas of different sizes, denoted in terms of 500-meter cell distances to each direction. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

D.2 Robustness of Selective In-Migration Results

TABLE D.6: Selective-In Migration: Not Controlling for Origin

	Dependent variable:					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A. Farming Experience (mean = 0.61, SD = 0.49)					
Soil Heterogeneity	−0.049***	−0.030***	−0.031***	−0.031***	−0.031***	−0.034***
	(0.012)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Observations	393,016	393,016	393,016	393,016	392,967	392,967
R^2	0.045	0.046	0.046	0.046	0.046	0.051
	Panel B. Children's Local Name Index (mean = 52.7, SD = 11.9)					
Soil Heterogeneity	−0.220	0.274	0.215	0.306	0.227	0.220
	(0.369)	(0.365)	(0.361)	(0.358)	(0.356)	(0.356)
Observations	570,121	570,121	570,121	570,121	566,731	566,731
R^2	0.012	0.013	0.013	0.014	0.014	0.014
	Panel C. Intra-Community Marriage (mean = 0.54, SD = 0.50)					
Soil Heterogeneity	−0.016	−0.009	−0.007	−0.009	−0.008	−0.009
	(0.022)	(0.021)	(0.020)	(0.020)	(0.021)	(0.021)
Observations	380,379	380,379	380,379	380,379	377,766	377,766
R^2	0.035	0.035	0.036	0.036	0.037	0.040
State × Year Fixed Effects	✓	✓	✓	✓	✓	✓
Geoclimatic Controls		✓	✓	✓	✓	✓
Smooth Location Controls			✓	✓	✓	✓
Agricultural Suitability Controls				✓	✓	✓
Higher Order Controls					✓	✓
Individual Controls						✓

Note: This table reports estimates of equation 2, when the dependent variable is prior farming experience (Panel A), the LNI score of children born before the migration, where "local" is defined as the state (Panel B), and intra-community marriage (Panel C). The sample in Panel A includes white native head of household farmers that migrated across counties, identified using the Census Linking Project (Abramitzky et al., 2022a,c,e), and in Panels B-C white native-born families with children between the ages of 0-10 that migrated once across states, identified using the year and state of birth of children. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. Agricultural suitability controls include average agricultural suitability indices for alfalfa, cotton, maize, oat, rye, sugarcane, sweet potato, tobacco, wheat, and white potato. Higher order controls include the standard deviations of temperature, precipitation, slope, elevation, flow accumulation, and the agricultural suitability indices. Individual controls include age fixed-effects in Panels A and C, and gender-by-year of birth and birth order fixed-effects in Panel B. See Appendix E for details on the data and variable construction. Standard errors clustered by destination county in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

D.3 Robustness of DID and DDD Results

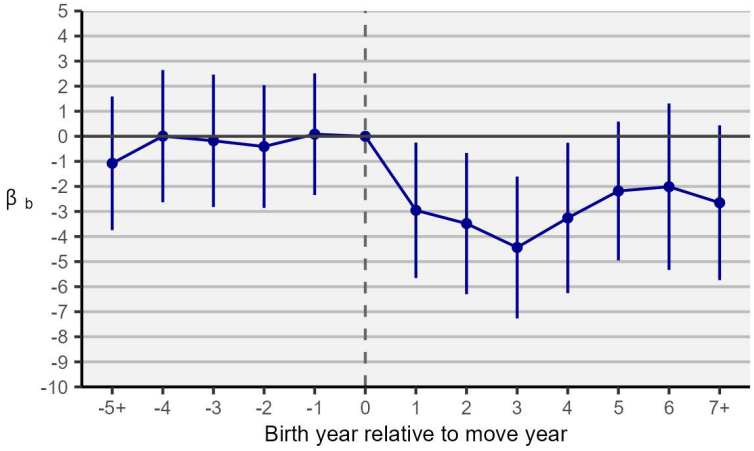
TABLE D.7: Adding Individual Controls to DID and DDD Estimates

	Dependent variable: Local Name Index			
	Difference-in-Differences			Triple-Difference
Sample:	All Households	Farmer′s Households	Non-Farmer′s Households	All Households
	(1)	(2)	(3)	(4)
Post Migration × Soil Heterogeneity	−1.741**	−3.086***	0.751	0.798
	(0.682)	(0.724)	(0.950)	(0.948)
Post Migration × Farmers′ Household × Soil Heterogeneity				−3.944***
				(0.969)
Dependent Variable Mean	54.22	54.40	53.95	54.22
Dependent Variable SD	13.67	13.42	14.01	13.67
Observations	1,203,721	713,850	489,866	1,203,716
R^2	0.365	0.350	0.386	0.365
Family Fixed Effects	✓	✓	✓	✓
Relative YOB Fixed Effects	✓	✓	✓	✓
Individual Controls	✓	✓	✓	✓

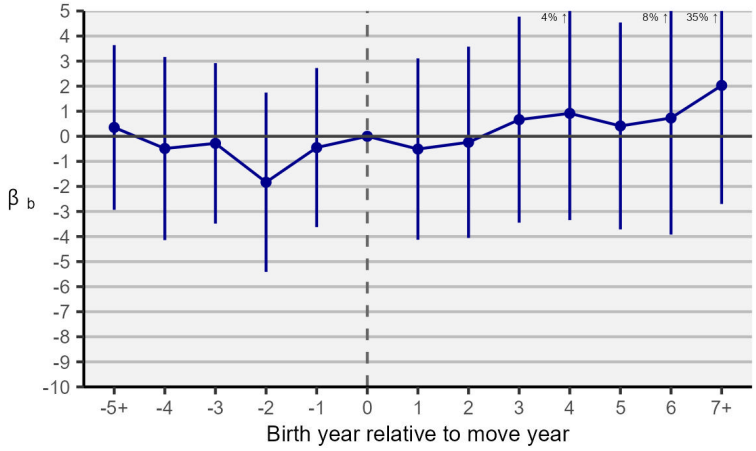
Note: This table presents estimates of the "static" versions of the Difference-in-Differences and Triple-Difference estimation frameworks (equations 3 and 4) when the dependent variable is children's LNI in which "local" is defined as the state. The sample includes white native-born children between the ages of 0 to 10 with native-born parents, whose families migrated once across states. Individual controls include gender, birth order, and a 5-year cohort fixed effects. See Appendix E for details on the data and variable construction. Standard errors clustered by county in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

FIGURE D.2: Adding Individual Controls to DID Estimates

(A) DID: Farmers' Families



(B) DID: Non-Farmers' Families



Note: This figure plots the estimates of β_b and 95% confidence intervals from equation 3 when the dependent variable is children's LNI where "local" is defined as the state. The sample includes white native-born children between the ages of 0 to 10 with native-born parents, whose families migrated once across states. The regressions also control for the child's gender, birth order, and a 5-year cohort fixed effect. See Appendix E for details on the data and variable construction. Standard errors are clustered at the destination county.

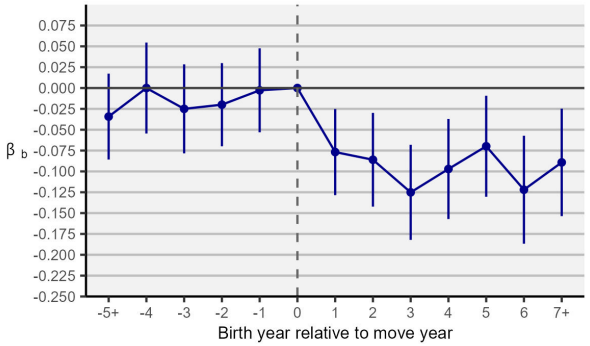
TABLE D.8: Robustness of DID Estimates to Sample Selection

Sample:	Dependent variable: Local Name Index		
	All Households	Farmer′s Households	Non-Farmer′s Households
	(1)	(2)	(3)
	Panel A: Include foreign-born parents		
Post Migration × Soil Heterogeneity	-2.167^{***}	-2.817^{***}	-1.016
	(0.664)	(0.703)	(1.103)
	Panel B: Include all races		
Post Migration × Soil Heterogeneity	-2.362^{***}	-3.714^{***}	-0.014
	(0.704)	(0.779)	(0.902)
	Panel C: Include multiple moves		
Post Migration × Soil Heterogeneity	-1.768^{***}	-3.218^{***}	0.853
	(0.685)	(0.734)	(0.938)
Family Fixed Effects	✓	✓	✓
Relative YOB Fixed Effects	✓	✓	✓

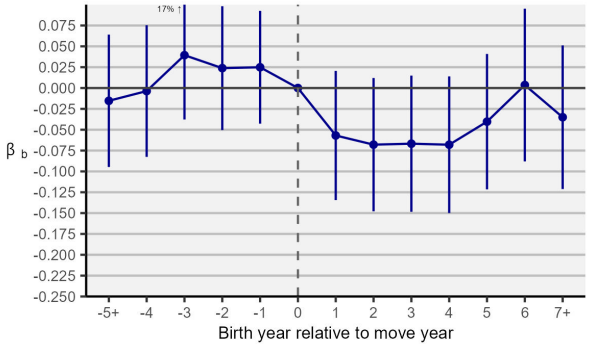
Note: This table presents estimates of the "static" version of the Difference-in-Differences estimation framework (equation 3) when the dependent variable is children's LNI in which "local" is defined as the state. See Appendix E for details on the data and variable construction. Standard errors clustered by destination county in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

FIGURE D.3: Heterogeneous Impact on Farmers by Prior Communal Identification, logs

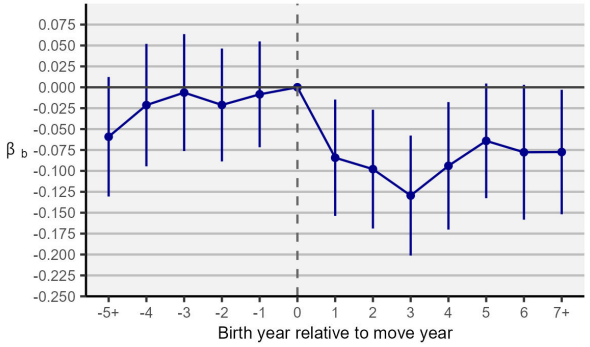
(A) DID: Farmers, High Prior Communal Identification (log)



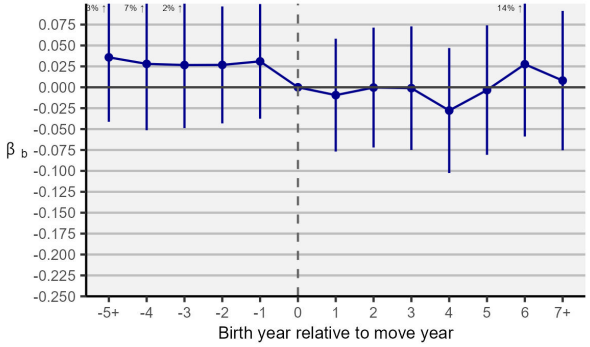
(B) DID: Farmers, Low Prior Communal Identification (log)



(C) DID: Farmers, Intra-Community Marriage (log)



(D) DID: Farmers, Extra-Community Marriage (log)



Note: This figure plots the estimates of β_b and 95% confidence intervals from equation 3 when the dependent variable is the natural logarithmic transformation of children's LNI, where "local" is defined as the state. The sample includes white native-born children between the ages of 0 to 10 with native-born parents, whose families migrated once across states and whose fathers are farmers. See Appendix E for details on the data and variable construction. Standard errors are clustered at the destination county.

D.4 Robustness of Migrants' Children Results

TABLE D.9: Additional Controls for Migrants' Children: Geographical Mobility (I)

	Dependent variable: Remained in the Destination County					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: All Migrants' Children (mean = 0.22, SD = 0.41)					
Soil Heterogeneity	−0.053	−0.044**	−0.043*	−0.042*	−0.052**	−0.059***
	(0.037)	(0.022)	(0.022)	(0.022)	(0.023)	(0.022)
Observations	134,233	134,233	134,233	134,233	134,233	133,533
R^2	0.0002	0.195	0.196	0.196	0.196	0.196
	Panel B: Farmers' Children (mean = 0.24, SD = 0.42)					
Soil Heterogeneity	−0.125***	−0.058**	−0.057**	−0.055**	−0.060**	−0.060**
	(0.041)	(0.026)	(0.026)	(0.026)	(0.025)	(0.025)
Observations	80,887	80,887	80,887	80,887	80,887	80,884
R^2	0.001	0.227	0.228	0.228	0.228	0.229
	Panel C: All Migrants' Children (mean = 0.22, SD = 0.41)					
Soil Heterogeneity	0.113**	0.018	0.019	0.019	0.007	−0.008
	(0.053)	(0.033)	(0.033)	(0.033)	(0.033)	(0.031)
Soil Heterogeneity × Farmers′ Household	−0.238***	−0.074*	−0.080**	−0.079**	−0.077**	−0.064*
	(0.059)	(0.039)	(0.038)	(0.038)	(0.038)	(0.036)
Observations	134,233	134,233	134,233	134,233	134,233	133,533
R^2	0.003	0.199	0.200	0.200	0.200	0.200
Cell Fixed Effects		✓	✓	✓	✓	✓
Geoclimatic Controls			✓	✓	✓	✓
Smooth Location Controls				✓	✓	✓
Agricultural Suitability Controls					✓	✓
Higher Order Controls						✓

Note: This table reports estimates of equation 5 when the dependent variable is remaining in the destination county as an adult. Cell fixed effects are fixed effects for the sextuple interaction of parental destination state × census year × birth state × relative-year-of-birth × high parental prior communal identification × parental intra-community marriage. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. Agricultural suitability controls include average agricultural suitability indices for alfalfa, cotton, maize, oat, rye, sugarcane, sweet potato, tobacco, wheat, and white potato. Higher order controls include the standard deviations of temperature, precipitation, slope, elevation, flow accumulation, and the agricultural suitability indices. All columns in Panel C also include the main effect of Farmers' Household. See Appendix E for details on the data and variable construction. Standard errors clustered by original destination county in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE D.10: Additional Controls for Migrants' Children: Geographical Mobility (II)

	Dependent variable: Remained in the Destination County					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: All Migrants' Children (mean = 0.22, SD = 0.41)					
Soil Heterogeneity	−0.093***	−0.043**	−0.044**	−0.038*	−0.040*	−0.042*
	(0.031)	(0.020)	(0.020)	(0.020)	(0.021)	(0.022)
Observations	134,233	134,233	134,233	134,233	134,233	134,233
R^2	0.010	0.019	0.023	0.050	0.106	0.196
	Panel B: Farmers' Children (mean = 0.24, SD = 0.42)					
Soil Heterogeneity	−0.136***	−0.052**	−0.050**	−0.046**	−0.050**	−0.055**
	(0.033)	(0.023)	(0.022)	(0.022)	(0.023)	(0.026)
Observations	80,887	80,887	80,887	80,887	80,887	80,887
R^2	0.012	0.021	0.026	0.057	0.127	0.228
	Panel C: All Migrants' Children (mean = 0.22, SD = 0.41)					
Soil Heterogeneity	0.033	0.028	0.024	0.027	0.030	0.019
	(0.037)	(0.029)	(0.029)	(0.030)	(0.031)	(0.033)
Soil Heterogeneity × Farmers' Household	−0.173***	−0.097***	−0.089***	−0.085**	−0.092***	−0.079**
	(0.042)	(0.034)	(0.033)	(0.033)	(0.035)	(0.038)
Observations	134,233	134,233	134,233	134,233	134,233	134,233
R^2	0.016	0.024	0.028	0.055	0.110	0.200
Destination State		✓	✓	✓	✓	✓
× Census year			✓	✓	✓	✓
× Birth State				✓	✓	✓
× Relative-year-of-birth					✓	✓
× High Parental Communalism						✓
Geoclimatic Controls	✓	✓	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓	✓	✓

Note: This table reports estimates of equation 5 when the dependent variable is remaining in the destination county as an adult. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. All columns in Panel C also include the main effect of Farmers' Household. See Appendix E for details on the data and variable construction. Standard errors clustered by original destination county in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE D.11: Additional Controls for Migrants' Children: Intra-Community Marriage (I)

	Dependent variable: Married a Spouse Born in the Destination State					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: All Migrants' Children (mean = 0.38, SD = 0.49)					
Soil Heterogeneity	−0.130**	−0.065***	−0.067***	−0.066***	−0.065***	−0.063**
	(0.057)	(0.025)	(0.025)	(0.025)	(0.024)	(0.024)
Observations	101,132	101,132	101,132	101,132	101,132	100,689
R^2	0.001	0.304	0.305	0.305	0.306	0.306
	Panel B: Farmers' Children (mean = 0.39, SD = 0.49)					
Soil Heterogeneity	−0.113	−0.027	−0.047	−0.050*	−0.049*	−0.047
	(0.070)	(0.030)	(0.030)	(0.030)	(0.029)	(0.029)
Observations	62,905	62,905	62,905	62,905	62,905	62,902
R^2	0.001	0.340	0.340	0.341	0.341	0.341
	Panel C: All Migrants' Children (mean = 0.38, SD = 0.49)					
Soil Heterogeneity	−0.126**	−0.075**	−0.081**	−0.078**	−0.078**	−0.080**
	(0.051)	(0.034)	(0.035)	(0.034)	(0.034)	(0.035)
Soil Heterogeneity × Farmers' Household	0.014	0.034	0.034	0.031	0.030	0.035
	(0.063)	(0.037)	(0.037)	(0.036)	(0.036)	(0.036)
Observations	101,132	101,132	101,132	101,132	101,132	100,689
R^2	0.002	0.306	0.307	0.307	0.308	0.308
Cell Fixed Effects		✓	✓	✓	✓	✓
Geoclimatic Controls			✓	✓	✓	✓
Smooth Location Controls				✓	✓	✓
Agricultural Suitability Controls					✓	✓
Higher Order Controls						✓

Note: This table reports estimates of equation 5 when the dependent variable is marrying a spouse born in the destination state. Cell fixed effects are fixed effects for the sextuple interaction of parental destination state × census year × birth state × relative-year-of-birth × high parental prior communal identification × parental intra-community marriage. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. Agricultural suitability controls include average agricultural suitability indices for alfalfa, cotton, maize, oat, rye, sugarcane, sweet potato, tobacco, wheat, and white potato. Higher order controls include the standard deviations of temperature, precipitation, slope, elevation, flow accumulation, and the agricultural suitability indices. All columns in Panel C also include the main effect of Farmers' Household. See Appendix E for details on the data and variable construction. Standard errors clustered by original destination county in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE D.12: Additional Controls for Migrants' Children: Intra-Community Marriage (II)

	Dependent variable: Married a Spouse Born in the Destination State					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: All Migrants' Children (mean = 0.38, SD = 0.49)					
Soil Heterogeneity	−0.134***	−0.072***	−0.064***	−0.049**	−0.055**	−0.066***
	(0.047)	(0.024)	(0.023)	(0.021)	(0.023)	(0.025)
Observations	101,132	101,132	101,132	101,132	101,132	101,132
R^2	0.019	0.083	0.091	0.157	0.216	0.305
	Panel B: Farmers' Children (mean = 0.39, SD = 0.49)					
Soil Heterogeneity	−0.102*	−0.068**	−0.057**	−0.040	−0.049*	−0.050*
	(0.053)	(0.028)	(0.027)	(0.025)	(0.027)	(0.030)
Observations	62,905	62,905	62,905	62,905	62,905	62,905
R^2	0.024	0.096	0.108	0.172	0.242	0.341
	Panel C: All Migrants' Children (mean = 0.38, SD = 0.49)					
Soil Heterogeneity	−0.153***	−0.032	−0.041	−0.040	−0.046	−0.078**
	(0.046)	(0.031)	(0.030)	(0.029)	(0.030)	(0.034)
Soil Heterogeneity × Farmers' Household	0.049	−0.049	−0.020	0.0002	−0.001	0.031
	(0.054)	(0.035)	(0.034)	(0.031)	(0.033)	(0.036)
Observations	101,132	101,132	101,132	101,132	101,132	101,132
R^2	0.022	0.085	0.094	0.160	0.218	0.307
Destination State		✓	✓	✓	✓	✓
× Census year			✓	✓	✓	✓
× Birth State				✓	✓	✓
× Relative-year-of-birth					✓	✓
× High Parental Communalism						✓
Geoclimatic Controls	✓	✓	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓	✓	✓

Note: This table reports estimates of equation 5 when the dependent variable is marrying a spouse born in the destination state. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. All columns in Panel C also include the main effect of Farmers' Household. See Appendix E for details on the data and variable construction. Standard errors clustered by original destination county in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

D.5 Robustness of Social Learning Results

TABLE D.13: Social Learning: Robustness to Different Measures

	Dependent variable:						
Sample:	All	All	All	Low Learning Potential	High Learning Potential	All	All
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: Fertilizer Adoption						
	Growth in Fertilizer Use						Learning Potential
Soil Heterogeneity	−0.771***		−0.774***	−1.184***	−0.356	−0.379	−0.061
	(0.223)		(0.221)	(0.292)	(0.246)	(0.248)	(0.062)
High Learning Potential		0.082	0.086			0.537**	
		(0.072)	(0.072)			(0.250)	
Soil Heterogeneity × High Learning Potential						−0.711*	
						(0.367)	
Observations	8,983	8,983	8,983	4,473	4,510	8,983	3,102
R^2	0.294	0.293	0.294	0.353	0.269	0.294	0.918
	Panel B: Wheat Adaptation						
	Growth in Wheat Production						Learning Potential
Soil Heterogeneity	−0.510***		−0.492***	−0.505***	−0.161	0.125	−0.084
	(0.169)		(0.166)	(0.146)	(0.249)	(0.289)	(0.140)
High Learning Potential		0.099	0.089			0.689***	
		(0.062)	(0.061)			(0.227)	
Soil Heterogeneity × High Learning Potential						−0.957***	
						(0.337)	
Observations	25,181	25,181	25,181	13,666	11,515	25,181	3,102
R^2	0.402	0.402	0.402	0.458	0.441	0.403	0.755
State × Year Fixed Effects	✓	✓	✓	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓	✓	✓	✓

Note: This table reports estimates of equation 1 when the dependent variables are different indicators for farmers' social learning. Panel A focuses on the growth rate of the share of farms using fertilizers, and Panel B on the growth rate of wheat production. Geoclimatic controls include average temperature, precipitation, slope, elevation, absolute agricultural productivity, flow accumulation, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE D.14: Fertilizers: Robustness to the Inclusion of Controls

	Dependent variable: Growth in Fertilizers Use						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: All Counties						
Soil Heterogeneity	-0.891^{***}	-0.857^{***}	-0.737^{***}	-0.771^{***}	-0.773^{***}	-0.713^{***}	-0.289^{*}
	(0.304)	(0.224)	(0.220)	(0.223)	(0.213)	(0.208)	(0.162)
Share Using Fertilizer$_{t-1}$							-2.615^{***}
							(0.133)
Oster δ for $\beta = 0$		7.32	3.62	4.06	3.88	3.06	1.00
Observations	8,983	8,983	8,983	8,983	8,983	8,971	8,971
R^2	0.002	0.278	0.292	0.294	0.295	0.300	0.333
	Panel B: High Learning Potential Counties						
Soil Heterogeneity	-1.455^{***}	-1.407^{***}	-1.101^{***}	-1.126^{***}	-1.079^{***}	-0.992^{***}	-0.542^{***}
	(0.380)	(0.294)	(0.266)	(0.267)	(0.269)	(0.249)	(0.198)
Share Using Fertilizer$_{t-1}$							-2.786^{***}
							(0.195)
Observations	4,472	4,472	4,472	4,472	4,472	4,460	4,460
R^2	0.007	0.347	0.365	0.366	0.369	0.373	0.404
State × Year Fixed Effects		✓	✓	✓	✓	✓	✓
Geoclimatic Controls			✓	✓	✓	✓	✓
Smooth Location Controls				✓	✓	✓	✓
Agricultural Suitability Controls					✓	✓	✓
Higher Order Controls						✓	✓

Note: This table reports estimates of equation 1 when the dependent variable is the county-level inverse hyperbolic sine (IHS) transformation of the 10-year growth rate in the share of farms using fertilizers. The sample in Panel A includes all counties, and in Panel B it is restricted to counties with a high learning potential. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. Agricultural suitability controls include average agricultural suitability indices for alfalfa, cotton, maize, oat, rye, sugarcane, sweet potato, tobacco, wheat, and white potato. Higher order controls include the standard deviations of temperature, precipitation, slope, elevation, flow accumulation, and the agricultural suitability indices. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE D.15: Wheat Land Share: Robustness to the Inclusion of Controls

	Dependent variable: Growth in Wheat Acreage Share						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: All Counties						
Soil Heterogeneity	-0.616^{**}	-0.418	-0.747^{***}	-0.600^{***}	-0.484^{***}	-0.408^{**}	-0.309^{**}
	(0.268)	(0.266)	(0.225)	(0.218)	(0.166)	(0.161)	(0.146)
Wheat Share$_{t-1}$							-2.784^{***}
							(0.185)
Oster δ for $\beta = 0$		2.72	34.90	7.30	3.76	2.51	1.87
Observations	19,531	19,531	19,531	19,531	19,531	19,502	19,502
R^2	0.001	0.369	0.387	0.393	0.398	0.401	0.420
	Panel B: High Learning Potential Counties						
Soil Heterogeneity	-1.399^{***}	-0.901^{***}	-0.677^{***}	-0.662^{***}	-0.504^{***}	-0.430^{**}	-0.340^{*}
	(0.262)	(0.228)	(0.228)	(0.223)	(0.188)	(0.187)	(0.177)
Wheat Share$_{t-1}$							-3.290^{***}
							(0.234)
Observations	10,472	10,472	10,472	10,472	10,472	10,443	10,443
R^2	0.010	0.378	0.387	0.390	0.394	0.399	0.439
State × Year Fixed Effects		✓	✓	✓	✓	✓	✓
Geoclimatic Controls			✓	✓	✓	✓	✓
Smooth Location Controls				✓	✓	✓	✓
Agricultural Suitability Controls					✓	✓	✓
Higher Order Controls						✓	✓

Note: This table reports estimates of equation 1 when the dependent variable is the county-level inverse hyperbolic sine (IHS) transformation of the 10-year growth rate in the share of land dedicated to cultivating wheat. The sample in Panel A includes all counties, and in Panel B it is restricted to counties with a high learning potential. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. Agricultural suitability controls include average agricultural suitability indices for alfalfa, cotton, maize, oat, rye, sugarcane, sweet potato, tobacco, wheat, and white potato. Higher order controls include the standard deviations of temperature, precipitation, slope, elevation, flow accumulation, and the agricultural suitability indices. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

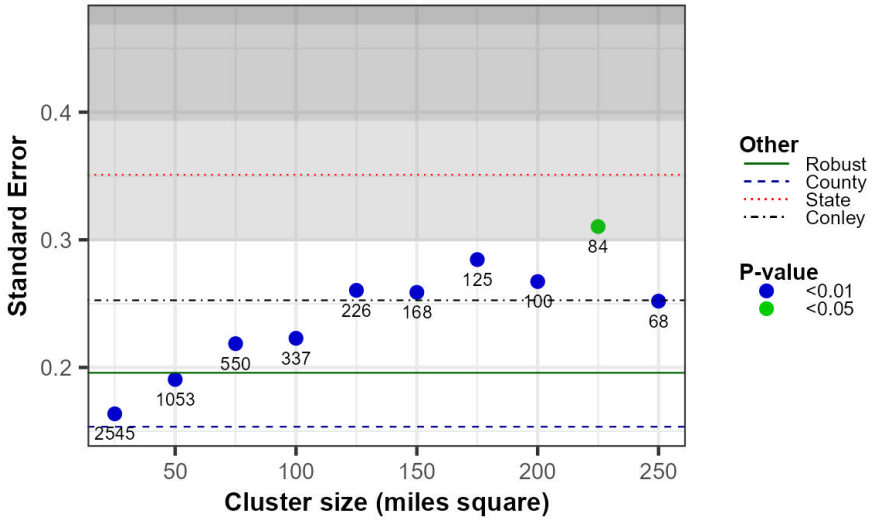
TABLE D.16: Robustness of Interaction with High Learning Potential

	Dependent variable:					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Growth in Fertilizer Use					
Soil Heterogeneity	−0.284	−0.245	−0.249	−0.277	−0.403	−0.325
	(0.375)	(0.256)	(0.263)	(0.260)	(0.258)	(0.248)
High Learning Potential	0.651*	0.847***	0.746***	0.699***	0.679***	0.678***
	(0.340)	(0.268)	(0.256)	(0.252)	(0.249)	(0.241)
Soil Heterogeneity × High Learning Potential	−1.171**	−1.137***	−0.888**	−0.858**	−0.605	−0.640*
	(0.497)	(0.404)	(0.373)	(0.371)	(0.370)	(0.356)
Observations	8,983	8,983	8,983	8,983	8,983	8,971
R^2	0.004	0.279	0.293	0.295	0.297	0.302
	Panel B: Growth in Wheat Share					
Soil Heterogeneity	0.053	0.282	−0.344	−0.064	−0.136	−0.109
	(0.435)	(0.428)	(0.289)	(0.261)	(0.234)	(0.230)
High Learning Potential	0.434	0.676**	0.368	0.626***	0.463**	0.392*
	(0.318)	(0.331)	(0.270)	(0.242)	(0.203)	(0.208)
Soil Heterogeneity × High Learning Potential	−1.453***	−1.147**	−0.644*	−0.805**	−0.513*	−0.451
	(0.505)	(0.498)	(0.391)	(0.352)	(0.290)	(0.299)
Observations	19,531	19,531	19,531	19,531	19,531	19,502
R^2	0.015	0.370	0.387	0.393	0.398	0.401
State × Year Fixed Effects		✓	✓	✓	✓	✓
Geoclimatic Controls			✓	✓	✓	✓
Smooth Location Controls				✓	✓	✓
Agricultural Suitability Controls					✓	✓
Higher Order Controls						✓

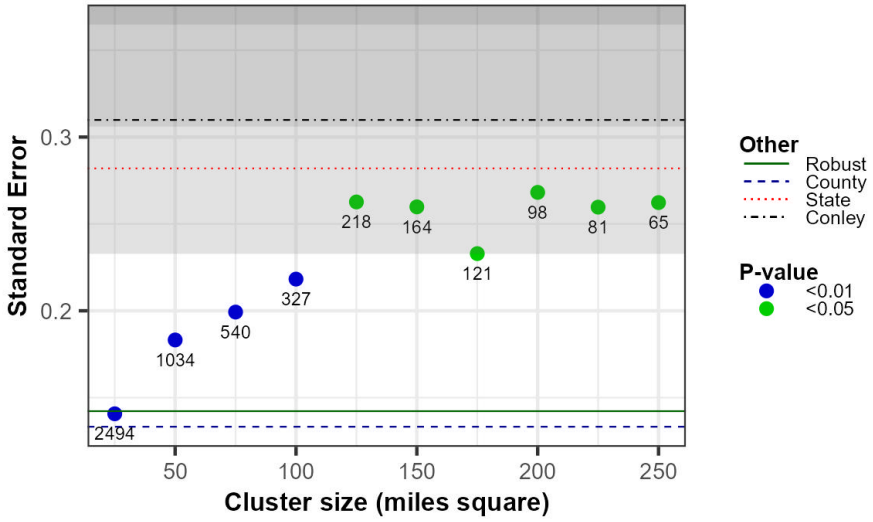
Note: This table reports estimates of equation 1 when SHI is interacted with high learning potential and the dependent variables are different indicators for farmers' social learning. Panel A focuses on the growth rate of the share of farms using fertilizers, and Panel B on the growth rate of the share of land devoted to wheat cultivation. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. Agricultural suitability controls include average agricultural suitability indices for alfalfa, cotton, maize, oat, rye, sugarcane, sweet potato, tobacco, wheat, and white potato. Higher order controls include the standard deviations of temperature, precipitation, slope, elevation, flow accumulation, and the agricultural suitability indices. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

FIGURE D.4: Social Learning: Robustness to Different Standard Errors

(A) Fertilizer Adoption



(B) Wheat Adaptation



Note: This figure plots the standard errors of β from the preferred specification of equation 1 using different approaches for inference. The blue dots represent the standard errors (on the y-axis) using arbitrary grid cells of different sizes (on the x-axis), as proposed by Bester et al. (2011). The numeric label under each dot indicates the number of spatial clusters. The solid black horizontal line plots Conley (1999) standard errors allowing a linearly decaying spatial correlation over 500 miles. The dotted dark green horizontal line plots the HC robust standard errors, the dashed dark blue horizontal line plots the standard errors when clustering at the county level, and the dash-dotted red horizontal line plots the standard errors when clustering at the state level. The background color is indicative of the level of statistical significance. The p-value is < 0.01 in the white area, and < 0.05, < 0.1, and > 0.1 in the light to dark shades of gray. See Appendix E for details on the data and variable construction.

TABLE D.17: Social Learning: Robustness to Dropping States Fixed Effects

	Dependent variable:						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: Fertilizer Adoption						
Soil Heterogeneity	-0.891^{***}	-0.890^{***}	-0.737^{***}	-0.741^{***}	-0.726^{***}	-0.638^{***}	-0.155
	(0.304)	(0.303)	(0.267)	(0.250)	(0.232)	(0.226)	(0.188)
Share Using Fertilizer$_{t-1}$							-2.602^{***}
							(0.146)
Observations	8,983	8,983	8,983	8,983	8,983	8,971	8,971
R^2	0.002	0.003	0.069	0.102	0.107	0.113	0.158
	Panel B: Wheat Adaptation						
Soil Heterogeneity	-0.616^{**}	-0.588^{**}	-0.734^{***}	-0.683^{***}	-0.607^{***}	-0.499^{***}	-0.386^{**}
	(0.268)	(0.271)	(0.238)	(0.240)	(0.182)	(0.177)	(0.154)
Wheat Share$_{t-1}$							-2.704^{***}
							(0.184)
Observations	19,531	19,531	19,531	19,531	19,531	19,502	19,502
R^2	0.001	0.121	0.183	0.203	0.213	0.217	0.244
Year Fixed Effects		✓	✓	✓	✓	✓	✓
Geoclimatic Controls			✓	✓	✓	✓	✓
Smooth Location Controls				✓	✓	✓	✓
Agricultural Suitability Controls					✓	✓	✓
Higher Order Controls						✓	✓

Note: This table reports estimates of equation 1 when the dependent variables are different indicators for farmers' social learning. Panel A focuses on the growth rate of the share of farms using fertilizers, and Panel B on the growth rate of wheat production. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. Agricultural suitability controls include average agricultural suitability indices for alfalfa, cotton, maize, oat, rye, sugarcane, sweet potato, tobacco, wheat, and white potato. Higher order controls include the standard deviations of temperature, precipitation, slope, elevation, flow accumulation, and the agricultural suitability indices. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE D.18: Social Learning: Robustness to Alternative Distances for SHI Calculation

SHI Cell Distance	Dependent variable:						
	Baseline 25	10	15	20	30	35	40
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: Fertilizer Adoption						
Soil Heterogeneity	-0.771^{***}	-0.590^{**}	-0.661^{***}	-0.721^{***}	-0.814^{***}	-0.852^{***}	-0.884^{***}
	(0.223)	(0.230)	(0.224)	(0.222)	(0.224)	(0.226)	(0.230)
Observations	8,983	8,983	8,983	8,983	8,983	8,983	8,983
R^2	0.294	0.293	0.294	0.294	0.294	0.294	0.294
	Panel B: Wheat Adaptation						
Soil Heterogeneity	-0.600^{***}	-0.483^{**}	-0.530^{**}	-0.570^{***}	-0.623^{***}	-0.639^{***}	-0.650^{***}
	(0.218)	(0.218)	(0.214)	(0.215)	(0.222)	(0.227)	(0.231)
Observations	19,531	19,531	19,531	19,531	19,531	19,531	19,531
R^2	0.393	0.392	0.392	0.393	0.393	0.393	0.393
State × Year Fixed Effects	✓	✓	✓	✓	✓	✓	✓
Geoclimatic Controls	✓	✓	✓	✓	✓	✓	✓
Smooth Location Controls	✓	✓	✓	✓	✓	✓	✓

Note: This table reports estimates of equation 1 when the dependent variables are different indicators for farmers' social learning. Panel A focuses on the growth rate of the share of farms using fertilizers, and Panel B on the growth rate of wheat production. In each column the SHI is calculated over areas of different sizes, denoted in terms of 500-meter cell distances to each direction. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. See Appendix E for details on the data and variable construction. Standard errors clustered at arbitrary grid cells of 100 miles square in parentheses (Bester et al., 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

D.6 Robustness of Selective Out-Migration Results

TABLE D.19: Additional Controls for Out-Migration Results (I): High Communal Identification

	Dependent variable: Remaining in the Same County (mean = 0.45, SD = 0.49)					
	(1)	(2)	(3)	(4)	(5)	(6)
Soil Heterogeneity	0.212^{***}	0.087^{**}	0.075^{*}	0.067	0.072	0.071
	(0.061)	(0.043)	(0.045)	(0.045)	(0.047)	(0.046)
Farmer	0.232^{***}	0.210^{***}	0.208^{***}	0.204^{***}	0.198^{***}	0.191^{***}
	(0.049)	(0.036)	(0.035)	(0.036)	(0.036)	(0.036)
Soil Heterogeneity × Farmer	-0.261^{***}	-0.178^{***}	-0.174^{***}	-0.168^{***}	-0.159^{***}	-0.159^{***}
	(0.074)	(0.055)	(0.054)	(0.054)	(0.055)	(0.055)
Observations	36,433	36,433	36,433	36,433	36,218	36,218
R^2	0.005	0.034	0.035	0.036	0.037	0.052
State × Year Fixed Effects		✓	✓	✓	✓	✓
Geoclimatic Controls			✓	✓	✓	✓
Smooth Location Controls				✓	✓	✓
Agricultural Suitability Controls				✓	✓	✓
Higher Order Controls					✓	✓
Individual Controls						✓

Note: This table reports estimates of equation 6 when the dependent variable is the probability of remaining in the same county in the following decade. The sample includes linked migrants with a high degree of prior communal identification. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. Agricultural suitability controls include average agricultural suitability indices for alfalfa, cotton, maize, oat, rye, sugarcane, sweet potato, tobacco, wheat, and white potato. Higher order controls include the standard deviations of temperature, precipitation, slope, elevation, flow accumulation, and the agricultural suitability indices. Individual controls include age and state-of-birth fixed effects. Standard errors clustered by original destination county in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE D.20: Additional Controls for Out-Migration Results (II): Intra-Community Marriage

	Dependent variable: Remaining in the Same County (mean = 0.46, SD = 0.49)					
	(1)	(2)	(3)	(4)	(5)	(6)
Soil Heterogeneity	0.129**	0.058	0.054	0.055	0.053	0.050
	(0.057)	(0.044)	(0.045)	(0.045)	(0.047)	(0.048)
Farmer	0.224***	0.183***	0.175***	0.173***	0.167***	0.158***
	(0.045)	(0.034)	(0.033)	(0.033)	(0.033)	(0.033)
Soil Heterogeneity × Farmer	−0.235***	−0.132**	−0.120**	−0.117**	−0.107**	−0.106**
	(0.069)	(0.052)	(0.050)	(0.050)	(0.051)	(0.051)
Observations	42,652	42,652	42,652	42,652	42,399	42,399
R^2	0.006	0.031	0.034	0.034	0.035	0.049
State × Year Fixed Effects		✓	✓	✓	✓	✓
Geoclimatic Controls			✓	✓	✓	✓
Smooth Location Controls				✓	✓	✓
Agricultural Suitability Controls				✓	✓	✓
Higher Order Controls					✓	✓
Individual Controls						✓

Note: This table reports estimates of equation 6 when the dependent variable is the probability of remaining in the same county in the following decade. The sample includes linked migrants whose wife was born in the same place as them. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. Agricultural suitability controls include average agricultural suitability indices for alfalfa, cotton, maize, oat, rye, sugarcane, sweet potato, tobacco, wheat, and white potato. Higher order controls include the standard deviations of temperature, precipitation, slope, elevation, flow accumulation, and the agricultural suitability indices. Individual controls include age and state-of-birth fixed effects. Standard errors clustered by original destination county in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE D.21: Additional Controls for Out-Migration Results (III): Interaction

	Dependent variable: Remaining in the Same County (mean = 0.45, SD = 0.490)					
	(1)	(2)	(3)	(4)	(5)	(6)
Soil Heterogeneity	0.082	−0.022	−0.032	−0.030	−0.039	−0.041
	(0.060)	(0.048)	(0.049)	(0.049)	(0.051)	(0.051)
Farmer	0.161***	0.124***	0.114***	0.114***	0.104***	0.096***
	(0.043)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)
Soil Heterogeneity × Farmer	−0.131**	−0.038	−0.026	−0.026	−0.009	−0.009
	(0.066)	(0.054)	(0.053)	(0.053)	(0.054)	(0.053)
Soil Heterogeneity × Farmer × High Communal Identification	−0.131*	−0.121*	−0.125*	−0.122*	−0.126*	−0.128*
	(0.071)	(0.068)	(0.068)	(0.068)	(0.070)	(0.069)
Observations	74,922	74,922	74,922	74,922	74,477	74,477
R^2	0.005	0.032	0.033	0.034	0.035	0.047
State × Year Fixed Effects		✓	✓	✓	✓	✓
Geoclimatic Controls			✓	✓	✓	✓
Smooth Location Controls				✓	✓	✓
Agricultural Suitability Controls				✓	✓	✓
Higher Order Controls					✓	✓
Individual Controls						✓

Note: This table reports estimates of equation 6 when the dependent variable is the probability of remaining in the same county in the following decade. The sample in includes all the linked migrants. Geoclimatic controls include average temperature, precipitation, slope, elevation, flow accumulation, absolute agricultural productivity, river density, total area, and distance to steamboat-navigated rivers, lakes, and the shore. Smooth location controls is a second-order polynomial in latitude and longitude. Agricultural suitability controls include average agricultural suitability indices for alfalfa, cotton, maize, oat, rye, sugarcane, sweet potato, tobacco, wheat, and white potato. Higher order controls include the standard deviations of temperature, precipitation, slope, elevation, flow accumulation, and the agricultural suitability indices. Individual controls include age and state-of-birth fixed effects. All columns also include the main effect of High Communal Identity and all the two-term interactions. Standard errors clustered by original destination county in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

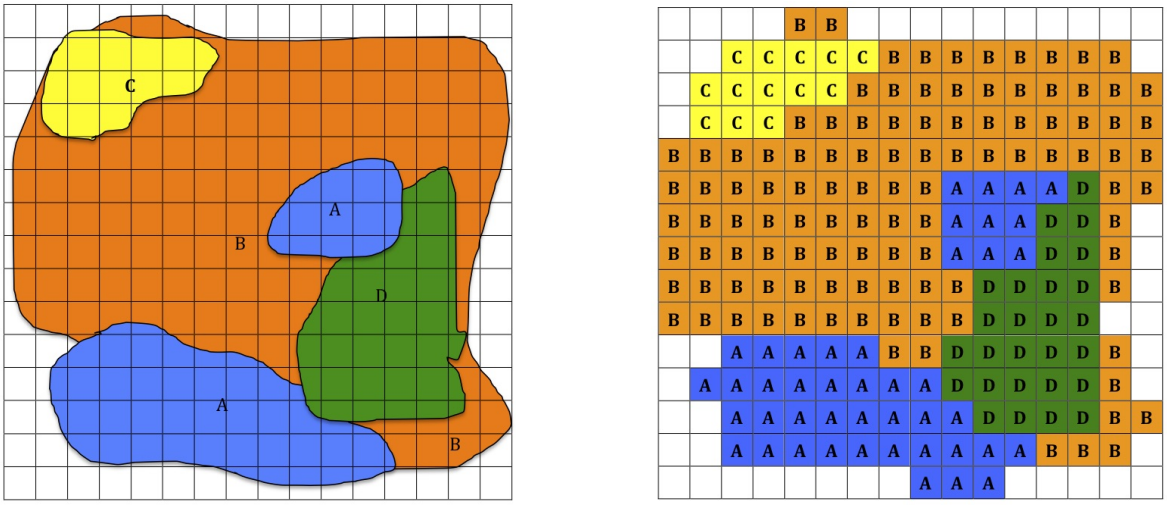
E Data Sources and Variables Construction

E.1 Soil Heterogeneity Index

I use detailed geo-referenced soil data from the Digital General Soil Map of the United States (STATSGO2) (Soil Survey Staff, 2017b), which maps soil areas that can be shown at a map of scale 1:250,000 in the continental U.S., to construct a county-level *"Soil Heterogeneity Index"* (SHI). The index is meant to capture the average dissimilarity of soil across neighboring farmers. I construct the SHI in the following steps:

1. I convert the STATSGO2 map containing soil map unit polygon features into a raster dataset containing fine-grid cells of size 500 meters square (illustrated in Figure E.1).
2. For each cell, I calculate the probability that a randomly selected neighboring cell is of a different soil map unit (illustrated in Figures E.2-E.3).
   I define neighboring cells as cells that fall within a square of a given size around each cell ("the considered area") rather than in terms of actual distance (i.e. fall within a given circle around each cell) to reduce computational demand. The size of the considered area affects the degree of heterogeneity. The bigger the area, the more likely it is to have cells of different soil types. My baseline SHI uses half of the mean size of U.S. counties in 2000 as a benchmark for the size of the considered area, which amounts to 25 cells in each direction (north, south, east, and west). I document robustness to using different sizes of the considered area.
3. Last, I aggregate the SHI at the fine grid level to the county level by taking the mean grid-level SHI within the county. The SHI is calculated for all contemporary U.S. counties in each decade using GIS and information on counties' contemporary borders from Manson et al., 2020.

FIGURE E.1: SHI CONSTRUCTION. STATSGO2 DATA TO GRID-CELLS

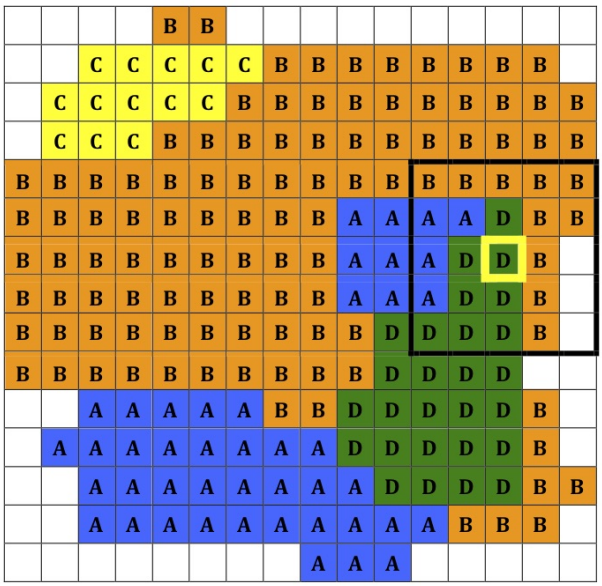


(A) Underlying STATSGO2 Data

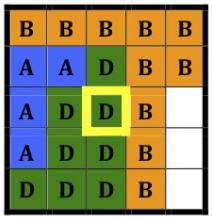
(B) Grid-cells data

Note: This figure illustrates step (1.) of the county-level SHI construction - converting the STATSGO2 map containing polygon features (Figure A) into a raster dataset containing fine-grid cells of size 500 meters square (Figure B).

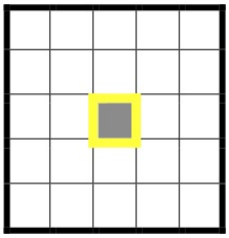
FIGURE E.2: SHI CONSTRUCTION. CALCULATING THE SHI FOR EACH CELL



(A) Neighboring cells (in black) around a given cell (in yellow)



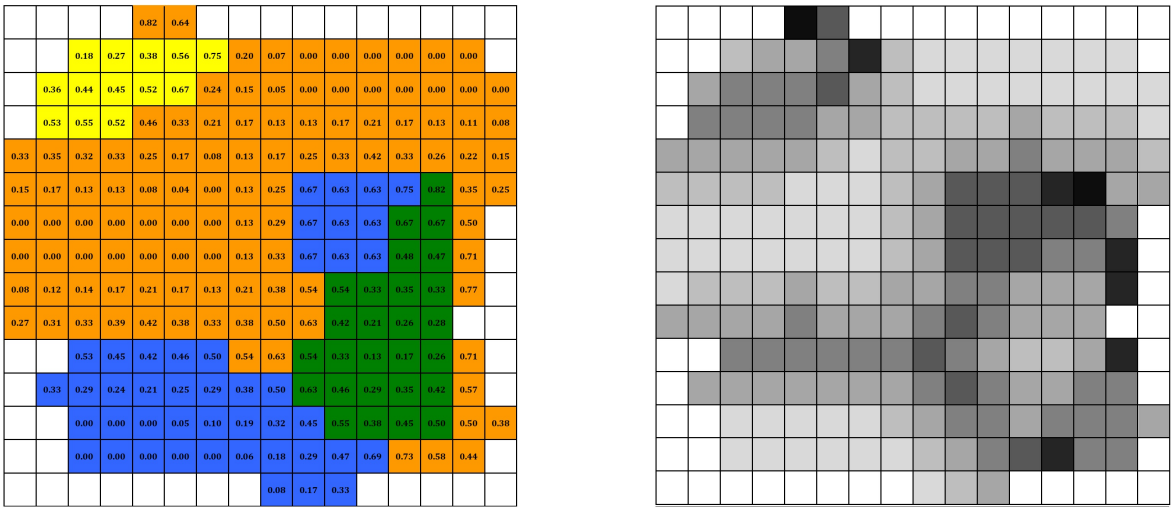
(B) 7 same-type cells out of 21 neighboring soil cells



(C) SHI for given cell is $1 - 7/24 = 0.67$

Note: This figure illustrates step (2.) of the county-level SHI construction - calculating the probability that a randomly selected neighboring cell (in bold-black frame) is of a different soil type than a given cell (in bold-yellow frame).

FIGURE E.3: SHI CONSTRUCTION. SHI FOR EACH CELL



(A) Numeric representation

(B) Heat map representation

Note: This figure illustrates the result of applying step (2.) of the county-level SHI construction to each cell. Figure (A) represents the cell-level SHI numerically, and Figure (B) represents it as a heat map.

E.2 Historical Measure of Close-Knit Communities

Local Name Index. The LNI is defined as

$$LNI_{nlgt} = 100 \times \frac{Pr(n|l,g,t)}{Pr(n|l,g,t) + Pr(n|-l,g,t)}$$

where n is a particular first name, l is the geographical level defined as "local" - the contemporaneous county or state, $-l$ is all the non-local locations, g is the child's gender, and t is the census year. The index ranges from 0 to 100, where a value of 100 reflects a distinctively local name and a value of zero reflects a distinctively "outsider's" name. Note that the LNI is invariant to the size of the population in different localities and to the general popularity of a given name.

I use restricted data on children's first names from the full count censuses between 1850-1940 (Ruggles et al., 2020). 1850 is the first year in which first names were recorded for children. 1940 is the last year for which full count data is available. Data is unavailable for 1890. Individuals living in group quarters are excluded. The sample is restricted to only include native-born children between the ages of 0 to 10. The baseline sample is further restricted to only include white children with native-born parents. Alternative samples add to the baseline sample (i) children with foreign-born parents (ii) non-white children, and (iii) both.

Intra-Community Marriage. An indicator that equals one if a married couple has a common birthplace, and zero otherwise. For native-born individuals, the state of birth is used and for foreign-born individuals, the country of birth is used. The county-level ICM is the share of ICM couples out of the total couples.

I use the full count censuses between 1850-1940 (Ruggles et al., 2020). Data is unavailable for 1890. Individuals living in group quarters are excluded, as well as married individuals for which data on the spouse is missing. The baseline sample is further restricted to only include white native-born couples. Alternative samples add to the baseline sample (i) foreign-born (ii) non-whites and (iii) both.

Tight Norms Index. For each census year between 1850-1940, I first use full-count census data (Ruggles et al., 2020) to calculate the county-level coefficient of variance of three key non-binary family-related choices that are observable in historical censuses: the mother's age at first birth, the total number of children, and the number of distinct families residing in the same house. For each of those variables, a lower coefficient of variance implies tighter norms. I limit the sample to households with married mothers who were between the ages of 35-44 when the census was taken and are the heads of the household or the wives of the head of the household, to avoid capturing variation that

originated from other factors, such as a different age composition across counties. Individuals living in group quarters are excluded. The baseline sample is further restricted to only include white native-born mothers. Alternative samples add to the baseline sample (i) foreign-born (ii) non-whites and (iii) both. Finally, for each year, I conduct a principal component analysis at the county level using these variables as inputs to construct a composite measure. The first eigenvector, which I refer to as the *Tight Norms Index* (TNI), explains between $42-56\%$ of the variance in the three variables, depending on the year and the sample. It is also the only component with an eigenvalue that is larger than one in all years and samples ($1.26-1.73$).[9] In all years and samples the loading on the three variables always has the same sign. Because there is no natural interpretation for the TNI units, I standardize it into z-scores within each year to ease the interpretation of estimated effects.

Religious Homogeneity Index. The RHI is defined as

$$Religious\ Homogeneity\ Index_{ct} = \sum_j s_{cjt}^2$$

where s_{cjt} is the share of members of religious denomination j, in county c, and in year t, out of the total number of members in religious institutions in county c year t. Note that the RHI equals the Herfindahl–Hirschman Index over the share of members of religious denominations. It measures the probability that two randomly drawn individuals from the population of members of religious institutions in a county belong to a different denomination.

The list of religious denominations for which data is collected varies across years. Therefore, and to ease the interpretation of estimated effects, in the empirical analysis I standardize the RHI into z-scores within each year. To calculate the index, I use county-level data on the number of members of religious institutions by denomination between the years 1850, 1860, 1870, 1890, 1906, 1916, 1926, and 1936 (Manson et al., 2020).

Median Kin Propinquity. I rely on data and methods from (Nelson, 2020) to measure kin propinquity. Using data on surnames and the distance between households on the enumeration form the probability that the surname match was not random and reflects kinship as is defined as

$$P(K_{rie}) = \left(1 - \frac{N_{rie}-1}{N_{re}-1}\right)^{D_i}$$

where N_{rie} is the number of the same-race (r) same-surname (i) households in the same enumera-

[9] In 1870, the second component has an eigenvalue of 1.02 and 1.00 in the sample that includes immigrants and non-whites, and in the sample that includes immigrants, respectively. In all other years and samples, all other components are strongly smaller than 1.

tion district (e), N_{re} is the total number of same-race households in the enumeration district, and D_i is the number of households with a different surname that are as close as the nearest same-surname household.

Using data shared by Nelson (2020) and the full count censuses between 1850-1940 (Ruggles et al., 2020), I calculate the county-level median distance between same surname households, weighted by the probability of a non-random match. I define *Median Kinship Propinquity* (MKP) as one over the median distance. Individuals living in group quarters are excluded. The baseline sample is further restricted to only include white native-born individuals. Alternative samples add to the baseline sample (i) foreign-born (ii) non-whites and (iii) both.

Kin Propinquity Rate. I follow Nelson (2020) and use the kin propinquity data and the full count censuses between 1850-1940 (Ruggles et al., 2020) to calculate a county-level *Kinship Propinquity Rate* (KPR), defined as the share of the population with kin residing in the same enumeration distinct, weighted by the probability of a non-random match. Individuals living in group quarters are excluded. The baseline sample is further restricted to only include white native-born individuals. Alternative samples add to the baseline sample (i) foreign-born (ii) non-whites and (iii) both.

The Strength of Family Ties Index. I use data from the full count censuses between 1860-1940 (Ruggles et al., 2020) to construct a county-level *"Strength of Family Ties Index"* (SFTI).[10] I focus on four key variables related to family structure and the choice of living arrangements that are observable in historical censuses and are associated with strong versus weak family ties. Individuals living in group quarters are excluded. The baseline sample is further restricted to only include white native-born households. Alternative samples add to the baseline sample (i) foreign-born (ii) non-whites and (iii) both. For each county-year, I calculate (i) the divorce-to-marriage ratio, (ii) the share of elderly people living without a relative, (iii) the share of people living with at least one person who is not their relative, and (iv) the mean size of families. Then, for each year, I conduct a principal component analysis at the county level using these variables as inputs. The first eigenvector, which I refer to as the SFTI, explains between $45-71\%$ of the variance in the four variables, depending on the year and sample. It is also the only component with an eigenvalue that is larger than one in all years and samples $(1.79-2.82)$.[11] In all years and samples, the loading on the four variables always has the same sign (negative on divorce-to-marriage ratio, the share of elderly people living without a relative, and the share of people living with a non-relative, and positive on family size). Because there is no

[10] 1850 is excluded because in this year information regarding material status was not recorded.

[11] The second component had an eigenvalue of 1.00 in 1860 in the sample that includes non-whites and in 1870 in the baseline sample. In all other years and samples, all other components are strongly smaller than 1.

natural interpretation for the SFTI units, I standardize it into z-scores within each year to ease the interpretation of estimated effects.

Children's Communal Ties. I use the Census Linking Project (Abramitzky et al., 2022b,d,f,g) to link the male children of families that migrated within the U.S. (identified using children's state of birth) to their record as adults. I link children observed in the 1850, 1870, and 1880 censuses with their records 30 years later, and children observed in the 1860 census with their record 40 years later.

Remained in the Destination County. Due to counties border changes between periods, remaining in the same county is defined as the share of the area of the county of destination (i.e., the migrant's county of residence observed in the earlier period) that overlaps with the county of residence of the migrant's child as an adult, calculated using GIS and information on counties' contemporary borders from Manson et al., 2020.

Intra-Community Marriage. An indicator that equals one if the migrant's child is married to a spouse born in the destination state, zero if he is married to a spouse born in a different state, and NA if unmarried or there is no data on his wife's state of birth.

E.3 Other Variables

Agricultural Heterogeneity

Agricultural Diversity. A county-level index defined as

$$Agricultural\ Diversity_{ct} = 1 - \sum_{j} s_{cjt}^2$$

where s_{cjt} is the share of acres used in the cultivation of agricultural product j, in county c, and in year t, out of the total number of acres under cultivation. Note that the index equals one minus the Herfindahl–Hirschman Index over the share of acres used in the cultivation of different agricultural products. The agricultural diversity index measures the probability that two randomly drawn acres used in farms in a county are used to grow different agricultural products. The list of agricultural products for which data is collected varies across years. Therefore, and to ease the interpretation of estimated effects, in the empirical analysis I standardize the index into z-scores within each year.

To calculate the index, I use county-level data on the number of acres used in the production of different agricultural products for each decade between the years 1880-1930, as well as 1925 and 1935 (Manson et al., 2020).

Variation of Agriculture Suitability. The county-level mean standardized standard deviation of agricultural suitability indices of the 10 most important crops according to their total output value in 1859 (Manson et al., 2020): alfalfa, cotton, maize, oat, rye, sugarcane, sweet potato, tobacco, wheat, and white potato. Calculated for all contemporary U.S. counties in each decade using GIS and information on counties' contemporary borders from Manson et al., 2020 and suitability index rasters, under intermediate input level and rain-fed management, baseline period, 1961-1990 from FAO and IIASA (2012).

Geoclimatic controls.

The following county-level variables are calculated for all contemporary U.S. counties in each decade using GIS and information on counties' contemporary borders from Manson et al., 2020 and geocliamtic information:

Average temperature. County-level mean annual temperature (in Celsius degrees), calculated for all contemporary U.S. counties in each decade using GIS and information on counties' contemporary borders from Manson et al., 2020 and a 5 arc-minutes mean annual temperature raster (baseline period, 1961-1990) from FAO and IIASA (2012).

Average precipitation. County-level mean annual precipitation (in mm), calculated for all contemporary U.S. counties in each decade using GIS and information on counties' contemporary borders from Manson et al., 2020 and a 5 arc-minutes mean annual precipitation raster (baseline period, 1961-1990) from FAO and IIASA (2012).

Average elevation. County-level mean elevation (in meters), calculated for all contemporary U.S. counties in each decade using GIS and information on counties' contemporary borders from Manson et al., 2020 and a 30 arc-second elevation raster (voild-filled DEM) from Lehner et al. (2008).

Average slope. County-level mean slope (in radians), calculated for all contemporary U.S. counties in each decade using GIS and information on counties' contemporary borders from Manson et al., 2020 and a 30 arc-second elevation raster (voild-filled DEM) from Lehner et al. (2008).

Average absolute agricultural productivity. County-level average maximal potential dollar value of agricultural products in 1860 prices for the following crops: alfalfa, barley, buckwheat, cotton, flax, maize, oat, rye, sugarcane, sweet potato, tobacco, rice, wheat, and white potato. Calculated for all contemporary U.S. counties in each decade using GIS and information on counties' contemporary borders from Manson et al., 2020, potential production capacity rasters (in t/ha), under intermediate input level and rain-fed management, baseline period, 1961-1990 from FAO and IIASA (2012), and

data on the price of farm products 1859/1860 from Manson et al., 2020.

Flow accumulation. County-level mean flow accumulation, defined as the amount of upstream area (in number of cells) draining into each cell and measuring the measure of the upstream catchment area, calculated for all contemporary U.S. counties in each decade using GIS and information on counties' contemporary borders from Manson et al., 2020 and a 30 arc-second flow accumulation raster (voild-filled ACC) from Lehner et al. (2008).

River density. County-level river density, defined as the share of county area in rivers or streams assuming a fixed 10-meter width for all rivers and streams, calculated for all contemporary U.S. counties in each decade using GIS and information on counties' contemporary borders from Manson et al., 2020 and a 30 arc-second river network raster (voild-filled RIV) from Lehner et al. (2008).

Land area. The total area in squared meters of all contemporary U.S. counties in each decade (Manson et al., 2020).

Distance to steamboat-navigated rivers. Natural logarithm of the minimal distance in meters between counties' centroid and steamboat-navigated rivers, calculated for all contemporary U.S. counties in each decade using GIS and information on counties' contemporary borders from Manson et al., 2020 and steamboat-navigated rivers from Atack (2015a).

Distance to lakes. Natural logarithm of the minimal distance in meters between counties' centroid and lakes, calculated for all contemporary U.S. counties in each decade using GIS and information on counties' contemporary borders from Manson et al., 2020 and U.S. lakes from Commission for Environmental Cooperation (CEC).

Distance to shore. Natural logarithm of the minimal distance in meters between counties' centroid and the shoreline, calculated for all contemporary U.S. counties in each decade using GIS and information on counties' contemporary borders from Manson et al., 2020 and U.S. shoreline from National Oceanic and Atmospheric Administration (NOAA).

Agricultural suitability.

Agricultural suitability indices. County-level mean agricultural suitability indices of the 10 most important crops according to their total output value in 1859 (Manson et al., 2020): alfalfa, cotton, maize, oat, rye, sugarcane, sweet potato, tobacco, wheat, and white potato. Calculated for all contemporary U.S. counties in each decade using GIS and information on counties' contemporary borders from Manson et al., 2020 and suitability index rasters, under intermediate input level and rain-fed management, baseline period, 1961-1990 from FAO and IIASA (2012).

Fertilizers.

Share using fertilizers. County-level data on the share of farms reporting expenditure on fertilizers out of the total number of farms for the years 1910-1930 (Manson et al., 2020). Data is harmonized holding 1930 county borders fixed (Hornbeck, 2010).

Growth of fertilizers use. Inverse hyperbolic sine (IHS) transformation of the 10-years growth rate of the share of farms using fertilizers.

Wheat production and share.

Wheat production. County-level data total wheat production (in bushel) for the years 1840-1935 (Manson et al., 2020). Data is harmonized holding 1930 county borders fixed (Hornbeck, 2010).

Growth of wheat production. Inverse hyperbolic sine (IHS) transformation of the 10-year growth rate of wheat production.

Wheat share. County-level data total share of land area (in acres) dedicated to cultivating wheat out of the total land area (in acres) dedicated to cultivating crops, for the years 1880-1935 (Manson et al., 2020). Data is harmonized holding 1930 county borders fixed (Hornbeck, 2010).

Growth of wheat share. Inverse hyperbolic sine (IHS) transformation of the 10-year growth rate of wheat share.

Learning Potential.

Learning Potential, fertilizer. County-level average standardized mean agro-climatic potential crop yield difference between low and high inputs for the five (or three) main crops in 1930 in terms of acreage (Manson et al., 2020): maize, wheat, cotton, oat, and various fodder crops (grass). Calculated for all contemporary U.S. counties in each decade using GIS and information on counties' contemporary borders from Manson et al., 2020 and agro-climatic potential yield rasters for the five (or three) crops, for the period 1961-1990, with an available water content of 200 mm/m (under irrigation conditions) from FAO and IIASA (2020).

Learning Potential, wheat. County-level agro-climatic potential crop yield difference between low and high inputs. Calculated for all contemporary U.S. counties in each decade using GIS and information on counties' contemporary borders from Manson et al., 2020 and agro-climatic potential yield rasters for wheat, for the period 1961-1990, with an available water content of 200 mm/m (under irrigation conditions) from FAO and IIASA (2020).

Confounding channels.

Climatic Risk. The county-level mean standardized standard deviation of annual precipitation and annual temperature. Calculated using data from GAEZ-FAO (FAO and IIASA, 2020) for the baseline period (1960-1990). Calculated for all contemporary U.S. counties in each decade using GIS and information on counties' contemporary borders from Manson et al., 2020.

Railroads. A county-level indicator for the existence of a railroad in the county, as well as within 10, 20, 30, 40, and 50 miles, for 1850-1910. Constructed using railroad shapefiles from Atack (2015b). Calculated for all contemporary U.S. counties in each decade using GIS and information on counties' contemporary borders from Manson et al., 2020.

Share Immigrants. A county-level share of household heads that were born outside the U.S. for the years 1850-1940, calculated using the full count censuses (Ruggles et al., 2020).

Birth Place Diversity. A county-level index that equals one minus the Herfindahl–Hirschman Index over the share of household heads that were born in different countries (or regions) for the years 1850-1940, calculated using the full count censuses (Ruggles et al., 2020).

Dominate Origin. Fixed effects for the the most common birthplace among household heads that that were born outside the U.S. for the years 1850-1940, calculated using the full count censuses (Ruggles et al., 2020).

Black Share. The county-level black population share, calculated for each decade 1850-1940 using data from Manson et al., 2020.

Historical Slave Share. The county-level slave-to-population ratio in 1850 and 1860, calculated using data from Manson et al., 2020. Data is harmonized holding 1940 county borders fixed (Hornbeck, 2010).

Farms' size Gini. The Gini coefficient on the county-level distribution of farm sizes for the years 1860-1940 (Manson et al., 2020). The coefficient is calculated using the midpoint of each category and 125% of the last (unbounded) category.

Share farmers. The share of households in a county in the relevant sample with an occupation code (1950 basis) of 100- "Farmers (owners and tenants)," or 123 - "Farm managers," out of the total number of households in the sample in the county with a valid occupation code. Data from the full count censuses between (Ruggles et al., 2020).

Literacy share. The share of household heads in a county in the relevant sample that are literate out of the total number of household heads in the sample in the county with a valid literacy code. Data

from the full count censuses between 1850-1930 (Ruggles et al., 2020). Data does not exist for 1940.

Share Urban. The share of the population in the county residing in urban locations, defined as locations with a population greater than 2,500 (Manson et al., 2020).

Manufacturing establishments. The number of manufacturing establishments in a county per 100 people (Manson et al., 2020). Data does not exist for 1850 and 1910.

Frontier Status. A county-level indicator for frontier status for 1850-1880, using the baseline frontier indicator from Bazzi et al. (2020) (variable name: "frontier100kmL6").

Total Frontier Experience. A county-level measure for 1940 counties of the amount of time in decades spent in a frontier status. Data from Bazzi et al. (2020) (variable name: "tye_tfe890_500kNI_100_l6").

Long-Run Close-Knit Communities

Clustering. County-level measures of the average fraction of an individual's friend pairs who are also friends with each other, calculated by Chetty et al. (2022a,b) using active U.S. Facebook users aged 25-44.

Support Ratio. County-level measures of the proportion of within-county friendships where the pair of friends share a third mutual friend within the same county, calculated by Chetty et al. (2022a,b) using active U.S. Facebook users aged 25-44.

Excess Support for Trump in 2016. A county-level measure of the difference between Trump's vote share in the 2016 presidential election and the average vote share of Romney in 2012 and McCain in 2008, standardized into z-scores. Data from Leip (2017).

Communal Moral Values. The first eigenvector from a principal component analysis on the five moral foundations *"Moral Foundations Theory"* (MFT) (Haidt and Graham, 2007): Harm / Care and Fairness / Reciprocity, In-group / Loyalty, and Authority / Respect and Purity / Sanctity. Data surveyed using the *"Moral Foundations Questionnaire"* (MFQ) (Graham et al., 2011) on `www.yourmorals.org` between 2008-2018, and includes the individual responses of approximately 242,000 Americans. For more information on the MFT and its measure, including the full MFQ texts, see `https://moralfoundations.org/`. Moral foundations scores range from 0 to 30, then standardized into z-scores.

Individuals are matched to counties using HUD USPS ZIP code to county crosswalk. Some zip code areas intersect more than one county. In those cases, I match the respondent to all possible counties (i.e. "duplicate" the respondent), and weight each observation by the ratio of residential addresses in

the ZIP-county to the total number of residential addresses in the entire ZIP code area (Wilson and Din, 2018), so that each respondent receives a total weight of one. Counties with fewer than five observatories are dropped.

The first eigenvector explains 46% of the variance in the five foundations and has an eigenvalue of 2.29. It is the only eigenvector for which the signs of the loadings on the five foundations correspond to the "communal" versus "individualizing" distinction (loads negatively on Harm / Care and Fairness / Reciprocity, and positively on In-group / Loyalty, Authority / Respect and Purity / Sanctity).

Relative importance of communal values. One minus the county-level variable named "universalist_vs_communal_values" which aggregates individual-level from the MFQ surveyed on `www.yourmorals.org` to the county-level to measure the relative local importance of universalist vs. communal moral values. Data from Enke (2020).

References

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, Myera Rashid, and Santiago Pérez, *Census Linking Project: 1850-1860 Crosswalk [dataset]*, Harvard Dataverse, 2022. https://doi.org/10.7910/DVN/KO5J44, V2.

_ , _ , _ , _ , **and** _ , *Census Linking Project: 1850-1880 Crosswalk [dataset]*, Harvard Dataverse, 2022. https://doi.org/10.7910/DVN/ZEBJRO, V2.

_ , _ , _ , _ , **and** _ , *Census Linking Project: 1860-1870 Crosswalk [dataset]*, Harvard Dataverse, 2022. https://doi.org/10.7910/DVN/TXNANS, V2.

_ , _ , _ , _ , **and** _ , *Census Linking Project: 1860-1900 Crosswalk [dataset]*, Harvard Dataverse, 2022. https://doi.org/10.7910/DVN/ZUKOFB, V2.

_ , _ , _ , _ , **and** _ , *Census Linking Project: 1870-1880 Crosswalk [dataset]*, Harvard Dataverse, 2022. https://doi.org/10.7910/DVN/OCWCFR, V2.

_ , _ , _ , _ , **and** _ , *Census Linking Project: 1870-1900 Crosswalk [dataset]*, Harvard Dataverse, 2022. https://doi.org/10.7910/DVN/8WMSZ7, V2.

_ , _ , _ , _ , **and** _ , *Census Linking Project: 1880-1910 Crosswalk [dataset]*, Harvard Dataverse, 2022. https://doi.org/10.7910/DVN/ISGRMS, V2.

Atack, Jeremy, *Historical geographic information systems (gis) database of steamboat-navigated rivers during the nineteenth century in the united states [dataset]* 2015. https://my.vanderbilt.edu/jeremyatack/data-downloads.

_ , *Historical Geographic Information Systems (GIS) database of U.S. Railroads [dataset]* 2015. https://my.vanderbilt.edu/jeremyatack/data-downloads.

Bazzi, Samuel, Martin Fiszbein, and Mesay Gebresilasse, "Frontier culture: The roots and persistence of "Rugged Individualism" in the United States," *Econometrica*, 2020, *88* (6), 2329–2368.

Beck-Knudsen, Anne Sofie, "Those Who Stayed: Selection and Cultural Change in the Age of Mass Migration," *Unpublished Manuscript*, 2024.

Bester, Alan C., Timothy G. Conley, and Christian B. Hansen, "Inference with dependent data using cluster covariance estimators," *Journal of Econometrics*, 2011, *165* (2), 137–151.

Buggle, Johannes C and Ruben Durante, "Climate risk, cooperation and the co-evolution of culture and institutions," *The Economic Journal*, 2021, *131* (637), 1947–1987.

Chetty, Raj, Matthew O. Jackson, Theresa Kuchler, Johannes Stroebel, Nathaniel Hendren, Robert Fluegge, Sara Gong, Federico Gonzalez, Armelle Grondin, Matthew Jacob, Drew Johnston, Martin Koenen, Eduardo Laguna-Muggenberg, Florian Mudekereza, Tom Rutter, Nicolaj Thor, Wilbur Townsend, Ruby Zhang, Mike Bailey, Pablo Barbera, Monica Bhole, and Nils Wernerfelt, "Social Capital I: Measurement and Associations with Economic Mobility," *Nature*, 2022, *608* (7921), 108––121.

_ , and _ , "Social Capital II: Determinants of Economic Connectedness," *Nature*, 2022, *608* (7921), 122––134.

Commission for Environmental Cooperation (CEC), *North American Atlas - Hydrography [dataset]* 2010. Natural Resources Canada (NRCan), Instituto Nacional de Estadística y Geografía (INEGI), U.S. Geological Survey (USGS). Ed. 2.0, Vector digital data [1:10,000,000]. http://www.cec.org/north-american-environmental-atlas/lakes-and-rivers-2009/. Accessed [30/06/2022].

Conley, Timothy, "GMM estimation with cross sectional dependence," *Journal of Econometrics*, 1999, *92*, 1–45.

Cook, Lisa D, Trevon D Logan, and John M Parman, "Distinctively black names in the American past," *Explorations in Economic History*, 2014, *53*, 64–82.

Enke, Benjamin, "Moral values and voting," *Journal of Political Economy*, 2020, *128* (10), 3679–3729.

Eruchimovitch, Israel, Moti Michaeli, and Assaf Sarid, "On the coevolution of individualism and institutions," *Journal of Economic Growth*, 2023, pp. 1–42.

FAO and IIASA, "Global Agro-Ecological Zones version 3 (GAEZ v3.0) [dataset]," 2012. IIASA, Laxenburg, Austria and FAO, Rome, Italy.

_ , "Global Agro-Ecological Zones version 4 (GAEZ v4) [dataset]," 2020. http://www.fao.org/gaez/.

Fenske, James, "Ecology, trade, and states in pre-colonial Africa," *Journal of the European Economic Association*, 2014, *12* (3), 612–640.

Fiszbein, Martin, "Agricultural Diversity, Structural Change, and Long-Run Development: Evidence from the United States," *American Economic Journal: Macroeconomics*, 2022, *14* (2), 1–43.

Fryer, Roland and S. Levitt, "The causes and consequences of distinctively Black names," *The Quarterly Journal of Economics*, 2004, *119* (3), 767–805.

Giuliano, Paola and Nathan Nunn, "Understanding cultural persistence and change," *The Review of Economic Studies*, 2021, *88* (4), 1541–1581.

Gorodnichenko, Yuriy and Gerard Roland, "Which dimensions of culture matter for long-run growth?," *American Economic Review*, 2011, *101* (3), 492–98.

_ **and** _ , "Culture, institutions, and the wealth of nations," *Review of Economics and Statistics*, 2017, *99* (3), 402–416.

Graham, Jesse, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto, "Mapping the moral domain.," *Journal of Personality and Social Psychology*, 2011, *101* (2), 366.

Greenfield, Patricia M, "Linking social change and developmental change: shifting pathways of human development.," *Developmental Psychology*, 2009, *45* (2), 401.

Haidt, Jonathan, "Morality," *Perspectives on Psychological Science*, 2008, *3* (1), 65–72.

_ **and Jesse Graham**, "When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize," *Social Justice Research*, 2007, *20* (1), 98–116.

Harrington, Jesse R and Michele J Gelfand, "Tightness–looseness across the 50 united states," *Proceedings of the National Academy of Sciences*, 2014, *111* (22), 7990–7995.

Henrich, Joseph, *The WEIRDest people in the world: How the west became psychologically peculiar and particularly prosperous*, Farrar, Straus and Giroux, 2020.

Hornbeck, Richard, "Barbed wire: Property rights and agricultural development," *The Quarterly Journal of Economics*, 2010, *125* (2), 767–810.

Lehner, Bernhard, Kris Verdin, and Andy Jarvis, "New global hydrography derived from spaceborne elevation data," *EOS, Transactions, American Geophysical Union*, 2008, *89* (10), 93–94.

Leip, Dave, "David Leip's atlas of U.S. Presidential elections, datasets," 2017. Harvard Dataverse.

Michalopoulos, Stelios, "The origins of ethnolinguistic diversity," *American Economic Review*, 2012, *102* (4), 1508–39.

National Oceanic and Atmospheric Administration (NOAA), *Medium Resolution Shoreline [dataset]* 2000. Available online https://shoreline.noaa.gov/data/datasheets/medres.html. Accessed [17/02/2022].

Nelson, Matt A, "The decline of patrilineal kin propinquity in the United States, 1790–1940," *Demographic research*, 2020, *43*, 501.

Obolensky, Marguerite, Marco Tabellini, and Charles Taylor, "Homeward Bound: How Migrants Seek Out Familiar Climates," *NBER Working Paper*, 2024.

Oster, Emily, "Unobservable selection and coefficient stability: Theory and evidence," *Journal of Business & Economic Statistics*, 2019, *37* (2), 187–204.

Schulz, Jonathan F, Duman Bahrami-Rad, Jonathan P Beauchamp, and Joseph Henrich, "The Church, intensive kinship, and global psychological variation," *Science*, 2019, *366* (6466).

Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture, *U.S. General Soil Map (STATSGO2) [dataset]* 2017. https://data.nal.usda.gov/dataset/united-states-general-soil-map-statsgo2. Accessed [06/09/2017].

Steven Manson, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles, "IPUMS National Historical Geographic Information System: Version 15.0 [dataset]," 2020. Minneapolis, MN: IPUMS.

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek, "IPUMS USA: Version 10.0 [dataset]," 2020. Minneapolis, MN: IPUMS.

Triandis, Harry C, Christopher McCusker, and C Harry Hui, "Multimethod probes of individualism and collectivism.," *Journal of Personality and Social Psychology*, 1990, *59* (5), 1006.

Vandello, Joseph A and Dov Cohen, "Patterns of individualism and collectivism across the United States.," *Journal of Personality and Social Psychology*, 1999, *77* (2), 279.

Wilson, Ron and Alexander Din, "Understanding and enhancing the U.S. Department of Housing and Urban Development's ZIP code crosswalk files," *Cityscape: A Journal of Policy Development and Research*, 2018, *20* (2), 277–294.