Online Appendix for

The Residential Segregation of Immigrants in the United States from 1850 to 1940

Table A1. Correlation matrix between neighbor-based, dissimilarity and isolation measures

	Overall			Urban County			Rural County		
	Neighbor	Diss.	Iso.	Neighbor	Diss.	Iso.	Neighbor	Diss.	Iso.
Panel A: Pooled 1880-1940									
Neighbor	1			1			1		
Dissimilarity	0.691	1		0.741	1		0.347	1	
Isolation	0.699	0.689	1	0.768	0.753	1	0.329	0.460	1
Panel B: 1880 Census									
Neighbor	1			1			1		
Dissimilarity	0.388	1		0.419	1		0.359	1	
Isolation	0.425	0.508	1	0.591	0.572	1	0.343	0.493	1
Panel C: 1900 Census									
Neighbor	1			1			1		
Dissimilarity	0.700	1		0.790	1		0.373	1	
Isolation	0.685	0.694	1	0.797	0.778	1	0.301	0.490	1
Panel D: 1910 Census									
Neighbor	1			1			1		
Dissimilarity	0.776	1		0.834	1		0.346	1	
Isolation	0.738	0.773	1	0.799	0.823	1	0.291	0.493	1
Panel E: 1920 Census									
Neighbor	1			1			1		
Dissimilarity	0.795	1		0.823	1		0.424	1	
Isolation	0.722	0.766	1	0.769	0.816	1	0.241	0.455	1
Panel F: 1930 Census									
Neighbor	1			1			1		
Dissimilarity	0.795	1		0.755	1		0.357	1	
Isolation	0.722	0.766	1	0.751	0.782	1	0.392	0.428	1
Panel G: 1940 Census									
Neighbor	1			1			1		
Dissimilarity	0.642	1		0.654	1		0.426	1	
Isolation	0.627	0.641	1	0.665	0.695	1	0.396	0.403	1

Source: 1880 to 1940 full-count censuses (Ruggles et al. 2018)

Notes: The table shows the correlation between the neighbor-based segregation measure, the dissimilarity index, and the isolation index. The measure is at the country of birth/county/year level. The correlation matrix is weighted by the number of immigrant households.

Table A2. The relationship between immigrant households and segregation levels

	All Sources			Northern and Western Europe			Southern and Eastern Europe		
	All	Rural	Urban	All	Rural	Urban	All	Rural	Urban
Fraction immigrant households	3.421***	3.279***	3.583***	3.045***	3.088***	2.797***	5.990***	6.038***	6.364***
	(0.058)	(0.078)	(0.120)	(0.063)	(0.085)	(0.121)	(0.333)	(0.515)	(0.507)
County by Year FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
County by Country of birth FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Country of birth by Year FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	198,886	107,630	83,399	129,397	77,615	47,399	31,513	12,398	17,505
R-squared	0.556	0.580	0.615	0.535	0.577	0.573	0.726	0.746	0.747
Log Immigrant households	0.055***	0.051***	0.061***	0.039***	0.041***	0.035***	0.078***	0.078***	0.081***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.004)	(0.003)
County by Year FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
County by Country of birth FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Country of birth by Year FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	198,883	107,628	83,402	129,397	77,615	47,399	31,513	12,397	17,503
R-squared	0.565	0.583	0.627	0.532	0.575	0.572	0.741	0.756	0.763

Sources: 1850 to 1940 full-count census (Ruggles et al. 2018).

Notes: The table shows the results of a regression of segregation levels on the fraction immigrant households in the top panel. The bottom panel shows the results when using the log number of immigrant households as the independent variable. The data is collapsed to the county/year/country of birth level. Fixed effects are included as listed in the table. Counties are set to their 1900 borders according to the County Longitudinal Template (ICPSR 6576; Horan and Hargis 1995); setting the borders to 1850 boundaries does not influence point estimates. In text, we interpret that doubling the number of immigrants increases the segregation level by 0.038. This is because doubling the number of immigrants is equivalent to a log increase of about 0.693.

Table A3. Segregation from 2nd-generation by country of birth

	Year								
Country	1850	1860	1870	1880	1900	1910	1920	1930	1940
Canada	0.144	0.118	0.146	0.120	0.078	0.038	0.010	0.008	0.009
Mexico	0.455	0.325	0.314	0.309	0.269	0.357	0.441	0.414	0.264
Cuba		-0.114	0.129	0.053	0.249	0.289	0.177	0.189	0.136
Denmark	0.163	0.279	0.314	0.303	0.171	0.096	0.039	0.022	0.016
Finland				0.510	0.563	0.497	0.398	0.278	0.163
Norway	0.632	0.590	0.541	0.489	0.252	0.159	0.086	0.053	0.039
Sweden	0.337	0.350	0.419	0.402	0.267	0.171	0.090	0.052	0.035
England	0.112	0.089	0.094	0.048	0.015	-0.012	-0.022	-0.008	0.006
Scotland	0.130	0.112	0.108	0.059	0.018	-0.009	-0.025	0.000	0.006
Ireland	0.383	0.365	0.337	0.263	0.107	0.041	-0.003	0.000	0.020
Belgium	0.395	0.331	0.362	0.310	0.211	0.193	0.145	0.107	0.071
France	0.261	0.278	0.202	0.165	0.075	0.069	0.051	0.048	0.039
Netherlands	0.490	0.427	0.392	0.334	0.224	0.161	0.101	0.062	0.041
Switzerland	0.362	0.351	0.268	0.220	0.100	0.059	0.029	0.031	0.028
Greece				0.155	0.185	0.370	0.293	0.203	0.139
Italy	0.175	0.293	0.349	0.395	0.568	0.586	0.505	0.361	0.217
Portugal	0.092	0.333	0.402	0.350	0.369	0.312	0.398	0.318	0.202
Spain	0.127	0.104	0.124	0.062	0.105	0.328	0.320	0.304	0.208
Austria/Hungary	0.236	0.481	0.490	0.491	0.476	0.499	0.393	0.250	0.159
Germany	0.421	0.400	0.310	0.262	0.129	0.083	0.017	0.019	0.023
Poland/Russia	0.213	0.230	0.333	0.520	0.605	0.559	0.478	0.318	0.199
China		0.652	0.666	0.601	0.353	0.265	0.244	0.247	0.261
Japan					0.694	0.608	0.442	0.399	
Turkey					0.274	0.411	0.390	0.279	0.204

Sources: 1850 to 1940 full-count United States Censuses (Ruggles et al. 2018).
Notes: See Figure 3 for graphical depiction for 12 selected countries.

Table A4. Segregation from 3rd-generation by country of birth

	1880	1900	1910	1920	1930
Canada	0.161	0.109	0.067	0.048	0.045
Mexico	0.425	0.395	0.483	0.525	0.510
Cuba	-0.008	0.206	0.277	0.206	0.251
Denmark	0.337	0.228	0.164	0.098	0.076
Finland	0.638	0.628	0.567	0.497	0.425
Norway	0.569	0.427	0.333	0.234	0.168
Sweden	0.436	0.343	0.256	0.186	0.137
England	0.071	0.040	0.005	-0.005	0.006
Scotland	0.086	0.040	0.006	-0.013	0.010
Wales	0.314	0.186	0.112	0.066	0.003
Ireland	0.377	0.274	0.192	0.122	0.077
Belgium	0.392	0.372	0.309	0.228	0.174
France	0.271	0.159	0.127	0.118	0.103
Netherlands	0.434	0.397	0.334	0.252	0.187
Switzerland	0.323	0.236	0.166	0.102	0.085
Greece	0.171	0.261	0.393	0.319	0.247
Italy	0.443	0.637	0.654	0.590	0.496
Portugal	0.398	0.476	0.477	0.530	0.474
Spain	0.071	0.149	0.368	0.366	0.377
Austria/Hungary	0.579	0.622	0.627	0.517	0.408
Germany	0.387	0.319	0.254	0.148	0.119
Poland/Russia	0.584	0.702	0.646	0.605	0.497
China	0.612	0.403	0.355	0.348	0.396
Japan		0.702	0.619	0.473	0.449
Turkey		0.322	0.447	0.426	0.334

Sources: 1850 to 1940 full-count United States Censuses (Ruggles et al. 2018).

Notes: See Figure 3 for graphical depiction for 12 selected countries.

Table A5. Segregation of the 1st and 2nd generation from the 3rd-plus generation

Country	1880	1900	1910	1920	1930
Canada	0.134	0.072	0.034	0.019	0.018
Mexico	0.388	0.344	0.430	0.487	0.477
Cuba	-0.021	0.153	0.179	0.154	0.204
Denmark	0.316	0.211	0.145	0.079	0.051
Finland	0.609	0.624	0.561	0.482	0.384
Iceland	0.610	0.623	0.712	0.423	0.176
Norway	0.550	0.399	0.297	0.197	0.123
Sweden	0.416	0.325	0.232	0.158	0.098
England	0.032	-0.019	-0.049	-0.048	-0.035
Scotland	0.017	-0.038	-0.061	-0.062	-0.034
Wales	0.229	0.100	0.031	0.004	-0.026
Ireland	0.307	0.179	0.090	0.033	0.003
Belgium	0.372	0.326	0.254	0.158	0.111
France	0.197	0.079	0.036	0.025	0.022
Netherlands	0.353	0.322	0.242	0.174	0.130
Switzerland	0.284	0.191	0.114	0.058	0.039
Greece	0.190	0.224	0.359	0.312	0.241
Italy	0.394	0.616	0.635	0.569	0.466
Portugal	0.363	0.430	0.422	0.461	0.400
Spain	0.033	0.050	0.186	0.239	0.270
Austria/Hungary	0.555	0.597	0.594	0.477	0.367
Bulgaria			0.498	0.305	0.204
Germany	0.350	0.251	0.171	0.087	0.060
Romania		0.732	0.672	0.530	0.405
Poland/Russia	0.550	0.688	0.629	0.576	0.463
China	0.588	0.392	0.328	0.345	0.377
Japan		0.700	0.616	0.471	0.446
Korea			0.679	0.395	0.191

Sources: 1850 to 1940 full-count United States Censuses (Ruggles et al. 2018).

Notes: We drop cells with less than 200 households.

Table A6. Rural and Urban country segregation

		Year								
Country		1850	1860	1870	1880	1900	1910	1920	1930	1940
Canada	Rural	0.162	0.128	0.115	0.078	0.048	0.036	0.025	0.006	0.033
Canada	Urban	0.082	0.090	0.181	0.161	0.089	0.038	0.008	0.008	0.007
Mexico	Rural	0.501	0.360	0.291	0.298	0.254	0.324	0.366	0.319	0.188
Mexico	Urban	0.147	0.239	0.342	0.328	0.288	0.382	0.472	0.434	0.274
Denmark	Rural	0.181	0.327	0.345	0.332	0.206	0.137	0.084	0.058	0.033
Denmark	Urban	0.159	0.225	0.273	0.264	0.148	0.073	0.020	0.010	0.012
Finland	Rural				0.559	0.573	0.519	0.429	0.360	0.212
Finland	Urban				0.357	0.560	0.490	0.390	0.263	0.156
Norway	Rural	0.644	0.612	0.562	0.523	0.268	0.176	0.095	0.059	0.038
Norway	Urban	0.578	0.467	0.455	0.365	0.229	0.143	0.080	0.050	0.039
Sweden	Rural	0.467	0.413	0.426	0.434	0.299	0.211	0.126	0.073	0.042
Sweden	Urban	0.246	0.217	0.405	0.350	0.250	0.156	0.079	0.048	0.033
England	Rural	0.143	0.115	0.103	0.068	0.036	0.017	0.013	0.010	0.021
England	Urban	0.077	0.062	0.087	0.034	0.010	-0.017	-0.026	-0.009	0.005
Scotland	Rural	0.158	0.135	0.131	0.088	0.059	0.050	0.030	0.023	0.017
Scotland	Urban	0.096	0.087	0.090	0.039	0.006	-0.020	-0.032	-0.001	0.006
Wales	Rural	0.383	0.344	0.318	0.253	0.098	0.055	0.027	0.033	0.008
Wales	Urban	0.155	0.216	0.276	0.208	0.085	0.014	-0.010	-0.018	-0.009
Ireland	Rural	0.289	0.277	0.244	0.178	0.063	0.046	0.040	0.016	0.030
Ireland	Urban	0.449	0.418	0.376	0.291	0.113	0.040	-0.005	0.000	0.019
France	Rural	0.293	0.296	0.198	0.157	0.080	0.104	0.058	0.044	0.047
France	Urban	0.232	0.262	0.205	0.171	0.074	0.063	0.050	0.048	0.039
Netherlands	Rural	0.607	0.485	0.450	0.396	0.199	0.167	0.121	0.074	0.055
Netherlands	Urban	0.295	0.360	0.337	0.284	0.232	0.159	0.096	0.060	0.039
Switzerland	Rural	0.371	0.342	0.264	0.210	0.115	0.080	0.047	0.039	0.033
Switzerland	Urban	0.347	0.364	0.272	0.229	0.092	0.052	0.025	0.029	0.027
Italy	Rural	0.191	0.339	0.328	0.309	0.435	0.476	0.381	0.240	0.151
Italy	Urban	0.171	0.270	0.357	0.416	0.585	0.595	0.512	0.364	0.219
Portugal	Rural		0.431	0.353	0.328	0.298	0.114	0.341	0.236	0.161
Portugal	Urban	0.115	0.234	0.419	0.361	0.378	0.362	0.406	0.321	0.204
Austria/Hungary	Rural		0.499	0.477	0.463	0.362	0.340	0.281	0.174	0.137
Austria/Hungary	Urban	0.222	0.467	0.498	0.509	0.508	0.522	0.406	0.257	0.160
Germany	Rural	0.365	0.349	0.295	0.238	0.124	0.081	0.043	0.029	0.024
Germany	Urban	0.469	0.441	0.319	0.274	0.130	0.084	0.012	0.017	0.023
Poland/Russia	Rural	0.216	0.221	0.316	0.538	0.463	0.354	0.257	0.177	0.115
Poland/Russia	Urban	0.212	0.246	0.339	0.509	0.626	0.579	0.493	0.325	0.203
China	Rural		0.642	0.661	0.567	0.366	0.207	0.193	0.248	0.131
China	Urban		0.732	0.674	0.660	0.366	0.283	0.250	0.247	0.265
Japan	Rural					0.410	0.664	0.370	0.260	
Japan	Urban					0.298	0.542	0.451	0.407	

Sources: 1850 to 1940 full-count United States Censuses (Ruggles et al. 2018).
Notes: The table shows the highest segregation levels for cities and source countries that have over 1,000 households. We drop values if they have less than 4,000 households in total in an urban or rural area.

Table A7. Spatial Assimilation using Segregation Measure

	1910	1920	1930	Change over Decade	N
Raw County-Level Segregation Measure					
Foreign-born Cohort of Arrival					
1900-1904	0.383	0.302		-0.080	50,385
1905-1909	0.386	0.306		-0.080	53,007
1910-1914		0.314	0.216	-0.099	100,641
1915-1919		0.253	0.180	-0.073	13,158

Sources: Linked samples between the 1910-1920 census and 1920-1930 census (Ward 2019).
Notes: The data reports the raw means of the main segregation measure in the panel data, merging at the county level.

Table A8. Fraction US Adults on page for those who switched enumeration districts, evidence from ten Northern cities

Cohort	1910	1920	1930	Change over decade	N
Panel A. Switched enumeration district					
1900-1904	0.277	0.422		0.145	13,079
1905-1909	0.257	0.410		0.152	13,435
1910-1914		0.361	0.498	0.137	23,098
1915-1919		0.419	0.520	0.101	3,245
Panel B. Same enumeration district					
1900-1904	0.323	0.309		-0.014	1,385
1905-1909	0.293	0.341		0.048	920
1910-1914		0.361	0.437	0.076	2,906
1915-1919		0.434	0.490	0.055	185

Sources: Linked samples between the 1910-1920 census and 1920-1930 census (Ward 2019).
Notes: Table shows the mean fraction of US-born on the census page for different arrival cohorts and year of observation. Table is split into panels by those who were in the same enumeration district ten years later and those who were not. The data is limited those who started in ten Northern cities when enumeration district maps are available. The cities are Baltimore, Boston, Chicago, Cincinnati, Cleveland, Detroit, Manhattan, Philadelphia, Pittsburgh and Saint Louis.

Table A9. Decomposition of fraction US adults on page for enumeration district switchers and stayers

	1st obs.	2nd obs.	Change over decade	N	Contribution to Growth (%)
Switched district	0.321	0.460	0.139	52,857	96.59
Same district	0.344	0.392	0.048	5,396	3.41
Overall	0.323	0.454	0.131	58,253	100

Sources: Linked samples between the 1910-1920 census and 1920-1930 census (Ward 2019).
Notes: Table shows the mean fraction of US-born on the census page at first observation and second observation for the linked sample. Table is split by those who were in the same enumeration district ten years later and those who were not. The data is limited those who started in ten Northern cities when enumeration district maps are available. The cities are Baltimore, Boston, Chicago, Cincinnati, Cleveland, Detroit, Manhattan, Philadelphia, Pittsburgh and Saint Louis.

Table A10. Spatial Assimilation regression estimates

	Fraction of page 2nd gen		Fraction of page 2nd gen		Next-door HH is 2nd gen		Next-door HH is 3rd gen	
Data Structure:	Panel	RCS	Panel	RCS	Panel	RCS	Panel	RCS
Years in US	-0.009	-0.004	-0.012	-0.011	-0.026	-0.061	-0.017	-0.053
	(0.006)	(0.001)	(0.005)	(0.000)	(0.016)	(0.002)	(0.020)	(0.002)
Years in US sq	0.002	0.003	0.002	0.003	0.005	0.010	0.004	0.009
	(0.001)	(0.000)	(0.001)	(0.000)	(0.003)	(0.000)	(0.004)	(0.000)
Years in US cub	0.000	0.000	0.000	0.000	0.000	-0.001	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Years in US quad	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Arrival Cohort 1900-1904	-0.114	-0.135	-0.074	-0.086	-0.144	-0.140	-0.113	-0.106
	(0.004)	(0.001)	(0.004)	(0.000)	(0.008)	(0.001)	(0.008)	(0.001)
Arrival Cohort 1905-1909	-0.110	-0.117	-0.073	-0.075	-0.135	-0.119	-0.110	-0.091
	(0.003)	(0.000)	(0.002)	(0.000)	(0.006)	(0.001)	(0.005)	(0.001)
Arrival Cohort 1910-1914	-0.056	-0.064	-0.039	-0.042	-0.083	-0.063	-0.068	-0.050
	(0.004)	(0.001)	(0.004)	(0.000)	(0.015)	(0.001)	(0.014)	(0.001)
Constant (Arrival Cohort 1915-1919)	-0.387	-0.449	-0.400	-0.426	-0.332	-0.308	-0.433	-0.377
	(0.008)	(0.001)	(0.007)	(0.001)	(0.021)	(0.004)	(0.024)	(0.003)
Observations	434,382	5,605,690	434,382	5,605,690	391,137	2,847,670	391,137	2,847,670
R-squared	0.043	0.077	0.013	0.027	0.016	0.021	0.011	0.013

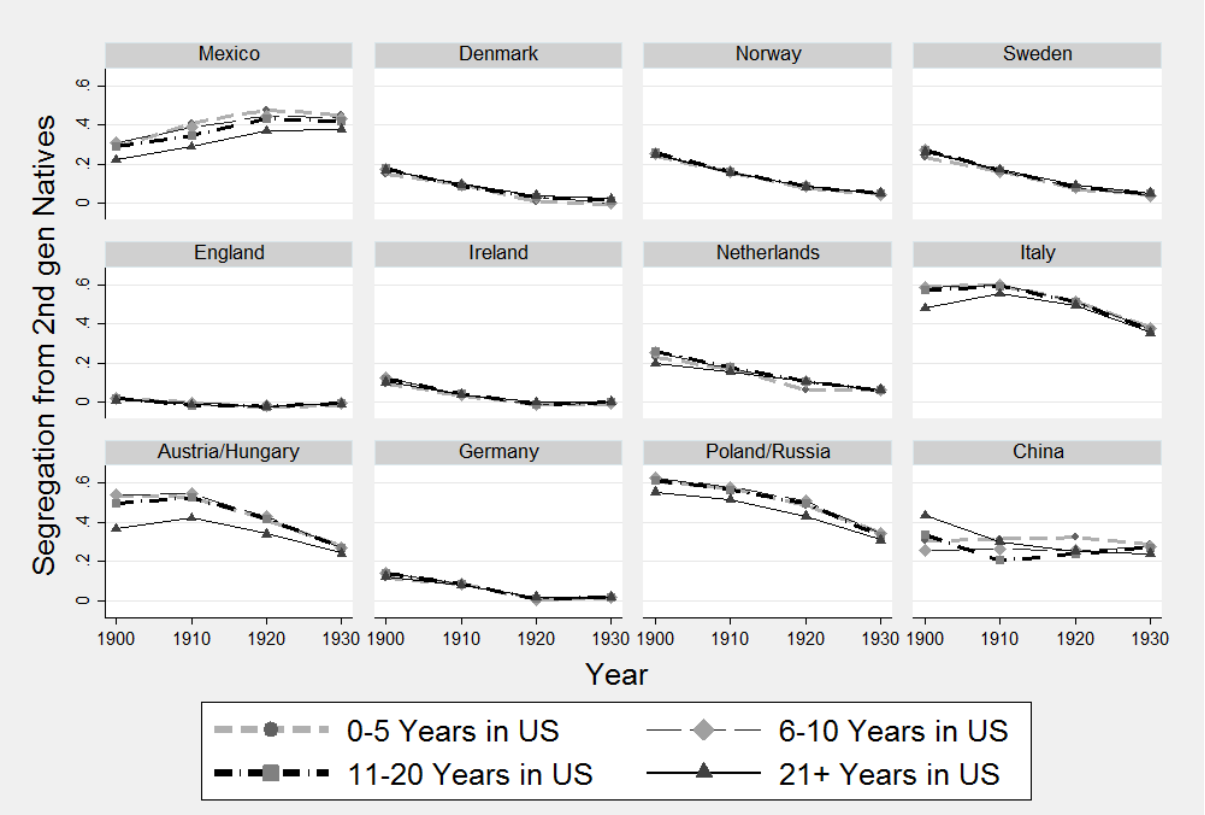
Sources: Linked samples between the 1910-1920 census and 1920-1930 census (Ward 2019) pooled with one percent random sample from 1910, 1920 and 1930 Censuses (Ruggles et al. 2018).
Notes: The dependent variable is the predicted gap between immigrants and natives after accounting for age and year effects.

Table A11. Fraction native-born on census page when accounting for geography

	Overall	Within State	Within County
Years in US	-0.00867	-0.00944	-0.00341
	(0.00563)	(0.00392)	(0.00404)
Years in US sq	0.00197	0.00232	0.00134
	(0.00103)	(0.000739)	(0.000744)
Years in US cub	-8.81e-05	-0.000125	-6.71e-05
	(6.86e-05)	(5.26e-05)	(5.10e-05)
Years in US quad	1.19e-06	2.28e-06*	1.12e-06
	(1.53e-06)	(1.26e-06)	(1.18e-06)
Arrival Cohort 1900-1904	-0.114	-0.112	-0.104
	(0.00381)	(0.00396)	(0.00350)
Arrival Cohort 1905-1909	-0.110	-0.111	-0.107
	(0.00273)	(0.00256)	(0.00251)
Arrival Cohort 1910-1914	-0.0560	-0.0644	-0.0657
	(0.00414)	(0.00340)	(0.00333)
Constant (Arrival Cohort 1915-1919)	-0.387	-0.295	-0.257
	(0.00767)	(0.00608)	(0.00613)
Observations	434,382	434,382	434,382
R-squared	0.043	0.043	0.042

Sources: Linked samples between the 1910-1920 census and 1920-1930 census (Ward 2019) pooled with one percent random sample from 1910, 1920 and 1930 Censuses (Ruggles et al. 2018).
Notes: The dependent variable is the predicted gap between immigrants and natives after accounting for age and year effects in the first column, including state fixed effects in the second columns, and including county fixed effects in the third column. See Figure A3 for estimated profiles.

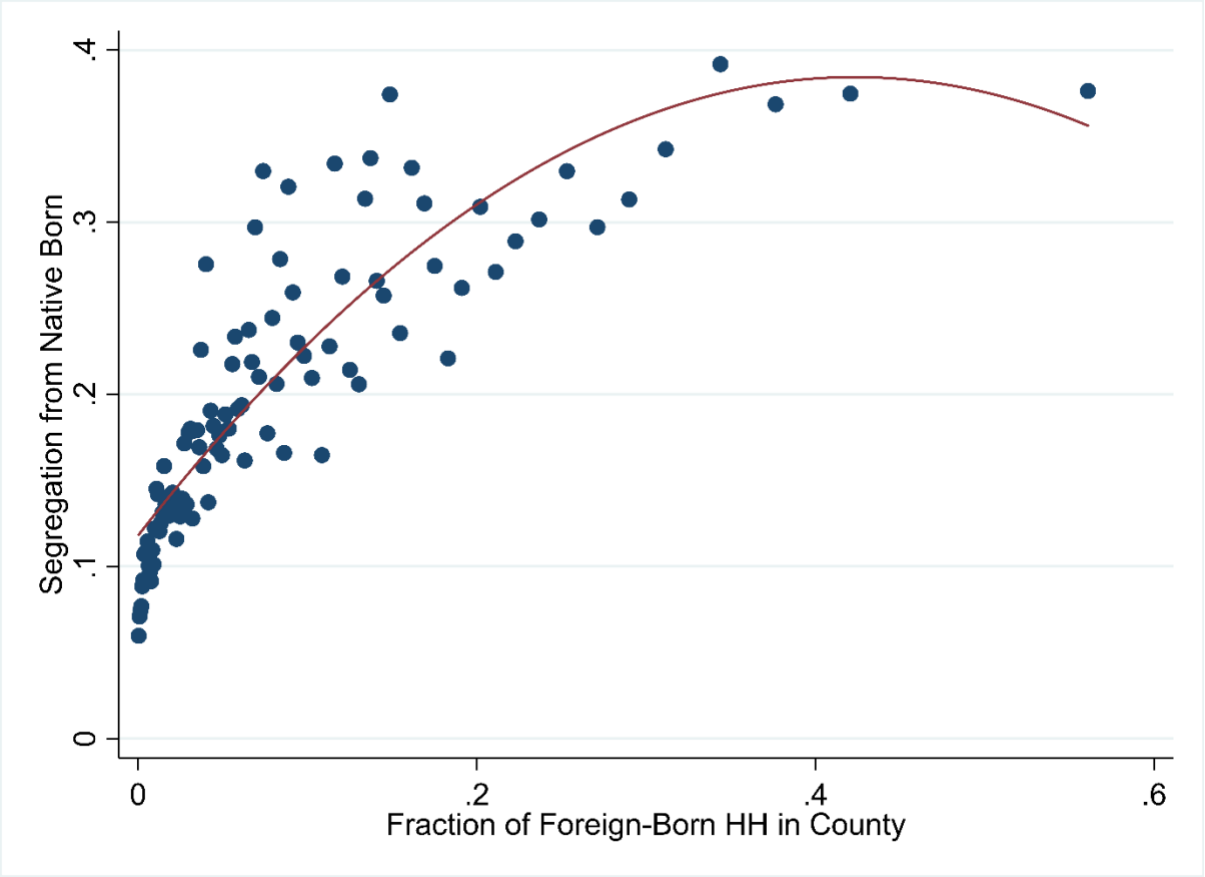
Figure A1. Segregation by years in the United States, by country of birth.



Sources: 1900 to 1930 full-count United States Censuses (Ruggles et al. 2018).

Notes: Segregation is calculated for each group from native-born households. The pattern shows little differences across years in the United States, suggesting little spatial assimilation. Little spatial assimilation is consistent with our estimates with panel data.

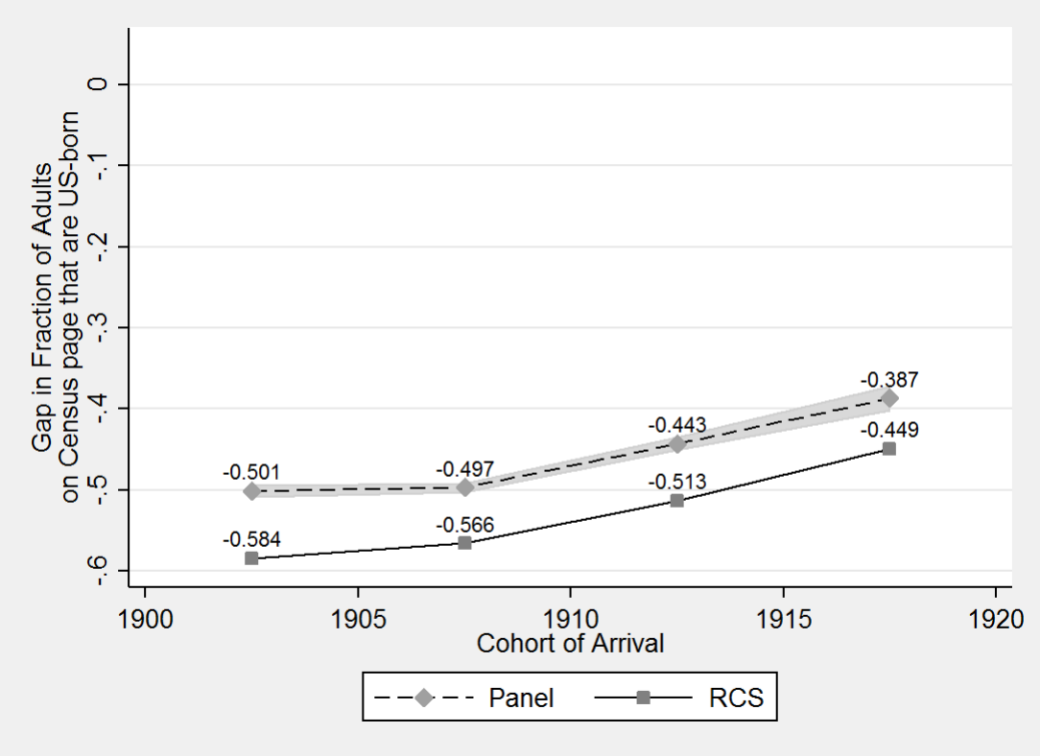
Figure A2. Relationship between Fraction Foreign-born in county and segregation



Sources: 1850 to 1940 full-count United States Censuses (Ruggles et al. 2018).

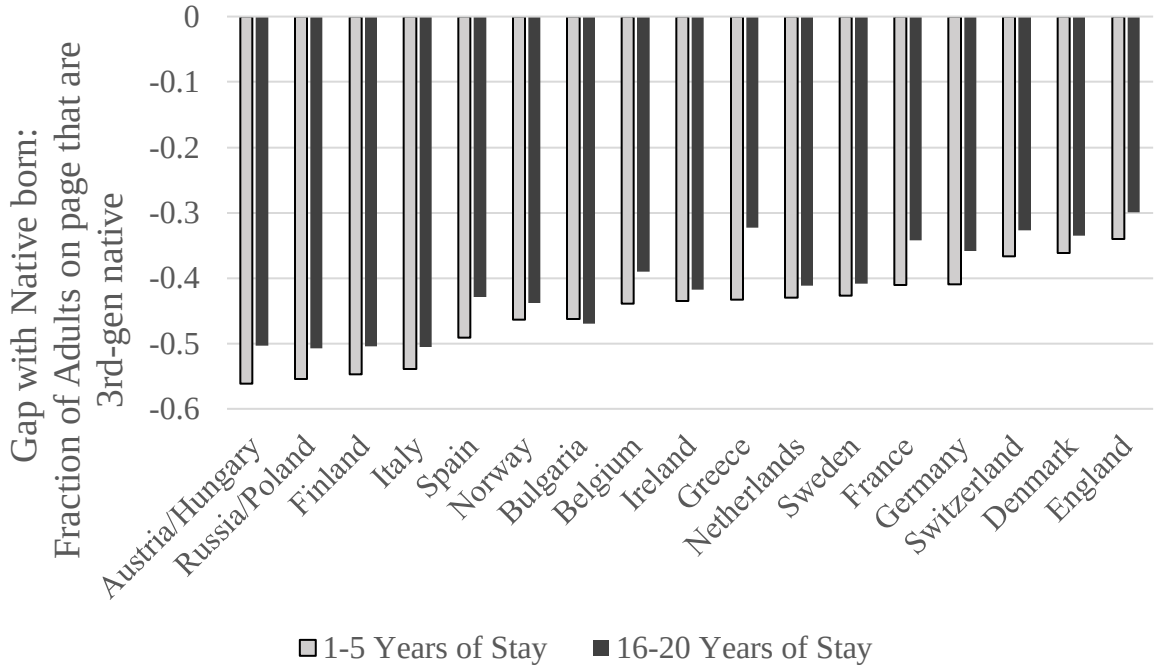
Notes: This is a bin scatter plot that shows the relationship between fraction foreign born in county with segregation at the county level. The underlying data are at the county-source country-year level.

Figure A3. Spatial assimilation profiles when accounting for geography



Sources: Linked samples between the 1910-1920 census and 1920-1930 census (Ward 2019) pooled with one percent random sample from 1910, 1920 and 1930 Censuses (Ruggles et al. 2018).
Notes: See Table A10 for underlying coefficients and regression

Figure A4. Spatial Assimilation by Country of Birth between 1910 and 1920



Sources: Linked samples between the 1910-1920 census and 1920-1930 census (Ward 2019) pooled with one percent random sample from 1910, 1920 and 1930 Censuses (Ruggles et al. 2018)
Notes: This is the same figure as Figure 5 from the main text, but this figure splits the sample by country of birth rather than by mother's tongue.

Appendix A. Further details on cleaning the data

We use the full-count data between 1850 and 1940 from the University of Minnesota Population Center. At the time of writing this paper, the 1850, 1880 and 1900-1940 censuses have been cleaned; the 1900 to 1940 are cleaned on a preliminary basis.[1] Therefore, we need to clean the 1860 and 1870 Censuses ourselves. The primary variables we are interested in cleaning are country of birth, county, city and household head. The process of cleaning the 1860 and 1870 datasets are described in further detail below.

- Country of birth

To clean the country of birth strings, we rely heavily on the strings already cleaned by the University of Minnesota Population Center for the 1850, 1880 and 1900 to 1940 full-count data. We create files that yield the most common country of birth codes (BPL) for each country of birth string (BPLSTR).

Armed with these files, we simply merge them to the uncleaned censuses starting with the nearest year – for example, the 1860 uncleaned census to the cleaned BPLSTR codes from 1850. For BPLSTR that are unmatched, we merge them onto later cleaned census files to update the BPL codes. For this process, we merge first to the 1880 or 1850, depending on closeness in time, and then to the 1900 to 1940 Census files. This is because border changes following World War I cause the pre-World War I censuses to be more reliable for assigning BPL codes. However, boundary changes do not bias results in text since we group countries by large region (i.e. Eastern Europe is one group).

After this initial pass, we have cleaned 99 percent of the country of birth strings. Following this, we tabulated a list of strings for each census and cleaned those which appeared more than 100 times. These were more common in the earlier censuses in the mid-19th century when individuals would sometimes list a town or a state within Germany. For country of birth strings which appeared less than 100 times, we left their country (bpl code) as missing and dropped them from the dataset.

[1] There is some evidence that group quarters variables have some inaccuracies in the full-count data, which may bias the household measure.

- Page indicators

We need to identify all immigrants who live next to each other on the same page. Rather than identify census page by NARA roll, reel and page, we used the codes for image id in the uncleaned data to determine whether an individual was on the same page. The image id is a code that Ancestry.com uses that combines string information from roll and page number, so it yields the same information but in one succinct variable. There are some instances in the Censuses where the page information was clearly inaccurate as there were over 50 households listed on a page. On the extreme end, there were 20,000 households listed on the same page in the 1860 Census, a problem that could not be fixed by resorting to information about the NARA roll or page number; however, this is not problematic for our main next-door measure. Moreover, the 1880 census include both sides of the census sheet to be on the same page, yielding of an average of about 100 individuals per sheet rather than the 50 in other censuses. While this does not strongly bias results, it may influence results in our robustness check of a "page-based" measure in Appendix C. Therefore, we sort by serial number and person number to ensure that we are capturing households in order and then create "synthetic pages" the start anew after 50 people.

- Relationship to head

We keep only the head of the household for our main segregation measure, but information about household head prior to the 1880 census was not explicitly listed in the Census. However, family numbers are provided within the raw data, which appears to separate individuals by household and not by nuclear family. Therefore, we keep the first family member listed in the 1860 and 1870 censuses to proxy for the household head.

- Identifying Households and Group Quarters

We do not have institutional or group quarters identifiers for the unclean censuses in 1860 and 1870. IPUMS codes group quarters based on the number of unrelated members in the household, typically if there are more than ten individuals who are unrelated to the household head. For 1860 and 1870, lacking relationship string data, we simply keep the first listed household member and drop households if there are more than twenty individuals in a family number who have different surnames.

- County

We merge the uncleaned county strings with the ICPSR county codes, which we referenced from the IPUMS website. https://usa.ipums.org/usa/volii/ICPSR.shtml

- City

For city, we merge the uncleaned strings with the IPUMS city codes. https://usa.ipums.org/usa-action/variables/CITY#codes_section. There are a few times where a city in earlier census years is part of a city in later years; for example, Northern Liberties, PA was coded as a separate city in 1850, but was later a part of Philadelphia. To consistently code cities, we include smaller cities as part of the main city; this occurs for Brooklyn as part of New York City, Georgetown as part of Washington DC, and Kensington, Mayamensing, Northern Liberties, Southwark and Spring Garden as part of Philadelphia.

- Urban

Urban status is not provided in the uncleaned census files. Following Logan and Parman (2017a), we define a county as urban as those with greater than 25 percent of the population living in an urban area, as defined by the IPUMS variable URBAN. We calculate the fraction of a county in an urban area using the 1850 to 1940 IPUMS samples.

- Country groupings

One issue when presenting results by country of birth is that countries change borders over time, especially before and after World War I. We make the following groupings

1. Russia / Poland includes Russia, Poland, Estonia, Latvia and Lithuania
2. Austria / Hungary includes Austria, Hungary and Czechoslovakia

Appendix B. Measuring immigrant segregation

We follow Logan and Parman (2017) for creating the segregation measure, but we make a few distinct changes to the formulas. The reason why we change the formula is because unlike black-white segregation which has two defined groups (black or white), immigrant segregation has multiple groups (Irish, German, Russian, etc.). Black and white are mutually exclusive sets where the union (mostly) forms the population prior to 1940; however, the union of immigrants from a certain country of birth and the native born do not form the entire population. Yet much of the following discussion closely follows Appendix 1 in Logan and Parman (2017).

The formula we use in the main results to calculate segregation measures is as follows:

$$\eta_c = \frac{E\left(\overline{native_c}\right) - native_c}{E\left(\overline{native_c}\right) - E(\underline{native_c})} \tag{1}$$

where $E\left(\overline{native_c}\right)$ is the expected number of immigrant households who have a native-born neighbor under random assignment, $native_c$ is the actual number of immigrant households from country c who are observed to have a native-born neighbor, and $E(\underline{native_c})$ is the expected number of immigrant households who have a native-born neighbor under complete segregation. Remember that immigration status or nativity status is defined by the household head. While the expected number of immigrant households with native-born neighbors seems like a straightforward concept, one must adjust for the fact we observe two neighboring households for those in the center of the census manuscript, but only one neighboring household for those at the top or bottom.

Let us define the following variables:

- $n_{c,N=2}$ – number of immigrants from country c with 2 observed neighbors
- $n_{c,N=1}$ – number of immigrants from country c with 1 observed neighbor
- n_{fb} – number of immigrants from all countries
- n_{all} – number of all households in area. Note that $n_{all} - n_{fb}$ is the number of native born

The expected number of immigrant households from a country of birth c with a native-born neighbor under random assignment is as follows:

$$E(\overline{native_c}) = n_{c,N=2} \cdot p(native\ neighbor|N=2) + n_{c,N=1} \cdot p(native\ neighbor|N=1) \quad \text{(B1)}$$
$$= n_{c,N=2}\left(1 - \left(\frac{n_{fb}-1}{n_{all}-1}\right)\left(\frac{n_{fb}-2}{n_{all}-2}\right)\right) + n_{c,N=1}\left(1 - \left(\frac{n_{fb}-1}{n_{all}-1}\right)\right)$$

The logic behind the formula is under random assignment and for those with two observed neighbors, the probability of having a foreign-born neighbor on one side is $\left(\frac{n_{fb}-1}{n_{all}-1}\right)$ and the probability of having foreign born neighbors on both sides is $\left(\frac{n_{fb}-1}{n_{all}-1}\right)\left(\frac{n_{fb}-2}{n_{all}-2}\right)$. Since we are interested in the case where an immigrant has at least one native-born neighbor, the probability of this occurring for an immigrant with two neighbors is simply one minus the probability of having two foreign-born neighbors, or $\left(1 - \left(\frac{n_{fb}-1}{n_{all}-1}\right)\left(\frac{n_{fb}-2}{n_{all}-2}\right)\right)$. It is straightforward to modify this formula where instead of measuring segregation of the foreign born of country c from natives, measuring their segregation from those outside the country of birth. This would change the formula to where instead of $\frac{n_{fb}-1}{n_{all}-1}$ measuring the likelihood a next-door neighbor was foreign-born, $\frac{n_c-1}{n_{all}-1}$ would measure the likelihood a next-door neighbor was from the same country of birth.

Now we turn to calculate the expected number of native-born neighbors under complete segregation, or $E(\underline{native_c})$. Complete segregation from natives would occur if all immigrants from an country of birth lived together along a line, leaving the two households on the sides of the neighborhood being either native-born or from a different country of birth. Complete segregation from the native born implies that the two households on either side are from different countries of birth; for example, an Irish neighborhood could be surrounded by German neighbors on both sides. Therefore, the lower bound for expected number of native-born neighbors $E(\underline{native_c})$ is equal to zero. Setting the lower bound equal to zero is not accurate for the special case when there are only one or no other foreign-born immigrants from another country living in the county. This event was uncommon, for example, not occurring in the 1880 Census. However, if one were to calculate the measure for smaller levels of geography, such as the enumeration district, then there may not be immigrants from other sources in the same enumeration district; if so, one should resort to the Logan and Parman (2017a) method of calculating the lower bound.

We also present estimates of the first generation from the third-plus generation, or of the first and second generations from the third-plus generation (which is a proxy for "ethnic segregation"). We can calculate these estimates for the 1880 and 1900-1930 censuses since both the mother and father's birthplaces are included in the data. The segregation measures can be conceptualized in the following table where the 2nd-generation can alternatively be conceptualized as "immigrants" or natives depending on the measure[2]:

Table B1. Different Segregation Measures

	1st generation	2nd generation	3rd-plus generation
1 v. 2nd plus (main measure)	Immigrant	Native	Native
1st v. 3rd-plus	Immigrant	-	Native
1st and 2nd v. 3rd-plus ("ethnic" segregation)	Immigrant	Immigrant	Native

When measuring segregation of the from the third-plus generation as in the second two rows, the probability of having a third-generation neighbor is now $\left(1-\left(\frac{n_{1st2nd}-1}{n_{all}-1}\right)\left(\frac{n_{1st2nd}-2}{n_{all}-2}\right)\right)$ where n_{1st2nd} is the number of first or second generation households in the area. Therefore, we can plug this equation into the formula for our segregation measures and measure the expected number of immigrants with 3rd-plus generation neighbors (or row 2 in Table B1) as:

$$E\left(\overline{native_c}\right) = n_{c,N=2}\cdot p(native\ neighbor|N=2) + n_{c,N=1}\cdot p(native\ neighbor|N=1) \quad \text{(B2)}$$
$$= n_{c,N=2}\left(1-\left(\frac{n_{1st2nd}-1}{n_{all}-1}\right)\left(\frac{n_{1st2nd}-2}{n_{all}-2}\right)\right) + n_{c,N=1}\left(1-\left(\frac{n_{1st2nd}-1}{n_{all}-1}\right)\right)$$

[2] Thanks to an anonymous referee for suggesting a table to show the different segregation measures.

The formula we use when measuring the expected number of 1st and 2nd-gen from the 3rd-plus generation (or row 3 in Table B1) is:

$$
\begin{aligned}
E\left(\overline{native_{1st2nd,c}}\right) &= n_{1st2nd,c,N=2} \cdot p(native\ neighbor|N=2) \\
&\quad + n_{1st2nd,c,N=1} \cdot p(native\ neighbor|N=1) \\
&= n_{1st2nd,c,N=2}\left(1-\left(\frac{n_{1st2nd}-1}{n_{all}-1}\right)\left(\frac{n_{1st2nd}-2}{n_{all}-2}\right)\right) \\
&\quad + n_{1st2nd,c,N=1}\left(1-\left(\frac{n_{1st2nd}-1}{n_{all}-1}\right)\right)
\end{aligned}
\tag{B3}
$$

Note that in Formula (B3), $n_{1st2nd,c}$ reflects either a 1st or 2nd generation household from country of birth *c*, instead of the main measure using only 1st-generation households.

Appendix C. Alternative ways to measure immigrant residential segregation

1. The page-based measure, which includes non-households and non-heads

The main measure of segregation is based on whether either of the next-door neighbor household heads are native born. This measure necessarily drops non-household heads, such as spouses, parents or servants. Moreover, the method drops non-households such as mining and railroad camps, poor houses and universities. Therefore, the household measure may provide an incomplete picture of interaction with the native born for the average immigrant.

We take an alternative approach to measuring segregation that does not require dropping non-households and non-household heads in the household. The approach is based on whether the foreign born are located on the same page as the native born, rather than whether the next-door head was native born. If the foreign born are not evenly spread throughout a county, then they will not appear on the same pages as the native born. Those on the census page are in close proximity since the census was taken on a line. The alternative segregation measure we use is the same basic formula for the main measure of segregation as in Equation (1):

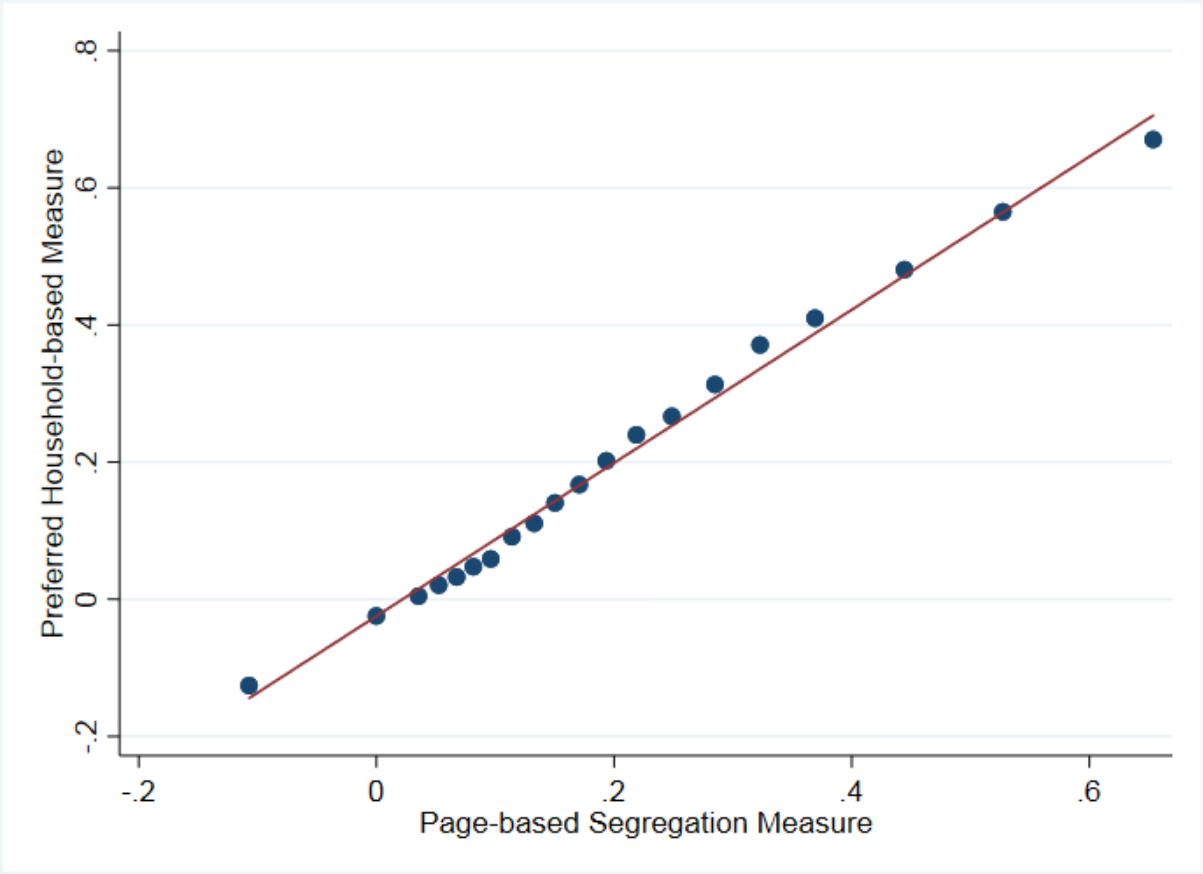
$$\delta_c = \frac{E\left(\overline{native_c}\right) - native_c}{E\left(\overline{native_c}\right) - E(\underline{native_c})} \tag{C1}$$

For this measure, now $E\left(\overline{native_c}\right)$ is the expected fraction native-born on the page within a county or city under random assignment. This is simply the total number of native-born in the city or county divided by the total number of pages. For this measure, we only include those aged 18 and older to reduce child-rearing bias. The variable $native_c$ is the observed fraction native-born individuals on the page for immigrants from source country *c*, and $E(\underline{native_c})$ is the expected fraction native-born on the page under complete segregation. Similar in the main section, we treat $E\left(\underline{native_c}\right) = 0$ since the foreign-born would be located either entirely on pages with other foreign born from the same country, or foreign born from a different country. Each foreign-born individual on a page has the same difference between the expected number and total number of native on the page; to aggregate the measure to the county level, we simply weight the measure by the number of foreign-born individuals on the page.

This "page-based" measure is similar in spirit to the next-door neighbor measure, but it captures segregation in a slightly different way. Besides the difference between using individuals instead of households, the page-based measure also measures segregation on the intensive margin of how many native-born does one live near, rather than just whether the individual lives near a native-born individual or not. We compare this page-based measure with the main household-based measure in Table C1 and show that the correlation between the two measures is 0.941. See Figure C1 for the binscatter relationship between the main household-based measure and this page-based measure. The difference between the measures could reflect a difference in measuring segregation, or the fact that we are able to include non-household heads and non-households in the measure; when calculate the "page-based" measure with only households, then the correlation with the main neighbor-based measure is 0.953. Therefore, the measures are closely related but do have slightly different results.

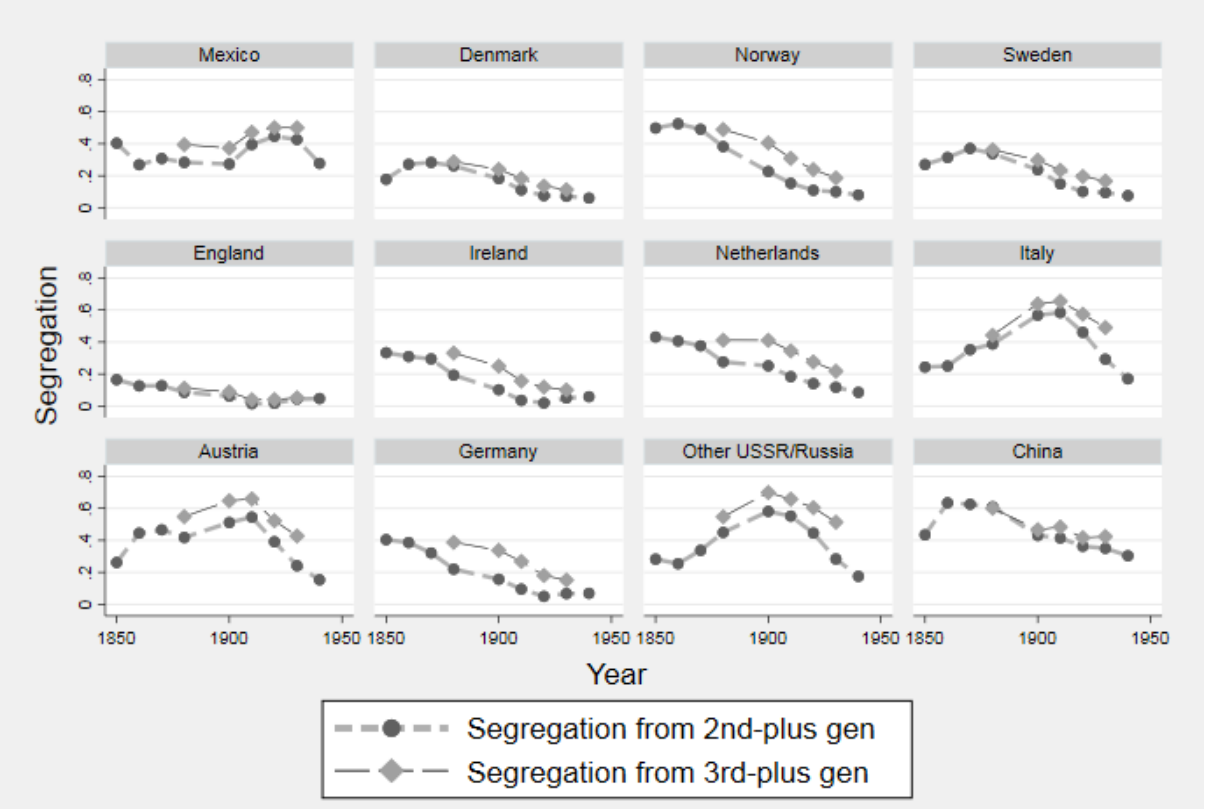
We present the page-based segregation trends by country of birth in Figure C2, which plots both trends for segregation from the 2nd-plus generation and segregation from the 3rd-plus generation. The broad relative levels and trends from the page-based measure are roughly the same as the neighbor-based measure. First, segregation levels are higher for Southern and Eastern Europeans at the turn of the 20th century relative to Western and Northern Europeans in the mid-19th century. Second, segregation trended to decrease past 1910 for all sources, and increased for Southern and Eastern Europeans between 1880 and 1910. Third, Chinese and Mexican segregation are relatively high, though the maximum levels are slightly lower than that of Southern and Eastern Europeans. The segregation trends by rural and urban areas are shown in Figure C3, which also demonstrate that trends were similar over time across rural and urban areas (except for Ireland). Moreover, segregation levels for Northern Europeans were very high in the mid-19th century. However, note that the levels of segregation across the page-based and neighbor-based measure are different.

Figure C1. Relationship between household-based measure and page-based measure



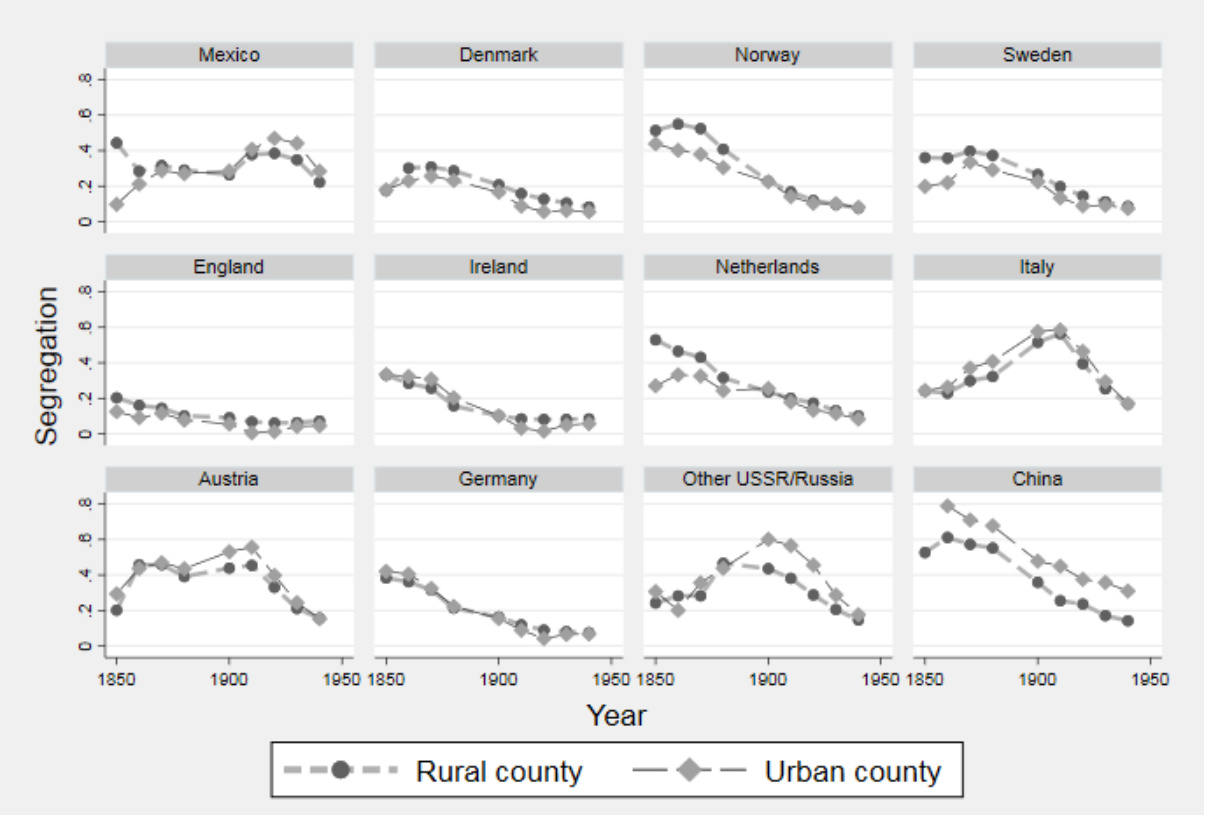
Sources: 1850 to 1940 full-count United States Censuses (Ruggles et al. 2018).

Figure C2. Trends in Segregation by Country of Birth, Page-Based Measure



Sources: 1850 to 1940 full-count United States Censuses (Ruggles et al. 2018).
Notes: The page-based measure is discussed in Appendix C. This figure mimic Figure 2 from the main text.

Figure C3. Trends in Segregation by Country of Birth and Urban/Rural Counties, Page-Based Measure



Sources: 1850 to 1940 full-count United States Censuses (Ruggles et al. 2018).
Notes: The page-based measure is discussed in Appendix C. This figure mimic Figure 3 from the main text.

Table C1. Correlation between preferred measure and page-based measure

	Household-based	Page-Based	Page-based, only HH
Main Household-based Measure	1		
Page-based Measure	0.9411	1	
Page-based Measure w/ only Household heads	0.953	0.9532	1

Notes: Correlation between measures when weighting for the number of households in county/year/country of birth cell.

II. Measuring Segregation from the Out-group

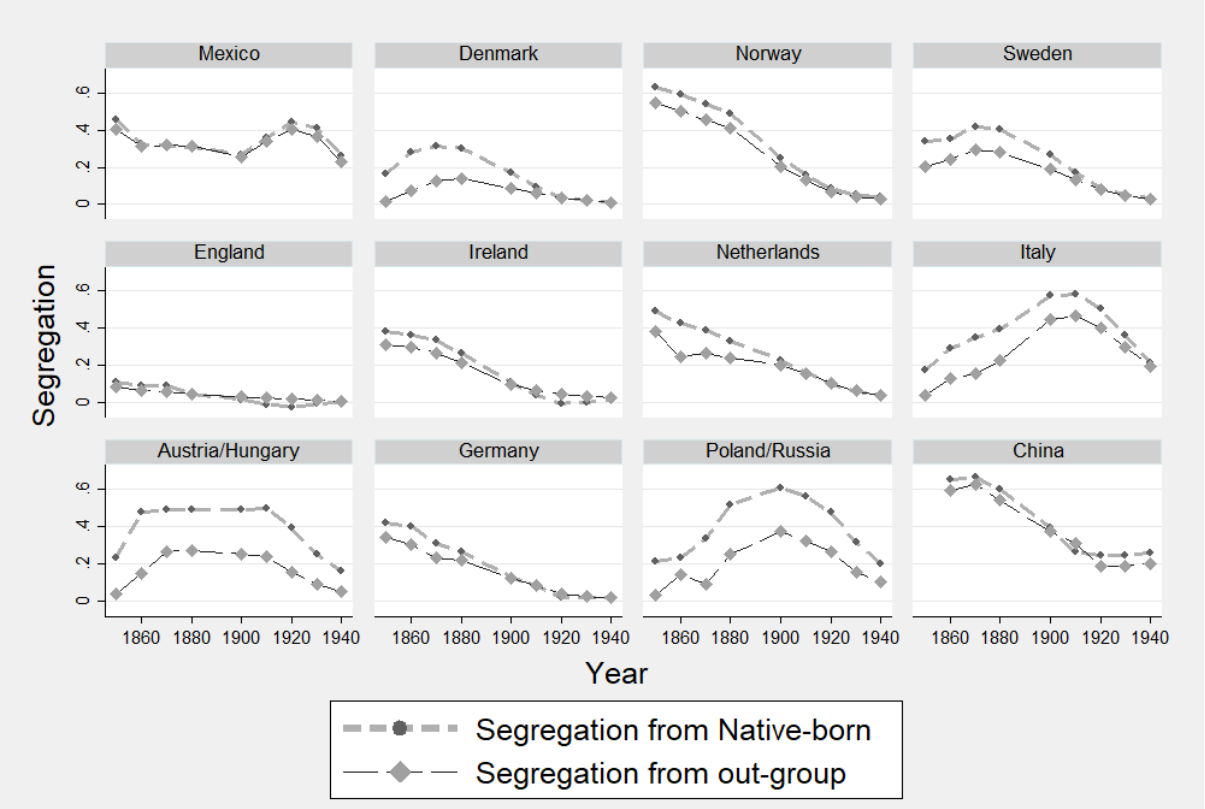
Our preferred measure of segregation is based on immigrants' (from a given country *c*) segregation from the *native-born*; that is, the in-group is based on country of birth, and the out-group are those born in the United States. Rather than using native-born as the out-group, one could use individuals from all other countries besides country *c* as the out-group. The fix for this in the formulas from Appendix B is simple: instead of counting the number of foreign-born with a *native-born* neighbor household head, we count the number of foreign-born with an *out-group* neighbor household head.

There are a few advantages for measuring segregation from other countries of birth rather than segregation from the native born; primarily, one does not measure negative levels of segregation for larger populations as we have done with our preferred measure. For example, we measure a negative level of segregation for Germans in New York City in 1940 because they were more likely to live next to a native-born household head than under random assignment. This is because Germans were more likely to next to US-born individuals rather than other non-German immigrants (e.g., from Southern or Eastern Europe). A negative level of segregation is not typical among standard segregation measures, such as the dissimilarity or isolation index. However, we prefer the main measure in text because we believe that living near the native-born is more relevant for measuring assimilation rather than segregation from the out-group; however, both measures are clearly informative for understanding immigrants' lived experience in the 19th and early 20th centuries.

In Figure C4, we present the segregation trends for our preferred measure and when measuring segregation from the out-group; this figure mirrors that of Figure 3. There are a few important differences in the trends and levels between the two measures. First, Eastern Europeans have a smaller level of segregation from the out-group than they do from the native born, and therefore were not as highly segregated from other individuals as southern Europeans. However, part of this may be because an immigrant from a given ethnicity or language may hail from different countries of birth; for example, Jewish immigrants from Russia/Poland, Germany, or Austria may live near each other and lower the measured level of segregation from other countries of birth. However, the level of segregation also falls for Southern Europeans, indicating that they also were less segregated from all others compared with segregated from the native born. Given

that segregation levels are lower for Southern and Eastern Europeans when measuring segregation from the out-group, this leaves Chinese immigrants as one of the most segregated sources, especially in the 19th century. Despite the level of segregation being lower for some sources, trends over time are largely similar.

Figure C4. Segregation from native-born versus segregation from out-group.



Sources: 1850 to 1940 full-count United States Censuses (Ruggles et al. 2018).
Notes: The out-group measure is discussed in Appendix C. Segregation is measured from the second-plus generation.

Table C2. Segregation from out-group (all other countries of birth)

	1850	1860	1870	1880	1900	1910	1920	1930	1940
Canada	0.121	0.106	0.114	0.117	0.101	0.076	0.053	0.036	0.023
Mexico	0.406	0.311	0.320	0.317	0.257	0.337	0.403	0.363	0.231
Cuba		0.012	0.197	0.117	0.223	0.205	0.106	0.059	0.020
Denmark	0.016	0.075	0.125	0.139	0.091	0.063	0.036	0.021	0.012
Finland				0.230	0.344	0.332	0.289	0.212	0.133
Norway	0.545	0.504	0.454	0.414	0.208	0.130	0.070	0.042	0.027
Sweden	0.204	0.241	0.295	0.280	0.189	0.132	0.082	0.048	0.028
England	0.083	0.066	0.060	0.046	0.030	0.025	0.020	0.013	0.010
Scotland	0.057	0.046	0.038	0.028	0.015	0.013	0.009	0.011	0.009
Wales	0.233	0.188	0.187	0.138	0.070	0.045	0.022	0.012	0.009
Ireland	0.312	0.301	0.269	0.214	0.096	0.065	0.047	0.034	0.030
Belgium	0.162	0.177	0.216	0.212	0.103	0.093	0.071	0.056	0.035
France	0.111	0.086	0.068	0.041	0.023	0.026	0.016	0.011	0.007
Netherlands	0.384	0.248	0.268	0.243	0.202	0.156	0.107	0.063	0.041
Switzerland	0.135	0.087	0.074	0.051	0.032	0.024	0.014	0.011	0.007
Greece				0.014	0.078	0.193	0.126	0.067	0.043
Italy	0.043	0.129	0.157	0.226	0.441	0.470	0.404	0.297	0.195
Portugal	0.048	0.185	0.223	0.207	0.285	0.287	0.280	0.216	0.142
Spain	0.043	0.039	0.020	0.018	0.037	0.132	0.114	0.069	0.047
Austria/Hungary	0.038	0.152	0.264	0.275	0.244	0.239	0.157	0.089	0.054
Germany	0.345	0.305	0.230	0.217	0.125	0.084	0.036	0.024	0.019
Poland/Russia	0.030	0.085	0.090	0.253	0.378	0.321	0.267	0.157	0.101
China		0.594	0.624	0.539	0.335	0.309	0.186	0.187	0.203
Japan					0.649	0.595	0.343	0.281	
Turkey					0.118	0.200	0.086	0.062	0.036

Sources: 1850 to 1940 full-count United States Censuses (Ruggles et al. 2018).

Table C3. Segregation from out-group (from non 1st or 2nd generation from same source), by country of birth

	1880	1900	1910	1920	1930
Canada	0.142	0.131	0.113	0.097	0.081
Mexico	0.414	0.366	0.446	0.476	0.448
Cuba	0.120	0.253	0.258	0.165	0.116
Denmark	0.144	0.108	0.090	0.066	0.047
Finland	0.255	0.357	0.350	0.321	0.273
Norway	0.455	0.300	0.231	0.168	0.120
Sweden	0.287	0.211	0.168	0.129	0.094
England	0.063	0.050	0.044	0.037	0.026
Scotland	0.041	0.026	0.022	0.017	0.016
Wales	0.177	0.115	0.088	0.055	0.035
Ireland	0.282	0.189	0.149	0.119	0.091
Belgium	0.230	0.148	0.124	0.088	0.071
France	0.062	0.036	0.035	0.024	0.017
Netherlands	0.281	0.296	0.259	0.206	0.147
Switzerland	0.067	0.049	0.041	0.029	0.024
Greece	0.019	0.080	0.195	0.128	0.068
Italy	0.232	0.448	0.484	0.432	0.363
Portugal	0.218	0.314	0.337	0.329	0.284
Spain	0.022	0.047	0.149	0.123	0.079
Austria/Hungary	0.283	0.275	0.268	0.187	0.127
Germany	0.281	0.241	0.194	0.121	0.094
Poland/Russia	0.258	0.395	0.341	0.311	0.218
China	0.544	0.354	0.377	0.258	0.298
Japan		0.650	0.599	0.345	0.294
Turkey		0.116	0.204	0.088	0.063

Sources: 1850 to 1940 full-count United States Censuses (Ruggles et al. 2018).

III. Counterfactual is perfect integration rather than random integration

The preferred measure of segregation compares the actual number of immigrant households with a US-born neighbor to the expected number of households under random assignment. However, an atypical feature about our preferred segregation measure is that the actual number of immigrant households with a US-born neighbor could be *more* than under random assignment in the case where immigrants are more likely to live near US-born households than next to immigrant households from other countries of birth. This leads to negative measure of segregation for some countries. While this is unusual, an advantage of our preferred measure is that it holds a consistent interpretation of a segregation values of 0 and 1, and it also provides information for when immigrants live closer to US-born households than expected.

An alternative way to calculate segregation is to use perfect integration with native-born households as the benchmark, rather than use random assignment of households. In this case, the following formula

$$\eta_c^{max} = \frac{E\left(\overline{native_c}\right) - native_c}{E\left(\overline{native_c}\right) - E(\underline{native_c})} \tag{C2}$$

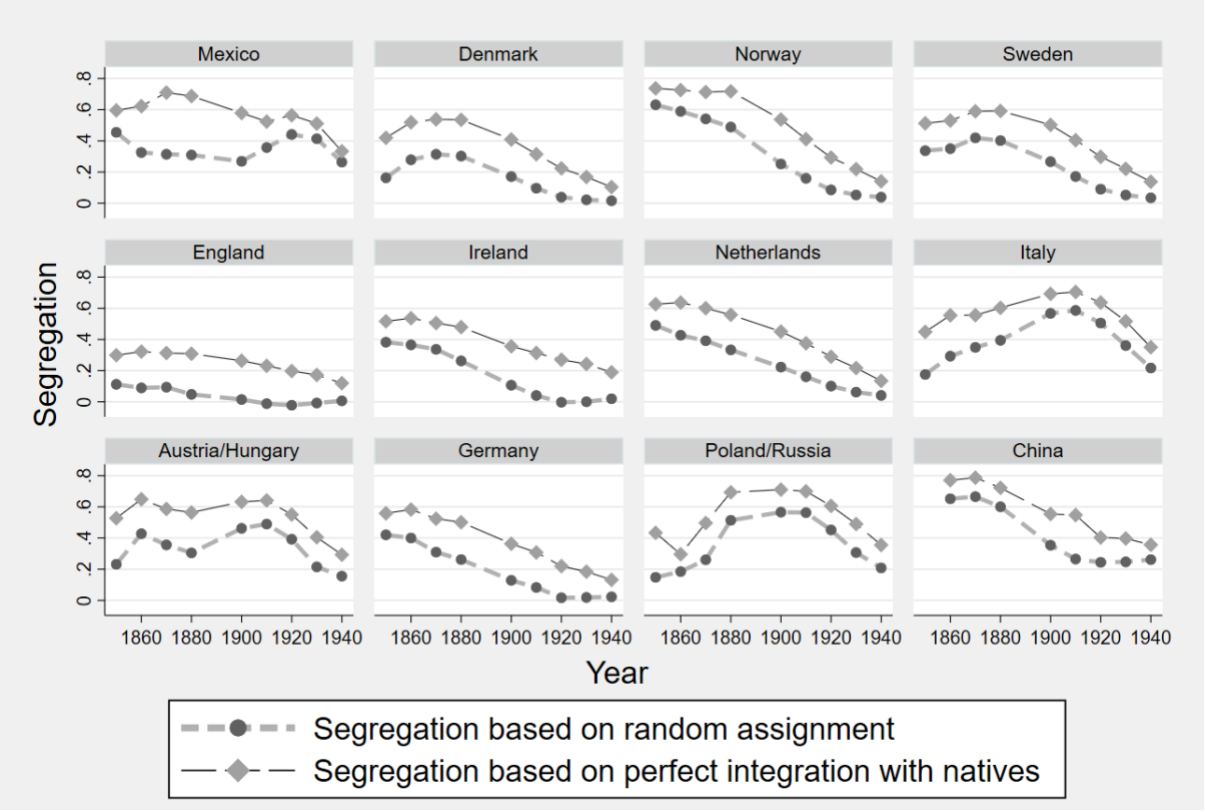
would calculate $E\left(\overline{native_c}\right)$ as the expected number of immigrant households with a US-born neighbor under perfect integration. To measure segregation in this alternative way, we allow $E\left(\overline{native_c}\right)$ to equal the number of a source's households, as long as there are more native-born households than foreign-born households.

The case is more complex if immigrant households outnumber native-born households in a county since not every immigrant households would have a US-born neighbor even under perfect integration. This occurred for less than 0.3% of the dataset, sometimes for counties along the US-Mexico border and in rural counties in the Midwest. When there are fewer native-born households than foreign-born households, it is possible to have a counterfactual neighborhood where one native-born household is placed in between two immigrant households such that each immigrant household has one US-born neighbor. In this case, the maximum number of immigrant households with a native-born neighbor is equal to twice the number of native-born households. Therefore, we allow $E\left(\overline{native_c}\right)$ to be equal to the number of immigrant households multiplied by two in cases

where the number of immigrant households is more than twice the number of native born households. In cases where the number of immigrant households is less than twice the number of native-born households, then $E(\overline{native_c})$ is equal to the number of foreign-born households.

The resulting estimates are shown in Figure C5 for 12 selected countries of birth. The segregation measures relative to perfect integration are mostly a level shift upward segregation based on random assignment. However, the measures trend similarly over time, and the relative comparisons are similar across most sources. Therefore, the interpretation of segregation from this measure is similar to the main one presented in text. The correlation coefficient between our preferred segregation measure and this new one is 0.83, showing that they capture similar information.

Figure C5. Segregation based on perfect integration with native-born.



Sources: 1850 to 1940 full-count United States Censuses (Ruggles et al. 2018).

References

Horan, Patrick M., and Hargis, Peggy G. County Longitudinal Template, 1840-1990. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1995-12-20. https://doi.org/10.3886/ICPSR06576.v1

Logan, Trevon D., and John M. Parman. "The national rise in residential segregation." *The Journal of Economic History* 77, no. 1 (2017a): 127-170.

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. *Integrated Public Use Microdata Series: Version 8.0* [dataset]. Minneapolis: University of Minnesota, 2018. https://doi.org/10.18128/D010.V8.0.

Ward, Zachary. "The low return to English fluency during the Age of Mass Migration." *European Review of Economic History.* 2019.