Online Appendix for

Internal Migration in the United States
Rates, Selection, and Destination Choice, 1850–1940

A Additional tables and figures

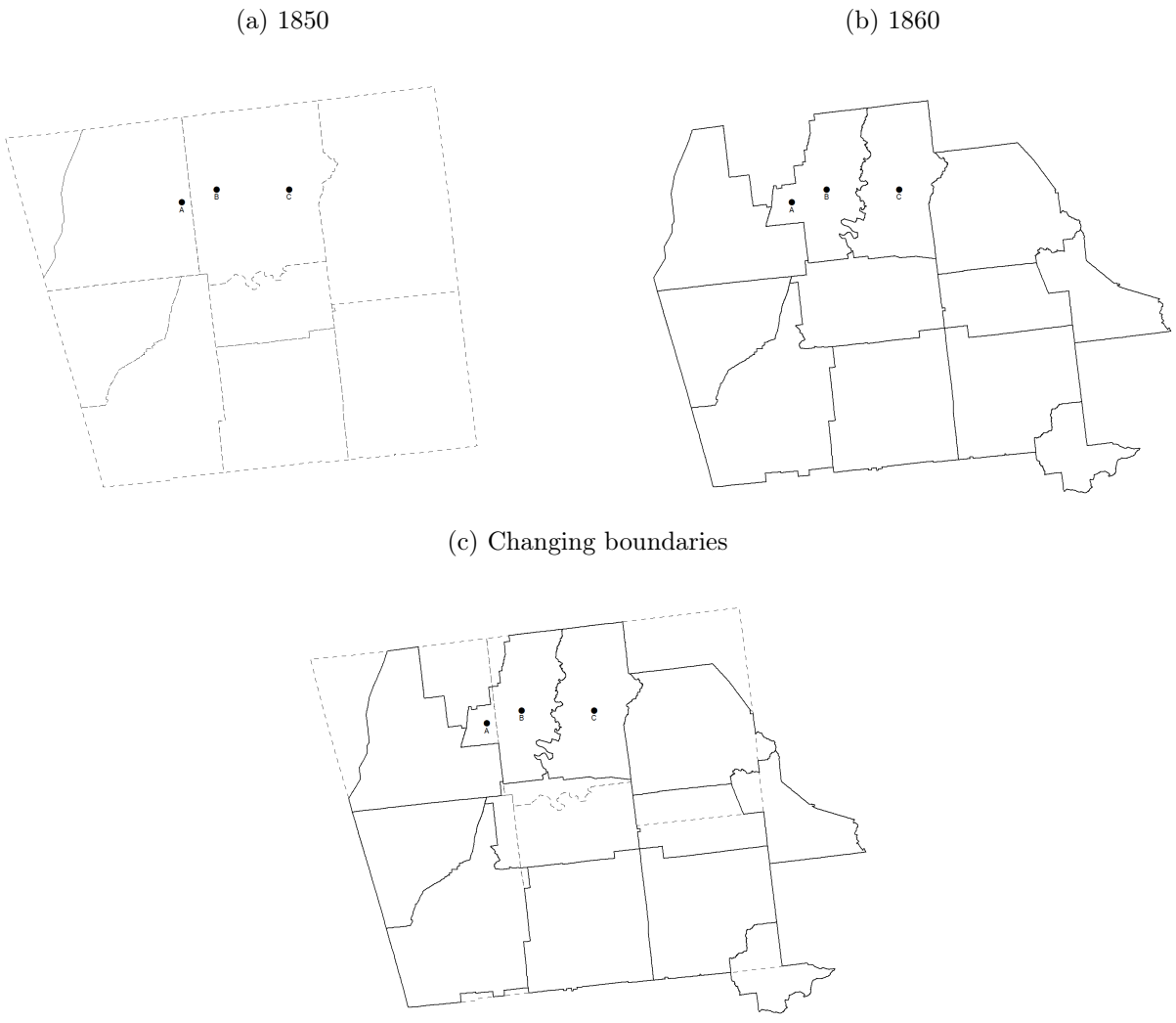


Figure A.1: Definition of inter-county migration

Note: Each map shows a group of counties in northwestern Georgia that experienced boundary changes 1850–1860. Points A and B are in the same county in 1860 but not in 1850. Points B and C are in the same county in 1860 but not 1850. Points A and C are never in the same county. Only moves between points A and C are classified as an inter-county move.

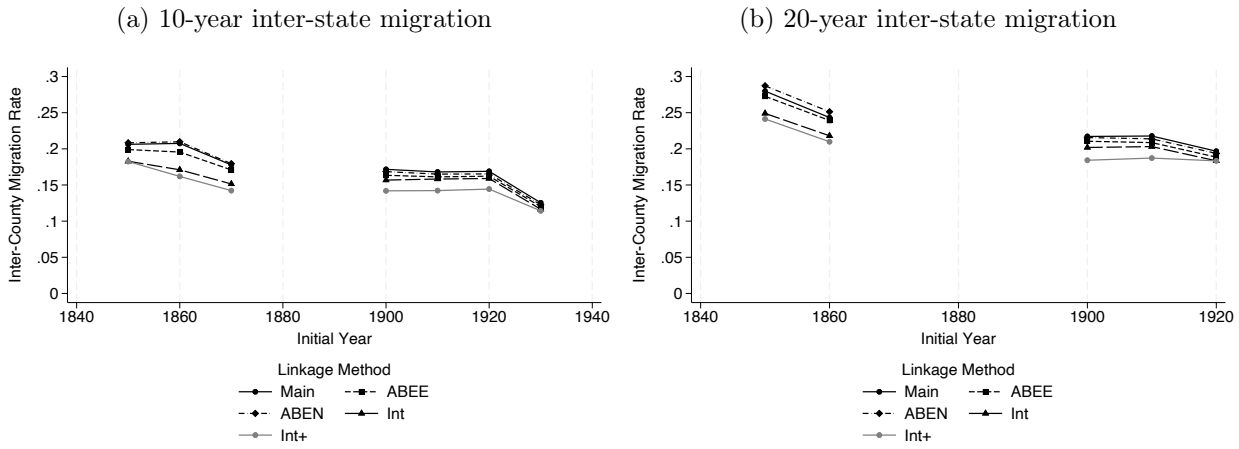


Figure A.2: Inter-state migration rates by linkage method and span, corrected for false matches

Note: Each figure shows the probability that an individual in the linked sample beginning in the year on the x-axis was observed living in a different non-overlapping state 10 or 20 years later, according to each linkage method. All observations are weighted by inverse linkage probability. The estimates are then adjusted for false matches according to the method presented in text.

Table A.1: Covariates available in each initial census

Variable	1850	1860	1870	1880	1900	1910	1920	1930
Age	X	X	X	X	X	X	X	X
Urban	X	X	X	X	X	X	X	X
City Population†	X	X	X	X	X	X	X	X
Farm	X	X	X	X	X	X	X	X
Literacy	X	X	X	X	X	X	X	X
Marital status				X	X	X	X	X
Occupation*	X	X	X	X	X	X	X	X
Household size	X	X	X	X	X	X	X	X
Household head	X	X	X	X	X	X	X	X
Birthplace	X	X	X	X	X	X	X	X
Place of residence	X	X	X	X	X	X	X	X
Real property	X	X	X					
Personal property		X	X					
Father's nativity				X	X	X	X	X
Owns home (mortgaged)					X	X	X	
Owns home (free and clear)					X	X	X	
Owns home								X
Value of home								X

†: used only to create measures of urbanization

*: used to create occupational categories, occupational rank measures, and the average occupational rank measure, all of which are available for all initial years.

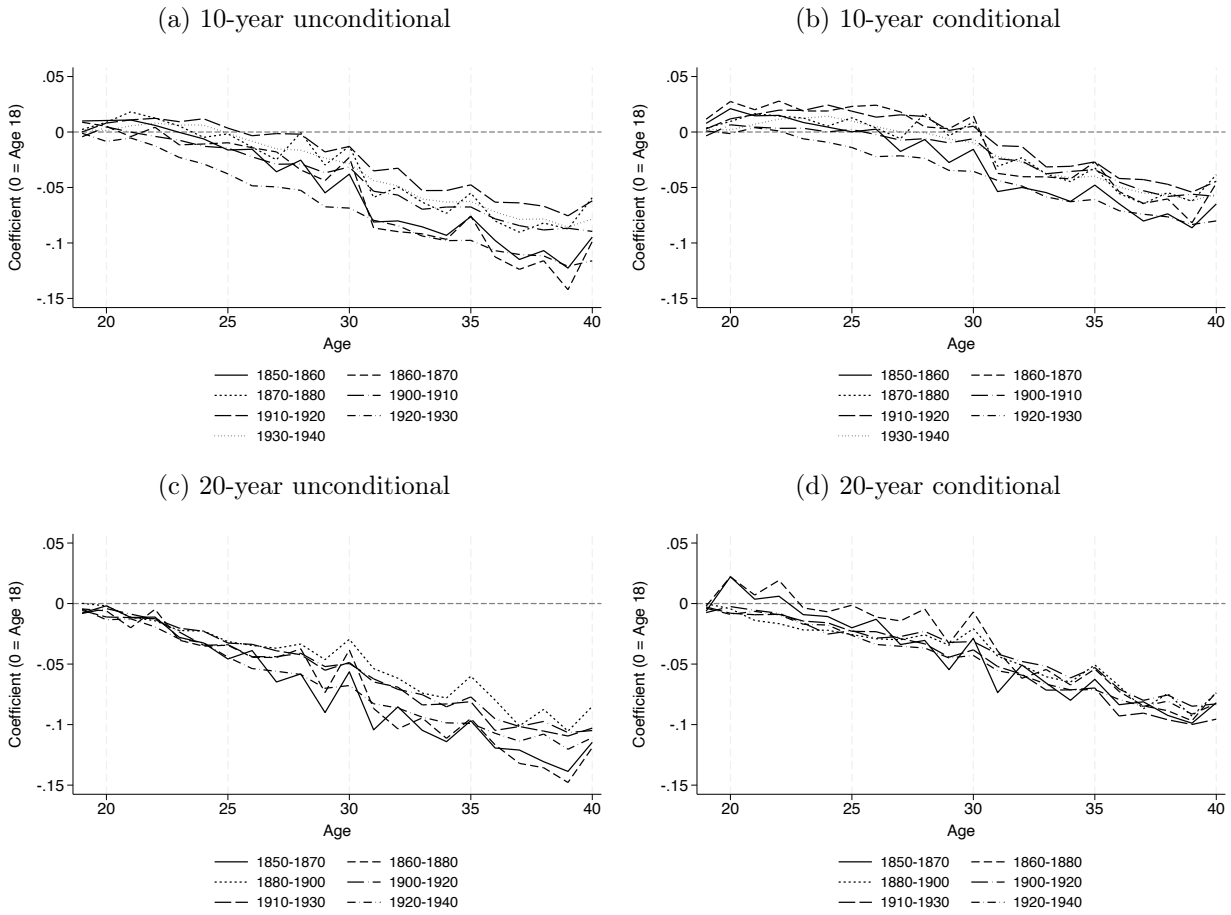
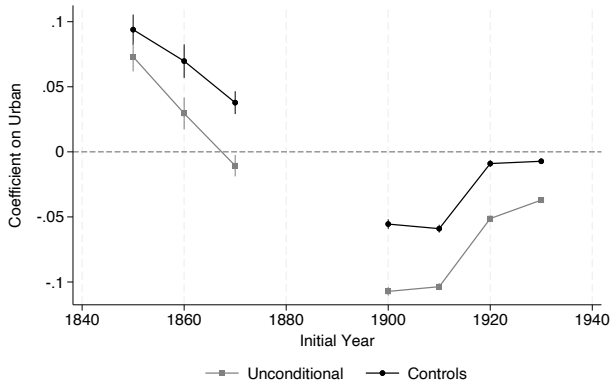


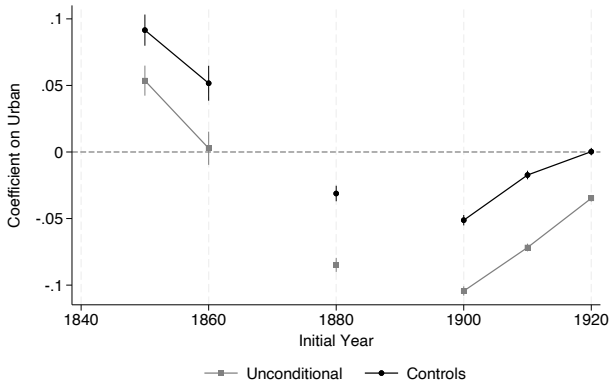
Figure A.3: Unconditional and conditional age profiles of migration

Note: This figure plots estimated age profiles (relative to age 18 in the initial year) of migration, either unconditionally, or from estimating equation (5) with the controls described in text. The age profiles are presented separately for each linkage period.

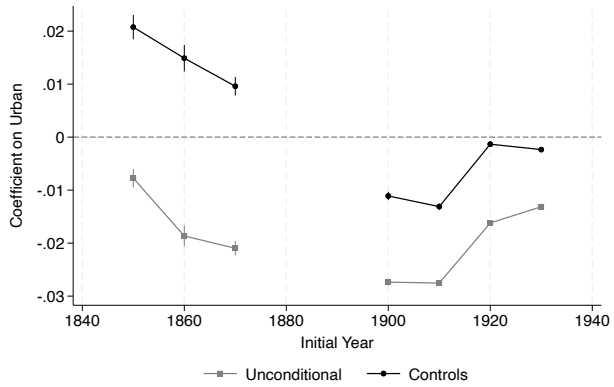
(a) Urban (20 Largest), 10-Year Spans



(b) Urban (20 Largest), 20-Year Spans



(c) Density, 10-Year Spans



(d) Density, 20-Year Spans

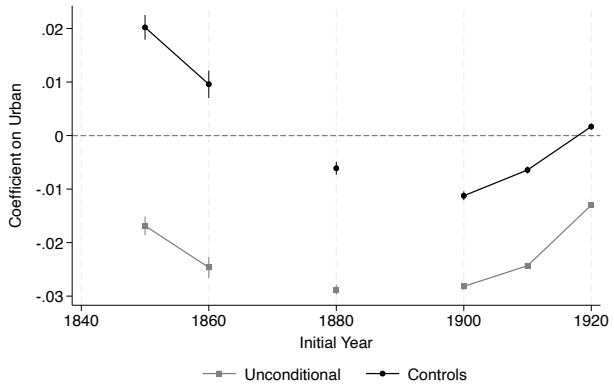
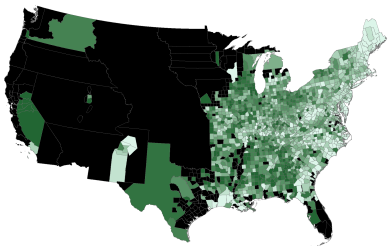


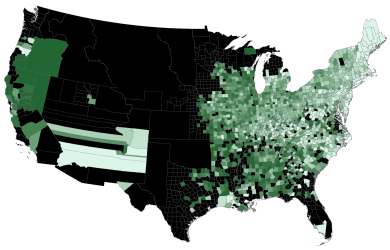
Figure A.4: Migrant selection by measure and span, alternate urban definitions

Note: Each figure shows the coefficient on the variable in question in a selection regression for migration over the stated span. All observations are weighted by inverse linkage probability. Vertical lines indicate 95-percent confidence intervals. Controls are all controls available in all census years, as well as state fixed effects.

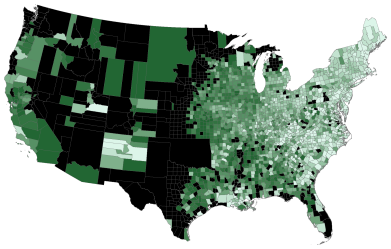
(a) 1850–1860



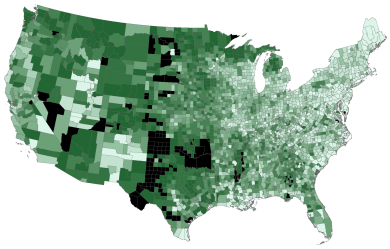
(b) 1860–1870



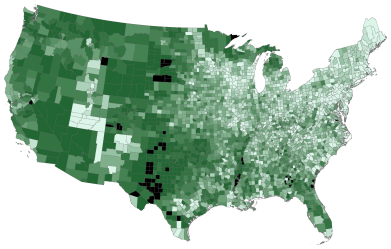
(c) 1870–1880



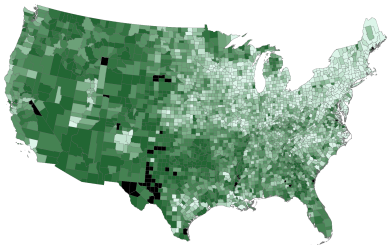
(d) 1900–1910



(e) 1910–1920



(f) 1920–1930



(g) 1930–1940

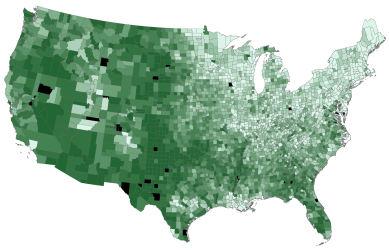
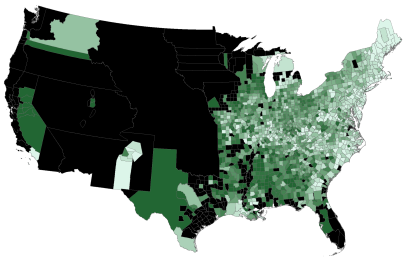


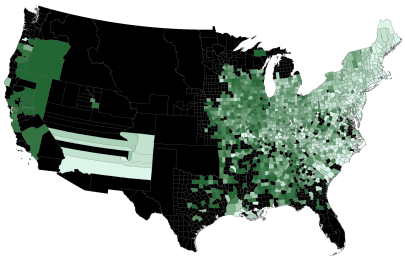
Figure A.5: County-level emigration rates, 10-year spans

Note: These are county-level emigration rates (i.e., the probability that an individual initially observed in that county would be observed in another county 10 years later) in the linked sample, weighted to correct for selection into linkage. Counties with fewer than 30 observations in the linked sample are dropped, and shown only in black. The scale is different in each year. It is based on deciles of migration rates across counties, with darker colors indicating higher migration rates.

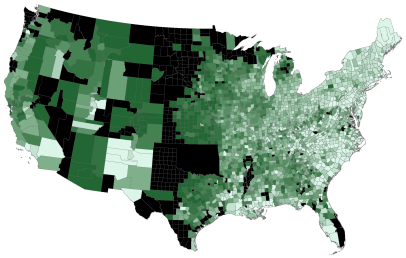
(a) 1850–1870



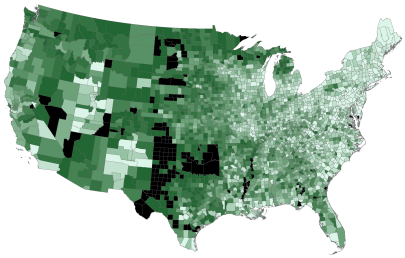
(b) 1860–1880



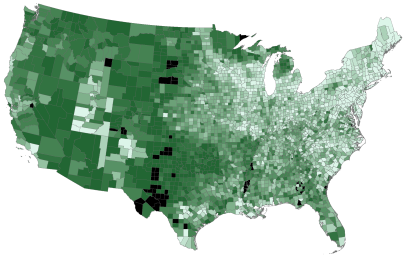
(c) 1880–1900



(d) 1900–1920



(e) 1910–1930



(f) 1920–1940

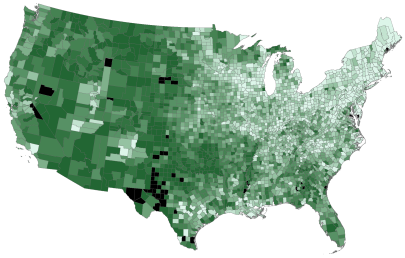


Figure A.6: County-level emigration rates, 20-year spans

Note: These are county-level emigration rates (i.e., the probability that an individual initially observed in that county would be observed in another county 20 years later) in the linked sample, weighted to correct for selection into linkage. Counties with fewer than 30 observations in the linked sample are dropped, and shown only in black. The scale is different in each year. It is based on deciles of migration rates across counties, with darker colors indicating higher migration rates.

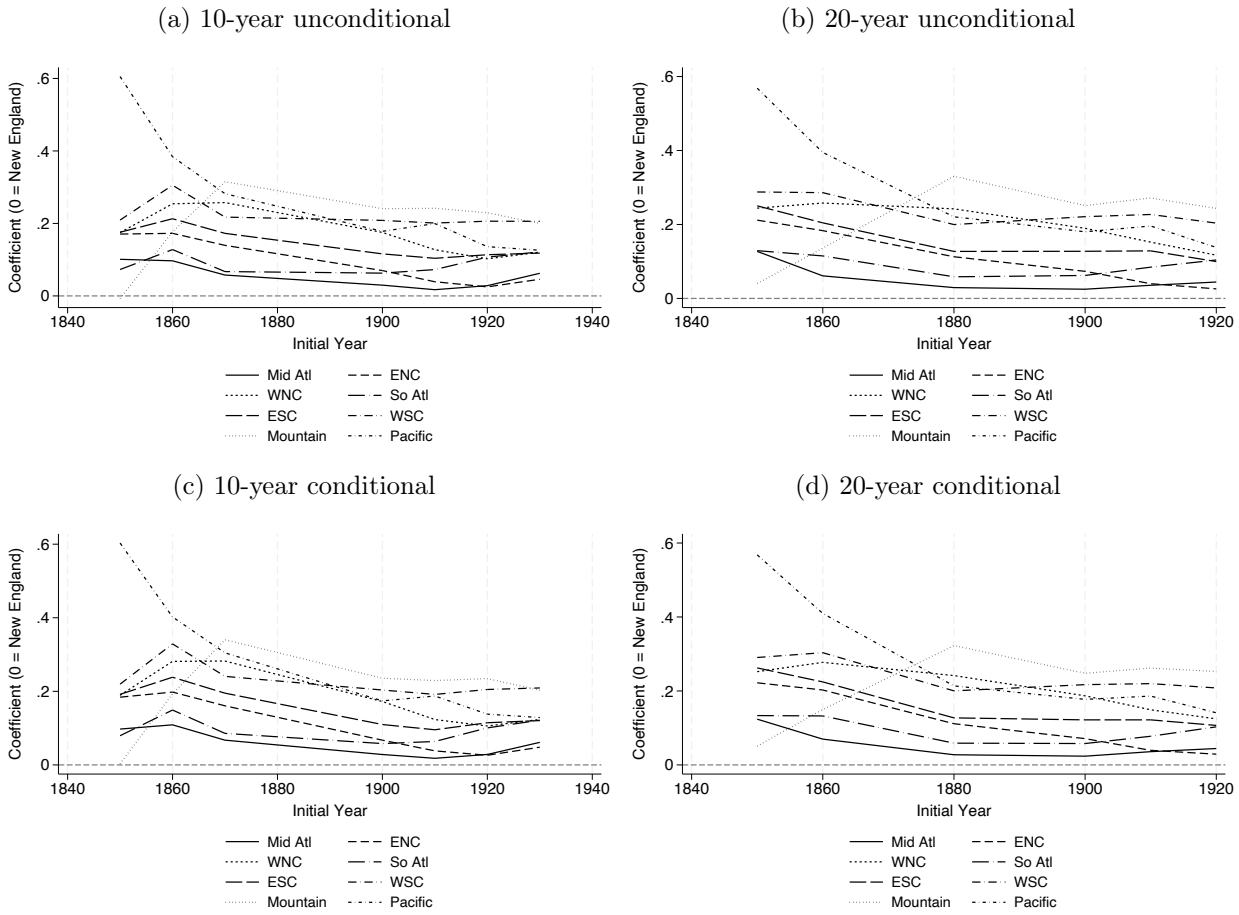
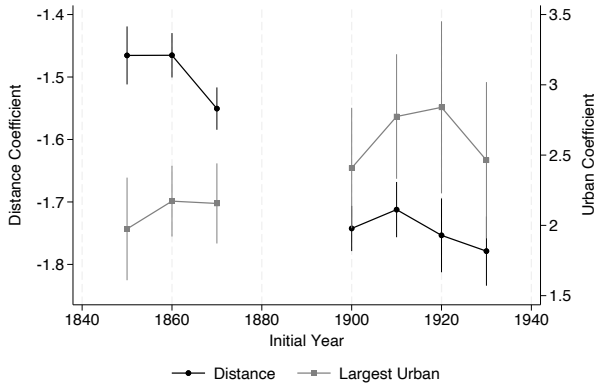


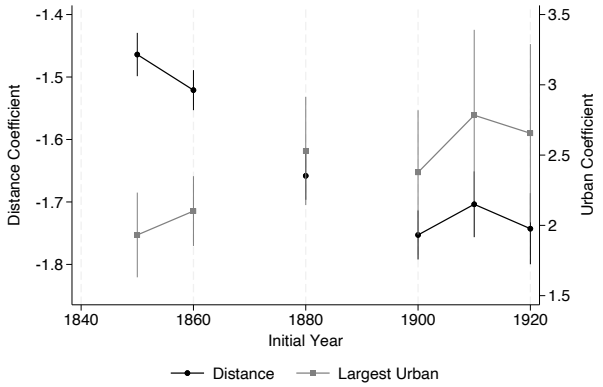
Figure A.7: Unconditional and conditional selection into migration by region

Note: This figure plots coefficients (with New England as the excluded region) for migrant selection regressions, either unconditionally, or from estimating equation (5) with the controls described in text. The coefficients indicate the premium in migration probability for each region over New England. Note that there is no data for spans that would include 1890; the lines span this period simply to enable clearer reading of the trend over time.

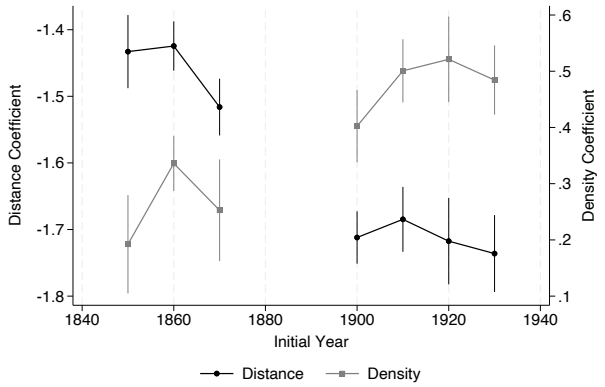
(a) Urban (20 Largest), 10-Year



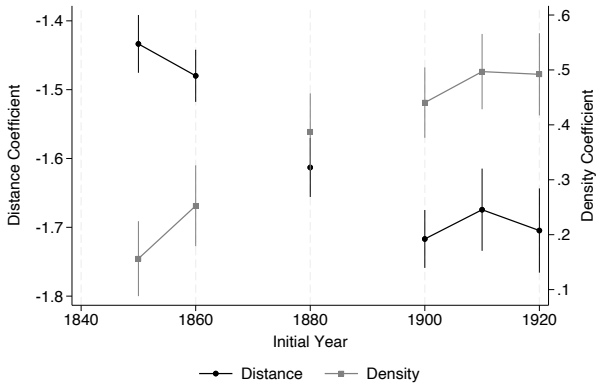
(b) Urban (20 Largest), 20-Year



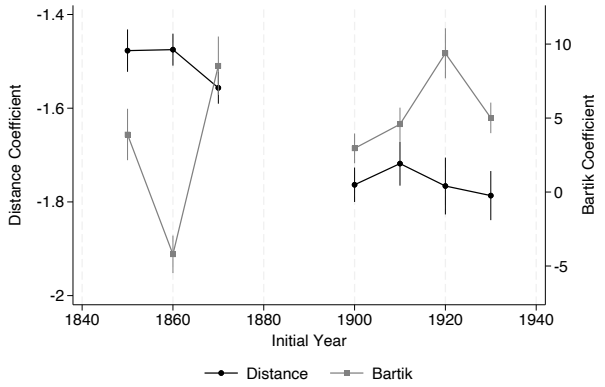
(c) Density, 10-Year



(d) Density, 20-Year



(e) Labor demand growth, 10-Year



(f) Labor demand growth, 20-Year

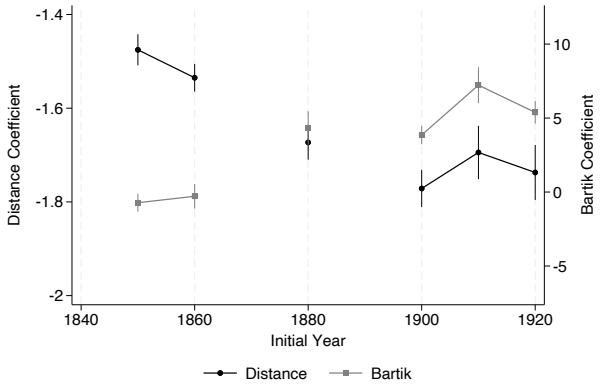
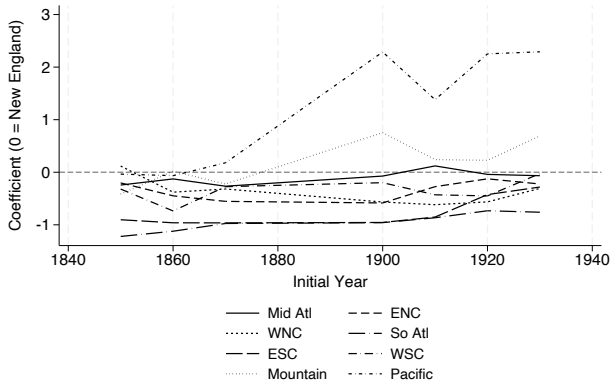


Figure A.8: Conditional logit results

Note: This figure presents coefficients on distance and the measure of urban residence or labor demand growth from the conditional logit estimation.

(a) Census divisions, 10-year



(b) Census divisions, 20-year

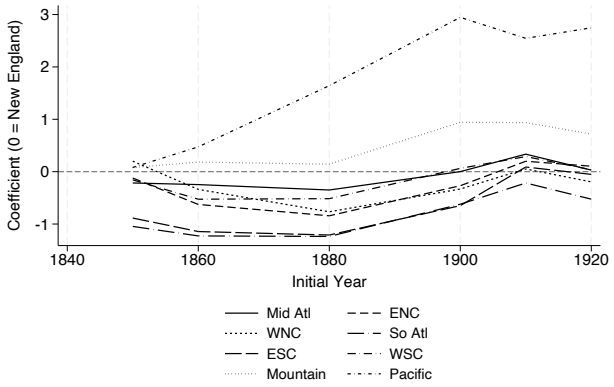


Figure A.9: Region indicators from conditional logit with labor demand growth

Note: These panels present the coefficients on the census division indicators, with New England excluded, controlling for labor demand growth. Note that there is no data for spans that would include 1890; the lines span this period simply to enable clearer reading of the trend over time.

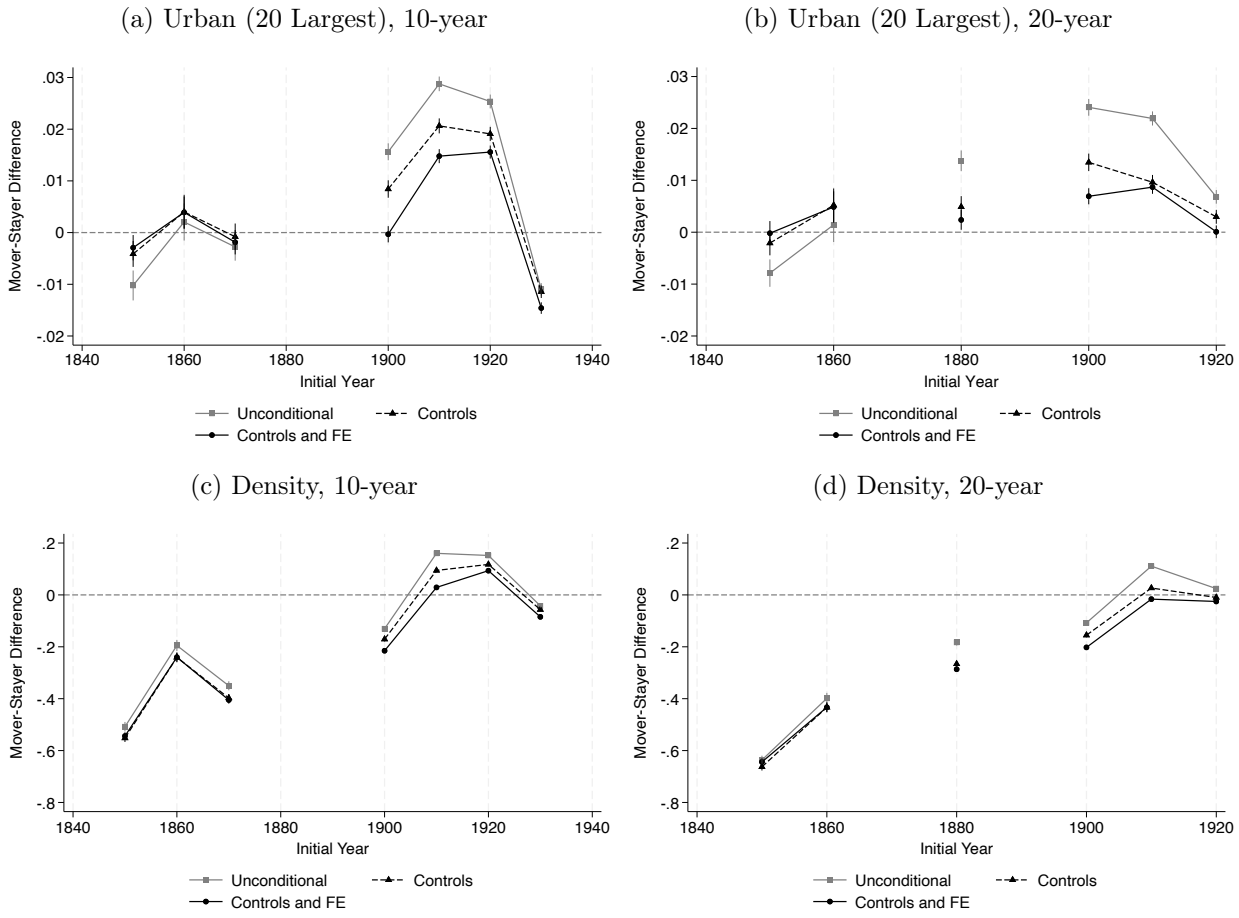


Figure A.10: Movers' change in urbanization and labor demand growth, alternate measures of urban residence

Note: This figure compares the growth in urbanization or density demand experienced by movers to that experienced by stayers, with either no controls, controlling for all observables, and controlling for all observables and county fixed effects.

B Additional details of data construction and estimation

B.1 Additional details of data linkage

I use a total of five linkage methods in my analysis. The main linkage method comes from the "basic" links provided by Zimran (2022a). Two of the other four methods are simply different in their linkage parameters: they are the ABE-Exact conservative method (*ABEE*) and the ABE-NYSIIS conservative method (*ABEN*) using crosswalks provided by Abramitzky et al. (2020). The other two methods are stricter. The first such method makes a match only when the Zimran (2022a) method and the two ABE methods agree;[53] I refer to this as the *intersection-of-matches method* (*Int*). The second, which I call the *intersection-of-matches-plus-corroboration* (*Int+*) method is even stricter, using only the subset of the intersection of matches in which the match is corroborated by all ostensibly time-invariant information not used in the linkage. For example, I require, where parents' birthplaces are provided in both censuses, that they agree across the sources.

Linkage rates for the main linkage method are presented in Table B.1. Figure B.1 compares the observable characteristics of the linked sample and the sample at risk for linkage for each linkage span. To reweight the data to adjust for selection into linkage on the basis of observable characteristics, I estimate, for each census linkage span, probit regressions for the probability of successful linkage as a function of all observables in the initial census. I then reweight the linked data by the inverse of the estimated conditional linkage probability, winsorized at the bottom 1 percent to avoid domination of the results by a small number of observations. In addition to the data in the census, I also use information on name length and commonality from Zimran (2022b). As is standard in such settings, this approach cannot address selection into linkage on the basis of unobservable characteristics.

B.2 Construction of occupational ranks

I create my occupational ranks as follows. First, following Collins and Zimran (2023), I assign to each individual an occupational score based either on the average wealth holdings of each occupation in 1870 or based on the average income of the occupation in 1900 (Preston and Haines 1991). Then, based on these two scores and the *occscore* variable provided by Ruggles et al. (2021), which is in turn based on data from the 1950 census, I determine the rank of each occupation relative to the white male population aged 18–64 in each census.[54] My occupational status measure is the simple average of these three ranks.

B.3 Representativeness of the corroboration sample

Figure B.2 compares the observable characteristics of the full linked sample to those of the corroboration sample. The most notable difference between the two is that, as is to be expected, individuals in the corroboration sample have substantially more children and are somewhat older than the full

[53] Abramitzky et al. (2021) show that the intersections of several sets of matches have a lower false positive rate than individual methods.

[54] Following Collins and Zimran (2023), I probabilistically reclassify the occupations of men in agriculture in 1850 who lived with a head of household who was a farmer and to whom they were related. This addresses the changing definition of a farmer from that census to future ones. I also follow Collins and Zimran (2023) in assigning family members of farmers the average of the occupational status of farmers and farm laborers.

linked sample. To get a sense of the severity of bias in false match rates induced by unrepresentativeness, I performed the following exercise, focusing on the number of children because it is the variable with the most severe selection. I compute the false match rates implied by limiting the corroboration sample to individuals with at least some number c of children, where c is incrementally increased from 0 to 9.[55] The intuition is that, if the false match estimates are sensitive to selection, then exacerbating selection by making the selection criterion more stringent should yield large, or at least meaningful, changes in the estimates. Figure B.3 shows, to the contrary, that the estimates are quite stable, suggesting that the estimates of false match rates, and especially of their changes over time, are, in fact, quite stable and thus that unrepresentativeness of the corroboration sample is unlikely to be an issue.

B.4 Conditional logit estimation

In principle, estimating the conditional logit model in equation (7) is computationally difficult because of the large number of options for individuals to choose from (all counties in the United States other than county o).[56] Fortunately, Guimarães, Figueirdo, and Woodward (2003) provide a method to simplify this estimation, making it tractable, though still computationally costly.[57] The cost of this simplification is that I am not able to control for different characteristics of internal migrants. Instead, the estimates must be thought of as providing a measure of the attractiveness of various location characteristics to internal migrants as a whole over time. In principle, it is possible to control for these characteristics by dividing the sample along any relevant characteristic and estimating the model separately, but the estimation burden is then effectively doubled for each binary variable added.

[55] The estimates for restriction to 0 and 1 children are the same because individuals must have at least one child in the household in order to be included in the corroboration sample.

[56] Any final-year county with borders overlapping those of county o must also be excluded because of how inter-county migration is defined.

[57] Guimarães, Figueirdo, and Woodward (2003) do not use weights. I adapt their method to the use of weights by replacing the number of individuals choosing a given option with the sum of normalized weights of individuals choosing a particular destination. The method is computationally costly because it requires the inclusion of origin-county fixed effects. I cluster standard errors by destination county for conservative inference.

Table B.1: Linkage rates according to the preferred linkage method

Span	(1) Start	(2) Linked
1850–1860	2,980,784	384,687 (0.129)
1850–1870	2,980,784	347,741 (0.117)
1860–1870	3,840,785	383,199 (0.100)
1860–1880	3,840,785	374,981 (0.098)
1870–1880	4,488,154	575,268 (0.128)
1880–1900	6,527,283	776,295 (0.119)
1900–1910	10,198,622	1,549,020 (0.152)
1900–1920	10,198,622	1,450,093 (0.142)
1910–1920	12,552,719	2,196,851 (0.175)
1910–1930	12,552,719	2,061,328 (0.164)
1920–1930	14,338,600	2,958,097 (0.206)
1920–1940	14,338,600	2,646,060 (0.185)
1930–1940	17,366,105	3,843,088 (0.221)

Numbers in parentheses are the fraction of individuals who were successfully linked.

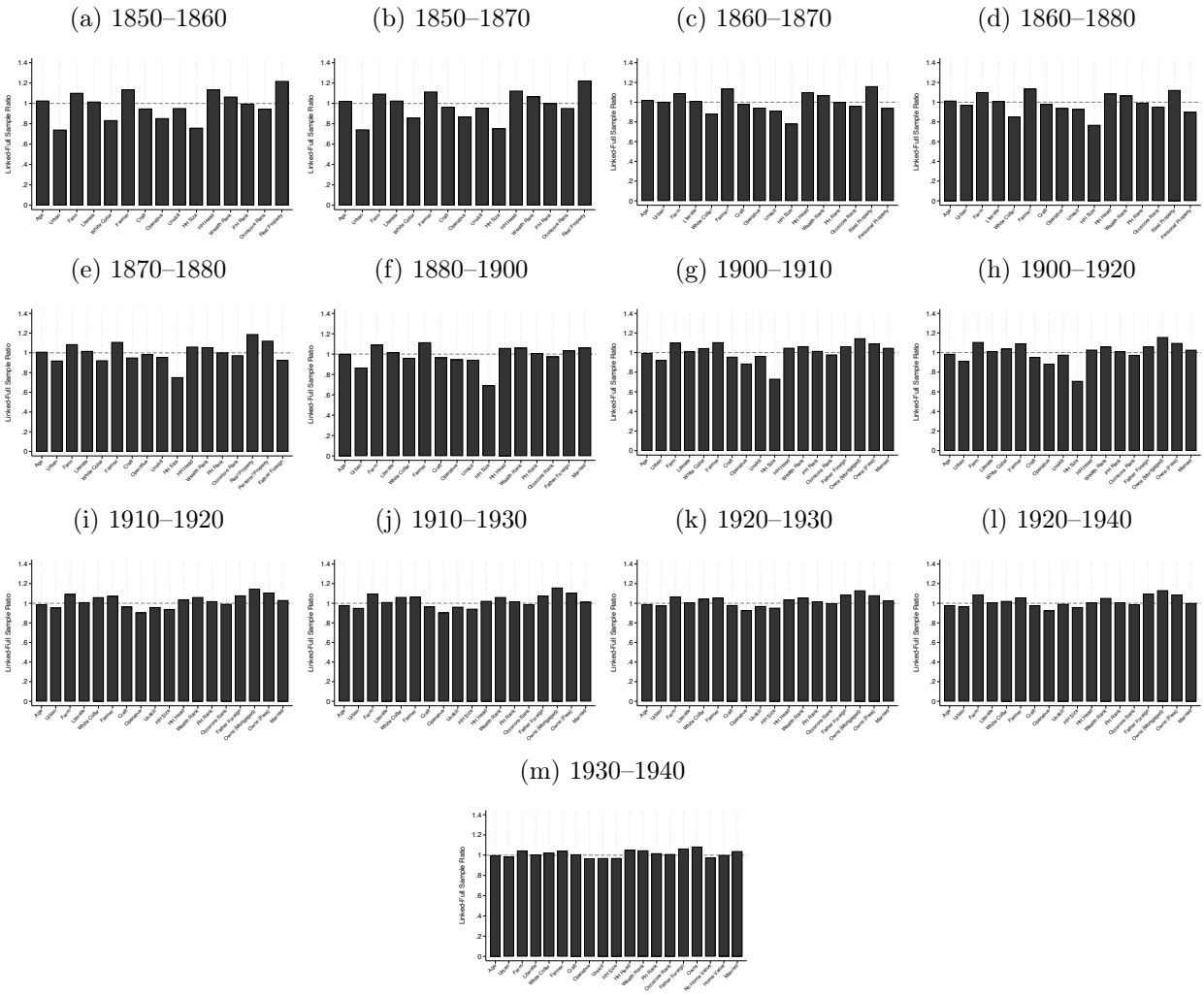


Figure B.1: Representativeness of the linked sample

Note: Each bar presents the ratio of the mean of each variable in the linked sample relative to the full sample at risk for linkage.

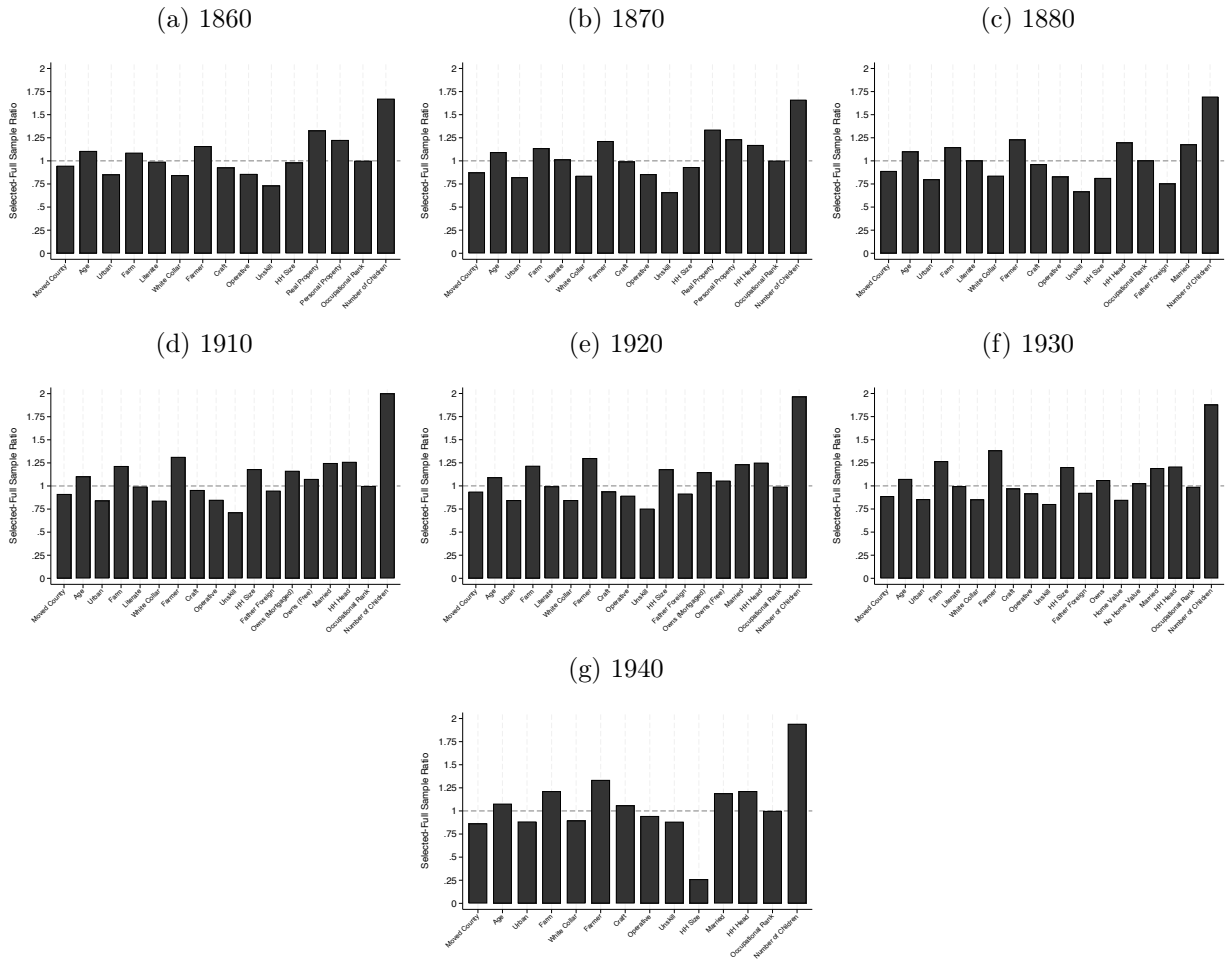


Figure B.2: Comparison of observables of the linked sample and the corroboration sample

Note: Each bar presents the ratio of the mean of each variable in the corroboration sample to the full linked sample. The indicated year refers to the latter year of a 10-year span. So, for instance, the 1860 figure compares the 1860 observables for the sample linked 1850–1860 to the subset of this sample that could also have its internal migration determined by family composition.

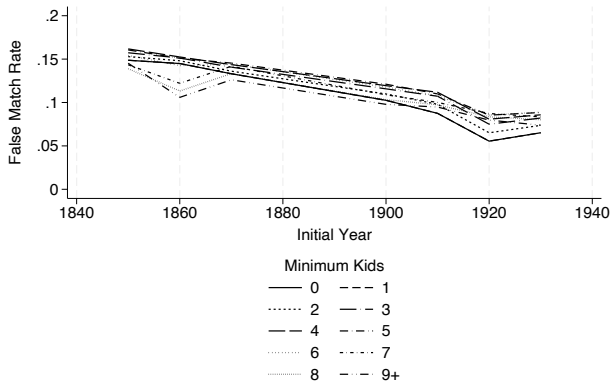


Figure B.3: Robustness of false match rate estimates to changes in the minimum number of children

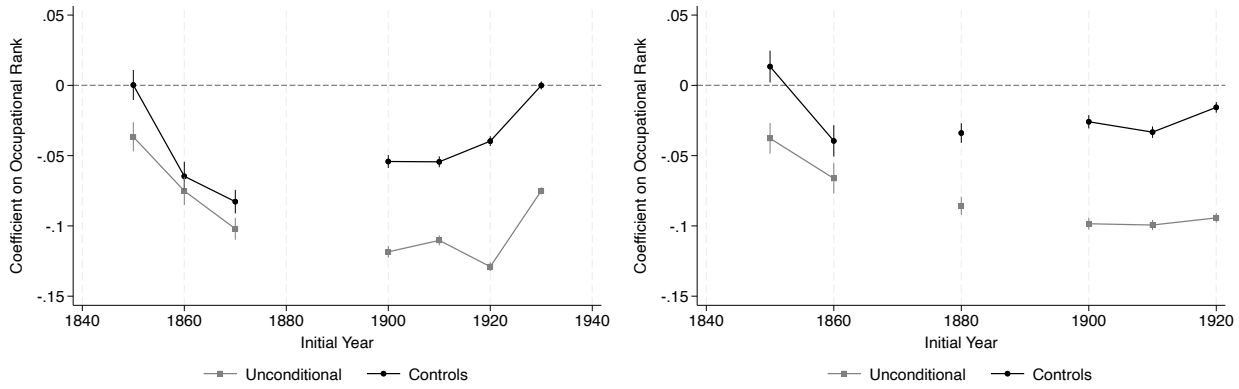
Note: Each line presents the time series of estimated false match rates where individuals are required to meet a threshold of the number of children in their household in order to be included in the corroboration sample.

C Selection and sorting results with alternative linkage methods

C.1 ABE-Exact

(a) Occupational rank, 10-year spans

(b) Occupational rank, 20-year spans



(c) Literacy, 10-year spans

(d) Literacy, 20-year spans

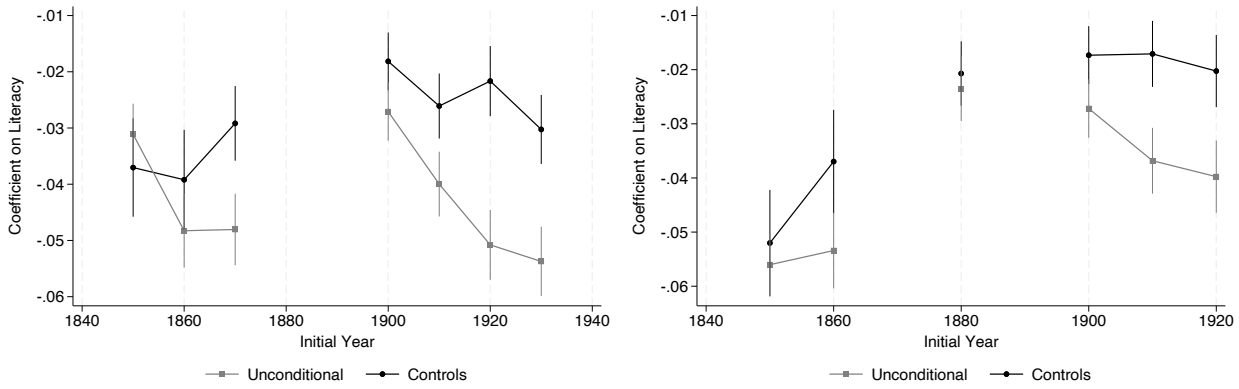
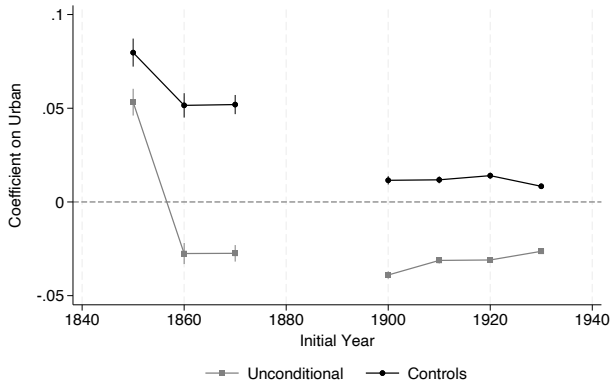


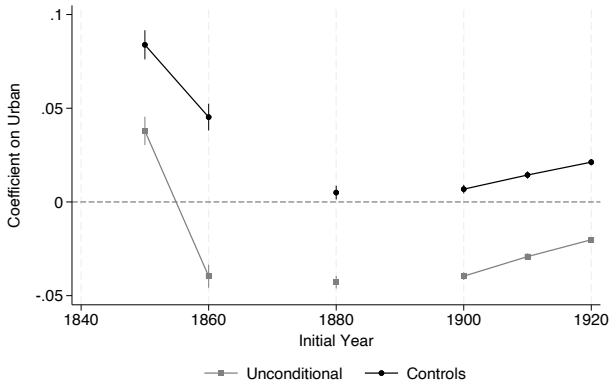
Figure C.1.1: Migrant selection by measure and span, individual characteristics

Note: Each figure shows the coefficient on the variable in question in a selection regression for migration over the stated span. All observations are weighted by inverse linkage probability. Vertical lines indicate 95-percent confidence intervals. Controls are all controls available in all census years, as well as county fixed effects.

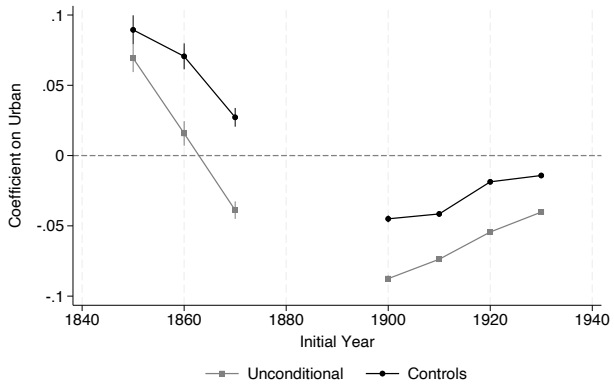
(a) Urban (2,500+), 10-year spans



(b) Urban (2,500+), 20-year spans



(c) Urban (25,000+), 10-year spans



(d) Urban (25,000+), 20-year spans

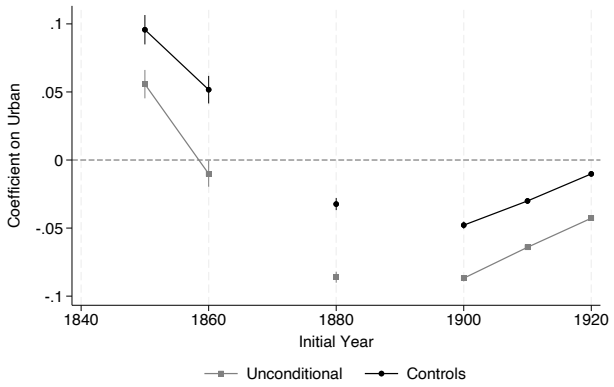
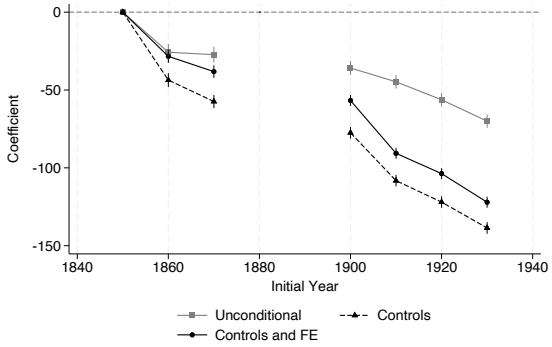


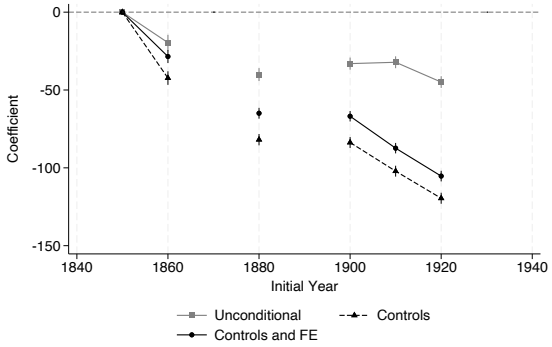
Figure C.1.2: Migrant selection by measure and span, urban residence

Note: Each figure shows the coefficient on the variable in question in a selection regression for migration over the stated span. All observations are weighted by inverse linkage probability. Vertical lines indicate 95-percent confidence intervals. Controls are all controls available in all census years, as well as state fixed effects.

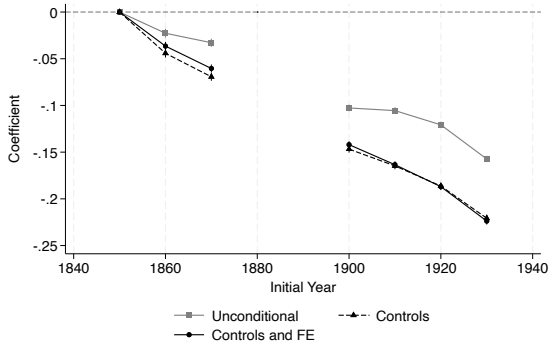
(a) Distance of move (miles), 10-year spans



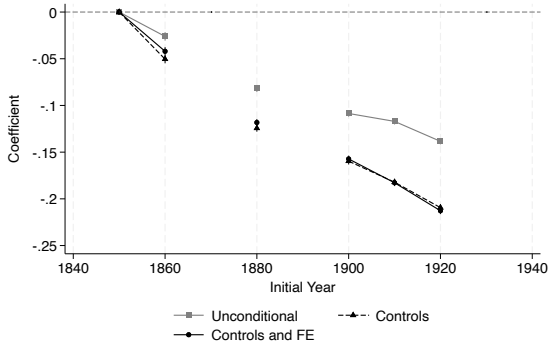
(b) Distance of move (miles), 20-year spans



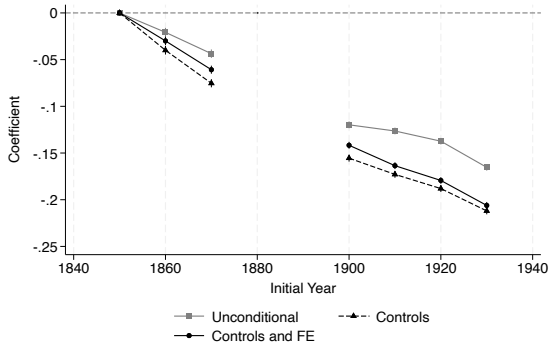
(c) Interstate move, 10-year spans



(d) Interstate move, 20-year spans



(e) Interregion move, 10-year spans



(f) Interregion move, 20-year spans

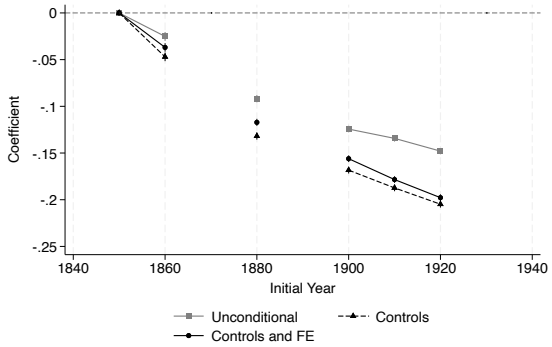


Figure C.1.3: Changes in distance of move

Note: These figures show the coefficients on initial-year fixed effects in regressions of destination characteristics with and without controls, with 1850 as the excluded year. "Controls" includes all controls available in all censuses, including state fixed effects. "Controls and FE' includes also county fixed effects. Vertical lines are 95-percent confidence intervals. Observations weighted to correct for selection into linkage.

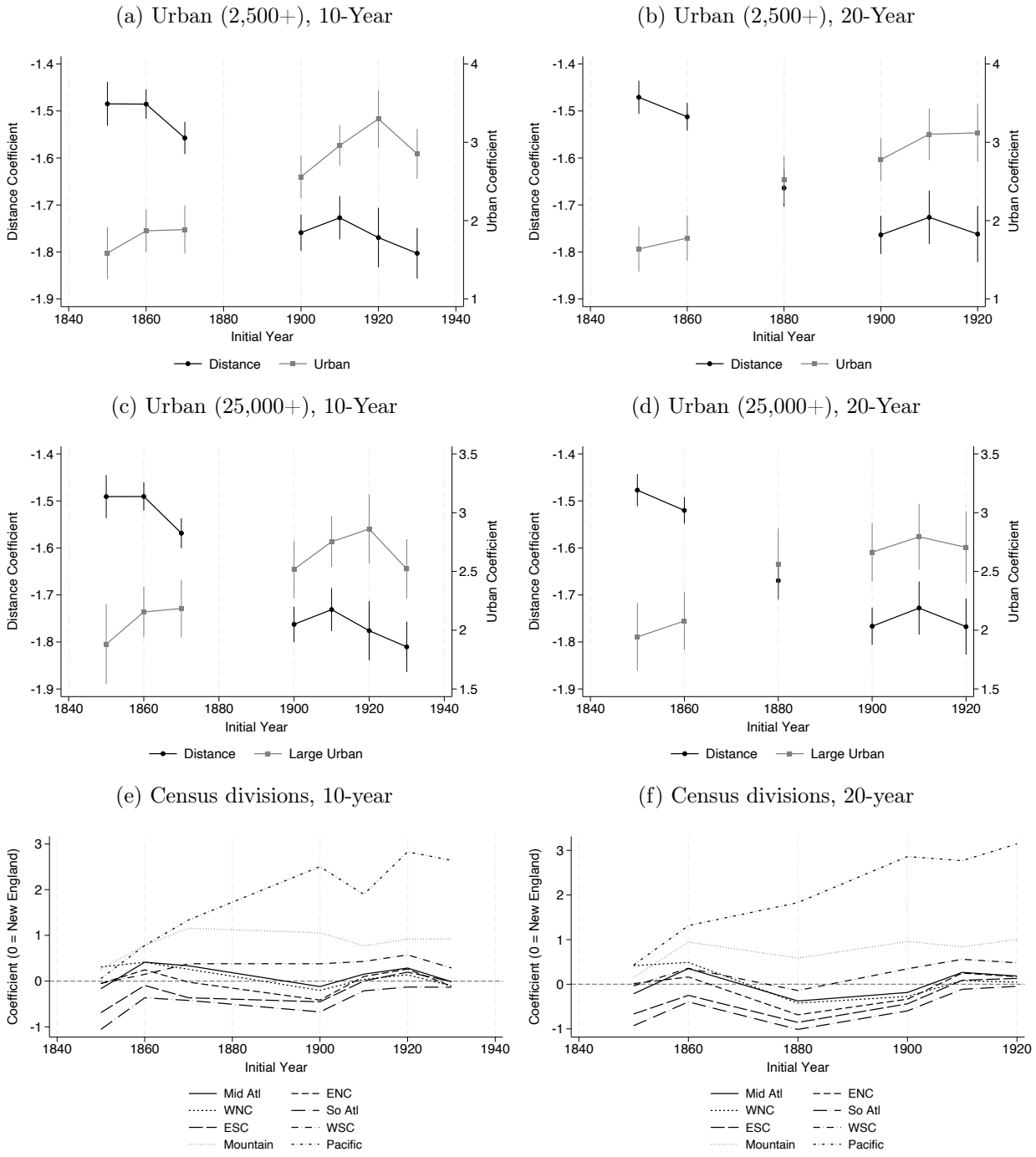


Figure C.1.4: Conditional Logit Results

Note: Panels (a)–(d) present coefficients on distance and the measure of urban residence from the conditional logit estimation. Panels (e) and (f) present the coefficients from the regressions of panels (a) and (b), respectively, on the census division indicators, with New England excluded. Note that there is no data for spans that would include 1890; the lines span this period simply to enable clearer reading of the trend over time.

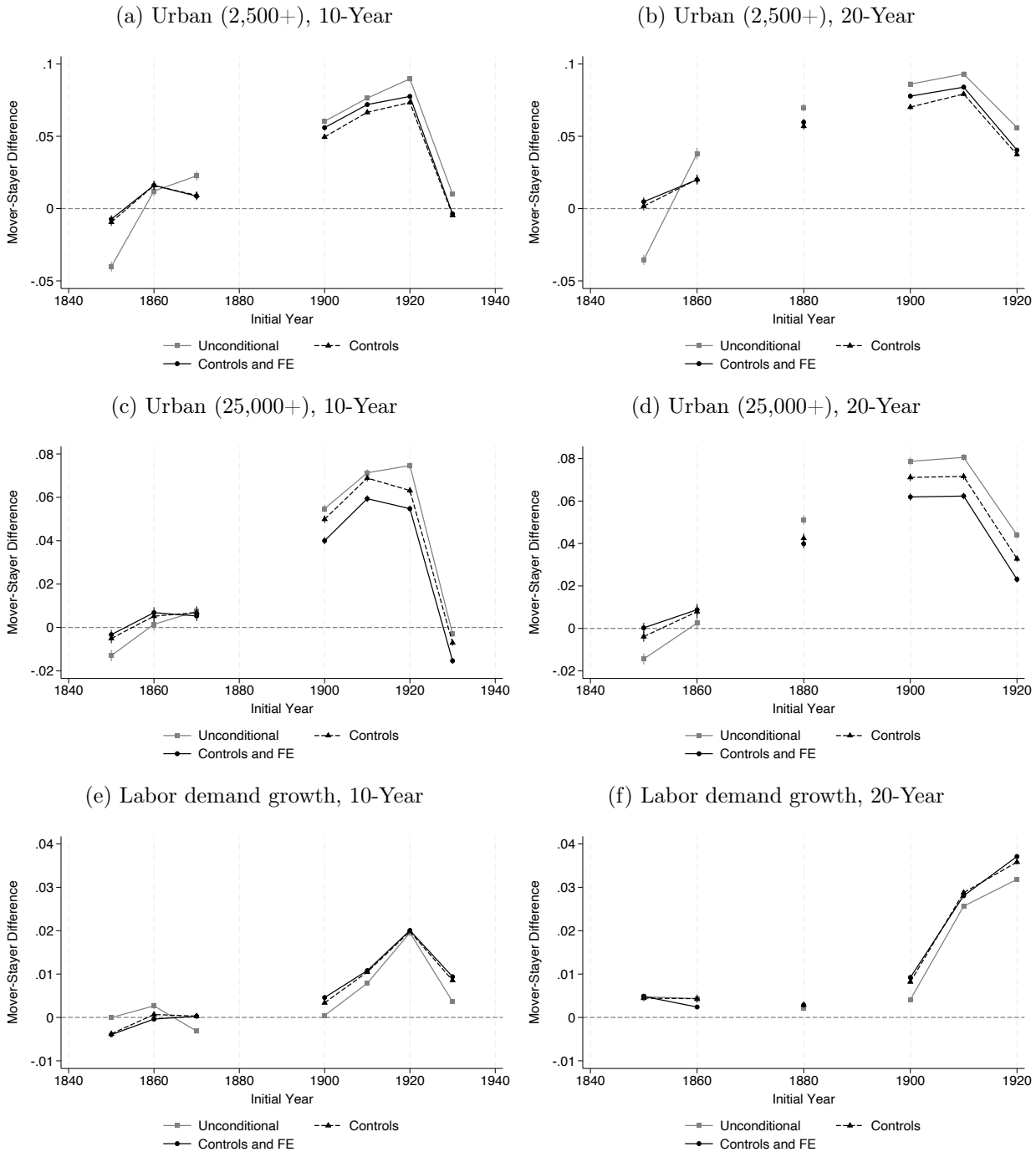
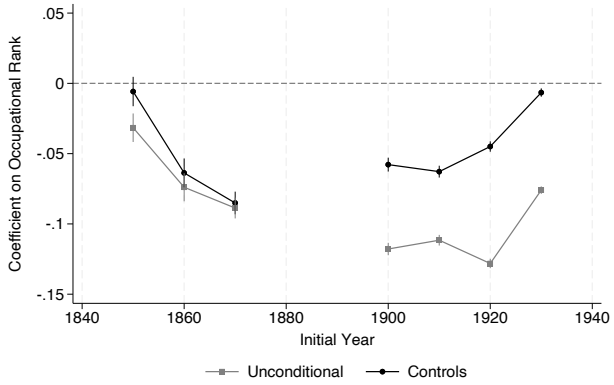


Figure C.1.5: Movers' change in urbanization and labor demand growth

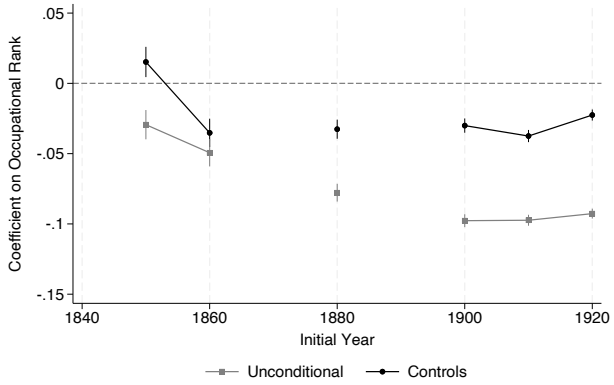
Note: This figure compares the growth in urbanization or labor demand experienced by movers to that experienced by stayers, with either no controls, controlling for all observables, and controlling for all observables and county fixed effects.

C.2 ABE-NYSIIS

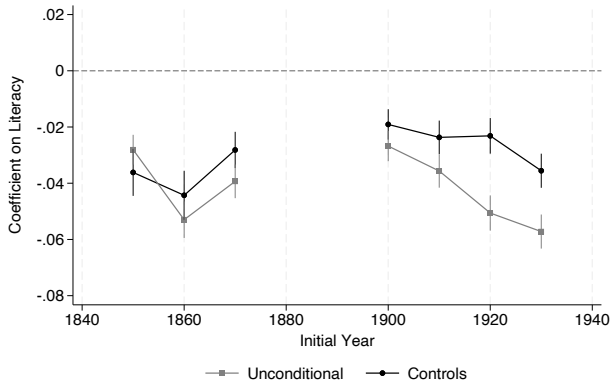
(a) Occupational rank, 10-year spans



(b) Occupational rank, 20-year spans



(c) Literacy, 10-year spans



(d) Literacy, 20-year spans

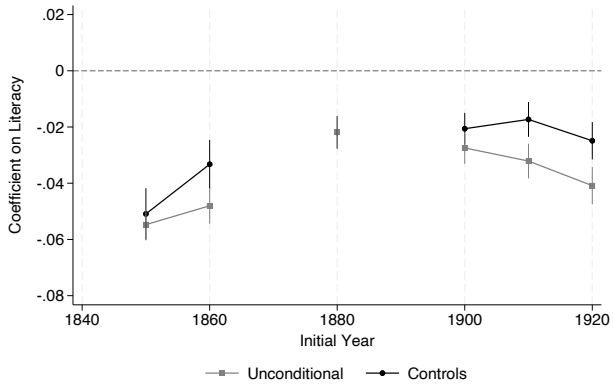
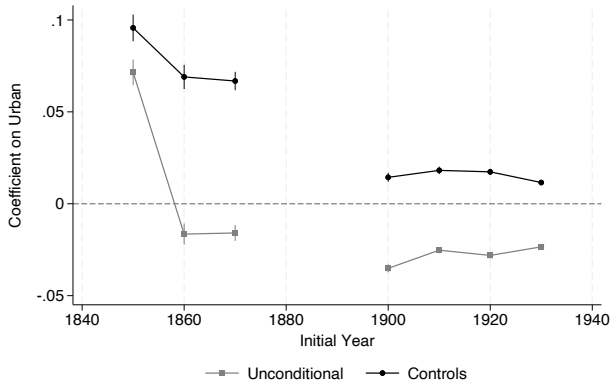


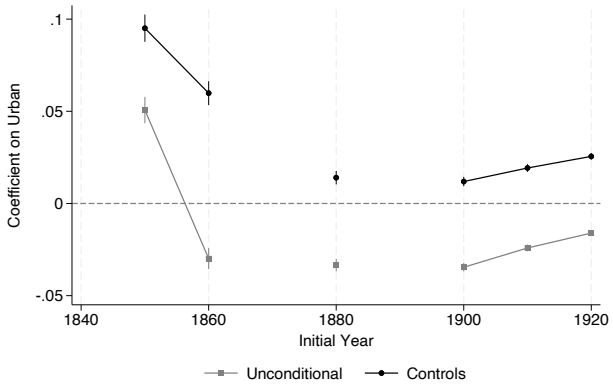
Figure C.2.1: Migrant selection by measure and span, individual characteristics

Note: Each figure shows the coefficient on the variable in question in a selection regression for migration over the stated span. All observations are weighted by inverse linkage probability. Vertical lines indicate 95-percent confidence intervals. Controls are all controls available in all census years, as well as county fixed effects.

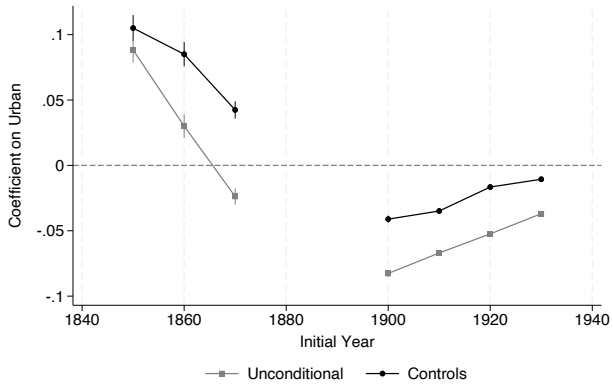
(a) Urban (2,500+), 10-year spans



(b) Urban (2,500+), 20-year spans



(c) Urban (25,000+), 10-year spans



(d) Urban (25,000+), 20-year spans

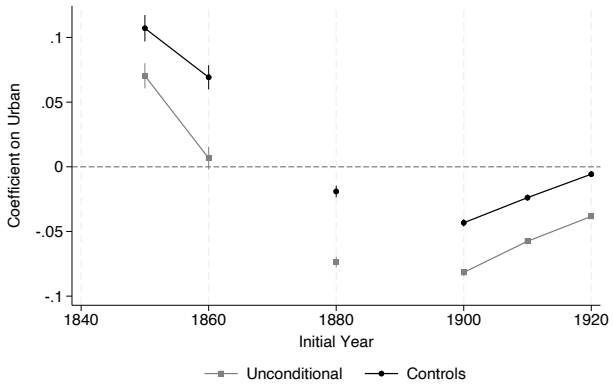
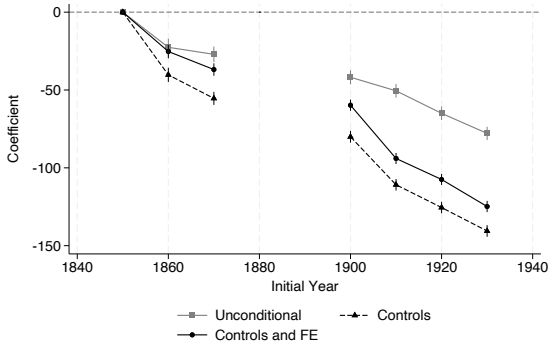


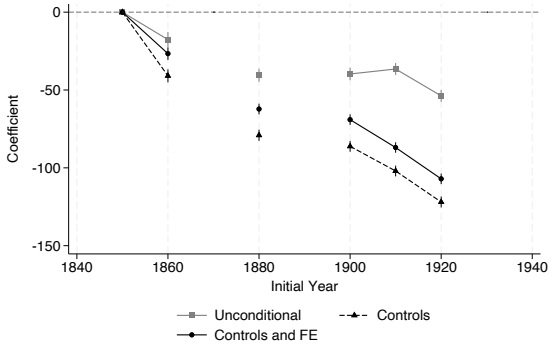
Figure C.2.2: Migrant selection by measure and span, urban residence

Note: Each figure shows the coefficient on the variable in question in a selection regression for migration over the stated span. All observations are weighted by inverse linkage probability. Vertical lines indicate 95-percent confidence intervals. Controls are all controls available in all census years, as well as state fixed effects.

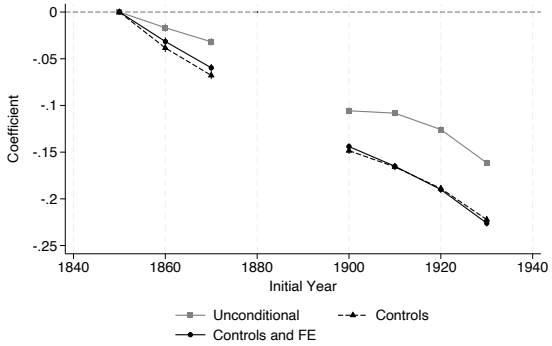
(a) Distance of move (miles), 10-year spans



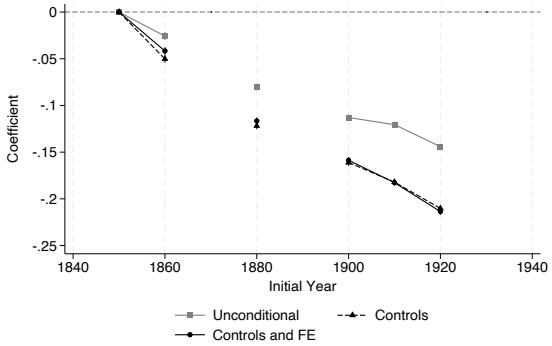
(b) Distance of move (miles), 20-year spans



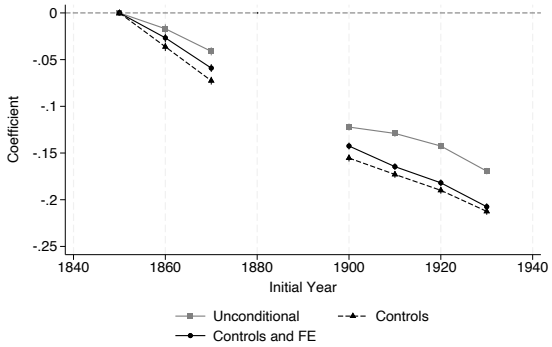
(c) Interstate move, 10-year spans



(d) Interstate move, 20-year spans



(e) Interregion move, 10-year spans



(f) Interregion move, 20-year spans

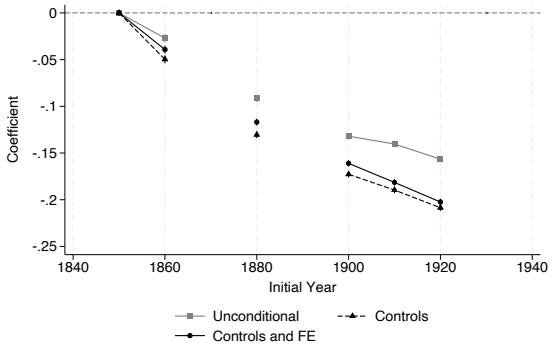


Figure C.2.3: Changes in distance of move

Note: These figures show the coefficients on initial-year fixed effects in regressions of destination characteristics with and without controls, with 1850 as the excluded year. "Controls" includes all controls available in all censuses, including state fixed effects. "Controls and FE' includes also county fixed effects. Vertical lines are 95-percent confidence intervals. Observations weighted to correct for selection into linkage.

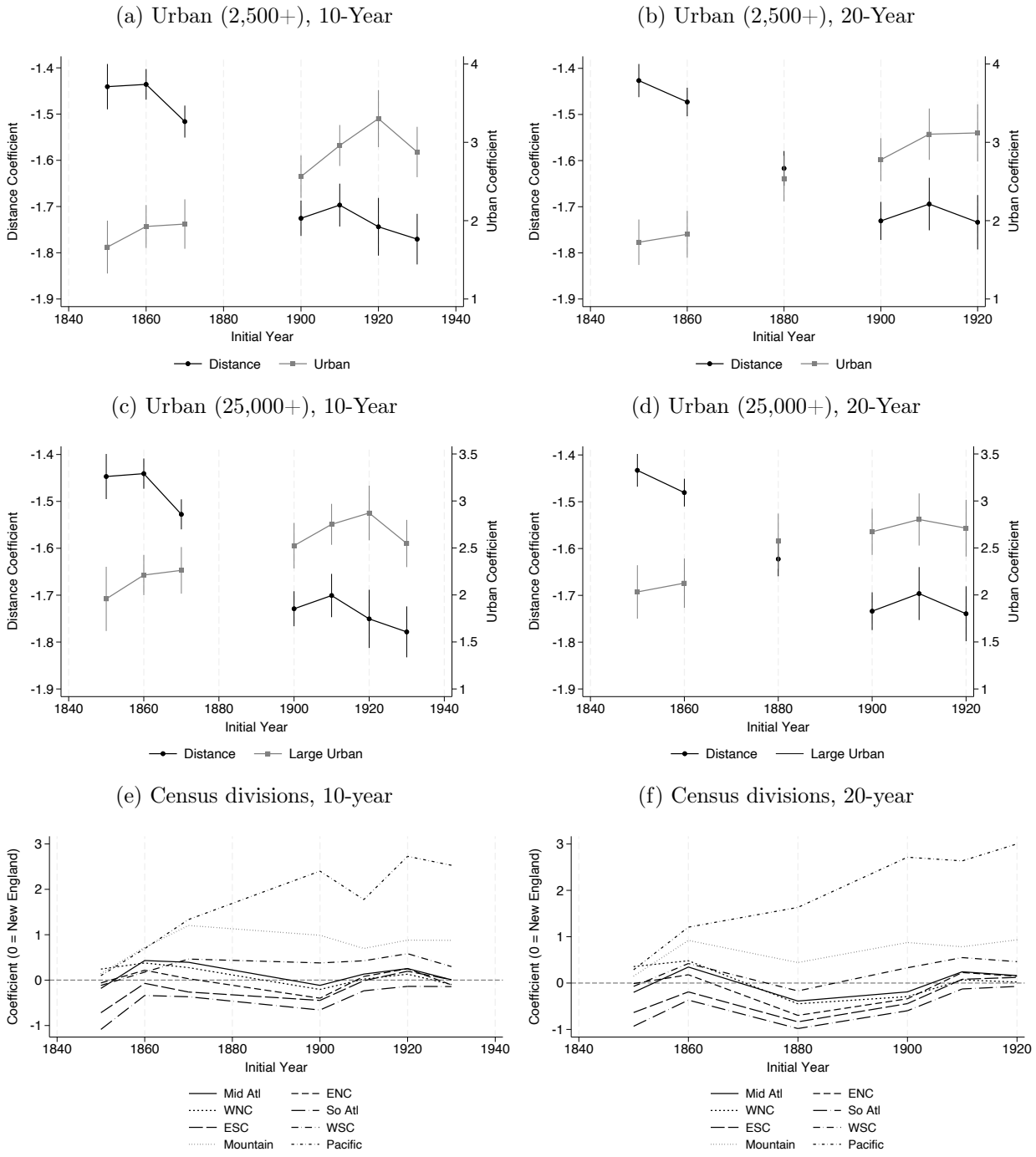
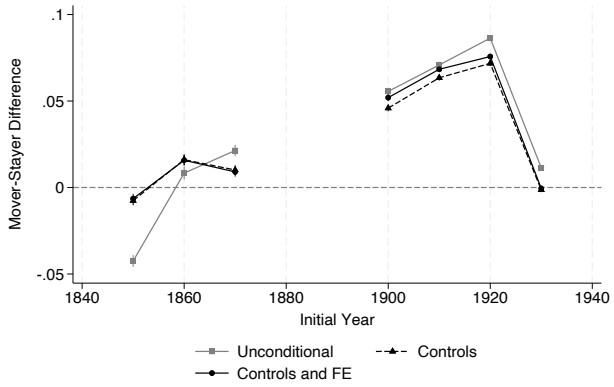


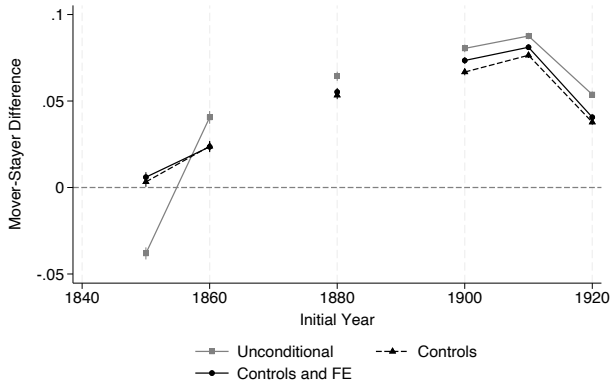
Figure C.2.4: Conditional Logit Results

Note: Panels (a)–(d) present coefficients on distance and the measure of urban residence from the conditional logit estimation. Panels (e) and (f) present the coefficients from the regressions of panels (a) and (b), respectively, on the census division indicators, with New England excluded. Note that there is no data for spans that would include 1890; the lines span this period simply to enable clearer reading of the trend over time.

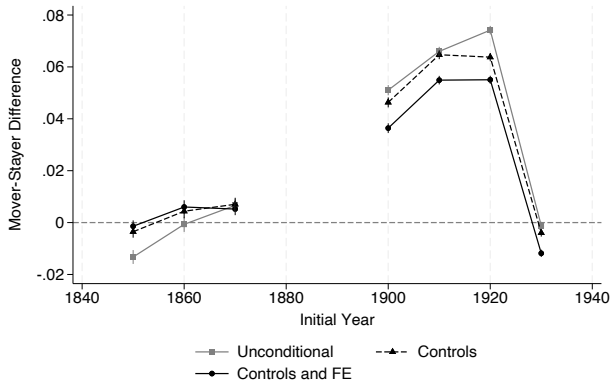
(a) Urban (2,500+), 10-Year



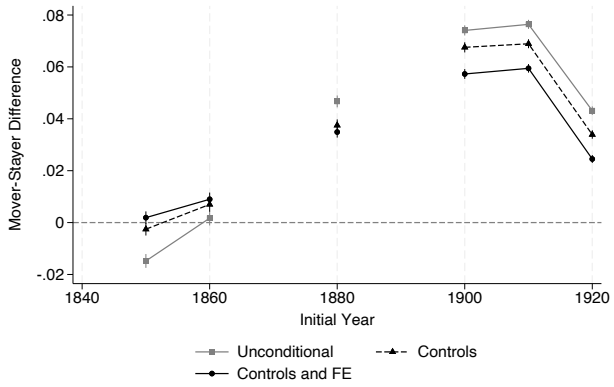
(b) Urban (2,500+), 20-Year



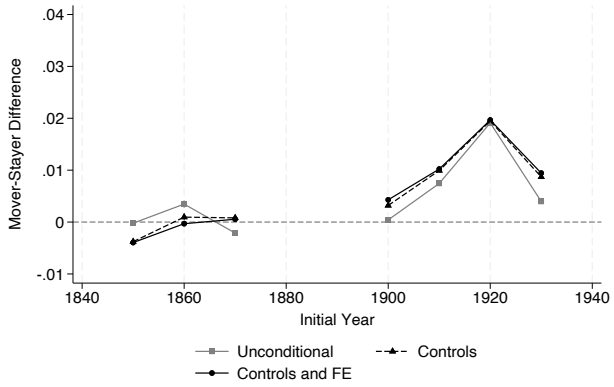
(c) Urban (25,000+), 10-Year



(d) Urban (25,000+), 20-Year



(e) Labor demand growth, 10-Year



(f) Labor demand growth, 20-Year

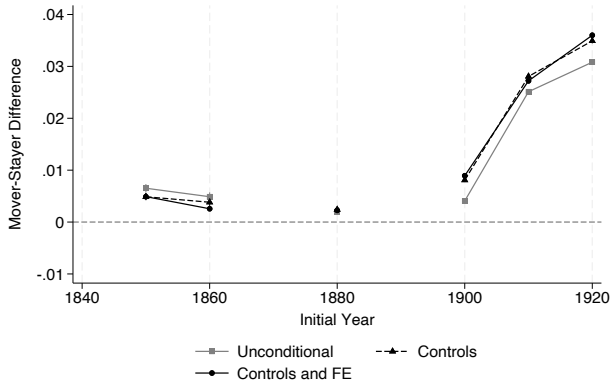


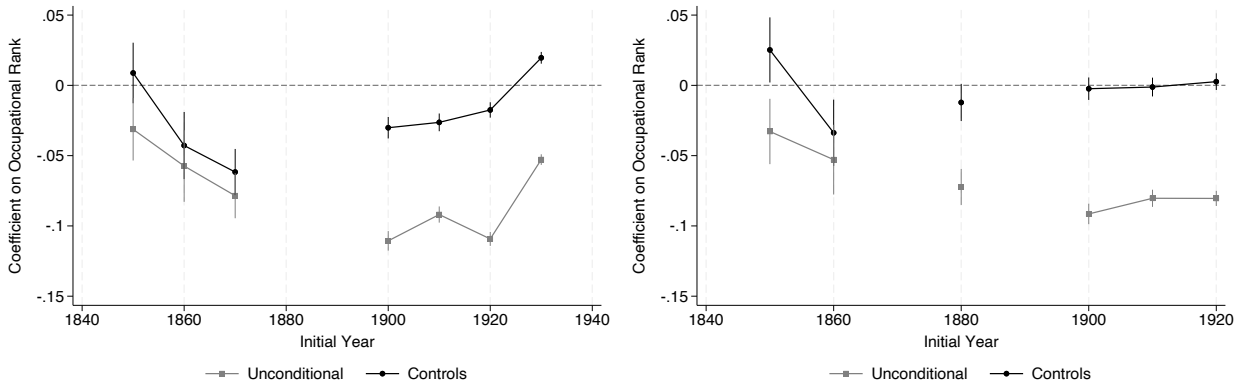
Figure C.2.5: Movers' change in urbanization and labor demand growth

Note: This figure compares the growth in urbanization or labor demand experienced by movers to that experienced by stayers, with either no controls, controlling for all observables, and controlling for all observables and county fixed effects.

C.3 Intersection of matches

(a) Occupational rank, 10-year spans

(b) Occupational rank, 20-year spans



(c) Literacy, 10-year spans

(d) Literacy, 20-year spans

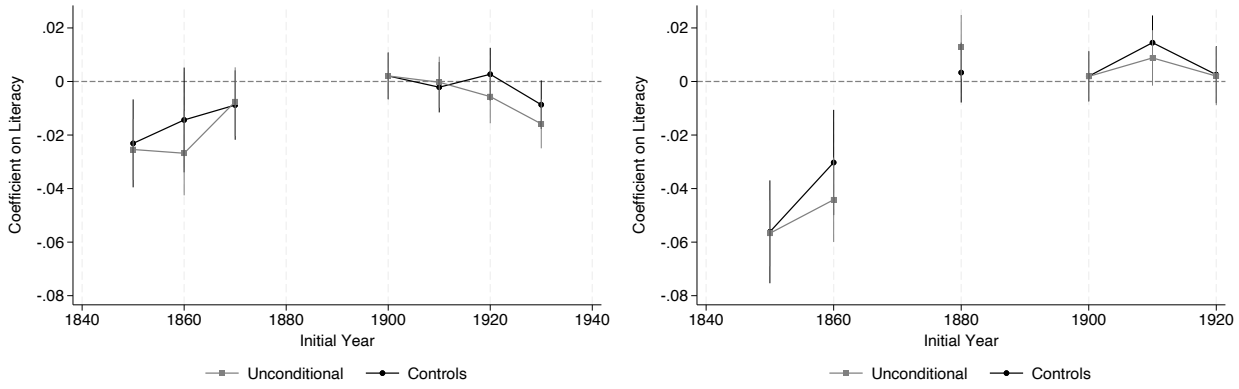
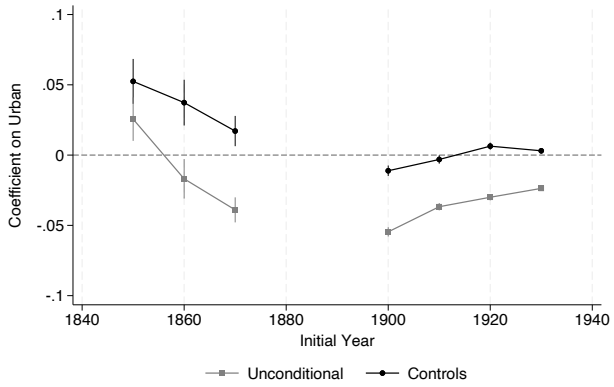


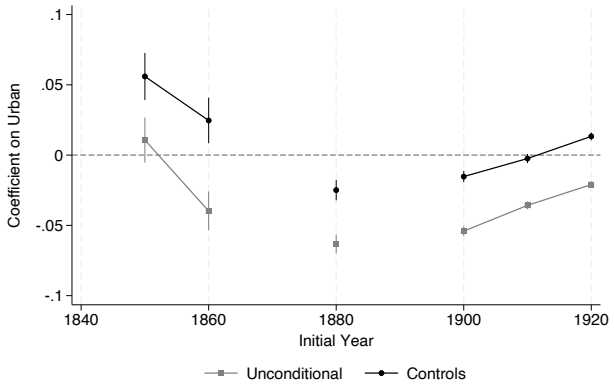
Figure C.3.1: Migrant selection by measure and span individual characteristics

Note: Each figure shows the coefficient on the variable in question in a selection regression for migration over the stated span. All observations are weighted by inverse linkage probability. Vertical lines indicate 95-percent confidence intervals. Controls are all controls available in all census years, as well as county fixed effects.

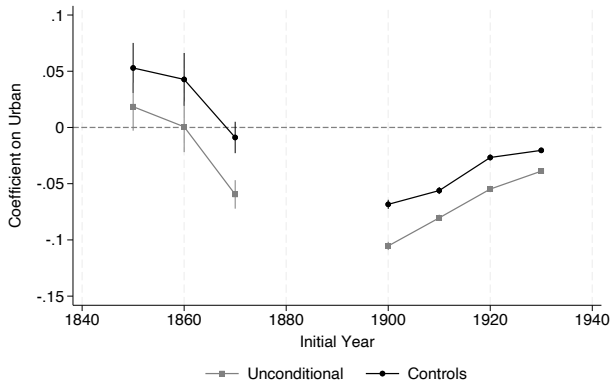
(a) Urban (2,500+), 10-year spans



(b) Urban (2,500+), 20-year spans



(c) Urban (25,000+), 10-year spans



(d) Urban (25,000+), 20-year spans

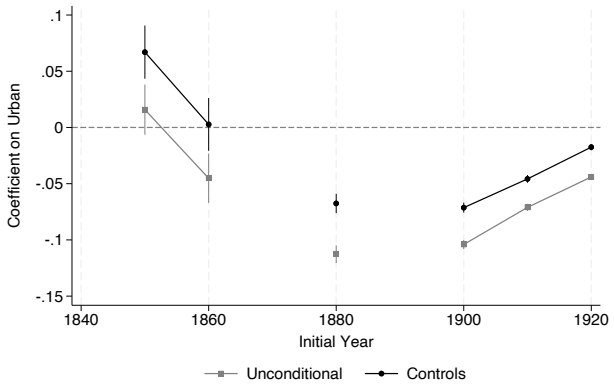
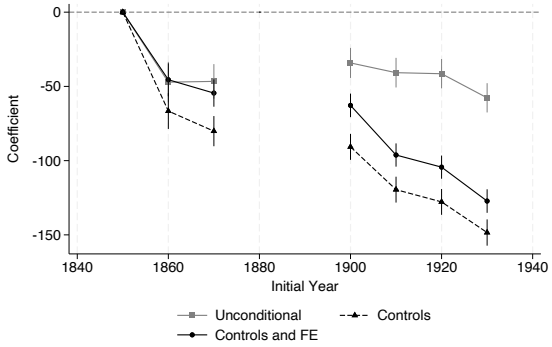


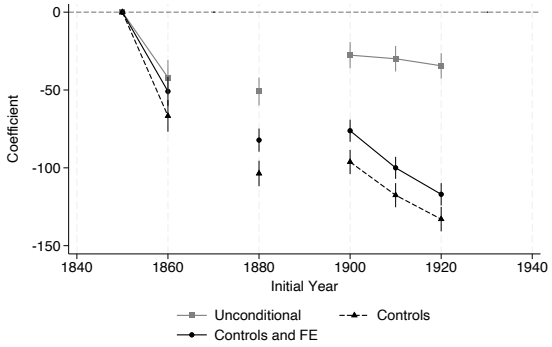
Figure C.3.2: Migrant selection by measure and span, urban residence

Note: Each figure shows the coefficient on the variable in question in a selection regression for migration over the stated span. All observations are weighted by inverse linkage probability. Vertical lines indicate 95-percent confidence intervals. Controls are all controls available in all census years, as well as state fixed effects.

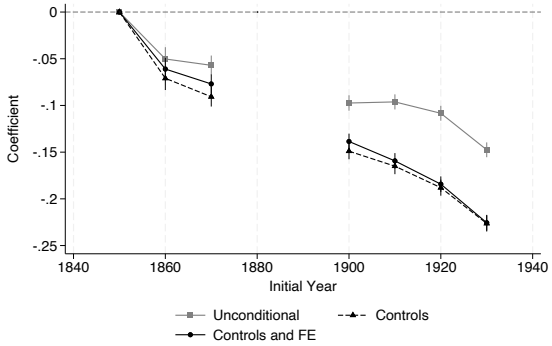
(a) Distance of move (miles), 10-year spans



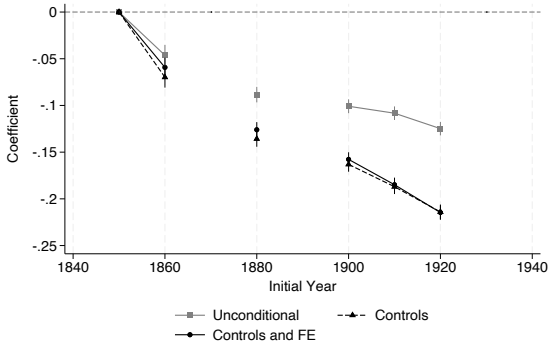
(b) Distance of move (miles), 20-year spans



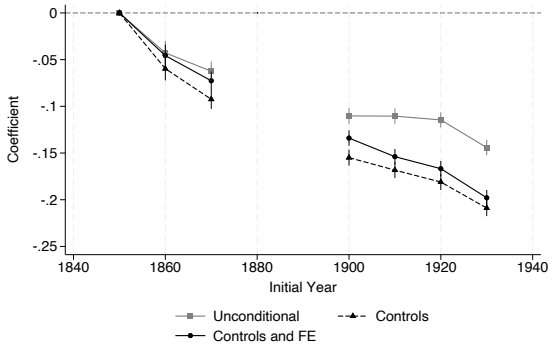
(c) Interstate move, 10-year spans



(d) Interstate move, 20-year spans



(e) Interregion move, 10-year spans



(f) Interregion move, 20-year spans

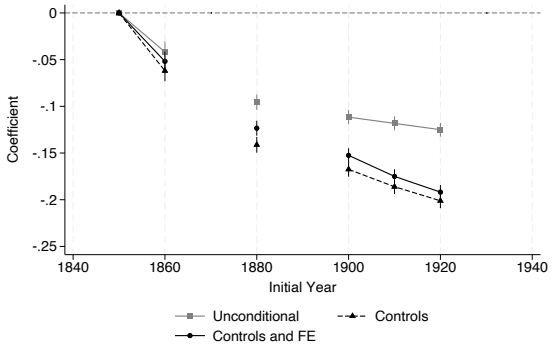
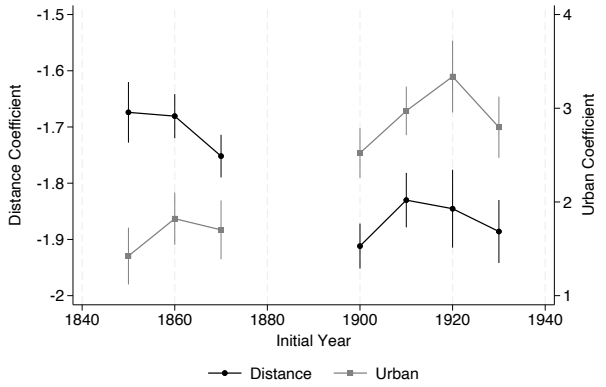


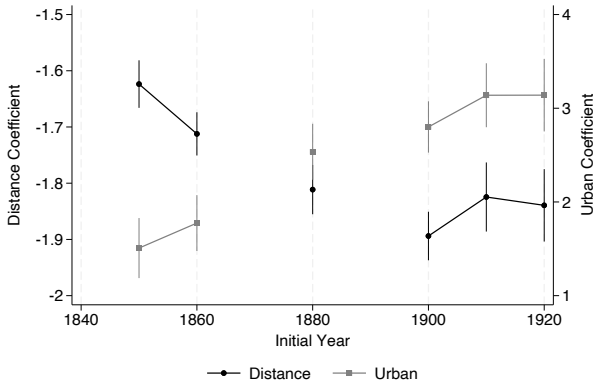
Figure C.3.3: Changes in distance of move

Note: These figures show the coefficients on initial-year fixed effects in regressions of destination characteristics with and without controls, with 1850 as the excluded year. "Controls" includes all controls available in all censuses, including state fixed effects. "Controls and FE' includes also county fixed effects. Vertical lines are 95-percent confidence intervals. Observations weighted to correct for selection into linkage.

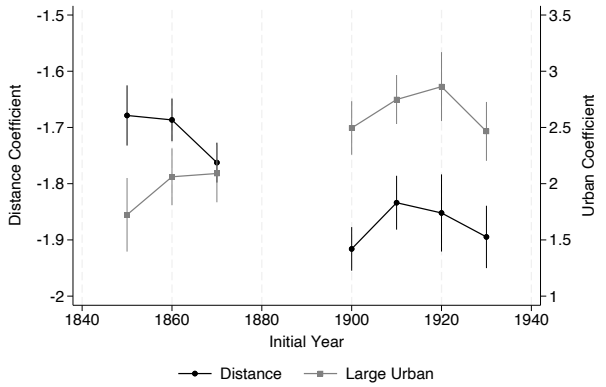
(a) Urban (2,500+), 10-Year



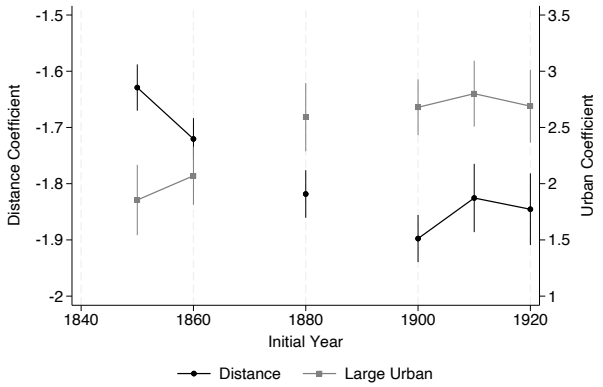
(b) Urban (2,500+), 20-Year



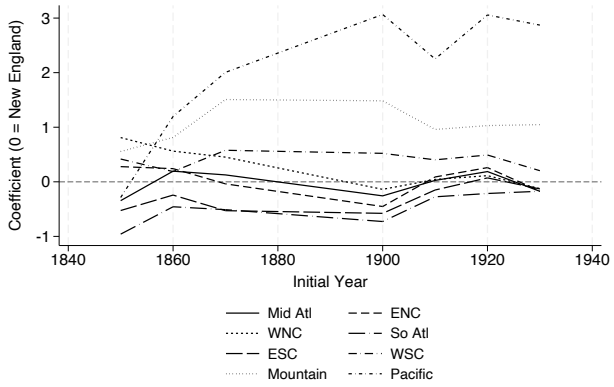
(c) Urban (25,000+), 10-Year



(d) Urban (25,000+), 20-Year



(e) Census divisions, 10-year



(f) Census divisions, 20-year

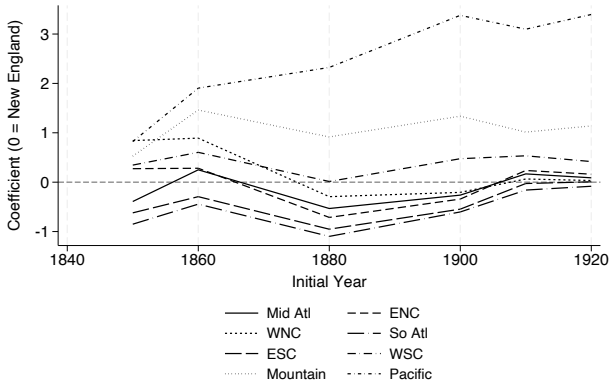


Figure C.3.4: Conditional Logit Results

Note: Panels (a)–(d) present coefficients on distance and the measure of urban residence from the conditional logit estimation. Panels (e) and (f) present the coefficients from the regressions of panels (a) and (b), respectively, on the census division indicators, with New England excluded. Note that there is no data for spans that would include 1890; the lines span this period simply to enable clearer reading of the trend over time.

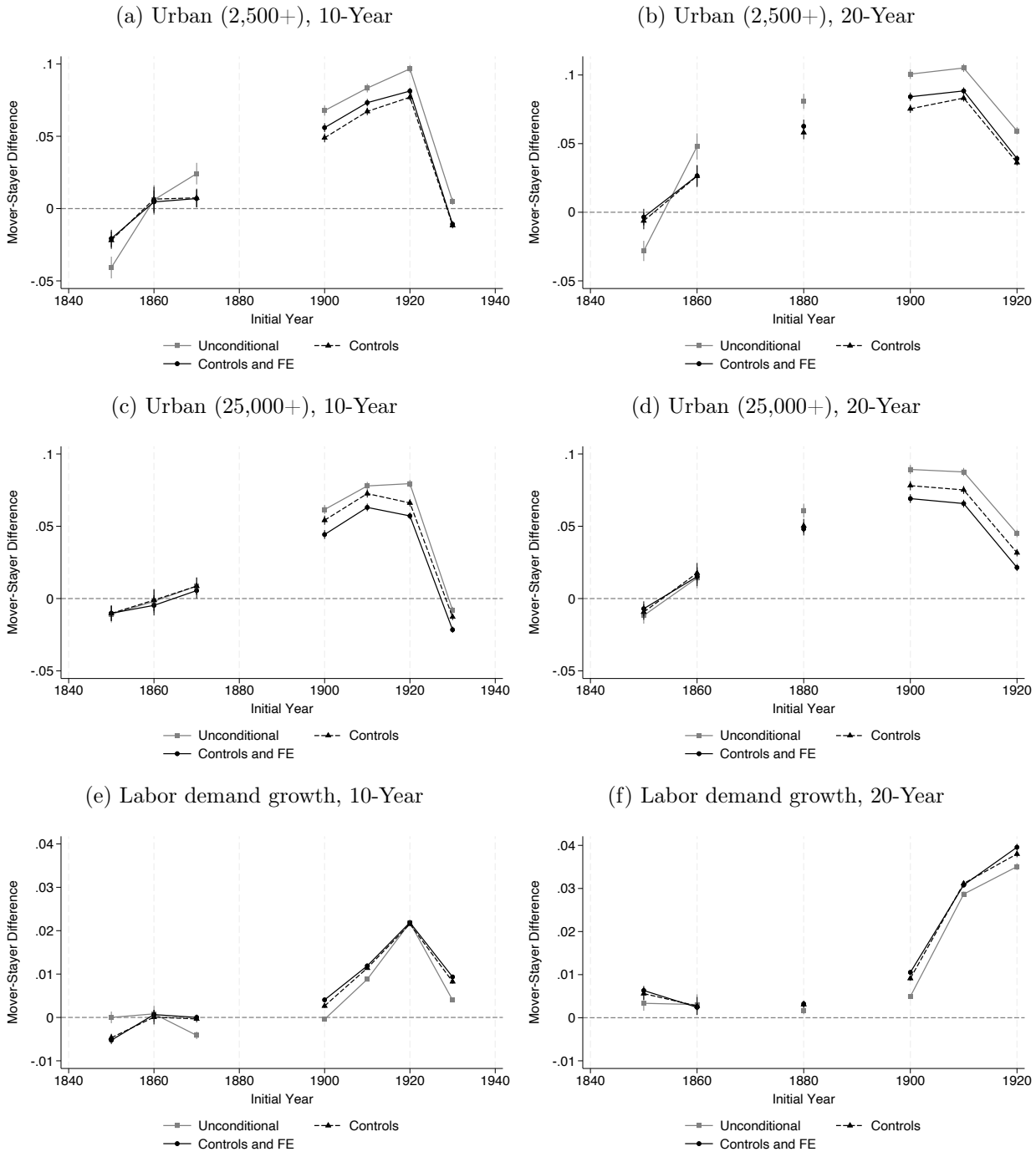


Figure C.3.5: Movers' change in urbanization and labor demand growth

Note: This figure compares the growth in urbanization or labor demand experienced by movers to that experienced by stayers, with either no controls, controlling for all observables, and controlling for all observables and county fixed effects.

C.4 Intersection of matches plus corroboration

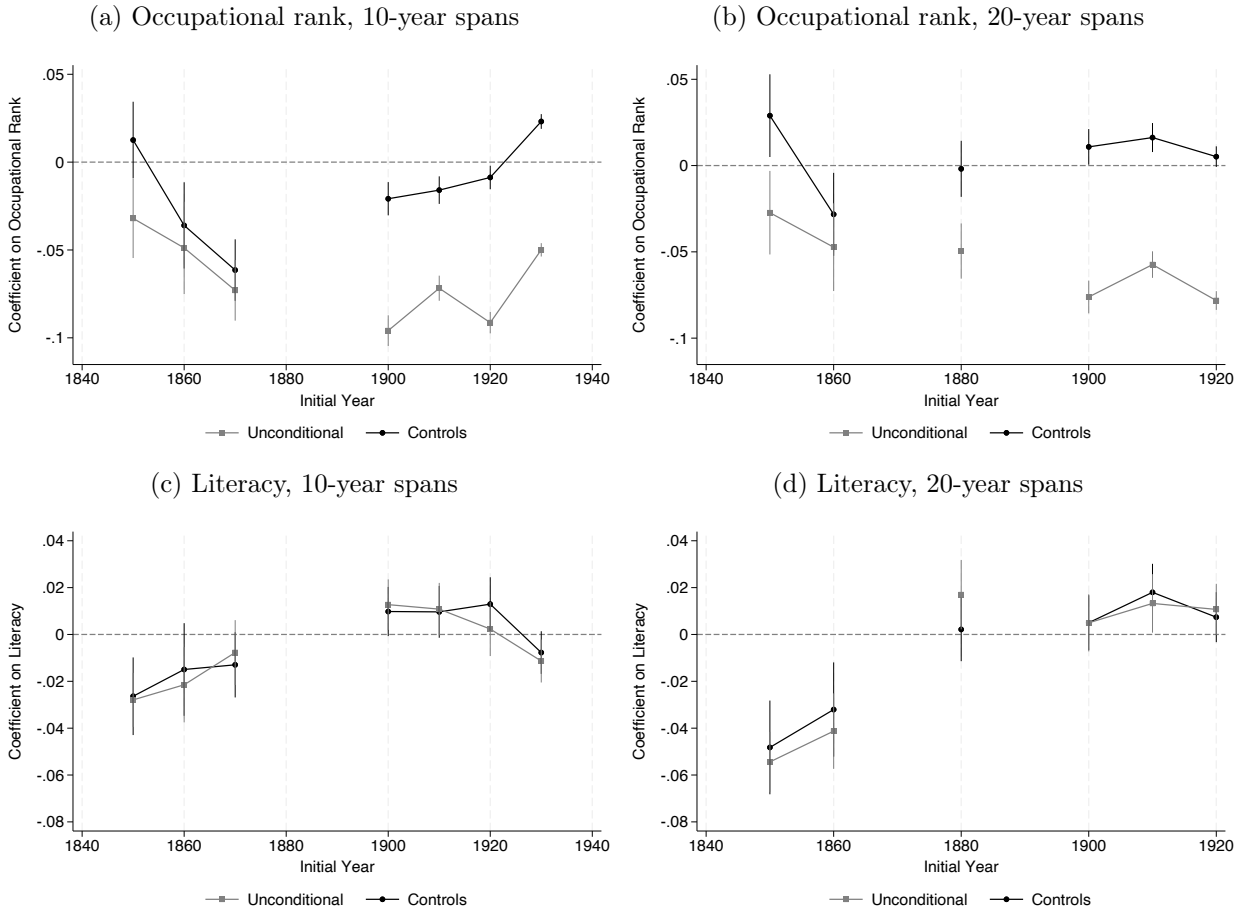
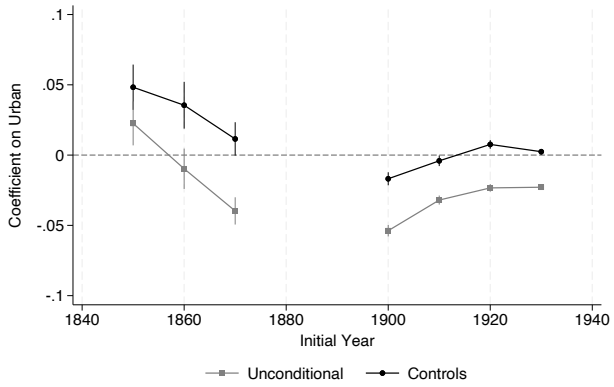


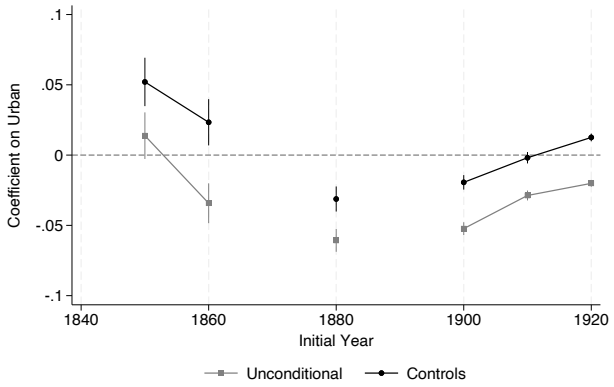
Figure C.4.6: Migrant selection by measure and span, individual characteristics

Note: Each figure shows the coefficient on the variable in question in a selection regression for migration over the stated span. All observations are weighted by inverse linkage probability. Vertical lines indicate 95-percent confidence intervals. Controls are all controls available in all census years, as well as county fixed effects.

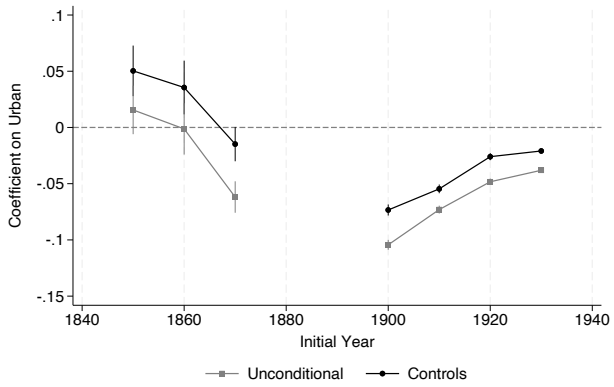
(a) Urban (2,500+), 10-year spans



(b) Urban (2,500+), 20-year spans



(c) Urban (25,000+), 10-year spans



(d) Urban (25,000+), 20-year spans

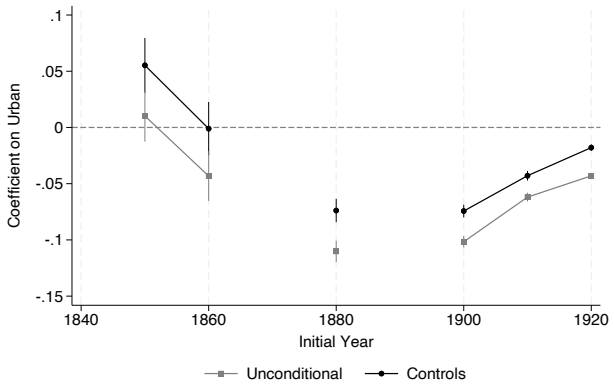
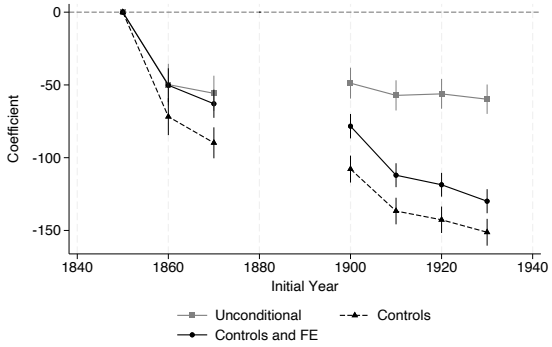


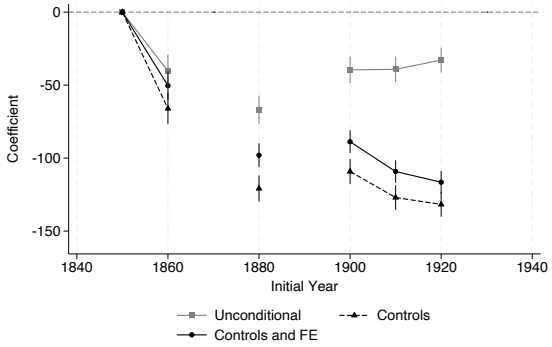
Figure C.4.7: Migrant selection by measure and span, urban residence

Note: Each figure shows the coefficient on the variable in question in a selection regression for migration over the stated span. All observations are weighted by inverse linkage probability. Vertical lines indicate 95-percent confidence intervals. Controls are all controls available in all census years, as well as state fixed effects.

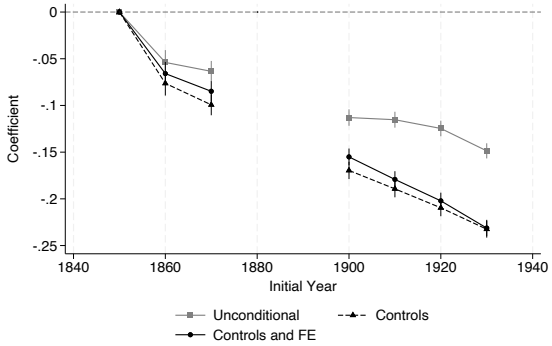
(a) Distance of move (miles), 10-year spans



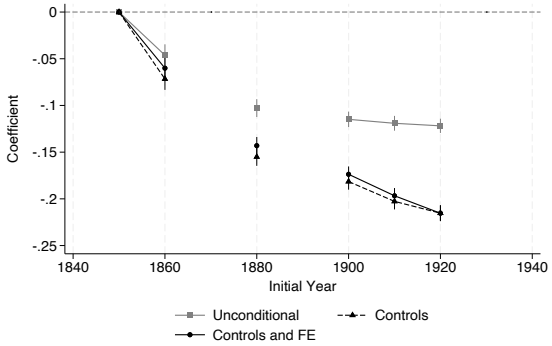
(b) Distance of move (miles), 20-year spans



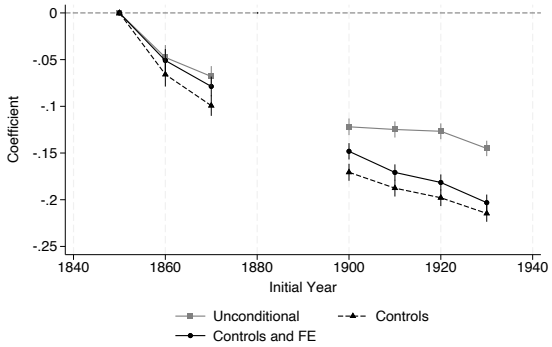
(c) Interstate move, 10-year spans



(d) Interstate move, 20-year spans



(e) Interregion move, 10-year spans



(f) Interregion move, 20-year spans

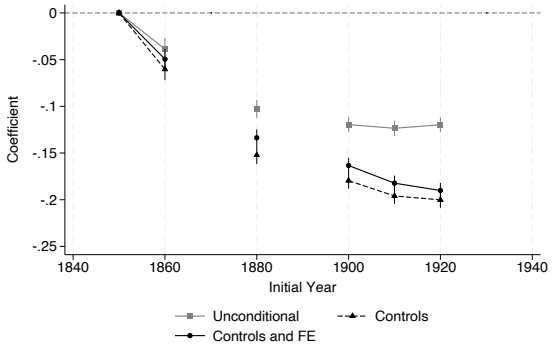
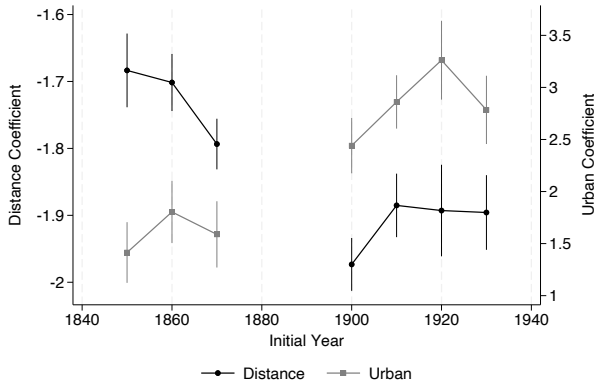


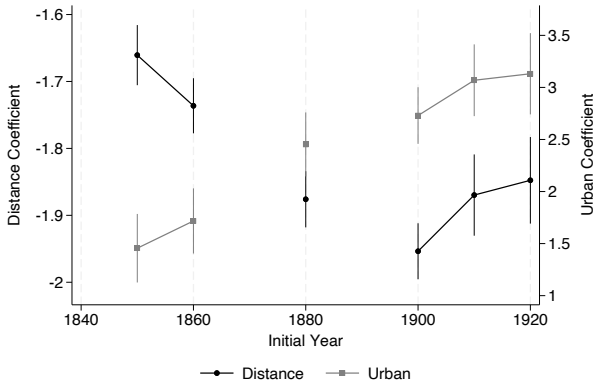
Figure C.4.8: Changes in distance of move

Note: These figures show the coefficients on initial-year fixed effects in regressions of destination characteristics with and without controls, with 1850 as the excluded year. "Controls" includes all controls available in all censuses, including state fixed effects. "Controls and FE' includes also county fixed effects. Vertical lines are 95-percent confidence intervals. Observations weighted to correct for selection into linkage.

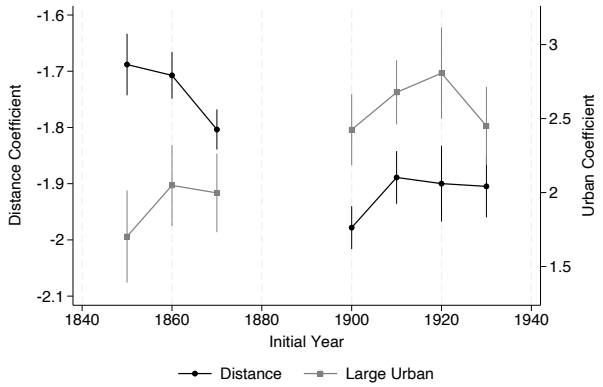
(a) Urban (2,500+), 10-Year



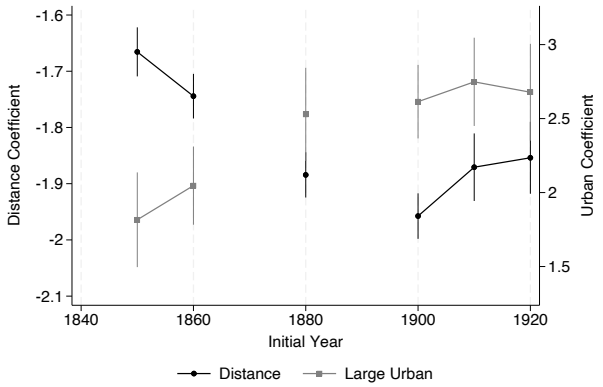
(b) Urban (2,500+), 20-Year



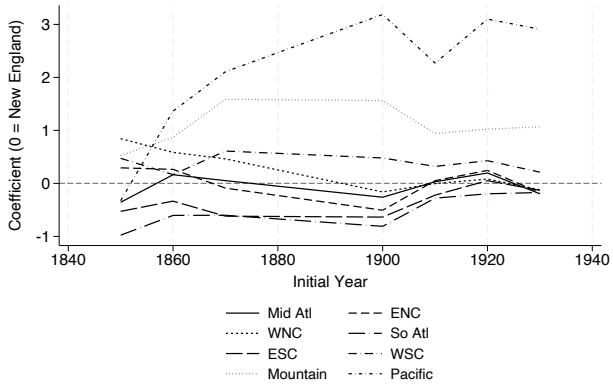
(c) Urban (25,000+), 10-Year



(d) Urban (25,000+), 20-Year



(e) Census divisions, 10-year



(f) Census divisions, 20-year

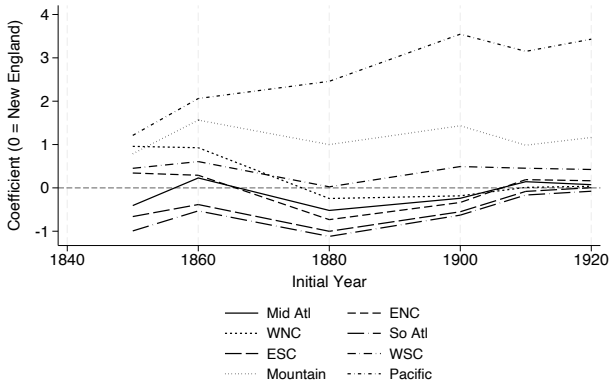
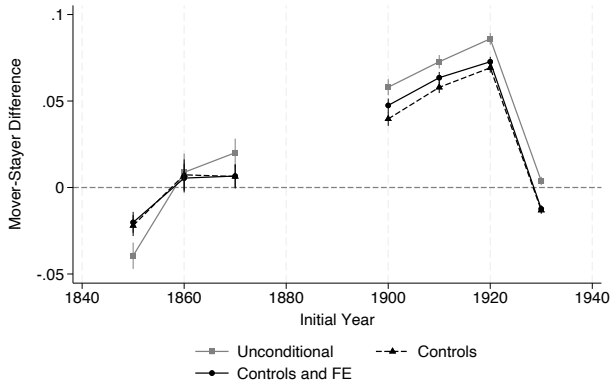


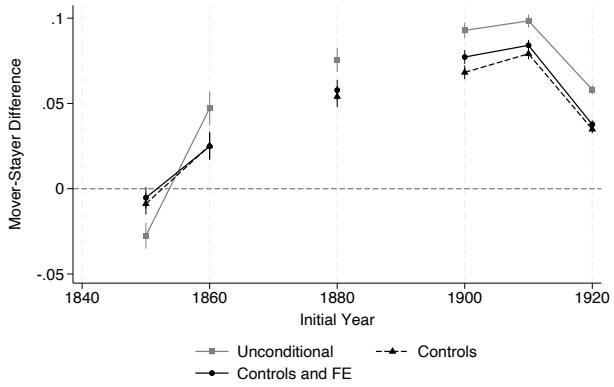
Figure C.4.9: Conditional Logit Results

Note: Panels (a)–(d) present coefficients on distance and the measure of urban residence from the conditional logit estimation. Panels (e) and (f) present the coefficients from the regressions of panels (a) and (b), respectively, on the census division indicators, with New England excluded. Note that there is no data for spans that would include 1890; the lines span this period simply to enable clearer reading of the trend over time.

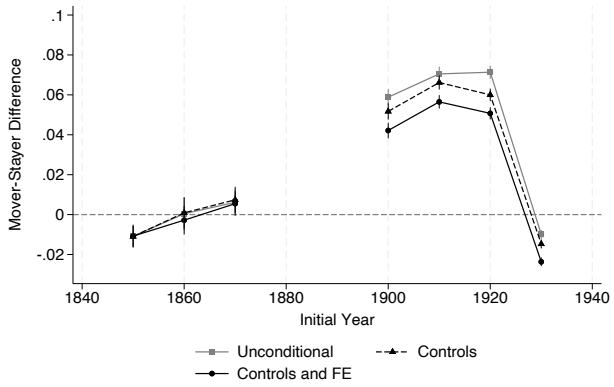
(a) Urban (2,500+), 10-Year



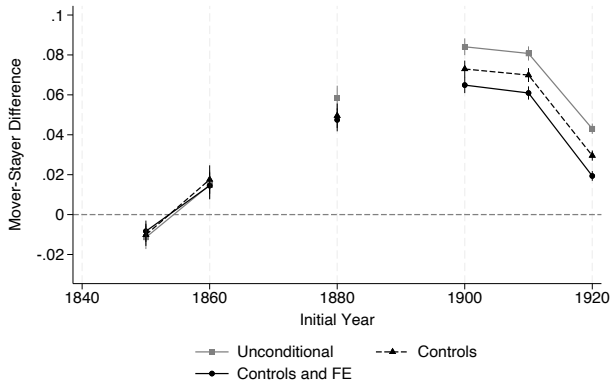
(b) Urban (2,500+), 20-Year



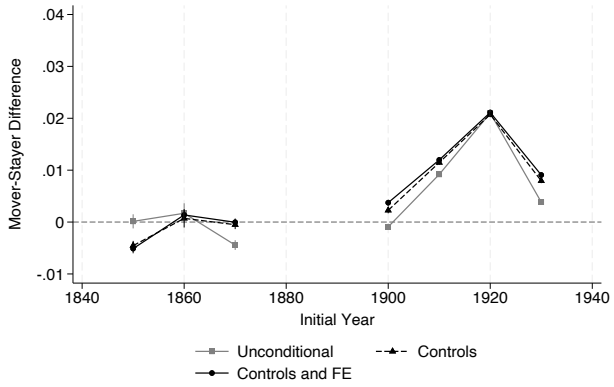
(c) Urban (25,000+), 10-Year



(d) Urban (25,000+), 20-Year



(e) Labor demand growth, 10-Year



(f) Labor demand growth, 20-Year

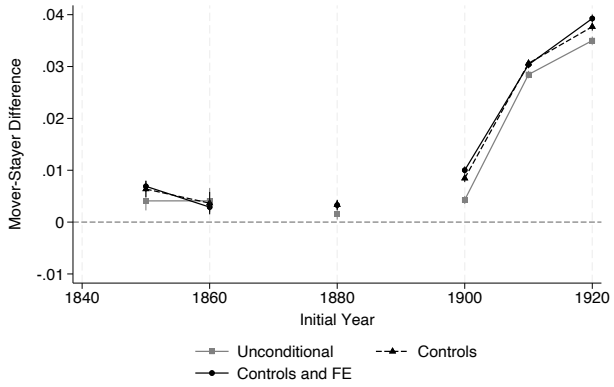


Figure C.4.10: Movers' change in urbanization and labor demand growth

Note: This figure compares the growth in urbanization or labor demand experienced by movers to that experienced by stayers, with either no controls, controlling for all observables, and controlling for all observables and county fixed effects.

D Results for longer-distance moves

In this appendix, I repeat the analysis of the main text, but redefine migration so that moves must be of at least 150 miles. The analysis is the same in all but two respects. First, I present migration rates only using the main linkage method, since the main text showed that estimates were insensitive to the method after correction. Second, for the correction of the estimated rates of migration for false linkage, I do not simplify equation (3) to equation (4) by setting $P(\text{county move}|\text{false match})$ to one. This is an accurate enough approximation in dealing with any inter-county move, but it is not in studying moves of at least 150 miles. The restriction of requiring a match to be to a person with the same birth state limits the probability that a false match is made to a distant person because individuals generally live in or near their state of birth. Thus, in this appendix, the correction is made using the unsimplified equation (3) in which $P(\text{county move}|\text{false match})$ is computed by determining, for each individual in the sample, the fraction of individuals to whom he could be linked (defined as those with the same birth place and an age-implied birthyear sufficiently similar that a match would be permitted) who in the second year of the span lived at least 150 miles away from the individual's county of residence in the first year of the span. The false match rate used in this calculation is the same as in the main text, since the same sample is being used.

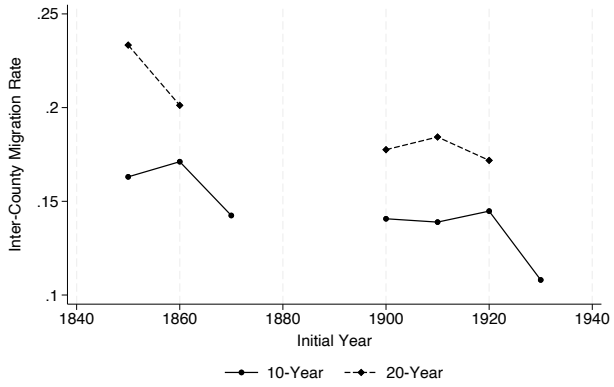
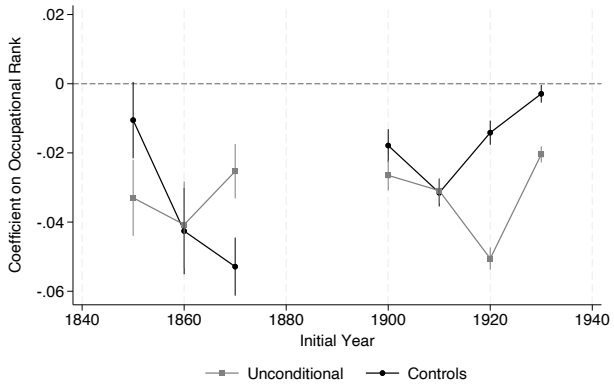


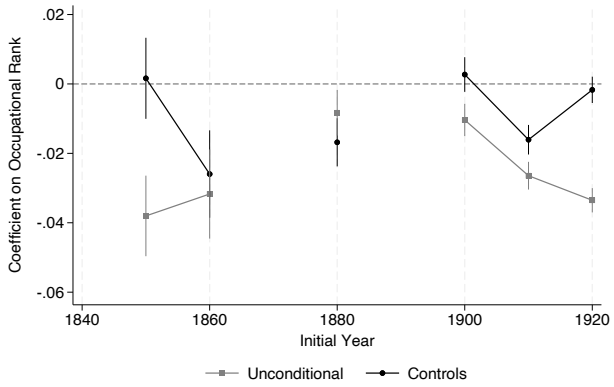
Figure D.1: Inter-county migration rates by span, corrected for false matches

Note: The figure shows the probability that an individual in the linked sample beginning in the year on the x-axis was observed living in a different non-overlapping county 10 or 20 years later, according to the main linkage method. All observations are weighted by inverse linkage probability. The estimates are then adjusted for false matches according to the method presented in text.

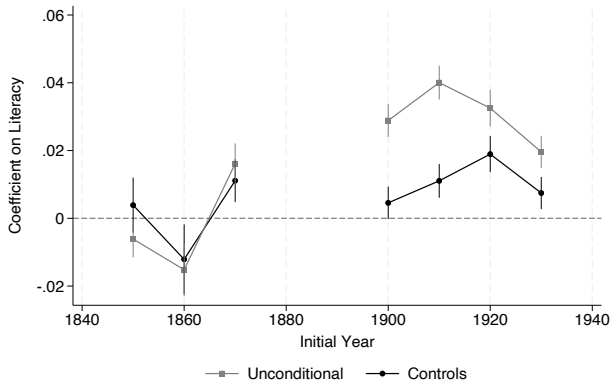
(a) Occupational rank, 10-year spans



(b) Occupational rank, 20-year spans



(c) Literacy, 10-year spans



(d) Literacy, 20-year spans

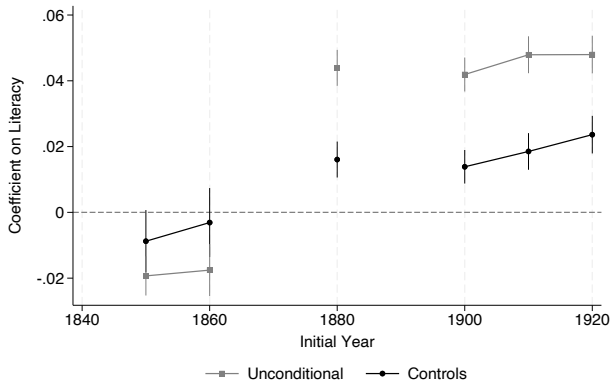
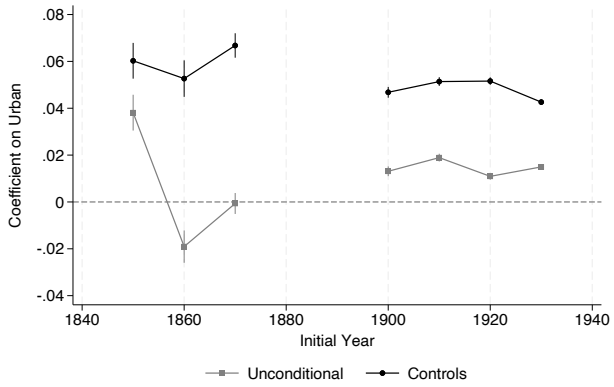


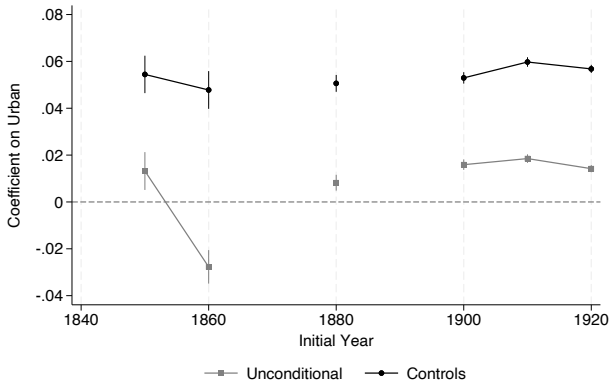
Figure D.2: Migrant selection by measure and span, individual characteristics

Note: Each figure shows the coefficient on the variable in question in a selection regression for migration over the stated span. All observations are weighted by inverse linkage probability. Vertical lines indicate 95-percent confidence intervals. Controls are all controls available in all census years, as well as county fixed effects.

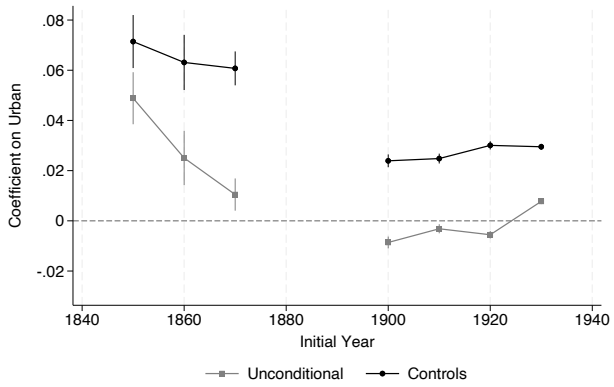
(a) Urban (2,500+), 10-year spans



(b) Urban (2,500+), 20-year spans



(c) Urban (25,000+), 10-year spans



(d) Urban (25,000+), 20-year spans

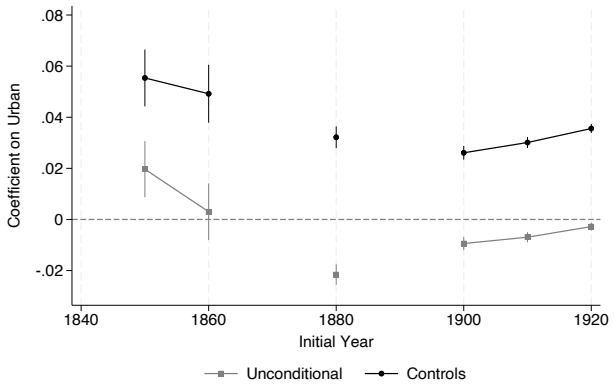
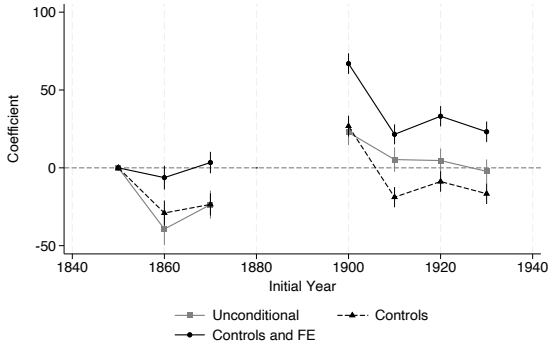


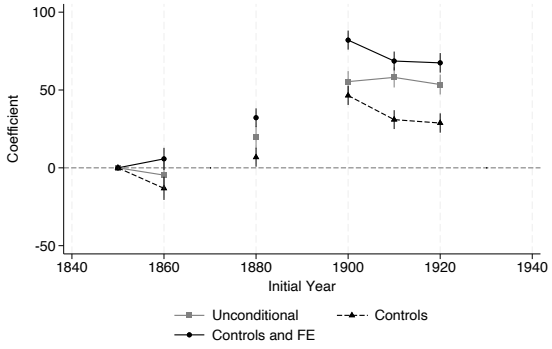
Figure D.3: Migrant selection by measure and span, urban residence

Note: Each figure shows the coefficient on the variable in question in a selection regression for migration over the stated span. All observations are weighted by inverse linkage probability. Vertical lines indicate 95-percent confidence intervals. Controls are all controls available in all census years, as well as state fixed effects.

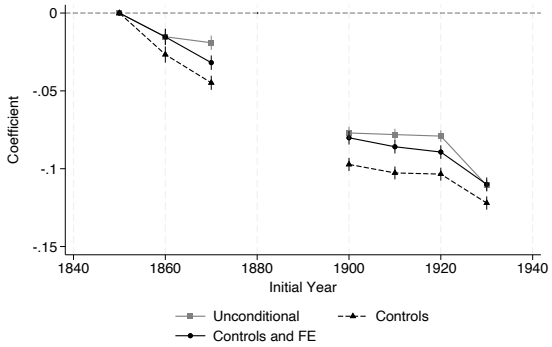
(a) Distance of move (miles), 10-year spans



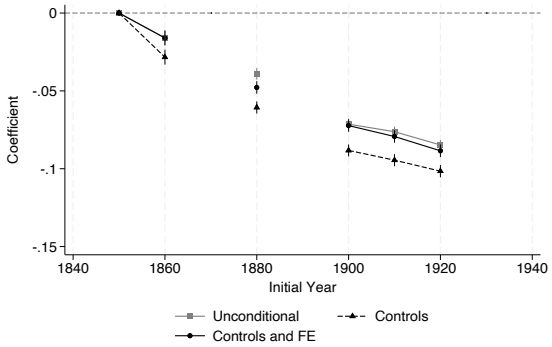
(b) Distance of move (miles), 20-year spans



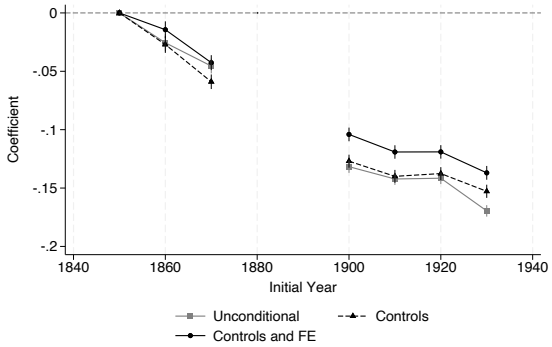
(c) Interstate move, 10-year spans



(d) Interstate move, 20-year spans



(e) Interregion move, 10-year spans



(f) Interregion move, 20-year spans

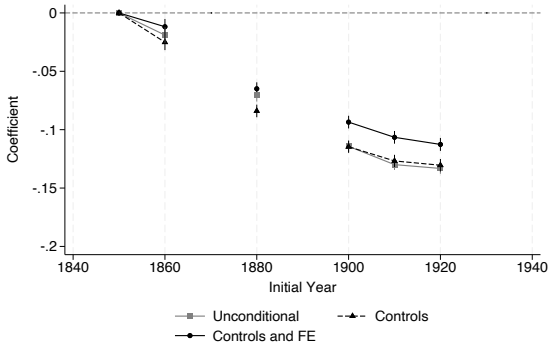


Figure D.4: Changes in distance of move

Note: These figures show the coefficients on initial-year fixed effects in regressions of destination characteristics with and without controls, with 1850 as the excluded year. "Controls" includes all controls available in all censuses, including state fixed effects. "Controls and FE' includes also county fixed effects. Vertical lines are 95-percent confidence intervals. Observations weighted to correct for selection into linkage.

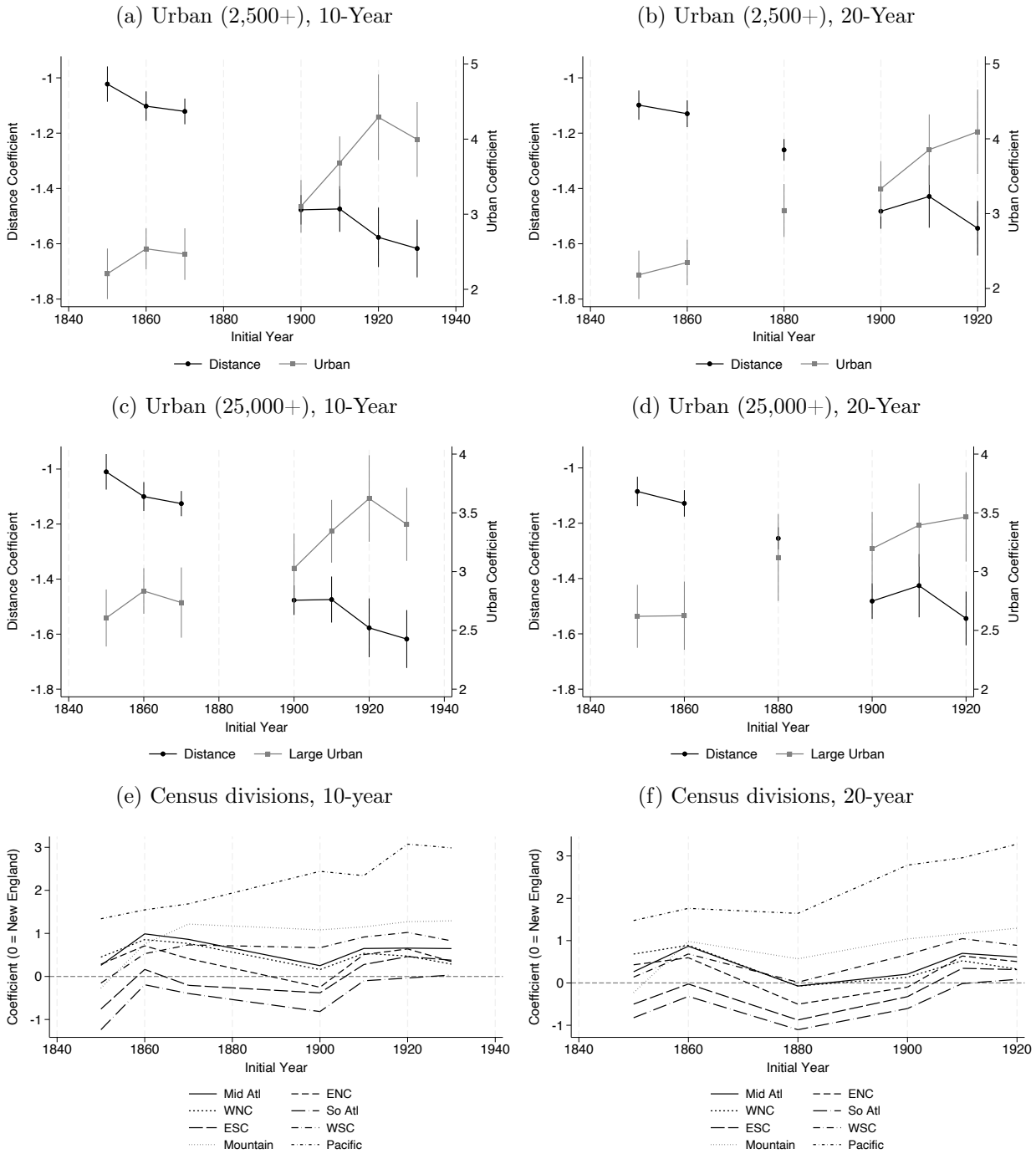
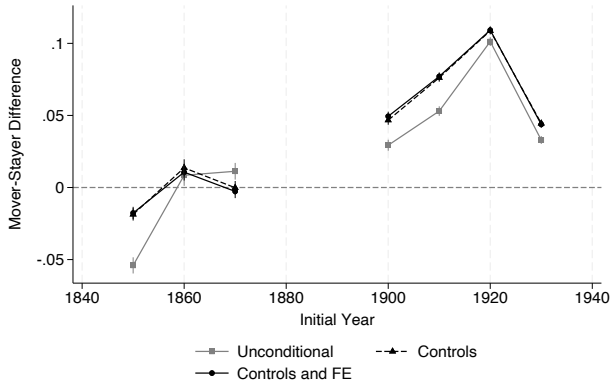


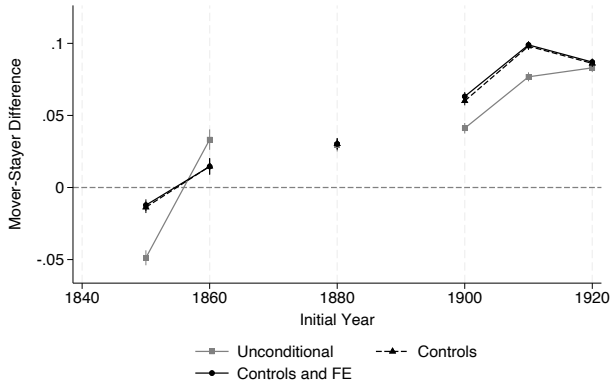
Figure D.5: Conditional Logit Results

Note: Panels (a)–(d) present coefficients on distance and the measure of urban residence from the conditional logit estimation. Panels (e) and (f) present the coefficients from the regressions of panels (a) and (b), respectively, on the census division indicators, with New England excluded. Note that there is no data for spans that would include 1890; the lines span this period simply to enable clearer reading of the trend over time.

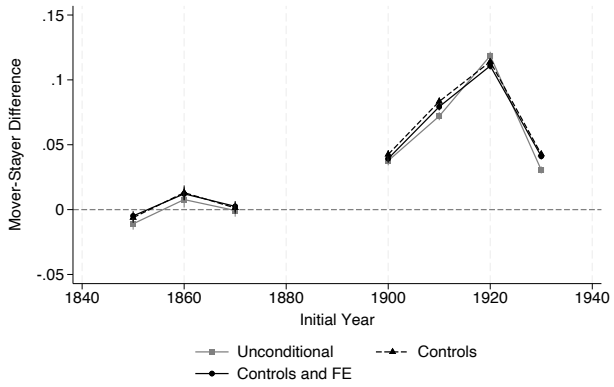
(a) Urban (2,500+), 10-Year



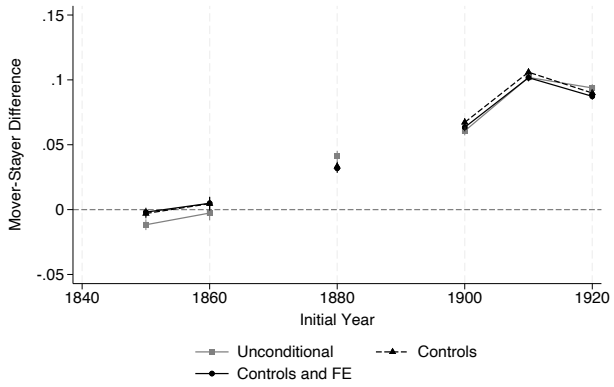
(b) Urban (2,500+), 20-Year



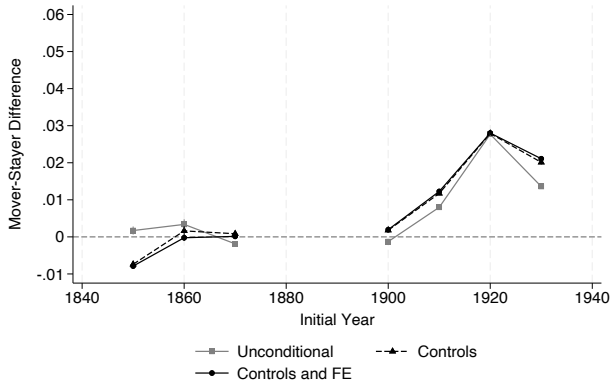
(c) Urban (25,000+), 10-Year



(d) Urban (25,000+), 20-Year



(e) Labor demand growth, 10-Year



(f) Labor demand growth, 20-Year

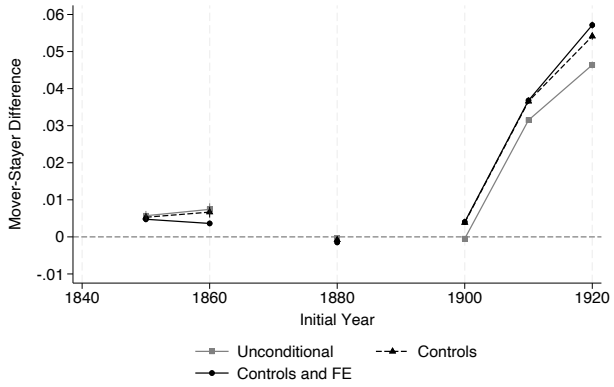


Figure D.6: Movers' change in urbanization and labor demand growth

Note: This figure compares the growth in urbanization or labor demand experienced by movers to that experienced by stayers, with either no controls, controlling for all observables, and controlling for all observables and county fixed effects.

E Results dropping second-generation immigrants

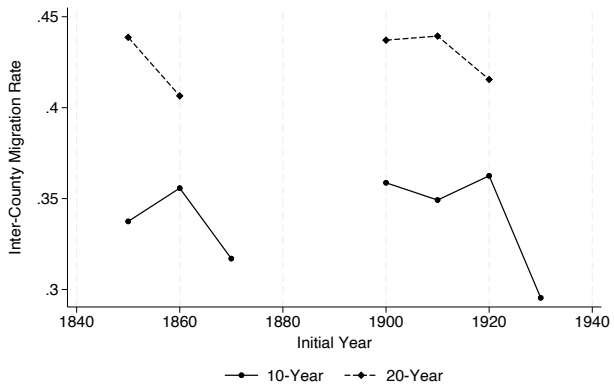
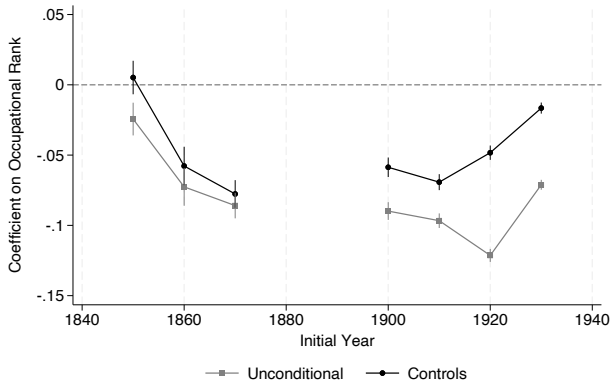


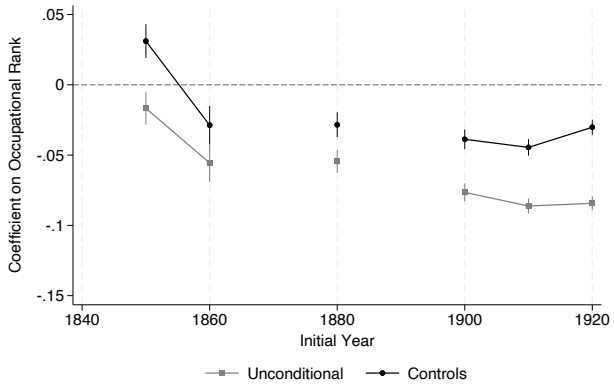
Figure E.1: Inter-county migration rates by span, corrected for false matches

Note: This figure shows the probability that an individual in the linked sample beginning in the year on the x-axis was observed living in a different non-overlapping county 10 or 20 years later, according to the main linkage method. All observations are weighted by inverse linkage probability. The estimates are then adjusted for false matches according to the method presented in text.

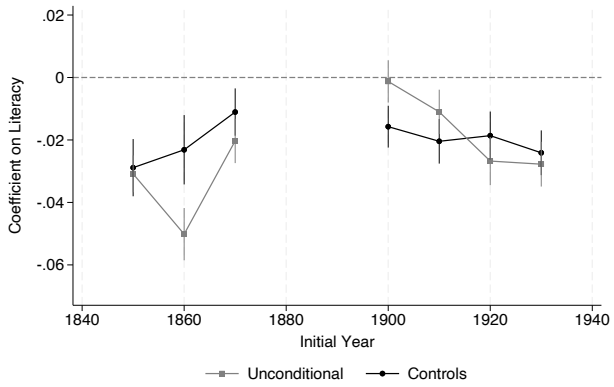
(a) Occupational rank, 10-year spans



(b) Occupational rank, 20-year spans



(c) Literacy, 10-year spans



(d) Literacy, 20-year spans

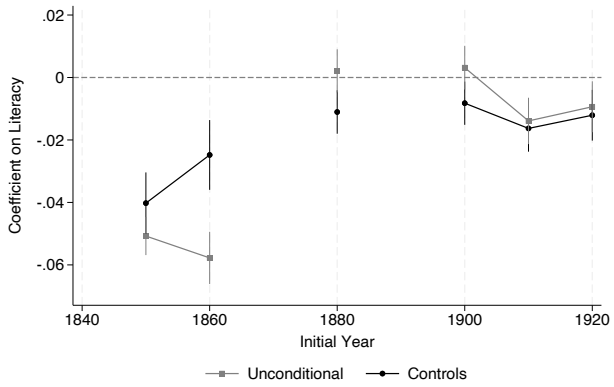
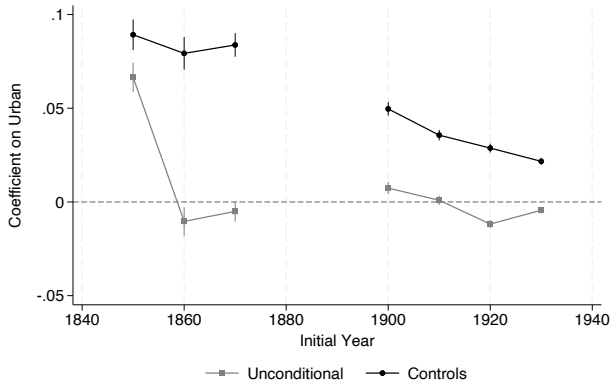


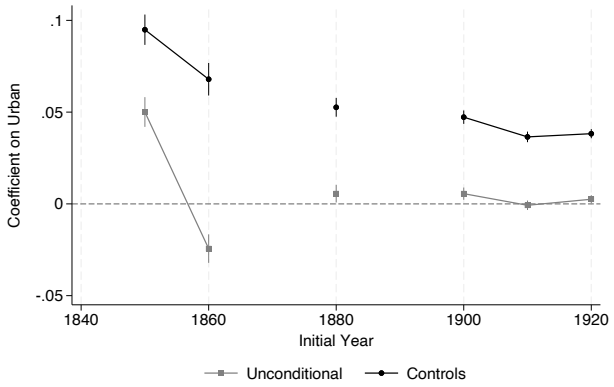
Figure E.2: Migrant selection by measure and span

Note: Each figure shows the coefficient on the variable in question in a selection regression for migration over the stated span. All observations are weighted by inverse linkage probability. Vertical lines indicate 95-percent confidence intervals. Controls are all controls available in all census years, as well as county fixed effects.

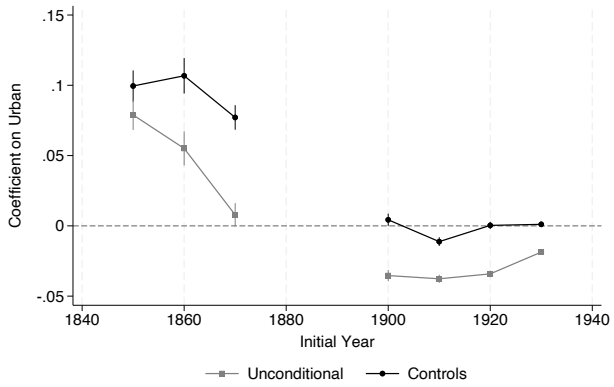
(a) Urban (2,500+), 10-year spans



(b) Urban (2,500+), 20-year spans



(c) Urban (25,000+), 10-year spans



(d) Urban (25,000+), 20-year spans

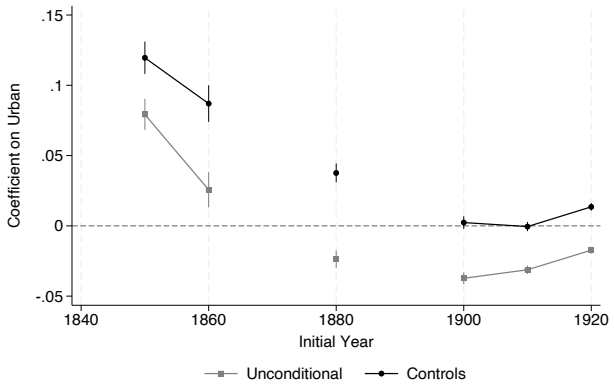
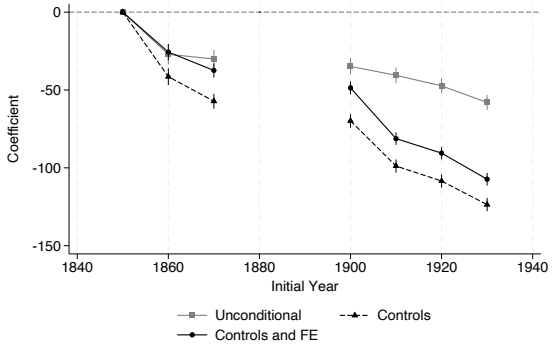


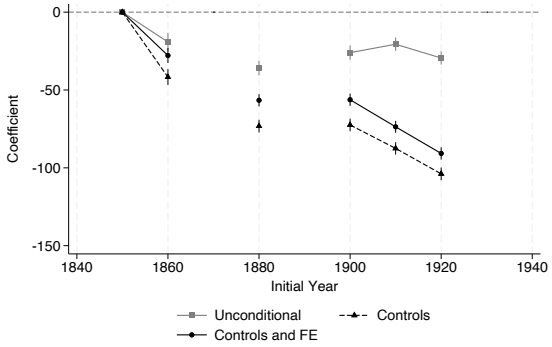
Figure E.3: Migrant selection by measure and span

Note: Each figure shows the coefficient on the variable in question in a selection regression for migration over the stated span. All observations are weighted by inverse linkage probability. Vertical lines indicate 95-percent confidence intervals. Controls are all controls available in all census years, as well as state fixed effects.

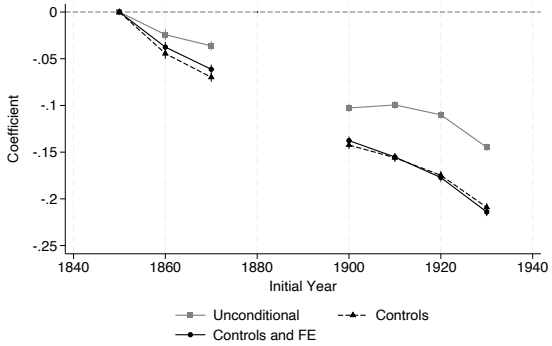
(a) Distance of move (miles), 10-year spans



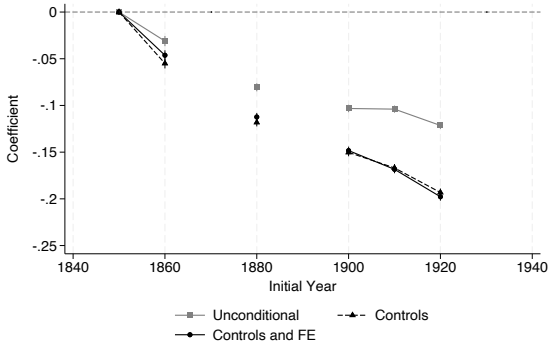
(b) Distance of move (miles), 20-year spans



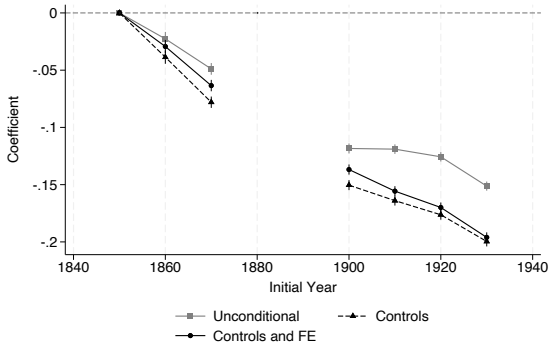
(c) Interstate move, 10-year spans



(d) Interstate move, 20-year spans



(e) Interregion move, 10-year spans



(f) Interregion move, 20-year spans

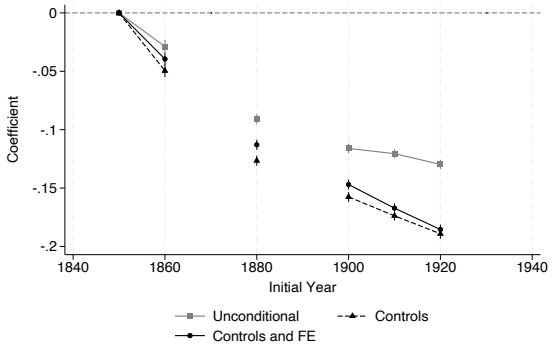
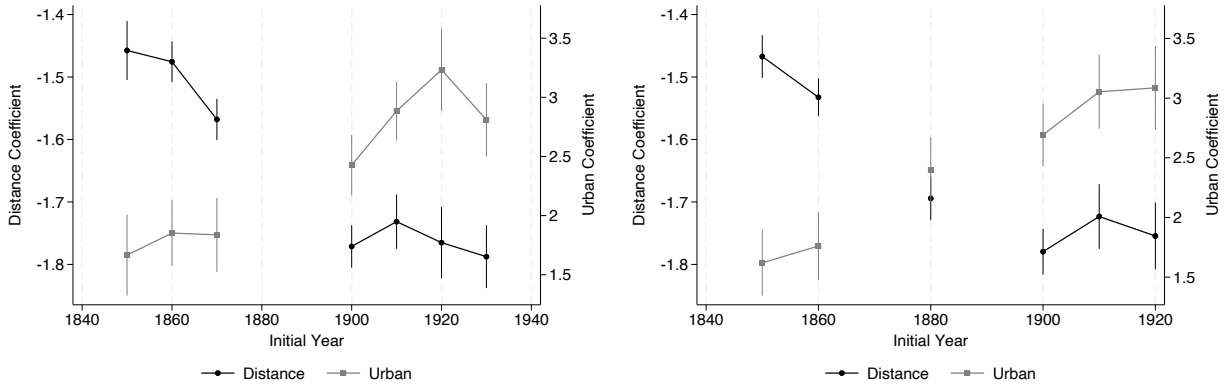


Figure E.4: Changes in distance of move

Note: These figures show the coefficients on initial-year fixed effects in regressions of destination characteristics with and without controls, with 1850 as the excluded year. "Controls" includes all controls available in all censuses, including state fixed effects. "Controls and FE' includes also county fixed effects. Vertical lines are 95-percent confidence intervals. Observations weighted to correct for selection into linkage.

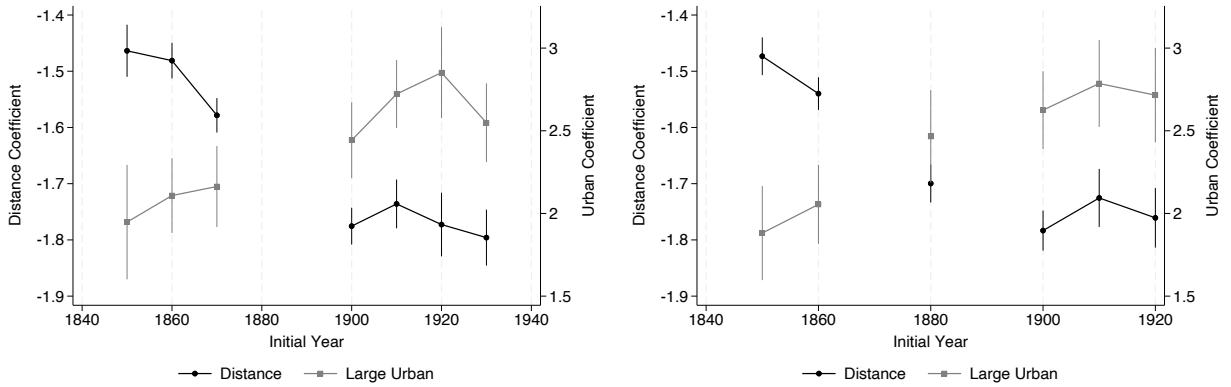
(a) Urban (2,500+), 10-Year

(b) Urban (2,500+), 20-Year



(c) Urban (25,000+), 10-Year

(d) Urban (25,000+), 20-Year



(e) Census divisions, 10-year

(f) Census divisions, 20-year

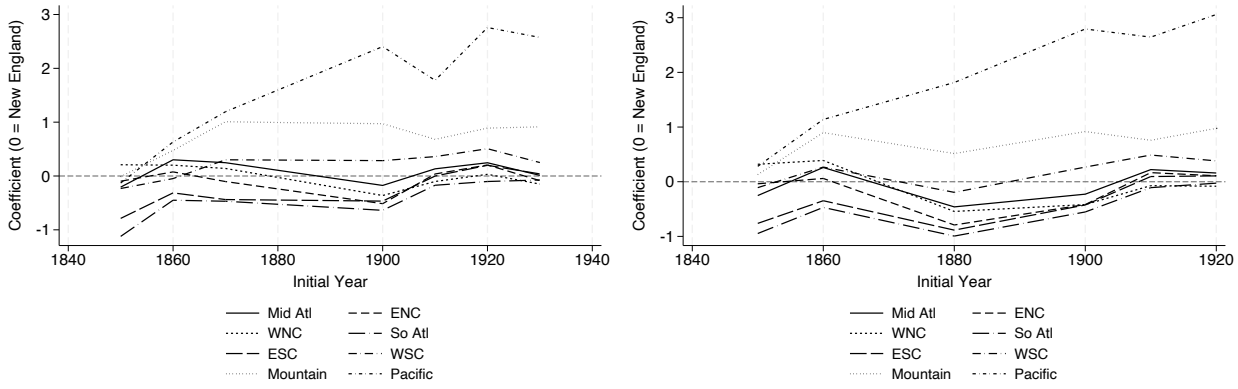
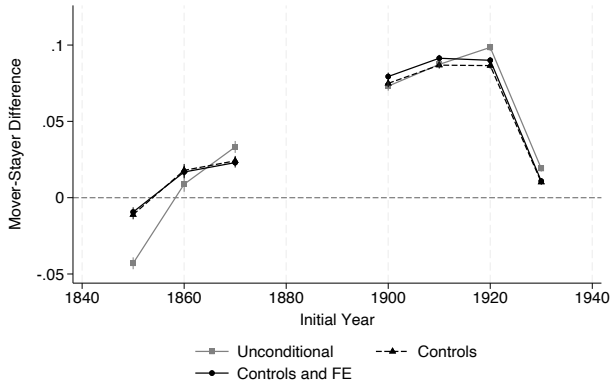


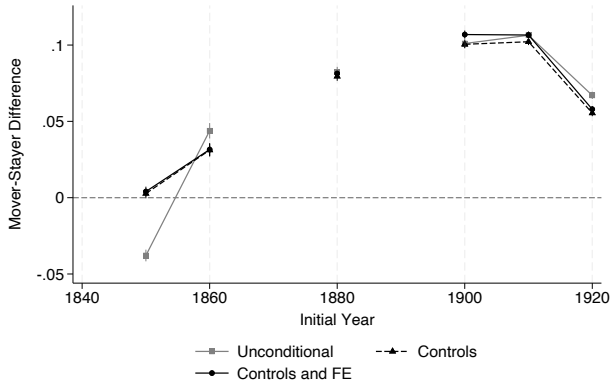
Figure E.5: Conditional Logit Results

Note: Panels (a)–(d) present coefficients on distance and the measure of urban residence from the conditional logit estimation. Panels (e) and (f) present the coefficients from the regressions of panels (a) and (b), respectively, on the census division indicators, with New England excluded. Note that there is no data for spans that would include 1890; the lines span this period simply to enable clearer reading of the trend over time.

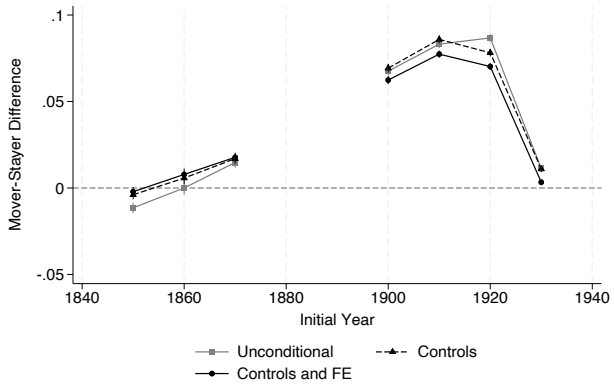
(a) Urban (2,500+), 10-Year



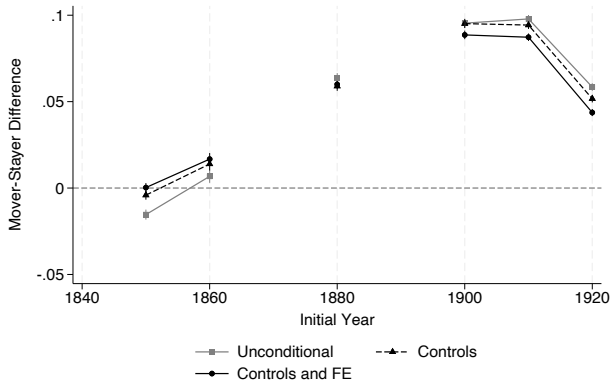
(b) Urban (2,500+), 20-Year



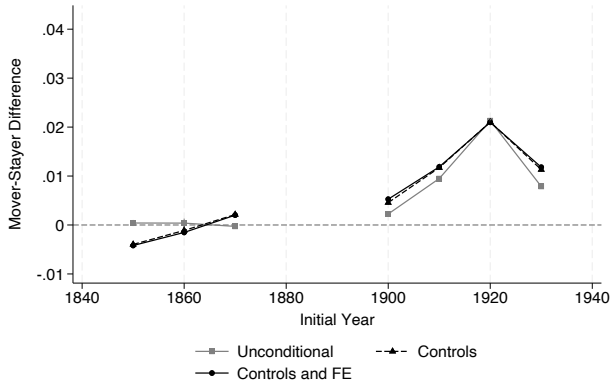
(c) Urban (25,000+), 10-Year



(d) Urban (25,000+), 20-Year



(e) Labor demand growth, 10-Year



(f) Labor demand growth, 20-Year

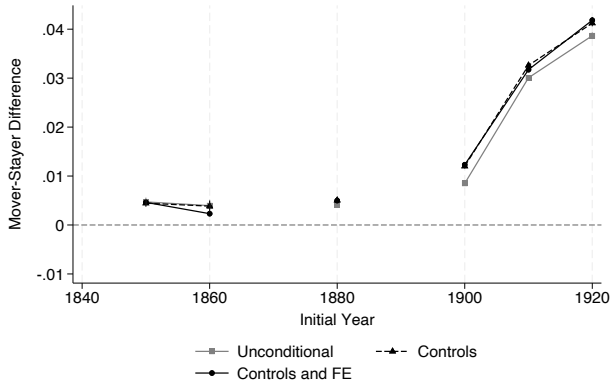


Figure E.6: Movers' change in urbanization and labor demand growth

Note: This figure compares the growth in urbanization or labor demand experienced by movers to that experienced by stayers, with either no controls, controlling for all observables, and controlling for all observables and county fixed effects.

F Results with imputed occupational codes

The census files provided by Ruggles et al. (2021) for census years 1900–1930 are preliminary. For the purposes of this paper, the main limitation is that in a significant number of cases, the occupational strings have not yet been assigned the three-digit codes that I use to assign them occupational ranks. Instead, they are listed as "Not Yet Classified" (occ1950=979). To bring these individuals back into analysis, I use Zimran's (2022b) crosswalks to assign each such individual an imputed occupational code based on the original occupational string. The results in this appendix repeat the analysis of the main text, but incorporate these individuals.

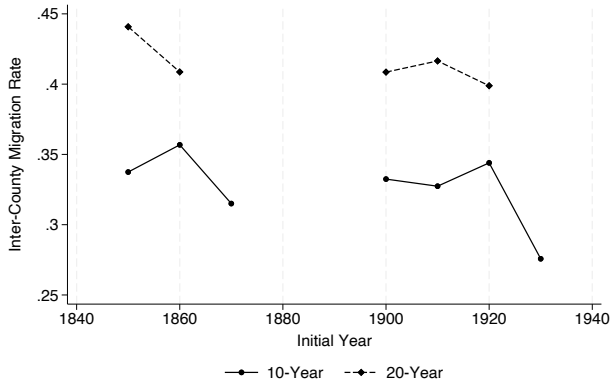
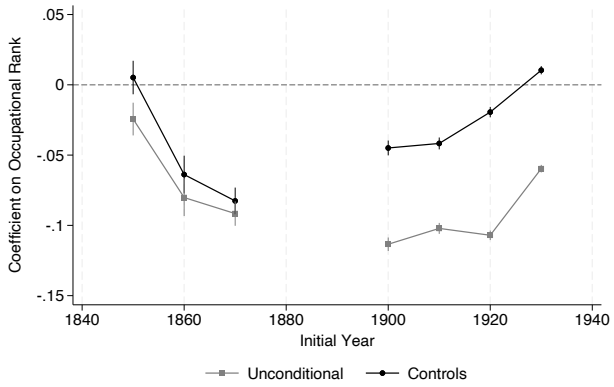


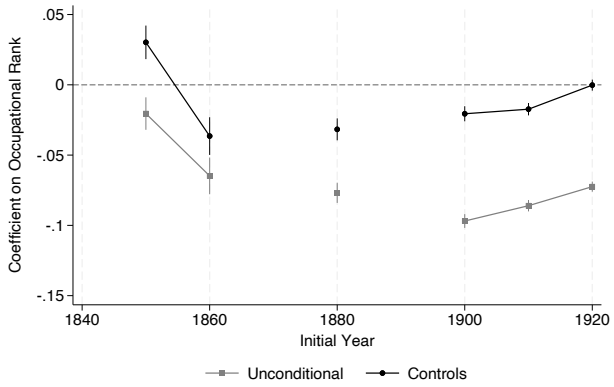
Figure F.1: Inter-county migration rates by linkage method and span, corrected for false matches

Note: This figure shows the probability that an individual in the linked sample beginning in the year on the x-axis was observed living in a different non-overlapping county 10 or 20 years later, according to the main linkage method. All observations are weighted by inverse linkage probability. The estimates are then adjusted for false matches according to the method presented in text.

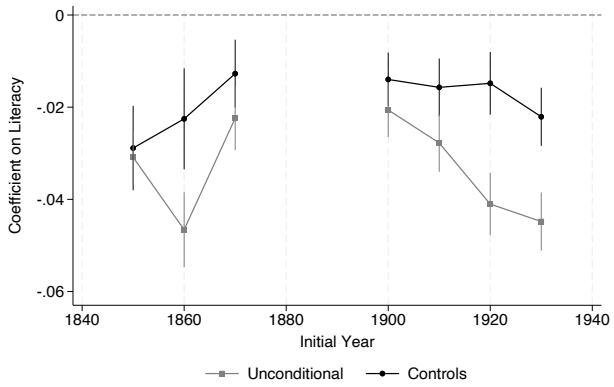
(a) Occupational rank, 10-year spans



(b) Occupational rank, 20-year spans



(c) Literacy, 10-year spans



(d) Literacy, 20-year spans

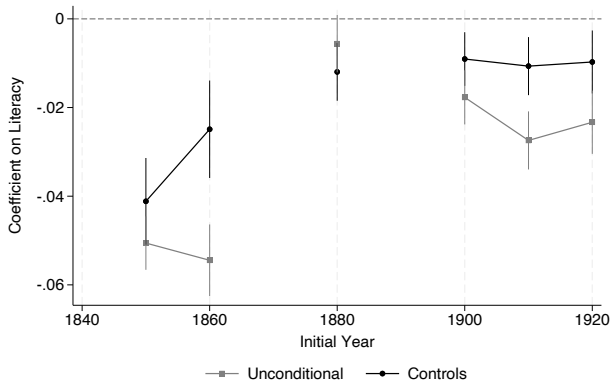
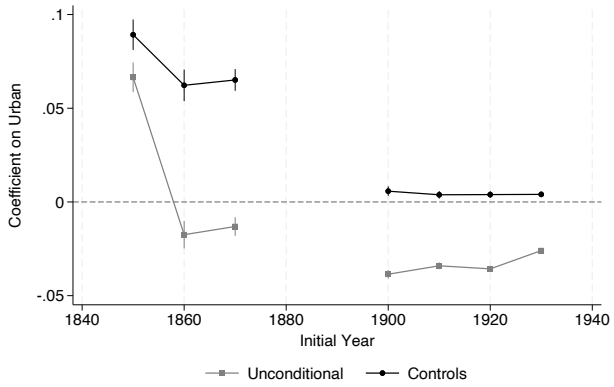


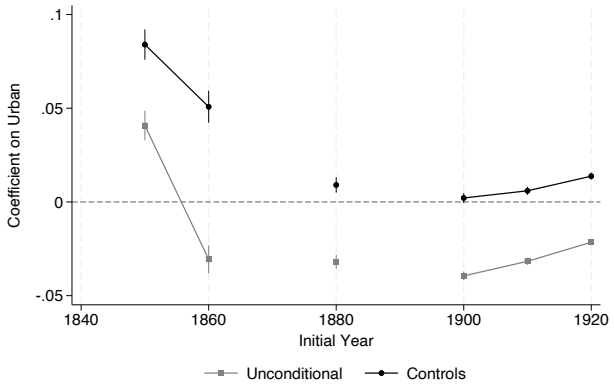
Figure F.2: Migrant selection by measure and span, individual characteristics

Note: Each figure shows the coefficient on the variable in question in a selection regression for migration over the stated span. All observations are weighted by inverse linkage probability. Vertical lines indicate 95-percent confidence intervals. Controls are all controls available in all census years, as well as county fixed effects.

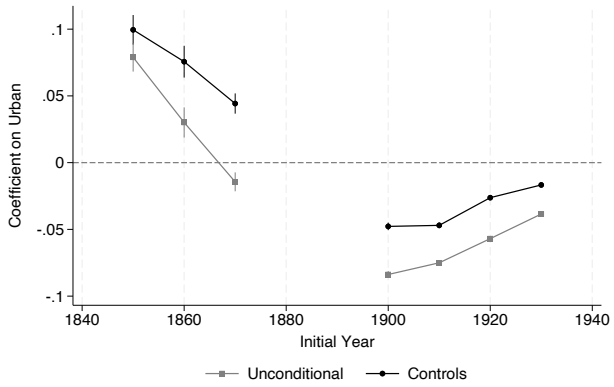
(a) Urban (2,500+), 10-year spans



(b) Urban (2,500+), 20-year spans



(c) Urban (25,000+), 10-year spans



(d) Urban (25,000+), 20-year spans

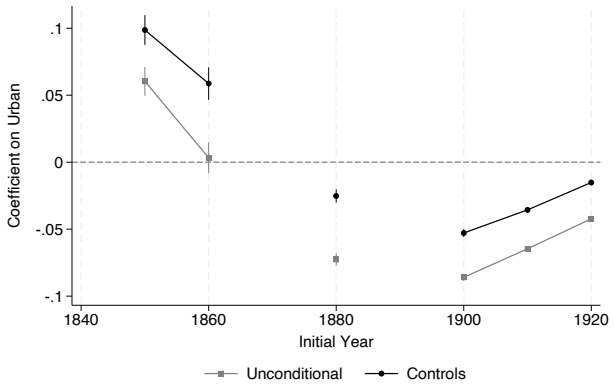
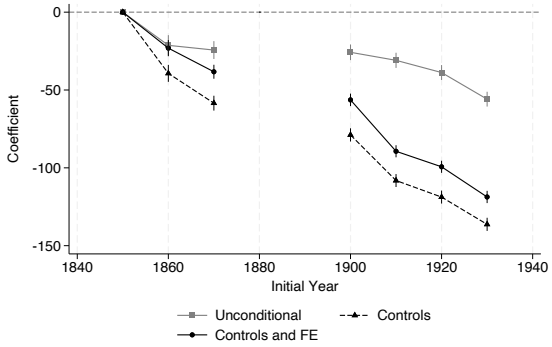


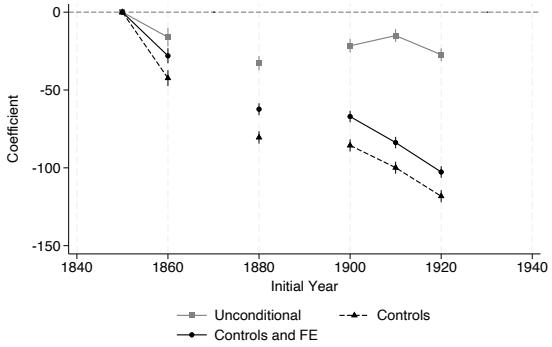
Figure F.3: Migrant selection by measure and span, urban residence

Note: Each figure shows the coefficient on the variable in question in a selection regression for migration over the stated span. All observations are weighted by inverse linkage probability. Vertical lines indicate 95-percent confidence intervals. Controls are all controls available in all census years, as well as state fixed effects.

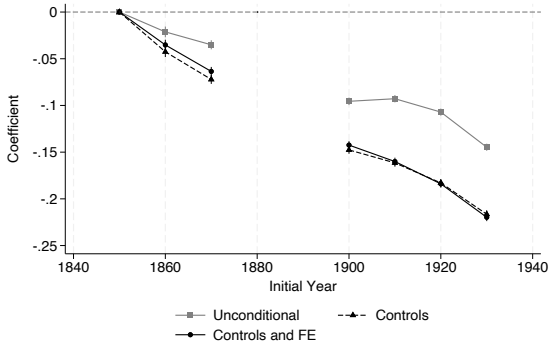
(a) Distance of move (miles), 10-year spans



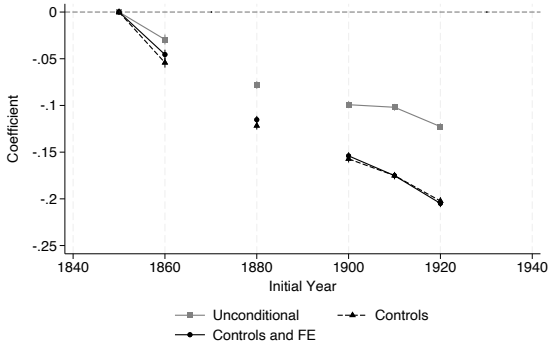
(b) Distance of move (miles), 20-year spans



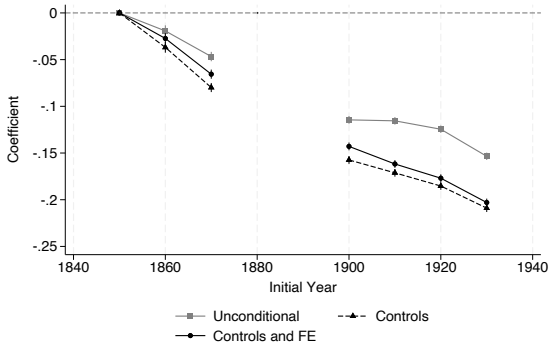
(c) Interstate move, 10-year spans



(d) Interstate move, 20-year spans



(e) Interregion move, 10-year spans



(f) Interregion move, 20-year spans

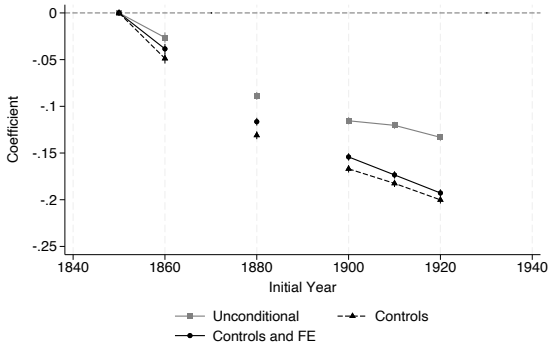


Figure F.4: Changes in distance of move

Note: These figures show the coefficients on initial-year fixed effects in regressions of destination characteristics with and without controls, with 1850 as the excluded year. "Controls" includes all controls available in all censuses, including state fixed effects. "Controls and FE' includes also county fixed effects. Vertical lines are 95-percent confidence intervals. Observations weighted to correct for selection into linkage.

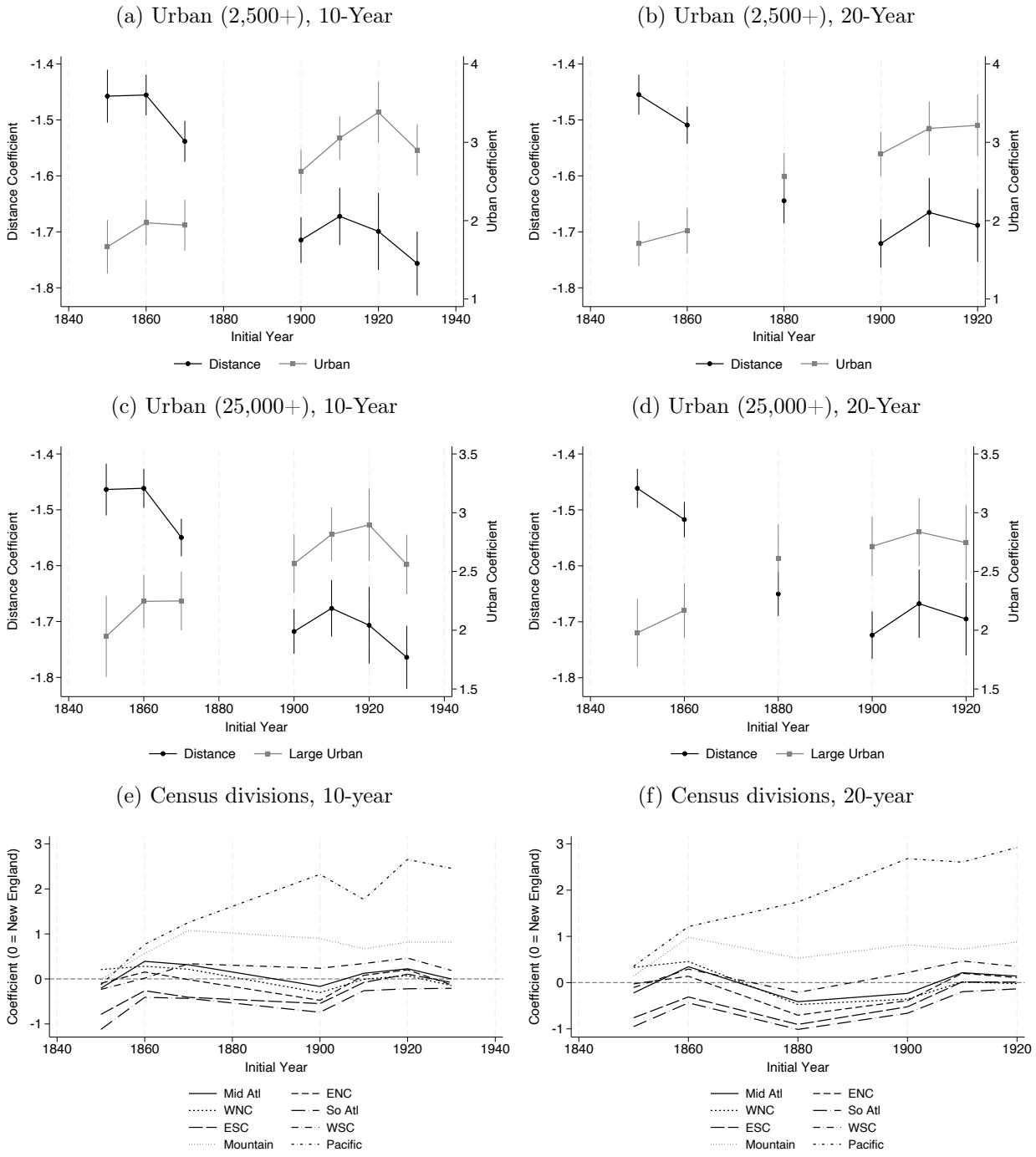


Figure F.5: Conditional Logit Results

Note: Panels (a)–(d) present coefficients on distance and the measure of urban residence from the conditional logit estimation. Panels (e) and (f) present the coefficients from the regressions of panels (a) and (b), respectively, on the census division indicators, with New England excluded. Note that there is no data for spans that would include 1890; the lines span this period simply to enable clearer reading of the trend over time.

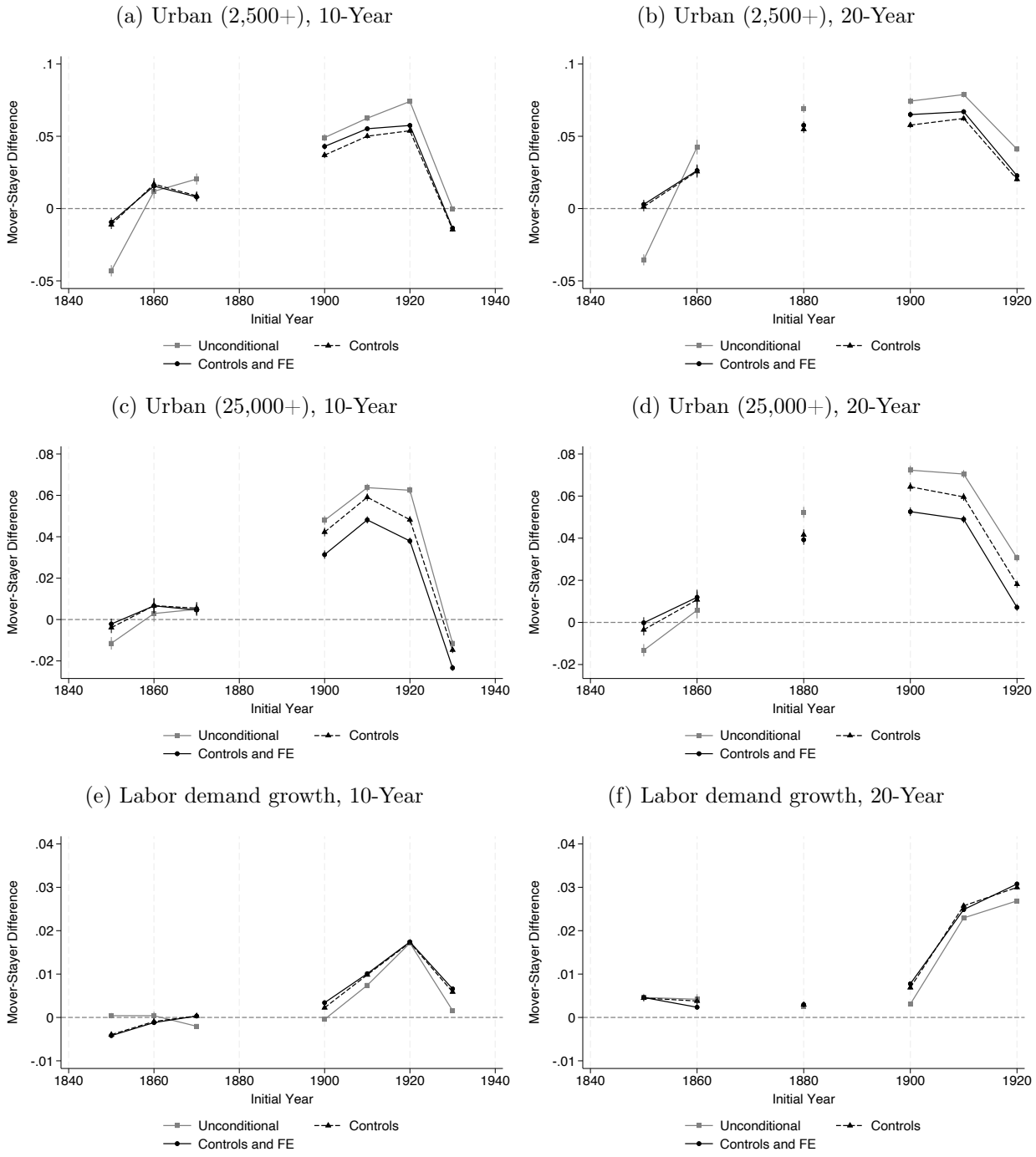


Figure F.6: Movers' change in urbanization and labor demand growth

Note: This figure compares the growth in urbanization or labor demand experienced by movers to that experienced by stayers, with either no controls, controlling for all observables, and controlling for all observables and county fixed effects.

References

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

Abramitzky, Ran, Leah Platt Boustan, Katherine Eriksson, Santiago Pérez, and Myera Rashid (2020). *Census Linking Project: Version 2.0 [dataset]*. https://censuslinkingproject.org.

Collins, William J. and Ariell Zimran (2023). "Working Their Way Up? US Immigrants' Changing Labor Market Assimilation in the Age of Mass Migration." *American Economic Journal: Applied Economics* 15:3, pp. 238–269.

Guimarães, Paulo, Octávio Figueirdo, and Douglas Woodward (2003). "A Tractable Approach to the Firm Location Decision Problem." *Review of Economics and Statistics* 85:1, pp. 201–204.

Preston, Samuel H. and Michael R. Haines (1991). *Fatal Years: Child Mortality in Late Nineteenth Century America*. Princeton: Princeton University Press.

Ruggles, Steven, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler, and Matthew Sobek (2021). *Integrated Public Use Microdata Series: Version 11.0* [dataset]. Minneapolis: University of Minnesota.

Zimran, Ariell (2022a). *Census Linkage Crosswalks for US Censuses 1850–1940*. Ann Arbor: Inter-university Consortium for Political and Social Research [distributor]. doi:10.3886/E171701V3.

——— (2022b). *Miscellaneous US Census Crosswalks, 1850–1930*. Ann Arbor: Inter-university Consortium for Political and Social Research [distributor]. doi:10.3886/E171861V2.