Can migration make deadly recessions look healthy? Evidence from large-scale linked microdata

Vellore Arthi
University of Essex

Brian Beach
William & Mary
and NBER

Walker Hanlon
NYU Stern
and NBER

Health and the business cycle

- Changes in economic conditions can have positive and negative health effects.
- Ruhm (2000): Mortality falls during recessions.
  - Introduces panel-data approach that compares state-level mortality and unemployment rates.
  - Methodology since applied to a variety of settings (developed, developing, modern, and historical)
    - Results largely consistent with pro-cyclicality.
  - Over 1600 cites and several mentions in popular press.

A puzzle in the literature

- Studies using individual panel data find evidence of counter-cyclicality or no relationship at all.
  - Gerdtham and Johannesson, 2005; Halliday, 2014.
- **This paper:** Could migration offer an explanation for these diverging findings?

Why might migration matter?

- Aggregate panel data approach observes measures of health in a given location at a point in time.
- Critical assumption:
  - There is no migratory response to changes in local economic conditions.
  - Or, if there is a migratory response, then:
    - Migration is accurately reflected in intercensal population estimates.
    - Migrants are not selected from at-risk population.

When migration occurs

- **Aggregate panel approach:** *Observed* death rate may change as a result of unobserved out-migration or selective migration.
  - We call this migration bias.
- **Micro data approach:** By tracking individuals over time, inference is less affected by migration.

This paper

- **Question:** Can migration bias meaningfully affect estimates of the recession-mortality relationship?
- **Setting:** Lancashire Cotton Famine (1861-65)
  - Plausibly exogenous shock in economic conditions.
    - Shock was geographic in nature, allowing for clean categorization of treatment and control locations.
    - Migration flows are also easily observed.
  - Comprehensive and publicly available death records and census records.

This paper

- **Empirical approach:**
  - Link individual death records (1861-65) back to 1861 census records to identify:
    - Where each individual died.
    - Where they lived at time of enumeration.
  - Repeat for 1851-55.
  - Construct mortality rate holding location fixed at place of enumeration.
  - Adopt a differences-in-differences specification to see how mortality in cotton districts changed during the recession.

This paper

- **Results:**
  - Mortality for those who lived in cotton districts at the onset of the shortage *increased*, particularly mortality among those over the age of 50.
  - Mortality in nearby (migrant-receiving) areas also increased, consistent with migration generating congestion effects.
  - Treating location of death as initial place of residence produces results that are statistically different.
  - Comparing these results with aggregate data results suggests migration bias is most pronounced for adults (15-54).

Migration and the standard approach

$$\ln\left(MR_{dt}\right) = \alpha + \beta E_{dt} + \phi_d + \eta_t + \varepsilon_{dt}$$

- MR_{dt} is the mortality rate (deaths/population) in region *d* during period *t*.
- E_{dt} is a measure of economic conditions (e.g., unemployment rate).
- Region and time fixed effects, ϕ_d and η_t.

Three problems:

1) E_{dt} and migration are related.
2) Size and composition of population denominator is now changing.
3) Potential spillovers between migrant-sending and migrant-receiving areas.

Fix #1

$$\ln\left(MR_{dt}\right) = \alpha + \beta SHOCK_{dt} + \phi_d + \eta_t + \varepsilon_{dt}$$

- Measures of economic conditions may be endogenously related to migration.
- **Solution:** Use a plausibly-exogenous measure of economic conditions.
  - E.g., global price shock interacted with region-specific vulnerability to that price shock (as in Miller and Urdinola, 2010).

Fix #2

$$\ln\left(\frac{Mort_{dt}}{Pop_{dt}}\right) = \alpha + \beta SHOCK_{dt} + \phi_d + \eta_t + \varepsilon_{dt}$$

- In aggregate panel approach, numerator and denominator are only observed at the district level.
  - Unobserved migration can cause population to be mis-measured.
  - Selective migration may mean observed deaths are not an accurate proxy for actual deaths.
- **Solution:** Construct a linked dataset to hold individual's locations fixed, regardless of where deaths ultimately occur.

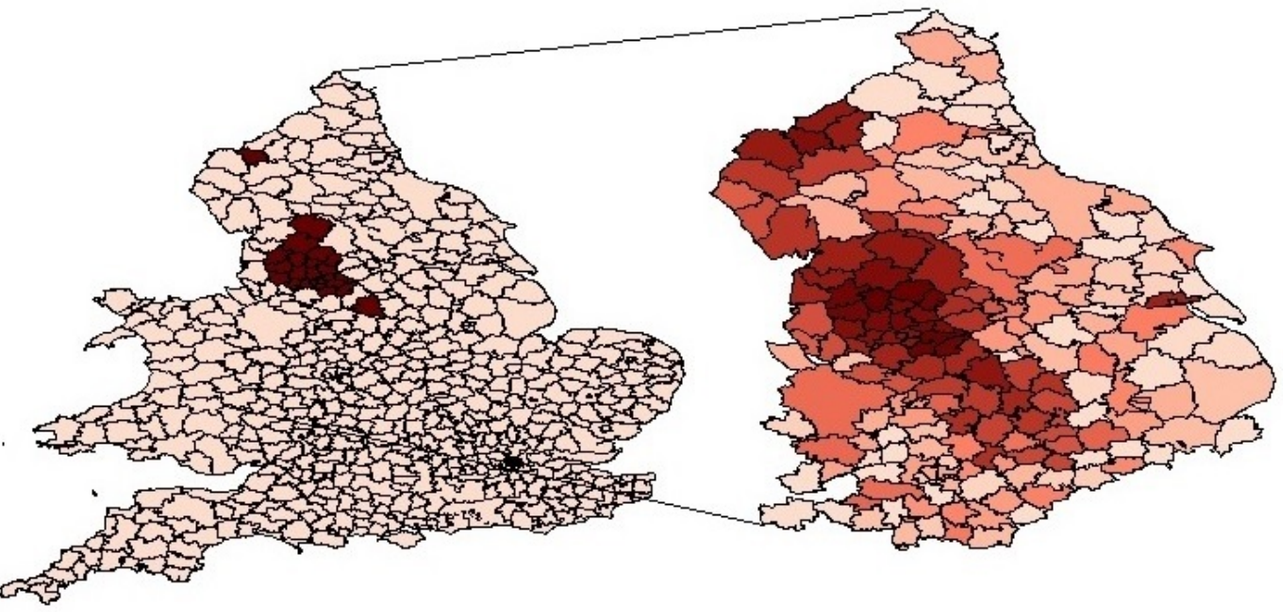
Fix #3

$$\ln\left(\frac{Mort_{dt}}{Pop_{dt}}\right) = \alpha + \beta_1 SHOCK_{dt} + \beta_2 RECEIVING_{dt} + \phi_d + \eta_t + \varepsilon_{dt}$$

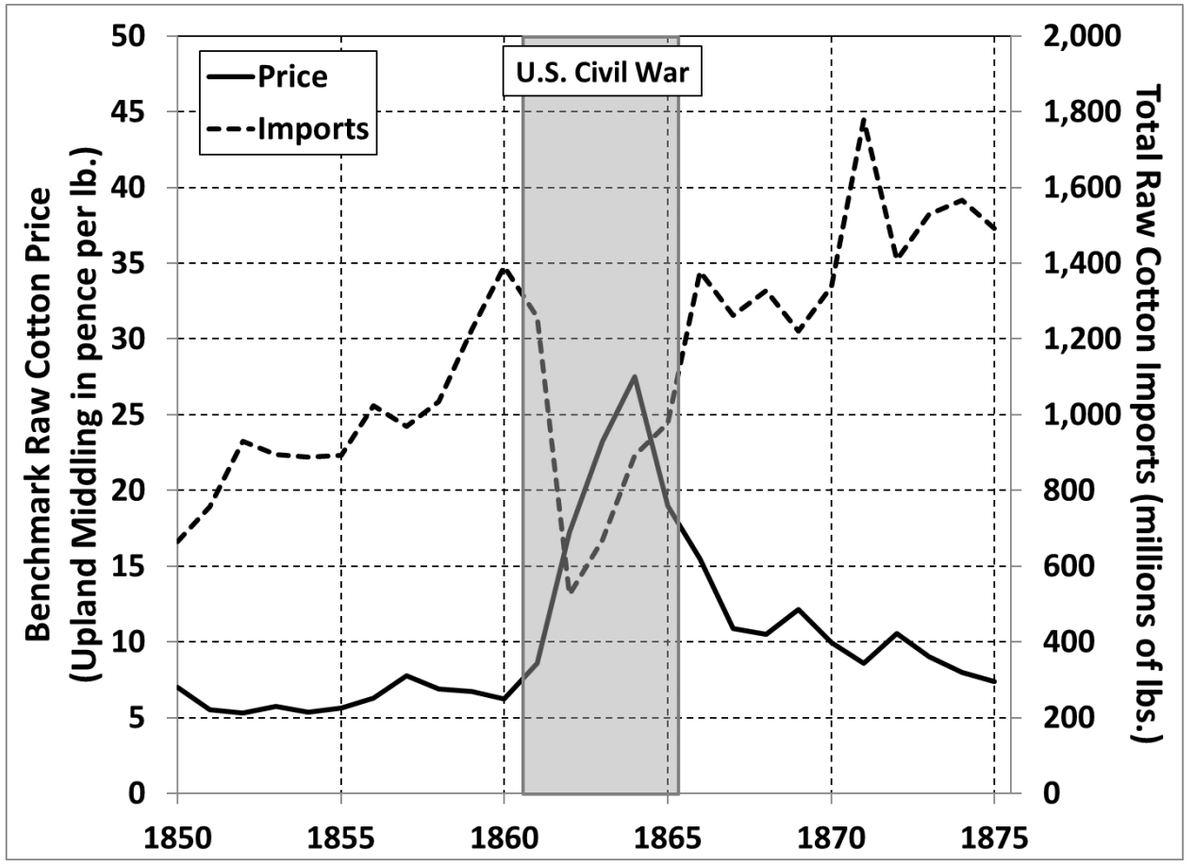
- If migrants influence mortality in migrant-receiving regions, then those spillovers will bias results in fixed effects regressions.
- **Solution:** Include an indicator that equals 1 for migrant-receiving regions during the shock period.

Empirical setting

Spatial distribution of cotton textile production

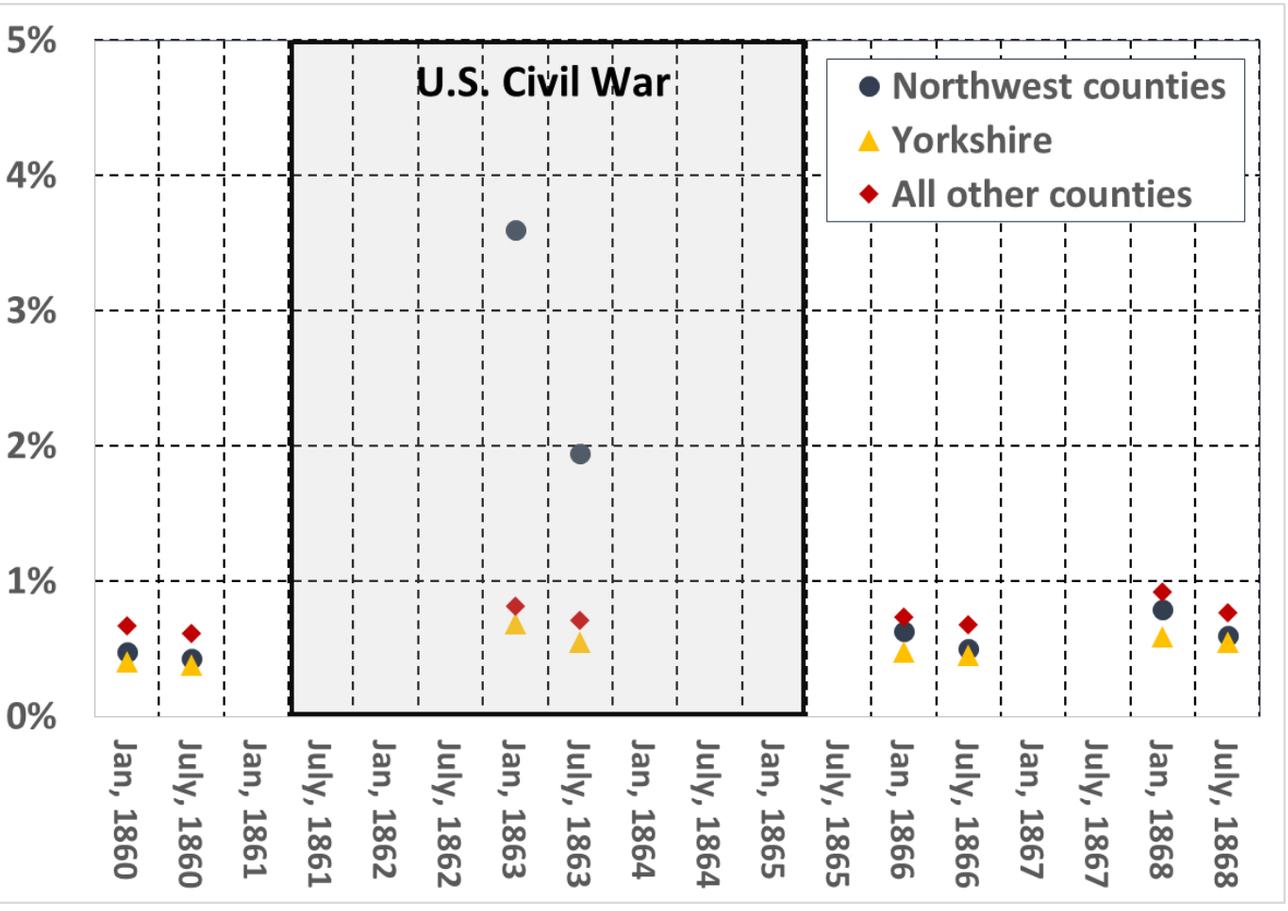


British Cotton Import Quantities and Prices

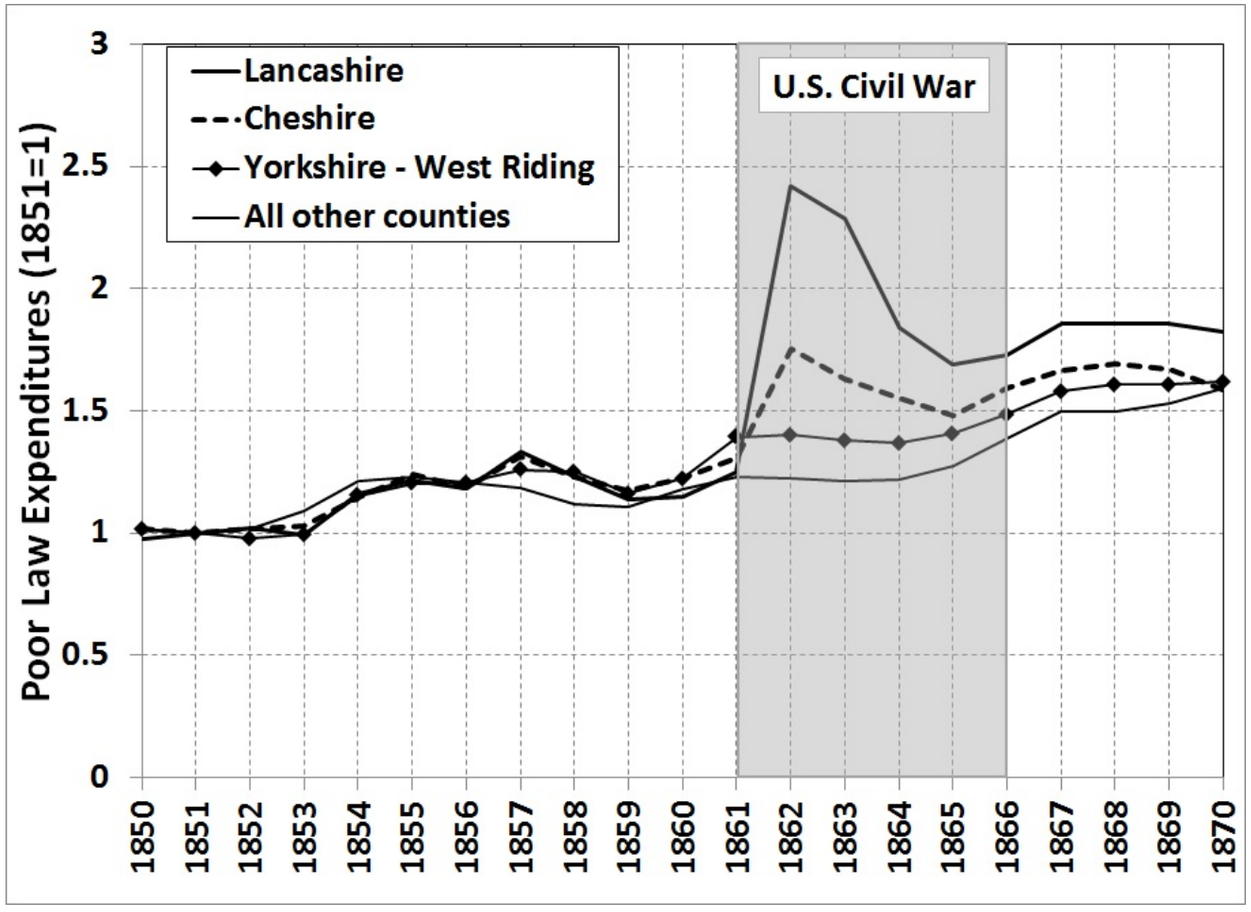


Empirical setting

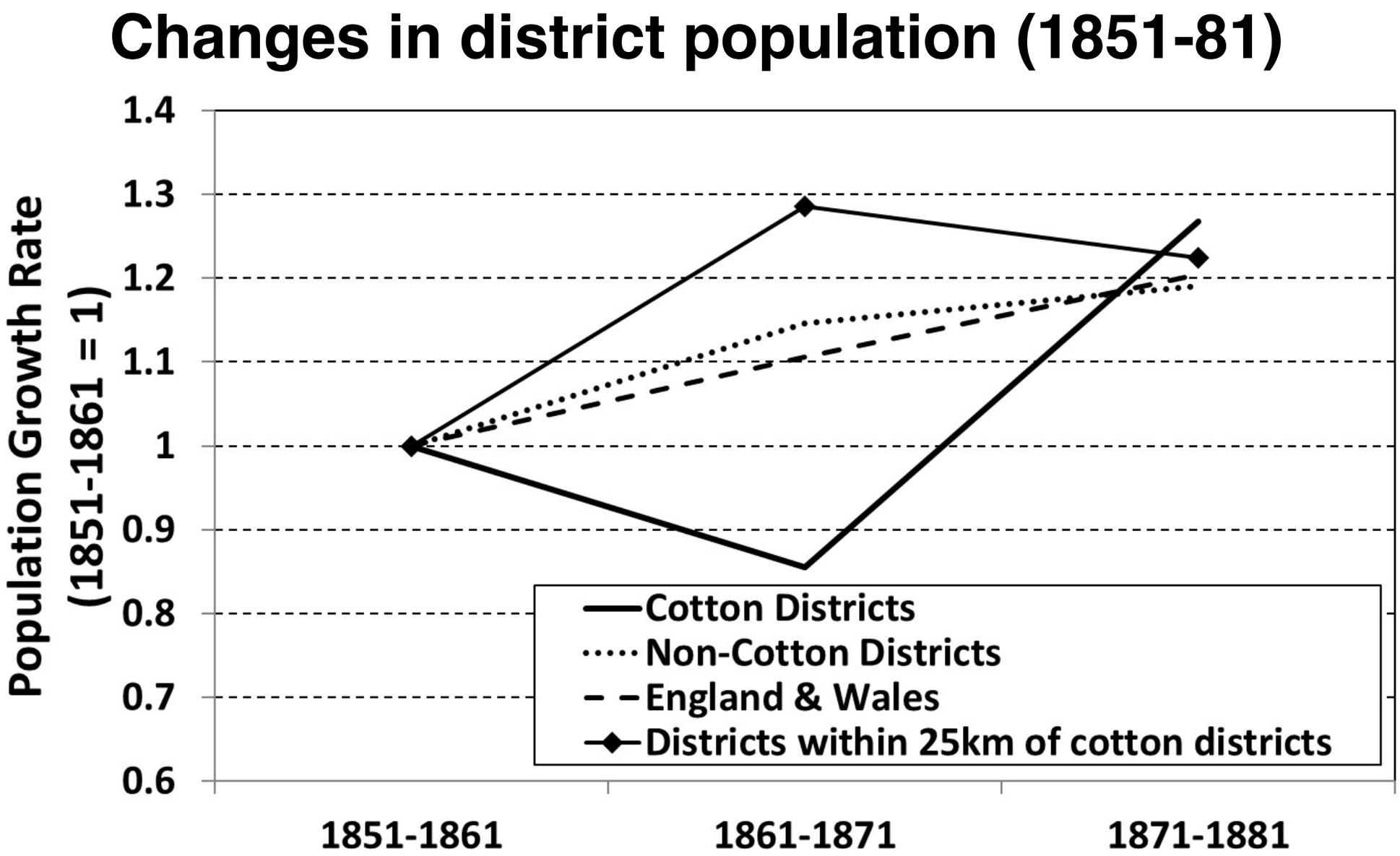
Able-bodied relief seekers (Share of pop.)



County-level poor law expenditures



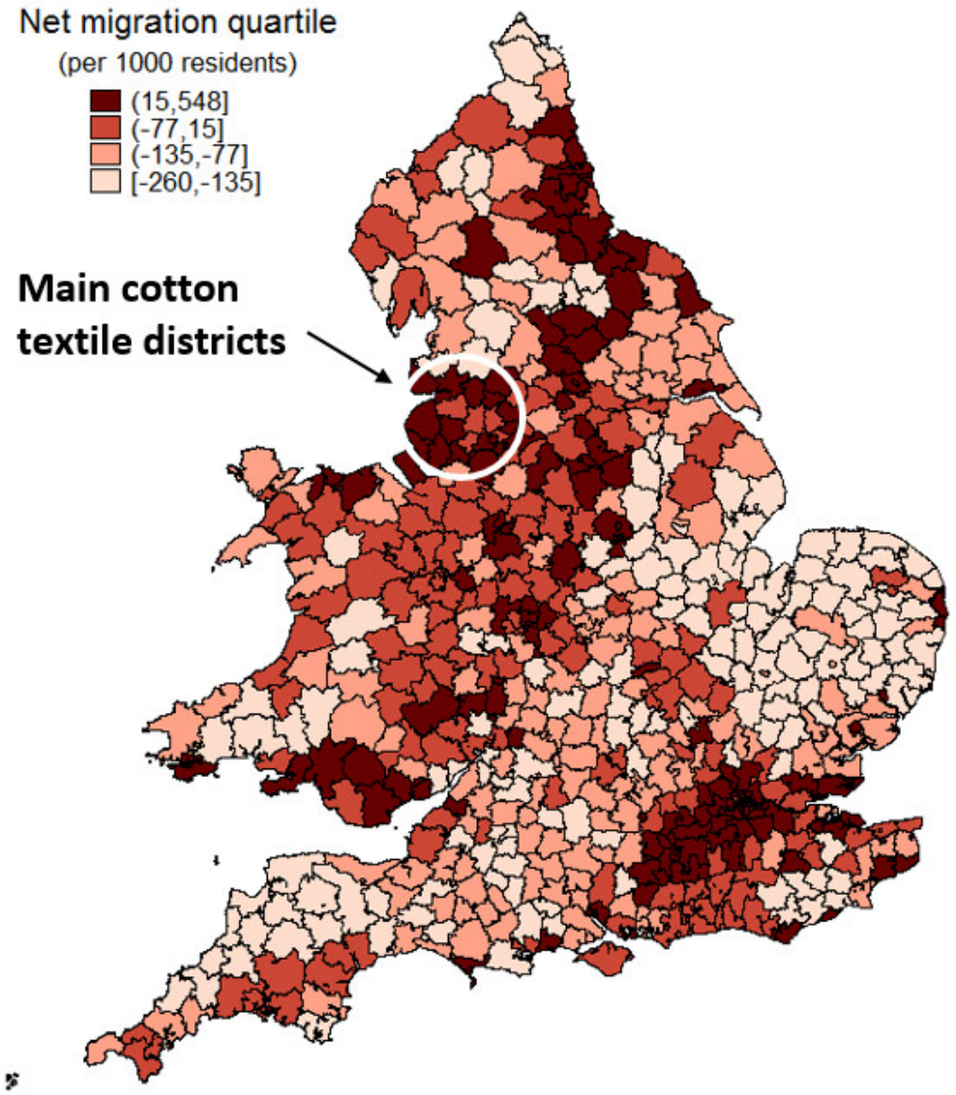
Migration results



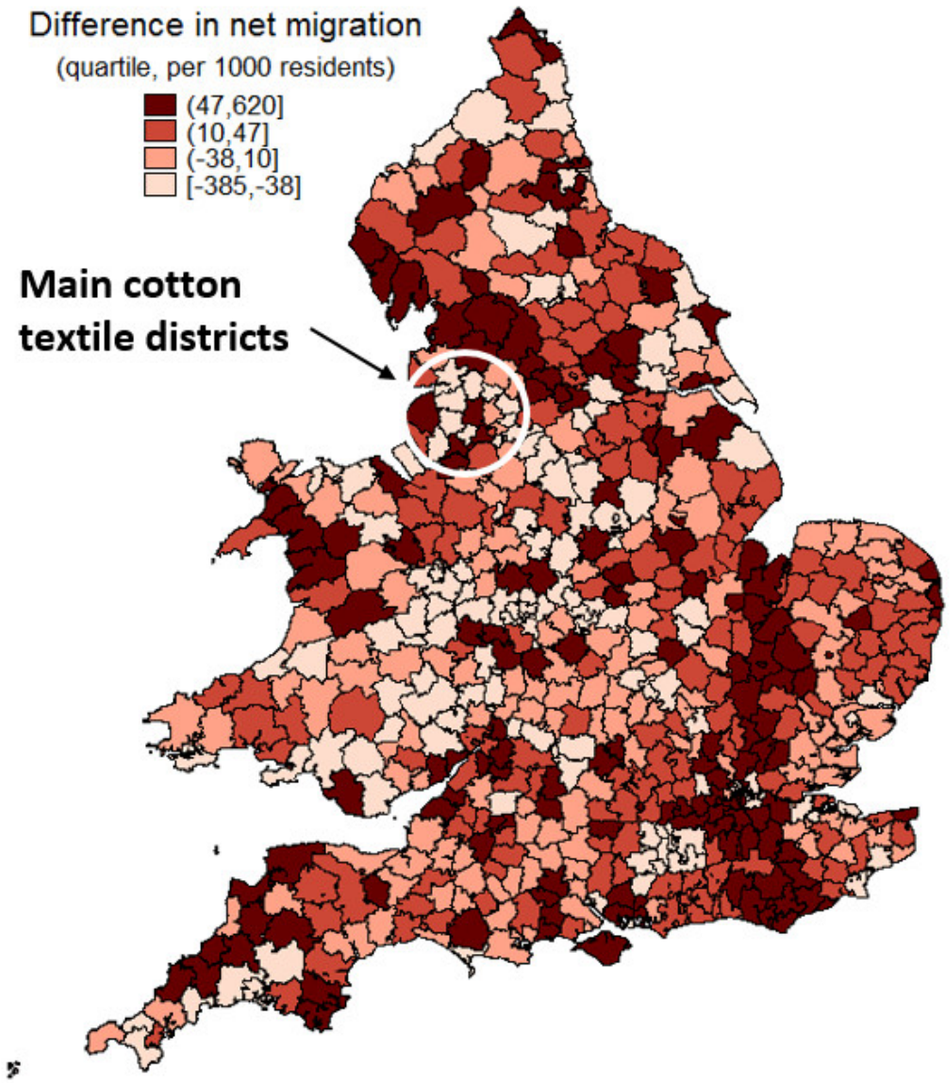
- If cotton districts continued to grow at their 51-61 rate, they would have had 50,000 additional residents in 1861 (2.3% increase)
- If nearby districts grew at their 51-61 rate they would have had 64,000 fewer residents (5% decrease).

Migration results

Net migration before the war (1851-61)



Implied change in net migration from 1851-1861 to 1861-1871



Net migration is the difference between two census population observations after adjusting for births and deaths.

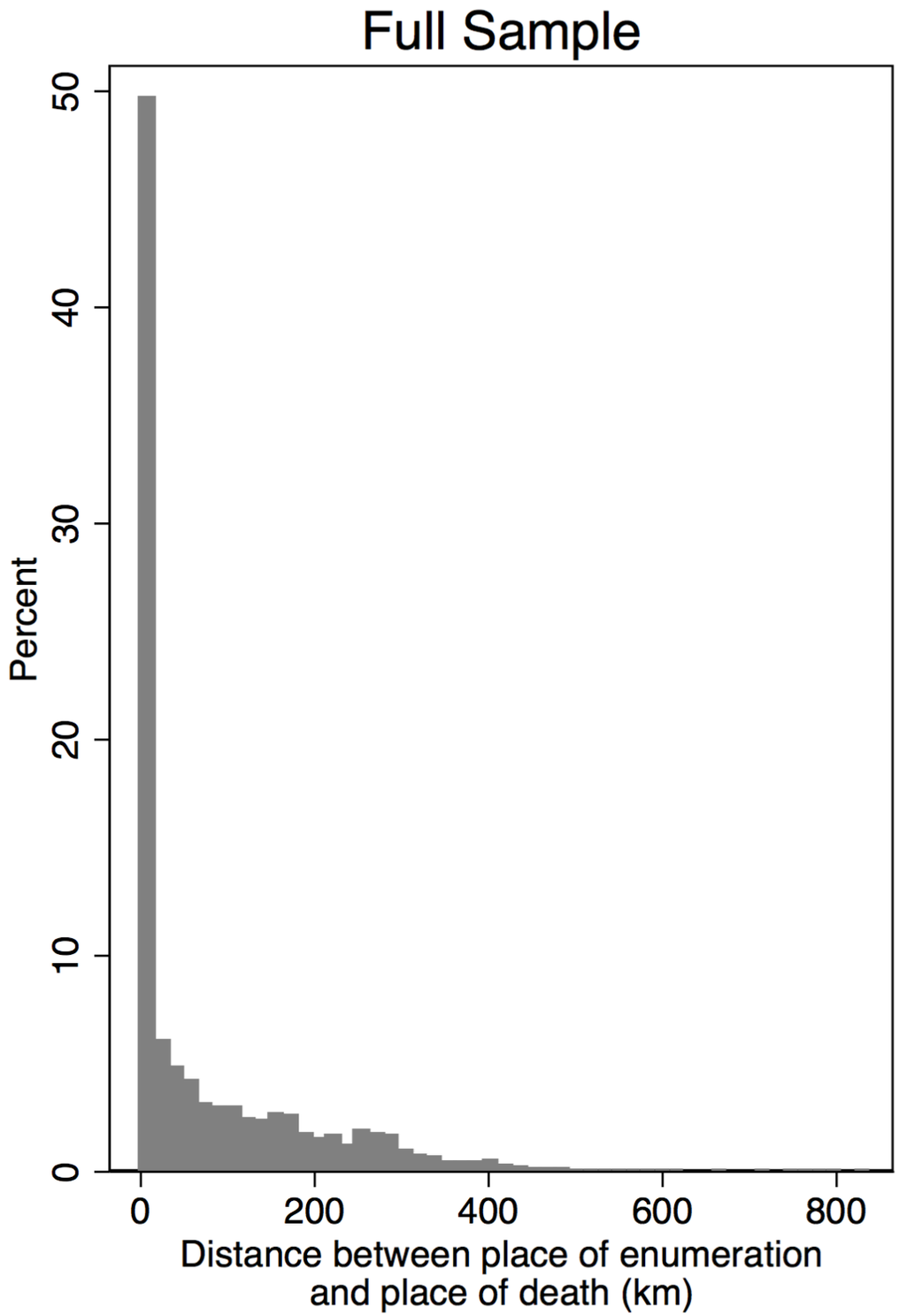
Data and linking procedure

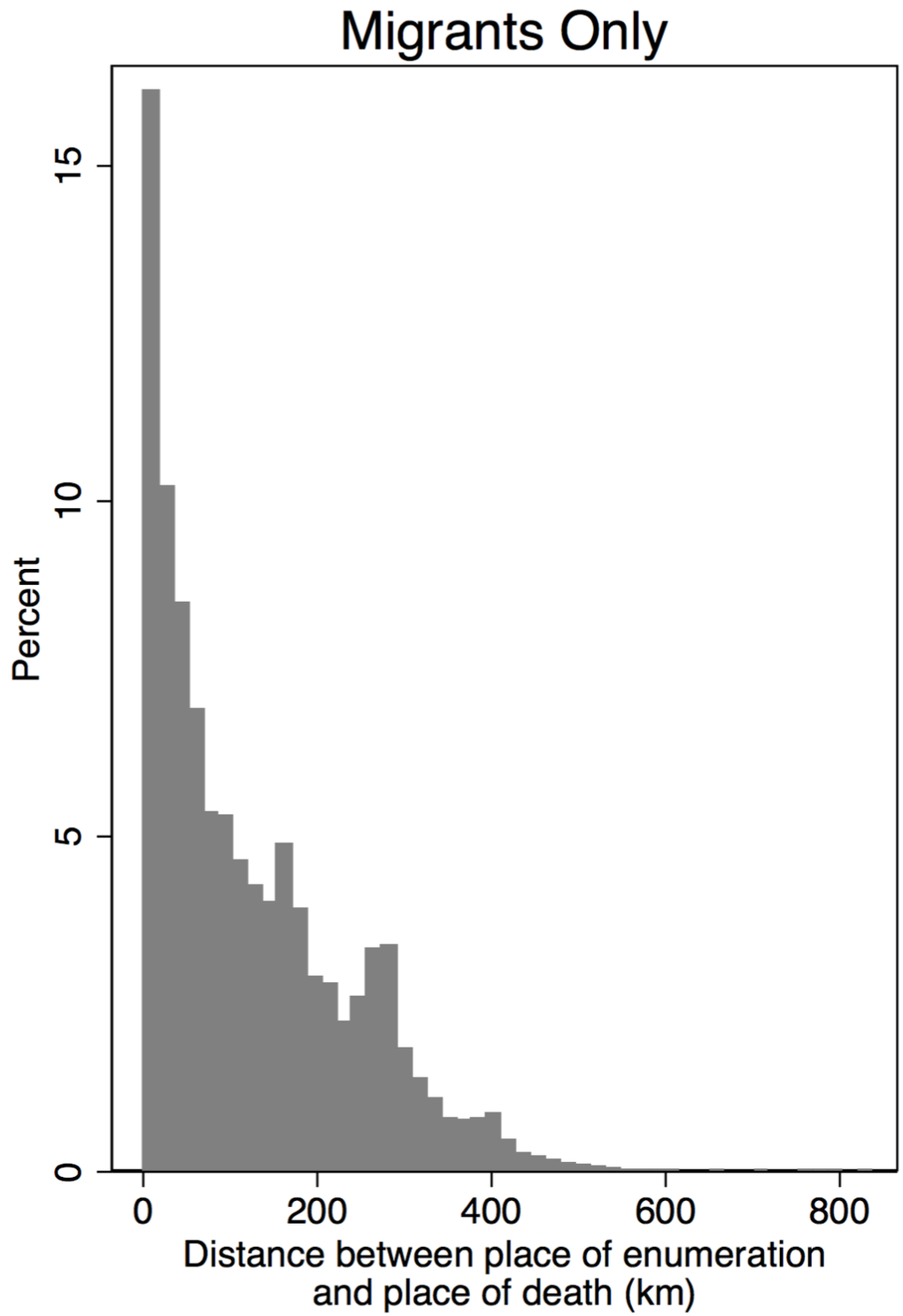
- Two comprehensive, individually-identified, and publicly available sources:
  - 1851 and 1861 Census microdata from I-CeM.
  - Death registry (1851-55 and 1861-65) from freeBMD.org.
- We take unique first and last name strings from death registry and link back to corresponding census.
  - 70,912 (3.6%) of deaths from 1851-55 period matched.
  - 80,646 (3.8%) of deaths from 1861-65 period matched.

A note about false positives

- Ferrie (1996) constructs a dataset linking unique names across censuses and obtains a match rate of 19%.
  - Our match rate among those appearing in the death registry with a unique name is 17%.
  - Similar false positive rates (as estimated in Bailey et al., 2018)
    - This suggests, among those with unique names, birth place and age aren't adding much.
- In our DD framework, false positives should attenuate results as long as they are not systematically related to treatment.

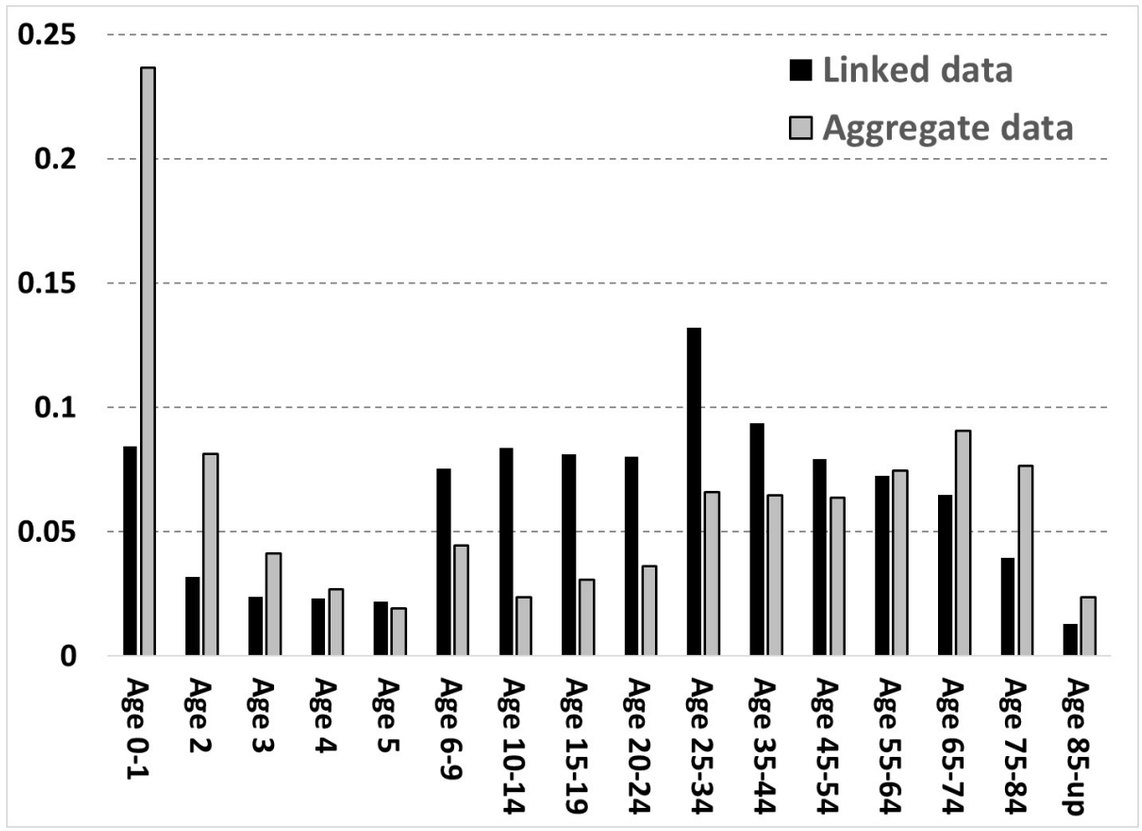
Sample check: Distribution of links by distance



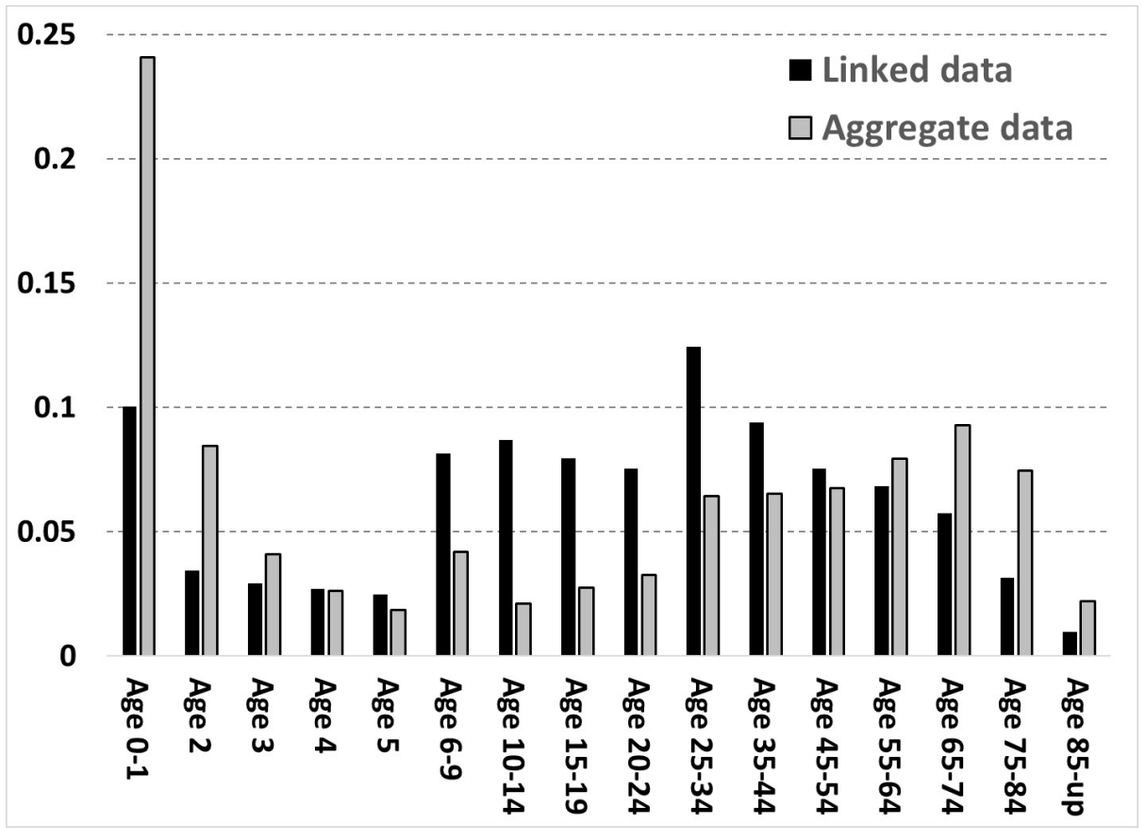


Sample representativeness

1851-55 Period



1861-65 Period



- Infant deaths and elderly deaths are underrepresented in linked data.
- Solutions:
  - Weight observations so that linked sample matches aggregate age distribution.
  - Split the same and estimate age-specific results.
- Sample is largely representative with respect to gender, occupation, and immigrant share.

Final tweaks to our estimating equation

$$\ln\left(\frac{Mort_{dt}}{Pop_{dt}}\right) = \alpha + \beta_1 SHOCK_{dt} + \beta_2 RECEIVING_{dt} + \phi_d + \eta_t + \varepsilon_{dt}$$

$$\widehat{Mort}_{dt} = Mort_{dt} * \lambda_{dt}$$

$$\sum_{i \in \{25,50,75\}} NEAR^{i}_{dt}$$

- λ_{dt} is the share of deaths from district that we successfully link.
- Thus, we must assume changes in λ_{dt} are uncorrelated with the shock.

Results: Mortality by initial place of residence

DV: Std. change in log mortality rate from 1851-5 to 1861-5				
	All ages	**Under 15**	**Age 15-54**	**Over 54**
	(1)	(2)	(3)	(4)
Cotton district indicator	0.823*	0.370	0.642	0.858***
	(0.437)	(0.351)	(0.406)	(0.319)
Nearby (0-25km)	0.446***	0.254*	0.360***	0.416***
	(0.125)	(0.143)	(0.0907)	(0.128)
Nearby (25-50km)	0.384**	0.106	0.295**	0.340***
	(0.153)	(0.137)	(0.126)	(0.129)
Nearby (50-75km)	-0.00220	0.00550	0.0936	0.0122
	(0.153)	(0.153)	(0.129)	(0.158)
Log initial pop.	-0.0762	-0.0928*	-0.122**	-0.115**
	(0.0879)	(0.0510)	(0.0504)	(0.0566)
Age share controls	Yes			
Observations	537	534	535	533
R-squared	0.090	0.020	0.056	0.080
Linked deaths	152,787	55,723	69,978	27,086

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. All-age regressions are weighted by district population in 1851. Age group controls include the share of district residents under 5, the share 5-15, and the share over 50 for both the 1851 and 1861 census.

Robust to: Not weighting, considering a continuous shock variable, excluding large cities, restricting to native born population only.

Results: Mortality by place of death

DV: Std. change in log mortality rate from 1851-5 to 1861-5				
	All ages	**Under 15**	**Age 15-54**	**Over 54**
	(1)	(2)	(3)	(4)
Cotton district indicator	0.00148	-0.266*	-0.180*	0.513***
	(0.131)	(0.160)	(0.101)	(0.123)
Nearby (0-25km)	0.247**	0.105	0.118	0.380***
	(0.104)	(0.116)	(0.114)	(0.112)
Nearby (25-50km)	0.100	-0.161	0.0697	0.393***
	(0.133)	(0.137)	(0.0969)	(0.136)
Nearby (50-75km)	-0.0366	0.237*	0.0275	-0.0226
	(0.203)	(0.134)	(0.199)	(0.143)
Log initial pop.	0.0252	-0.0347	-0.0768**	-0.0492
	(0.0519)	(0.0431)	(0.0360)	(0.0473)
Age controls	Yes			
Observations	538	531	534	529
R-squared	0.038	0.027	0.020	0.049
Linked deaths	152,787	55,723	69,978	27,086
Testing difference relative to results by residence				
Chi-sq. stat.	3.30	3.30	3.63	8.57
P-value	0.0693	0.0695	0.0569	0.0034

*** p<0.01, ** p<0.05, * p<0.1. Bottom panel presents tests comparing the cotton district coefficients estimated in this table to those obtained when deaths are assigned to the initial residence. Remaining table notes as before.

How well does the aggregate approach perform?

- Apples-to-apples comparison:
  - Obtain annual mortality data from Registrar General Reports.
  - Population data from the census
  - For comparability, we generate mortality rates by collapsing deaths by period (1851-55 vs 1861-65).
    - As a robustness check we run the more standard panel approach and find similar patterns.

Aggregate-to-linked comparison: By place of death

DV: Std. change in log mortality rate from 1851-5 to 1861-5						
	Ages under 15		**Ages 15 to 54**		**Ages 55 and up**	
	Agg. data	Linked data	Agg. data	Linked data	Agg. data	Linked data
	(1)	(2)	(3)	(4)	(5)	(6)
Cotton dist. ind.	-0.604***	-0.266*	-0.231	-0.180*	0.508***	0.513***
	(0.136)	(0.160)	(0.179)	(0.101)	(0.141)	(0.123)
Nearby dist.	Yes	Yes	Yes	Yes	Yes	Yes
Observations	531	531	534	534	529	529
R-squared	0.089	0.027	0.120	0.020	0.133	0.049
Testing difference between linked and aggregate cotton dist. coef.						
Chi-sq stat.	2.46		0.10		0.00	
P-value	0.1166		0.7547		0.9728	

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. Regressions are weighted by district population in 1851. All regressions include controls for nearby districts 0-25km, 25-50km and 50-75km away from the cotton districts. The sample size changes slightly across specifications if there are small districts where we do not have links when cutting by age group. At the bottom of the table we present tests comparing the cotton district coefficients estimated in this table to those obtained when deaths are assigned to each person's initial residence location. This test is implemented by running seemingly unrelated regressions.

Aggregate-to-linked comparison: By place of enumeration

DV: Std. change in log mortality rate from 1851-5 to 1861-5						
	Ages under 15		Ages 15 to 54		Ages 55 and up	
	Agg. data	Linked data	Agg. data	Linked data	Agg. data	Linked data
	(1)	(2)	(3)	(4)	(5)	(6)
Cotton district ind.	-0.605***	0.370	-0.231	0.642	0.510***	0.858***
	(0.136)	(0.351)	(0.178)	(0.406)	(0.138)	(0.319)
Nearby dist.	Yes	Yes	Yes	Yes	Yes	Yes
Observations	534	534	535	535	533	533
R-squared	0.087	0.020	0.120	0.056	0.137	0.080
Testing difference between linked and aggregate cotton dist. coef.						
Chi-sq stat.	7.45		4.20		0.85	
P-value	0.0064		0.0404		0.3559	

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. Regressions are weighted by district population in 1851. All regressions include controls for nearby districts 0-25km, 25-50km and 50-75km away from the cotton districts. The sample size changes slightly across specifications if there are small districts where we do not have links when cutting by age group. At the bottom of the table we present tests comparing the cotton district coefficients estimated in this table to those obtained when deaths are assigned to each person's initial residence location. This test is implemented by running seemingly unrelated regressions.

Standard aggregate panel results

Sample:	**Dependent variable is log mortality rate** **1851-1871**			
	(1)	(2)	(3)	(4)
Cotton district ind.	-0.0437***	-0.0460***	-0.0423***	-0.0468***
	(0.00785)	(0.00717)	(0.00789)	(0.00724)
Nearby (0-25km)			0.00314	-0.00484
			(0.0112)	(0.00873)
Nearby (25-50km)			0.00114	-0.0120
			(0.0140)	(0.0109)
Nearby (50-75km)			0.0176**	0.0143*
			(0.00826)	(0.00754)
Time Trends	N	Y	N	Y
Observations	10,780	10,780	10,780	10,780
R-squared	0.139	0.256	0.140	0.256
No. districts	539	539	539	539

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses

Microdata advantage: Incidence

- Census data allow us to identify household structure and occupation.
- Given that the shock was specific to the cotton industry, how much of the effect is driven by cotton workers?

Share of deaths accounted for by cotton households

	Cotton workers	Households with at least one cotton worker	Non-cotton workers living in a household with a cotton worker
1851	0.0085	0.0356	0.0271
1861	0.0149	0.0449	0.0300
Increase (p.p.)	0.00638	0.009304	0.002925
Percent increase	75.21%	26.16%	10.80%

Mortality results removing cotton workers and households

DV: Std. change in log mortality rate from 1851-5 to 1861-5	Full linked sample (1)	Dropping cotton workers (2)	Dropping all cotton hh members (3)
Cotton district indicator	0.823*	0.726	0.586
	(0.437)	(0.443)	(0.591)
Nearby (0-25km)	0.446***	0.436***	0.425***
	(0.125)	(0.127)	(0.160)
Nearby (25-50km)	0.384**	0.380**	0.424***
	(0.153)	(0.154)	(0.141)
Nearby (50-75km)	-0.00220	-0.00283	-0.0130
	(0.153)	(0.155)	(0.161)
Log initial pop.	-0.0762	-0.0813	-0.115
	(0.0879)	(0.0887)	(0.0949)
Age share controls	Yes	Yes	Yes
Observations	537	537	537
R-squared	0.090	0.079	0.058

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. Regressions are weighted by district population in 1851. Age group controls include the share of district residents under 5, the share 5-15, and the share over 50 for both the 1851 and 1861 census.

Conclusion

- The Lancashire Cotton Famine offers a unique setting for analyzing whether this migration bias can undermine inference.
  - Comprehensive and publicly available microdata.
  - Place of residence identified just before the recession.
  - Geographically concentrated shock that allows us to cleanly identify migration flows.
- We show that: unobserved migration can lead to incorrect conclusions when using aggregate data.

Conclusion

- These findings offer an explanation for the aggregate vs microdata puzzle.
- Relatedly, studies that analyze the relationship at higher geographic levels also find weaker evidence of pro-cyclicality (Lindo, 2015 and Cutler et al., 2016).
- Declining rates of internal migration may help explain decline in pro-cyclicality documented in Ruhm (2015).