Urban Migration, Public Health Amenities, and Local Newspapers in 1870-1940 U.S.

Anaïs Galdin and Quan Le[†]

Click for Latest Version

December 16, 2024

Abstract

We study the role of newspapers in the migration decisions of rural individuals in the 1870-1940 United States. Contributing to the urban and health economics literature, we offer new insights about how information, specifically newspaper portrayals of public health advancements like water filtration and sewage systems, shapes rural-urban migration patterns. We show that access to information about urban health conditions through newspapers played a crucial role in encouraging rural individuals to move to cities over the turn of the 20th century. By exploiting a novel linkage between full-count U.S. census data and a unique historical newspaper dataset at the county level, which leverages text-mining and Natural Language Processing (NLP) algorithms, we provide novel evidence that rural individuals responded to newspaper narratives on public health investments and disease occurrences. Rural migrants with access to information about public health investments migrated in higher proportion to sanitation-adopting cities compared to rural migrants with no access to such information, and avoided cities affected by epidemics more than their non-informed counterparts. This finding is consistent with the literature suggesting that such investments significantly reduced mortality from waterborne diseases (typhoid and diphtheria) and improved quality of life, making cities more attractive places to live. As policymakers consider strategies to revitalize urban areas in the post-pandemic era, our study highlights the potential importance of information dissemination in promoting urban growth and development.

Keywords: Urban Migration, Public Health Amenities, Local Newspapers

JEL Codes: J61, I18, R23, L82

[†]Department of Economics, Princeton University, agaldin@princeton.edu and quanle@princeton.edu. We are grateful to Leah Boustan, Janet Currie, Kate Ho, Eduardo Morales and Steve Redding for their continual guidance and support on this project. We also thank Nick Buchholz, Jacob Dorn, Gene Grossman, Jakub Kastl, Ilyana Kuziemko, Alessandro Lizzeri, Chris Neilson and Tianyi Wang for helpful discussions and comments. This research benefited from financial support from the International Economics Section (IES) at Princeton University, whose support is gratefully acknowledged. All errors are our own.

Contents

1 Introduction

Potential migrants likely do not have all relevant information about each possible destination. This was particularly true for most of history, when search and information acquisition was relatively costly. This friction suggests a potentially important role for the news media in historical migration: learning about the characteristics of other places from the news may change one's beliefs about those places and subsequently the decision of whether and where to migrate.

We study the effects of local newspapers on urbanization via the provision of information to rural migrants about urban amenities. We focus on the U.S. between 1870 and 1940 for two reasons. First, this was a period of rapid urbanization in the U.S., when the percentage of the population living in urban areas increased from 22% to 55%. Second, local newspapers were by far the primary available news media in our setting, as TV had not yet been popularized and radio rarely engaged in news programming. This setting provides us ample variation in exposure to local news and in urban migration.

The urban amenity of interest to our study is the public health conditions in a city, in particular the incidence of waterborne diseases. Waterborne diseases were an important concern for urban life at the time due to their high death rates, severe symptoms, and their ability to disrupt life and cause massive income loss (Anderson et al., 2022; Beach, 2022b). In 1900, diseases such as typhoid, dysentery, and diarrhea were responsible for nearly 200 deaths per 100,000 residents in the U.S., with most deaths happening during mass outbreaks. Our study period also coincides with a wave of investments in public sanitation projects aimed at eradicating these diseases, such as a water filtration plant or sewage system. Sanitation projects mostly eliminated outbreaks, greatly decreased the incidence of disease, and reduced urban mortality. In this paper, we explore the possibility that rural migrants who learned of improved urban conditions due to these projects from local newspapers could have been induced to move there.

We micro-found the migration decision of a individual using a discrete-choice framework inspired by Berry (1994). To evaluate the importance of different urban characteristics, we explicitly specify the utility of the individual across potential locations to include characteristics such as wages, sanitation, and disease rates. We allow individuals' salience about characteristics to vary depending on whether they have access to a local newspaper. Our empirical implementation includes origin-time and destination-time fixed effects and so uses bilateral variation across origin-destination pairs to identify the importance of characteristics and information access. Specifically, we identify the role of information in migrants' decision by comparing migration rates across destina-

tions with different characteristics *within* an origin at a given time, after controlling for unobservable destination characteristics that are constant across time.

We apply this framework to historical migration flows in the 1870-1940 U.S., using full-count U.S. censuses from IPUMS (Ruggles et al., 2024a; Ruggles et al., 2024b). We use record linkage techniques from Abramitzky et al. (2020) to match people over time in the decennial census data. Our data on public health conditions include typhoid death rates from Beach (2022b) and the implementation of water filtration plants as well as other public health intervention such as water chlorination, tuberculosis testing, and sewage systems from Ao (2015) and Anderson et al. (2022). To proxy for the availability of information to rural migrants, we use the United States Newspaper Panel from Gentzkow et al. (2011) to detect the presence of a newspaper. We directly augment this data with keyword searches on the Chronicling America database from Dell et al. (2023) to measure instances of news coverage of waterborne diseases and changes in public health infrastructure.

First, we present evidence linking the significance of sanitation systems, news coverage of urban health, and rural-urban migration. Utilizing our data on migration flows, we support earlier findings that migration is primarily local (Eckert and Peters, 2018), noting the average mover relocated just 9 miles from their origin. We also find that a significant driver of urbanization during this period was rural-urban migration, rather than direct growth in urban areas. We then demonstrate that a wave of investments in sanitation systems coincided with urbanization between 1870 and 1940. Together, these results indicate that rural migrants' access to information about urban destinations' sanitation systems is a plausible factor in their migration decisions.

We further provide evidence that water filtration meaningfully changed how the news cover an urban destination's quality of life. An informal audit of our news corpus reveals that direct coverage of water filtration systems that reflects improvement in urban amenity was minimal. On the other hand, these improvements were reflected in news coverage via a large drop in the amount of news coverage of waterborne diseases. In an event study, we find that water filtration systems significantly decreased the number of mentions of typhoid related to an urban destination. We repeat this exercise for a set of non-waterborne diseases and find that, conversely, these diseases received increased coverage after the implementation of water filtration. This suggests that newspapers shifted their coverage away from waterborne diseases after water filtration.

We then implement our primary empirical model, motivated from our discrete choice model of an individual's migration decision. We find that on average, the implementation of a water filtration plant at an urban destination increases the share of residents

from any origin moving to that destination by 74%, while the joint presence of a newspaper at an origin and a filtration plant at the destination increases this share by an additional 25%. This result supports our hypothesis that while water filtration and public health reflect an important amenity for rural-urban migrants, migrants with access to newspapers at their local origin are potentially more salient to this amenity change and thus more likely to move there.

We augment our primary results with two additional empirical specifications: instead of using the implementation of a water filtration plant as a proxy for urban amenity, we use first, a direct measure of destination typhoid death rate and second, a measure of the intensity of news coverage about typhoid at an urban destination. Since migration toward an urban destination can also cause increased news coverage, we use the variation in disease news predicted by the implementation of a water filtration system for our preferred specification. We find that on average, a standard deviation increase in typhoid mortality rate leads to a lower in-migration rate for a destination and even lower from origins with newspaper access, although this result is not statistically significant in our preferred specification. On the other hand, we find that individuals strongly respond to typhoid news coverage: an increase of one standard deviation leads to a 61% lower migration share from an origin with newspapers.

Throughout our empirical results, we study the bilateral variation in migration between origin-destination pairs in response to bilateral variation in destination amenities and origin newspaper presence, after controlling for origin-time and destination-time fixed effects. This helps us overcome potential concerns about unobserved push factors that are common across destinations (or pull factors that are common across origins). However, concerns about pair-specific unobservables remain. For example, if origins that have newspapers and destinations that have water filtration systems are more similar in characteristics, they might experience higher migration rates. We partially alleviate these concerns by including pair-level congruence measures of observable demographics in the Census. Data limitations mean that potentially unobservable congruence among characteristics that are not captured by Census demographics can still bias our results.

Another important concern is that we cannot rule out other urban attributes that may have changed at the same time as public health investments. These include changes in productivity, job opportunities, or leisure amenities at a destination city, to which individuals could be responding rather than to public health conditions directly. This concern limits our ability to make claims about the precise mechanisms through which public health and information interact to increase urbanization. Nonetheless, our results indicate that public health was an important component of urban amenities and that

migrants' response to amenity changes depends on access to information through local newspapers.

Related Literature Our paper builds on a literature that found relatively low migration response to local shocks, suggesting the existence of significant barriers to migration (Porcher, 2020 WP). A set of recent papers have found evidence for information frictions to be a major cause (e.g. Wilson (2022) on migration responses to fracking in the U.S., Porcher (2020 WP) on broadband internet in Brazil). In addition, we show that information frictions hinder the extent to which individuals can migrate in response to changes in amenities. While we do not provide a model of endogenous amenity investment in our setting (e.g. Almagro and Lino (2022 WP) or Diamond (2016)), our results suggest that incorporating information can be a valuable addition for future work in that direction.

Our work also relates to a literature on historical urbanization in the U.S. (see Boustan et al. (2018) for a survey) and that on the importance of sanitation systems and public health (Beach, 2022b; Anderson et al., 2022; Cutler and Miller, 2005). We show that the implementation of a sanitation system in a city likely induced in-migration from rural areas and that this effect is heterogeneous depending on the information environment experienced by potential migrants. Finally, we contribute to a literature that studies the role of historical newspaper content in influencing societal outcomes. In particular, we corroborate the findings of Costa and Kahn (2017) that the historical news media focused more on "negative" news of diseases rather than "positive" news of decreased death rate and improved conditions in the city. We extend this finding by showing that the migration decisions of individuals in fact respond to the absence of disease news, suggesting that individuals can rationally update on the conditions of the city despite the improvement not being directly covered.

Our paper proceeds as followed. Section 2 describes the data. Section 3 describes the historical background for our study. Section 4 describes our theoretical and empirical model and a discussion of our identification strategy. Section 5 describes our results and robustness tests. Section 6 concludes.

2 Data and Descriptives: Sanitation Systems and Urbanization

In this paper, we measure the salience of health-related events in local newspapers and capture migration responses to health-related destination amenities across different in-

formation environments. We first construct a novel dataset that leverages full-count census data in order to build historical migration flows across U.S. counties and cities. We then pair rural-urban migration flows with data on public health investment and typhoid death rates in cities, as well as a large corpus of digitized local newspaper text from 1870 to 1940.

2.1 Rural-Urban Migration Flows

This section first describes how we constructed bilateral migration flows from historical U.S. censuses, before drawing key statistics related to rural-urban migrations around the turn of the 20^{th} century.

2.1.1 Building Migration Flows from Historical U.S. Censuses

In order to create a measure of bilateral migration flows within the U.S., we use historical full-count decennial censuses for years 1870 to 1940 from the Integrated Public Use Microdata Series (IPUMS).[1] Each digitized full-count census provides geocoded individual-level data, including details on education, race, income, occupation, family, health and residence. For the period 1850-1940, the full-count censuses contain over 650 million individual-level records.

Census Linking. In order to build intra-U.S. migration flows, we rely on the Census Linking Project (Abramitzky et al., 2020, CLP)[2] to link individuals between adjacent pair of complete-count census years. Matches between adjacent census years are made based on an individual's first and last name, age and state/country of birth. In order to limit false positives in migration events, our main results exploit the most conservative matching algorithm from the CLP (ABE-exact), which requires a unique and exact match of name and age within a 5-year age band.[3]

Our results are robust to the use of the use of the three alternative automated matching methods from the CLP. Depending on the specific matching algorithm, the linked samples used in our analysis include between 10 and 20% of the total U.S. population compared to the full-count census.[4] Appendix A further details sampling bias from

[1] At the time of this study, the latest full-count census being published for the U.S. is for the year 1950, as full-count censuses have to wait 72 years before becoming publicly available. The 1960s one should be released in 2032.

[2] https://censuslinkingproject.org provides a series of fully anonymous crosswalks between each pair of censuses.

[3] The approach is described in more detail in Abramitzky et al. (2021).

[4] The sample only includes male individuals due to the nature of the data, as tracking females through adjacent census decades is complicated by changes of family names post marriage. The Census Linking

IPUMS and provides several sampling statistics that support that the distribution of the population in our matched census data is closely aligned to the original full-count census from IPUMS. The final sample contains 617,483,404 unique matched individuals over the 1870-1940 period.

Bilateral Migration Flows. We measure migration flows from an origin o to a destination d for an intercensal period (10 years) as the share of individuals living in o at the beginning of the period that we find in location d at the end of the period. Our primary outcome for migration flows is thus equivalent to computing the empirical probability that a person moves from an origin location o to a destination d during an intercensal period t, allowing for people to stay put:

$$m_{odt} = \Pr[\text{individual moves from } o \text{ to } d \text{ in period } t]$$

For instance, if o is Mercer county and d is Trenton city in New Jersey, and t captures the intercensal period 1900-1910, m_{odt} captures the share of Mercer county residents in 1900 found in Trenton in 1910.

As our measure of migration is constructed from moves across linked full-count U.S. censuses, it is limited to intra-U.S. migrants, and the set of origin locations o only captures U.S.-based locations and do not include international migrants. The data, however, captures the birth location of individuals and classifies them into foreign-born, native born, first-generation and native born. We use these individual-specific observables as controls in our regression framework.

Rural-Urban Migrants. We define a rural-urban migrant as an individual who moved from a rural county to an IPUMS location identified as a "city".[5] Figure 1 plots the distribution of the population in the full-count U.S. census for different type of geographical units (county, city, urban area). The accompany table to Figure 1 details the share of all individuals from the full-count census living in an area that is not identified as a city. This share decreases from around 80% of the population in 1870 to 50% in 1940.

Figure 2 compares our newly created measure of urban migration flows to the share of population that is urban in the matched full-count census data. The urban population

Project thus only provides links for men on the site and our our analysis is consistent with the literature using the linked census data. It moreover does not allow to capture international migration flows.

[5]IPUMS defines a city based on its population: a place is referred to as a city census if it has more than 10,000 contemporary inhabitants in the 1870 and 1880 censuses, and more than 25,000 contemporary inhabitants from 1900 to 1940. Our results are qualitatively similar if we use the less restrictive IPUMS definition of an "urban" county. We prefer the city definition as it allows us to generates a fixed and comparable set of cities across our sample. Appendix C.2 describes in details the definitions of "cities" and "urban/rural areas" used by the full-count U.S. census.

Year	Share of Rural Pop.
1870	0.814
1880	0.784
1900	0.671
1910	0.630
1920	0.576
1930	0.525
1940	0.543

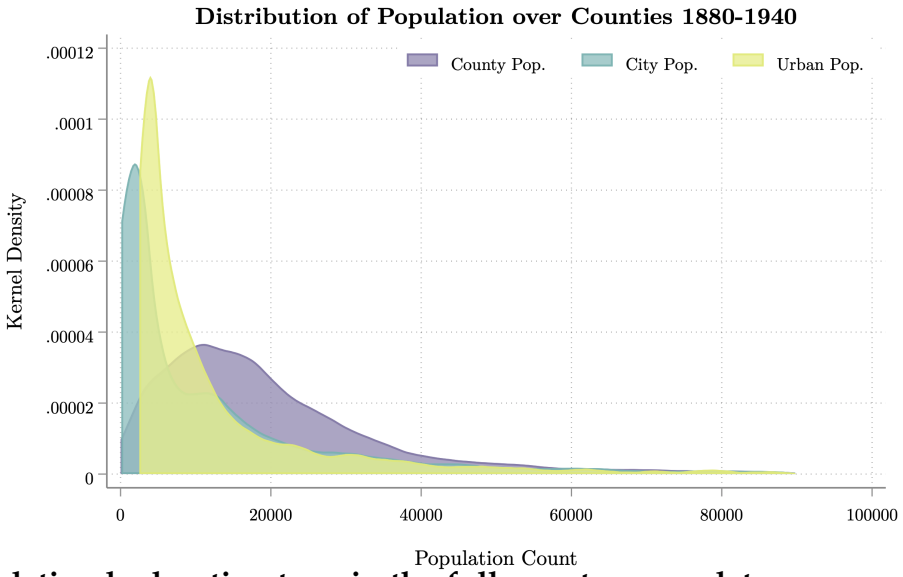

Figure 1: **Population by location type in the full-count census data**

Notes: Figure 1 is based on the unlinked version of the full-count U.S. census data from IPUMS. A "city" is (mostly) defined as any location with 10,000+ inhabitants. "Urban" areas denote all cities and incorporated places of 2,500+ inhabitants. All areas not classified as urban are designated rural. Individuals living in smaller, non-urban places are assigned to a state, a county, and a "non-identifiable city" code. The left table details the share of IPUMS Population living in a "non-city" area. Appendix C.2 describes in details the definitions of "cities" and "urban/rural areas" used by the full-count U.S. census.

and the urban migration shares are closely related, both in terms of magnitude and time trend, which supports the adequacy of our urban migration measure. Appendix figure A.1 further supports that our matched census dataset is representative of the population distribution of the full-count census, with a correlation coefficient above 0.93 both at the county and city levels. Appendix Section A provides several summary statistics about sample selection.

2.1.2 Statistics on Migration Flows

Migration is a local phenomenon. Acquiring information about different potential destinations has historically been costly. It is thus without surprises that migrations in the U.S. primarily remained a local phenomenon between 1870 and 1940 (Figure 5.1). The average mover in our linked census dataset only moved 9 miles away from their origin location between two census years, with a standard deviation of 10 miles. This is consistent with the findings of Eckert and Peters (2018) who finds that American industrialization was also highly local, and suggests that access to information may be an important barrier to migration and integration of labor forces.

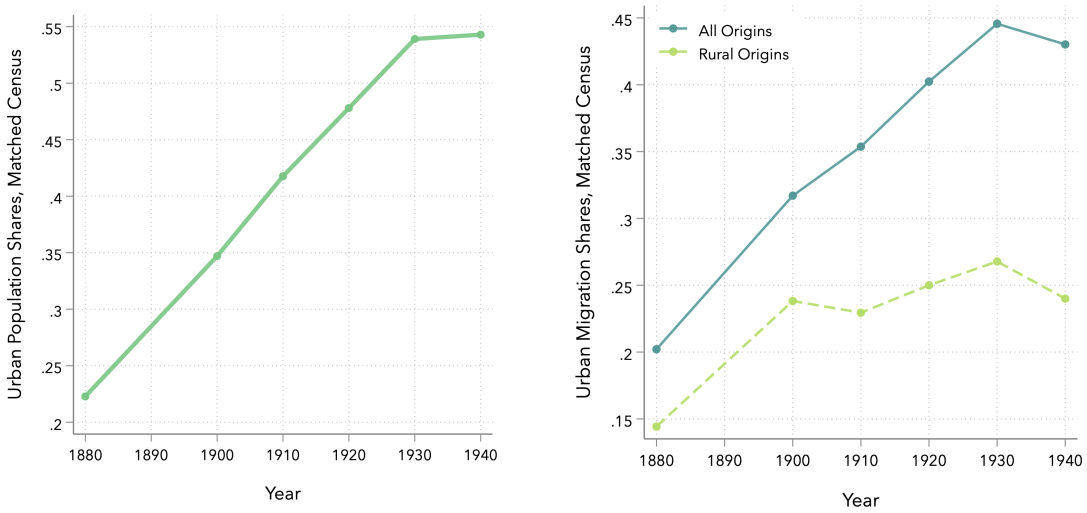

Figure 2: **Urbanization is closely tied to urban migration**

Notes: The left figure plots the share of the entire U.S. population living in urban areas, as defined by IPUMS. The right figure plots the share of urban migrants in our data, splitting migrants by type of origin location. Both measures of urban population shares and urban migrations are computed based on the linked version of the full-count U.S. census, using the Census Linking Project. "Urban" areas denote all cities and incorporated places of 2,500+ inhabitants. All areas not classified as urban are designated rural. Before 1950, the urban share only includes residents living in incorporated places. Appendix C.2 describes in details the definitions of "cities" and "urban/rural areas" used by the full-count U.S. census. The slight flattening of the trend from 1930 to 1940 may be due to the rise of unemployment in cities during the Great Depression. Looking at urban population shares over longer time series (see e.g. Boustan et al. (2018)) shows that the upward trend continues after 1940 and this flattening is temporary.

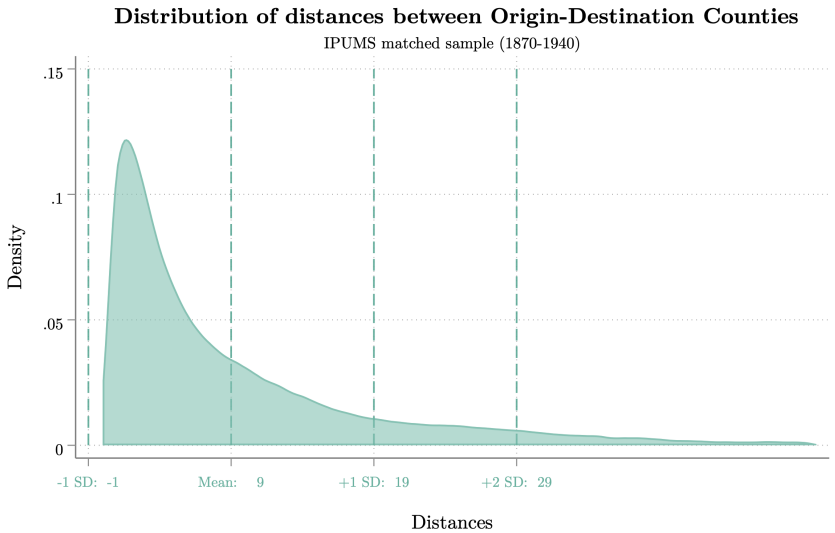

Figure 3: Migration is mostly a local phenomenon

Notes: Figure 5.1 reports the distribution of distance between origin and destination counties/cities for individuals identified as movers in the matched census data. The average mover in the linked census only moved 9 miles away from their origin location between two census years, with a standard deviation of 10 miles. Intra-U.S. migrations primarily remained a local phenomenon between 1870 and 1940.

Share of Urban Movers. Migration flows to cities accelerated over the turn of the 20th century (Table 1). Most individuals who moved before 1920 where going from rural areas to urban areas, or from non-cities to cities. Starting around 1920-1930, as more people live in urban counties, this trend reverses and migration flows become mostly urban-to-urban and city-to-city. An important caveat is that since we do not have international migrants in our data, these figures may be understating the extent of urbanization in this period and international arrivals' contribution to that process.

Table 1: **Number and Share of Movers to Urban and City Areas, by Year**

Year	Urban Areas	Rural to Urban	Urban to Urban	City Areas	Non-City to City	City to City
			Total Number of Movers (Volumes)			
1880	227554	123403	104151	178351	113305	65046
1900	549116	360296	188820	446615	316655	129960
1910	928953	480752	448201	718322	416061	302261
1920	1.420e+06	669103	751136	1.138e+06	639412	499087
1930	1.979e+06	867559	1.111e+06	1.635e+06	872126	762589
1940	1.753e+06	661979	1.091e+06	1.372e+06	636419	735502
			Share of Movers (Among all Movers)			
1880	0.253	0.137	0.116	0.198	0.126	0.0723
1900	0.396	0.260	0.136	0.322	0.228	0.0938
1910	0.437	0.226	0.211	0.338	0.196	0.142
1920	0.512	0.241	0.271	0.411	0.231	0.180
1930	0.572	0.251	0.321	0.473	0.252	0.221
1940	0.521	0.197	0.324	0.407	0.189	0.218
			Share of Movers (Among the whole IPUMS Population)			
1880	0.0868	0.0470	0.0397	0.0680	0.0432	0.0248
1900	0.176	0.116	0.0606	0.143	0.102	0.0417
1910	0.146	0.0756	0.0705	0.113	0.0654	0.0475
1920	0.169	0.0797	0.0894	0.136	0.0761	0.0594
1930	0.186	0.0817	0.105	0.154	0.0821	0.0718
1940	0.140	0.0527	0.0868	0.109	0.0507	0.0585

Notes: The second (middle) part of the table reports the share of all movers that goes from a given type of origin to a given type of destination (e.g.: total number of movers from rural to urban areas in census year t divided by the total number of movers in year t, for all possible origins-destinations). The share of movers going to urban areas increase over the period, and while the early 20^{th} century migrations were mostly from rural to urban areas, migrations become increasingly urban-to-urban starting around 1920-1930. The third (bottom) part of the reports the share of movers as a percentage of the whole U.S. population from census (for instance, total number of movers to cities in year t divided by the total U.S. population in year t).

2.2 Public Health Investments and Waterborne Diseases

In this subsection, we present the historical background that motivates our empirical analysis. First, we describe the state of urban public health in the U.S. and investments in public health infrastructure during our sample period. Second, we describe our public health investment and waterborne disease data. Third, we describe the relationship between urbanization and public health investments.

2.2.1 Waterborne Diseases in the 1870-1840

For most of American history, cities were deadly places. Important killers of urban dwellers include waterborne diseases which spread via water contaminated with bacteria or parasites such as typhoid, cholera, and dysentery. In 1900, these diseases combined caused approximately 186 deaths per 100,000 inhabitants in the U.S. (Troesken, 2001).[6]

These diseases were salient for reasons beyond their high death rates. For example, consider typhoid which accounted for roughly 2.5% of total mortality during this period (Cutler and Miller, 2005). Infection typically occurs via drinking water tainted with infected feces. Symptoms include muscle ache, diarrhea, severe fever, delirium, internal hemorrhage, and increase the risks of pneumonia and tuberculosis during the disease. The brutality of the symptoms is intensified by their duration: the disease's active period can last from three to four weeks, but the recovery period can be as long as four months. Reinfections were common and there was no effective treatment or vaccine. Furthermore, epidemics were common (Troesken, 1999). Given the severity, length, and frequency of the disease, typhoid was often a signpost for the filthy environmental conditions of the city.

2.2.2 Public Health Developments in the 1870-1840

As of the late 19th century, public health officials understood that cleaner water, e.g. via water filtration or chlorination, could eradicate typhoid and other waterborne diseases (Troesken, 1999). This led to a wave of investments in these structures around the turn of the 20th century in the United States (Figure 4). There is an active debate in the academic literature on just how much of the decrease in mortality in the U.S. during this period was due to these investments, with estimates ranging from 20 to 50%. (Cutler and Miller, 2005; Beach, 2022b; Anderson et al., 2022).

[6]For comparison, COVID-19 caused roughly 200 deaths per 100,000 inhabitants in the U.S. between February 2020 to November 2021. (CDC 2021)

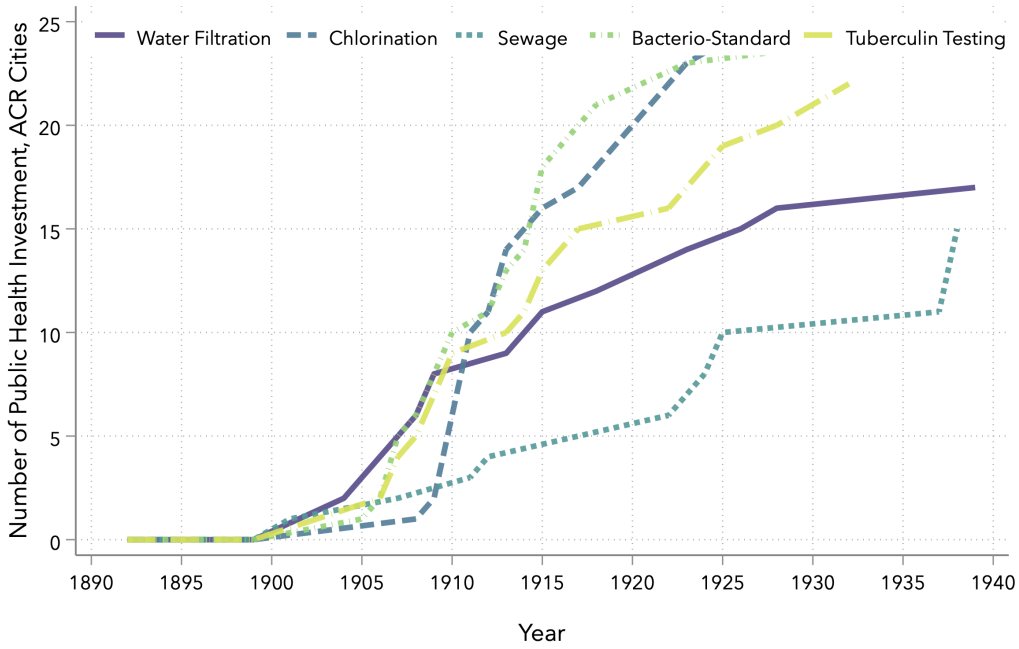

Figure 4: Public Health Investment over the beginning of the 20^{th} century

Notes: This figure reports the cumulative sum of U.S. cities adopting public health investment (PHI) systems in the U.S. The subset is limited to the 25 cities studied by Anderson et al. (2022), splitting the data by type of PHI. The figure emphasizes that most of the changes occurred during the first 20 years of the 20^{th} century. Nearly all cities in the study had adopted bacterio-standard, chlorination and tuberculin testing by 1940. The timing of water filtration and sewage systems was more idiosyncratic, with almost half of major U.S. cities not having implemented such systems by 1940.

Several key public health developments, combined with advancements in microbiology and epidemiology, helped significantly reduce the incidence of waterborne diseases like typhoid or cholera in the United States during the 1870-1940 period. The introduction of water filtration systems was one of the most significant step in reducing waterborne diseases. In the late 19^{th} century, cities began implementing sand filtration systems, which helped remove bacteria and other contaminants from the water supply. For example, in 1872, Poughkeepsie, New York, became the first U.S. city to use a sand filtration system for its public water supply. Chlorination also proved to be an effective method for disinfecting water and reducing the spread of waterborne diseases like typhoid fever and cholera. In 1908, Jersey City, New Jersey, became the first U.S. city to use chlorination for water treatment. The development of piped water systems, such as the Croton Aqueduct in New York City (completed in 1842), played another crucial role in providing clean water to urban populations. These systems helped reduce the reliance on contaminated water sources, such as wells and rivers, which

were often breeding grounds for waterborne diseases. Finally, the implementation of sewage treatment systems was another important factor in reducing waterborne diseases. By treating sewage before releasing it into water bodies, cities could reduce the contamination of drinking water sources. In the early 20th century, cities began adopting activated sludge systems and other advanced sewage treatment methods. For a literature review and history of the politics of water infrastructures in the U.S., starting with early waterworks and later sanitation efforts, see Beach (2022a).

Timing of PHI Adoption. The literature on public health argues that the timing of adoption of these water filtration systems in U.S. cities is *exogenous* to the health status of the city, due to political, financial, and legal situations (Anderson et al., 2022; Cutler and Miller, 2005). The average central water filtration system costed one to two times the average municipal budget to implement. The link between waterborne diseases and contaminated water was also not widely acknowledge before the end of the 19th century. For most cities, water filtration systems thus took a long time to plan and yet could suddenly be implemented, so that among all potential candidate cities for public health investment, the timing of installation is exogenous to health outcomes.

Due to these frictions in adoption, it is difficult to generalize a geographic pattern according to which cities adopted water filtration systems and, more generally, clean water supply. Previous work has identified factors that leads to higher probability adoption including segregation (Beach et al., 2022) and privatization of waterworks (Troesken, 1999). While a city's lack of access to alternative, untainted water sources in principle could drive filtration efforts, but cities also often filtered clean water sources for amenity values (such as taste, Beach (2022a)). The lack of a clear geographic pattern of adoption further motivates us to use destination fixed-effects in our analysis as we cannot provide plausible geographic control variables for why one city would have invested in clean water infrastructure over another.

2.2.3 Data on Public Health Investment and Typhoid Outbreaks

Public Health Investments Data. We use two main sets of data to measure public health investments in the United States over the period 1870-194. The first dataset comes from Anderson et al. (2022), which provides several public health investments measures for 25 U.S. cities. These investments include water filtration, water chlorination, clean water projects such as aqueducts, milk standards (such as tuberculin testing for cows and bacteriological standards), and sewage systems.

As this initial dataset is limited to 25 U.S. cities, we complement the dataset with

data related to water filtration systems from Ao (2015). This dataset records the year of installation of these systems for 74 cities located across 30 U.S. states. The creation of the dataset involves compiling data from historical surveys and censuses of filtration plants published by American Water Works Association as well as supplementary textbooks on water filtration practices. The sample of the cities was conditional on having a population of over 30,000 in 1920, having information on water filtration plant installation dates available, and having existed in the 1880 census.

Table 2 describes these two data sources across counties as well as the characteristics of those counties, using covariates from ICPSR. By the end of the 1870-1940 period, 59 cities out of the 74 had installed a filtration system. This represents approximately 1.5% of our county-year observations.

County-level Covariates. We complement the data on public health investment with historical county-level covariates from the Inter-university Consortium for Political and Social Research (ICPSR, Haines and Inter-university Consortium for Political and Social Research (2010)). The covariates include the population share that is white, male, foreign-born, urban, working in the manufacturing sector, literate, and the log manufacturing output per capita.

Table 2: **Descriptive Statistics: Public Health Investments**

	Mean	SD	Description
Public Health Investment			
Sewage Treatment/Diversion	0.024	0.152	= 1 if municipality had a sewage treatment plant or diverted sewage away from drinking water supply, = 0 otherwise
Filtration	0.015	0.123	= 1 if municipality had a water filtration plant, = 0 otherwise
Chlorination	0.033	0.179	= 1 if municipality chemically treated water supply, = 0 otherwise
Clean Water Project	0.022	0.146	= 1 if municipality had completed a clean water project, = 0 otherwise
Bacteriological Standard	0.024	0.154	= 1 if municipality set bacteriological standard for milk supply, = 0 otherwise
TB Test	0.023	0.151	= 1 if municipality required tuberculin testing of cows, = 0 otherwise
County-level Covariates			
Diff. in % Foreign, Dest.	-0.267	0.120	Difference in share of County population foreign born between year t and t-4
Diff. in % Manufacturing, Dest.	-0.210	0.204	Difference in share manufacturing in the County between year t and t-4
Diff. in % Male, Dest.	-0.163	0.128	Difference in share of male population in the County between year t and t-4
Diff. in % White, Dest.	-0.084	0.171	Difference in share of white population in the County between year t and t-4
Diff. in % Literacy, Dest.	-0.515	0.124	Difference in share of literate population in the County between year t and t-4
County $\times$ **Year**	4122317		

Typhoid data. Our data on typhoid epidemics come from Brian Beach (Beach, 2022b; Beach et al., 2016). Typhoid rates are measured as deaths per thousand persons. The typhoid rate measures is a consistent yearly series (1880-1930) that combines information

from Whipple with information from the mortality statistics. The data is at the city level, for 76 U.S. cities. Figure 5 reports the evolution of typhoid rates over time and space.

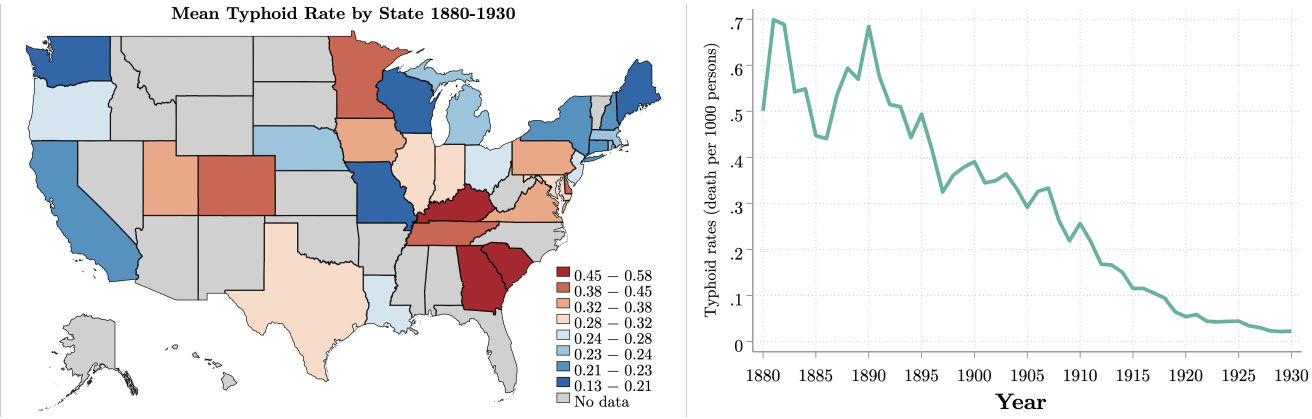

Figure 5: **Typhoid death rates over time and space**

Notes: These figures plot the evolution of typhoid death rates over time and space. The map plots the average typhoid rates per state over all years in the sample (1880-1930). The typhoid rates are defined as deaths per 1000 persons. The measure of typhoid rates is a consistent and balanced series (1880-1930) that combines information from Whipple with the mortality statistics. The data were obtained from Brian Beach (Beach, 2022b; Beach et al., 2016).

2.2.4 Sanitation Systems and Urbanization

Given the deadliness of the urban conditions, eradicating waterborne diseases was an important goal for cities at the time as they grew. Cities' investment in clean water works closely tracked the rate of urbanization (Beach, 2022b). Municipal ownership of water companies and public investment into filtration systems were viewed as vote-winners by political candidates (Troesken, 1999). Furthermore, industrialization brought in migrants from the rural area which did not have any immunity to typhoid, increased population density, and was considered a contributor to typhoid epidemics (Wohl, 1983). Thus, investments into water infrastructure and urbanization were closely linked at the time.

We show that investment in water filtration systems closely tracked urbanization. In Figure 6, we plot the number of cities with water filtration systems from Ao (2015) data against the share of population living in urban counties, which we computed from the U.S. decennial censuses. While the first water filtration systems were only being installed in the 1890s, by 1920, nearly 60 of the cities in our sample had an operational filtration plant. This closely tracks the linear increase in urbanization in the U.S. from 1900 to 1930, which tapered off between 1930 and 1940.

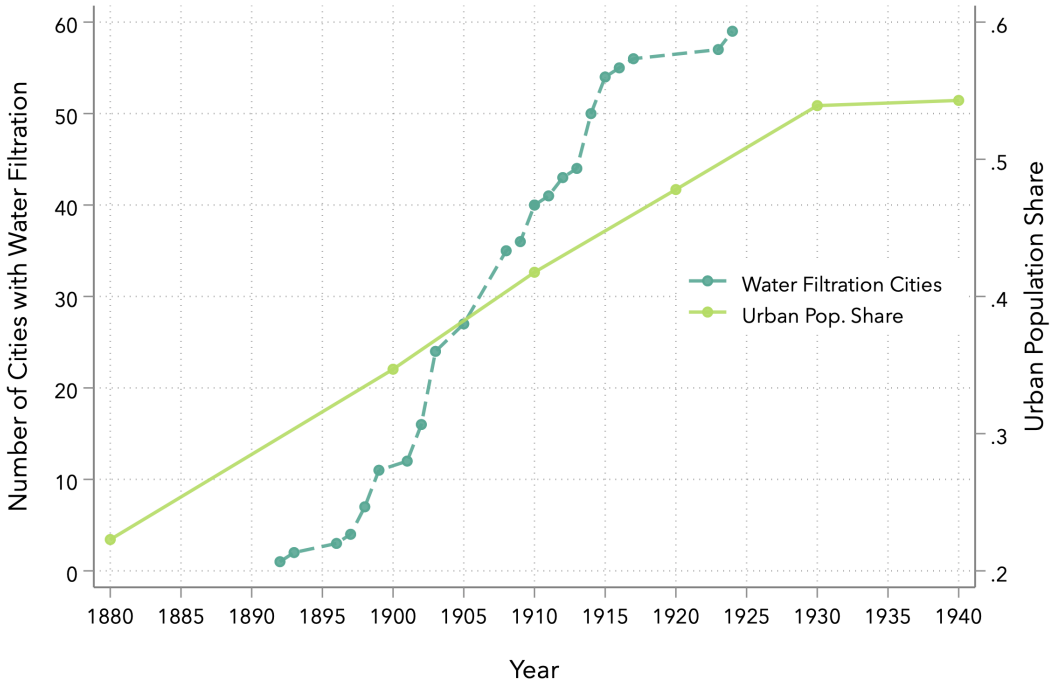

Figure 6: Urbanization & public health coincided

Notes: This figure reports the number of cities with water filtration systems from Ao (2015) data (in dark green) against the share of population living in urban counties (in light green), which we computed from the U.S. census. We find that the timing of urbanization and the implementation of water filtration systems coincided.

3 Newspaper Coverage of Urban Public Health

In this section, we document our primary data source of historical local newspapers for the U.S.. We then present descriptive evidence related to newspaper coverage of urban public health.

3.1 Newspaper Text Corpus

Our primary data source for the presence of local newspapers is the United States Newspaper Panel from Gentzkow et al. (2011). The dataset contains detailed characteristics for newspapers for each election year in the US, including readership size, subscription prices, posted advertising prices, political affiliations, and dates of entry and exit. We use this dataset to construct a measure for the availability of a local daily newspaper for each county in the US during our study period. There were approximately 300 counties

in the U.S. with at least one newspaper in 1880 and more than 1,000 of such counties by 1940. Detailed summary statistics of this dataset are available in Gentzkow et al. (2011). To match the decennial census data level of observation, we aggregate the United States Newspaper Panel data at the decennial level. For each U.S. county and census year, we compute the mean number of newspaper dailies, and the net change in number of newspapers (capturing entry/exit).[7]

Our data on newspaper coverage comes from the Chronicling America database compiled by the Library of Congress (Dell et al., 2023). To construct a measure of newspaper coverage of urban public health, we use a simple keyword search on the Chronicling America database. For each newspaper page, we detect a "hit" when a phrase related to public health appears near the name of one of our sample cities on the page.

We perform this exercise for two sets of public health phrases. First, we look for mentions of specific diseases. Among waterborne diseases, we include "typhoid," "dysentery," and "diarrhea." We also include a set of "placebo" diseases for which we do not expect to be affected by investments into clean water, which include "plague," "influenza," and "tuberculosis." For mentions of public health infrastructures, we look for mentions of a "water filtration" project.

Our measure of newspaper coverage related to public health is the total count of hits for each phrase in any given year divided by the total number of mentions for any city that year. This gives us the *share* of mentions of a city that also pertains to the public health topic of interest.[8]

3.2 Capturing Local News about Urban Public Health

The turn of the 20^{th} century witnessed radical changes in the American press. Newspapers became more independent from political parties and more informative (Petrova, 2011). Increased reach of newspapers also led to an increase in both in-depth investigative coverage as well as sensational journalism (Gentzkow et al., 2004). Our informal audit of the newspaper content suggests that news about public health can fit both categories.

First, we find that the vast majority of news coverage related to water filtration

[7]We focus on the presence of a newspaper headquarter in a county as our measure for information access. Newspapers could circulate outside of their headquarter county via carriers and mail. Gentzkow et al. (2014) find that in 1924, home-market papers are responsible for 90% of circulation.

[8]Here, we measure intensity of news coverage related to a place, as salience likely increases with intensity.For example, if there are 1,000 pages across the database that mention "Milwaukee" in 1900 and there are 5 pages that have "Milwaukee" and "typhoid" in proximity of each other, we compute this share to be 0.005.

projects revolves around their funding. Since these projects were massive municipal projects at the time, there were often political debates and referenda (e.g. on whether to issue municipal bonds) as well as sensational stories regarding government corruption in implementing these projects. Mentions were not always regarding coverage in the newspaper's own city as there were also discussion of example projects in other cities. Perhaps surprisingly, there was very little coverage celebrating the completion of a project or to advertise the improvement in public health as an amenity.

Second, we find that coverage surrounding diseases at the time broadly fit two types. The first type delivered hard information about local conditions such as case counts, death counts and obituaries. The second type of coverage was more sensational, which included features of urban epidemics and stories of local residents who have a given disease. We note that the majority of disease news was negative and focused on increases in death rates rather than decreases, consistent with Costa and Kahn (2017).

In Figure 7, we report the average share of newspaper articles mentioning either water filtration (top) or waterborne diseases (bottom) in association with a city in our dataset around the time of implementation of a water filtration system. We find that the number of mentions of diseases vastly swamped that of the implementation of a water filtration system. Coverage of water filtration peaks around the time of implementation and shortly after. On the other hand, coverage of waterborne diseases sharply decline after the implementation of filtration systems, consistent with evidence showing that death rates from these diseases also declined after water filtration (Alsan and Goldin, 2019; Anderson et al., 2022; Beach, 2022b). Disaggregating the second series across waterborne diseases (Appendix Figures B.2) shows that the highest level of coverage (and highest corresponding drop) was due to coverage of typhoid as opposed to dysentery of diarrhea. In Section 5, we show these results formally within a difference-in-differences regression alongside that for placebo diseases.

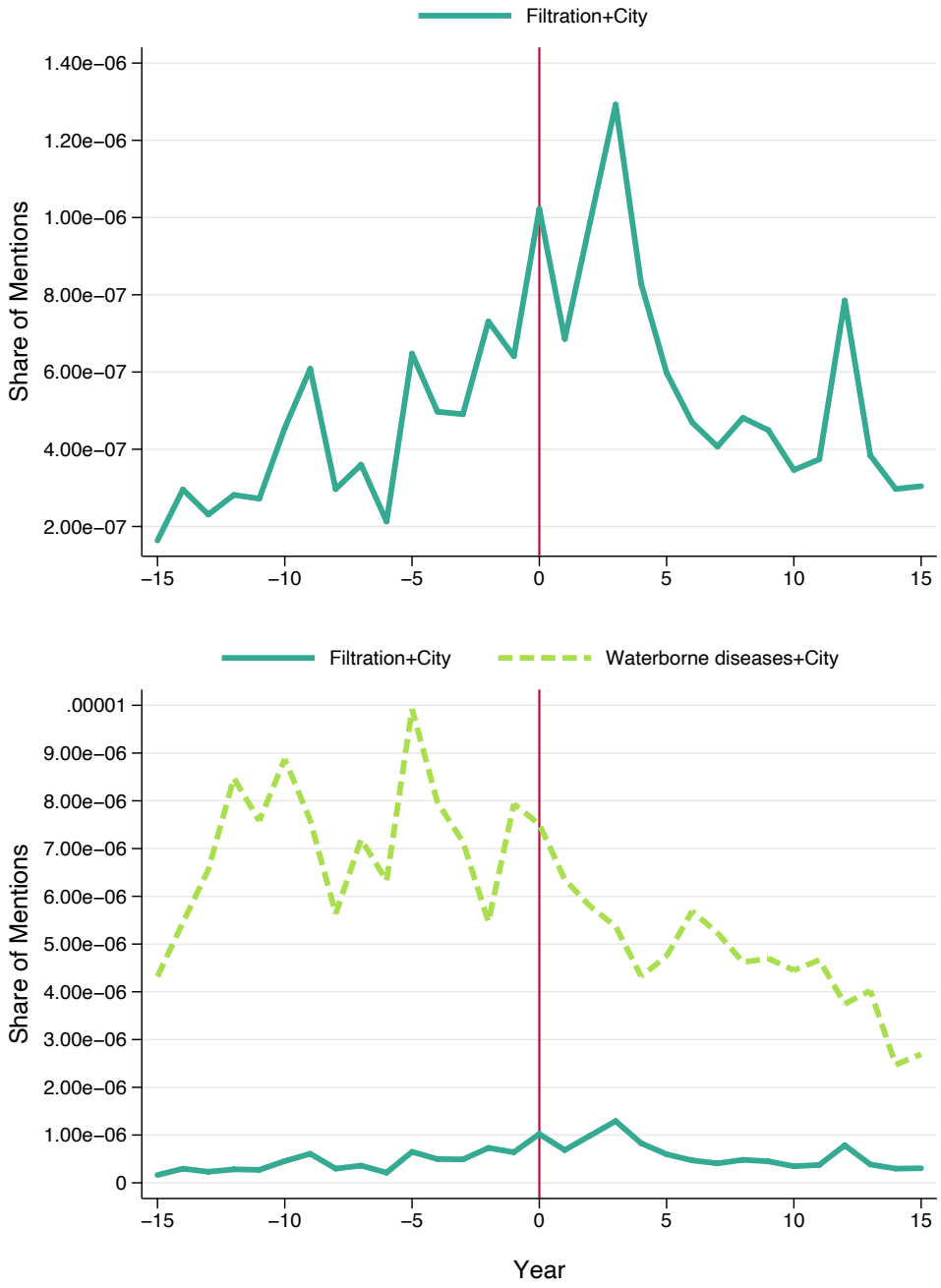

Figure 7: Capturing News coverage

Notes: These graphs report the share of newspaper articles mentioning either water filtration (top) or waterborne diseases (bottom) that are associated to a city in our dataset. The sample is currently limited to the 74 cities in our public health investment dataset, some of these cities not having an implemented water filtration system over the period. In both cases, we compute the number of keyword hits having both a city name and either filtration or typhoid phrases appearing next to each other in an article. Each graph is centered around the year of implementation of a water filtration system in the city. In the top graph, news related to these sanitation events peak around the time of the public health investment. In the bottom graph, we superpose the share of mentions to waterborne diseases (including typhoid, dysentry, and diarrhea) for these cities. The timing of water filtration investments correspond with a decrease in reports of diseases outbreaks in the newspapers. These negative health shocks are also more likely to peak in the news than positive news about health investment.

3.3 Change in coverage with public health infrastructure

In this section, we present evidence that news coverage of diseases changed at the same time of the implementation of a water filtration system. We do so by regressing the normalized number of hits for "typhoid" linked to a city name per thousand articles (averaged over a decade) on time (in decades) elapsed since the implementation of a water filtration plant. Formally, we perform the following regression

$$Hits_{dt} = \sum_{t=-3,\ldots,0,\ldots 3} \phi_t Filtration_{dt} + \kappa_d + \kappa_t + \eta_{dt},$$

where t denotes decades since the filtration system was put into place and κ denotes city and decade fixed effects.

In figure 8, we plot ϕ_t for the left-hand-side variable of hits of "typhoid" linked to the city name . We find that mentions of typhoid decrease after the city implemented a water filtration system, consistent with the disease being gradually eradicated and leading to less news coverage overall (F statistic $> 1,000$). The effect is gradual and pronounced, consistent with works showing a similar trend effect in the eradication of the disease itself. (Beach, 2022b; Alsan and Goldin, 2019).

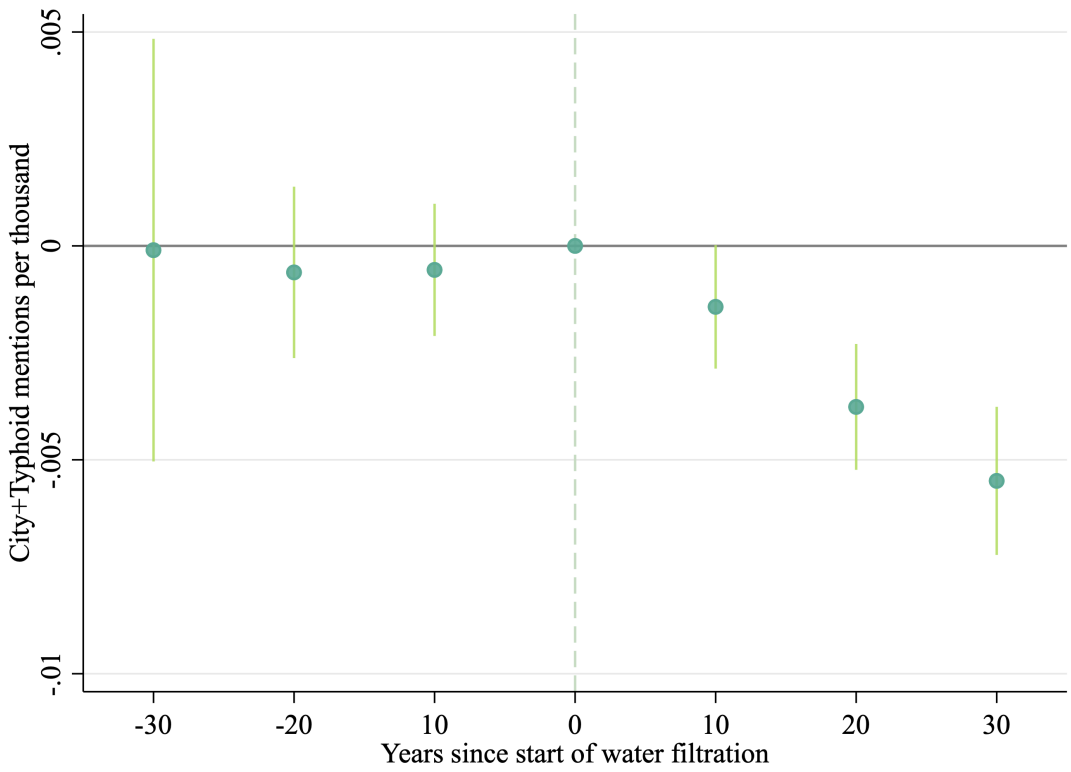

Figure 8: Effect of water filtration system on typhoid news coverage

Notes: This figure presents the changes in newspaper coverage of typhoid in response to the implementation of a water filtration system, estimated via an event study specification after controlling for place and time fixed effects. The y-axis denotes the number of hits for a destination city and "typhoid" on a newspaper page divided by the total number of hits for a destination city during that time period. The x-axis denotes years since the implementation of the city's water filtration system, grouped into decades.

Figure 8 highlights that the implementation of a water filtration system had direct effects on news coverage of disease. While we cannot rule out that migrants' potential responses could be due to other factors that accompany the implementation of a water filtration system (e.g. job opportunities, leisure amenities, or productivity), here we established that water filtration caused meaningful changes in news coverage of urban public health.

To establish that this was not a coincidence or due to secular improvement in urban life quality or the media production function, we repeat this analysis for a set of diseases likely unaffected by the implementation of a water filtration plant: plague, influenza and tuberculosis. Figure 9 reports the resulting event study graphs. The pronounced decrease in coverage of waterborne diseases reported in Figure 8 is replaced by an increase in coverage of placebo diseases. This pattern can be consistent with two different hypotheses. First, newspapers, having allocated a fixed amount of space for health and

disease coverage, could have shifted coverage to another type of disease. Second, as waterborne diseases became less of a concern, readers demand more news related to other diseases for practical purposes.

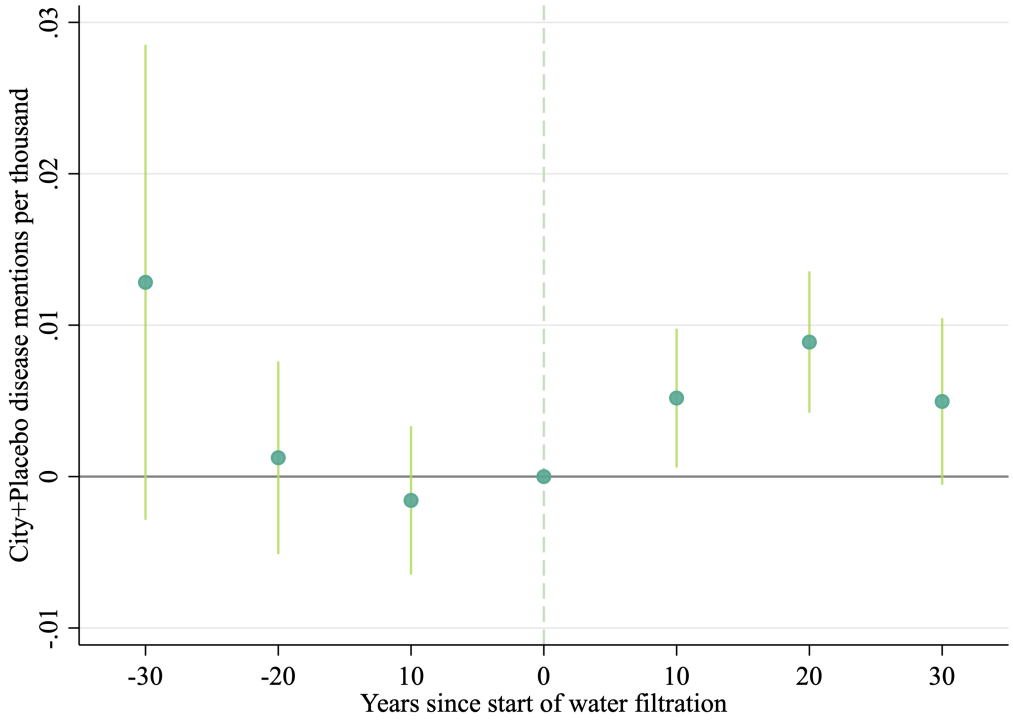

Figure 9: Effect of water filtration system on placebo disease news coverage

Notes: This figure presents the changes in newspaper coverage of tuberculosis, influenza, and plague in response to the implementation of a water filtration system, estimated via an event study specification after controlling for place and time fixed effects. The y-axis denotes the number of hits for a destination city and tuberculosis, influenza, and plague on a newspaper page divided by the total number of hits for a destination city during that time period. The x-axis denotes years since the implementation of the city's water filtration system, grouped into decades.

4 Model of Migration Choices and Information Acquisitions about Destination Amenities

This section presents our model. First, we describe a simple theoretical framework of location choice that emphasizes the role of information. Second, we use the theoretical framework to motivate our empirical exercises. The goal of the model is to make clear the identification assumptions for our empirical exercises, interpret their results and discuss their limitations.

4.1 Theoretical Framework

This section builds a discrete choice model of individual location decisions that emphasizes the role of information acquisition and micro-founds our estimating equations.

4.1.1 Set up: Players and Destination-Specific Amenities

Let $t \in \{1, \ldots, T\}$ index periods in time and $\{1, \ldots, D\}$ be the set of geographical divisions. For each period t, a individual i's location at the beginning of the period is referred to as its *origin* $o \in \{1, \ldots, D\}$ and its location at the end of the period is referred to as its *destination* $d \in \{1, \ldots, D\}$. We say that a individual i has migrated during the 10-year intercensal period t if $d \neq o$.

Let H_{dt} denote a vector of characteristics of destination location d at the end of the period t that is relevant to the individual and observable by the researcher. For expositional purposes, this subsection assumes that H_{dt} is a scalar (e.g. the typhoid mortality rate at a destination). The individual may not know H_{dt} and instead have a perception $\tilde{H}_{dt}$ of what H_{dt} may be.

4.1.2 Individual's Perception of Location-specific Amenities

We make several assumptions on the individual's *perception* of the characteristics of different locations.

First, individuals have perfect information on the characteristics of the place they currently live in, so that $\tilde{H}_{ot} = H_{ot}$. Second, there are two information regimes possible for each origin location, captured by the indicator C_{ot} for *access to local information*. $C_{ot} = 1$ indicates that a resident at the origin would have had access to information about potential destinations. Finally, the perceived location characteristic for a destination $d \neq o$ depends on access to information in the origin location and can be written as

$$\tilde{H}_{dt} = \begin{cases} \overline{H}_{dt} & \text{if } C_{ot} = 0 \\ \rho H_{dt} + (1 - \rho)\, \overline{H}_{dt} & \text{if } C_{ot} = 1 \end{cases}$$

where $\overline{H}_{dt}$ is the prior belief about the characteristic of destination d common to individuals across all origins o. $\overline{H}_{dt}$ thus captures a population-weighted average belief about d over all origin locations o. The parameter ρ captures the extent of updating to the real value H_{dt} for individuals living in an origin location with more information (i.e. for individuals living in location such that $C_{ot} = 1$).[9]

[9] An implicit assumption is that the prior for a given destination is common across origins.

Note that this framework nests the strong assumption of perfect updating (e.g. $\rho = 1$ implies that $\tilde{H}_{dt} = H_{dt}$) or completely discarding the information ($\rho = 0$ implies that $\tilde{H}_{dt} = \overline{H}_{dt}$).

4.1.3 Individual's Static Utility from Moving

Suppose a potential migrant i from origin location o makes a one-shot, static decision about whether to move to location d during the 10-years period t.[10] We specify the static, myopic location decision of a individual i over a ten-year periods using the following utility specification for each possible destination d:[11]

$$d_{iot} = \underset{d' \in \{1,\ldots,D\}}{\arg\max}\ u_{iodt}$$

$$u_{iodt} = \begin{cases} \tilde{\beta}\left(\tilde{H}_{dt} - \tilde{H}_{ot}\right) + X'_{odt}\alpha + \xi_{dt} - \xi_{ot} + \xi_{odt} + \eta_{iodt} & d \neq o \\ 0 & d = o \end{cases}$$

We are assuming that the individual's outside option is to stay put (i.e. to choose $d = o$) and we normalize the utility from this outside option to zero. As a result of this normalization, preferences are interpreted as relative to the outside option of not moving. $\tilde{\beta}$ denotes the moving individual's preference over the perceived difference in destination and origin characteristic, $\tilde{H}_{dt} - \tilde{H}_{ot}$. X_{odt} is a vector of migration costs between o and d for intercensal period t. ξ_{dt} is a shifter of utility for destination location d at time t, ξ_{ot} captures the disutility from moving away from location o in period t, and ξ_{odt} is an o-d-pair level shifter of utility. η_{iodt} is an idiosyncratic shock to individual i. The individual observes ξ_{dt}, ξ_{ot}, ξ_{odt} and η_{iodt}, while the researcher does not.

We can rewrite the utility specification as following, accounting for the presence of information C_{ot},

[10] Under this setting, u_{iodt} is not a continuation value. In a dynamic model of migration, locations may have option values that capture the potential future benefits of moving to a location, and influence migration decisions. Our model does not incorporate option values and a dynamic model may yield different insights. We believe these concerns are attenuated by the long 10 years time-frame of the migration decisions in our historical context. The static, one-shot decision of the individuals thus spread a 10-years period. Future research could investigate an extension of this model to a dynamic setting with option values in the spirit of Caliendo et al. (2019).

[11] While a myopic model may not capture all aspects of migration decisions, it can still provide valuable insights into the role of information and other factors in shaping migration patterns over these decennial periods.

$$u_{iodt} = \begin{cases} \underbrace{\tilde{\beta}\left[\rho\frac{\left(H_{dt}-\overline{H}_{dt}\right)}{H_{dt}}\right]C_{ot}H_{dt} + X'_{odt}\alpha + \left(\tilde{\beta}\overline{H}_{dt}+\xi_{dt}\right) - \left(\tilde{\beta}H_{ot}+\xi_{ot}\right) + \xi_{odt}}_{\delta_{odt}} + \eta_{iodt} & d \neq o \\ 0 & d = o \end{cases} \tag{1}$$

where δ_{odt} denotes the mean utility of destination d for individuals from origin o in period t.

Assuming that the error term η_{iodt} is identically and independently distributed across destinations and individuals and it follows an extreme value distribution, we can write the predicted share of individuals from o that migrate to d in period t as a function of the mean utility δ_{odt},

$$s_{odt}\left(\vec{\delta}\right) = \frac{\exp\left(\delta_{odt}\right)}{1+\sum_{d'\neq o}\exp\left(\delta_{od't}\right)} \quad \text{and} \quad s_{oot}\left(\vec{\delta}\right) = \frac{1}{1+\sum_{d'\neq o}\exp\left(\delta_{od't}\right)} \tag{2}$$

and obtain the familiar "market share regression" as in Berry (1994):

$$\log\left(s_{odt}\right) - \log\left(s_{oot}\right) = \delta_{odt}.$$

4.2 Empirical model

Our main estimating equation is directly micro-founded by the discrete choice logit model (1) and market share equation (2),

$$\begin{aligned} \log\left(s_{odtg}\right) - \log\left(s_{ootg}\right) = \beta C_{ot} \times H_{dt} + \underbrace{\alpha_1 \log\left(\text{dist}_{\text{od}}\right) + \alpha_2 \ln\left(\tau_{\text{odt}}\right)}_{\text{migration costs}} + \alpha_w w_{gdt} \\ + \underbrace{X'_{odt}\alpha + \gamma_{dt} - \gamma_{otg}}_{\text{location-specific cov.}} + \gamma_{\text{occup}} + \varepsilon_{odgt} \end{aligned} \tag{3}$$

In this equation, o indexes origins, d indexes destinations, t index time periods, and g index demographic groups (age, race and literacy). The left-hand side captures the log probability of moving from o to d by the end of the intercensal period t, for individuals belonging to group g. C_{ot} captures access to information from local newspapers and is a proxy for information available at the origin. In practice, C_{ot} captures the fraction of period t during which origin location o has a local daily newspaper. H_{dt} is a destination characteristic of interest. In our setting, it will be related to health investments and either

measure (i) the fraction of time during period t that destination d has implemented a water filtration system, or (ii) mortality due to typhoid outbreaks. The interaction term $C_{ot} \times H_{dt}$ represents preference over public health amenities, while the coefficients α_1 and α_2 capture moving costs.

The term X_{odt} is a vector of origin-destination-time specific covariates and γ_{occup} is a work occupation fixed effect. ε_{odtg} is a regression error term. The two γ terms represent preferences over various location characteristics: γ_{dt} and γ_{otg} are destination-time and origin-time-demographic group fixed effects, respectively. The origin-year-demographic group fixed effect ensures that we are comparing two destinations d and d' for a given origin o, while adding a destination-year fixed effect ensures that what is left in the unobserved term is changes in moving costs from o to d over time. The γ_{dt} fixed-effect thus control for any unobserved factors that might vary across destinations and over time, but remain constant within each destination-year combination.

By incorporating both sets of origin-demographic group-year and destination-year fixed effects, we are effectively controlling for unobserved factors that could be correlated with the interaction term $C_{ot} \times H_{dt}$, both at the origin location-demographic group level and the destination level, as well as any time-specific shocks that affect all groups or destinations equally. Under this setting, identification comes solely from bilateral variation in the interaction term $C_{ot} \times H_{dt}$. The residual variation in migration after origin-year-group and destination-year fixed effects is an origin-destination-time specific shock. We discuss identification in more details in section 4.3 below.

Location-time specific control variables include share white, male, foreign-born, literate and share working in the manufacturing sector. Each control variable is measured as the difference between origin and destination locations.

The parameter of interest is β, which is a product of three distinct objects,

$$\beta = \tilde{\beta}\rho\frac{\left(H_{dt} - \overline{H}_{dt}\right)}{H_{dt}}$$

In our model, $\tilde{\beta}$ represents the direct value of a characteristic on a individual's utility, ρ is the individual's extent of updating of their beliefs given the information available in C_{ot}, and $\frac{\left(H_{dt}-\overline{H}_{dt}\right)}{H_{dt}}$ is the extent of correction of the individual's belief given their prior if they fully discard their prior. We cannot separately identify any of these objects without data on the other two. We will instead present a range of possible values for each of these objects given our estimate of β and calibrated values for the remaining parameters.

Our use of fixed effects γ_{dt} and γ_{otg} directly corresponds to objects in a individual's utility that are constant at the destination-time level $\left(\tilde{\beta}\overline{H}_{dt} + \xi_{dt}\right)$ and those at the origin-

time level $\left(\tilde{\beta} H_{ot} + \xi_{ot}\right)$. As in Berry (1994), our regression error term corresponds to the origin-destination-time-occupation level unobservable characteristic ξ_{odtg}.

4.3 Identification assumptions

Two-way Fixed Effects. In this setting, the main coefficient of interest is the one on the interaction term $C_{ot} \times H_{dt}$, which represents the effect of information (news coverage at origin) interacted with public health amenities (water filtration or typhoid mortality at destination) on migration flows. The goal is to identify this effect while controlling for potential confounding factors. The two-way fixed effects play a crucial role in this identification strategy:

1. The origin-year fixed effects (γ_{ot}) control for any time-varying factors specific to the origin location that might affect out-migration, such as local economic conditions, population changes, or other shocks. They also absorb the main effect of C_{ot}, the news coverage variable.

2. The destination-year fixed effects (γ_{dt}) control for any time-varying factors specific to the destination location that might affect in-migration, such as local economic conditions, amenities, or other shocks. They also absorb the main effect of H_{dt}, the water filtration or typhoid mortality variable.

By including these fixed effects, the regression effectively controls for all factors that vary across origins and years (but are constant across destinations for a given origin-year) and all factors that vary across destinations and years (but are constant across origins for a given destination-year). This leaves only the bilateral variation in the interaction term $C_{ot} \times H_{dt}$ to identify the effect of interest.

Our primary identification assumption is thus that the co-variation in origin information environment (e.g., the presence of a newspaper) and the destination characteristic of interest (e.g., typhoid rate) is exogenous to the flow of migrants between those two places after controlling for unobservable characteristics specific to each place. The key assumption for this identification strategy to be valid is that there are no other bilateral factors (varying at the origin-destination-year level) that are correlated with both the interaction term and the migration flows, after controlling for the two-way fixed effects. If this assumption holds, then the coefficient on the interaction term can be interpreted as the causal effect of information interacted with public health amenities on migration decisions.

Identification Assumptions and Potential Violations. Any variation in the flow of migrants from o to d that correlates with origin information and destination characteristic other than through preference for that characteristic and belief updating violates this assumption. Formally, we assume that

$$\mathbb{E}\left[\xi_{odgt}|C_{ot}H_{dt}, X_{odt}, \gamma_{dt}, \gamma_{otg}, \gamma_g\right] = 0.$$

Importantly, we do not assume that the origin information environment is exogenous to migration or that the destination characteristic is exogenous to migration. To make this discussion concrete, let us consider C_{ot} as the presence of a local newspaper at the origin and H_{dt} as the presence of a sanitation structure at the destination.

We do not assume that a local newspaper's presence is uncorrelated with migration. In fact, the entry of local newspapers is highly correlated with the predicted profit in that market and thus population growth. We only require that it does not differentially influence migration to places with sanitation structures as opposed to places without, other than through preference for sanitation and updating beliefs about the presence of sanitation structures. To the extent that the presence of local newspapers correlates with unobservable characteristics specific to that origin location in that time period, we can control for those unobservables with the use of origin-time fixed effects.

Similarly, we do not assume that a sanitation structure's presence at the destination is exogenous to in-migration. Sanitation structures are more likely to be implemented in cities that are able to fund large public works projects and in the aftermath of deadly outbreaks, and we are able to control for these factors with the use of destination-time fixed effects. We only require that it does not differentially influence in-migration from places with a local newspaper as opposed to places without, other than through preference for sanitation and updating beliefs about the presence of sanitation structures.

Discussion of Potential Caveats. There are potential caveats to our identification strategy that warrant discussion. First, there is a possibility of omitted variable bias if there are time-varying origin-destination specific factors that are correlated with both the interaction term and the migration flows, even after controlling for the two-way fixed effects. For example, if there are bilateral trade or cultural ties between specific origin-destination pairs that evolve over time and are correlated with both information flows and migration, this could bias our estimates. While we believe such factors are likely to be limited in our historical setting, we cannot entirely rule out this possibility.

Second, there may be concerns about reverse causality, such as the possibility that

migration flows might influence the entry of newspapers or the adoption of sanitation structures. While the two-way fixed effects help mitigate this concern by controlling for time-varying factors at the origin and destination level, it is worth acknowledging this possibility. However, given the historical context of the United States in the years 1870-1940, we argue that newspapers were a prime way of obtaining information about potential destinations, and their entry was likely driven by factors such as local economic conditions and population growth rather than migration flows per se.

Third, there is a possibility that sanitation investments at the destination could be made in response to expected future migration flows from population origins, which in turn could be more likely to have newspapers. If this were the case, our estimates could be biased, as the interaction term $C_{ot} \times H_{dt}$ would be correlated with unobserved factors driving both sanitation investments and migration flows. While this concern is partially addressed by the inclusion of destination-time fixed effects, which control for any time-varying factors that affect all origins equally for a given destination-year, we cannot completely rule out the possibility of origin-specific expectations shaping destination investments. On the other hand, political complications related to investments into public health discussed in Section 2.2.2 make it highly unlikely that local government officials would have expended political capital toward these investments on behalf of future residents.

Fourth, it is important to consider whether information about disease could be carried through migrants moving back and forth between an origin and destination. If this were the case, our interaction term $C_{ot} \times H_{dt}$ might not only capture the effect of newspaper-based information but also the effect of information transmitted through personal networks and repeat migration. While we believe newspapers were the primary source of information about distant locations during our study period, we acknowledge that interpersonal information flows could have played a role as well. To the extent that these interpersonal flows are correlated with newspaper presence at the origin and sanitation investments at the destination, our estimates could be biased. Disentangling these different information channels poses an interesting challenge for future research, which could utilize data on repeat migration or social networks to shed light on the relative importance of various information sources.

Fifth, as we do not observe international migrants, there are potential concerns is that we are either understating the importance of international migrants toward public health or not allowing for international migrants to have different responses to newspapers than native-borns. This concern is potentially important: we document in Appendix A.2 that foreign-born populations are mostly located in urban areas (63.2% of foreign-borns

are in urban areas against 36.8% in rural areas). Over the 1870-1940 period, foreign-born individuals only represent 8.3% of the population living in rural areas, against 19.7% of the population living in urban areas. Given that our empirical implementation controls for destination-time fixed effects, we are particularly concerned if international migrants can induce rural-urban migration from one origin over another. Since most rural population is native-born (75.5%), we believe that this particular form of bias is unlikely.

Finally, it is important to consider the external validity of our findings, given the specific historical and geographic context of our study. While we believe our results provide valuable insights into the role of information in shaping migration decisions during a period of rapid urbanization and public health improvements in the United States, the generalizability of our findings to other settings or time periods may be limited.

5 Empirical estimates

5.1 Newspaper Presence and Destination Amenities

Bilateral Flows Sample. Tables 3 and 4 report the estimates from our primary estimating equation (4), where the destination characteristic of interest H_{dt} stand for the presence of water filtration plants and typhoid outbreaks, respectively. In these tables, we restrict origin locations to rural areas and destination locations to cities. We do not condition on the presence of PHI or a typhoid outbreak at destination. Note that restricting the sample to origin-destination pairs within 50 miles do not change the results (these pairs in fact represent 99% of our sample, see discussion from Section 2.1.2 and Figure). Appendix D lists robustness test results restricting to different bilateral flows samples.

Regression Framework. In Table 4 , H_{dt} is measured as average typhoid rates in destination location d over the decennial period t, in deaths per thousand persons. In Table 3, H_{dt} is measured as a percentage; it captures the fraction of the 10-years intercensal period t during which a water filtration plant was present at d. For instance, if a water filtration plant was implemented in location d in 1902, H_{dt} takes a value of zero for all census years up to 1900. For intercensal period 1900-1010, H_{dt} takes a value of 0.8. From 1910 onwards, it takes a value of 1. Similarly, in both tables, C_{ot} captures the fraction of the decennial period t during which origin location o had a local daily newspaper.

The dependent variable $\log\left(s_{odtg}\right) - \log\left(s_{ootg}\right)$ captures bilateral migration flows

from the logit model, in logs, subtracting the log share of people choosing the outside option of staying put (i.e. choosing $d = o$) to the share of people migrating to a destination $d \neq o$. s_{odtg} captures the probability that an individual from demographic group g moves from location o to d in intercensal period t. Within the model, this dependent variable is interpreted as the utility of destination d for a migrant from o in time t of group g relative to his outside option. Since we control for origin-time fixed efects (and so the level of $\log\left(s_{ootg}\right)$) we can interpret the remaining coefficients as shifters of the log probability of moving from o to d, $\log\left(s_{odtg}\right)$.

The final sample contains 617,483,404 unique matched individuals over the 1870-1940 period. The share is measured for each demographic group g (age, race and literacy) and census time period t. Age, race and literacy are categorical variables. "Race" is split between white and non white and"age" between '16-40' and '41-65' years old.[12] "Literacy" is binary and is split between "literate" and "illiterate". We consider the ability to read as enough to be literate. We do not condition literacy on the ability to write, as reading is the relevant margin of literacy for our newspaper information channel.

All regressions control for occupation fixed effect. We define five occupation groups that aggregate the original IPUMS occupation categories: "Professionals & Managers", "High-skilled Service", "Low-skilled Service", "Farmers and laborers" and "N/A or unemployed".[13] All regressions also include log distance (in miles) between individuals' origin and destination, the interaction of log distance and public health investment (presence of a water filtration plant), log transportation costs between origin o an destination d from Donaldson and Hornbeck (2016) and the mean wage of demographic group g in destination d.[14] Appendix section C.1 discuss how we construct the wage variable. County-level control variables include share male, white, literate, foreign-born and share working in the manufacturing sector, from ICPSR. All controls are bilateral measures and capture the difference between an origin and destination pair, as they are

[12] We drop people below 16 and above 65. The mean age in our data is 27, the median age is 24, and the 75^{th} is 4

[13] See https://usa.ipums.org/usa/resources/volii/Occupations1950.pdf for a description of the IPUMS' classification of occupations. The original data includes 13 categories: "Professional, technical and kindred workers", "Farmers and farm managers", "Managers, officials, and proprietors, except farm", "Clerical and kindred workers", "Sales workers", "Craftsmen, foremen, and kindred workers", "Operatives and kindred workers", "Private household workers", "Farm labors and foremen","Laborers, except farm and mine", "Service workers, except private household", "Housekeeping/housewife","N/A, not classified or unemployed".

[14] We use Donaldson and Hornbeck (2016)'s predicted county-to-county transportation cost matrices, in logs. Donaldson and Hornbeck (2016) calculate these costs using a newly constructed geographic information system (GIS) network database, which relies on available railroads,canals, natural waterways, and wagons.

otherwise not separately identified from the fixed effects.

Table 3: **Regressions of Migration Flows to Cities on Newspaper Presence at Origin and Water Filtration Plants at Destination**

	Dependent Var.: Destination Utility			
Fixed Effects:	**Origin-Year-Group**		**Origin-Year-Group & Destination-Year**	
	(1)	(2)	(3)	(4)
Newspaper Presence × Water Filtration	0.2551***	0.2531***	0.1431***	0.1286**
	[0.08973]	[0.0809]	[0.0489]	[0.05236]
Water Filtration	0.9123***	0.7409***		
	[0.07325]	[0.06707]		
Log Distance	-0.6609***	-0.8293***	-0.975***	-0.9838***
	[0.05364]	[0.04094]	[0.01491]	[0.01704]
Mean wage of group g in d	1.7e-07***	1.4e-07***	2.5e-08***	1.5e-08**
	[2.9e-08]	[2.5e-08]	[5.7e-09]	[6.6e-09]
Log Distance × Water Filtration	-0.3192***	-0.3281***	-0.3292***	-0.33***
	[0.0335]	[0.02589]	[0.02034]	[0.02125]
Log Transportation Cost	0.5374***	0.7662***	0.9861***	0.9939***
	[0.1355]	[0.1245]	[0.04301]	[0.05056]
County Covariates, in Diff. $(d - o)$		✓		✓
Occupation FE	✓	✓	✓	✓
Constant	-4.956***	-5.492***	-5.116***	-5.085***
	[.1592]	[.1485]	[.05147]	[.07239]
R2	0.6686	0.7014	0.8069	0.8046
F-Stat	1222	912.9	3110	1567
Dep. Var.: Mean	-4.7	-4.724	-4.7	-4.724
Dep. Var.: Std. Dev.	1.54	1.543	1.54	1.543
N Obs.	9.5e+05	7.6e+05	9.5e+05	7.6e+05

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: All coefficients are standardized beta coefficients. Standard errors are clustered at the origin-group-year and destination-year level. "Water Filtration" captures the implementation of water filtration plants at the destination city. All regressions include the interaction of log distance and public health investment (water filtration plant), log distance, log transportation costs from Donaldson and Hornbeck (2016) and mean wage of deographic group g in destination d. County-level control variables include share male, white, literate, foreign-born and share in the manufacturing sector. All controls are bilateral measures and capture the difference between an origin and destination pair, as they are otherwise not separately identified from the fixed effects.

Regression results: Water Filtration Plants at Destination. The first two columns

of Table 3 report results from a version of model (4) that only includes origin-year-group fixed effect without destination-year fixed effects. Column (1) does not include county-level covariates, while column (2) includes these covariates, measured in difference between a destination and origin location. All coefficients are standardized beta coefficients. On average, the presence of a water filtration plant increases the share of residents from o moving to d by 74% compared to a location d' without water filtration systems, while the joint presence of a newspaper at the origin location o and the presence of water plants at d increases this probability by an additional 25% relative to origins without newspaper coverage (Column 2). Taken together, for origin-destination pairs where the origin o has news coverage and the destination has a water filtration systems, the share of residents moving across that pair is $(25 + 74) = 99\%$ larger than for pairs without either.

Columns (3) and (4) implement the full model (4) including a destination-year fixed effect. This controls for any destination-year specific heterogeneity, so we no longer need to control for the sole presence of water filtration plants at the destination. Under this specification, the fixed effects capture unobserved time-invariant factors specific to each destination and thus control for any unobserved factors that might vary across destinations and over time but remain constant within each destination-year combination. The occupation fixed effects further control for variations across job-specific occupations. We find that on average, the share of residents moving from origin-destination pairs where o has news coverage and d has a water filtration plant is 13% higher than other origin-destination pairs. The results suggest that, while migrants generally respond to better public health conditions in urban destinations, those with access to information about these conditions exhibit significantly stronger responses. The differential migration flows based on news coverage at the origin location provide evidence for the importance of information in shaping migrants' decisions and their ability to accurately assess the benefits of public health investments at potential destinations.

Regression results: Typhoid Outbreaks at Destination. Similarly, in columns (1) and (2) of Table 4, we establish that origin-destination pairs where the origin has a local newspaper and the destination has higher typhoid rates experience a lower share of migrants than origin-destination pairs not sharing these characteristics. The coefficient for interaction term News Coverage $\times$ Typhoid rate is negative and statistically significant in columns (1), (3), and (4), suggesting that higher news coverage at the origin interacted with higher typhoid mortality rates at the destination is associated with reduced migration flows.

Table 4: **Regressions of Migration Flows to Cities on Newspaper Presence at Origin and Typhoid Outbreaks at Destination**

	Dependent Var.: Destination Utility			
Fixed Effects:	**Origin-Year-Group**		**Origin-Year-Group & Destination-Year**	
	(1)	(2)	(3)	(4)
Newspaper Presence × Typhoid rate	-0.5806***	-0.3071	-0.3199***	-0.2805**
	[0.192]	[0.1897]	[0.1117]	[0.1375]
Typhoid rate	-0.8369***	-0.1383		
	[0.1695]	[0.2316]		
Log Distance	-0.9487***	-1.02***	-1.153***	-1.136***
	[0.06677]	[0.06379]	[0.02861]	[0.03277]
Mean wage of group g in d	9.6e-08***	3.7e-08**	3.4e-08***	2.8e-08***
	[2.2e-08]	[1.7e-08]	[8.1e-09]	[9.5e-09]
Log Distance × Water Filtration	-0.09105	-0.1126***	-0.2426***	-0.2456***
	[0.05621]	[0.04188]	[0.02454]	[0.02678]
Log Transportation Cost	0.744***	0.769***	1.043***	0.9935***
	[0.1655]	[0.1745]	[0.06335]	[0.07208]
County Covariates, in Diff. ($d - o$)		✓		✓
Occupation FE	✓	✓	✓	✓
Constant	-4.052***	-4.648***	-4.335***	-4.006***
	[0.1912]	[0.2217]	[0.06546]	[0.1187]
R2	0.6708	0.7071	0.8037	0.8032
F-Stat	338.2	318.7	879.6	512.3
Dep. Var.: Mean	-4.421	-4.412	-4.421	-4.412
Dep. Var.: Std. Dev.	1.462	1.472	1.462	1.472
N Obs.	3.8e+05	3.0e+05	3.8e+05	3.0e+05

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: All coefficients are standardized beta coefficients. Standard errors are clustered at the origin-group-year and destination-year level. Typhoid rates are measured as deaths by thousand. All regressions include the interaction of log distance and public health investment (water filtration plant), log distance, log transportation costs from Donaldson and Hornbeck (2016) and mean wage of deographic group g in destination d. County-level control variables include share male, white, literate, foreign-born and share in the manufacturing sector. All controls are bilateral measures and capture the difference between an origin and destination pair, as they are otherwise not separately identified from the fixed effects.

Given an origin o with news coverage, the share of residents moving from location o to destination d with a one standard deviation higher typhoid mortality rate is 30% lower than for a location d' with a one standard deviation lower typhoid mortality

rate (Column 2). However, this effect is not statistically significant when controlling for county-level covariates and origin-year-group fixed effects only. Columns (3) and (4) include destination-year fixed effects, controlling for any destination-year specific heterogeneity. Under this specification, we find that the share of residents moving from origin-destination pairs where o has news coverage and d has a one standard deviation higher typhoid mortality rate is 28% lower than other origin-destination pairs, significant at the 10% level (Column 4).

Positive versus Negative Destination Characteristics. Comparing the results from Table 3 and Table 4 suggests that newspaper presence at the origin interacts with both positive and negative destination characteristics to influence migration decisions. The magnitude of the interaction effect is larger for typhoid mortality than for water filtration. In the most comprehensive specification (Column 4), a one standard deviation increase in water filtration is associated with a 13% increase in migration flows for origin-destination pairs with news coverage, while a one standard deviation increase in typhoid mortality is associated with a 28% decrease in migration flows.

Finally, the main effect of the destination characteristic (second row of the tables) is significant for water filtration but not for typhoid mortality when controlling for origin-year-group fixed effects only (Column 2). This suggests that the presence of water filtration has a positive impact on migration flows even without considering information, while the effect of typhoid mortality is more dependent on the interaction with newspaper presence.

Section E presents additional in-migration regression results based on placebo tests that use past migration flows.

5.2 News Coverage of Destination Amenities

Our results in the previous section demonstrate that rural migrants differentially respond to changes in public health conditions of urban destinations depending on whether their rural origin has a local newspaper. It could still be that migrants were responding to other amenities that improved at the same time as public health conditions (e.g. employment opportunities, productivity, or leisure). In this section, we augment the previous results and show evidence that news coverage of public health changed with public health conditions and was likely important in driving migration responses.

Regression framework: Response to Typhoid News. We keep our primary estimating equation (4), but change the implementation of H_{dt} to be the degree of coverage of waterborne diseases at an urban destination rather than actual case count. We interpret the results of this regression as evidence that migrants potentially respond to news coverage of waterborne diseases rather than the prevalence of the disease itself.

Using our measure of news coverage constructed in Section 2.3 as a regressor within our primary estimating equation (4) presents two primary challenges.

The first challenge is a data limitation: we only observe disease coverage for large urban destinations but not for all destination places (such as potential rural destinations or places not named in the census). We take two approaches to this problem. First, we impute the missing amount of coverage to be 0 for rural areas, while unnamed urban areas get the average coverage measure for cities without typhoid outbreaks/PHI investments during that period and within the same state. We believe that this is justified: the rate of typhoid in rural areas was very low (and likely did not receive any coverage). For robustness, we present qualitatively similar results in the appendix where we limit our sample to only urban areas where we have data on news coverage of diseases.

The second challenge is related to pair-level reverse causality. News about diseases at a given destination d should correlate with mentions of this destination in a newspaper at the origin o, since migrations from o to d would increase demand for news of d at o. Furthermore, previous migrants from an origin o to a destination d could carry information back to o.

Due to this problem, we predict changes in coverage of typhoid at a destination by using the timing implementation of water filtration systems in each city. Formally, we construct $\hat{H}_{dt}$ as the fitted values from the following regression:

$$H_{dt} = \phi \text{Filtration}_{dt} + \gamma_d + \eta_{dt}$$

where H_{dt} denotes the amount of news coverage of typhoid at a migrant's potential destination and γ_d denotes destination fixed effects. We then estimate the following regression:

$$\log\left(s_{odtg}\right) - \log\left(s_{ootg}\right) = \beta C_{ot} \times \hat{H}_{dt} + \underbrace{\alpha_1 \log\left(\text{dist}_{\text{od}}\right) + \alpha_2 \ln\left(\tau_{\text{odt}}\right)}_{\text{migration costs}} + \alpha_w w_{gdt} + \underbrace{X'_{odt}\alpha + \gamma_{dt} - \gamma_{otg}}_{\text{location-specific cov.}} + \gamma_{\text{occup}} + \varepsilon_{odgt}. \tag{4}$$

where the remaining components are the same as that in our primary estimating equation (4).

Regression results: Response to Typhoid News. Table 5 describes our results. When we do not use the predicted values for typhoid news and instead use the direct value H_{dt} (Columns 1 and 2), we find that origins with a newspaper have more migrants toward destinations that have a higher incidence of typhoid news. This likely represents the previously discussed bias: news coverage of a destination likely increases as in-migration from origins with newspapers increases.

When we instead predict the amount of coverage of typhoid using the implementation of a water filtration system, we find that migrants from origins with newspapers respond especially negatively to the increase in coverage of typhoid news. In our preferred specification (Column 4), an increase in one standard deviation of typhoid news coverage of a destination leads to a 61% lower migration share from an origin with a newspaper.

Table 5: **Predicted Typhoid News**

	Actual Coverage		**Predicted Coverage**	
	(1)	(2)	(3)	(4)
Newspaper Presence × Typhoid Coverage (std.)	0.085***	0.11***	-0.50**	-0.61**
	[0.014]	[0.021]	[0.20]	[0.27]
Newspaper Presence × Urban/rural dest.	0.050***	0.039*	0.84***	0.83***
	[0.018]	[0.022]	[0.25]	[0.29]
Log Distance	-0.98***	-0.99***	-0.98***	-0.99***
	[0.0073]	[0.0087]	[0.0078]	[0.0095]
County Covariates, in Diff. $(d - o)$		✓		✓
Origin-Year FE	✓	✓	✓	✓
Destination-Year FE	✓	✓	✓	✓
R2	.756	.758	.5386	.5416
Dep. Var.: Mean	-7.00	-7.02	-7.00	-7.02
N Obs.	2.7e+06	2.0e+06	2.7e+06	2.0e+06

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table describes results for our regression of migration flow on the typhoid news coverage. Columns 1 and 2 describe the regression using the actual amount of coverage measured. Columns 3 and 4 describe the two-stage least squares specification where we predict news coverage using the implementation of a water filtration system at the urban destination, destination fixed effects, and time fixed effects. Each column includes origin-year and destination-year fixed effects, and columns 2 and 4 include controls for demographic congruence between the origin and the destination. We report coefficients for three variables. The first is the interaction term between whether the rural origin has a newspaper with the standardized amount of typhoid coverage of a destination. The second is the interaction term between whether the rural origin has a newspaper and whether the destination is classified as a rural or urban place in the census. The third is the natural logarithm of the distance between the origin and the destination. All standard errors are clustered at the origin-year and destination-year level.

We interpret this result as evidence that migrants not only respond to changes in typhoid levels but that news coverage is a plausible channel. This is intuitive given our result in Figure 8 showing that news coverage of typhoid decreases sharply given implementation of a filtration system. The fact that migration increase when news coverage of typhoid decreases and news coverage of airborne diseases such as influenza and tuberculosis increases (Figure 9) suggests that waterborne diseases were far more detrimental to perception of urban conditions than other. It is possible that waterborne diseases are more salient as a factor of urban conditions due to their concentrated and predictably seasonal outbreaks, and that tuberculosis was already treatable in 1900 with drugs that

were widely used (Smith, 1988).

6 Conclusion

This paper demonstrates the significant role that daily newspapers in rural areas played in the process of urbanization in the United States between 1870 and 1940. Through the analysis of a new dataset that combines full-count U.S. census information with historical newspaper archives, we provide compelling evidence that the dissemination of information about urban amenities, particularly public health improvements such as water filtration and sewage systems, directly influenced the migratory decisions of rural individuals. Our findings highlight the critical impact of local newspapers in bridging the informational gap about the benefits of living in cities that were investing in sanitation infrastructure, thereby making these urban centers more attractive to potential migrants. Notably, we observe that rural migrants were significantly more likely to relocate to areas boasting public health investments (PHI) when they had access to a local newspaper that reported on these urban advancements.

The implications of our study extend beyond the historical context of the United States, offering potential insights into the ongoing urbanization trends in regions such as Latin America, Asia, and Africa. These areas are currently experiencing rapid urban growth in mega-cities like São Paulo, Jakarta, and Mumbai. One plausible explanation for this phenomenon, as suggested by our research, is the increased availability of information. In today's digital age, where information dissemination is even more widespread and accessible than in the past, our findings underscore the importance of information availability in shaping migration patterns. This is particularly relevant in the context of epidemics or pandemics, where real-time information about public health conditions can significantly influence the reallocation of populations.

As the world continues to urbanize at a rapid rate, the lessons drawn from our study emphasize the need for policymakers to focus on enhancing information dissemination about urban amenities and public health investments. This strategy could be instrumental in promoting urban growth and development, especially in developing regions where the rapid expansion of cities presents both challenges and opportunities. In conclusion, our research not only contributes to the academic discourse on spatial and urban economics but also offers practical insights for contemporary urban planning and policy-making, highlighting the enduring relevance of information in shaping human migratory behaviors.

.

References

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and **Santiago Pérez** (2021). "Automated Linking of Historical Data". In: *Journal of Economic Literature* 59 (3), pp. 865–918 (Cited on page 8).

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, Santiago Perez, and **Myera Rashid** (2020). "Census Linking Project: Version 2.0 [dataset]". In: (Cited on pages 5, 8, 46–50, 52).

Almagro, Milena and **Tomas Domiguez Lino** (2022 WP). "Location Sorting and Endogenous Amenities: Evidence from Amsterdam". In: (Cited on page 7).

Alsan, Marcella and **Claudia Goldin** (Apr. 2019). "Watersheds in Child Mortality: The Role of Effective Water and Sewerage Infrastructure, 1880–1920". In: *J. Polit. Econ.* 127 (2), pp. 586–638 (Cited on pages 20, 22).

Anderson, D. Mark, Kerwin Kofi Charles, and **Daniel I. Rees** (2022). "Reexamining the Contribution of Public Health Efforts to the Decline in Urban Mortality". In: *American Economic Journal: Applied Economics* 14 (2), pp. 126–57 (Cited on pages 4, 5, 7, 13–15, 20).

Ao, Chon-Kit (2015). "Clean Water and Human Capital Investment". In: *Working Paper* (Cited on pages 5, 16–18).

Beach, Brian (May 2022a). "Water and Waste: A History of Reluctant Policymaking in U.S. Cities". Invited contribution to proposed volume: *Managing Water: Lessons from History* for University of Pittsburgh Press' History of the Urban Environment Series (Edited by: Joel Tarr and Martin Melosi) (Cited on page 15).

— (2022b). "Water infrastructure and health in U.S. cities". In: *Regional Science and Urban Economics* 94. Urban Economics and History, p. 103674. ISSN: 0166-0462 (Cited on pages 4, 5, 7, 13, 16, 17, 20, 22).

Beach, Brian, Joseph Ferrie, Martin Saavedra, and **Werner Troesken** (Mar. 2016). "Typhoid Fever, Water Quality, and Human Capital Formation". In: *J. Econ. Hist.* 76 (1), pp. 41–75 (Cited on pages 16, 17).

Beach, Brian, John Parman, and **Martin Saavedra** (May 2022). "Segregation and the Initial Provision of Water in the United States". In: *AEA Papers and Proceedings* 112, pp. 193–198 (Cited on page 15).

Berry, Steven T. (1994). "Estimating Discrete-Choice Models of Product Differentiation". In: *RAND Journal of Economics* 25 (2), pp. 242–262 (Cited on pages 4, 27, 29).

Boustan, Leah, Devin Bunten, and **Owen Hearey** (Jan. 2018). "Urbanization in American economic history, 1800–2000". In: 2, pp. 75–99 (Cited on pages 7, 11).

Caliendo, Lorenzo, Maximiliano Dvorkin, and **Fernando Parro** (2019). "Trade and Labor Market Dynamics: General Equilibrium Analysis of the China Trade Shock". In: *Econometrica* 87 (3), pp. 741–835 (Cited on page 26).

Costa, Dora L and **Matthew E Kahn** (July 2017). "Death and the media: infectious disease reporting during the health transition". en. In: *Economica* 84 (335), pp. 393–416 (Cited on pages 7, 20).

Cutler, David and **Grant Miller** (2005). "The Role of Public Health Improvements in Health Advances: The Twentieth-Century United States". In: *Demography* 42 (1), pp. 1–22. ISSN: 00703370, 15337790 (Cited on pages 7, 13, 15).

Dell, Melissa, Jacob Carlson, Tom Bryan, Emily Silcock, Abhishek Arora, Zejiang Shen, Luca D'Amico-Wong, Quan Le, Pablo Querubin, and **Leander Heldring** (Aug. 2023). "American Stories: A Large-Scale Structured Text Dataset of Historical U.S. Newspapers". In: (Cited on pages 5, 19).

Diamond, Rebecca (2016). "The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980-2000". In: *American Economic Review* 106 (3), pp. 479–524 (Cited on page 7).

Dickstein, Michael J and **Eduardo Morales** (July 2018). "What do Exporters Know?*". In: *The Quarterly Journal of Economics* 133 (4), pp. 1753–1801. ISSN: 0033-5533.

Donaldson, Dave and **Richard Hornbeck** (2016). " Railroads and American Economic Growth: A "Market Access" Approach *". In: *The Quarterly Journal of Economics* 131 (2), pp. 799–858 (Cited on pages 33, 34, 36, 63–66, 68–71).

Eckert, Fabian and **Michael Peters** (2018). "Spatial Structural Change". In: (98) (Cited on pages 5, 10).

Fee, Elizabeth (1987). *Disease and Discovery: A History of the Johns Hopkins School of Hygiene and Public Health, 1916–1939*. Baltimore: Johns Hopkins University Press.

Fee, Elizabeth and **F B Smith** (Oct. 1990). "The Retreat of tuberculosis, 1850-1950". In: *Am. Hist. Rev.* 95 (4), p. 1197.

Fujiwara, Thomas, Eduardo Morales, and **Charly Porcher** (Work-in-progress). "A Revealed-Preference Approach to Measuring Information Frictions in Migration Decisions". In.

Gentzkow, Matthew, Edward L Glaeser, and **Claudia Goldin** (Sept. 2004). "The Rise of the Fourth Estate: How Newspapers Became Informative and Why It Mattered" (Cited on page 19).

Gentzkow, Matthew, Jesse M. Shapiro, and **Michael Sinkinson** (2011). "The Effect of Newspaper Entry and Exit on Electoral Politics". In: *American Economic Review* 101 (7), pp. 2980–3018 (Cited on pages 5, 18, 19).

Gentzkow, **Matthew**, **Jesse M Shapiro**, and **Michael Sinkinson** (Oct. 2014). "Competition and Ideological Diversity: Historical Evidence from US Newspapers". In: *Am. Econ. Rev.* 104 (10), pp. 3073–3114 (Cited on page 19).

Glaeser, **Edward** and **David Cutler** (2022). *Survival of the City: Living and Thriving in an Age of Isolation*. New York: Penguin Press. ISBN: 978-0593297681.

Haines, **Michael R.** and **Inter-university Consortium for Political and Social Research** (May 2010). *Historical, Demographic, Economic, and Social Data: The United States, 1790-2002* (Cited on page 16).

Institute of Medicine (US) Committee for the Study of the Future of Public Health (1988). *The Future of Public Health*. Washington (DC): National Academies Press (US). Chap. 3.

Kuziemko, **Ilyana** and **Ebonya Washington** (2018). "Why Did the Democrats Lose the South? Bringing New Data to an Old Debate". In: *American Economic Review* 108 (10), pp. 2830–67.

Moretti, **Enrico** (2013). "Real Wage Inequality". In: *American Economic Journal: Applied Economics* 5 (1), pp. 65–103.

Petrova, **Maria** (2011). "Newspapers and Parties: How Advertising Revenues Created an Independent Press". In: *Am. Polit. Sci. Rev.* 105 (4), pp. 790–808 (Cited on page 19).

Porcher, **Charly** (2020 WP). "Migration with Costly Information". In: (Cited on page 7).

Ruggles, **Steven**, **Sarah Flood**, **Matthew Sobek**, **Daniel Backman**, **Annie Chen**, **Grace Cooper**, **Stephanie Richards**, **Renae Rodgers**, and **Megan Schouweiler** (2024a). *IPUMS USA: Version 15.0 [dataset]*. https://doi.org/10.18128/D010.V15.0. Minneapolis, MN (Cited on page 5).

Ruggles, **Steven**, **Matt A. Nelson**, **Matthew Sobek**, **Catherine A. Fitch**, **Ronald Goeken**, **J. David Hacker**, **Evan Roberts**, and **J. Robert Warren** (2024b). *IPUMS Ancestry Full Count Data: Version 4.0 [dataset]*. https://doi.org/10.18128/D014.V4.0. Minneapolis, MN (Cited on page 5).

Smith, **Francis Barrymore** (1988). *The Retreat of Tuberculosis, 1850-1950*. en. Croom Helm (Cited on page 41).

Steinwender, **Claudia** (2018). "Real Effects of Information Frictions: When the States and the Kingdom Became United". In: *American Economic Review* 108 (3), pp. 657–96.

Troesken, **Werner** (1999). "Typhoid Rates and the Public Acquisition of Private Waterworks, 1880-1920". In: *J. Econ. Hist.* 59 (4), pp. 927–948 (Cited on pages 13, 15, 17).

— (2001). "Race, Disease, and the Provision of Water in American Cities, 1889-1921". In: *The Journal of Economic History* 61 (3), pp. 750–776. ISSN: 00220507, 14716372 (Cited on page 13).

Wilson, **Riley** (2022). "Moving to Economic Opportunity". In: *Journal of Human Resources* 57 (3), pp. 918–955. ISSN: 0022-166X (Cited on page 7).

Winslow, **C. E. A.** (1923). "The Evolution and Significance of the Modern Public Health Campaign". In: *Journal of Public Health Policy*.

Wohl, **Anthony S.** (1983). *Endangered Lives: Public Health in Victorian Britain*. Cambridge, MA: Harvard University Press (Cited on page 17).

Appendices

A Sample Selection: Sampling from Censuses

Matching full-count decennial censuses from IPUMS comes with some sampling bias, as detailed in (Abramitzky et al., 2020). Some demographic groups are more challenging to accurately match and link across census years due to various factors. For instance, women frequently change their surnames upon marriage, making it difficult to establish reliable connections. As a result, the Abramitzky et al. (2020) only provides linkage information for men. Additionally, match rates for African Americans and certain immigrant populations tend to be lower compared to those of white, U.S.-born men. Finally, when dealing with small samples of foreign-born individuals, transcription errors can lead to an increased likelihood of false positive matches.

The following statistics compar the districution of population in the full-count U.S. census and in our matched census data.

A.1 Sampling of Population in Matched Census

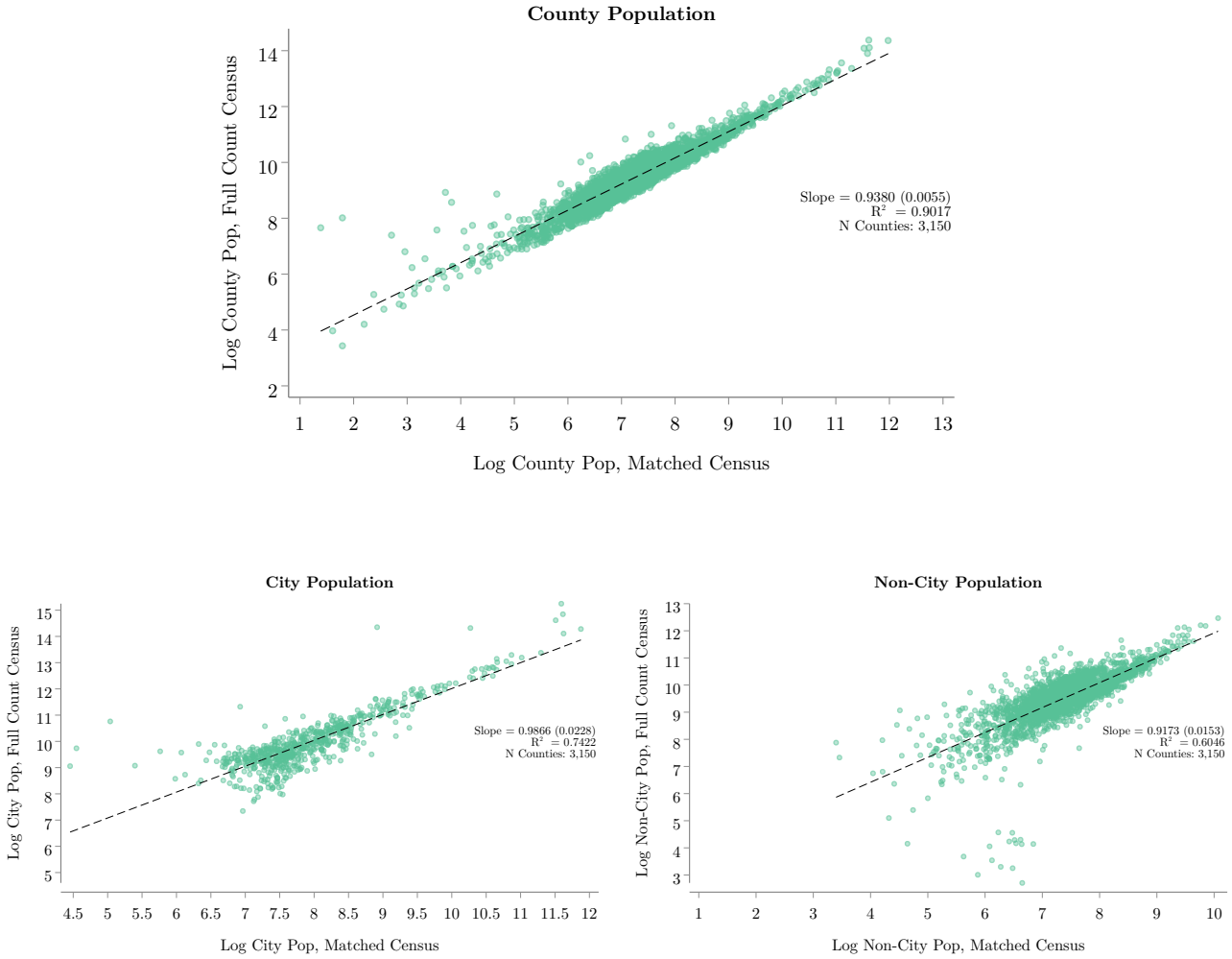

Figure A.1: **Population Distribution in Matched versus Full-Count U.S. censuses**

Notes: These graphs plot county (top), city (bottom left) or non-city (bottom right) level population (in logs), comparing the distribution of population in the full-count U.S. census (1870 to 1940) to our matched version of this full-count census. The matched version is a subset of the full-count census in which individuals have been matched between two census years. We use the most conservative matching method from Abramitzky et al. (2020) to match individuals in adjacent census, using unique exact name and age in a 5-year age band.

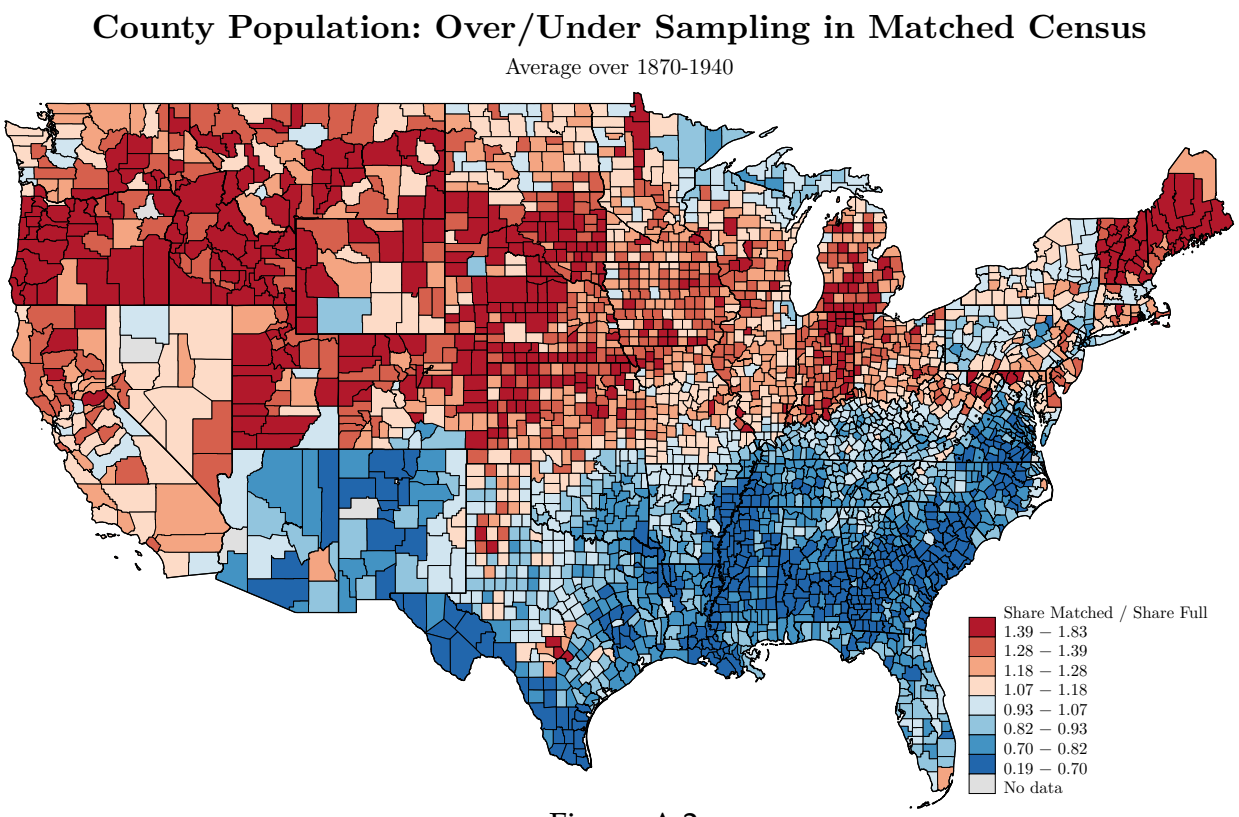

Figure A.2

Notes: This map compares the distribution of county-level population in the full-count U.S. census to our matched version of this full-count census. County populations are averaged over the period 1870 to 1940. The matched-census version is a subset of the full-count census in which individuals have been matched across two adjacent census years, in order to compute migration responses. We use the most conservative matching method from Abramitzky et al. (2020) to match individuals in adjacent census, using unique exact name and age in a 5-year age band (ABE-exact). For each county, we compute the population share compared to the entire U.S.. We then divide the population share of each county in our matched census data by the population share in the full-count census. Counties in red are oversampled in our matched census dataset compared to the full-count census, while counties in blue are undersampled. The South of the United States is more likely to suffer from undersampling, which may be largely driven by higher shares of African-American population in the south. The Census Linking Project mostly captures white males and African-American populations are less likely to be matched across census years (see Abramitzky et al. (2020) for more details).

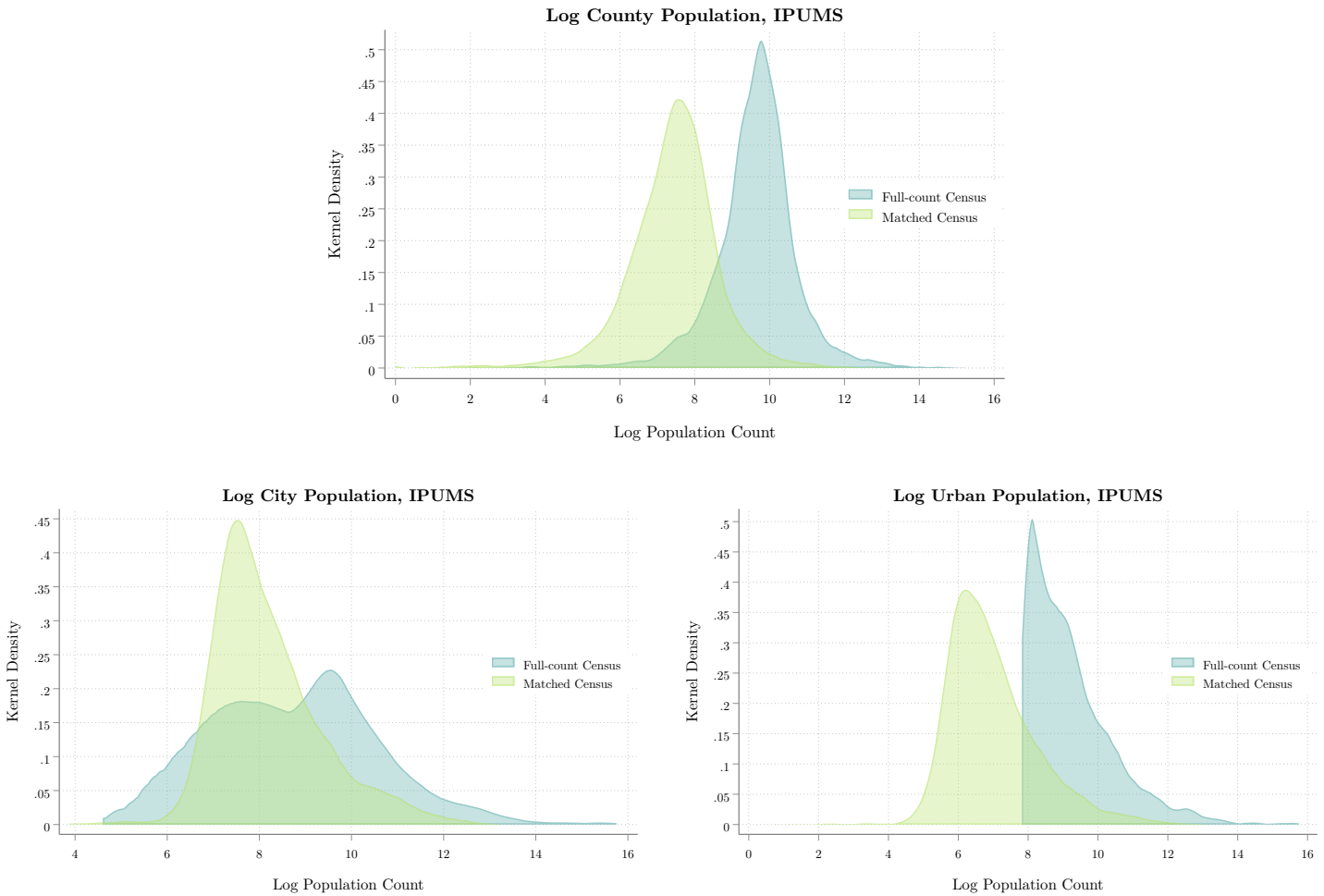

Figure A.3: Distribution of Population (in logs) by County, Matched versus Full-Count census

Notes: These graphs plot county (top), city (bottom left) or non-city (bottom right) level population (in logs), comparing the distribution of population in the full-count U.S. census (1870 to 1940) to our matched version of this full-count census. The matched version is a subset of the full-count census in which individuals have been matched between two census years. We use the most conservative matching method from Abramitzky et al. (2020) to match individuals in adjacent census, using unique exact name and age in a 5-year age band (ABE-exact).

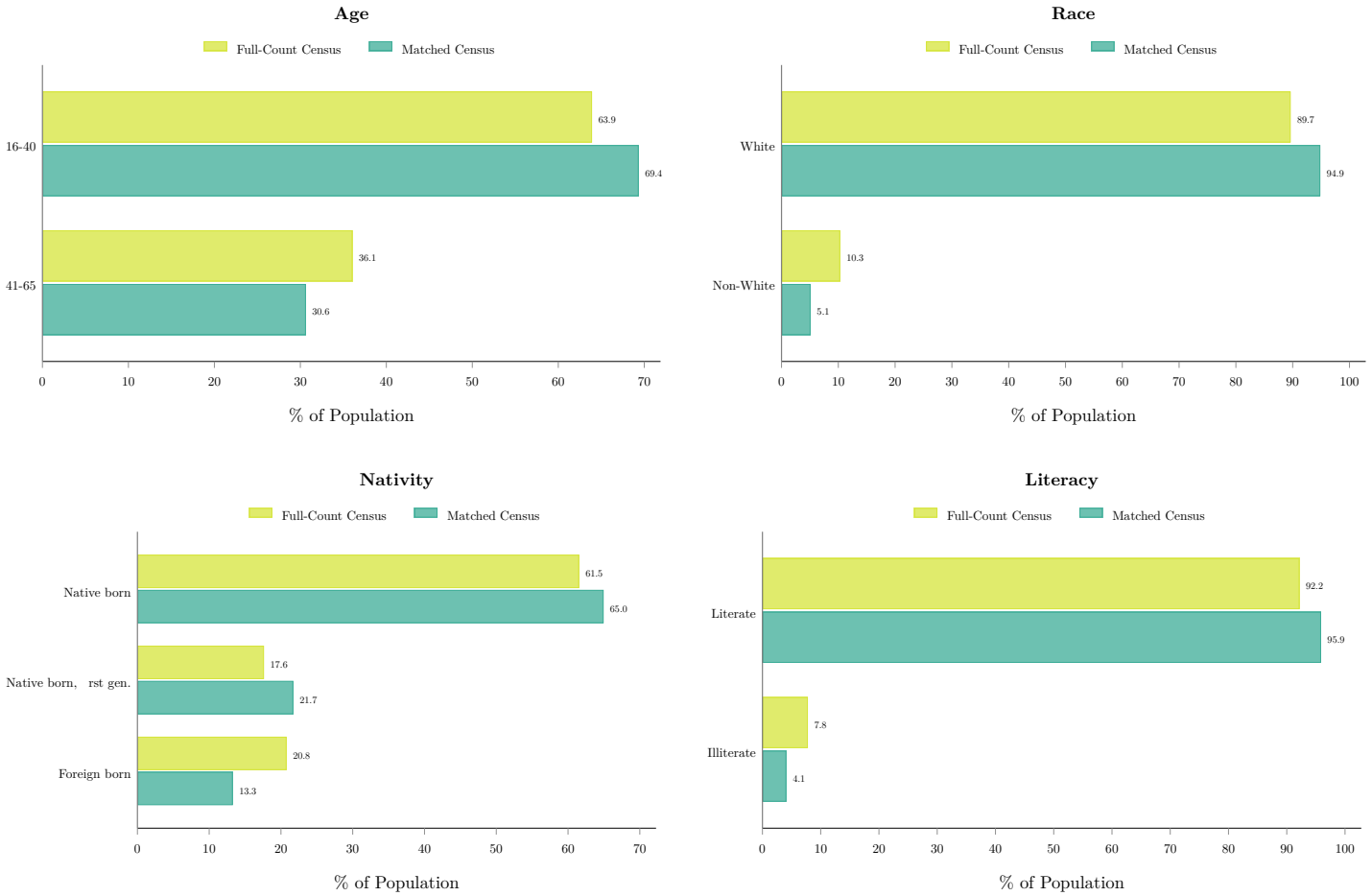

Figure A.4: **Distribution of Demographics, Full-Count vs Matched Census**

Notes: These graphs compare the distribution of demographics in the full-count U.S. census and our matched version of the full-count census. The matched version is a subset of the full-count census in which individuals have been matched between two census years. We use the most conservative matching method from Abramitzky et al. (2020) to match individuals in adjacent census, using unique exact name and age in a 5-year age band (ABE-exact). The light green bars represent the full-count census, while the dark green bars represent our matched-version of the census.

A.2 Documenting Birthplace (Origin Locations)

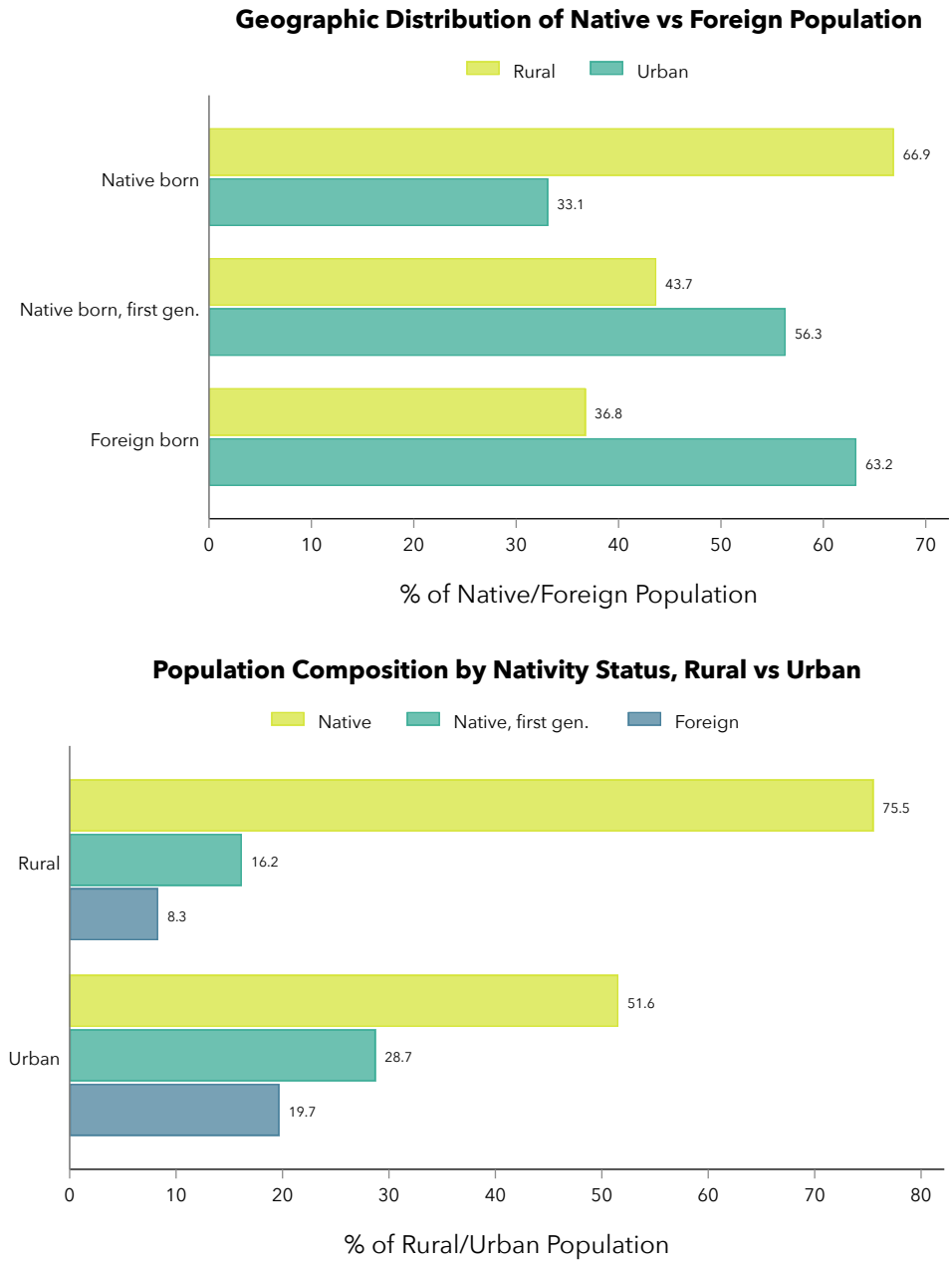

Figure A.5: **Documenting Birthplace (Origin Locations)**

Notes: These graphs document the birthplace of individuals in our matched census data. The top graph compares rural and urban population share across nativity status. For each nativity status (native born, native born first generation or foreign born), the total population share sums to 100. The vast majority of individuals hat are native born live in rural areas (66.9%). On the contrary, only about half of native born that are first-generation U.S. citizens are living in rural places (43.7%) and most foreign-born individuals live in urban ares (62.2%). The bottom graph compares nativity status *within* a location. For each location type (rural or urban), the sum of bar plots sum to 100. 75% of individuals in rural areas are native, 16.2% are native but first-generation and 8.3% only are foreign-born.Urban areas count 51.6% of native-born, 28.7% of first-generation native born and 19.7% of foreign-born. These two graphs underline that rural areas are mostly composed of populations that immigrated in the U.S. earlier than those located in urban areas.

A.3 Occupations & Sectors Switching Rates

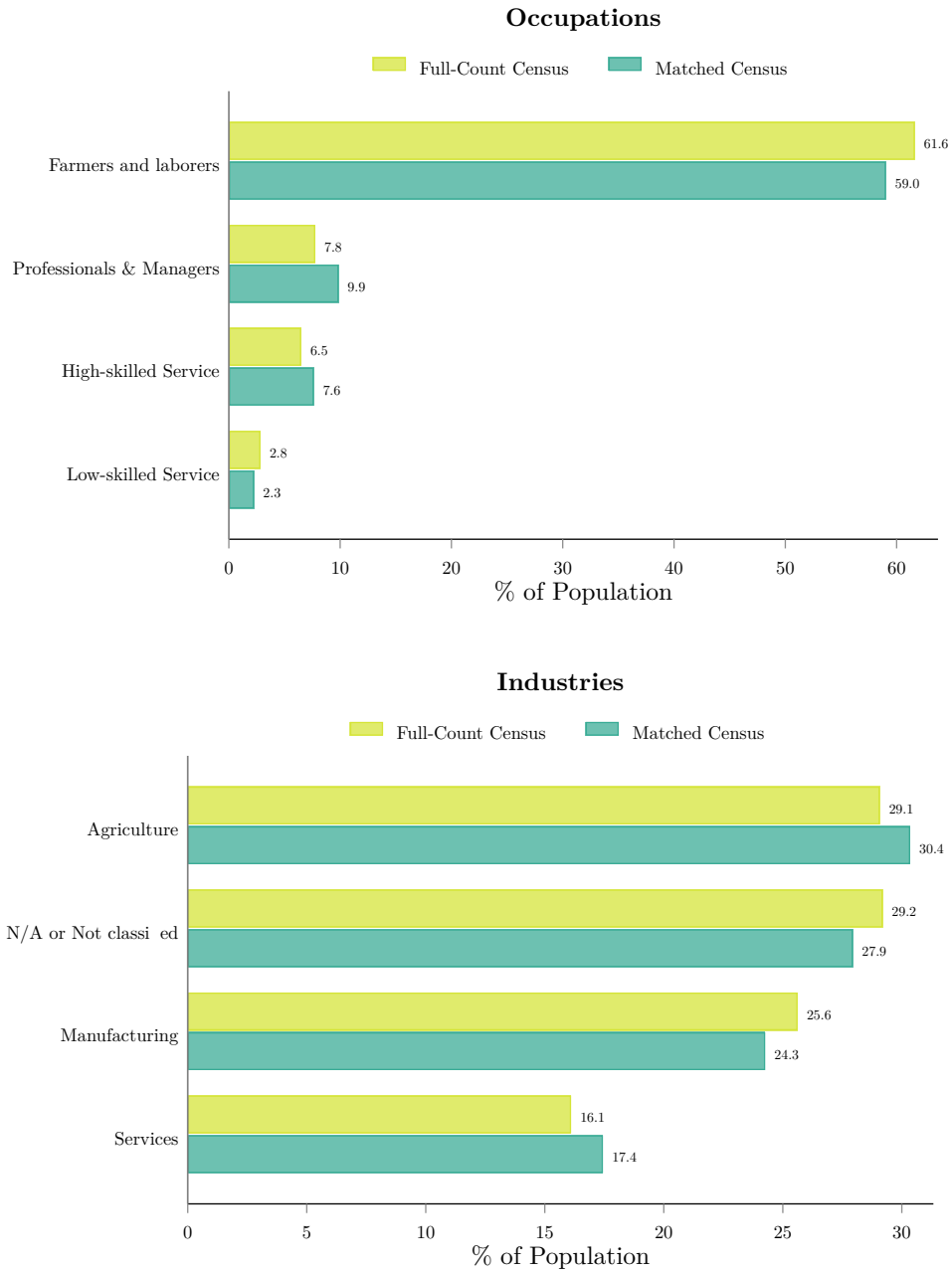

Figure A.6: **Distribution of Individuals across Occupations and Indsutries**

Notes: These graphs compare the distribution of populations by occupations and industries in the full-count U.S. census and our matched version of the full-count census. The matched version is a subset of the full-count census in which individuals have been matched between two census years. We use the most conservative matching method from Abramitzky et al. (2020) to match individuals in adjacent census, using unique exact name and age in a 5-year age band (ABE-exact). The light green bars represent the full-count census, while the dark green bars represent our matched-version of the census.

Table A.1: **Yearly Share of Movers Across Occupations and Sectors**

Year	Occupation 1 digit	Industry 1 Digit	Industry 3 Groups	Occupation 4 Groups	Occupation-Industry 1×1 Digit	Occupation-Industry 3× 4 Groups
1880	0.544	0.439	0.401	0.127	0.576	0.309
1900	0.636	0.531	0.489	0.146	0.667	0.322
1910	0.610	0.525	0.486	0.118	0.658	0.264
1920	0.600	0.524	0.477	0.110	0.662	0.244
1930	0.612	0.555	0.500	0.122	0.696	0.266
1940	0.623	0.603	0.511	0.162	0.728	0.326

Notes: This table details individuals' switching rates between intercensal years in terms of occupation and industry. The number for the first row denotes the share of people switching occupation or industry between 1870 and 1880 in the matched census data. Note that there is no 1890 full-count U.S. census. For robustness, we display switching rates for different defitions of occupation and industries. The first column refers to 1-digit occupation categories as defined by the IPUMS full-count census. The second column does the same for 1-digit industry. The two following columns aggregate industry and occupations into 3 or 4 groups. The two last columns look at switching rates across occupation-industry bins.

Table A.2: **Aggregate Share of Movers Across Occupations and Sectors**

Category	Value
Occupation 1 digit	0.611
Industry 1 Digit	0.552
Industry 3 Groups	0.491
Occupation 4 Groups	0.133
Occupation-Industry 1×1 Digit	0.686
Occupation-Industry 3× 4 Groups	0.286

Notes: This table details individuals' switching rates in terms of occupation and industry. Movers are defined as individuals moving at least once over the 1870-1940 period. For robustness, we display switching rates for different defitions of occupation and industries. The first row refers to 1-digit occupation categories as defined by the IPUMS full-count census. The second row does the same for 1-digit industry. The two subsequent rows aggregate industry and occupations into 3 or 4 groups. The two last rows look at switching rates across occupation-industry bins.

Table A.3: **Aggregate Share of Movers Across Geography Units and Industries**

Geography	Share Movers	Industry/Sector Def.	Share Movers
City	0.335	Occupation Groups	0.133
Urban	0.378	Sector Groups	0.391
County	0.323	Occupations × Sectors	0.286

Notes: These tables display the share of movers in out matched census data across geography units, across industry sectors and across occupations. Industries and sectors are not aggregagted into groups. Movers are defined as individuals moving at least once over the 1870-1940 period.

B Descriptive Statistics: Newspaper Coverage of Diseases

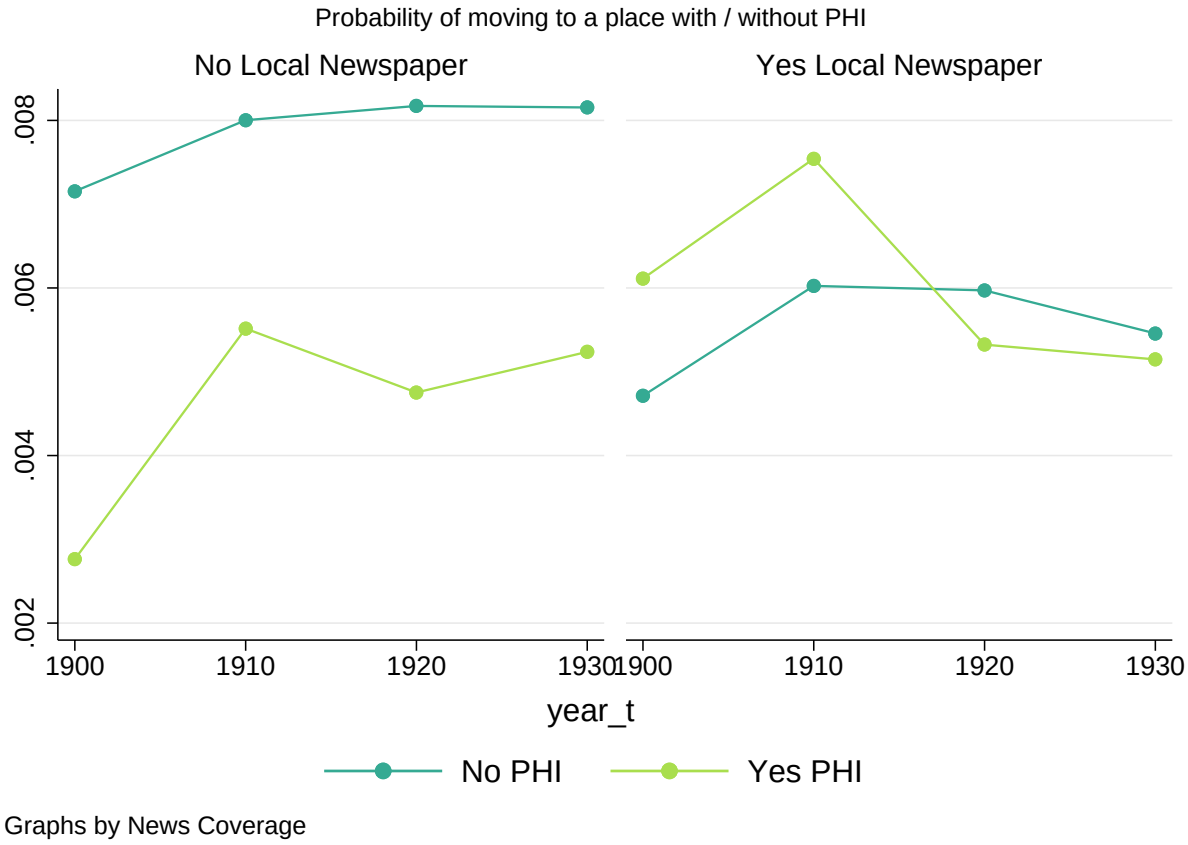

Figure B.1: Differential migration to destinations with water filtration, depending on the presence of local newspaper at origin

Notes: This figure depicts the average migration rates across two types of origins and destinations respectively. The left panel depicts average rates from origins without a local newspaper and the right panel depicts that from origins with a local newspaper. Within each panel, the dark green line depicts average rate from that origin type to destinations without a water filtration plant and the light green line depicts that to destinations with a water filtration plant. The y-axis denotes the share of individuals from an origin type who migrated to a destination type for each census year. The x-axis denotes the census years.

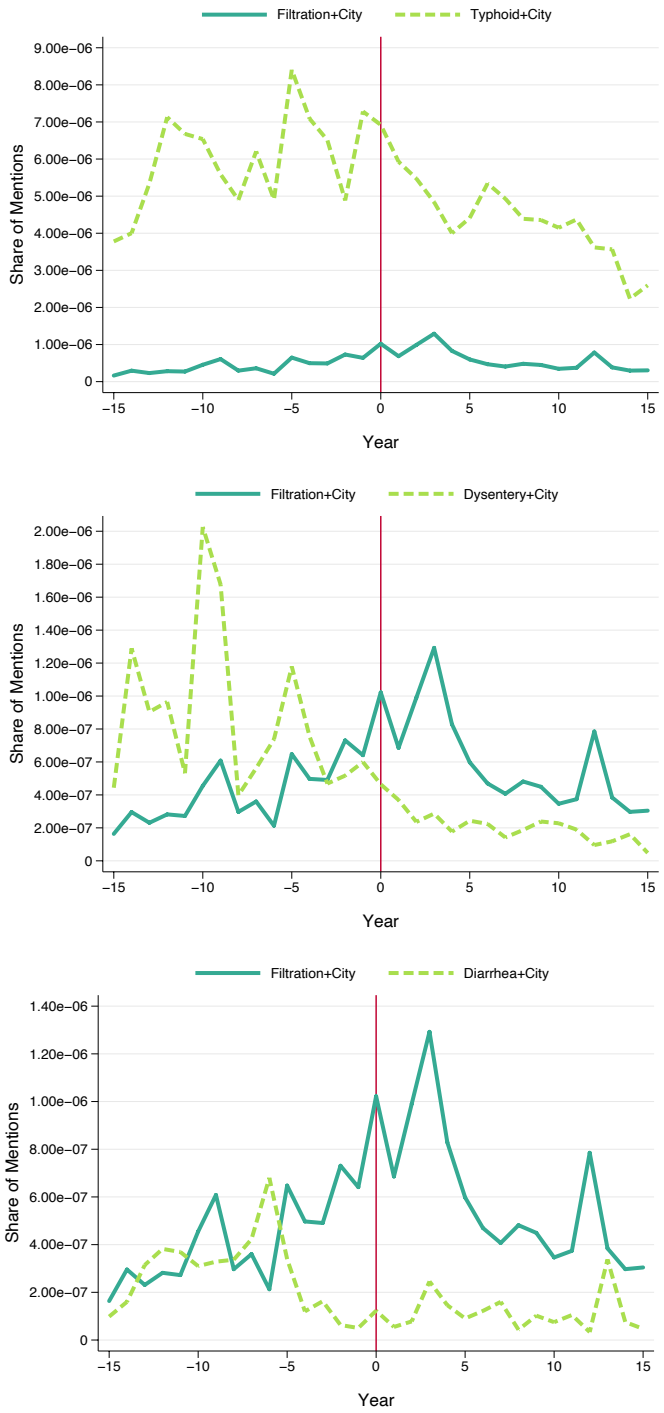

Figure B.2: **Most of the drop in diseases outbreaks reporting comes from typhoid news**

Notes: These graphs report the share of newspaper articles mentioning waterborne diseases that are associated to a city in our dataset. The sample is currently limited to the 74 cities in our public health investment dataset, some of these cities not having an implemented water filtration system over the period. We compute the number of keyword hits having both a city name and either typhoid, dysentery or diarrhea sentences appearing next to each other in an article. Each graph is centered around the year of implementation of a water filtration system in the city. In each graph, we superpose the share of mentions to waterborne diseases (including typhoid, dysentry, and diarrhea) for these cities. The timing of water filtration investments correspond with a decrease in reports of diseases outbreaks in the newspapers, and this decrease is mostly driven by a decrease in typhoid reporting.

C Data Appendix

C.1 Constructing Historical Wage Measures

C.1.1 Wage Imputation Framework

The full-count historical censuses from IPUMS do not include wages or income measures at the individual level for the 1870-1930 period. Individuals' wages appear for the first time in the 1940 full-count census.

We detail in this section the pre-1940 wage imputation technique we use to predict individual-level real wages for the period before 1940, for each potential destination that that individual could end up in.

An individual in our data is differentiated by their census **t**ime period, **o**rigin county, **d**estination county, occu**p**ation, and demographic **g**roup (defined by an age category, a race category and a literacy group). We make the assumption throughout this section that the origin of a person's wage is irrelevant to their wage at destination d, so we don't have an origin subscript.

An individual i at destination d's real wage is written as followed:

$$\tilde{w}_i = \tilde{w}_{tdpg} = \frac{w_{tdpg}}{\Pi_{td}}$$

where $\tilde{w}$ refers to the real wage, w refers to the nominal wage, and Π refers to the price index. By having a single price index, we hold the share of food and non-food expenditure constant.

The data we need to perform this exercise is:

- Nominal wages measured at the final period (1940), which varies by destination county, occupation, and demographic group w^*_{dpg}.
- Average nominal wage for each occupation-period $\overline{w}_{tp}$
- Average nominal wage for each destination-period $\overline{w}_{td}$
- Average nominal wage for each demographic group-period $\overline{w}_{tg}$
- The distribution of occupation by time-destination $\phi_{td}(p)$ where $\sum_p \phi_{td}(p) = 1$
- The distribution of demographics by time-destination $\gamma_{td}(g)$ where $\sum_g \gamma_{td}(g) = 1$
- The distribution of population across destinations by time $\kappa_t(d)$ where $\sum_d \kappa_t(d) = 1$
- Price index for each period, not for each destination, but varied by urban vs. rural

destination: Π_t^{urban} and Π_t^{rural}

We want to predict w_{tdpg}, which represents $T \times D \times P \times G$ numbers. Suppose we have 6 census years, 50 destinations (49 states with urban destinations + 1 common rural wage), 10 occupation-sector pairs, and 2 demographic groups (white/non-white), combining into a total of 6,000 numbers.

With $\overline{w}_{td}$, $\overline{w}_{tp}$, ϕ_{td} and γ_{td} available, note that

$$\overline{w}_{td} = \sum_p \phi_{td}(p) \left[\sum_g \gamma_{td}(g) w_{tdpg} \right] \quad \text{[6×50=300 moments]}$$

$$\overline{w}_{tp} = \sum_d \phi_{td}(p) \left[\sum_g \gamma_{td}(g) w_{tdpg} \right] \quad \text{[6×10=60 moments]}$$

$$\overline{w}_{tg} = \sum_d \kappa_t(d) \left[\sum_p \phi_{td}(p) w_{tdpg} \right] \quad \text{[6×2=12 moments]}$$

The final equation is leveraging the information from the nominal wages measured at the final period. If we assume that

$$w_{tdpg} = \alpha_t \times w^*_{dpg}$$

then we assume that the distribution of wages across counties, occupations, and demographic groups have stayed constant over time. We do not want to make this assumption – we are making a statement about migration across counties, and therefore it would be unwise to assume that the wage distribution across counties over time is constant and the same as the terminal distribution.

Suppose that we are comfortable instead with assuming that while the distribution across destinations have changed, the distribution across occupation-sector and demographic groups have not. We can then write:

$$w_{tdpg} = \alpha_{td} \times w^*_{dpg}$$

This adds 300 numbers that we need to estimate: α_{td}, but on the other hand we earn 6,000 moments.

C.1.2 Estimation Process

Time-scalar only

$$w_{tdpg} = \alpha_t \times w^*_{dpg}$$

In this case, we just need to compute w^*_{dpg} and then merge by dpg (destination-occupation-demographic group) cell to each individual in our data.

- An issue that we would get from using the 1% census sample is that the number of individuals per cell can get quite small. We aim to get to a minimum of 5 individuals per cell and median of about 15 individuals per cell. A cell must be at least a county, occupation, and demographic.
- We make the decision to group all rural areas within a state together. This could be changed in future research by enriching the data via the digitization of the Census of Agriculture rural wages.
- Given that 10% of the population is black in our sample, cells for black individuals end up being too thin. To solve this issue, we calculate a average of wage gap for southern states and non-southern states in 1940, and assume that the wage gap was the same going backward. Table C.1 reports the racial wage gap by occupation group and region (south versus North)

Table C.1: Racial Wage Gap by Occupation Group and Region

U.S. Region	Occupation Group	Racial Wage Gap
North	Professionals & Managers	0.575
South	Professionals & Managers	0.350
North	High-skilled Service	0.714
South	High-skilled Service	0.538
North	Low-skilled Service	0.800
South	Low-skilled Service	0.625
North	Farmers and laborers	0.636
South	Farmers and laborers	0.556

- In term of geography, we are eventually left with:
  - Within each state, we group all rural counties together to have the same wage.
  - Then we group all counties within a "city" designation together for the same wage.
  - e.g., for the state of Washington, there are four geographic divisions: Rural, Seattle counties, Spokane counties, and Tacoma counties.

– In future iterations, we could limit the definition of "Seattle counties" here to only include the urban core of Seattle, and drop surrounding counties into the Rural designation, or perhaps have a Suburban designation.

Just-identification

Suppose that we do not use the information from $\overline{w}_{tp}$ and $\overline{w}_{tg}$, and only use the mean wage by each destination-time $\overline{w}_{td}$.

Under this specification, we end up with the following system of equations:

$$\overline{w}_{td} = \sum_{p,g} \phi_{td}(p)\, \gamma_{td}(g)\, w_{tdpg}$$
$$\frac{1}{\alpha_{td}} w_{tdpg} = w^*_{dpg}$$

Which means that we are now trying to solve 6,300 equations for 6,300 parameters. We can derive exact solutions:

$$\overline{w}_{td} = \alpha_{td} \left[\sum_{p,g} \phi_{td}(p)\, \gamma_{td}(g)\, w^*_{dpg} \right]$$
$$\implies \alpha_{td} = \frac{\overline{w}_{td}}{\left[\sum_{p,g} \phi_{td}(p)\, \gamma_{td}(g)\, w^*_{dpg} \right]}. \tag{5}$$

Equation 5 evidences that in order to construct the mean wage of a county-period, we first need to compute the counterfactual mean wage of this county in the final period if the occupation and demographics distributions are the same as in the current period:

$$\sum_{p,g} \phi_{td}(p)\, \gamma_{td}(g)\, w^*_{dpg}$$

and since the only thing that changed is some destination-time-specific factor α_{td}, the ratio between the two mean wages gives us the scaling factor.

Data

The final imputation exercise eventually requires the following data:

- w^*_{dpg} comes from computing a median or mean of wages at the end of the sample using the 1940 Census.

- $\phi_{td}(p)$ comes directly from our data.
- $\gamma_{td}(g)$ comes directly from our data.
- $\overline{w}_{td}$:

Table C.2: $\overline{w}_{td}$

	Urban state	Rural state	Urban County	Rural County
1890			Census RoM., tab.1	COA (ICPSR 35206)
1900	CoM1905, tab.69		Census1900 RoM, v7p3 Tab. XXVIII	
1910	BLS history (sec/occ). Roy (CoM)	BLS History, p.562	Census1910 RoM, v8p2 pg.92, 93. v9 (state-level)	
1920	BLS history (sec/occ). Census1920 RoM tab.48		Census1920 RoM tab.50 tab. 51	
1930	BLS history (sec/occ)		COM (ICPSR 37114)	

Note: Roy (CoM) is limited to only 15-ish states.

C.2 IPUMS Definition of Geographical Units

URBAN indicates whether a individual's location was urban or rural. Definitions of "urban" vary from year to year, but the term generally denotes all cities and incorporated places of 2,500+ inhabitants. All areas not classified as urban are designated rural. More details are given below.

CITY identifies the city of residence for individuals located in identifiable cities. This variable is essentially comparable across years, but not all cities are identified in all years, and there are some variations in the exact correspondence between CITY codes and city residents. A year-by-year discussion follows. All rank and size measurements refer to contemporary (not current) population figures, unless otherwise noted:

- 1850 and 1880: The city of residence is given if the individual was in one of the nation's 98 largest cities.
- 1860 and 1870: The city of residence is given for individuals in any city with 10,000+ inhabitants.
- 1900 - 1930: The city of residence is given for individuals in any city with 25,000+ inhabitants.

CITYPOP reports the population, in hundreds, for all identifiable cities.

URBPOP gives the population, in hundreds, of places considered "urban" according to the census Bureau's 1930-1940 definition. All incorporated municipalities with a population of 2,500 or more were defined as urban, but this definition of urban also included some places that were not incorporated municipalities.

Details on the definition of Urban areas. The definition of "urban" areas in the U.S. censuses from 1870 to 1940 initially relied on population size and incorporation status, then incorporated population density, and finally included unincorporated areas and adjacent areas meeting specific criteria. Below we provide more details on the changes in this definition:

- From 1870 to 1900, the census distinguished between "'urban" and "rural" based on the population size of incorporated places. The threshold varied:
  - 1860-1880: Places with 8,000 or more inhabitants were considered urban.
  - 1890: Places with 4,000 or more inhabitants were considered urban.
  - 1900: Places with 2,500 or more inhabitants were considered urban.

- For census years 1910-192, the census defined "urban" as incorporated places with 2,500 or more inhabitants, as well as towns and townships with a total population of 10,000 or more and a density of 1,000 persons per square mile.

- For the 1930 census year, the definition of "urban" was expanded to include:
  - Incorporated places with 2,500 or more inhabitants.
  - Towns and townships with a total population of 10,000 or more and a density of 1,000 persons per square mile.
  - Unincorporated places with a population of 2,500 or more.

- In 1940, the census further expanded the definition of "urban" to include:
  - Incorporated places with 2,500 or more inhabitants.
  - Unincorporated places with 2,500 or more inhabitants.
  - Areas adjacent to incorporated places with a population of 2,500 or more, if the adjacent area had a population of 500 or more and a density of 1,000 persons per square mile.

D In-Migrations Regressions Framework - Additional Empirical Results

In all regressions below, for the Typhoid version, the set of destination cities is restricted to the 73 cities for which we have typhoid rates, from Brian Beach's dataset (other cities have missing values for Typhoid). We do not consider individuals who did not move. The y variable is the logit log share (i.e. log share of people moving from O to D minus log share of people staying put - the outside option).

Table D.1: **All origin locations, city destinations only**

	Dependent Var.: Probability of moving							
Fixed Effects:	**Origin-Year-Group**				**Origin-Year-Group & Dest-Year**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
News Coverage × Typhoid rate	-.9694***	-.5241*			-.5352***	-.4197**		
	[.3118]	[.2753]			[.1712]	[.206]		
Typhoid rate	-.7254***	.1174						
	[.1591]	[.2843]						
News Coverage × Water Filtration			.4144***	.39***			.2398***	.2256***
			[.1452]	[.1276]			[.07764]	[.08223]
Water Filtration			.904***	.7123***				
			[.07147]	[.06052]				
Log Distance	-.8374***	-.8717***	-.624***	-.7582***	-1.044***	-1.033***	-.9329***	-.939***
	[.0853]	[.08351]	[.0651]	[.05496]	[.02535]	[.03006]	[.01239]	[.01419]
Mean wage of group g in d	1.6e-07***	7.5e-08***	2.3e-07***	1.8e-07***	5.0e-08***	4.1e-08***	5.3e-08***	4.1e-08***
	[3.2e-08]	[2.1e-08]	[4.0e-08]	[3.4e-08]	[9.7e-09]	[1.2e-08]	[6.4e-09]	[7.2e-09]
Log Distance × Water Filtration	-.07372	-.08288	-.2946***	-.2871***	-.1732***	-.1776***	-.2255***	-.224***
	[.07525]	[.05857]	[.03454]	[.02859]	[.02309]	[.02645]	[.01682]	[.01874]
Log Transportation Cost	.6309***	.5248**	.5419***	.6776***	.7943***	.7538***	.8398***	.8347***
	[.2142]	[.2374]	[.165]	[.1607]	[.05643]	[.06611]	[.03392]	[.03905]
County-level controls ($d - o$)		✓		✓		✓		✓
Occupation FE	✓	✓	✓	✓	✓	✓	✓	✓
Constant	-4.141***	-4.541***	-5.301***	-5.63***	-4.192***	-4.199***	-5.188***	-5.224***
	[.2352]	[.2615]	[.182]	[.1788]	[.05596]	[.1342]	[.03938]	[.06666]
R2	.5994	.6681	.6341	.6849	.7986	.8001	.8092	.8103
F-Stat	369.8	381.6	804.6	630.4	866.6	453.8	1301	683.2
Dep. Var.: Mean	-4.623	-4.626	-5.152	-5.182	-4.623	-4.626	-5.152	-5.182
Dep. Var.: Std. Dev.	1.506	1.52	1.628	1.644	1.506	1.52	1.628	1.644
N Obs.	6.1e+05	4.8e+05	1.4e+06	1.1e+06	6.1e+05	4.8e+05	1.4e+06	1.1e+06

Notes: * (p<0.1), ** (p<0.05), *** (p<0.01). All coefficients are standardized beta coefficients. Standard errors are clustered at the origin-group-year and destination-year level. Typhoid rates are measured as deaths by thousand. All regressions include the interaction of log distance and public health investment (water filtration plant), log distance, log transportation costs from Donaldson and Hornbeck (2016) and mean wage of demographic group g in destination d. County-level control variables include share male, white, literate, foreign-born and share in the manufacturing sector. All controls are bilateral measures and capture the difference between an origin and destination pair, as they are otherwise not separately identified from the fixed effects.

Table D.2: **All origin locations, city destinations only- Robustness: alternative sample keeping only $o-d$ pairs with migration above zero for more than one year**

	Dependent Var.: Probability of moving							
Fixed Effects:	**Origin-Year-Group**				**Origin-Year-Group & Dest-Year**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
News Coverage × Typhoid rate	-.9845***	-.4583			-.5981***	-.4548**		
	[.3242]	[.2847]			[.1762]	[.2113]		
Typhoid rate	-.7489***	.1212						
	[.1653]	[.2816]						
News Coverage × Water Filtration			.4196***	.3831***			.182**	.1658*
			[.1518]	[.1395]			[.0857]	[.09294]
Water Filtration			.803***	.6047***				
			[.07059]	[.05956]				
Log Distance	-.8163***	-.8764***	-.555***	-.7559***	-1.094***	-1.082***	-1.054***	-1.062***
	[.08031]	[.07764]	[.06909]	[.05345]	[.02451]	[.02927]	[.01374]	[.01608]
Mean wage of group g in d	1.7e-07***	7.7e-08***	2.5e-07***	1.7e-07***	4.2e-08***	3.6e-08***	3.9e-08***	2.9e-08***
	[3.3e-08]	[2.3e-08]	[4.0e-08]	[3.4e-08]	[1.1e-08]	[1.4e-08]	[7.6e-09]	[9.0e-09]
Log Distance × Water Filtration	-.07073	-.09237	-.3089***	-.3003***	-.1493***	-.1581***	-.141***	-.141***
	[.07806]	[.06055]	[.03973]	[.03394]	[.02726]	[.03156]	[.01934]	[.02201]
Log Transportation Cost	.6173***	.5407**	.4611***	.6627***	.7988***	.7611***	.8217***	.8177***
	[.2024]	[.2266]	[.1749]	[.1679]	[.05399]	[.06393]	[.03179]	[.03762]
County-level controls ($d-o$)		✓		✓		✓		✓
Occupation FE	✓	✓	✓	✓	✓	✓	✓	✓
Constant	-4.153***	-4.599***	-5.184***	-5.594***	-4.154***	-4.164***	-5***	-5.059***
	[.2273]	[.2516]	[.1933]	[.1882]	[.05229]	[.1394]	[.03684]	[.08387]
R2	.601	.6707	.6225	.6826	.8085	.8095	.8218	.8224
F-Stat	329.7	351	560.2	578.8	953.6	472.3	1478	741
Dep. Var.: Mean	-4.583	-4.583	-4.997	-5.01	-4.583	-4.583	-4.997	-5.01
Dep. Var.: Std. Dev.	1.519	1.533	1.668	1.683	1.519	1.533	1.668	1.683
N Obs.	5.5e+05	4.3e+05	1.1e+06	8.6e+05	5.5e+05	4.3e+05	1.1e+06	8.6e+05

Notes: * (p<0.1), ** (p<0.05), *** (p<0.01). All coefficients are standardized beta coefficients. Standard errors are clustered at the origin-group-year and destination-year level. Typhoid rates are measured as deaths by thousand. All regressions include the interaction of log distance and public health investment (water filtration plant), log distance, log transportation costs from Donaldson and Hornbeck (2016) and mean wage of demographic group g in destination d. County-level control variables include share male, white, literate, foreign-born and share in the manufacturing sector. All controls are bilateral measures and capture the difference between an origin and destination pair, as they are otherwise not separately identified from the fixed effects.

Table D.3: **Rural origins and city destinations only**

	Dependent Var.: Probability of moving							
Fixed Effects:	**Origin-Year-Group**				**Origin-Year-Group & Dest-Year**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
News Coverage × Typhoid rate	-.5806***	-.3071			-.3199***	-.2805**		
	[.192]	[.1897]			[.1117]	[.1375]		
Typhoid rate	-.8369***	-.1383						
	[.1695]	[.2316]						
News Coverage × Water Filtration			.2551***	.2531***			.1431***	.1286**
			[.08973]	[.0809]			[.0489]	[.05236]
Water Filtration			.9123***	.7409***				
			[.07325]	[.06707]				
Log Distance	-.9487***	-1.02***	-.6609***	-.8293***	-1.153***	-1.136***	-.975***	-.9838***
	[.06677]	[.06379]	[.05364]	[.04094]	[.02861]	[.03277]	[.01491]	[.01704]
Mean wage of group g in d	9.6e-08***	3.7e-08**	1.7e-07***	1.4e-07***	3.4e-08***	2.8e-08***	2.5e-08***	1.5e-08**
	[2.2e-08]	[1.7e-08]	[2.9e-08]	[2.5e-08]	[8.1e-09]	[9.5e-09]	[5.7e-09]	[6.6e-09]
Log Distance × Water Filtration	-.09105	-.1126***	-.3192***	-.3281***	-.2426***	-.2456***	-.3292***	-.33***
	[.05621]	[.04188]	[.0335]	[.02589]	[.02454]	[.02678]	[.02034]	[.02125]
Log Transportation Cost	.744***	.769***	.5374***	.7662***	1.043***	.9935***	.9861***	.9939***
	[.1655]	[.1745]	[.1355]	[.1245]	[.06335]	[.07208]	[.04301]	[.05056]
County-level controls ($d - o$)		✓		✓		✓		✓
Occupation FE	✓	✓	✓	✓	✓	✓	✓	✓
Constant	-4.052***	-4.648***	-4.956***	-5.492***	-4.335***	-4.006***	-5.116***	-5.085***
	[.1912]	[.2217]	[.1592]	[.1485]	[.06546]	[.1187]	[.05147]	[.07239]
R2	.6708	.7071	.6686	.7014	.8037	.8032	.8069	.8046
F-Stat	338.2	318.7	1222	912.9	879.6	512.3	3110	1567
Dep. Var.: Mean	-4.421	-4.412	-4.7	-4.724	-4.421	-4.412	-4.7	-4.724
Dep. Var.: Std. Dev.	1.462	1.472	1.54	1.543	1.462	1.472	1.54	1.543
N Obs.	3.8e+05	3.0e+05	9.5e+05	7.6e+05	3.8e+05	3.0e+05	9.5e+05	7.6e+05

Notes: * (p<0.1), ** (p<0.05), *** (p<0.01). All coefficients are standardized beta coefficients. Standard errors are clustered at the origin-group-year and destination-year level. Typhoid rates are measured as deaths by thousand. All regressions include the interaction of log distance and public health investment (water filtration plant), log distance, log transportation costs from Donaldson and Hornbeck (2016) and mean wage of demographic group g in destination d. County-level control variables include share male, white, literate, foreign-born and share in the manufacturing sector. All controls are bilateral measures and capture the difference between an origin and destination pair, as they are otherwise not separately identified from the fixed effects.

Table D.4: **Rural origins and city destinations only - Robustness: alternative sample keeping only O-D pairs with migration above zero for more than one year.**

	Dependent Var.: Probability of moving							
Fixed Effects:	**Origin-Year-Group**				**Origin-Year-Group & Dest-Year**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
News Coverage × Typhoid rate	-.5665***	-.2352			-.3354***	-.2803**		
	[.1988]	[.1986]			[.1101]	[.1324]		
Typhoid rate	-.8764***	-.1614						
	[.1793]	[.2321]						
News Coverage × Water Filtration			.2458***	.2424***			.09742*	.08406
			[.09051]	[.08391]			[.05169]	[.05654]
Water Filtration			.8106***	.6222***				
			[.071]	[.0615]				
Log Distance	-.9345***	-1.03***	-.5925***	-.829***	-1.198***	-1.18***	-1.069***	-1.08***
	[.06292]	[.05846]	[.05675]	[.03837]	[.02795]	[.0324]	[.0149]	[.01731]
Mean wage of group g in d	9.9e-08***	3.8e-08**	1.8e-07***	1.3e-07***	2.8e-08***	2.4e-08**	1.7e-08**	8.5e-09
	[2.2e-08]	[1.8e-08]	[2.9e-08]	[2.5e-08]	[9.6e-09]	[1.2e-08]	[7.2e-09]	[8.6e-09]
Log Distance × Water Filtration	-.0918	-.1285***	-.3416***	-.3506***	-.2293***	-.235***	-.2641***	-.2667***
	[.05842]	[.04311]	[.03975]	[.02994]	[.02464]	[.02763]	[.0197]	[.02174]
Log Transportation Cost	.7436***	.8056***	.4485***	.7543***	1.063***	1.016***	.9394***	.9514***
	[.1559]	[.1631]	[.1426]	[.1264]	[.06296]	[.07302]	[.04203]	[.05052]
County-level controls ($d - o$)		✓		✓		✓		✓
Occupation FE	✓	✓	✓	✓	✓	✓	✓	✓
Constant	-4.068***	-4.73***	-4.79***	-5.472***	-4.327***	-4.008***	-4.92***	-4.914***
	[.1845]	[.2169]	[.1674]	[.1545]	[.06364]	[.1169]	[.04938]	[.08162]
R2	.6707	.7071	.6515	.6905	.8113	.8102	.8127	.8096
F-Stat	301.2	286	903.3	774.6	880.8	468.6	2601	1278
Dep. Var.: Mean	-4.348	-4.337	-4.458	-4.467	-4.348	-4.337	-4.458	-4.467
Dep. Var.: Std. Dev.	1.468	1.477	1.552	1.553	1.468	1.477	1.552	1.553
N Obs.	3.4e+05	2.7e+05	7.2e+05	5.7e+05	3.4e+05	2.7e+05	7.2e+05	5.7e+05

Notes: * (p<0.1), ** (p<0.05), *** (p<0.01). All coefficients are standardized beta coefficients. Standard errors are clustered at the origin-group-year and destination-year level. Typhoid rates are measured as deaths by thousand. All regressions include the interaction of log distance and public health investment (water filtration plant), log distance, log transportation costs from Donaldson and Hornbeck (2016) and mean wage of demographic group g in destination d. County-level control variables include share male, white, literate, foreign-born and share in the manufacturing sector. All controls are bilateral measures and capture the difference between an origin and destination pair, as they are otherwise not separately identified from the fixed effects.

E In-Migrations: Placebo Tests

This section presents additional in-migration regression results based on placebo tests that use past migration flows. The first two tables restrict the sample to city destinations, but no restrictions is made on origin locations. The two following tables restrict the sample to both rural origins and city destinations.

We present two series of place tests for each specification: one in which we use as a dependent variable migration shares from 10 years ago, and one in which we use migrations from 20 years ago.

Table E.1: **Placebo test, dependent variable: migrations 10 years ago. Restricting the sample to city destinations only. No restrictions on origins.**

	Dependent Var.: Probability of moving							
Fixed Effects:	**Origin-Year-Group**				**Origin-Year-Group & Dest-Year**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
News Coverage × Typhoid rate	-.4901	.1448			.3786	.39		
	[.5222]	[.436]			[.3567]	[.3654]		
Typhoid rate	.2348	.4518						
	[.9182]	[1.065]						
News Coverage × Water Filtration			-.1393	.02432			-.3805**	-.3933**
			[.1931]	[.187]			[.1714]	[.1892]
Water Filtration			-.1792	-.2596				
			[.162]	[.1949]				
Log Distance	-.5007***	-.4953***	-.08921	-.01018	-.5663***	-.499***	.121**	.1035*
	[.08821]	[.09394]	[.08694]	[.09205]	[.06493]	[.07079]	[.05248]	[.05757]
Mean wage of group g in d	1.6e-07***	1.2e-07***	-1.1e-07	-5.4e-08	1.2e-07***	1.1e-07***	1.4e-07***	1.4e-07***
	[4.6e-08]	[4.2e-08]	[6.7e-08]	[7.3e-08]	[2.7e-08]	[3.1e-08]	[1.8e-08]	[2.0e-08]
Log Distance × Water Filtration	.05712	.06356	-.259***	-.202***	-.2261***	-.2128***	-.5676***	-.5458***
	[.05476]	[.05121]	[.06719]	[.06869]	[.06739]	[.07064]	[.07045]	[.07311]
Log Transportation Cost	.5993***	.493**	.5387**	.3224	.9411***	.784***	.6988***	.7181***
	[.1977]	[.2141]	[.2181]	[.2244]	[.1459]	[.1655]	[.1301]	[.1467]
County-level controls ($d - o$)		✓		✓		✓		✓
Occupation FE	✓	✓	✓	✓	✓	✓	✓	✓
Constant	-4.321***	-4.442***	-4.107***	-3.975***	-4.674***	-7.516***	-4.582***	-5.303***
	[.2379]	[.2573]	[.211]	[.22]	[.1565]	[1.144]	[.1406]	[.3532]
R2	.3495	.3517	.2479	.2528	.4206	.4297	.4662	.4762
F-Stat	23.26	18.93	21.55	20.6	34.88	25.54	41.99	28.97
Dep. Var.: Mean	-4.178	-4.311	-3.622	-3.717	-4.178	-4.311	-3.622	-3.717
Dep. Var.: Std. Dev.	2.771	2.784	3.52	3.561	2.771	2.784	3.52	3.561
N Obs.	2.8e+05	2.2e+05	6.3e+05	5.1e+05	2.8e+05	2.2e+05	6.3e+05	5.1e+05

Notes: * (p<0.1), ** (p<0.05), *** (p<0.01). This table presents additional in-migration regression results based on placebo tests that use past migration flows. All coefficients are standardized beta coefficients. Standard errors are clustered at the origin-group-year and destination-year level. Typhoid rates are measured as deaths by thousand. All regressions include the interaction of log distance and public health investment (water filtration plant), log distance, log transportation costs from Donaldson and Hornbeck (2016) and mean wage of demographic group g in destination d. County-level control variables include share male, white, literate, foreign-born and share in the manufacturing sector. All controls are bilateral measures and capture the difference between an origin and destination pair, as they are otherwise not separately identified from the fixed effects.

Table E.2: **Placebo test, dependent variable: migrations 20 years ago. Restricting the sample to city destinations only. No restrictions on origins.**

	Dependent Var.: Probability of moving							
Fixed Effects:	**Origin-Year-Group**				**Origin-Year-Group & Dest-Year**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
News Coverage × Typhoid rate	.2003	.9623***			1.213***	1.273***		
	[.6155]	[.348]			[.3111]	[.3225]		
Typhoid rate	-.1384	.1276						
	[.7322]	[.7467]						
News Coverage × Water Filtration			.16	.1145			-.1444	-.3351**
			[.2514]	[.2717]			[.1676]	[.159]
Water Filtration			-.8935***	-.726***				
			[.2348]	[.2801]				
Log Distance	-.06423	.109	.06194	.3368***	-.2636***	-.1705	.2072***	.2531***
	[.1165]	[.1132]	[.103]	[.1084]	[.08821]	[.1067]	[.05208]	[.06168]
Mean wage of group g in d	2.2e-07***	1.2e-07**	-2.1e-07***	-1.6e-07*	1.1e-07***	9.7e-08**	9.7e-08***	9.2e-08***
	[7.2e-08]	[5.0e-08]	[8.1e-08]	[8.2e-08]	[3.4e-08]	[3.9e-08]	[1.9e-08]	[2.1e-08]
Log Distance × Water Filtration	.03189	-.029	-.2531***	-.2698***	-.3851***	-.3684***	-.5407***	-.5625***
	[.06791]	[.05263]	[.07965]	[.07819]	[.08539]	[.09188]	[.07934]	[.08255]
Log Transportation Cost	.1359	-.369*	.3607	-.2351	.883***	.6476***	.6365***	.606***
	[.2772]	[.2172]	[.2585]	[.2688]	[.1846]	[.2138]	[.1311]	[.1562]
County-level controls ($d - o$)		✓		✓		✓		✓
Occupation FE	✓	✓	✓	✓	✓	✓	✓	✓
Constant	-3.951***	-3.604***	-3.442***	-3.078***	-4.7***	-7.981***	-4.127***	-4.937***
	[.2815]	[.2333]	[.2507]	[.2566]	[.1889]	[.5578]	[.1365]	[.2683]
R2	.3034	.3242	.2263	.2355	.4029	.4028	.4741	.4548
F-Stat	3.725	7.008	29	26.44	12.5	180.3	35.49	25.6
Dep. Var.: Mean	-3.8	-3.845	-3.069	-3.215	-3.8	-3.845	-3.069	-3.215
Dep. Var.: Std. Dev.	3.192	3.187	3.641	3.664	3.192	3.187	3.641	3.664
N Obs.	2.5e+05	2.0e+05	5.6e+05	4.6e+05	2.5e+05	2.0e+05	5.6e+05	4.6e+05

Notes: * (p<0.1), ** (p<0.05), *** (p<0.01). This table presents additional in-migration regression results based on placebo tests that use past migration flows. All coefficients are standardized beta coefficients. Standard errors are clustered at the origin-group-year and destination-year level. Typhoid rates are measured as deaths by thousand. All regressions include the interaction of log distance and public health investment (water filtration plant), log distance, log transportation costs from Donaldson and Hornbeck (2016) and mean wage of demographic group g in destination d. County-level control variables include share male, white, literate, foreign-born and share in the manufacturing sector. All controls are bilateral measures and capture the difference between an origin and destination pair, as they are otherwise not separately identified from the fixed effects.

Table E.3: **Placebo test, dependent variable: migrations 10 years ago. Restricting the sample to rural origins and city destinations.**

	Dependent Var.: Probability of moving							
Fixed Effects:	**Origin-Year-Group**				**Origin-Year-Group & Dest-Year**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
News Coverage × Typhoid rate	.01428	.4191			.4384	.6666*		
	[.5569]	[.3471]			[.3398]	[.3775]		
Typhoid rate	.2355	-.2518						
	[1.054]	[.9195]						
News Coverage × Water Filtration			-.00147	-.06773			.1333	-.01761
			[.2266]	[.2862]			[.1945]	[.2296]
Water Filtration			-.2882	-.01781				
			[.1963]	[.2602]				
Log Distance	-.6624***	-.6477***	-.1238	.08426	-.7671***	-.6688***	.07975	.06477
	[.1629]	[.1972]	[.1398]	[.1421]	[.1151]	[.1473]	[.07619]	[.09294]
Mean wage of group g in d	2.5e-08	2.0e-08	-2.0e-07**	-1.1e-07	8.6e-08*	9.1e-08*	9.6e-08***	9.7e-08***
	[6.6e-08]	[7.3e-08]	[8.4e-08]	[9.1e-08]	[4.7e-08]	[5.5e-08]	[3.0e-08]	[3.6e-08]
Log Distance × Water Filtration	.1474	.117	-.239	-.1577	-.3461*	-.3251	-.6513***	-.5687***
	[.0913]	[.1424]	[.156]	[.178]	[.1761]	[.2124]	[.1664]	[.2025]
Log Transportation Cost	1.164***	1.259**	.9144**	.526	1.842***	1.747***	1.187***	1.252***
	[.4193]	[.5573]	[.3646]	[.409]	[.2677]	[.3638]	[.2212]	[.2839]
County-level controls ($d - o$)		✓		✓		✓		✓
Occupation FE	✓	✓	✓	✓	✓	✓	✓	✓
Constant	-4***	-3.943***	-3.349***	-2.628***	-4.741***	-15.87***	-3.996***	-5.24***
	[.4438]	[.6893]	[.3693]	[.4356]	[.2885]	[1.686]	[.2451]	[.4319]
R2	.2852	.2889	.2181	.2557	.3778	.4145	.4396	.4736
F-Stat	4.944	6.017	17.21	27.9	18.7	1347	29.84	1097
Dep. Var.: Mean	-3.163	-3.3	-2.393	-2.432	-3.163	-3.3	-2.393	-2.432
Dep. Var.: Std. Dev.	2.911	3.019	3.348	3.432	2.911	3.019	3.348	3.432
N Obs.	83774	58504	1.7e+05	1.2e+05	83774	58504	1.7e+05	1.2e+05

Notes: * (p<0.1), ** (p<0.05), *** (p<0.01). This table presents additional in-migration regression results based on placebo tests that use past migration flows. All coefficients are standardized beta coefficients. Standard errors are clustered at the origin-group-year and destination-year level. Typhoid rates are measured as deaths by thousand. All regressions include the interaction of log distance and public health investment (water filtration plant), log distance, log transportation costs from Donaldson and Hornbeck (2016) and mean wage of demographic group g in destination d. County-level control variables include share male, white, literate, foreign-born and share in the manufacturing sector. All controls are bilateral measures and capture the difference between an origin and destination pair, as they are otherwise not separately identified from the fixed effects.

Table E.4: **Placebo test, dependent variable: migrations 20 years ago. Restricting the sample to rural origins and city destinations.**

	Dependent Var.: Probability of moving							
Fixed Effects:	**Origin-Year-Group**				**Origin-Year-Group & Dest-Year**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
News Coverage × Typhoid rate	-.2638	.2949			.5583	.6702		
	[.8717]	[.649]			[.5013]	[.5113]		
Typhoid rate	.09452	.1928						
	[.8247]	[.738]						
News Coverage × Water Filtration			.05484	-.171			.1285	-.1335
			[.2889]	[.343]			[.2563]	[.2826]
Water Filtration			-.5917**	-.04986				
			[.2339]	[.2088]				
Log Distance	.06129	.4135**	.2254	.7535***	-.1786	-.00995	.2899***	.4078***
	[.2004]	[.2084]	[.1417]	[.1428]	[.1432]	[.1819]	[.07169]	[.09069]
Mean wage of group g in d	6.2e-08	-2.7e-08	-2.4e-07***	-1.6e-07*	2.9e-08	-1.8e-08	3.0e-08	1.5e-08
	[8.7e-08]	[8.2e-08]	[8.4e-08]	[8.1e-08]	[5.4e-08]	[6.9e-08]	[3.1e-08]	[3.8e-08]
Log Distance × Water Filtration	-.0497	-.2196*	-.3266**	-.4669***	-.7254***	-.7432***	-.7089***	-.7675***
	[.1167]	[.115]	[.1532]	[.1557]	[.1702]	[.1937]	[.16]	[.1819]
Log Transportation Cost	.3613	-.4887	.4856	-.5503	1.387***	1.075***	.9187***	.8352***
	[.5075]	[.4713]	[.3547]	[.4076]	[.297]	[.3732]	[.2027]	[.2716]
County-level controls ($d - o$)		✓		✓		✓		✓
Occupation FE	✓	✓	✓	✓	✓	✓	✓	✓
Constant	-3.621***	-2.708***	-3.109***	-2.157***	-4.689***	-12.73***	-3.85***	-4.954***
	[.4895]	[.5336]	[.3639]	[.4245]	[.3027]	[.5442]	[.2166]	[.4044]
R2	.2448	.2406	.2368	.2431	.3661	.3464	.4524	.4094
F-Stat	1.486	3.957	24	29.54	9.675	.	38.54	261
Dep. Var.: Mean	-3.082	-3.014	-2.349	-2.499	-3.082	-3.014	-2.349	-2.499
Dep. Var.: Std. Dev.	3.285	3.3	3.507	3.533	3.285	3.3	3.507	3.533
N Obs.	82657	57859	1.7e+05	1.2e+05	82657	57859	1.7e+05	1.2e+05

Notes: * (p<0.1), ** (p<0.05), *** (p<0.01). This table presents additional in-migration regression results based on placebo tests that use past migration flows. All coefficients are standardized beta coefficients. Standard errors are clustered at the origin-group-year and destination-year level. Typhoid rates are measured as deaths by thousand. All regressions include the interaction of log distance and public health investment (water filtration plant), log distance, log transportation costs from Donaldson and Hornbeck (2016) and mean wage of demographic group g in destination d. County-level control variables include share male, white, literate, foreign-born and share in the manufacturing sector. All controls are bilateral measures and capture the difference between an origin and destination pair, as they are otherwise not separately identified from the fixed effects.