Railroads and the Rural to Urban Transition: Evidence from 19th-Century Argentina*

Santiago Pérez†

August 30, 2018

Abstract

I provide microeconomic evidence on the role of reductions in transport costs in shaping the movement of labor out of agriculture. I exploit the expansion of the railroad network in 19th-century Argentina and individual-level longitudinal data. I construct an instrumental variable exploiting that districts along the route of province capitals were more likely to be connected. Adults in connected districts remained in farming. By contrast, their children moved out of farming toward more modern and higher paying occupations. This movement reflected increased outmigration and higher urbanization rates in connected districts, and was stronger in districts with low agricultural suitability.

*This is a shortened version of the third chapter of my dissertation completed at Stanford University. I thank my advisor Ran Abramitzky, as well as Arun Chandrasekhar, Victor Lavy, Melanie Morten and Gavin Wright for outstanding guidance as part of my dissertation committee. I am grateful to Dan Bogart, Raj Chetty, William Collins, Melissa Dell, Dave Donaldson, Pascaline Dupas, Xavier Duran, Marcel Fafchamps, James Feigenbaum, Joseph Ferrie, Price Fishback, Michela Giorcelli, Avner Greif, Caroline Hoxby, Claudia Rei, Mark Rosenzweig, Noam Yutchman and Ariell Zimran for detailed feedback. I am indebted to Stephen Redding and Pablo Fajgelbaum for sharing their GIS data on historical department borders of Argentina. I acknowledge the financial support of the Stanford University Economics Department, the Economic History Association, the Stanford Center for International Development and the Leonard W. Ely and Shirley R. Ely Graduate Student Fund Fellowship.

†Assistant Professor of Economics, University of California, Davis, Department of Economics. email: seperez@ucdavis.edu

Economists going back to North 1958 and Rostow 1959 have argued that improvements in transportation technology were a key driver of the transition from self-sufficient and mostly agrarian to modern economies. In this paper, I provide microeconomic evidence on the role of reductions in transport costs in shaping the movement of labor out of agriculture. To do so, I exploit the expansion of the railroad network in 19th-century Argentina and new longitudinal data following individuals before and after this expansion took place. I show how households responded to railroad construction through investments in human capital, occupational choices and migration decisions.

To conduct the analysis, I collect longitudinal data linking more than 30,000 males across the 1869 and 1895 Argentine censuses. In these data, I observe two groups of individuals. First, I observe children in the 1869 census and then link them to their long-term adult outcomes in 1895. Second, I observe working-age individuals in two points of their adult life. These data enable me to follow individuals over a long period of time, and as they moved across places and sectors of the economy. I combine these data with digitized maps of the historical transportation network of Argentina.

To address the potentially endogenous placement of railroads, I use an instrumental variables approach. As a byproduct of connecting province capitals, some intermediate districts not explicitly targeted were nevertheless connected to the railroad network. My identification strategy exploits that, among these districts, those located along a convenient route – from a cost minimizing perspective – were more likely to be connected to the network. More precisely, I instrument railroad access with access to a hypothetical network that connects all province capitals to each other in a cost-minimizing way, a similar strategy to that in Chandra and Thompson 2000, Michaels 2008, Banerjee, Duflo, and Qian 2012, Faber 2014 and Morten and Oliveira 2015.

By 1869, the year of the first national census, almost 70 percent of the Argentine population resided in rural areas and was employed in the primary sector. Because of the absence of navigable rivers in the interior of the country, transportation was conducted

using wagon carts (Conde 1979). By 1895, the railroad network had expanded from less than 700 kilometers in 1869 to more than 14,000 kilometers. Railroads brought a dramatic decline in the cost of moving goods and people. In 1874, traveling from Córdoba to Rosario (a 250-mile trip) took 15 hours by train – a trip that would have taken half a month in the pre-railroads era (Lewis 1983).

Railroad expansion could have influenced the exit out of agricultural occupations for three main reasons. First, railroads might have enabled individuals in rural areas to move physically. Railroads dramatically reduced travel times and also likely increased information flows across connected districts. Second, railroads might have enabled districts with relatively low productivity in farming to buy agricultural products from other districts. Third, the agglomeration of economic activity around train stations might have translated into higher urbanization rates in connected areas.

On the other hand, standard trade theory predicts that districts with a comparative advantage in the production of agricultural commodities should have increased agricultural production and employment after being connected to the railroad network. This distinction is especially relevant in the context of late 19th-century Argentina, a period characterized by the integration of the country into world markets as an exporter of agricultural commodities (Fajgelbaum and Redding 2014).

I first show that railroads had limited impacts on the probability that adults employed in farming or as farm workers in 1869 would transition out of these occupations. However, railroads had large effects on the probability that the *children* of farmers and farm workers would exit farming. Children of farmers and farm workers in connected districts were 13 percentage points more likely to work outside of farming in adulthood, relative to a baseline probability of 32 percent. As they exited farming, these children entered white-collar and skilled blue-collar jobs in adulthood.

How did railroads increase the probability of children exiting farming occupations? Children in connected districts might have transitioned out of farming by: (1) exiting

farming occupations but staying in their 1869 district of residence, or (2) both exiting farming and leaving their 1869 district. I find that at least 25 percent of the increase in the probability of exiting farming can be attributed to the fact that children in connected districts were more likely to *both* migrate and exit farming. This finding implies that railroads increased migration propensities and highlights the importance of tracking internal migrants to obtain a comprehensive assessment of the impacts of transport infrastructure.

At the same time, the exit out of farming was also driven by increased urbanization in connected districts: individuals who stayed in their 1869 district of residence accounted for most of the decline of farming in the children's generation. In particular, this decline was largely explained by children who in 1869 resided in districts where the soil was less suitable for agriculture. By contrast, there is no evidence of higher exit out of farming in connected districts with high agricultural suitability.

Why did the transition out of farming occur in the children's generation but not among adults? First, imperfections in the land market might have prevented individuals who worked their own land from exiting farming. Second, adults likely had a higher opportunity cost than children of acquiring the skills required to work outside of farming.[1]

The evidence is more consistent with a mismatch between the skills of adults and the skills required in non-farming occupations than with frictions in the land market. First, adults in connected districts were also more likely to migrate internally, which suggests that they were not stuck to their land. However, migration was less strongly associated with movements out of farming in the adults' generation than in the children's generation. Second, the effects of railroads on adults were similarly small regardless of whether I focus on those who likely owned land in 1869 or on those who likely did not. Third, consistent with the fact that transitioning into non-farming occupations required

[1]The returns to human capital in farming have likely increased in recent times due to skill-biased technological change in agriculture (Foster and Rosenzweig 1996). Yet, the allocation of workers across sectors strongly suggests lower returns to human capital in the farming sector than in other sectors of the economy. Gollin, Lagakos, and Waugh 2014 shows that in virtually every country in the world average years of schooling are higher outside of agriculture.

a different set of skills, I find that children in connected districts were more likely to be literate in adulthood.

One concern when using linked individual-level data to conduct such an analysis is that some individuals might be incorrectly matched (Bailey et al. 2017). In the context of this paper, if the share of incorrect matches was higher in districts connected to the railroad network, I would mechanically find higher rates of mobility out of farming in connected districts. To address this concern, I show that the results are similar when using a more conservatively linked sample *only* for connected districts, thus biasing the samples against the main findings of the paper.

This paper contributes to three main strands of the literature. First, it contributes to the literature on the economic impacts of transport infrastructure projects. Recent work on railroads has shown that their construction led to price convergence and increased regional trade (Donaldson 2015; Keller and Shiue 2008), higher agricultural land values (Donaldson and Hornbeck 2015), urbanization and city growth (Atack, Haines, and R. A. Margo 2009; Berger and Enflo 2015; Hornung 2015), higher school enrollment (Atack, R. Margo, and Perlman 2012), and the spread of factories (Atack, Haines, and R. A. Margo 2011). These studies have focused either on relatively aggregate outcomes or on cross-sectional individual-level outcomes. I complement the existing literature with micro-level evidence on how individual households respond to these interventions.[2]

A crucial innovation relative to these studies is in the use of individual-level longitudinal data. As improvements in transportation can make individuals more mobile geographically, longitudinal data enable me to disentangle individual-level responses from compositional changes in the population. Indeed, I empirically show that in my context railroads indeed induced these compositional changes.

Second, the findings of this paper contribute to the understanding of the movement of labor from farming into non-farming activities. A large literature in macroeconomics

[2]The broader literature on the impacts of transport infrastructure is summarized in Redding and Turner 2014.

describes theoretically and empirically how this transformation occurs along the process of development.[3] However, as pointed out by Foster and Rosenzweig 2007, the microeconomic aspects of this transition are not well understood. My findings provide empirical evidence consistent with the theoretical mechanisms underlying structural transformation in Matsuyama 1992 – that only relatively young workers are able to switch economic sectors – and Caselli and Coleman II 2001 – that exiting farming requires acquiring a new set of skills.

In related research, Adamopoulos 2011, Asher and Novosad 2016, Gollin and Rogerson 2014 and Herrendorf, Schmitz Jr, and Teixeira 2012 show that low productivity in the transportation sector can partially account for the high proportion of individuals employed in agriculture in developing countries. Similarly, Fajgelbaum and Redding 2014 demonstrate, also in the context of late 19th-century Argentina, that railroad construction resulted in structural transformation away from agriculture. My results shed light on the microeconomic mechanisms behind the relationship between improvements in transport infrastructure and structural change.

More broadly, this paper relates to the literature on the labor market adjustments to external factors such as trade shocks or technological change.[4] I use longitudinal data following workers and their children and consider the labor market impacts of one major technological change: the transportation revolution.

1 Brief Historical Background

The first railroad line of Argentina opened in 1857, connecting the city of Buenos Aires to the nearby town of San José de Flores. Figures A.2 and A.3 show the expansion of the railroad network that took place from 1870 to 1895. By 1869, the year of the first national census, the network was very limited, extending for less than 700 kilometers and

[3] This literature is summarized in Herrendorf, Rogerson, and Valentinyi 2014.

[4] There is a vast empirical literature on this issue. See, for instance Dix-Carneiro 2014 on labor market adjustments to trade shocks and Walker 2013 on adjustments to environmental regulations.

reaching only a handful of districts; less than 20 percent of the country's population lived within 10 kilometers of railroad lines, most of them in the city of Buenos Aires. After a moderate expansion during the 1870s, the 1880s featured a boom in railroad construction. By 1895, the network extended for more than 14,000 kilometers and was the longest in Latin America and the ninth longest in the world.[5]

To understand the location of railroad lines and the extent of government involvement in their financing and construction, it is useful to distinguish between Buenos Aires and the provinces of the *Interior*. Railroads in Buenos Aires were rapidly profitable and attracted the interest of private capital, mainly of British origin. The chief goal of these lines was to connect highly fertile land to ports so as to facilitate the exports of agricultural goods. By contrast, railroads in the Interior served areas with less agricultural exporting potential, as well as lower levels of population density and income.[6] The central government was heavily involved in the development of the railroad network in these areas of the country.

Extending the railroad network to the interior provinces was part of an effort to unify the country politically and economically. The first line of the Interior was the Central Argentine Railway, which connected the two main cities of Argentina outside of Buenos Aires: Rosario and Córdoba. By 1891, all province capitals of 1869 Argentina were connected to the railroad network.[7]

Prior to the railroads era, wagons constituted the main option for land travel. Wagons were substantially slower than trains; they traveled 3 to 4 kilometers an hour depending on terrain, compared to the 30 to 40 kilometers an hour of the early trains (Zalduendo

[5]This section is partly based on Lewis 1983.

[6]Differences existed within the interior provinces. The province of Santa Fe also experienced rapid railroad expansion and substantial involvement of private capital.

[7]There were only 14 official provinces in 1869 Argentina. The remaining areas of the country were considered national "territories" and had minimal state presence and non-native population. The capitals of the 14 provinces of 1869 Argentina were founded during the 16th century. The colonization of contemporary Argentina occurred in three waves: East, West and North. The colonizing wave of the East settled Buenos Aires (first in 1536, then in 1580), Santa Fe (1573) and Corrientes (1588) The settlers entering through the North funded the cities of Santiago Del Estero (1553), Tucumán (1565), Córdoba (1573), Salta (1581), La Rioja (1591) and Jujuy (1593). The colonizing wave of the West settled the cities of Mendoza (1561), San Juan (1562) and San Luis (1594).

1975). In addition to being slow, wagon travel was inconvenient and dangerous (Cárcano 1893).[8]

The advantage of railroads was more modest in terms of direct monetary costs. I estimate that the average monetary cost per passenger-kilometer of railroads was about half the cost of wagons (Ferrocarriles Nacionales 1896; Zalduendo 1975; Herranz-Loncán 2014). The direct monetary cost – excluding foregone earnings and expenses for shelter and food during the trip – of 100 miles of wagon travel was equivalent to about 3 days of pay for a laborer. Differences in the direct monetary cost of transporting goods were also substantial. Herranz-Loncán 2011 estimates that the price per ton-kilometer of transporting goods in 19th-century Argentina was seven to eight times higher using wagons than using trains.

2 Data

2.1 Creating the Linked Sample

I constructed a sample following males[9] through Argentina's 1869 and 1895 national censuses.[10]To create this sample, I identified two groups of individuals in the 1869 census full count: (1) males 0 to 16 years old with their father present in the household ("sons"),[11] and (2) males 18 to 35 years old ("adults"). These two groups included a total of 457,842 individuals.

[8]For instance, in describing the preparations undertaken before a wagon journey, Cárcano 1893 writes that: "Everything for the journey had to be prepared, from water to firewood, from shelter against the inclemencies of weather, to weapons to defend from the assaults of the road." Accounts of these assaults are common in the chronicles of the travelers of the period. In a letter to another priest quoted in Cárcano 1893, a Jesuit describes how a group of *abipones* – a indigenous tribe – killed 24 priests on their way from Córdoba to Santa Fe.

[9]Women were employed in an extremely limited set of occupations in this time period. I follow recent economic history papers (Abramitzky, Boustan, and Eriksson 2012; Abramitzky, Boustan, and Eriksson 2013; Abramitzky, Boustan, and Eriksson 2014; Collins and Wanamaker 2014; Collins and Wanamaker 2015; Feigenbaum 2016; Long and Ferrie 2013) and limit the analysis to males.

[10]These are the only two national censuses of Argentina for which individual records – including names – are currently available.

[11]See section A.3 for details on the procedure used to identify fathers and sons in the data.

I then searched the 1895 census for a set of potential matches for each of these individuals. Based on the similarity in their reported names and years of birth, I calculated a linking score ranging from 0 to 1 for each pair of potential matches: higher scores represented pairs of records that were more similar to each other. I provide details on the procedure used to compute the linking score in Online Appendix section A.1. The linking algorithm is also described in detail in Abramitzky, Mill, and Pérez 2018.

I used these linking scores to inform the decision rule on which records to incorporate to the analysis. To be considered a unique match for an individual in the 1869 census, a record in the 1895 census had to satisfy three conditions: (1) being the record with the highest linking score among all the potential matches for that individual, (2) having a linking score above a threshold ($p_1 > \underline{p}$) and (3) having a linking score sufficiently higher than the second highest linking score ($\frac{p_1}{p_2} > l$).

Table A.2 shows the matching rates, disaggregated by an individual age group and own or father's place of birth. I uniquely linked approximately 11 percent of sons and 10 percent of working-age individuals.[12] In section A.2 in the Online Appendix, I provide details on the reasons for match failure, as well as a comparison of the matching rates in this paper to those in other economic history papers linking US censuses.[13]

While the census manuscripts are available online in familysearch.org, the digitized information includes only the name, age and place of birth for each individual. Hence, after completing the linking procedure, I manually digitized the economic outcome variables available in each of the censuses. In the case of working-age immigrants, children of immigrants and children of natives, I digitized the economic outcome variables for every

[12]The baseline results are based on a sample created using a relatively conservative choice of the parameters $\underline{p}$ and l. As a result, my matching rates are lower than those typically found in recent economic history papers using US census data. For instance, Abramitzky, Boustan, and Eriksson 2014 report a matching rate of 12 percent when linking the 1900 to both the 1910 and 1920 US censuses. Feigenbaum 2016 reports matching rates above 50 percent when linking the Bureau of Labor Statistics sample to the 1940 census. Mill and Stein 2012. which is the closest paper to mine in terms of linking strategy report matching rates in the order of 10 percent.

[13]Mortality in the intercensal period is the main reason for match failure. Other reasons included census underenumeration and mistakes in the reported identifying information that are too severe to be accommodated by the linking strategy.

individual in the linked sample. In the case of working-age natives, I digitized the economic outcomes only for a random sample of the linked individuals. The final sample includes about 6,000 working-age natives, 5,000 working-age immigrants, 18,000 sons of natives and 2,500 native-born sons of immigrants.[14]

2.2 Comparing the Linked Sample to the Population

The likelihood of uniquely linking an individual depends both on the commonness of his name and on how accurately his name was recorded and transcribed. The dependence on names could lead to a biased sample if having a name that is both uncommon and accurately recorded is correlated with social and economic characteristics.

Sample of Sons. I use the 1869 census full count of sons aged 0 to 16 years old to estimate a probit model of the probability of being uniquely linked to an observation in the 1895 census. I consider four groups of variables: (1) name characteristics, (2) proxies for enumerator quality, (3) demographic characteristics, and (4) measures of railroad access. Table 1 shows the marginal effects of the probit model. In the even columns, I control for province of birth fixed effects – or country of birth in the case of the foreign born. These controls account for the fact that matching rates could be mechanically correlated with the size of a province of birth because, for a given distribution of names, individuals born in smaller provinces are more likely to be uniquely linked.[15]

I first assess the association between name characteristics and linkage probability. I measure first and last name commonness as the fraction of individuals with a given first or last name in the 1869 cross-section. Not surprisingly, individuals with more common first and last names are less likely to be uniquely matched.

If certain enumerators were less careful in recording the information, then individuals

[14]Note that for most of the analysis I restrict the sample to individuals outside of the province of Buenos Aires and outside of province capitals. As a result, the baseline sample that I use for the empirical analysis is smaller both for sons and for adults.

[15]An additional reason why the within-province results are relevant is that the analysis below conditions on province fixed effects.

whose information was collected by those enumerators will be less likely to be found in 1895. To assess this possibility, I proxy for enumerator quality using the leave-one-out matching rate of each individual census enumerator. I find that enumerator quality is a strong predictor of matching status.

Table 1 also shows that there are a number of individual-level demographic characteristics that predict a successful match. Older individuals and those with more siblings are more likely to be matched.[16] In addition, individuals with older fathers and fathers born abroad are more likely to be linked. Finally, urban status in 1869 is also positively associated with linking.

Despite sons in the linked sample differ from sons in the population in these dimensions, in all cases the marginal effects are small relative to the baseline linkage probability. For instance, being one year older in 1869 increases the probability of matching by 0.2 percentage points, relative to a baseline linkage probability of about 11 percent. Similarly, one additional sibling is associated with an increase of 0.2 percentage points in the probability of linking, and urban status increases the probability of linkage by 0.9-1.1 percentage points. In section 6, I show that the results are similar when I re-weight the sample to account for selection on observable characteristics.

Important to the analysis, I find a small association between matching probability and proximity to the railroad network of an individual's place of residence in 1869. I measure railroad proximity as the log distance from the centroid of an individual district of residence in 1869 to the closest operating railroad line. The estimated marginal effects are small regardless of the year in which I measure proximity – column 3. Moreover, the marginal effects are close to zero after I include province of birth fixed effects in the regression – column 4. The point estimates are similarly small after I simultaneously include the full set of covariates in the regression – columns 5 and 6.

Finally, in figure 1, I show the bivariate relationship between the linkage probability

[16] This finding likely reflects the high-mortality rate among individuals below the age of three in the sample.

and proximity to railroads, by year of construction. I present binned scatterplots of the distance to the closest railroad – x-axis – on the probability of a successful match – y-axis – , controlling for province/country of birth fixed effects. The figure shows a nearly flat relationship between the linkage probability and distance to the closest railroad line. The highest absolute value of the slope is -0.0034, when I measure railroad access in 1870. This slope implies that doubling the distance to the railroad network decreases the probability of a match by 0.24 percentage points, relative to a matching rate of about 11 percent. In section 6, I use the approach in Lee 2009 to show that the results are robust to this differential attrition.

Sample of Adults. In table A.1 and figure A.6 in the Online Appendix, I repeat the analysis above but focusing on the sample of adults. The results show a similar pattern. There are some relatively small differences in demographic characteristics with respect to the cross-section. Importantly, there is also a close to zero correlation between distance to railroad lines and the probability of a match.

2.3 Creating a Railroad Network Database

To create year-by-year digitized maps of the historical railroad network, I began from a geo-referenced map of the modern railroad network of Argentina constructed by the Argentine National Institute of Geography.[17] Then, I erased railroad segments built after 1895 by overlaying the map of the modern network with maps of the historical network from Randle 1981. Having built a digitized map of the 1895 network, I coded the opening year of each segment based on the official information in Ferrocarriles Nacionales 1896. Finally, I checked the consistency of the digitized network by collecting data on the geographic coordinates and opening date of each train stop in operation in 1895. I discuss further details on the procedure used to geo-reference the data in Online Appendix section A.4.

[17] The shape files are available for download from the following website: http://www.ign.gob.ar/sig.

2.4 Measuring Labor Market Outcomes

Occupational Categories. I classified the more than 100 different occupational titles in the sample into occupational categories using the Historical International Social Class Scheme (HISCLASS) (Leeuwen, Maas, and Miles 2002). One occupational category of special interest for the analysis is employment in farming. I classified individuals as employed in farming if they reported farmer or rural laborer as their occupation. Table A.5 shows the list of original – in Spanish – occupational titles that are classified as belonging to the farming sector. The two most common farming occupations are those of *labradores* and *jornaleros*. The former translates as "those who cultivate the land" and the latter corresponds to rural day laborers. *Labradores* typically worked a small plot of land that they owned, did not hire outside labor and produced primarily for self-consumption. *Jornaleros* were landless individuals who worked for a wage under short-term contracts.

3 Empirical Strategy

In the baseline strategy, I estimate the following equation:

$$y_{ijp} = \alpha_p + \beta Connected_{pj} + \gamma X_{ijp} + \epsilon_{ijp} \tag{1}$$

where y_{ijp} is an outcome of individual i in district j in province p, α_p is a province fixed effect – as determined by place of residence in 1869 – and X_{ijp} is a vector of individual-level controls and district characteristics described in detail below. In the baseline specification, $Connected_{pj}$ is an indicator that takes a value of one if an individual resided in 1869 in a district that would be connected to the railroad network by 1885. I defined a district as being connected in year t if its centroid was within 10 kilometers of a railroad line that was functioning in that year.[18] By 1869, only 1.5 percent of the individuals residing

[18]The distance buffer was introduced to accommodate two factors. First, train stops might have been away from the district centroid. Second, there are likely some imprecisions in the exact position of both the railroad lines and the districts, which the distance buffer helps to accommodate. All the results are

outside province capitals lived in a district connected to the railroad network. As a result, $Connected_{pj}$ captures the extent to which an individual's district of residence in 1869 *gained* access to the network during the 1869 to 1885 period. In the robustness section, I report results using the distance to the closest railroad line as the variable of interest. Throughout the paper, I cluster the standard errors at the level of the district of residence in 1869.[19]

Estimating equation 1 by OLS would imply the assumption that, conditional on X_{ijp}, railroads were randomly assigned within provinces. To test the plausibility of this assumption, in table 2 I compare individuals who in 1869 resided in districts that would be connected to the railroad network by 1885 to individuals who resided in districts that would not. In panel (a), I focus on the sample of sons, whereas in panel (b) I focus on the sample of adults. In all cases, I focus on individuals residing outside of province capitals, as these are the individuals that I include in the regressions for reasons discussed in detail below. In column 1, I report the sample mean for each of the background characteristics. In columns 2 and 3, I report the means among individuals who resided in connected and unconnected districts, respectively. In column 4, I report the difference in the means across the two groups.

Table 2 shows that railroads appeared to target districts in which individuals were slightly older, had fewer brothers and were less likely to live in urban areas. In contrast, the occupational structure was relatively similar in both groups of districts. While I am able to directly control for these background characteristics in the regressions, there is still the concern that individuals in connected districts might have differed in terms of other unobservable characteristics correlated with occupational outcomes.

robust to the choice of the radius of the distance buffer.

[19] One limitation of the data is that I observe an individuals' place of residence both in 1869 and 1895, but I have no information on his place of residence in the intercensal period. Because individuals moved even in the absence of railroads, the later in time a district got connected the less likely that the individual will still be residing there by the time that the train arrived.

3.1 Euclidean Network IV

I instrument railroad access with access to a hypothetical optimal – "Euclidean" – network. As a byproduct of connecting province capitals to each other, some intermediate districts not explicitly targeted had to be connected to the railroad network. This strategy exploits that, among these districts, those located along a convenient route – from a cost minimizing perspective – were more likely to be connected. This strategy is similar to that used in Chandra and Thompson 2000; Banerjee, Duflo, and Qian 2012; Faber 2014; Michaels 2008; Morten and Oliveira 2015.[20]

To build the Euclidean network, I implemented a procedure that closely follows Faber 2014. I started from a GIS file with the coordinates of all province capitals of 1869 Argentina, plus the city of Rosario. I then found the shortest network connecting all these cities on a continuous graph. This network is the one that a social planner would have built if her only objective was to connect all the targeted cities while minimizing the total length of the network.Figure 2 shows the actual (in red) and the predicted (in black) networks. The black circles represent the set of targeted cities.[21]

Identification Assumption. The baseline identification assumption is that, conditional on province fixed effects and distance to the nearest targeted city, location along the optimal Euclidean network influences economic outcomes only through its effect on railroad placement.

Threats to Identification. Targeted cities were connected both by the Euclidean and by the actual network. Hence, if residing in a targeted city had a direct impact on economic outcomes, the exclusion restriction would be invalid. To address this concern, I restrict the sample to individuals who in 1869 lived outside of the departments containing targeted cities.

[20] Redding and Turner 2014 refers to this strategy as the "inconsequential units" approach.

[21] On top of the Euclidean network, Faber 2014 also creates a network that takes into account differences in terrain, favoring routes with less slope and less water coverage. I chose to focus on the results based on the Euclidean network for two reasons. First, it is unclear how to exactly weight factors other than distance in the least cost calculations. Second, terrain might have an independent effect on geographic mobility, thus making the exclusion restriction less likely to hold.

In addition, because districts closer to targeted cities are mechanically more likely to lie along the Euclidean network, the exclusion restriction would be invalid if proximity to targeted cities had an independent effect on economic outcomes. To address this concern, all the regressions include the distance to the nearest targeted city as an additional control variable.

I discuss further concerns with the exclusion restriction in section 6. In particular, I discuss the possibility that location along the Euclidean network is correlated with historical trade routes and show that my results are robust to controlling for proximity tof these routes.

4 The Impact of Railroads on Occupational Outcomes

4.1 Railroads and the Transition Out of Farming Occupations

In figure 3, I plot the fraction of individuals employed in farming in connected and unconnected districts, by census year. In panel (a), I include the full sample of districts, whereas in panels (b) and (c) I restrict the sample to districts with below and above median agricultural suitability, respectively. To do so, I use the index of agricultural potential of the National Institute of Agricultural Technology of Argentina (Cruzate et al. 2012).[22] The figure shows that, by 1869 (before railroad construction), the fraction of individuals employed as either farmers or farm workers was relatively similar in districts that would be connected to the railroad network by 1885 than in districts that would not. By 1895, the fraction of workers in farming was lower in connected districts when focusing on the full sample, but no such differential decline is apparent for districts with high agricultural suitability.

Fixed Effects Regressions. Despite the longitudinal nature of my data, I observe whether an individual transitioned out of farming only once. However, it is possible

[22] Figure B.1 shows the spatial distribution of agricultural suitability: darker colors correspond to areas with higher agricultural potential.

to estimate the following model of occupational choice:

$$\mathbb{1}(Occupation = Farming)_{it} = \alpha_i + \alpha_t + \beta Connected_{it} + \gamma X_{it} + \epsilon_{it} \tag{2}$$

where $\mathbb{1}(Occupation = Farming)_{it}$ is an indicator that takes a value of one if an individual in district i worked in farming in year t, α_i is a district fixed effect and α_t is a census year fixed effect. This specification controls for fixed district characteristics correlated with the propensity of farming employment in the district. To control for differential trends across regions of the country in the decline of farming employment, I also report specifications in which I control for an interaction between the 1895 census indicator and region fixed effects. In all cases, I assign individuals to districts based on their 1869 location.

Panel (a) in table B.4 shows the results of this specification. Consistent with the previous models, the results also show a decline in the probability of farming employment. Finally, in panel (b) I present results in which I combine the district fixed effects and the IV models. In particular, the IV takes a value of zero for every district in 1869, and a value of one for districts that are connected to the Euclidean network in 1895. These results also show a decline in the probability of farming employment.

To more formally asses the relationship between railroads and the probability of exiting farming. I estimate:

$$\begin{aligned}\mathbb{1}(Not\,in\,Farming^{1895} | Father\,in\,Farming^{1869})_{ijp} = \\ \alpha_p + \beta Connected_{pj} + \gamma X_{ijp} + \epsilon_{ijp}\end{aligned} \tag{3}$$

where i indexes the father-son pair, j indexes the district and p indexes the province. In all cases, I control for a quartic in son's and father's age, as well as for province of residence fixed effects – α_p – and distance to the nearest targeted city (based on place of residence in 1869). In the even columns, I also control for a vector of household characteristics: the number of siblings in the household, whether the father was literate

and whether the father was foreign born. Because in this section of the analysis I am interested in *exit* out of farming, I restrict the sample to father-son pairs in which the father was employed in this sector in 1869.

Panel (a) of table 3 reports the results of this estimation. The OLS results show that sons in connected districts whose fathers were employed in farming were about 5 percentage points more likely to exit these occupations. The IV results suggest a larger increase in the probability of exiting farming: sons in connected districts were 13 percentage points more likely to exit farming, relative to a baseline probability of 31 percent. The first stage F-statistic indicates that the placement of the Euclidean network is a strong predictor of the actual network. In both the OLS and IV estimates, the results exhibit little sensitivity to controlling for household background characteristics in 1869.

In panel (b) of table 3, I focus instead on the probability that farmers and farm workers would themselves exit these occupations. To do so, I focus on adults – aged 18 to 35 years old in 1869 – employed as farmers or farm workers in 1869. In contrast to the results for children, I find that there is no increase in the probability that adults would exit farming in connected districts. Both the OLS and the IV point estimates are much smaller than those of the children's generation, and not statistically different from zero.

Why are the IV Estimates Larger than the OLS Estimates? First, the IV estimates identify a local average treatment effect among the set of compliers. In this case, the compliers are individuals residing in districts that were connected to the railroad network because of their location along a convenient route, but would not have been connected otherwise. In particular, districts with low agricultural suitability are overrepresented among compliers. Below, I show that the effects were indeed on average stronger in these districts.[23]

Second, the IV estimates correct for classical measurement error in the measure of

[23] I estimate the set of compliers to correspond to approximately 30 percent of the total number of individuals in connected districts.This proportion is estimated as the first stage, times the ratio between the fraction of individuals connected to the Euclidean network to the fraction of individuals connected to the actual network (Angrist and Pischke 2008).

railroad access, which will tend to increase the absolute value of the IV estimates relative to OLS. Third, OLS estimates could be biased due to the non-random placement of railroads. The direction of this bias is ex-ante unclear. On the one hand, railroad construction might have targeted locations whose residents had high ex-ante propensities for occupational mobility. On the other hand, railroads might have targeted economically struggling districts, where the potential for transitioning out of farming might have been more limited.

Occupational Transitions. I next use the more detailed occupational categories data to answer two questions: (1) Into which occupations did sons of farmers and farm workers transition? (2) Were sons of workers outside of farming less likely to enter these occupations? I estimate:

$$\mathbb{1}(Son\,Occupation = k/Father\,Occupation = m)_{ijp} = \alpha_p + \beta Connected_{pj} + \gamma X_{ij} + \epsilon_{ijp} \tag{4}$$

where $\mathbb{1}(Son\,Occupation = k/Father\,Occupation = m)_{ijp}$ is an indicator that takes a value of one if a son with a father in occupational category m in 1869 worked in occupational category k in 1895. As before, i indexes the father-son pair, j indexes the district and p indexes the province. The set of control variables is the same as in the previous specification. I divide occupations into four mutually exclusive categories: white-collar, farmer and farm workers, skilled/semi-skilled and unskilled urban workers. Table A.6 shows the ten most common occupations for fathers (in 1869) and for sons (in 1895), and their corresponding occupational category.

Table 4 shows the results of estimating equation 4 by OLS. I focus on the OLS results in this subsection as the IV estimates when looking at these detailed transitions are very noisy. Rows in this table represent the occupation of the father in 1869, whereas columns represent the occupation of the son in 1895. Each cell in the table corresponds to the value of β for each subsample – defined by parental occupation – and for each occupational outcome. The last row of the table shows the results for the full sample of sons. Positive

values of β imply that a given transition was more likely in connected districts, and vice versa. In panel (a) of table B.2, I report the baseline father-son occupational transitions in this sample.

First, regardless of father's occupation and in line with the results described above, railroads were associated with a lower probability of sons working in farming in adulthood. On average, sons in connected districts were approximately 6 percentage points less likely to be farmers or farm workers in adulthood – last row in panel (a) of table 4. On the other hand, railroads were associated with an increase in the propensity of performing white-collar and skilled/semi-skilled blue-collar jobs.

The overall reduction in the propensity for farming in the sons' generation was driven not only by a higher probability of exiting farming, but also by a lower probability of entering it. Sons of white-collar workers in connected districts were about 9 percentage points more likely to be white-collar workers themselves. This increase came mostly at the expense of a reduction in the prevalence of farming. There is a similar pattern among the sons of skilled/semi-skilled workers.

In panel (b) of table 4, I re-estimate a version of equation 4 but focusing instead on *intragenerational* occupational mobility. First, similar to the results on children, I find that railroads are also associated with a lower overall probability of working in farming in 1895. However, this decline is just driven by a decreased likelihood of *entering* farming, with no change in the probability of *exiting* the sector. This evidence suggests the existence of some friction preventing adults employed in agriculture from exiting this sector.

5 Mechanisms

5.1 How Did Children Transition Out of Farming Occupations?

First, railroads might have facilitated occupational mobility by facilitating physical mobility. To test this possibility, I estimate:

$$Mover_{ijp} = \alpha_p + \beta Connected_{pj} + \gamma X_{ij} + \epsilon_{ijp} \tag{5}$$

where $Mover_{ijp}$ is an indicator that takes a value of one if an individual resided 200 miles away or more from his 1869 district, and the rest of the variables are defined as in the previous section.

Panel (a) of table **??** shows the results of this regression in the sample of sons. In columns 1 and 3, I report the OLS and IV results controlling only for province of residence fixed effects, distance to the nearest targeted city and a quartic in sons' age. In columns 2 and 4, I control for a vector of individual-level background characteristics: the number of siblings, the occupational category of the father, urban status of the family, whether the father was literate and whether the father was foreign born.

Sons in connected districts had a substantially higher probability of moving long-distance. The OLS estimates suggest a difference of approximately 4 percentage points, with little difference depending on whether I control for background characteristics or not. The IV estimates are of similar size but not precisely estimated. These results suggest a large influence of railroads on migration propensities: just about 8.5 percent of the sons in the sample resided 200 miles or more away from their 1869 district of residence.

The large impact of railroad availability on migration propensities might result surprising given that the cost of migrating was perhaps small relative to the potential lifetime gains from moving.[24] A number of factors might explain the large impact. First, railroads

[24]Large responses to relatively small interventions have been documented elsewhere in the context of migration. Bryan, Chowdhury, and Mobarak 2014 find strong responses to an intervention subsidizing the cost of bus tickets for seasonal migrants in Bangladesh. Black et al. 2015 document that individuals

dramatically reduced the cost of traveling internally, both in terms of reduced travel times and in terms of increased safety. This reduction in the cost might have had an amplified effect if liquidity constraints prevented individuals from moving in the pre-railroads era. Second, railroads also likely improved information flows among connected districts, increasing individuals' awareness about opportunities outside of the local economy. Third, railroads also made it easier for individuals to return home when needed.

Decomposing the Transition Out of Farming Occupations Between Movers and Stayers. To quantify the role of geographic mobility in explaining the results of the previous section, I next present a decomposition exercise to measure the proportion of the overall shift away from farming that can be attributed to: (1) those who stayed in their 1869 district of residence, and (2) those who left their 1869 district of residence.

To do so, I define two indicator variables. The first takes a value of one if an individual exited farming but stayed in his 1869 location. The second takes a value of one if the individual both exited farming and left his 1869 location. I then make use of the fact that:

$$Pr(Out\,of\,Farming) = Pr(Out\,Farming \cap Stay) + Pr(Out\,Farming \cap Leave) \quad (6)$$

and estimate:

$$\mathbb{1}(Out\,Farming \cap Stay) = \alpha_p + \beta_s Connected_{pj} + \gamma X_{ij} + \epsilon_{ijp} \quad (7)$$

and:

$$\mathbb{1}(Out\,Farming \cap Leave) = \alpha_p + \beta_l Connected_{pj} + \gamma X_{ij} + \epsilon_{ijp} \quad (8)$$

born in railroad towns were 6 percentage points more likely to move from the US South to the North during the Great Migration.

The fraction of the overall shift away from farming that is driven by movers is given by:

$$\frac{\beta_l}{\beta_l + \beta_s} \tag{9}$$

Panel (a) of table **??** shows the results of this exercise. In columns 1 and 2, the outcome variable is an indicator that takes a value of one if the son was employed in a non-farming occupation in adulthood but stayed in his 1869 district of residence. In columns 3 and 4, the outcome variable is an indicator that takes a value of one if the son was employed in a non-farming occupation in adulthood and had left his 1869 district of residence. The OLS estimates indicate that 39 percent[25] of the transition out of farming occupations was driven by sons who physically moved, whereas the rest was explained by those who stayed. The IV estimates also suggest that an important fraction – 25 percent[26] – of the overall effect is driven by internal migrants.

These results lend support to the hypotheses that migration was an intermediate channel through which railroads led to mobility out of farming occupations in the sons' generation. An important implication of these findings is that focusing on the sample of *stayers* when analyzing the impacts of transport infrastructure on individual level outcomes could be problematic if one way that transport infrastructure helps individuals is by facilitating geographic mobility. In this case, focusing on stayers alone would miss 25 to 39 percent of the overall shift out of farming occupations.

At the same time, most of the decline of farming in the sons' generation is explained by sons who stayed in their 1869 districts of residence. This finding suggests that railroads resulted in a reduction of the local demand for agricultural labor in connected districts. Why did this reduction take place? One possibility is that the concentration of economic activity around train stations resulted in higher urbanization, thus increasing the relative demand for non-agricultural occupations. To test this possibility I estimate:

[25] $\frac{0.0176}{0.0274 + 0.0176}$

[26] $\frac{0.0336}{0.10 + 0.0336}$

$$Urban_{ijp} = \alpha_p + \beta Connected_{pj} + \gamma X_{ij} + \epsilon_{ijp} \tag{10}$$

where $Urban_{it}$ takes a value of one if an individual lived in an urban area in year 1895, and X_{ij} includes the same variables as above (including urban status in 1869). Indeed, table 5 shows that individuals who lived in districts that were connected to the railroad network became more likely to live in an urban area in 1895.

As railroads connected districts with different degrees of agricultural productivity, railroads might have allowed districts with relatively low productivity in farming to buy agricultural products from other districts. Figure B.2 shows evidence consistent with limited specialization across departments in 1869 Argentina – the pre-railroad era. The figure shows a binned scatterplot of the proportion of the adult population employed in farming occupations (y-axis) on an index of agricultural suitability (x-axis): departments with low agricultural potential had a *higher* proportion of the working force employed in farming.

Panel (b) of table **??** shows that the shift away from farming in the children's generation was largely driven by sons who in 1869 resided in departments with low agricultural suitability. In this table, I stratify the sample based on suitability for agriculture. Sons in districts with low agricultural suitability were about 20 percentage points more likely to exit farming, whereas those in high suitability districts were at most 5 percentage points more likely. Note that the average probability of exiting farming occupations is similar across the two groups – 32 percent in districts with low suitability versus 30 percent in districts with high suitability. These results indicate that the transition out of farming took place mainly in districts where farming was relatively unproductive. Note that, consistent with the low agricultural suitability districts being overrepresented among compliers, the OLS and IV results are very close to each other in this sample. The overrepresentation of these districts stems from the fact that districts with low agricultural suitability were connected to the network *only* when they were located along a convenient route. Also

note that the instrument has weak predictive power in the high-agricultural suitability sample.

Table B.4 shows that this pattern of heterogeneity is also present when estimating fixed effects models of occupational choice. Using this alternative specification, I also find that the effects are close to zero and not statistically significant in the sample of districts with high suitability for agriculture, and large and significant in the sample with low agricultural suitability.

5.2 Why Did the Transition Out of Farming Occupations Occur in the Children's Generation?

The effects on adults might have differed from the effects on children for two main reasons. First, imperfections in the land market might have prevented individuals who worked their own land from exiting farming. Second, adults likely had a higher opportunity cost than children of acquiring the skills required to work outside of farming.

Two pieces of evidence suggest that imperfections in the land market were unlikely to be the reason for the limited response among adults. First, panel (b) of table **??** shows that railroads led to increased migration propensities also in the sample of adults, thus suggesting that adults were not stuck to their land. However, there was a relatively weaker association between long-distance migration and mobility out of farming occupations among adults: adults who migrated long-distance stayed in agricultural occupations to a greater extent than their children.

Second, table B.1 shows that the results for adults are similar for those who were likely landowners and for those who were likely landless. Since the 1869 census lacks direct information on ownership status, I use the information contained in occupational titles, in combination with the data on property ownership by occupational title from the 1895 census. In particular, I compute a "property index" based on the average ownership rate for each occupational title in 1895. For instance, 16 percent of rural laborers – *jornaleros*

– owned property in 1895, whereas 76 percent of *estancieros* did. I then stratify the sample of farmers and farm workers into those with occupational titles with below and above-median values on the property index. The effects are close to zero and statistically insignificant in both groups.

If jobs outside of farming required higher skills, the shift away from farming occupations might have required investing in human capital. The data indeed show that workers in farming had lower skills, as proxied by literacy rates. Figure B.3 shows literacy rates by occupational category and by census year. By 1869, white-collar workers had a literacy rate that was close to 80 percent, whereas the literacy rate for farmers was slightly above 30 percent.[27]

Did families in connected districts invest more in the human capital of their children? In the linked sample, I observe information on literacy rates both for fathers – in 1869 – and for sons – in 1895.[28] Hence, I am able to test whether literacy rates changed as a result of access to railroads. I estimate:

$$Literacy_{ijp} = \alpha_p + \beta Connected_{pj} + \gamma X_{ij} + \epsilon_{ijp} \tag{11}$$

where $Literacy_{ijp}$ is the literacy of the son in father-son pair i – measured in 1895 – and the rest of the covariates are defined as above. Table 6 shows the results of estimating equation 11. In columns 1 and 2, I report the OLS estimates and in column 3 and 4 I report the IV estimates. In the even columns, I control for a vector of household background characteristics, including a set of indicators of a father's occupational category and whether the father himself was literate.

The results show that sons in connected districts were more likely to be literate in adulthood. The OLS estimates imply that railroads were associated with a 4 percentage

[27] Literacy was still far from universal in late 19th-century Argentina, especially in the interior provinces. Among native-born males aged 18 years old or more in 1895, the literacy rate was approximately 50 percent in the country as a whole and 43 percent excluding the province of Buenos Aires.

[28] Unlike modern censuses, early censuses did not contain information on completed years of schooling.

point increase in literacy rates, relative to an average of approximately 48 percent. The IV estimates suggest a similar increase, although they are not precisely estimated.

5.3 Spillover effects

.

As discussed in Donaldson and Hornbeck 2015, railroads likely had aggregate effects. My identification strategy can only identify relative effects.

Changes in market access.

As an alternative exercise, , I rerun my main specification on a sample that excludes individuals in unconnected districts that were nearby connected districts. Intuitively, spillover effects should be smaller for districts that were far from connected districts. Specifically, I progressively exclude individuals located at 50, 100, 200 and 400 kilometers of connected districts Figure B.4 shows that the estimates are actually larger when excluding unconnected districts that were close to connected ones. This evidence is consistent with spillover effects biasing the estimated effects downwards.

6 Alternative Specifications and Robustness

I show that the main results are robust to: (1) the procedure used to create the linked sample, (2) accounting for differential attrition, (3) controlling for the location of historical trade routes, (4) accounting for spillover effects, (5) excluding one province at a time, and (6) using alternative definitions of the treatment variable.

Linking Procedure. There are two main concerns related to the linking procedure. First, as discussed in section 2.2, the linked samples are not fully representative of the population. To alleviate this concern, in the second row of each of the panels in figure 4 I show that the results are similar when I re-weight the sample to account for selection into the linked sample based on observable characteristics.[29] The point estimates are close

[29] To compute the sample weights, I used the estimates from table 1 to compute an estimated linkage

to those obtained with the unweighted sample, suggesting that selection on observable characteristics into the linked sample does not drive the results.

Second, some of the links might be incorrect. These errors would be particularly problematic in the context of this paper If the fraction of incorrect matches was correlated with railroad availability, as I would then mechanically find a correlation between occupational and geographic mobility and access to railroads. I perform two separate exercises to alleviate this concern. In both exercises, I adopt a more conservative linking strategy only in *connected* districts, so as to bias the sample towards less spurious mobility in connected districts. First, I replicate the analysis but focusing on the sample of individuals in connected districts who match perfectly in terms of their identifying information, while keeping the rest of the sample.[30] Second, as false positives are likely more prevalent among individuals with common names, I drop from the sample all individuals in connected districts with first names in the top 25 percent in terms of first name commonness within their province of birth. In all cases, I find that the results in this alternative samples are relatively similar (but less precisely estimated reflecting the smaller sample size) to the baseline.

Differential Attrition. Section 2.2 shows that there is a small correlation between measures of railroad access and the probability of successfully linking an individual. However, note that in most of the analysis my independent variable of interest is an indicator that takes a value of one when a district is connected to the railroad network. Among individuals in the baseline father-sons sample, the linkage probability is equal to 11.9% for those initially living in unconnected districts and 11.7% for those in connected districts. Given this differential attrition, It is possible to construct bounds for the effects using the approach in Lee 2009. This approach requires trimming the distribution of outcomes for

probability for each observation in the 1869 census cross-section. I then re-weighted the sample by the inverse of this linkage probability.

[30] I define a perfect match as one in which both the first name and the last name agree perfectly, but I allow the year of birth to differ by at most one year. Because the two censuses took place in different moments of the year, the difference in estimated year of birth could be one, even if an individual accurately reported his age in both censuses.

the treatment group by the difference in attrition between the treatment and the control groups. Figure B.5 shows the results of this exercise, where the outcome is the probability of working outside of farming among children of farmers. Not surprisingly given the small differences in attrition, both the upper and the lower bounds are very close to the baseline result.

Historical Trade Routes. One concern with the IV strategy is that location along the Euclidean network might be correlated with historical trade routes. This correlation could arise if regional trade among province capitals took place along shortest routes. If location along trade routes had a direct influence on geographic and occupational mobility, then the IV strategy would violate the exclusion restriction. The main trade route during the colonial period was "El Camino Real", connecting the cities of Buenos Aires and Lima, Peru. To address the concern that trade routes might have had an independent effect on mobility, I include proximity to "El Camino Real" as an additional control variable. The last row in each of the panels in figure 4 shows that the point estimates remain similar after the inclusion of this control variable. In addition, note that for the effects of historical trade routes to explain my results, it would need to be the case that historical trades routes affected occupational outcomes only in districts with low agricultural suitability and only among relatively young individuals.

Excluding One Province at a Time. In figure B.7, I show the point estimates and confidence intervals for each of the main specifications in the paper after excluding one province at a time based on a family's place of residence in 1869. This figure suggests that the effects are not driven by any single province.

Measures of Railroad Access. In table B.3 I replace the baseline measure of railroad access – an indicator that takes a value of one if the location was connected to the network – with a continuous measure of railroad exposure: the distance to the closest operating line. The results show the same pattern as in the baseline analysis: sons in districts closer to railroad lines were less likely to work in agriculture in adulthood, but no such effects

are present for adults.

7 Conclusion

Mobility out of farming and rural areas is one of the defining characteristics of modern economic growth. In this paper, I studied how improvements in transportation infrastructure can shape this transition. To do so, I exploited the expansion of the railroad network in 19th-century Argentina and newly-assembled individual-level longitudinal data.

I documented that adults who were employed in farming before railroad expansion were not more likely to transition out of these occupations once the train arrived to their district. By contrast, the children of these workers exited the agricultural sector and transitioned into the modern sector of the economy at higher rates. These children were less likely to enter farming occupations in adulthood not only because they moved physically, but also because railroads changed the nature of local economic activity: connected districts, in particular those with lower agricultural suitability, became more urbanized after being connected to the railroad network. The difference between adults and children suggests that the transition away from predominantly farming jobs took more than one generation: transitioning out of these occupations might have required complementary investments in human capital.

What are the broader implications of this research? During this period, Argentina was undergoing a process of structural change of similar intensity to that being observed in some developing countries today. Urbanization rates grew from 28.6 percent in 1869 to 37.4 percent in 1895 – a level that is similar to that of contemporary developing countries.[31] This paper shows that improvements in transport infrastructure were an important driver of this transition: they enabled individuals in districts with low agricultural potential to exit farming and also facilitated geographic mobility into the urban sector. However,

[31]For instance, urbanization rates were 33 percent in India and 34 percent in Vietnam in 2015. See http://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZShttp://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS.

the effects of transport infrastructure might not be immediate; if transitioning into the modern sector of the economy requires acquiring new skills, then only relatively young individuals might be able to take advantage of the increased opportunities. More broadly, the results of this paper show that human capital can be slow to adjust to technological change.

From a methodological point of view, this paper highlights the importance of longitudinal data in assessing the effects of localized interventions in settings with high population mobility. Because selection into migration is unlikely to be random, estimates based on a sample of stayers could provide an inaccurate picture.

References

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson (2012). "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration". In: *American Economic Review* 102.5, pp. 1832–1856.

— (2013). "Have the poor always been less likely to migrate? Evidence from inheritance practices during the Age of Mass Migration". In: *Journal of Development Economics* 102, pp. 2–14.

— (2014). "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration". In: *Journal of Political Economy* 122.3, pp. 467–506.

Abramitzky, Ran, Roy Mill, and Santiago Pérez (2018). *Linking Individuals Across Historical Sources: a Fully Automated Approach*. Tech. rep. National Bureau of Economic Research.

Adamopoulos, Tasso (2011). "Transportation Costs, Agricultural Productivity, And Cross-Country Income Differences". In: *International Economic Review* 52.2, pp. 489–521.

Angrist, Joshua D and Jörn-Steffen Pischke (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Asher, Sam and Paul Novosad (2016). "Market Access and Structural Transformation: Evidence from Rural Roads in India". In: *Manuscript: Department of Economics, University of Oxford*.

Atack, Jeremy, Michael Haines, and Robert A Margo (2009). "Did railroads induce or follow economic growth? Urbanization and population growth in the American Midwest, 1850-60". In: *NBER Working Paper* 14640.

— (2011). "Railroads and the Rise of the Factory: Evidence for the United States, 1850–1870". In: *Economic Evolution and Revolution in Historical Time*.

Atack, Jeremy, Robert Margo, and Elisabeth Perlman (2012). "The impact of railroads on school enrollment in nineteenth century America". In:

Bailey, Martha et al. (2017). *How Well Do Automated Linking Methods Perform in Historical Samples? Evidence from New Ground Truth*. Tech. rep. Working Paper.

Banerjee, Abhijit, Esther Duflo, and Nancy Qian (Mar. 2012). *On the Road: Access to Transportation Infrastructure and Economic Growth in China*. Working Paper 17897. National Bureau of Economic Research. DOI: 10.3386/w17897. URL: http://www.nber.org/papers/w17897.

Berger, Thor and Kerstin Enflo (2015). "Locomotives of local growth: The short-and long-term impact of railroads in Sweden". In: *Journal of Urban Economics*.

Black, Dan A et al. (2015). "The impact of the Great Migration on mortality of African Americans: Evidence from the Deep South". In: *The American Economic Review* 105.2, pp. 477–503.

Bryan, Gharad, Shyamal Chowdhury, and Ahmed Mushfiq Mobarak (2014). "Underinvestment in a profitable technology: The case of seasonal migration in Bangladesh". In: *Econometrica* 82.5, pp. 1671–1748.

Cárcano, Ramón José (1893). *Historia de los medios de communicación y transporte en la República Argentina*. Vol. 2. F. Lajouane.

Caselli, Francesco and Wilbur John Coleman II (2001). "The US structural transformation and regional convergence: A reinterpretation". In: *Journal of political Economy* 109.3, pp. 584–616.

Chandra, Amitabh and Eric Thompson (2000). "Does public infrastructure affect economic activity?: Evidence from the rural interstate highway system". In: *Regional Science and Urban Economics* 30.4, pp. 457–490.

Collins, William J and Marianne H Wanamaker (2014). "Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data". In: *American Economic Journal: Applied Economics* 6.1, pp. 220–252.

— (2015). "The Great Migration in Black and White: New Evidence on the Selection and Sorting of Southern Migrants". In: *The Journal of Economic History* 75.04, pp. 947–992.

Conde, Roberto Cortés (1979). "El progreso argentino: 1880-1914". In:

Cruzate, G et al. (2012). *Suelos de la República Argentina 1: 500000 y 1: 1000000*.

Dix-Carneiro, Rafael (2014). "Trade liberalization and labor market dynamics". In: *Econometrica* 82.3, pp. 825–885.

Donaldson, Dave (2015). "Railroads of the Raj: Estimating the impact of transportation infrastructure." In: *American Economic Review*.

Donaldson, Dave and Richard Hornbeck (2015). *Railroads and American economic growth: A "market access" approach*. Tech. rep. National Bureau of Economic Research.

Faber, Benjamin (2014). "Trade Integration, Market Size, and Industrialization: Evidence from China's National Trunk Highway System*". In: *The Review of Economic Studies*.

Fajgelbaum, Pablo and Stephen J Redding (2014). *External integration, structural transformation and economic development: Evidence from argentina 1870-1914*. Tech. rep. National Bureau of Economic Research.

Feigenbaum, James J. (2016). "Intergenerational Mobility during the Great Depression". In: *mimeo*.

Ferrocarriles Nacionales, Dirección General de (1896). *Estadıstica de los ferrocarriles en explotación*.

Foster, Andrew D and Mark R Rosenzweig (1996). "Technical change and human-capital returns and investments: evidence from the green revolution". In: *The American economic review*, pp. 931–953.

— (2007). "Economic development and the decline of agricultural employment". In: *Handbook of development economics* 4, pp. 3051–3083.

Gollin, Douglas, David Lagakos, and Michael E Waugh (2014). "The Agricultural Productivity Gap". In: *The Quarterly Journal of Economics* 129.2, pp. 939–993.

Gollin, Douglas and Richard Rogerson (2014). "Productivity, transport costs and subsistence agriculture". In: *Journal of Development Economics* 107, pp. 38–48.

Herranz-Loncán, Alfonso (2011). "El impacto directo del ferrocarril sobre el crecimiento económico argentino durante la Primera Globalización". In: *Revista Uruguaya de Historia Económica* 1.1, pp. 34–52.

— (2014). "Transport technology and economic expansion: The growth contribution of railways in Latin America before 1914". In: *Revista de Historia Económica/Journal of Iberian and Latin American Economic History (New Series)* 32.01, pp. 13–45.

Herrendorf, Berthold, Richard Rogerson, and Ákos Valentinyi (2014). "Growth and Structural Transformation". In: *Handbook of Economic Growth* 2, pp. 855–941.

Herrendorf, Berthold, James A Schmitz Jr, and Arilton Teixeira (2012). "The role of transportation in US economic development: 1840–1860". In: *International Economic Review* 53.3, pp. 693–716.

Hornung, Erik (2015). "Railroads and growth in Prussia". In: *Journal of the European Economic Association* 13.4, pp. 699–736.

Keller, Wolfgang and Carol H Shiue (2008). *Institutions, technology, and trade*. Tech. rep. National Bureau of Economic Research.

Lee, David S (2009). "Training, wages, and sample selection: Estimating sharp bounds on treatment effects". In: *The Review of Economic Studies* 76.3, pp. 1071–1102.

Leeuwen, MHD van, Ineke Maas, and Andrew Miles (2002). *HISCO: Historical international standard classification of occupations*. Leuven: Leuven University Press.

Lewis, Colin M (1983). *British railways in Argentina, 1857-1914: a case study of foreign investment*. Vol. 12. Burns & Oates.

Long, Jason and Joseph Ferrie (2013). "Intergenerational occupational mobility in Great Britain and the United States since 1850". In: *The American Economic Review* 103.4, pp. 1109–1137.

Matsuyama, Kiminori (1992). "A simple model of sectoral adjustment". In: *The Review of Economic Studies* 59.2, pp. 375–387.

Michaels, Guy (2008). "The effect of trade on the demand for skill: Evidence from the interstate highway system". In: *The Review of Economics and Statistics* 90.4, pp. 683–701.

Mill, Roy and Luke CD Stein (2012). *Race, Skin Color, and Economic Outcomes in Early Twentieth-Century America*. Tech. rep. Working Paper, Stanford University December.

Morten, Melanie and Jaqueline Oliveira (2015). "Migration, roads and labor market integration: Evidence from a planned capital city". In:

North, Douglass (1958). "Ocean Freight Rates and Economic Development 1730-1913". In: *The Journal of Economic History* 18.04, pp. 537–555.

Randle, Patricio H (1981). *Atlas del desarrollo territorial de la Argentina*. Oikos.

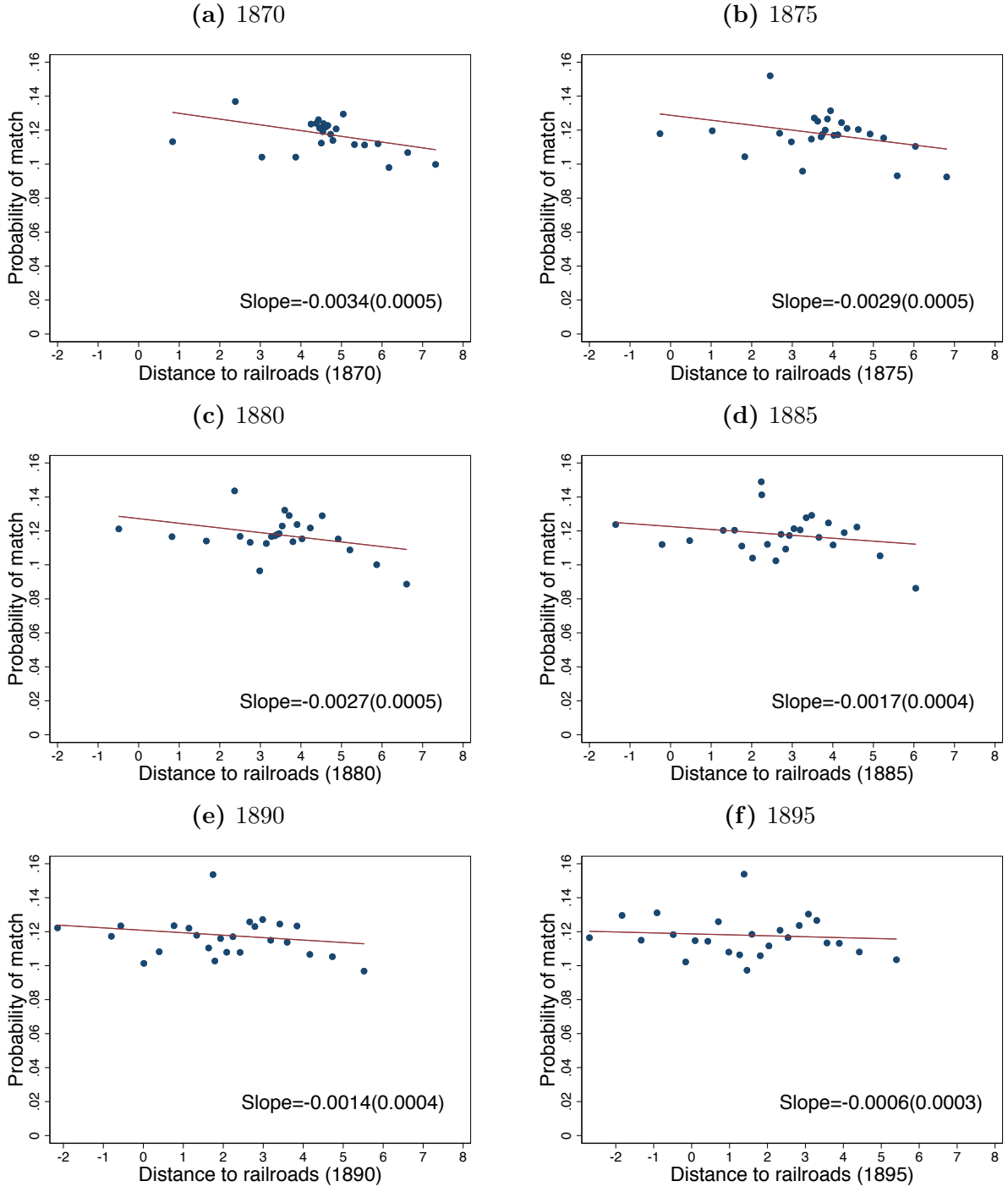
Redding, Stephen J and Matthew A Turner (2014). *Transportation costs and the spatial organization of economic activity*. Tech. rep. National Bureau of Economic Research.

Rostow, Walt Whitman (1959). "The stages of economic growth". In: *The Economic History Review* 12.1, pp. 1–16.

Walker, W Reed (2013). "The transitional costs of sectoral reallocation: Evidence from the Clean Air Act and the workforce". In: *The Quarterly journal of economics* 128.4, pp. 1787–1835.

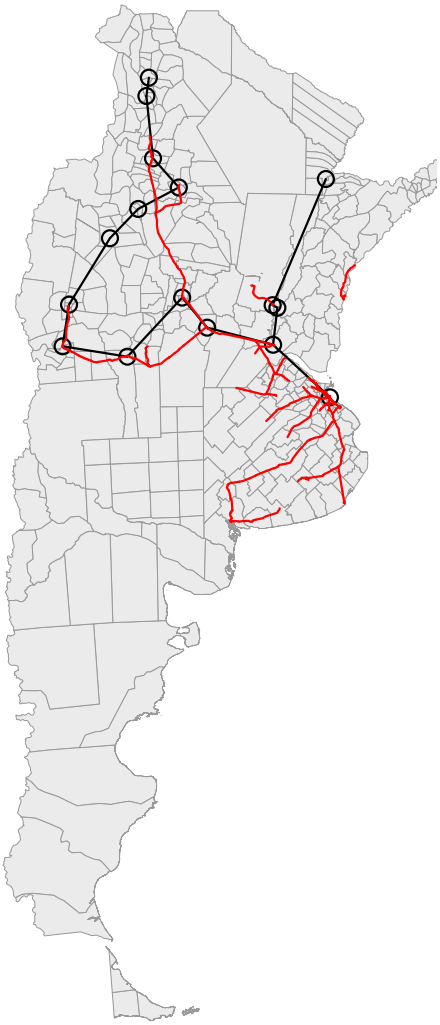
Zalduendo, Eduardo A (1975). *Libras y rieles: las inversiones británicas para el desarrollo de los ferrocarriles en Argentina, Brasil, Canadá e India durante el siglo XIX*. Editorial El Coloquio.

Figure 1: Probability of a match and distance to closest railroad line, sample of sons



Notes: This figure shows a binned scatterplot of the probability of being on the linked sample (y-axis) on the log distance to the closest railroad line (x-axis) by year of construction, net of province or country of birth fixed effects. I divide the data into 25 equally sized bins based on distance to the closest railroad line. The y-axis shows the average probability of a match within each of the bins. Distance is measured in kilometers from the district centroid. For each of the binned scatterplots, I report the slope coefficient of an OLS regression of the probability of a match on distance to the closest railroad line, controlling for province or country of birth fixed effects (standard errors in parentheses).

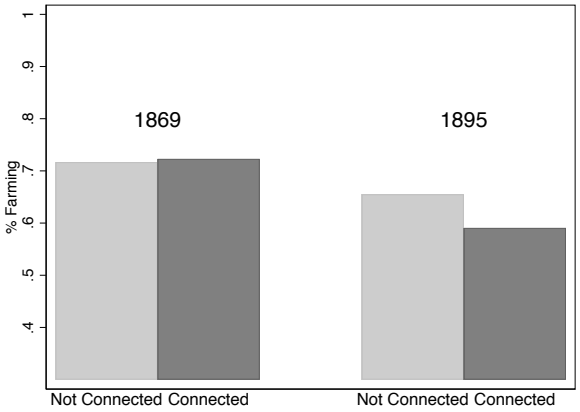
Figure 2: Euclidean Network



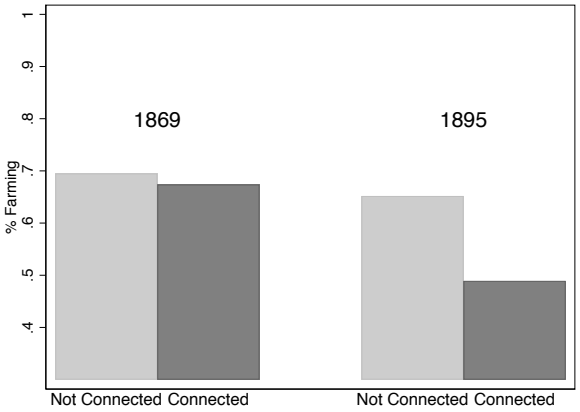
Notes: The above map depicts the Euclidean (in black) and actual (in red) railroad networks in 1885. The Euclidean network was found by constructing the shortest network connecting all province capitals of 1869 Argentina plus Rosario in a continuous graph.

Figure 3: Fraction employed in farming, connected and unconnected districts

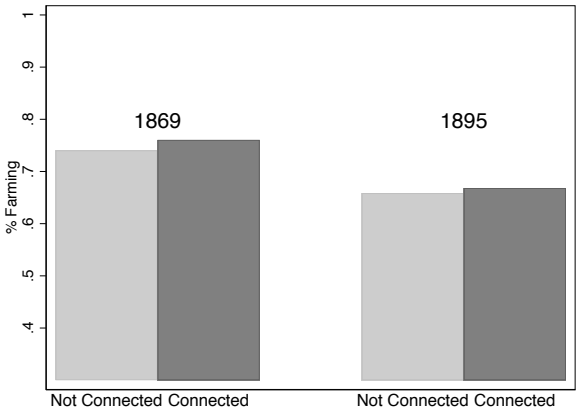
(a) % in farming



(b) % in farming (low agricultural suitability)



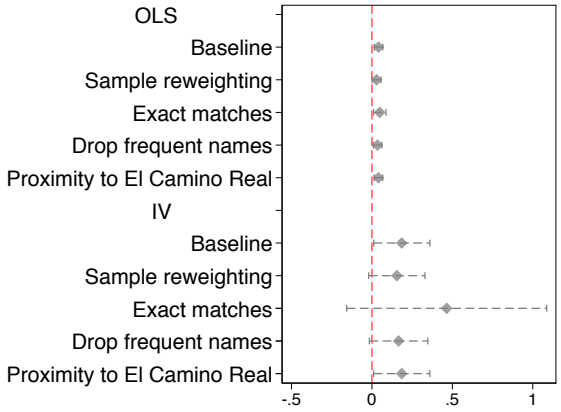
(c) % in farming (high agricultural suitability)



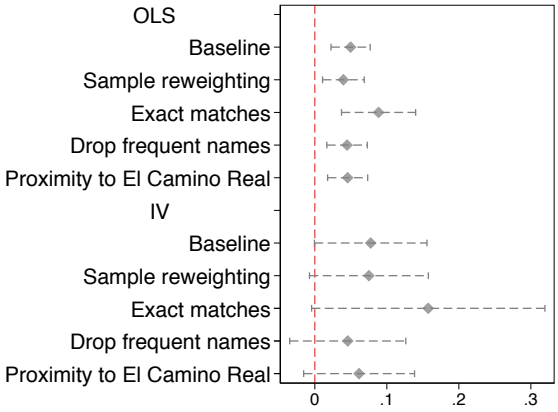
Notes: These figures show the fraction of individuals employed in farming in connected and unconnected districts, by census year. Districts are classified as connected based on railroad availability in 1885. In panel (a), I include the full sample of districts. In panel (b), I only include districts with below median agricultural suitability. In panel (c), I only include districts with above median agricultural suitability.

Figure 4: Robustness to alternative empirical specifications

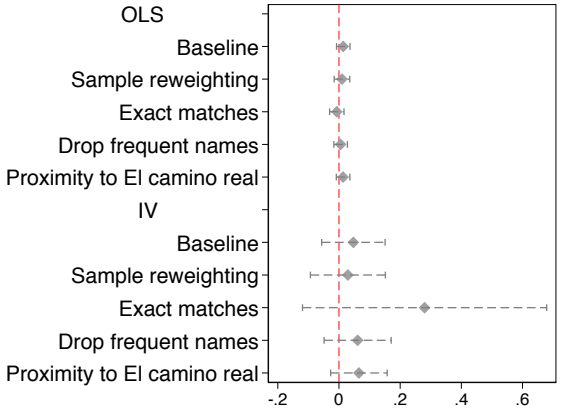
(a) Pr(Not Farmer in 1895 / Father Farmer in 1869)



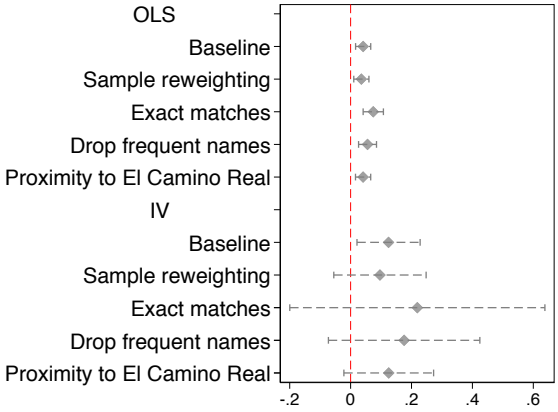
(b) $\log(Occ.\,Score_{son})$



(c) Pr(Long-distance migration)



(d) Literacy



Notes: Each of the figures shows the robustness of the results to features of the linking procedure (rows 2 to 4 in each of the figures) and to controlling for the placement of historical trade routes (row 5). Panel (a) shows the robustness of the results on the probability of exiting farming. Panel (b) shows the results on log occupational earning in adulthood. Panel (c) shows the results on the probability of moving long-distance. Finally, panel (d) shows the results on literacy rates in adulthood.

	Low suitability		High suitability	
	(1)	(2)	(3)	(4)
	OLS	IV	OLS	IV
Connected	0.176***	0.233**	0.0141	0.189
	(0.0519)	(0.111)	(0.0143)	(0.567)
Controls	Yes	Yes	Yes	Yes
Observations	5142	5142	7270	7270
Mean of dependent variable	0.321	0.321	0.321	0.321
First-stage F-stat				

Notes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Robust standard errors clustered at the district level in parentheses. All specifications control for a quartic in age, province fixed effects and distance to the nearest targeted city, urban status of the family in 1869, whether the father is literate and whether the father is foreign born. Panel (a): the dependent variable in columns 1 and 2 is an indicator that takes a value of one if an individual worked outside of farming in 1895 and stayed within 200 miles away from his 1869 district of residence. The dependent variable in columns 3 and 4 is an indicator that takes a value of one if the individual worked outside of farming in 1895 and moved at least 200 miles away of his 1869 district of residence. Panel (b): the dependent variable is an indicator that takes a value of one if the individual works outside of farming in 1895. In columns 1 and 2, the sample is restricted to districts with below-median agricultural suitability. In columns 3 and 4, the sample is restricted to districts with above-median suitability.

Table 1: Which factors predict matching? Marginal effects, probit model, sample of sons

	Match					
	(1)	(2)	(3)	(4)	(5)	(6)
First name commonness	-2.282***	-2.295***			-2.285***	-2.296***
	(0.1142)	(0.1102)			(0.1107)	(0.1110)
Last name commonness	-6.791***	-6.915***			-6.798***	-6.911***
	(0.5495)	(0.5885)			(0.5718)	(0.5833)
Leave-one-out enum. rate	0.492***	0.437***			0.486***	0.435***
	(0.0819)	(0.0688)			(0.0781)	(0.0701)
Number of siblings	0.002***	0.002***			0.002***	0.002***
	(0.0005)	(0.0005)			(0.0005)	(0.0005)
Age	0.002***	0.002***			0.002***	0.002***
	(0.0003)	(0.0003)			(0.0003)	(0.0003)
Father's Age	0.000***	0.000***			0.000***	0.000***
	(0.0001)	(0.0001)			(0.0001)	(0.0001)
Father is foreign born	0.011***	0.017***			0.014***	0.016***
	(0.0038)	(0.0038)			(0.0041)	(0.0042)
Urban	0.009***	0.012***			0.011***	0.011***
	(0.0030)	(0.0042)			(0.0034)	(0.0038)
Distance to railroads (1870)			-0.001	-0.002	0.002	0.000
			(0.0025)	(0.0030)	(0.0012)	(0.0017)
Distance to railroads (1875)			-0.001	-0.000	-0.001	-0.001
			(0.0030)	(0.0032)	(0.0020)	(0.0024)
Distance to railroads (1880)			0.002	-0.002	0.001	-0.001
			(0.0027)	(0.0032)	(0.0018)	(0.0024)
Distance to railroads (1885)			0.001	0.001	0.000	0.001
			(0.0020)	(0.0019)	(0.0013)	(0.0013)
Distance to railroads (1890)			-0.004**	-0.001	-0.002	-0.001
			(0.0019)	(0.0015)	(0.0014)	(0.0011)
Distance to railroads (1895)			0.002	0.001	0.001	0.001
			(0.0015)	(0.0012)	(0.0011)	(0.0008)
Place of birth FE	No	Yes	No	Yes	No	Yes
Observations	211925	211925	211925	211925	211925	211925
Match rate	.118	.118	.118	.118	.118	.118

Notes: $***p < 0.01$, $**p < 0.05$, $*p < 0.1$. Robust standard errors clustered at the district level. Table reports average marginal effects of a probit model of the probability that a son in the 1869 census is linked to the 1895 census. Distance to railroads is measured as the log distance from the centroid of a sons' 1869 district to the closest railroad line.

Table 2: Comparing connected and unconnected districts

(a) Sons

		Connected		
	All	Yes	No	Diff.
Variable	(1)	(2)	(3)	(4)
I: Demographic				
Age	7.08	7.16	7.04	0.12
Number of brothers	2.24	2.05	2.32	-0.26***
Father is foreign	0.13	0.25	0.08	0.16***
Father is literate	0.33	0.39	0.31	0.08***
II: Father's Occupation				
White-collar	0.10	0.13	0.08	0.04***
Farmer, farm worker	0.69	0.66	0.71	-0.05*
Skilled/semi-skilled	0.11	0.11	0.11	-0.00
Unskilled urban	0.04	0.05	0.04	0.01
III: Place of residence				
Urban	0.21	0.21	0.21	-0.00
Observations	18643	5386	13257	.

(b) Adults

		Connected		
	All	Yes	No	Diff.
Variable	(1)	(2)	(3)	(4)
I: Demographic				
Age	25.11	25.33	24.96	0.37***
Foreign born	0.31	0.55	0.16	0.39***
Literacy	0.44	0.50	0.41	0.09***
II: Occupation				
White-collar	0.13	0.15	0.12	0.03**
Farmer, farm worker	0.58	0.52	0.62	-0.09***
Skilled/semi-skilled	0.14	0.16	0.12	0.04***
Unskilled urban	0.08	0.09	0.07	0.02
III: Place of residence				
Urban	0.28	0.30	0.27	0.04
Observations	8159	3230	4929	.

Notes: In column 1, I compute the mean of each of the variables for individuals in the baseline sample in 1869. In column 2, I focus on individuals who in 1869 resided in districts that would be connected to the railroad network by 1885. Column 3 focuses on individuals in unconnected districts. Column 4 reports the difference between the groups in columns 2 and 3. Standard errors of the differences in column 4 are clustered at the district level. Significance levels are indicated by: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Sample is restricted to individuals residing outside of targeted cities and outside the province of Buenos Aires. Panel (a) shows the results in the sample of sons and panel (b) shows the results in the sample of adults.

Table 3: Railroads and the probability of transitioning out of farming occupations

(a) Sons

	OLS		IV	
	(1)	(2)	(3)	(4)
Connected	0.0436***	0.0411**	0.188*	0.185*
	(0.0164)	(0.0159)	(0.108)	(0.106)
Controls	No	Yes	No	Yes
Observations	12412	12412	12412	12412
Mean of dependent variable	0.321	0.321	0.321	0.321
First-stage F-stat				

(b) Adults

	OLS		IV	
	(1)	(2)	(3)	(4)
Connected	0.0471**	0.0398*	-0.0668	-0.0520
	(0.0217)	(0.0217)	(0.215)	(0.211)
Controls	No	Yes	No	Yes
Observations	4412	4412	4412	4412
Mean of dependent variable	0.280	0.280	0.280	0.280
First-stage F-stat				

Notes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Robust standard errors clustered at the district level in parentheses. The dependent variable is an indicator that takes a value of one if the individual worked outside of farming in 1895. All specifications control for province fixed effects, distance to the nearest targeted city and a quartic in age. In the controlled specification in panel (a), I further control for urban status, number of siblings in the household in 1869, whether the father was literate and whether the father was foreign born. In the controlled specification in panel (b), I further control for urban status, literacy and immigration status.

Table 4: Railroads and occupational transition probabilities

(a) Sons

	Son's occupation, 1895				
Father's occ., 1869	White collar	Farmer, farm worker	Skilled/Semi-skilled	Unskilled urban	N
White collar	0.039	-0.083**	0.023	0.021	1786
	(0.031)	(0.035)	(0.026)	(0.015)	
Farmer, farm worker	0.021**	-0.047***	0.022**	0.004	12865
	(0.009)	(0.016)	(0.009)	(0.006)	
Skilled/semi skilled	0.042	-0.066	0.016	0.008	2011
	(0.031)	(0.041)	(0.032)	(0.021)	
Unskilled urban	0.043	0.016	0.012	-0.072**	752
	(0.044)	(0.078)	(0.062)	(0.028)	
All	0.035***	-0.064***	0.026**	0.003	18417
	(0.012)	(0.018)	(0.011)	(0.006)	

(b) Adults

	Occupation, 1895				
Occupation, 1869	White collar	Farmer, farm worker	Skilled/Semi-skilled	Unskilled urban	N
White collar	0.072**	-0.029	-0.014	-0.016	1116
	(0.035)	(0.042)	(0.021)	(0.019)	
Farmer, farm worker	0.022*	-0.047**	0.003	0.010	4914
	(0.013)	(0.023)	(0.010)	(0.009)	
Skilled/semi skilled	-0.046	-0.011	0.043	0.010	1166
	(0.031)	(0.040)	(0.042)	(0.024)	
Unskilled urban	-0.031	-0.093	0.095**	-0.001	634
	(0.041)	(0.072)	(0.043)	(0.049)	
All	0.017	-0.042**	0.017	0.005	8462
	(0.012)	(0.020)	(0.010)	(0.008)	

Notes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Robust standard errors clustered at the district level in parentheses. In panel (a), each cell represents a different OLS regression of the probability that a son with a father in occupational i worked in occupational category j in adulthood. In panel (b), each cell represents a different OLS regressions of the probability that an adult employed in occupational category i in 1869 worked in occupational category j in 1895. The last column of each of the matrices shows the number of observations in each of the regressions.

Table 5: Railroads and urbanization

	OLS		IV	
	(1)	(2)	(3)	(4)
Connected	0.0551***	0.0532***	0.182***	0.180***
	(0.0187)	(0.0182)	(0.0664)	(0.0668)
Controls	No	Yes	No	Yes
Observations	10397	10397	10397	10397
Mean of dependent variable	0.174	0.174	0.174	0.174
First-stage F-stat			14.50	14.42

Notes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Robust standard errors clustered at the district level. The dependent variable is an indicator that takes a value of one if the individual resided in an urban location. All regressions control for a quartic in age, province fixed effects and distance to the nearest targeted city. In columns 2 and 4, I further control for a vector of background household characteristics: number of siblings in the household in 1869, urban status of the family, whether the father was literate and whether the father was foreign born.

Table 6: Sons in connected districts were more likely to be literate in adulthood

	OLS		IV	
	(1)	(2)	(3)	(4)
Connected	0.0482***	0.0412***	0.199*	0.125
	(0.0164)	(0.0151)	(0.115)	(0.0892)
Controls	No	Yes	No	Yes
Observations	17700	17700	17700	17700
Mean of dependent variable	0.520	0.520	0.520	0.520
FIrs-stage F-stat				

Notes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Robust standard errors clustered at the district level. All regressions control for a quartic in age, province fixed effects and distance to the nearest targeted city. In columns 2 and 4, I further control for a vector of background household characteristics: number of siblings in the household in 1869, the occupational category of the father, urban status of the family, whether the father was literate and whether the father was foreign born.

Online Appendix - Not for Publication

A Data Appendix

A.1 Linking Algorithm

The procedure for linking males from the 1869 to the 1895 national censuses is analogous to the one proposed in Winkler 1988 and first used in economic history by Mill and Stein 2012.

I started by identifying a set of potential matches for each individual in the 1869 census. To be considered a potential match, an individual in the 1895 census had to satisfy the following criteria: (1) being born in the same province of birth (country of birth in the case of the foreign born), (2) a first name with the same first letter, (3) a last name with the same first letter, (4) a predicted age difference within five years.

For those pairs of individuals that met these criteria, I measured their similarity in terms of three identifying variables: first name, last name and age.[32] To calculate the similarity in first and last names, I used the Jaro-Winkler string distance. The Jaro-Winkler distance (Winkler 1990) is a measure of the similarity between two strings. The measure is normalized such that a score of zero represents two identical strings and a score of one represents two strings with no common characters. I measured the similarity between the ages by computing the absolute value of their difference – that is, $abs(Age^{1869}+26-Age^{1895})$.

This procedure generated a vector with three distance measures (first name, last name, age) for each pair of potential matches. The next step of the procedure is to summarize this vector into a single score representing the probability that these records constitute a true match – that is, belong to the same individual. There are two key relationships – derived from Bayes rule – that allow me to estimate this probability.

First, the probability that a pair i is a match ($i \in M$) conditional on the distances in the identifying variables can be derived from Bayes Rule as:

$$P(i \in M/Distance) = \frac{P(Distance/i \in M)P(i \in M)}{P(Distance)} \tag{1}$$

[32] Theoretically, it would be possible to compute this vector of distances for each possible pair of records in the two censuses. However, this computation is impractical from as the number of calculations is proportional to the product of the observations in the two datasets. Even with moderately sized datasets, the number of calculations involved soon becomes too large.

Second, the probability of observing a given vector of distances for each pair i can be written as:

$$P(Distance) = P(Distance/i \in M)P(i \in M) + P(Distance/i \notin M)P(i \notin M) \tag{2}$$

Equation 2 implies that the distances can be modeled as a mixture of two statistical distributions, with unknown mixture parameter $p_m = P(i \in M)$. That is, for each pair i in the data, I observe the vector of distances but I do not know whether the observation belongs to the set M or to its complement – i.e. corresponds to the same individual or not.

To estimate this mixture model, I made two simplifying assumptions following Winkler 1988. First, I assumed that the distances in each of the identifying variables are conditionally independent of each other. This assumption implies that it is possible to write:

$$P(Distance) = \prod_{k=1}^{3} P(Distance_k/i \in M)P(i \in M) + \prod_{k=1}^{3} P(Distance_k/i \notin M)P(i \notin M) \tag{3}$$

where k indexes the three identifying variables (first name, last name, age).

Second, I assumed that each component of the mixture model follows a multinomial distribution. The goal is then to estimate:

$$\theta_M = [\theta_M^{First\,Name}, \theta_M^{Last\,Name}, \theta_M^{Age}] \tag{4}$$

where θ_M is the vector of parameters of the multinomial distribution when $i \in M$ and:

$$\theta_U = [\theta_U^{First\,Name}, \theta_U^{Last\,Name}, \theta_U^{Age}] \tag{5}$$

where θ_U is the vector of parameters of the multinomial distribution when $i \notin M$.

I further assumed that the absolute difference in age between two records can adopt any value in the zero to five range. Finally, I discretized the name distances by dividing it into four mutually exclusive groups following Winkler 1988: [0,0.1],[(0.1,0.12], (0.12,0.25], (0.25,1].

The final step is then to choose θ_M, θ_U and p_M so as to maximize the probability of generating the distances that are actually observed in the data, that is:

$$\max_{\theta_M, \theta_U, p_m} \prod_{k=1}^{3} P(Distance_k/i \in M)P(i \in M) + \prod_{k=1}^{3} P(Distance_k/i \notin M)P(i \notin M) \tag{6}$$

The solution to this maximization problem does not have a closed form solution. I solve it instead using the EM algorithm. The EM algorithm is a well-known numerical procedure for finding maximum

likelihood estimates when the likelihood function lacks a closed form solution and is a standard procedure for estimating mixture models.

A.2 Accounting for Match Failure

There is a trade-off in the linking procedure between efficiency – matching a large fraction of the observations – and accuracy – avoiding incorrect matches. In the baseline sample, I find a match – defined as a potential match with a linking score above the $\underline{p}$ threshold – for about 38 percent and a unique match – defined as a potential match with a linking score that is both above the $\underline{p}$ threshold *and* sufficiently better than the second best match – for about 11 percent of the sons in the sample. When linking adults, I find a potential candidate for 28 percent of the sample and a unique candidate for approximately 10 percent (table A.2).

Table A.3 compares the observed matching rates with the predicted matching rates after subtracting mortality, census underenumeration and return migration. The main reason for match failure is mortality during the intercensal period. Based on the censuses full count, I estimated that about 56 percent of the individuals in the 1869 cross-section had survived by 1895.[33] Hence, the matching rate is capped at 56 percent. In addition, despite censuses are intended to be a full count of the population, there is non-trivial underenumeration. I am not aware of estimates of underenumeration in 19th-century Argentina censuses, but contemporary estimates using US census data find underenumeration rates ranging from 7.4 percent to as high as 23 percent (King and Magnuson 1995). An additional source of match failure for the foreign born is return migration. Assuming independence among these three sources of underenumeration, the predicted matching rate ranges from 45 percent to 53 percent for sons of natives and working-age natives, and 29 percent to 48 percent for working-age foreigners. The remaining difference between the predicted and the observed matching rate corresponds to errors in the enumeration process that are too severe to be accommodated by my linking procedure. For instance, individuals that misreport their age by more than five years or that have the first letter of their first or last names misspelled will be missed by my linking procedure.

A.3 Identifying Fathers and Sons in the Data

There are two challenges in identifying fathers and sons in the 1869 census. The first is that not all fathers and sons lived in the same household by the time of the census. To minimize this possibility, I

[33] As a comparison, 69 percent of white, native-born males under age 25 survived from 1850 to 1880 in the US (Long and Ferrie 2013).

restrict the analysis to children who are young enough (16 years old or younger) in 1869 to presumably co-reside with their parents in that census year. However, even if the children were relatively young at the moment of the census, the father might nevertheless have have been absent from the household – for instance, due to mortality.

A second challenge is that even if fathers co-reside with their sons, the data lack household identifiers. Similar to US censuses of the mid 19th-century, the 1869 census does not include a question on the relationship of each household member to the head of the household. However, because members of the same household were recorded in the census forms consecutively and father and sons share their last names, it is possible to identify for each individual a set of potential fathers. More precisely, for each male under the age of 16, I identified the set of potential fathers as anyone who met all of the following criteria: (1) same last name, (2) recorded consecutively in the census forms (either on the same page or on the one immediately before or on the one immediately after) and (3) an age difference of at least 15 years but no more than 55 years.

The procedure for identifying fathers and sons is similar to the one used by IPUMS to impute relationships among different household members in the 1850, 1860 and 1870 US censuses, where the question on relationship to head of household is also unavailable (Ruggles et al. 1997). Table A.4 summarizes the results of this procedure. For about 53.4 percent of individuals aged 16 years old or less, I found a candidate father. In particular, 42 percent of sons had exactly one person satisfying all of the above criteria, 8 percent had two and about 2 percent had three or more. Finally, for 46.5 percent of individuals aged 16 years old or less it was not possible to find a candidate father.

In those cases in which there were two or more potential candidates, I ranked the fathers based on the following criteria (from more to less likely): (1) recorded on the same page of the census, (2) recorded on the previous page of the census, (3) age difference with respect to son more or equal than 20 and less or equal than 40, (4) age difference closer to median age difference among those with a unique candidate father.

The fraction of sons not co-residing with their father in this time period is high relative to the US. Cacopardo et al. 1997 describe a number of reasons for this phenomenon.[34] First, mortality rates were higher in Argentina than in the US. Second, there is a significant number of single mothers in the data. Argentina in the time period was still characterized by a high prevalence of non-traditional households. Indeed, near 20 percent of the children in the 1869 census cross-section were classified as illegitimate -that is, born outside of a marriage-.

[34]In a comment article to Long and Ferrie 2013, Xie and Killewald 2013 show that in the US 1850 census about 70 percent of sons aged 0-25 years old co-resided with their fathers.

One concern with the analysis is that the likelihood of observing a father in the household might be correlated with railroad access. To investigate this issue, in figure A.5 I show a binned scatter plot of the probability of having a father present in the household (y-axis) and distance to the closest railroad line (x-axis). Regardless of the year in which access to railroads is measured, there is a nearly flat relationship between the presence of a father in the household and distance to the closest railroad line. The slope of the relationship between distance to the railroad network and the probability of a present father is at most 0.01. This magnitude implies that doubling the distance to the railroad network is associated with a 0.7 percentage point increase in the probability of a present father in the household, relative to a baseline of 53 percent.

A.4 Geocoding the 1869 and 1895 Censuses

I geocoded each individual in the 1869 census at the census enumeration district-level.[35] The census enumeration district typically corresponds to a town – in the case of small towns – or to a section of a town – in the case of larger towns. In cases where the district had changed its name, I used information from a historical dictionary of Argentine geography (Latzina 1906) to identify its location. This dictionary contains the name of the district, together with a verbal description of its geographic location that includes its province and department, as well as information on nearby towns and rivers. The district-level information enables me to more precisely measure the extent to which an individual resided in a location that was connected to the railroad network.

Figure A.1 shows the location of each of the districts – blue dots – included in the linked sample. Individuals in the linked sample resided in 583 different districts in 1869.Note that, at the time of the first national census, a large portion of contemporary Argentina was outside of the central government's control and hence was not included in the census.[36]

[35]The first page of each of the books used during the enumeration process indicated both the department and section in which the enumeration took place. While the published census tabulations report information at the department level (Fuente 1872), it is possible to recover the finer geographic detail by looking into the original census manuscripts that are available in familysearch.org.

[36]The census authorities estimated that, by 1869, the indigenous population excluded from the census was approximately 93,000, or about 5 percent of the total population of Argentina at the time.

A.5 Occupational and Earnings Data

A.5.1 Salaried Workers

The data on daily wages of blue-collar workers comes from two main sources: Buchanan 1898 and the 1881 census of the province of Buenos Aires (Provincia de Buenos Aires 1883). Buchanan was the economic aggregate of the US Embassy in Buenos Aires and systematically collected wage data for workers in this city. His report contains yearly information on the typical wages on 95 occupations from 1886 to 1896. These data have been used extensively in historical research (Dorfman 1942; Panettieri 1965; Panettieri 1998) and are considered to be accurate (Conde 1979). For each of the occupations in the Buchanan's data, I computed a simple average of the wages in 1894 and 1896. I complement these data with the 1881 census of the Province of Buenos Aires, which contains information on the wages in 65 different occupations.

Because of its greater level of detail and the availability of data closer to 1895, my baseline results use the information from Buchanan 1898 to assign a wage to those occupations in which this information is available from both Buchanan 1898 and the 1881 census. In those cases where the information is only available in the 1881 census, I use wages in the census scaled by the mean wage in Buchanan 1898.

The data on the wages of public employees comes from the 1893 national census of public employees, which contains the full roster of public employees and their corresponding monthly wages (Argentina, Dirección General de Estadistica 1895). I computed average wages in the public sector by dividing the sum of the wages of male public employees by their total number.

Wages in the above sources were sometimes reported on a monthly basis. In those cases, I converted the data to daily values by dividing the monthly wage by 25 working days, following the assumption in Álvarez and Nicolini 2010.

In a small number of occupations, the above sources report salaries that include both a monetary and an in-kind payment, typically in terms of food and lodging. In those cases, I followed Álvarez and Nicolini 2010 and assumed that earnings were 50 percent higher if they included an allowance for food and 100 percent higher if they included an allowance for both food and lodging.

A.5.2 Business Owners

To estimate the average income of storekeepers – "comerciantes" – , I complemented the wage data with information on the size of the capital stock in the commercial sector, obtained from the third volume of the 1895 national census (Fuente 1898). Conceptually, the earnings of a storekeeper could be decomposed into the returns to capital and the returns to labor. Based on this insight, average occupational earnings were

computed as the sum of the earnings of a store clerk – the returns to labor – and the estimated per capita returns to capital, assuming a 8 percent net annual return on capital. I computed per capita capital as the ratio between the total capital stock in the commercial sector by province – as reported in the census – and the total number of individuals who declared working as shopkeepers in the census by province. More precisely, earnings of shopkeepers s were estimated as: $Earnings_s = Labor\,Income_s + 0.08\frac{Capital_s}{\#Storekeepers}$.

Average earnings of the owners of industrial firms – "industriales" and "fabricantes" in the census – were similarly estimated by adding the labor income – in this case, the earnings of a foreman – and the returns to capital in the industrial sector. Data on capital in the industrial sector are also from the 1895 national census.

A.5.3 Farmers

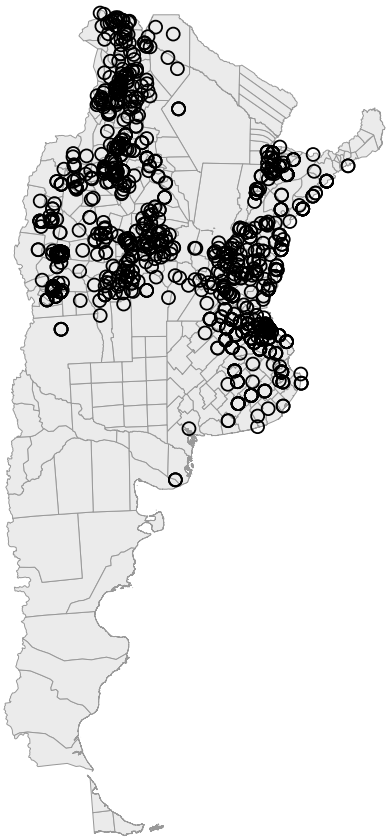
I estimated the income of farmers using the following procedure. I first divided farmers in the agricultural sector into two groups: *hacendados* and *agricultores*. The *hacendados* category corresponds to those holding the largest plots of land. Overall, less than 10 percent of all the farmers in my sample are in the *hacendados* category. The second group in the agricultural sector -*agricultores*. encompasses the vast majority of farmers. I then estimated the earnings of farmers using the information provided in the Congressional report of the farming sector prepared by Correa and Lahitte 1898. This report includes information on the typical revenue and expenditure in inputs of farms of different size.

I similarly divided those in the cattle ranching sector into two groups: *estancieros* and *criadores*. The *estancieros* category corresponds to those holding the largest plots.

A.5.4 Others

Finally, I assigned the mean earnings within their corresponding HISCLASS to those occupations for which I could not find information in the above sources.

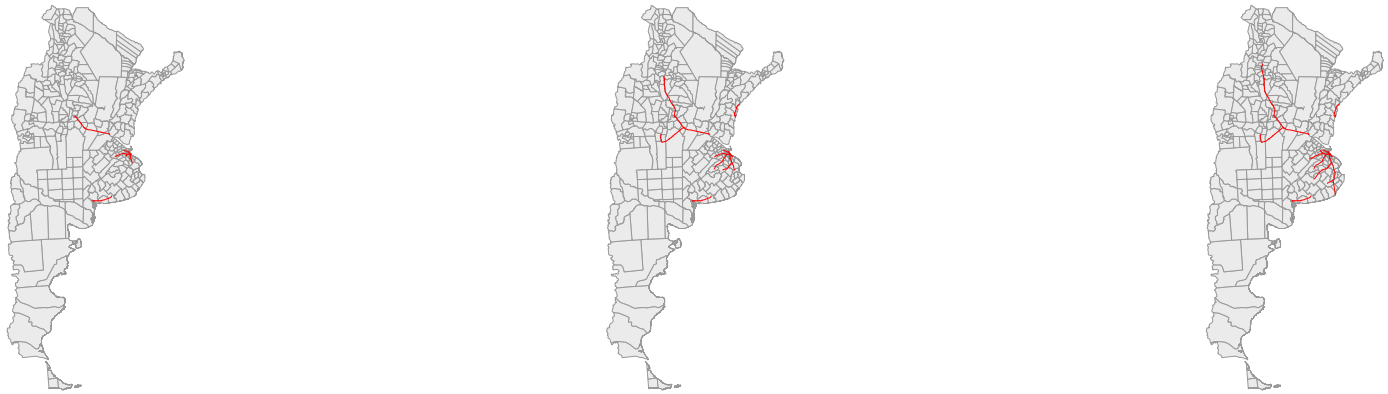
Figure A.1: Location of Individuals in 1869



Notes: This map shows the 1869 location of individuals in the linked sample. The polygons represent departments of 1869 Argentina. The dots represent the centroids of the 1869 districts.

Figure A.2: Expansion of the railroad network (1870-1895)

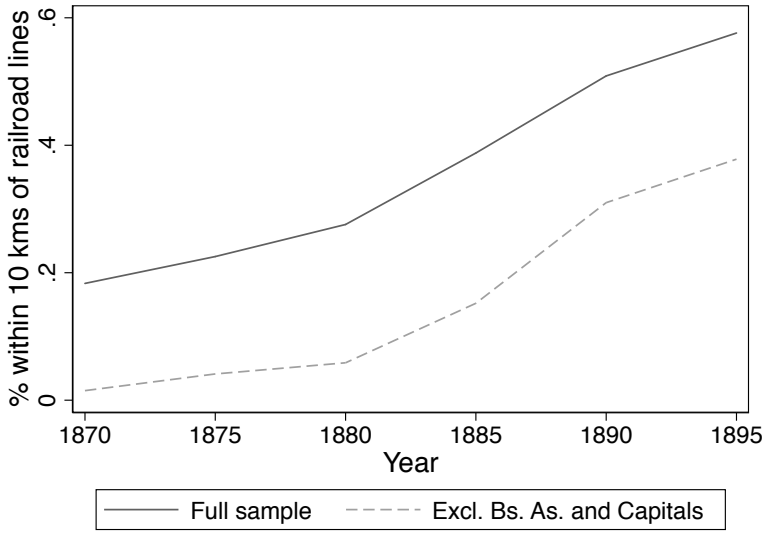
(a) 1870 **(b)** 1875 **(c)** 1880



(d) 1885 **(e)** 1890 **(f)** 1895

Notes: This figure shows the expansion of the railroad network that took place from 1870 to 1895.

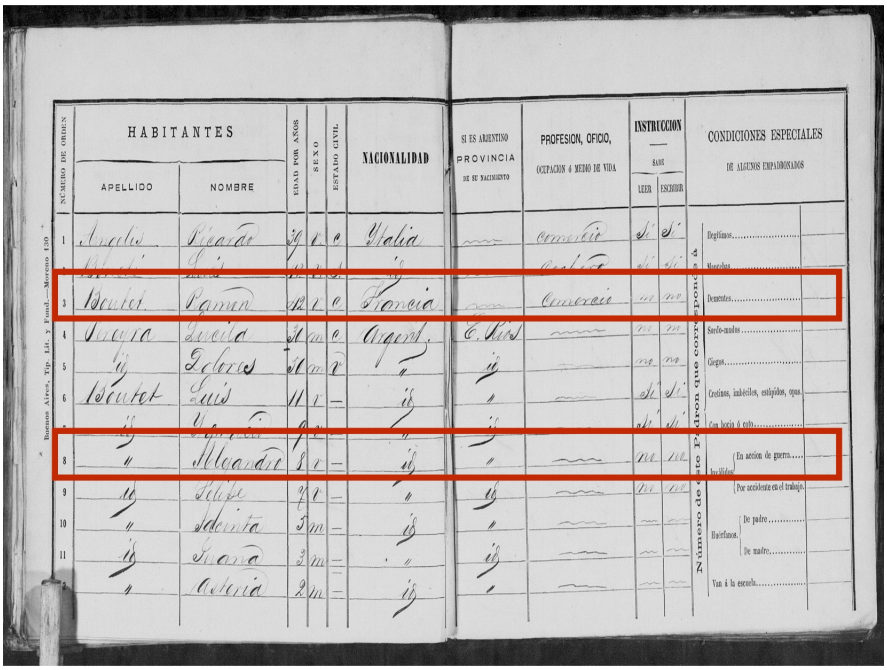
Figure A.3: Fraction of individuals within 10 kilometers of railroad lines



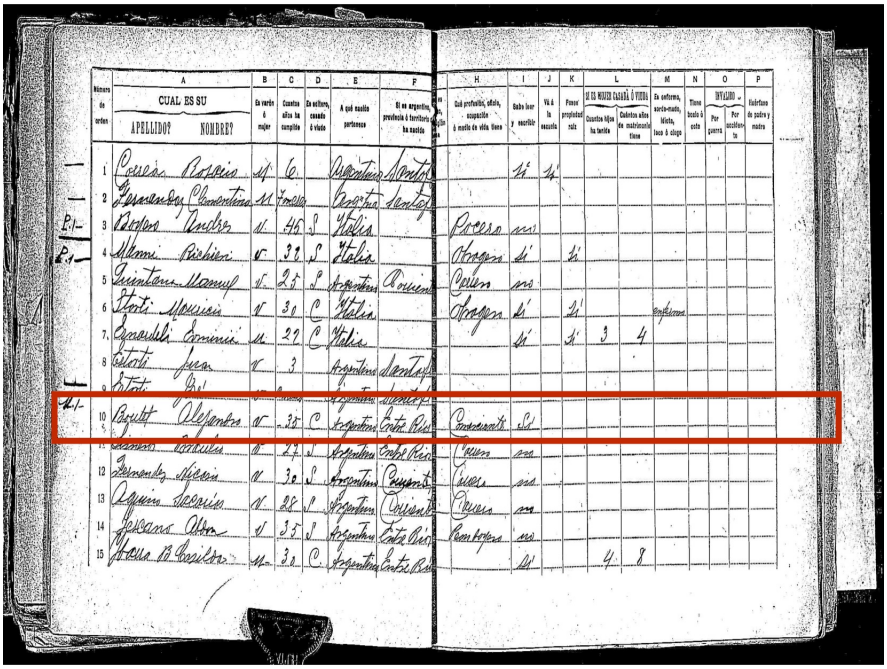
Notes: This figure shows the fraction of individuals residing within 10 kilometers of railroad lines, based on place of residence in 1869. Distance from railroad lines is based on a district's centroid.

Figure A.4: Illustration of the linking procedure

(a) 1869 census

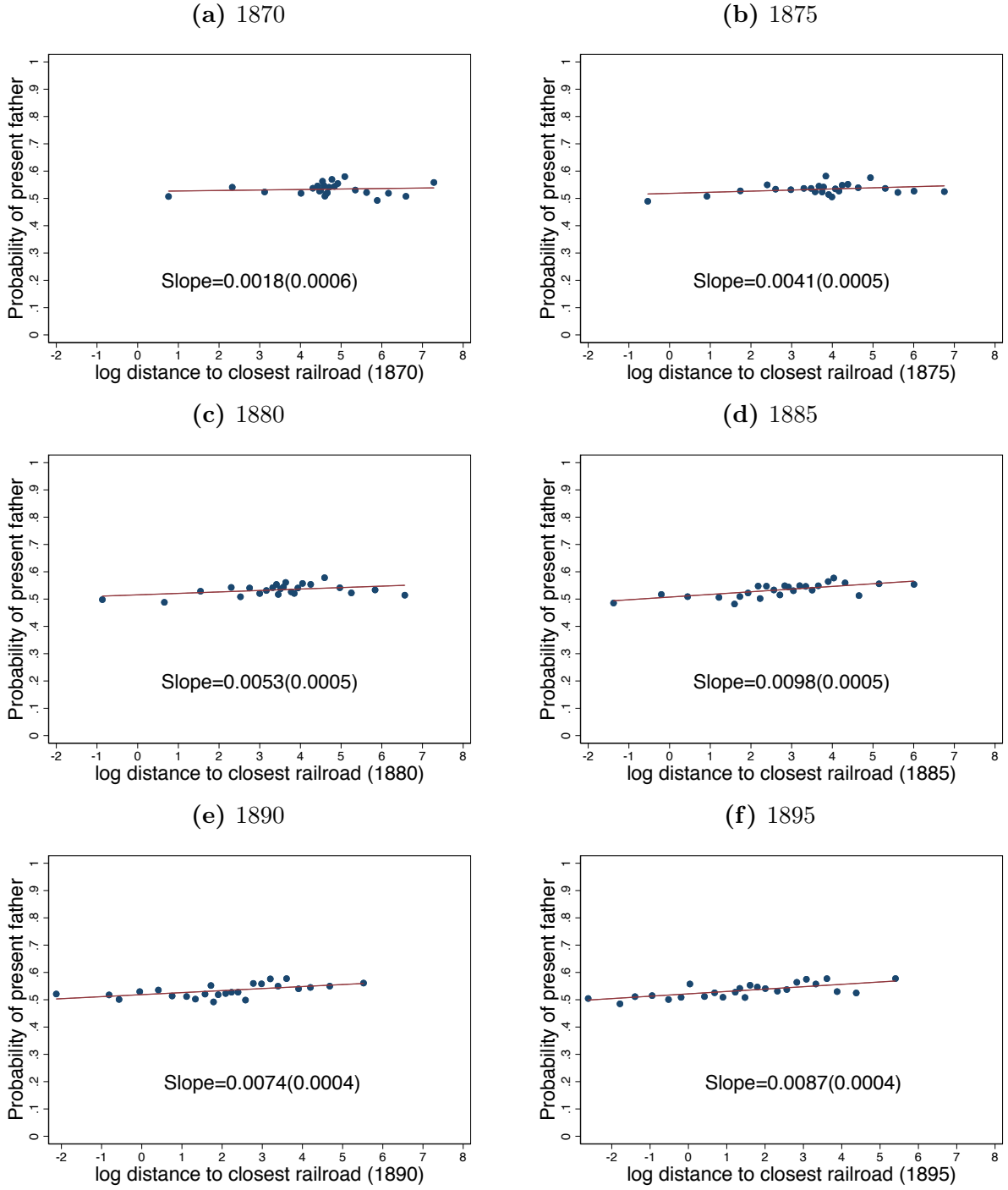


(b) 1895 census



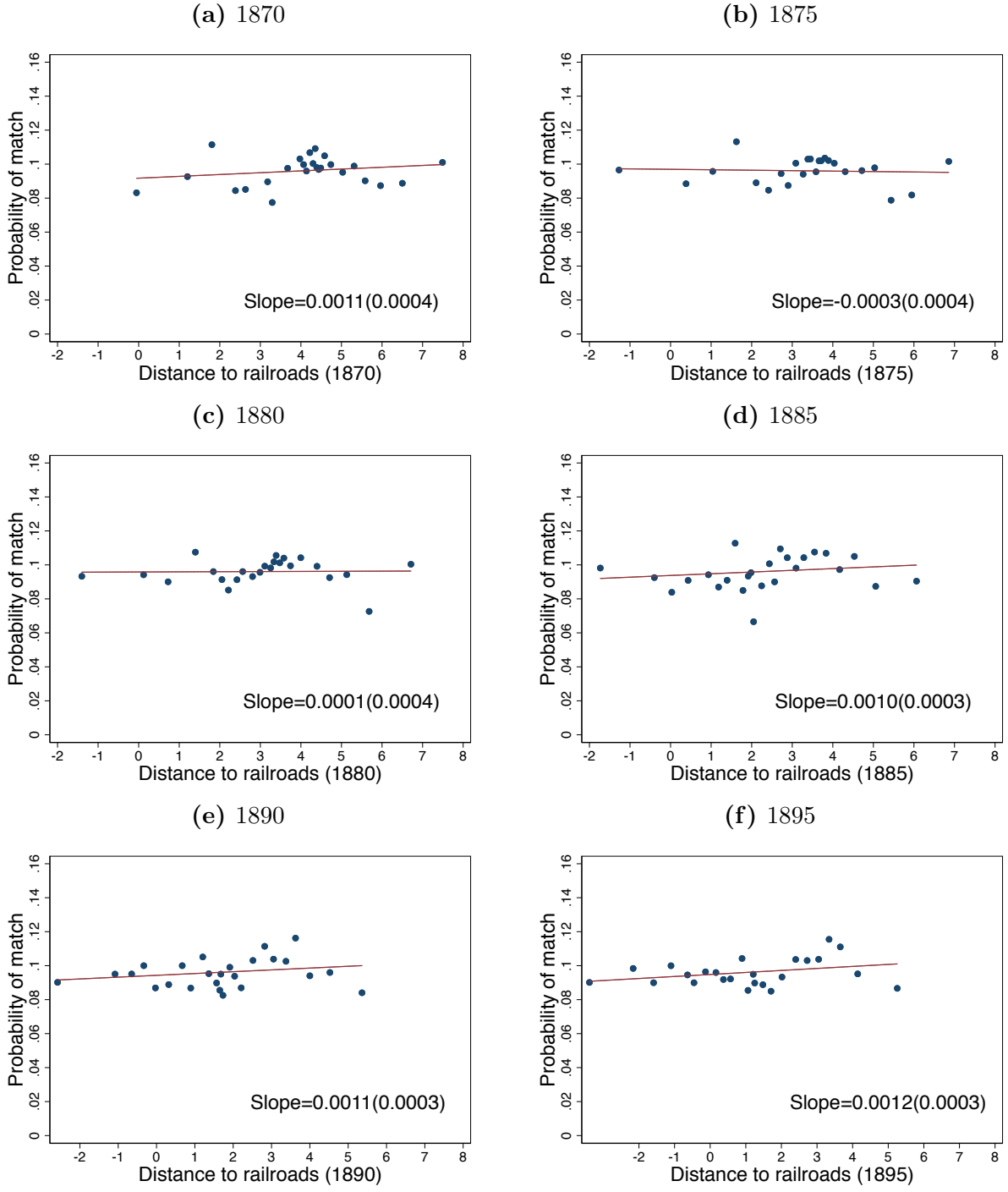
Notes: The first panel shows a family in the 1869 census. In this figure, the father ("Ramon Boutet") and the son ("Alejandro Boutet") are observed living in the same household. The second panel shows the son ("Alejandro") in the 1895 census.

Figure A.5: Probability of present father and distance to closest railroad line



Notes: This figure shows a binned scatterplot of the probability that the father is present in the household (y-axis) on the log distance to the closest railroad line, by year of construction (x-axis). I divide the data into 25 equally sized bins based on distance to the closest railroad line. The y-axis shows the average probability of a present father within each of these bins. Distance is measured in kilometers from the district's centroid. For each of the binned scatterplots, I report the slope coefficient of an OLS regression of the probability of a present father on distance to the closest railroad line, controlling for province of birth fixed effects (standard errors in parentheses). Sample is restricted to all males aged 0 to 16 years old in 1869.

Figure A.6: Probability of a match and distance to closest railroad line, sample of adults



Notes: This figure shows a binned scatterplot of the probability of being on the linked sample (y-axis) on the log distance to the closest railroad line, by year of construction. I divide the data into 25 equally sized bins based on distance on the closest railroad line. The y-axis shows the average probability of a match within each of the bins. Distance is measured in kilometers from the district's centroid. For each of the binned scatterplots, I report the slope coefficient of an OLS regression of the probability of a match on distance to the closest railroad line, controlling for province of birth fixed effects (standard errors in parentheses).

Table A.1: Which factors predict matching? Marginal effects, probit model, sample of adults

	Match					
	(1)	(2)	(3)	(4)	(5)	(6)
First name commonness	-0.998***	-1.013***			-0.999***	-1.012***
	(0.0684)	(0.0674)			(0.0682)	(0.0659)
Last name commonness	-3.205***	-3.402***			-3.186***	-3.386***
	(0.5895)	(0.5842)			(0.5983)	(0.5931)
Leave-one-out enum. rate	0.582***	0.485***			0.579***	0.484***
	(0.0820)	(0.0720)			(0.0838)	(0.0711)
Age	-0.001***	-0.001***			-0.001***	-0.001***
	(0.0001)	(0.0001)			(0.0001)	(0.0001)
Foreign born	0.003	0.004			0.004	0.005
	(0.0060)	(0.0147)			(0.0080)	(0.0146)
Urban	0.003	0.000			0.004	0.002
	(0.0027)	(0.0022)			(0.0024)	(0.0024)
Distance to railroads (1870)			0.001	0.003***	0.002*	0.003***
			(0.0011)	(0.0012)	(0.0009)	(0.0008)
Distance to railroads (1875)			-0.004	-0.004*	-0.003	-0.004**
			(0.0026)	(0.0025)	(0.0016)	(0.0017)
Distance to railroads (1880)			0.000	0.000	0.001	0.001
			(0.0026)	(0.0026)	(0.0017)	(0.0019)
Distance to railroads (1885)			0.003	0.002	0.001	0.001
			(0.0017)	(0.0016)	(0.0011)	(0.0012)
Distance to railroads (1890)			-0.002	-0.000	-0.001	0.000
			(0.0022)	(0.0014)	(0.0014)	(0.0010)
Distance to railroads (1895)			0.001	0.001	0.001	0.001
			(0.0019)	(0.0011)	(0.0013)	(0.0008)
Place of birth FE	No	Yes	No	Yes	No	Yes
Observations	236258	236258	236258	236258	236258	236258
Match rate	.0961	.0961	.0961	.0961	.0961	.0961

Notes: $***p < 0.01$, $**p < 0.05$, $*p < 0.1$. Robust standard errors clustered at the district level in parentheses. This table shows the average marginal effects of a probit model of the probability that an adult – aged 18-35 years old in the 1869 census – is uniquely linked to the 1895 census. Distance to railroads is measured as the log distance from the centroid of an individual's 1869 district to the closest operating railroad line.

Table A.2: Matching rates by age group and nativity status

(a) Sons

Country of Origin	N	Matched	Uniquely Matched	Matched (%)	Uniquely matched (%)
Immigrants	35334	12491	4584	0.354	0.130
Natives	180771	70094	20895	0.388	0.116
Total	216105	82585	25479	0.382	0.118

(b) Working-age individuals

Country of Origin	N	Matched	Uniquely Matched	Matched (%)	Uniquely matched (%)
Immigrants	58755	18934	5824	0.322	0.099
Natives	182982	48593	17352	0.266	0.095
Total	241737	67527	23176	0.279	0.096

Notes: This table shows the matching rates obtained in the linking procedure. Panel (a) shows the matching rates for males aged 0 to 16 years old with father present in the household in the 1869 census. Panel (b) shows the matching rates for males aged 18-35 years old in the 1869 census. In panel (a), I distinguish between sons of immigrants and sons of natives. In panel (b), I distinguish between foreign born and natives.

Table A.3: Accounting for match failure

	Natives		Foreigners
	$\leq$ 16 years old	$\geq$ 18 years old	
	(1)	(2)	(3)
Mortality	0.42	0.42	0.25
Census underenumeration	0.9-0.22	0.9-0.22	0.9-0.22
Return migration	.	.	0.3-0.5
Predicted matching rate	0.45 -0.53	0.45 -0.53	0.29 -0.48
Share matched	0.36	0.27	0.32
Share uniquely matched	0.11	0.09	0.10

Notes: This table reports the reasons for match failure in the sample linking the 1869 and 1895 censuses. The predicted matching rate is computed assuming independence among the factors leading to match failure. Mortality is estimated based on the census data. Return migration estimates are from (Alsina 1898). Estimates of census underenumeration are based on estimates from the US spanning the same time period (Knights 1991), as no references were found for the case of Argentina.

Table A.4: Identifying fathers and sons in the 1869 census

	N	%
Males aged 16 or less in 1869	405453	100
Potential father	216420	53.38
i. One	174133	42.95
ii. Two	33773	8.33
iii. Three or more	8514	2.10
No potential father	189033	46.62

Notes: This table shows the number and fraction of males aged 16 years old or less who were co-residing with their father in the 1869 census. The procedure for identifying fathers and sons in the data is described in detail in section A.3.

Table A.5: Most frequent farming occupations: original occupational titles

(a) Fathers (1869)

Occupation	Frequency
Labrador	6838
Jornalero	2392
Estanciero	1804
Criador	964
Hacendado	894
Agricultor	691
Pastor	404
Acarreador	338
Peon Rural	316
Puestero	153
Ganadero	111
Chacarero	58
Rural	29
Ganan	29
Quintero	28

(b) Sons (1895)

Occupation	Frequency
Jornalero	4472
Labrador	2792
Agricultor	2764
Criador	1252
Estanciero	970
Hacendado	578
Pastor	196
Ganadero	130
Puestero	122
Acarreador	121
Peon Rural	108
Chacarero	63
Marinero	26
Domador	19
Lenador	18

Notes: This table shows the most frequent occupations of fathers (in 1869) and sons (in 1895) among those individuals that I classify as employed in farming.

Table A.6: Most frequent occupations of individuals in the linked sample

(a) Fathers (1869)

Occupation	Frequency	%	Occupational group
Farmer	9872	40.29	Farmer
Laborer	3161	12.90	Unskilled
Storekeeper	1730	7.06	White collar
Breeder	1051	4.29	Farmer
Carpenter	790	3.22	Skilled/semiskilled
Shoemaker	532	2.17	Skilled/semiskilled
Shepherd	446	1.82	Unskilled
Foreman	371	1.51	White collar
Construction worker	321	1.31	Skilled/semiskilled
Carter	189	0.77	Skilled/semiskilled
Total Top 10	18463	75.36	
Total	24501	100	

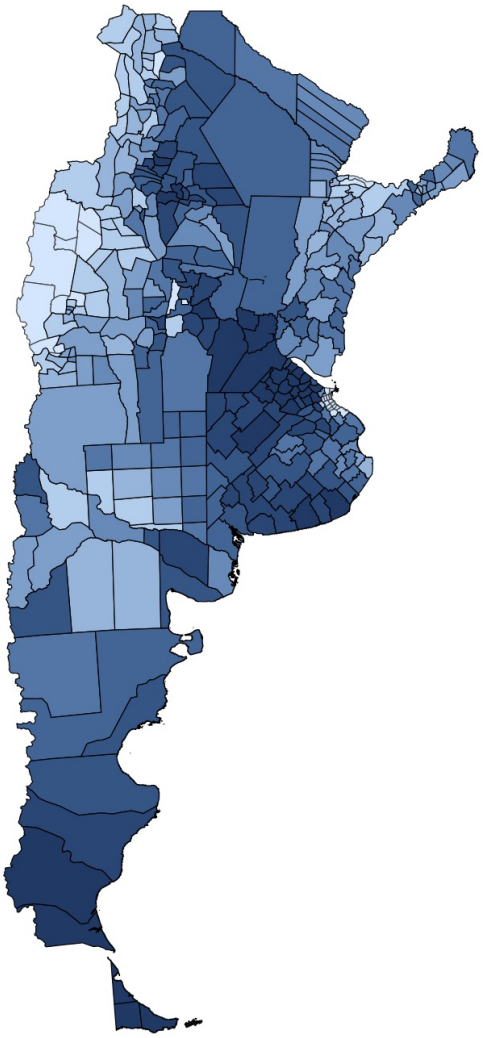
(b) Sons (1895)

Occupation	Frequency	%	Occupational group
Farmer	6845	27.94	Farmer
Laborer	5481	22.37	Unskilled
Storekeeper	1675	6.84	White collar
Breeder	1319	5.38	Farmer
Carter	539	2.20	Skilled/semiskilled
Public employee	467	1.91	White collar
Carpenter	410	1.67	Skilled/semiskilled
Independent means	387	1.58	White collar
Construction worker	372	1.52	Skilled/semiskilled
Soldier	348	1.42	Skilled/semiskilled
Total Top 10	17843	72.83	
Total	24501	100	

Notes: This table shows the ten most frequent occupations among fathers (in 1869) and sons (in1896) in the linked sample.

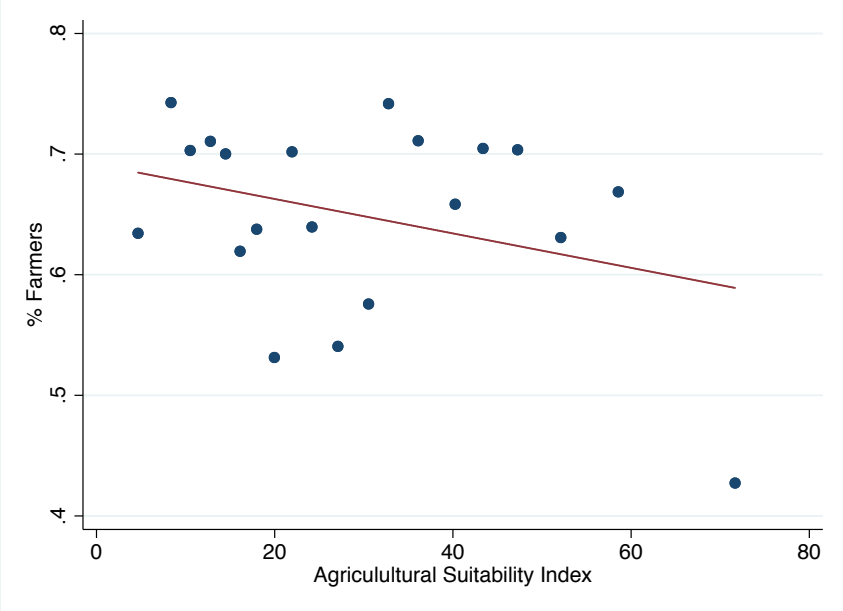
B Robustness and Additional Results

Figure B.1: Agricultural suitability index



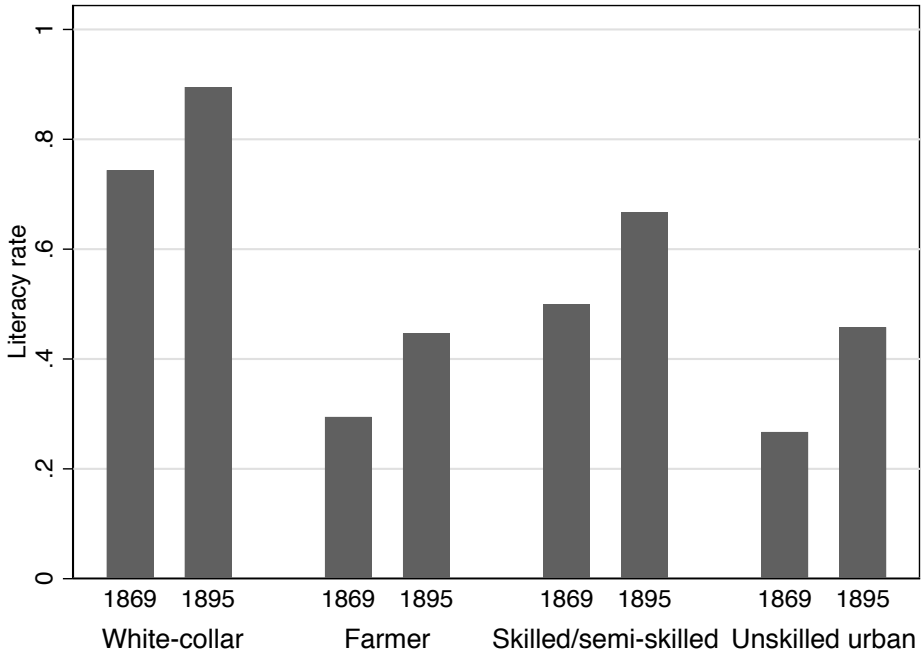
Notes: In this map, I group departments into ten equally sized groups based on their agricultural suitability, as computed by the National institute of Agricultural Technology of Argentina (Cruzate et al. 2012). Darker areas represent departments with higher agricultural suitability.

Figure B.2: Departments with lower agricultural suitability had a higher share of the population employed in farming in 1869



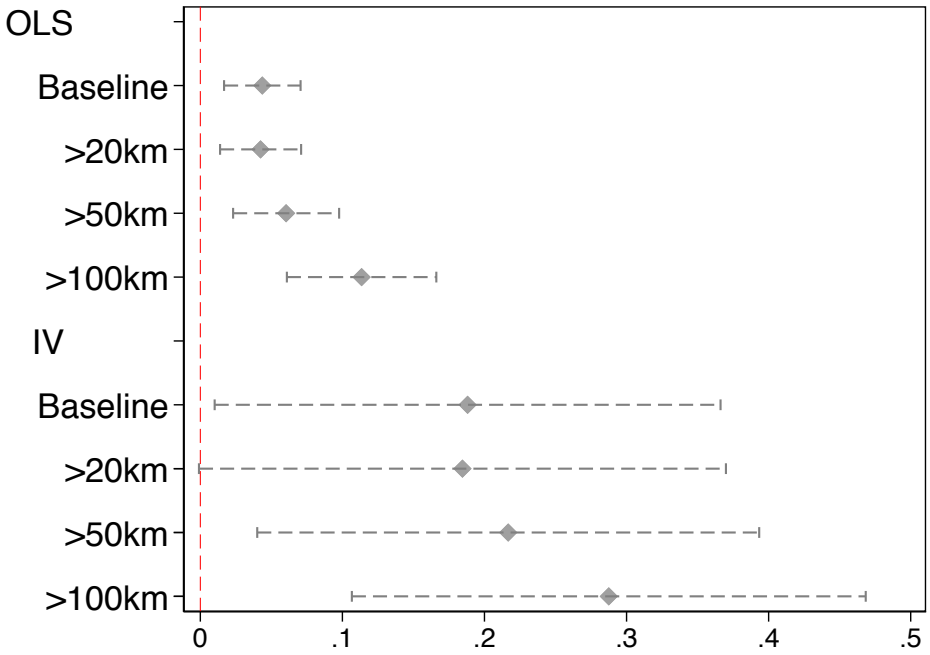
Notes: This figure shows a binned scatterplot of a department's agricultural suitability (x-axis) and the fraction of a department's workforce (males 18 to 60 years old) that was employed in farming or as farm workers in 1869. I divide departments into 20 equally sized bins based on their agricultural suitability index. The y-axis shows the average fraction of the workforce employed in farming or as farm workers within each of the bins. The slope reports the coefficient of an OLS regression (standard error in parentheses). Agricultural suitability is computed as the area weighted average suitability in a department, based on the GIS files described in Cruzate et al. 2012.

Figure B.3: Literacy rate by occupational category



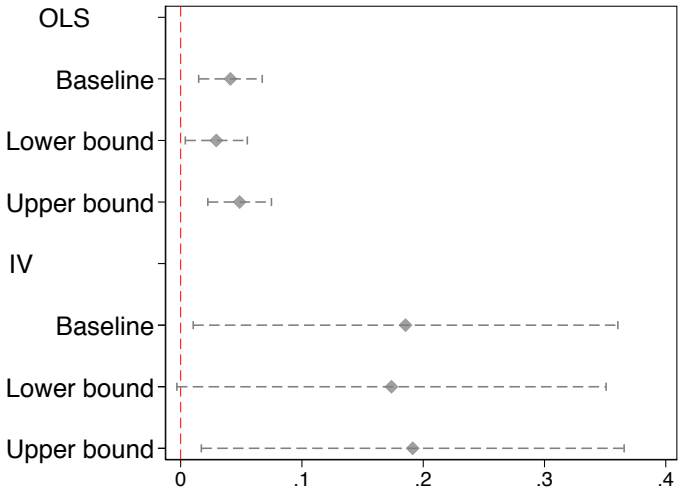
Notes: This figure shows the average literacy rate of fathers (1869) and sons (1895) in the linked sample, by occupational category.

Figure B.4: Spillover effects, probability of exiting farming



Notes: This figure shows the estimated coefficient corresponding to the *Connected* indicator in a regression in which the dependent variable is an indicator that takes a value of one if a son was employed outside of farming by 1895. Each row shows the results of a specification in which the sample is restricted to (1) connected districts or (2) unconnected districts that were at least x kilometers of a connected district.

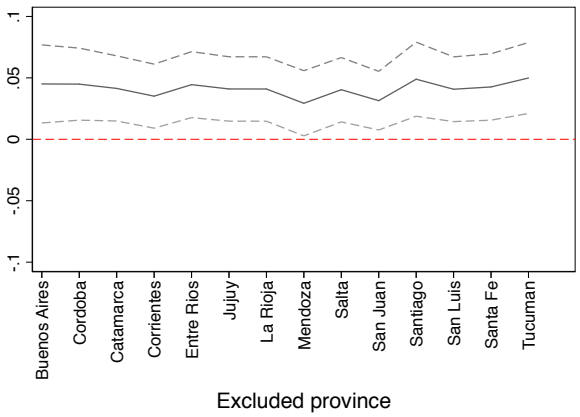
Figure B.5: Lee 2009 bounds, probability of exiting farming



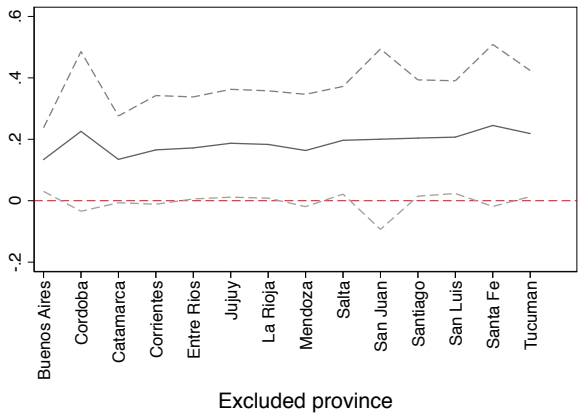
Notes: This figure shows the Lee 2009 bounds of the effects on the probability of exiting farming, after accounting for differential attrition between connected and unconnected districts. See main text for description of the exercise.

Figure B.6: Excluding one province at a time

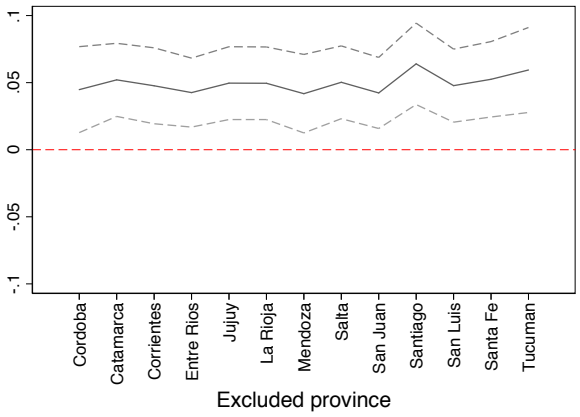
(a) Out of farming, OLS



(b) Out of farming, IV



(c) $\log(Occ.\,Score_{son})$, OLS



(d) $\log(Occ.\,Score_{son})$, IV

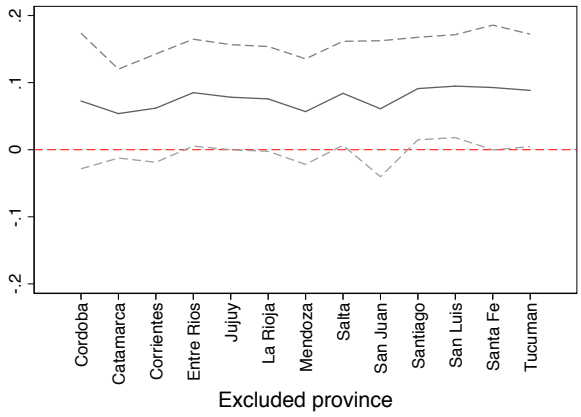
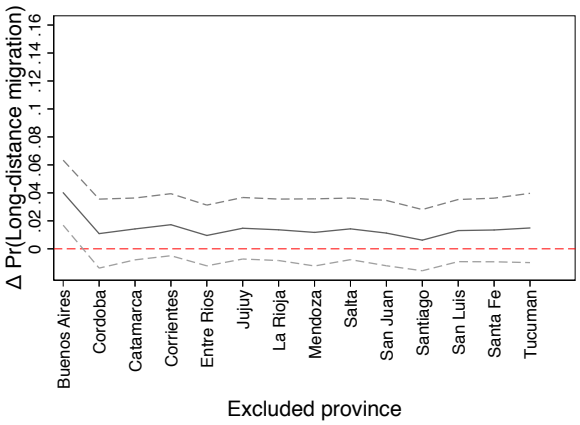
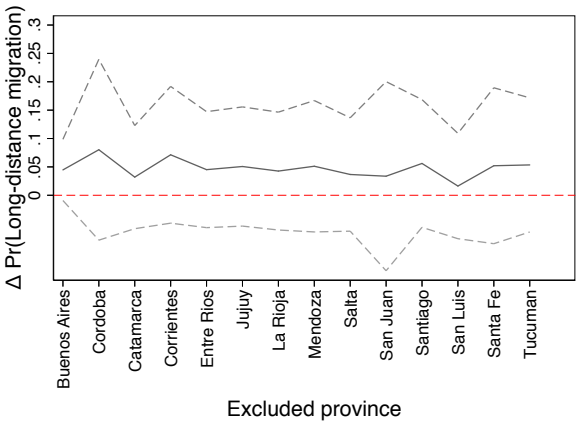


Figure B.7: Excluding one province at a time (cont.)

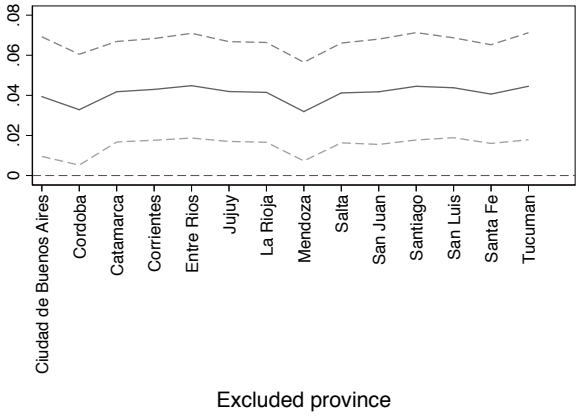
(a) Long-distance migration, OLS



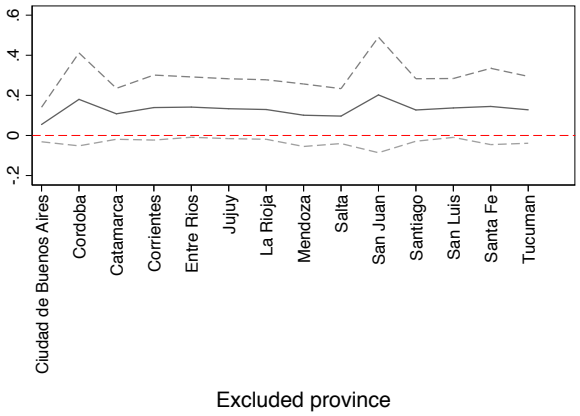
(b) Long-distance migration, IV



(c) Literacy, OLS



(d) Literacy, IV



Notes: In this figure, I show the sensitivity of the results to excluding one province at a time based on an individual place of residence in 1869. In all cases, I show the results for the sample of sons -aged 0 to 16 years old in 1869-.

Table B.1: Railroads and the probability of moving out of farming occupations, by likely ownership status

	Low property		High property	
	(1) OLS	(2) IV	(3) OLS	(4) IV
Connected	0.0332	-0.365	0.0368	-0.0631
	(0.0298)	(0.627)	(0.0277)	(0.165)
Controls	Yes	Yes	Yes	Yes
Observations	2086	2086	2326	2326
Mean of dependent variable	0.310	0.310	0.254	0.254
First-stage F-stat				

Notes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Robust standard errors clustered at the district level in parentheses. All specifications control for a quartic in age, province fixed effects and distance to the nearest targeted city. In the controlled specification, I further control for urban status, literacy and immigrant status. In the "low property" columns, I restrict the sample to adults who likely did not own their farm in 1869. In the "high property" columns, I restrict the sample to adults who were likely to be owners in 1869. See main text for details on this exercise.

Table B.2: Occupational transitions

(a) Intergenerational

Father's occup., 1869	Son's occupation, 1895				
	White collar	Farmer, farm worker	Skilled/semi-skilled	Unskilled urban	Row Total
White collar	**0.34**	0.51	0.12	0.04	1
	(**376**)	(562)	(132)	(42)	(1112)
Farmer, farm worker	0.10	**0.75**	0.11	0.05	1
	(1008)	(**7376**)	(1041)	(449)	(9874)
Skilled/semi-skilled	0.14	0.60	**0.22**	0.05	1
	(212)	(930)	(**338**)	(83)	(1563)
Unskilled urban	0.12	0.62	0.15	**0.10**	1
	(58)	(289)	(71)	(**47**)	(465)
Total	0.13	0.70	0.12	0.05	100
	(1759)	(9673)	(1689)	(664)	(13014)

(b) Intragenerational

Occupation, 1869	Occupation, 1895				
	White collar	Farmer, farm worker	Skilled/semi-skilled	Unskilled urban	Row Total
White collar	**0.42**	0.46	0.08	0.04	1
	(**214**)	(236)	(42)	(19)	(511)
Farmer, farm worker	0.08	**0.82**	0.07	0.03	1
	(252)	(**2587**)	(237)	(92)	(3168)
Skilled/semi-skilled	0.12	0.49	**0.35**	0.05	1
	(79)	(323)	(**232**)	(31)	(665)
Unskilled urban	0.13	0.68	0.12	**0.07**	1
	(37)	(195)	(33)	(**21**)	(286)
Total	0.13	0.72	0.12	0.04	100
	(582)	(3341)	(544)	(163)	(4630)

Notes: Panel (a) shows an intergenerational occupational transition matrix. Rows represent the occupation of the father in 1869. Columns represent the occupation of the son in 1895. Panel (b) shows intragenerational occupational transitions. Rows represent the occupation of an adult in 1869 and columns represent his occupation in 1895. Occupations were classified based on the HISCLASS scheme. White-collar (HISCLASS 1-5), farmer (HISCLASS 8), skilled/semi-skilled (HISCLASS 6-7,9) and unskilled (HISCLASS 10-12) (Leeuwen, Maas, and Miles 2002).

Table B.3: Railroads and the probability of transitioning out of farming occupations, continuous measure

(a) Sons

	OLS		IV	
	(1)	(2)	(3)	(4)
log(Distance to Network)	-0.0140***	-0.0136***	-0.0680	-0.0707
	(0.00399)	(0.00387)	(0.0579)	(0.0584)
Controls	No	Yes	No	Yes
Observations	12412	12412	12412	12412
Mean of dependent variable	0.321	0.321	0.321	0.321
First-stage F-stat				

(b) Adults

	OLS		IV	
	(1)	(2)	(3)	(4)
log(Distance to Network)	-0.0107**	-0.00769	0.0500	0.0264
	(0.00492)	(0.00505)	(0.324)	(0.260)
Controls	No	Yes	No	Yes
Observations	4412	4412	4412	4412
Mean of dependent variable	0.280	0.280	0.280	0.280
First-stage F-stat				

Notes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Robust standard errors clustered at the district level. The dependent variable is an indicator that takes a value of one if the individual worked outside of farming in 1895. All specifications control for province fixed effects, distance to the nearest targeted city and a quartic in age. In the controlled specification in panel (a), I further control for urban status, whether the father was literate and whether the father was foreign born. In the controlled specification in panel (b). I further control for urban status, literacy and immigrant status.

Table B.4: Railroads and the probability of farming employment

(a) Father-Sons, OLS

	(1) Full	(2) Full	(3) High suitability	(4) High suitability	(5) Low suitability	(6) Low suitability
Connected	-0.0655***	-0.0281	-0.00963	-0.00199	-0.154***	-0.0857**
	(0.0213)	(0.0182)	(0.0234)	(0.0204)	(0.0378)	(0.0420)
1869 District FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Region X Year FE	No	Yes	No	Yes	No	Yes
Observations	31104	31104	15704	15704	15400	15400
Mean of dep. variable	0.708	0.709	0.734	0.734	0.684	0.684

(b) Father-Sons. IV

	Full sample		Low suitability		High suitability	
	(1) Full	(2) Full	(3) High suitability	(4) High suitability	(5) Low suitability	(6) Low suitability
Connected	-0.293*	-0.271*	0.921	0.904	-0.193**	-0.168*
	(0.151)	(0.157)	(1.849)	(1.938)	(0.0797)	(0.0893)
1869 District FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Region X Year FE	No	Yes	No	Yes	No	Yes
Observations	24898	24898	12554	12554	12344	12344
Mean of dep. variable	0.731	0.732	0.747	0.747	0.716	0.716

Notes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Robust standard errors clustered at the district level. The dependent variable is an indicator that takes a value of one if the individual worked outside of farming. In panel (b), I instrument railroad access with access to the Euclidean network as described in the main text.

Appendix References

Alsina, Juan A (1898). "La inmigración europea en la República Argentina..." In:

Álvarez, Beatriz and Esteban Alberto Nicolini Esteban A (2010). "Income Inequality in the North-West of Argentina during the first globalization. Methodology and Preliminary Results". In: *Ponencia presentada en II Encuentro Anual de la Asociación Española de Historia Económica. Madrid* 8.

Argentina, Dirección General de Estadistica (1895). *Censo De Los Empleados Administrativos, Funcionarios Judiciales Personal Docente De La República Argentina Correspondiente Al 31 De Diciembre De 1894*. Compañia sud-americana de billetes de banco.

Buchanan, William I (1898). "La moneda y la vida en la República Argentina". In: *Cuadernos del CISH* 3.

Cacopardo, Maria Cristina et al. (1997). "Cuando los hombres estaban ausentes: la familia del interior de la Argentina decimonónica". In: *Poblaciones argentinas: estudios de demografia diferencial. Tandil: PROPIEP*, pp. 13–28.

Conde, Roberto Cortés (1979). "El progreso argentino: 1880-1914". In:

Correa, A and Emilio Lahitte (1898). "Investigación parlamentaria sobre agricultura, ganaderia, industrias derivadas y colonización". In: *Anexo B, Buenos Aires*.

Cruzate, G et al. (2012). *Suelos de la República Argentina 1: 500000 y 1: 1000000*.

Dorfman, Adolfo (1942). *Evolución industrial argentina*. Losada.

Fuente, Diego Gregorio de la (1872). *Primer censo de la República Argentina: Verificado en los dias 15, 16 y 17 de setiembre de 1869*. Impr. del Porvenir.

— (1898). *Segundo censo de la República argentina, mayo 10 de 1895*. Vol. 3. Taller tip. de la Penitenciaria nacional.

King, Miriam L and Diana L Magnuson (1995). "Perspectives on historical US census undercounts". In: *Social Science History*, pp. 455–466.

Knights, Peter R. (1991). "Potholes in the Road of Improvement? Estimating Census Underenumeration by Longitudinal Tracing: U.S. Censuses, 1850-1880". In: *Social Science History* 15.4, pp. 517-526.

Latzina, Francisco (1906). *Diccionario geográfico argentino*. Compañıa Sud-Americana de Billetes de Banco.

Lee, David S (2009). "Training, wages, and sample selection: Estimating sharp bounds on treatment effects". In: *The Review of Economic Studies* 76.3, pp. 1071–1102.

Leeuwen, MHD van, Ineke Maas, and Andrew Miles (2002). *HISCO: Historical international standard classification of occupations*. Leuven: Leuven University Press.

Long, Jason and Joseph Ferrie (2013). "Intergenerational occupational mobility in Great Britain and the United States since 1850". In: *The American Economic Review* 103.4, pp. 1109–1137.

Mill, Roy and Luke CD Stein (2012). *Race, Skin Color, and Economic Outcomes in Early Twentieth-Century America*. Tech. rep. Working Paper, Stanford University December.

Panettieri, José (1965). "Los trabajadores en tiempos de la inmigración masiva en Argentina 1870-1910". PhD thesis. Facultad de Humanidades y Ciencias de la Educación.

— (1998). "El Informe Buchanan: Primer estudio sobre salarios y precios en la Argentina, 1886/1896". In: *Sociohistórica* 3.4.

Provincia de Buenos Aires (1883). "Censo General de la provincia de Buenos Aires, 1881". In: *Buenos Aires*, p. 239.

Ruggles, Steven et al. (1997). *Integrated public use microdata series: Version 2.0*. Historical Census Projects, Department of History, University of Minnesota.

Winkler, William E (1988). "Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage". In: *Proceedings of the Section on Survey Research Methods, American Statistical Association*. Vol. 667, p. 671.

Winkler, William E (1990). "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage." In:

Xie, Yu and Alexandra Killewald (2013). "Intergenerational occupational mobility in Great Britain and the United States since 1850: Comment". In: *The American economic review* 103.5, pp. 2003–2020.