ASSESSING THE IMPACTS OF STATE-SUPPORTED RAIL SERVICES ON LOCAL POPULATION AND EMPLOYMENT: A CALIFORNIA CASE STUDY

Ahmadreza Talebian[a,1], Bo Zou[a], Mark Hansen[b]

[a] Department of Civil and Materials Engineering, University of Illinois at Chicago, United States

[b] Department of Civil and Environmental Engineering, University of California at Berkeley, United States

Abstract: The State of California has been financially supporting Amtrak intercity passenger rail services since 1976. This paper studies the impacts of this support on local population and employment at both county and city levels. We use datasets which include geographic, transport, and socioeconomic characteristics of California counties and cities from 1950 to 2010. Propensity score, one-to-one matching models are employed to draw units from the control group, which are counties/cities that do not have a state-supported Amtrak station, to match with units from the treatment group, which are counties/cities that do. Using regression analysis, we find that state-support Amtrak stations have significant effect on local population in the long term, and the effect increases with time. However, the effect on civilian employment is almost non-existent. This suggests that state-supported Amtrak services can provide quality rail mobility and accessibility, which attract people to live in a rail-accessible region. However, the economic influence seems limited.

Key words: Amtrak station, state-support rail services, California, multivariate normal imputation, matching, regression

1 Introduction

Passenger rail transportation in the US has experienced major upheavals in the past century. Until about 1920, intercity travel in the US had been completely dominated by rail transportation. The services were historically provided by private freight railroads that owned and maintained rail tracks and managed the operations. From 1920, rail ridership started to diminish, and this trend continued until 1934, mainly due to the rise of automobiles and increased popularity of intercity bus services (Thompson, 1993). Although railroads enhanced services in the mid-1930s with new diesel-powered streamliners, rail ridership decline continued. The share of rail transport in total passenger miles decreased to 67% in 1940, and further to only 15% in 1965. In the late 1960s, most of the passenger rail services were not able to break even, and some major rail companies became insolvent. The US federal government ultimately stepped in 1970 and President Richard Nixon signed the Rail Passenger Service Act, based on which the National Railroad Passenger Corporation, known as Amtrak, was formed to take over passenger rail operations on May 1, 1971. The total number of

[1] Corresponding author.

Email addresses: ataleb2@uic.edu (A. Talebian), bzou@uic.edu (B. Zou), mhansen@ce.berkeley.edu (M. Hansen).

services was pruned from 364 to 182. Since 1971, Amtrak has been the only provider of intercity passenger rail services in the United States (Nice, 1998; Pan, 2010).

As in other states, Amtrak discontinued multiple rail services in the State of California in 1971, including Redwood, Sacramento daylight, Jan Joaquin Daylight, San Francisco Chief, El Capitan, and Del Monte. On the other hand, to foster intercity passenger rail services, the California Department of Transportation (Caltrans) has been providing Amtrak with financial support since 1976, which has helped Amtrak initiate new services, extend existing services, and improve service quality. However, the impact of this support on regional economic development is not yet well known. To fill this gap, this paper employs propensity score matching and regression modeling to study the decade-by-decade effects of state-supported Amtrak services on population and employment of California counties and cities.

Studying the economic impacts of public capital has been of interest to the academic community for an extended period of time. By and large, public capital significantly stimulates the economic growth of a region (Munnell and Cook, 1990). Aschauer (1989) investigates the relationship between aggregate productivity and public/private capital stock. He shows that core infrastructures, including highways, mass transit, airports, etc., account for 55% of aggregate productivity; whereas the total share of hospitals, office buildings, courthouses, garages, etc. holds no more than 10%.

Many researchers have investigated the relationship between highways and economic development (e.g. Banerjee et al. (2012), Duranton and Turner (2012), Duranton et al. (2014), Faber (2014), Garcia-Milà and Montalvo (2007), and Gibbons et al. (2012)). Baum-Snow (2007) considers the 1947 US highway plan as an instrument and develops regression models to understand how construction of new, limited-access highways has influenced central city populations between 1950 and 1990. The study finds that central city population in each metropolitan statistical area was reduced by about 18% if a new highway passed through a city. However, population would increase by 8% should the highway be absent. In a similar study, Michaels (2008) investigates the impacts on domestic trade of the US interstate highway system. Highways are found to significantly impact the demand for highly-skilled, nonproduction workers in counties. Chi (2010) studies the relationships between interstate highway expansion and population change in the 1980s and 1990s in Wisconsin. Two effects of economic growth are recognized: spreading and backwash effects. However, as argued by the authors caveats should be exerted when estimating highway impacts. Population growth in one location could lead to population decline in the surrounding areas.

In the aviation arena, it is widely believed that air transport services, by connecting urban regions, attract new business activities, thereby stimulating local population and economic development. By developing instrumental variable regression models, Brueckner (2003) finds that 10% increase in passenger enplanements elevates employment in service-related industries by about 1%. He finds no significant effect of airline traffic on manufacturing and other goods-related employment. Green (2007) develops instrumental variable regression models to study the impacts of airports on regional growth. Different measures of airport activity, including boardings, originations, hub status, and cargo volume are investigated. The author concludes that passenger activity is a statistically significant predictor of regional growth; whereas cargo activity is not. The results indicate that increasing boardings per capita by one standard deviation will result in 8% increase in regional population in a decade. To investigate how small- and mid-size commercial airports affect local economies over the post-World War II period, McGraw (2014) develops instrumental variable regression models, and finds that existence of an airport in a Core Based Statistical Area (CBSA) results in 14.6% to 29% population growth, and 17.4% to 36.6% total employment growth. In addition, airports impact tradable industry employment more than non-tradable industry employment. Other insights about the relationship between airports and economic development are obtained in Percoco (2010), Mukkala and Tervo (2013), Cidell (2014), Sheard (2014), and Blonigen and Cristea (2015).

Because of the long-standing position of rail in the transportation system, a large body of the literature exists on assessing the economic impacts of rail transport. Building on the general equilibrium trade theory, Donaldson and Hornbeck (2013) study how railroads have influenced America's economic growth. In the study, a change in "market access" represents aggregate impact of a change in the rail network. Removing all railroads is found to reduce average market access of counties by 63%, which in turn would decrease gross national product by 6.3%. The authors find that rail access has small positive impact on population density and boosts urbanization. On the grounds that Swedish railroads have been extended quasi-randomly, Berger and Enflo (2014) use two-stages least squares (2SLS) and limited information maximum likelihood (LIML) methods to estimate the extent to which railroads contributed to town-level growth over the last 150 years. Compared to cities with no access to rail, towns with rail access experienced large population increase in the short run. Population further spills overs to nearby towns. However, the relative differences in population among towns is largely stable in the long term despite continuous expansion of the rail network. Hornung (2012) studies the causal effects of rail station access in the German state of Prussia during the 1840-1871 period using instrumental variable and fixed-effects estimation techniques. Urban population growth is considered as a proxy for economic growth, and it is found that economic growth of cities with rail access is roughly 1-2% greater than cities with no rail access. Gregory and Henneberg (2010) examine whether acquisition of a rail station had significantly driven population growth in England and Wales parishes in the pre-World War I period. They find that parishes with a station early grew substantially faster than those without. Parishes gaining a station earlier had faster growth rates than gaining a station later.

For more recent passenger rail systems, Wang and Wu (2015) apply the difference-in-difference method to estimating local economic impacts of China's Qinghai-Tibet rail line. Results indicate that the rail line stimulates annual GDP per capita by about 33%. The impact is focused on manufacturing, with almost no effect on agriculture and service industries. Nordstrom (2015) uses ridership data, surveys, corridor development information, and property value assessment to explore the role and impact of commuter rail on local geography. Elkind et al. (2015) study and grade the neighborhood within 1/2 -mile radius of 489 existing stations in 6 district California rail transit systems. Sperry et al. (2013) investigate the economic impact of the Michigan Amtrak service including traveler savings, passenger spending at local businesses, and Amtrak-related expenditures in 22 communities. For further understanding of the economic impacts of rail transport, readers may refer to Atack and Margo (2009), Atack et al. (2010), Franch et al. (2013), Koopmans et al. (2012), Pereira et al. (2015), and van den Heuvel et al., (2014).

Despite the rich literature on estimating the economic impacts of rail transport, no effort is made to investigate how the state-level support of Amtrak services affects regional socioeconomic characteristics. In this paper, we make the first attempt to fill this gap. We use historic data from California to empirically investigate to what extent the presence of an Amtrak station(s) in a county or a city affects population and civilian employment of the county/city. In investigating this plausible causal relationship, two challenges need to be overcome. First, the dataset required for causal inference includes missing values, which is an important issue as the number of observations (which correspond to counties or cities) in our study is limited. Second, rail services, like other transportation services, are not randomly assigned to counties and cities (McGraw, 2014). Characteristics of the counties/cities with rail services may differ systematically from those without. As a consequence, estimating the economic effects of rail services using regression would yield biased results if no adjustment is made. To tackle these challenges, we employ the multivariate normal imputation method to fill in the missing values in the dataset. One-to-one propensity score matching models are employed to match counties/cities without rail services with counties/cities with rail services. We then perform ordinary least squares (OLS) regressions to quantify the impacts on local population and employment of state-supported Amtrak services. Figure 1 illustrates the modeling framework.

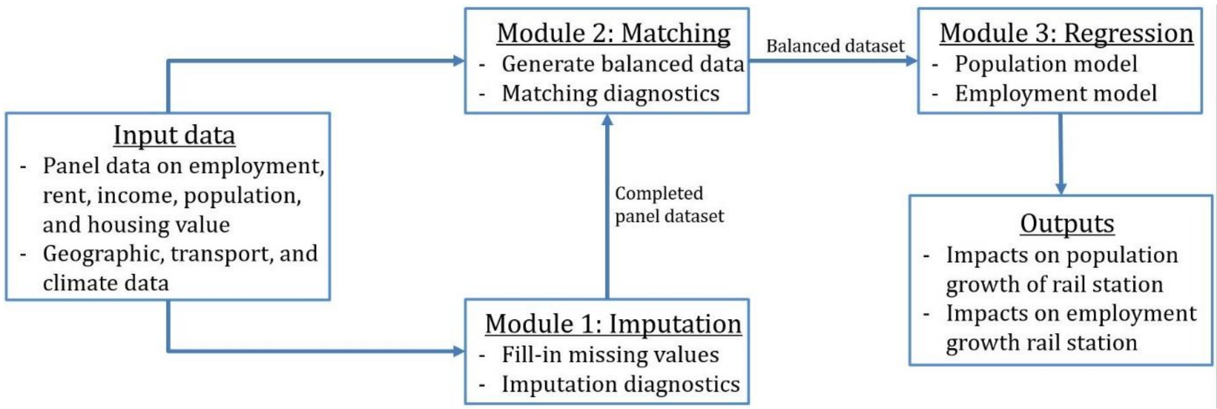

Figure 1: Modelling framework for assessing the economic impacts of state-supported rail services

The reminder of the paper proceeds as follows. In Section 2, we provide details on data preparation. Section 3 is dedicated to describing the theoretical background and results of multivariate normal imputation. Section 4 discusses on the principles of the causal inference framework and the matching models used. Section 5 presents the OLS estimation of the impacts on population and employment of rail station access. Summary of major findings and directions for future research are given in Section 6.

2 Data preparation

2.1 State-supported rail services in California

Currently, Caltrans provides financial support for three Amtrak rail corridors in California: Pacific Surfliner, San Joaquin, and Capitol Corridor (Figure 2). The length, number of stations, ridership, and on-time performance of each corridor are presented in Table 1. The Pacific Surfliner serves the coastal strip between San Diego and San Luis Obispo. The portion connecting San Diego to Los Angeles has been in place since 1938 under the *San Diegan* brand. The service extended to Santa Barbara in 1988 and to San Luis Obispo in 1995. Later in 2000, the service was renamed Pacific Surfliner. The Capitol Corridor connects San Jose to Auburn. The portion between Martinez and Sacramento was served by the Southern Pacific's Senator service until 1962. In 1990, California passed two propositions to support resurging rail services along this corridor. As a result, Capitol Corridor service began a year after that. Previously, the San Joaquin Daylight served Los Angeles-Oakland Pier corridor from 1941 to 1971. The new San Joaquin service debuted in 1974, and has been receiving state funding support since 1979.

Figure 2: California intercity passenger rail services (source: AECOM (2013))

Table 1**: On-time performance, the number of stations served, line mileage, and ridership of the three state-supported Amtrak services in California**

Line	On-time Performance	Num. of Station	Mileage	Ridership (in thousand passengers)				
				2002	2005	2008	2011	2014
Pacific Surfliner	73.2%	31	350	1725	2625	2835	2786	2681
Capitol Corridor	93.6%	17	168	1080	1260	1694	1708,	1419
San Joaquin	76.1%	18	282 (Sacramento)	733	743	894	1067	1188
			315 (Oakland)					

Note: On-time performance is for August 2015.
Data sources: http://www.amtrak.com/, http://www.dot.ca.gov/, http://www.capitolcorridor.org/, and https://www.acerail.com

2.2 Data sources

2.2.1 County-level data

A county-level dataset is compiled which contains socioeconomic, demographic, geographic, and transportation information for California counties in 1950-2010. The dataset includes the following items:

1- The 2010 geographic boundaries of 58 California counties, obtained from National Historical Geographic Information System (NHGIS) (MPC, 2011).

2- Total highway mileage in the National Highway System (NHS mileage) for each county (NTAD, 2015).

3- List of ports in California obtained from NOAA (2000). Counties having port(s) are identified with a dummy variable.

4- List of public-use airports in each county, based on NTAD (2015). For an airport to be listed, it should have a control tower and non-zero aircraft operations. Similar to item 3, counties having airport(s) are identified with a dummy variable.

5- List of coastal counties, based on NOAA. According to NOAA, a county meeting one of the following two criteria is viewed as a coastal county: 1) at least 15% of a county's land is located within the Nation's coastal watershed; 2) a portion of a county accounts for at least 15% of a coastal cataloging unit" (NOAA, 2012).

6- Amtrak state-supported routes and station locations, obtained from Caltrans (Caltrans, 2015). Based on the location information, we construct a dummy variable for each county which takes value 1 for having at least one such station in the county. In total 20 counties have value 1 for this dummy variable.

7- Commuter rail service network, obtained from Caltrans (2015). This network includes Altamont Corridor Express (ACE), Caltrain, Coaster, Metro Blue, Gold & Green Line, METROLINK, and BART. We consider a dummy variable for commuter rail services. This variable is equal to 1 if a county has at least one station served by commuter rail service(s), and 0 otherwise.

8- Freight rail network, obtained from the Oak Ridge National Lab (CTA, 2003). Rail network mileage information is aggregated to the county level.

9- County characteristics, collected from County Characteristics 2000-2007 (ICPSR, 2015b). They include mean January temperature (Jan. temp.), mean July temperature (Jul. temp.), land area, and water area.

10- County population data (Pop) for 1950-2010, obtained from NHGIS.

11- Median family income (Income), median gross rent (GRent), and median housing value (Housing) for 1950-1970 and 1980-2010, obtained from CCDB (ICPSR, 2015a) and NHGIS, respectively.

12- Civilian employment data (Civilian), obtained from NHGIS for 1970-2010 and from CCDB for 1960-1970.

Overall, the total number of entries in the panel data set on population, housing value, gross rent, income, and civilian employment from 1950 to 2010 in a 10-year increment across 58 counties is 2030 (5 variables × 7 years × 58 counties). Among these values, seven are missing. To fill in the missing entries, we employ the multivariate normal imputation method (Section 3), which is shown to perform better than other missing-value imputation methods such as complete-case analysis and ad-hoc mean imputation (King et al., 2001; Lee and Carlin, 2010).

2.2.2 City-level data

The city-level dataset is also compiled which includes:

1- The 2014 geographic boundaries of 482 California cities, obtained from Caltrans (2015). Due to data limitation, we only consider 84 cities which continuously have had a population greater than 25,000 since 1960. To reduce data heterogeneity, the two largest cities, Los Angeles and San Diego with more than 1 million population in 2005, are removed. This leaves 82 cities in the dataset.

2- Amtrak station locations, obtained again from Caltrans (2015). Following Murakami and Cervero (2010), we assume that a rail station impacts a circular area with a 5-km radius. Using ArcGIS, we identify cities whose jurisdiction is within the 5-km radius of an Amtrak station on a state-supported rail route. By doing so, 90 cities are identified. Among these, only 26 cities are among the cities with recorded population data continuously in 1960-2005. Therefore, the rail service dummy variable for a city equals 1 if the city jurisdiction intersects a circular area with a 5-km radius around an Amtrak station on a state-supported rail route.

3- Commuter rail service network, obtained from Caltrans (2015). Similar to the county level data set, the value of the commuter rail variable is 1 if a city has at least one station on a commuter rail line, and 0 otherwise.

4- Presence of a reachable airport for a city. According to Lieshout (2012) and Marcucci and Gatta (2011), the catchment area of an airport is any place within 2-hour driving by car. Assuming an average car speed of 25 mph, it results in all 82 cities being within 50 miles from an airport (we also experiment with a much smaller, 15-mile radius for defining reachable airports. In this case, still, 78 cities will have a reachable airport). Therefore, airport catchment area is not used. The value of the airport dummy variable of a city equals 1 if the city has airport(s) within its jurisdiction, and 0 otherwise.

5- Total highway mileage in the National Highway System (NHS mileage) for each city (NTAD, 2015).

6- List of ports, based on NOAA (2000). We consider a dummy variable for having port(s) in a city.

7- Freight rail network data, obtained from the Oak Ridge National Lab (CTA, 2003). Rail network mileage information (Rail mileage) is aggregated to the city level.

8- City characteristics, including population (Pop), civilian employment (Civilian), median gross rent (GRent), median family income (Income), and median housing value (Housing) for 1960, 1970, 1980, 1990, 2000, and 2005. The information is collected from County and City Data Book (CCDB) series (US Census Bureau, 2010; ICPSR, 2015a).

9- Land area and climate data, including mean January and July temperatures (Jan. temp. and Jul. temp.), and annual precipitation (Ann. prec.). The information is collected from 2007 CCDB (US Census Bureau, 2010).

Table 2 presents descriptive statistics of the city-level data. Note that the coefficients of variation for population and civilian employment vary between 0.99 and 2.01; whereas for income, rent, and housing value the coefficients of variation are between 0.13 and 0.31. This suggests that population and income are more dispersed across cities than rent, income, and housing value.

Table 2: Descriptive statistics of the city-level dataset

Variable	Mean	Std. Dev.	Min	Max	Variable	Mean	Std. Dev.	Min	Max
Land Area (sq miles)	25.4	28.5	3	174.9	Civilian1960	29958	43276	8491	352858
Jan. temp. (°F)	53.9	3.9	46	58.8	Civilian1970	39551	44134	11199	340075
Jul. temp. (°F)	71.2	5	57.3	83.1	Civilian1980	51954	55156	13390	364689
Ann. prec. (inches)	16.1	4.6	6.5	31	Civilian1990	65966	68717	15086	434202

NHS mileage (miles)	38.4	40.8	7.8	234.9	Civilian2000	72008	78138	16890	489677
Rail mileage (miles)	8.4	10.3	0	46.8	Civilian2005	73234	72807	15392	430431
Pop1960	46802	94182	25136	740316	Housing1960 ($)	16095	3773	10900	35000
Pop1970	93635	99800	26826	715674	Housing1970 ($)	24672	7549	15409	71336
Pop1980	103642	106596	26438	678974	GRent1960 ($)	88	13	55	122
Pop1990	126404	127866	28696	783233	GRent1970 ($)	137	22	93	190
Pop2000	141270	143718	30785	895279	Income1960 ($)	7221	939	5292	11977
Pop2005	146161	146674	29661	912332	Income1970 ($)	11331	1859	8029	20303

2.3 Treatment group vs. control group

In a comparative experiment whose aim is to assess the outcome of a treatment program, study units are categorized into two groups: treatment and control. The units which receive the treatment form the treatment group; the control group is made of non-participants of the treatment program. In our study, the treated units are cities/counties which have at least one station served by state-supported passenger rail services. Cities/counties with no intercity passenger rail services or with long-distance Amtrak rail services form the control group. At the city level, 26 cities fall in the treatment group and the remaining 56 cities form the control group. At the county level, 20 counties take advantage of state-supported rail services. As mentioned before, the two counties encompassing the cities of Los Angeles and San Diego are removed to make the county-level dataset consistent with that at city-level. In total 18 counties are in the treatment group and 38 counties in the control group. Since the State of California began to support rail services in 1976, we consider 1950-1970 as the pre-treatment period.

3 Imputation

Recall that some missing values appear in the county-level dataset. This section deals with imputing these missing values. To impute (or fill in, or rectangularize) our county-level dataset, we use the multivariate normal imputation method (Honaker and King, 2010; Honaker et al., 2011). The basic idea of the multivariate normal imputation method is that the distribution of the dataset, including both observed and missing entries, is multivariate normal. In this study, we implement the multivariate normal imputation method using the Amelia II package in the statistical software R. We impute m values for each missing entry in the data set, thus generating m complete datasets in which observed values are the same but missing entries are completed. Then, we take the mean of the m datasets as the final filled-in dataset for subsequent matching. For theoretical background of multivariate normal imputation, readers are referred to Honaker et al. (2011). Our county-level dataset has 5 missing values, including 2 rent values, 2 income values, and 3 housing values. For the choice of m value, m=5 will be sufficient as long as the percentage of missing observations is not very high (Honaker et al., 2011). We use m=10 in this study. To enhance the prediction power, we also include population and civilian employment variables. To further improve imputation, the following strategies are adopted:

1- **Log transformation:** Kolmogorov-Smirnov and Shapiro-Wilk tests show that distributions of the available entries for population and civilian employment are not normal; whereas the common logarithm (i.e., with base 10) of each variable is normally distributed. Therefore, we use the common logarithm of population and civilian employment.

2- **Time-series data:** Amelia is capable of identifying time-series patterns of observed data. By estimating a sequence of polynomials of the time index, the package creates a model of patterns within variables across time. Amelia then adds the generated patterns as new covariates to the existing dataset and conducts imputation. The new covariates improve predictability of the imputation model. We will take advantage of this ability in our imputation.

3- **Logical bounds:** for each missing value, we provide appropriate lower and upper bounds. For example, the lower and upper bounds for income in a county are set to 0 and income of the next decade. Although sometimes these bounds are not tight, they improve the imputation.

To check the quality of the missing value imputation, we conduct overimputation. In this process, we sequentially assume that one of the observed entries is missing and perform imputation. Imputed values are then plotted against observed values. Imputation would be perfect if all points lie on the 45° line of the imputed value – observed value plot. For each observed entry, we impute a large number of values, based on which we construct a 90% confidence interval. Figure 3 shows the confidence intervals on the imputed value – observed value plots for housing value, family income, and gross rent. Most of the intervals are centered around the 45° line, indicating that the multivariate normal imputation method performs well in filling-in missing values for all four variables. Note that a red confidence interval in the plots (i.e., at the lower end of the family income and gross rent plots) indicates that the corresponding observation has fewer covariates available for imputation, resulting in greater variance of the imputed values.

The descriptive statistics of the complete county-level dataset are presented in Table 3. We find that values for the population and income variables are more dispersed across counties, with coefficients of variation between 1.3 and 1.6. For income, rent, and housing value, the coefficients of variation are between 0.13 and 0.3. This is in line with the coefficient of variation values at the city level and thus a further validation of the imputed values.

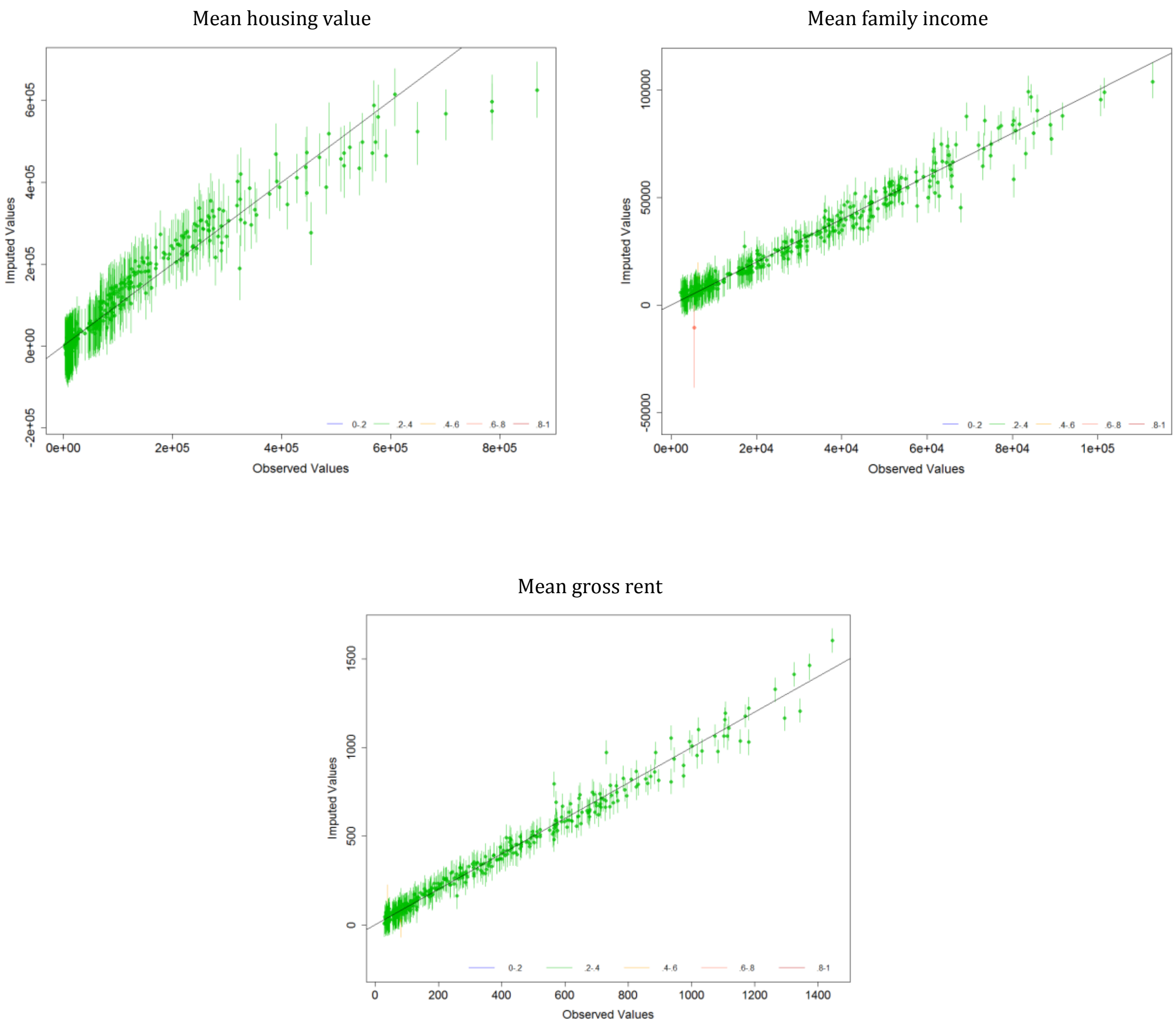

Figure 3: Overimputation diagnostic graphs

Table 3: Descriptive statistics of the completed county-level dataset

Variable	Mean	Std. dev.	Min	Max	Variable	Mean	Std. dev.	Min	Max
Land area (sq miles)	2637.4	3147.2	46.7	20052.5	NHS mileage (miles)	295.3	270.8	64.4	1594.7
Water area (sq miles)	120	177.1	0.5	1052.1	Housing 1950 ($)	7072	2171	3428	12547
Water pct. (%)	7.4	12.8	0.1	75	Housing 1960 ($)	12098	2804	7200	20200
Jan. temp. (°F)	44.1	6.7	28.6	54.2	Housing 1970 ($)	18562	5055	11227	33852
Jul. temp (°F)	71.6	7.6	56.3	92.0	Income 1950 ($)	3263	432	2256	4467
Pop. 1950	104959	154441	241	775357	Income 1960 ($)	5927	782	4438	8110
Pop. 1960	154383	212749	397	908209	Income 1970 ($)	9372	1459	6551	13931
Pop. 1970	206486	303310	484	1420386	Civilian 1950	41261	66285	86	359060
Pop. 1980	255867	371579	1097	1932709	Civilian 1960	58614	85094	129	360427
Pop. 1990	328551	472601	1113	2410556	Civilian 1970	80685	124350	217	575570
Pop. 2000	384616	557150	1208	2846289	Civilian 1980	122828	190468	612	1016754
Pop. 2010	434644	629605	1175	3010232	Civilian 1990	164905	253088	591	1357847
GRent 1950 ($)	39	6	16	56.26	Civilian 2000	182177	270295	683	1409897
GRent 1960 ($)	70	12	33	107	Civilian 2010	216061	319657	604	1592219
GRent 1970 ($)	109	21	73	172					

Figure 4 illustrates how county-level population, civilian employment over population ratio, income, and housing value, by control vs. treatment group, have evolved between 1950 and 2010. Population, income, and housing value have steadily increased since 1950. Compared to counties with no state-supported Amtrak services (i.e., counties in the control group), average population, income, and housing value are always greater for counties having state-supported services (i.e., counties in the treatment group). This difference is most significant for population. We further observe a diverging trend over time between the treatment and control curves in population, income, and housing values. The lower-right panel shows the county-average value of civil employment over population. For both control and treatment groups, this ratio slightly drops from 1950 to 1960. Greater ratio mostly appears in the treated counties in the post treatment period.

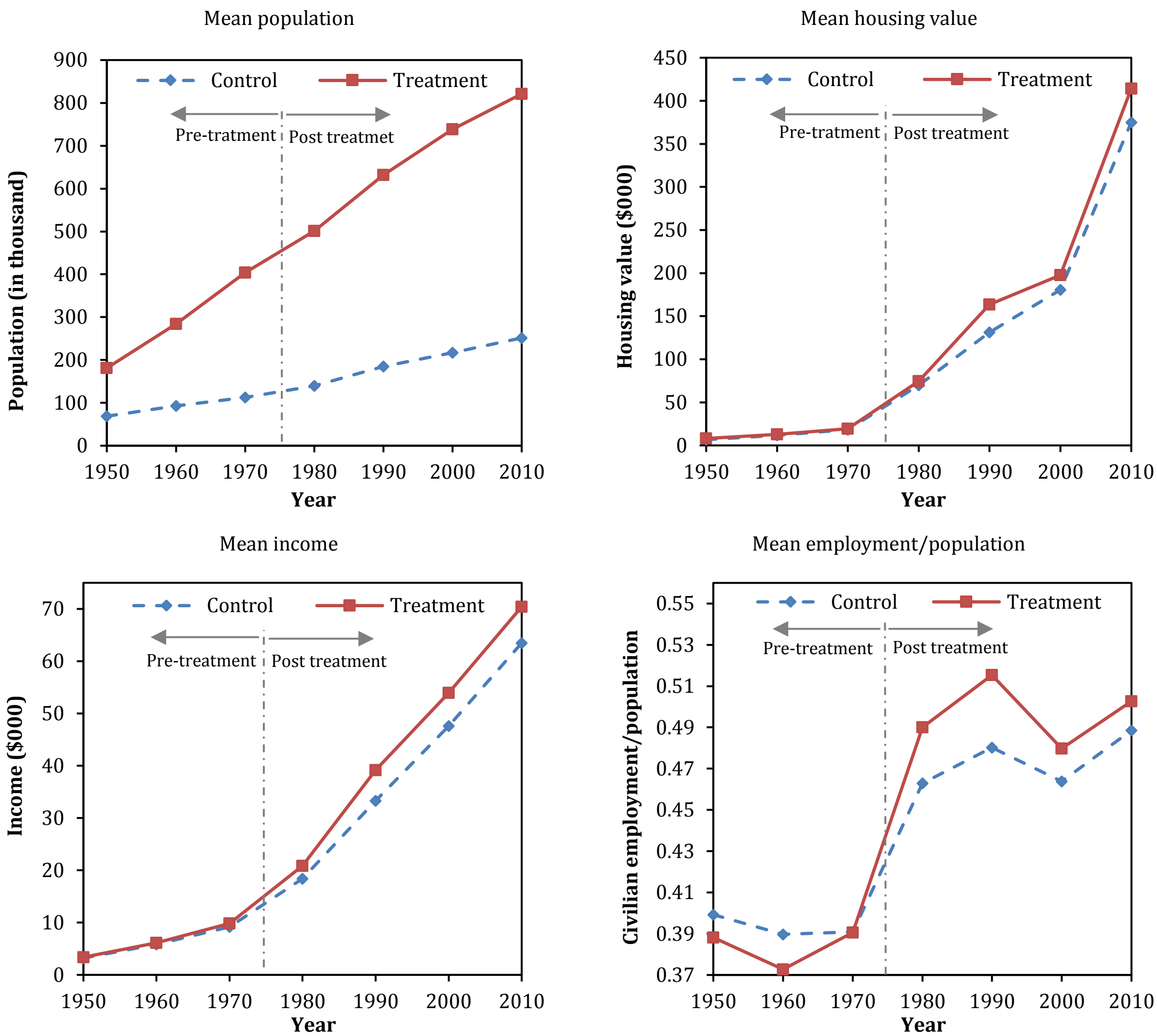

Figure 4: Illustration of population, income, civilian employment/population ratio, and housing value evolution in 1950-2010

4 Causal inference

In general, causal inference in a randomized investigation, in which treatment is randomly assigned to units, is straightforward. Estimation of the treatment effects is given by the difference in the outcome of interest between treated and control units. Unfortunately, Amtrak stations are not randomly assigned to counties and cities. In this case, baseline characteristics, known as baseline covariates, of the treatment group can systematically differ from those of the control group. One approach to correct for this systematic difference in baseline covariates is to use a matching model. In this study, we employ one-to-one propensity score matching methods to draw units from the control group and match them with units from the treatment group. Below we start with the Rubin's causal inference model (Rubin, 1973; 1974) and the propensity score matching method. We then present the matching results.

4.1 Theoretical background

Let Y_{i1} denote the outcome for unit *i* if the unit receives treatment, i.e., it is in the treatment group; Y_{i0} the outcome for unit *i* if it is not treated, i.e., it is in the control group. The effect of treatment for unit *i* is then $\tau_i = Y_{i1} - Y_{i0}$. Note that only one of Y_{i1} and Y_{i0} is observed in reality. Let the treatment indicator T_i be 1 if unit *i* receives treatment and 0 otherwise. Then, the observed outcome of interest for unit *i* is $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$. Assuming that the treatment assignment T_i is independent of Y_{i0} and Y_{i1}, the average treatment effect (ATE), τ, is estimated as:

$$\tau = E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 0) = E(Y_i|T_i = 1) - E(Y_i|T_i = 0) \tag{1}$$

ATE is in fact the average effect of assigning treatment to the entire population. Another quantity of interest is the average effect of treatment on the subjects who receive treatment (ATT), $\tau|(T = 1)$, which is estimated by:

$$\tau|(T = 1) = E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 1) \tag{2}$$

In a randomized experiment, ATT and ATE are the same, because treated units are randomly selected and not systematically different from control units. In an observational experiment, ATT cannot be estimated because Y_{i0} is not measured for treated units. Based on strong ignorability assumption that there is no difference between treated and non-treated units conditional on the observed covariates, Rubin (1974) shows that

$$\tau|(T = 1) = E\{E(Y_i|X_i, T_i = 1) - E(Y_i|X_i, T_i = 0) \mid T_i = 1\} \tag{3}$$

where X is the set of covariates, and the outer expectation is taken over the distribution of baseline covariates in the treatment group ($X_i| \, T_i = 1$). To condition on X, the most straightforward, nonparametric method is to match on covariates (Sekhon, 2011). The ultimate goal of this matching is to find a dataset consisting of a set of observations for which the mean of treated units is close to that of non-treated units across all covariates. Such dataset is defined as a balanced dataset.

In this paper, we use two variants of the one-to-one nearest neighbor matching model: optimal matching and caliper matching. The general idea of nearest neighbor matching can be described as pairing each observation in the treatment group with an observation from the control group which has the lowest distance from the treated unit. Simple nearest neighborhood matching is not considered, because the order of matching treated units may change the overall distance between units in the control and treatment groups. To overcome this, we use optimal matching, a variant of the nearest neighbor matching (Rosenbaum, 1989), in which a global distance measure is minimized when choosing individual matches.

Any matching method requires a distance model to measure the closeness between each pair of observations. In this study, we employ propensity score to quantify the distance between two observations:

$$D_{ij} = e_i - e_j \tag{4}$$

where D_{ij} is the distance between observation *i* and observation *j*; and e_k $(k = i, j)$ is the propensity score for observation k. Propensity score for observation k is the probability of being treated: $e_k(X_k) = Pr(T_k = 1|X_k)$. In fact, propensity score summarizes the values of all covariates for observation *k* (i.e., X_k) into one scalar (e_k), i.e., the probability of being treated. If observation *j* (from the control group) is perfectly matched with observation *i* (from the treatment group), it means that $e_j = e_i$. Any model relating a binary variable, i.e., the variable indicating if an observation received treatment (T), to a set of covariates (X) can be employed to estimate propensity score. In this study, we use logistic regression to estimate propensity scores.

When optimal matching results in poor matches, i.e., large distances between observations of each matched pair, one remedy is to impose a caliper. In doing so, we only select a matched pair if it is within a pre-specified caliper:

$$\left|e_i - e_j\right| \leq \delta \tag{5}$$

where δ is the width of the caliper. This width is usually set to a multiplier of standard deviation of propensity scores across all observations. In this study, we use caliper matching with a 0.3 standard deviation of propensity score used for city-level data. At the county level, however, we still stick to optimal matching because imposing a caliper excludes most observations and leaves very few match pairs for regression.

To examine how well the matching model balances the treatment and control groups, we use the standardized difference of means (SDM):

$$SDM = \frac{\bar{X}_t - \bar{X}_c}{\sigma_t} \tag{6}$$

where, $\bar{X}_t$ is the mean of covariates in the treatment group; $\bar{X}_c$ the mean of covariates in the control group; and σ_t the standard deviation of covariates in the treatment group. We compute standardized difference of means before and after matching. Percentage of balance improvement (PBI) is then defined as:

$$PBI = \frac{\left|SDM_{Before}\right| - \left|SDM_{After}\right|}{\left|SDM_{Before}\right|} * 100\% \tag{7}$$

4.2 Results

We use the MatchIt package in R to implement optimal and caliper matching models (Ho et al., 2007). Recall that the State of California began to support rail services in 1976, we consider 1950-1970 as the pre-treatment period. The first step in implementing a matching model is to determine what variables to be included. The key concept in this inclusion is strong ignorability assumption. To best satisfy this assumption, it is suggested that analysts be liberal in including potentially associated variables and add as many as possible variables which have some relevance to both the treatment assignment and the outcome of interest (Stuart, 2010). This is because when using a propensity score matching model, inclusion of a variable that is not so relevant to the outcome variable may only slightly increase the variance of other covariates among matched control units. But ignoring a variable that has close relevance to the outcome variable can substantially elevate the bias (Stuart, 2010).

At the county level, we match on the natural logarithm of population in 1950-1970, housing value in 1950-1970, percentage of land in total area, the presence of a reachable airport, the presence of a port, mean January temperature, and NHS mileage. To account for population and housing growth rates (i.e., 1950-1960 and 1960-1970), we follow McGraw (2014) and incorporate interaction terms into the matching models. In generating the matching model for total civilian employment, log of civilian employment replaces log of population. This set of covariates covers a wide range of county socioeconomic, geographic, climate, and transportation characteristics. At the same time, it results in the highest possible balance improvement across most covariates.

Table 4 documents standardized difference of means before and after matching at the county level. For population matching, we observe that the model substantially improves the balance across all the variables. On average, matching improves the balance across the variables by about 63%. For civilian employment matching model, log of NHS mileage results in very poor balance; thus we match on NHS mileage. The highest balance improvements are achieved for housing variables. The average balance improvement across all variables is by about 50.3%.

Table 5 shows standardized difference of means before and after matching at the city level. We match on pre-treatment population, mean January temperature, NHS mileage, land area, total freight railroad mileage, the presence of a reachable airport, median gross rent, and median family income. We prefer median gross rent to mean housing value as a covariate in matching because median gross rent results in greater balance improvements. We include median family income so that the number of covariates of city-level matching is comparable to county-level matching. On average, matching improves SDM of population and civilian employment by 73% and 70% respectively.

Table 4: Standardized difference of means before and after matching at the county level

Population matching				Civilian employment matching			
Variable	Std. Mean Diff.		Bal. imp. (%)	Variable	Std. Mean Diff.		Bal. imp. (%)
	Bef. Mat.	Aft. Mat.			Bef. Mat.	Aft. Mat.	
Log(Pop1950)	1.890	0.638	66.24	Log(Civil1950)	1.791	1.149	35.86
Log(Pop1960)	1.861	0.653	64.92	Log(Civil1960)	1.741	1.071	38.48
Log(Pop1970)	1.849	0.689	62.77	Log(Civil1970)	1.719	1.058	38.44
I(Log(Pop1950)*Log(Pop1960))	1.683	0.614	63.51	I(Log(Civil1950)*Log(Civil1960))	1.570	1.021	34.99
I(Log(Pop1960)*Log(Pop1970))	1.656	0.637	61.51	I(Log(Civil1960)*Log(Civil1970))	1.534	0.975	36.44
House1950	0.825	0.216	73.82	House1950	0.825	0.403	51.16
House1960	0.469	0.067	85.63	House1960	0.469	0.051	89.09
House1970	0.221	-0.047	78.88	House1970	0.221	0.017	92.11
I(House1950*House1960)	0.581	0.063	89.12	I(House1950*House1960)	0.581	0.193	66.86
I(House1960*House1970)	0.294	-0.063	78.69	I(House1960*House1970)	0.294	-0.032	89.11
Port	0.335	0.130	61.22	Port	0.335	0.260	22.45
Airport	0.940	0.458	51.28	Airport	0.940	0.573	39.10
Jan. temp.	1.471	0.391	73.43	Log (Jan. temp.)	1.782	0.683	61.64
Log(Land area)	-0.084	0.076	8.86	Log(Land area)	-0.084	-0.066	21.50
Log(NHS mileage)	0.999	0.738	26.11	NHS mileage	0.598	0.380	36.54

Table 5: Standardized difference of means before and after matching at the city level

Population matching				Civilian employment matching			
Variable	Std. Mean Diff.		Bal. imp. (%)	Variable	Std. Mean Diff.		Bal. imp. (%)
	Bef. Mat.	Aft. Mat.			Bef. Mat.	Aft. Mat.	
Log(Pop1960)	0.680	-0.391	42.55	Log(Civil1960)	0.606	0.049	91.93
Log(Pop1970)	0.951	-0.350	63.24	Log(Civil1970)	0.925	-0.013	98.62
I(Log(Pop1960)*Log(Pop1970))	0.792	-0.383	51.70	I(Log(Civil1960)*Log(Civil1970))	0.733	0.025	96.56
Income1960	-0.556	-0.053	90.44	Income1960	-0.556	-0.198	64.50
Income1970	-0.337	-0.112	66.67	Income1970	-0.337	-0.164	51.44
I(Income1960* Income1970)	-0.561	-0.082	85.30	I(Income1960* Income1970)	-0.561	-0.207	63.03
Rent1960	-0.557	0.000	100.00	Rent1960	-0.557	-0.145	73.92
Rent1970	-0.191	0.061	68.13	Rent1970	-0.191	-0.158	17.37
I(Rent1960* Rent1970)	-0.430	-0.005	98.86	I(Rent1960* Rent1970)	-0.430	-0.207	51.81
Log(Land area)	1.131	-0.352	68.88	Log(Land area)	1.131	-0.190	83.18
Jan. temp.	-0.669	0.063	90.64	Jan. temp.	-0.669	0.172	74.21
Airport	0.437	0.221	49.44	Airport	0.437	0.246	43.83
NHS mileage	0.673	-0.267	60.33	NHS mileage	0.673	-0.096	85.75
Railroad mileage	0.876	-0.111	87.33	Railroad mileage	0.876	-0.178	79.73

5 Estimation results

Given units from the control group which match treated units, we now estimate the effects of the treatment program, i.e., providing financial assistance to Amtrak in the State of California since 1976. We examine the effects on population and civilian employment, which we consider as two proxies of economic development.

5.1 County level

This subsection presents the decade-by-decade effects of the treatment program at the county level. For each year, log (natural logarithm) of population in that year is regressed against a set of control variables including log of population in the baseline year (1970), transportation variables (log of NHS mileage and airport), geographic variables (water area), climate variable (mean January temperature), and the dummy variable indicating if the county has station(s) served by state-funded rail services. While not available to us, additional variables such as economic measures of productivity (e.g., county GDP) and availability of other forms of public transit (e.g., subway, light rail, bus networks) can be further considered provided that such information is made available for the years under investigation.

Table 6 reports the estimation results. All models have high goodness-of-fits. Three points are worth mentioning. First, all variables have expected signs. Not surprisingly, the log of population in 1970 contributes to future population more than any other variables. Greater values of NHS mileage and mean January temperature, and presence of an airport seem to encourage population growth. In contrast, water area discourages population growth, though the effect is very small. Second, rail service is statistically insignificant in the short term (1980 and 1990), but significant in the long term (2000 and 2010), which is intuitive because it takes time for state-supported rail services to take effect on local socioeconomic development. The effect of rail service increases as time goes by. Third, the constant term, which is statistically significant, entails the effect of a variety of factors such as economy, increase in industrial employment, development of infrastructure facilities, etc. that are not captured by the explanatory variables.

Table 6: OLS estimates for impacts of rail services on population at county-level

	1980		1990		2000		2010	
	Estimate	Std. err.	Estimate	Std. err.	Estimate	Std. err.	Estimate	Std. err.
(Intercept)	0.972**	0.276	1.581***	0.418	1.736**	0.501	2.045**	0.571
Treat (Rail service)	0.033	0.043	0.103	0.065	0.160*	0.078	0.205*	0.089
Log(Pop1970)	0.896***	0.030	0.833***	0.045	0.800***	0.054	0.754***	0.061
NHS mileage	9.14E-05	0.000	3.06E-04*	0.000	3.72E-04*	0.000	5.05E-04**	0.000
Water area	-2.21E-04†	0.000	-3.78E-04*	0.000	-5.49E-04*	0.000	-7.23E-04**	0.000
Jan. temp.	0.010	0.006	0.016†	0.009	0.024*	0.010	0.031*	0.012
Airport	0.099†	0.056	0.136	0.084	0.124	0.101	0.106	0.115
Adjusted R^2	0.989		0.975		0.965		0.954	
Sample size	36		36		36		36	

† Significant at 10% level, * Significant at 5% level, ** Significant at 1% level, *** Significant at 0.1% level.

The results of civilian employment regression are reported in Table 7. These models also have high goodness-of-fit and expected signs for the estimated coefficients. Again, the log of civilian employment in 1970 is highly collinear with civilian employment. Higher NHS mileage and mean January temperature lead to greater population growth starting from 1990. The effect of water area is negative but very limited. The presence of an airport does not have significant impact on civilian employment. The rail service variable is (marginally) statistically significant only in 2010, when the longest time has passed in our dataset. Looking at the point estimates, the effect of rail station presence on employment seems to increase over time, which is similar to the impact on population in Table 6.

Table 7: OLS estimates for impacts of rail services on civilian employment at county-level

	1980		1990		2000		2010	
	Estimate	Std. err.	Estimate	Std. err.	Estimate	Std. err.	Estimate	Std. err.
(Intercept)	0.786*	0.321	0.843†	0.457	1.210*	0.539	1.351*	0.600
Treat (Rail service)	0.057	0.057	0.122	0.081	0.157	0.095	0.214†	0.106
Log(Civil1970)	0.923***	0.034	0.880***	0.048	0.841***	0.057	0.812***	0.063
NHS mileage	9.35E-05	0.000	2.70E-04†	0.000	3.00E-04†	0.000	4.48E-04*	0.000
Water area	-1.58E-04	0.000	-3.12E-04	0.000	-4.71E-04*	0.000	-6.66E-04*	0.000
Jan. temp.	0.010	0.007	0.023*	0.011	0.027*	0.012	0.033*	0.014
Airport	0.041	0.077	0.063	0.109	0.063	0.129	0.052	0.143
Adjusted R^2	0.988		0.977		0.967		0.961	
Sample size	36		36		36		36	

† Significant at 10% level, * Significant at 5% level, ** Significant at 1% level, *** Significant at 0.1% level.

One may argue that commuter rail service might affect population and employment in California. To investigate this, we include a dummy variable indicating the presence of commuter rail service (defined in Section 2.2) in the population and employment models. However, we find no statistically significant effect.

All employment and population models presented in Table 6 and Table 7 have high R^2. To understand how much each explanatory variable contributes to R^2, we employ the Shapley value to distribute R^2 of each model among the explanatory variables. To do this, we use STATA software module REGO (Huettner and Sunder, 2015) to implement Shapley value decomposition. For details on the theoretical backgrounds of Shapley value decomposition, readers may refer to Huettner and Sunder (2012).

Figure 5 illustrates the contribution of each variable to R^2 in the 2010 population model. Other years follow similar patterns. Bootstrapped confidence intervals at 90% level for the contribution are also plotted. The confidence interval for the log of 1970 population does not overlap with the confidence interval of any other variable, implying the dominance of the variable in the explained variance of the 2010 population (as measured by R^2). Water area, which is significant in all models, has the lowest share (< 2%) in the explained variance. For the rail service (treatment) variable, the point estimate shows a 5% contribution to the explained variance. In addition, the lower bound of the confidence interval is around 1%, which suggests that the ability of an Amtrak station on a state-supported rail line to promote population could be marginally substantive.

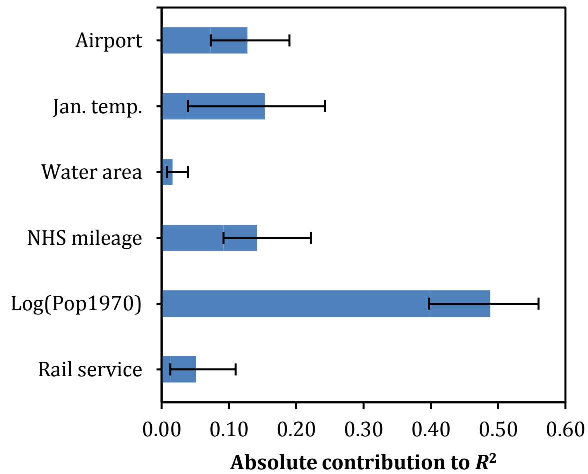

Figure 5: Decomposition of R^2 for 2010 county population model, with 90% bootstrap confidence interval, based on 5000 bootstrap replications

In an attempt to further testify the effect of state financial support for Amtrak passenger rail services on local population and employment, we also collect data and perform similar analysis for another state, Illinois. Details about the data collection, matching, and regression are presented in the Appendix. The results show that, like the California case, state support for Amtrak passenger rail services has statistically significant effects on population and employment only in part of the years. Compared to the California case, however, the effects are generally small and in short term, which might be attributed to different strengths of support of the two states over time.

5.2 City level

This subsection presents the estimated impact on city-level population and civil employment of Amtrak stations on state-supported rail lines. There are two differences in the choice of explanatory variables between the city- and county-level models. First, we use land area, which is expected to have the opposite effect of water area (in fact, the water area variable is not available at the city level). Second, since all cities in the dataset have a reachable airport located in less than 50 miles from the city's boundary, we do not include an airport variable.

Table 8 reports the estimated results for the population model. As expected, population in 1970 is still highly significant, although the coefficient suggests that the impact is smaller than at the county level. The coefficients for NHS mileage are positive and statistically significant across all four models. However, neither the land area nor the January temperature variable has a significant effect on population.

Turning to the treatment variable, we find that the presence of an Amtrak station on a state-supported line does not have significant effect on population in 1980; however, the effect becomes significant starting from 1990. The magnitude of the point estimate is similar to the county-level estimate: forty years after the establishment of the state support (2010), an Amtrak station on a state-supported line increases a city's population by 17%.

Table 8: OLS estimates for impacts of rail services on population at city-level

	1980		1990		2000		2005	
	Estimate	Std. err.	Estimate	Std. err.	Estimate	Std. err.	Estimate	Std. err.
(Intercept)	2.158	1.351	2.955	1.969	2.793	2.082	2.435	2.290
Treat (Rail service)	0.087	0.053	0.146†	0.077	0.180*	0.082	0.170†	0.090
Log(Pop1970)	0.738***	0.078	0.570***	0.113	0.487**	0.120	0.445**	0.132
Log(NHS Mileage)	0.283**	0.090	0.483**	0.132	0.561**	0.139	0.569**	0.153
Land area	8.38E-04	0.003	2.04E-03	0.004	2.28E-03	0.004	4.40E-03	0.004
Log(Jan. temp.)	-0.037	0.387	0.100	0.564	0.325	0.597	0.522	0.656
Adjusted R^2	0.942		0.894		0.887		0.875	
Sample size	20		20		20		20	

† Significant at 10% level, * Significant at 5% level, ** Significant at 1% level, *** Significant at 0.1% level.

Estimates for the civilian employment models are documented in Table 9. Compared to the population models, the civilian employment models have lower R^2 values. Again, land area and January temperature variables do not have significant coefficients. As in the county-level models, civilian employment in 1970 and NHS mileage have positive coefficients, which are mostly significant. We find that the coefficients of the treatment variable are statistically insignificant in all four models. This confirms the finding at the county level that the presence of an Amtrak station on a state-supported line has little effect on civilian employment.

Similar to the county-level case, we investigate the impact of commuter rail service on population and employment at the city level. Again, the estimates are not statistically significant.

Table 9: OLS estimates for impacts of rail services on civilian employment at city-level

	1980		1990		2000		2005	
	Estimate	Std. err.	Estimate	Std. err.	Estimate	Std. err.	Estimate	Std. err.
(Intercept)	2.973	2.734	4.477	2.925	5.854†	2.998	3.850	2.759
Treat (Rail service)	0.001	0.080	0.0170	0.085	0.047	0.087	0.059	0.081
Log(Civil1970)	0.826***	0.137	0.653***	0.147	0.669***	0.150	0.597**	0.138
Log(NHS Mileage)	0.170	0.111	0.300*	0.119	0.251†	0.122	0.282*	0.112
Land area	-2.71E-04	0.003	1.78E-03	0.004	2.69E-03	0.004	5.93E-03	0.003
Log(Jan. temp.)	-0.368	0.628	-0.363	0.672	-0.702	0.689	-0.049	0.634
Adjusted R^2	0.838		0.842		0.825		0.877	
Sample size	18		18		18		18	

† Significant at 10% level, * Significant at 5% level, ** Significant at 1% level, *** Significant at 0.1% level.

We also investigate how much each of the explanatory variables contributes to R^2 in the population model. We do not present the results for the civilian employment models, as the treatment variable is not significant. Figure 6 shows the R^2 decomposition results for the 2005 city population model. Again, models for other years offer similar results. As in the county-level case, the baseline population and NHS mileage have the greatest mean contributions to R^2, with about 40% and 32%

shares respectively. The treatment variable contributes about 2% to the explained variance. The wide confidence interval of the treatment variable, in particular the lower bound which is again close to 1%, reaffirms our earlier argument at the county level that the effect on population of an Amtrak station on a state-supported rail line could be small.

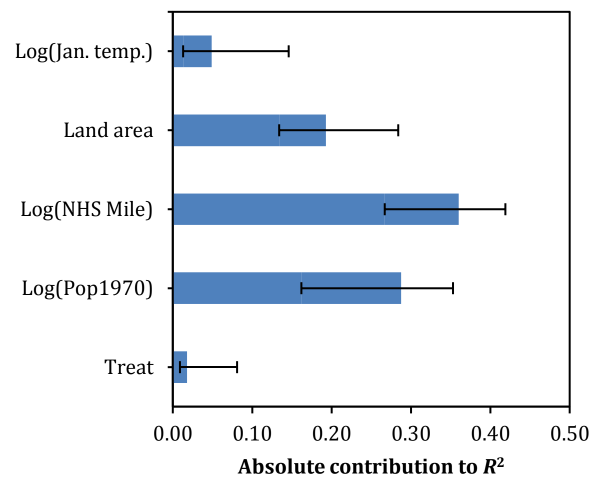

Figure 6: Decomposition of R^2 for the 2005 city population model,

with 90% bootstrap confidence interval, based on 5000 bootstrap replications

6 Concluding remarks

Passenger rail played a vital role in US intercity travel and economy in the early 20th century. However, due to the advent of automobiles and airplanes, passenger rail lost much of its dominance. The establishment of Amtrak in 1971 and subsequently state-level support for parts of Amtrak's services helped preserve the passenger rail system in the country and revitalize services that remain an integral part in the national multimodal transportation system. Given that the local socioeconomic impact of such services is largely unknown in the literature, this paper intends to fill the gap by empirically investigating how Amtrak stations on state-support rail lines have affected population and employment at county and city levels.

We compile two panel datasets for the state of California which include various county- and city-level geographic, transportation, and socioeconomic characteristics. In view of the missing values in the datasets, multivariate normal imputation is used to fill in the missing values. We then employ a propensity score based one-to-one matching model to draw units from the control group, which are counties/cities that do not have a state-supported Amtrak station, to match with units from the treatment group, which are counties/cities that do. Using the matched data, we perform ordinary least square regressions to estimate the effect of a state-supported Amtrak station on local population and employment.

The estimation results suggest a positive effect on population at both city and county levels. The effect is more prominent as time goes by. The population growth in turn spurs Amtrak's ridership, whose growth is more than double the population growth between 2009 and 2015 in the state (Amtrak, 2015). At the county level, the effects on population and civilian employments have similar point estimates. However, for the effect on civilian employment most of them are statistically insignificant (only significant at 10% level in 2010). At the city level, the point estimates for the civilian employment effect are much smaller than for the population effect, and none of them is found statistically significant.

Thus overall, we are more confident about the role an average Amtrak station on a state-funded line plays to promote population growth than to encourage civilian employment, although the effect on population growth could still be small, given the confidence intervals of the rail service variable's contribution to the overall goodness-of-fit of the population models. One plausible explanation may be that state-supported Amtrak services provide quality mobility and accessibility by rail, which attract people to live in rail-accessible regions. On the other hand, the weak effect on employment of state-supported Amtrak service is not surprising, with two possible explanations. First, a train station has only marginal impact on overall accessibility, and therefore can induce limited direct or indirect economic activities. Second, although Amtrak is reported to support thousands of jobs (Amtrak, 2015), it is not clear how these jobs are distributed between state-supported and non-state-supported routes, and among counties and cities. For example, jobs on purchase of supplies and materials are not necessarily correlated with the geographic coverage of transportation service. Thus, it is hard to draw concrete conclusions on the benefits of state-supported Amtrak stations on local employment.

This study can be extended in a few directions. First, this study only investigates the effect of state-supported rail service on total employment in aggregate at county and city levels. It would be interesting to examine the impact on specific sectors (e.g., tradable, non-tradable, and transportation sectors). Second, the impact of an Amtrak station on local development may vary with county/city size. Future research may look into the economic impact of state-supported rail services for different county/city sizes. Third, in addition to the state-supported rail service, several other factors, such as fertility rate, mortality rate (life expectancy), and migration, could be responsible for population growth. In the current models, these factors are only implicitly and indirectly captured through the constant, the 1970 population / employment variable, and the error term. Pin-pointing specific factors is challenging. Future efforts should be directed to identifying such factors, collecting relevant data, and testing their effects on population growth. Finally, this study focuses on California. Application of the methodology developed in this study to other states will lend a more comprehensive understanding of the socioeconomic impact of state-supported Amtrak services. Such understanding can help inform future policies and development of intercity passenger rail in the US. For example, as mentioned in Sperry et al. (2013), the understanding can be incorporated into state rail plans and applications for federal grants for passenger rail. The understanding can also be part of the material for public outreach to local community leaders and businesses, and for educating legislators on the socioeconomic impact of Amtrak service as decision are made about continuing/strengthening/reducing the financial support for passenger rail service in California and elsewhere.

Acknowledgment

This research was partially supported by California Department of Transportation (Caltrans). We would like to thank Dr. Marquise McGraw from the University of California at Berkeley for initial discussions, and Professor David Brownstone from the University of California at Irvine and Dr. Joshua Seeherman from the University of California at Berkeley for their helpful comments. The views are those of the authors alone.

Appendix: Analysis for the State of Illinois

In this appendix, we present our data collection and modeling efforts for the State of Illinois. Illinois has four state-supported services: Zephyr Service (Chicago – Quincy, receiving state support since 1971), Lincoln Service (Chicago – Springfield – St. Louis, receiving state support since 1973), Illini Service (Chicago – Champaign, receiving state support since 1973), and Hiawatha Service (Chicago – Milwaukee, receiving state support since 1989). As state support for intercity passenger rail services in Illinois started in the early 1970s, we consider 1950-1970 as the pre-treatment period (the same as in California case).

Similar to what we do for California, a dataset is developed which contains socioeconomic, demographic, geographic, and transportation information of Illinois counties between 1950 and 2010. The locations of Amtrak stations on state-supported routes are obtained from the Illinois Department of Transportation (IDOT, 2017). In total, 24 counties have such Amtrak stations. Cook County, whose county seat is Chicago, is removed from the dataset because of its very different characteristics from other counties in the state. After the removal, the dataset has 23 counties in the treatment group. The state's remaining 79 counties form the control group.

Most of the information for Illinois is obtained from the same data sources as for California (see Section 2.4), except that commuter rail information draws from Metra (2017). No imputation is needed for Illinois as the collected information is complete. Figure A.1 shows the mean population and civilian employment in the control and treatment groups between 1950 and 2010.

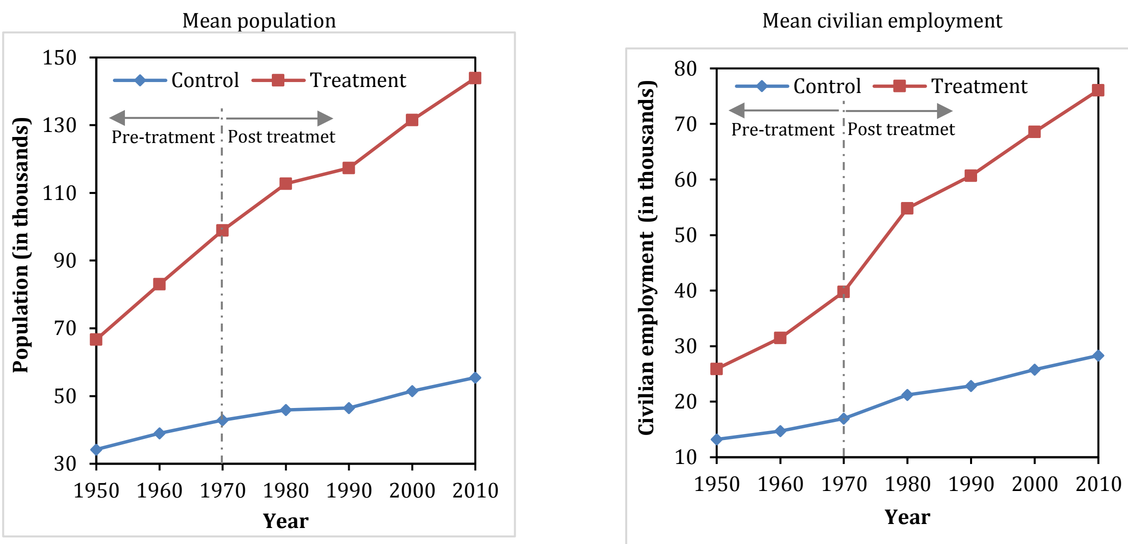

Figure A.1: county-level mean population and civilian employment in the control and the treatment groups

The explanatory variables considered in the regression models are the same as those in the California models, with two differences: 1) the portion of water area in the total area of a county is used rather than the water area itself; 2) two dummy variables related to the Chicago metropolitan area are added. For the water area portion variable, the hypothesis is that a larger portion of water area has a positive effect on local population/employment. This is because in Illinois, unlike in California, population conglomeration outside the Chicago metropolitan area is often nearby rivers (e.g., Mississippi River, Illinois River, Kaskaskia River, Ohio River, Wabash River, Kankakee River).

Also note that Cook and Lake counties, which are the only two counties in the state bordered with Lake Michigan, are not in the dataset after matching.

The two Chicago metro dummies intends to capture the Chicago effect in dominating the state's population and the suburbanization trend within the metropolitan area. The first dummy (Chicago metro1) takes value 1 if a county is DuPage, Kane, or Will, and 0 otherwise. These counties are in the immediate surroundings of the city of Chicago. The second dummy (Chicago metro2) takes value 1 if a county is from the following list: DeKalb, Grundy, Kendall, and McHenry, and 0 otherwise. These counties are farther away from the city of Chicago but still belong to the metropolitan area according to US Census Bureau (2016). Our hypothesis is that population and employment in these counties – especially those in the outer region of the Chicago metropolitan area – have grown faster than the rest of the state due to the continuous suburbanization after 1970.

The regression results for population and employment are presented in Tables A1 and A2. In the population models, the coefficients for the rail service variable consistently have a positive sign, but only significant in 1980. The point estimates are generally smaller than those for California (see Table 6). This suggests that the effect of state support for Amtrak service on local population is in shorter term and weaker in Illinois than in California.

For the other variables, the log of population in 1970 again contributes most to the variation in future population. The commuter dummy variable has an expected positive coefficient in all models, but only significant in 1990. The coefficients for NHS mileage and water area portion are consistently non-negative as well, significant only for 2010. January temperature and airport presence do not show statistically significant effect on population. The large and significant coefficients for the outer counties of the Chicago metropolitan area (Chicago metro2) suggest stronger population growth in these counties in the study period than in the inner counties (Chicago metro1) and the rest of the state.

Table A1: OLS estimates for impacts of rail services on population at the county level for Illinois

	1980		1990		2000		2010	
	Estimate	Std. err.	Estimate	Std. err.	Estimate	Std. err.	Estimate	Std. err.
(Intercept)	0.249	0.333	0.425	0.464	1.054	0.685	1.848	0.981
Treat (Rail service)	0.049*	0.021	0.042	0.030	0.049	0.044	0.061	0.063
Log(Civil1970)	0.970***	0.027	0.945***	0.037	0.892***	0.055	0.822***	0.078
Commuter	0.143	0.094	0.286*	0.130	0.323	0.193	-0.070	0.276
NHS Mileage	-8.14E-05	3.30E-04	1.68E-04	4.60E-04	9.13E-04	6.79E-04	0.002†	0.001
Water area portion	2.111	1.257	2.522	1.752	4.267	2.588	7.163†	3.706
Jan. temp.	0.003	0.004	0.004	0.006	-3.76E-04	0.008	-0.006	0.012
Airport	0.019	0.036	0.077	0.051	0.084	0.075	0.128	0.108
Chicago metro1	0.097	0.091	0.098	0.127	0.198	0.188	0.653*	0.269
Chicago metro2	0.170**	0.051	0.280***	0.071	0.485***	0.105	0.942***	0.150
Adjusted R^2	0.996		0.992		0.984		0.970	
Sample size	46		46		46		46	

† Significant at 10% level, * Significant at 5% level, ** Significant at 1% level, *** Significant at 0.1% level.

The results for the civil employment model yield similar insights. The rail service dummy is only significant in 1980. The most explanatory power of the models comes from the log of civil employment in 1970. The commuter variable has a positive coefficient, which is significant for 1990. NHS mileage and water area portion again have positive coefficients, significant only for 2010. The Chicago metro2 variable shows strong and significant effect on civil employment. For the coefficients of the other variables, most of them have expected signs but are statistically insignificant (we note that Jan. Temp. has two unexpected negative coefficients. However, they are also highly insignificant).

Table A2: OLS estimates for impacts of rail services on civilian employment at the county level for Illinois

	1980		1990		2000		2010	
	Estimate	Std. err.	Estimate	Std. err.	Estimate	Std. err.	Estimate	Std. err.
(Intercept)	0.583	0.408	0.782	0.587	1.445†	0.794	1.910†	1.126
Treat (Rail service)	0.066*	0.027	0.054	0.039	0.072	0.053	0.075	0.076
Log(Civil1970)	0.952***	0.034	0.921***	0.049	0.862***	0.067	0.818***	0.094
Commuter	0.180	0.121	0.314†	0.174	0.339	0.235	-0.100	0.333
NHS Mileage	-3.16E-06	4.25E-04	3.78E-04	6.11E-04	1.13E-03	8.26E-04	0.002†	0.001
Water area portion	2.018	1.615	2.584	2.321	4.626	3.137	8.860†	4.451
Jan. temp.	0.001	0.005	0.003	0.008	-0.001	0.010	-0.007	0.015
Airport	0.056	0.048	0.123†	0.069	0.122	0.093	0.128	0.132
Chicago metro1	0.109	0.118	0.125	0.169	0.181	0.229	0.660*	0.325
Chicago metro2	0.180***	0.065	0.327***	0.094	0.510***	0.127	0.989***	0.180
Adjusted R^2	0.993		0.987		0.978		0.961	
Sample size	46		46		46		46	

† Significant at 10% level, * Significant at 5% level, ** Significant at 1% level, *** Significant at 0.1% level.

References

1. AECOM, 2013. 2013 California State Rail Plan. California Department of Transportation, Sacramento, CA.
2. Aschauer, D.A., 1989. Is public expenditure productive? Journal of Monetary Economics 23 (2), 177-200.
3. Atack, J., Bateman, F., Haines, M., Margo, R.A., 2010. Did railroads induce or follow economic growth. Social Science History 34 (2), 171-197.
4. Atack, J., Margo, R.A., 2009. Agricultural Improvements and Access to Rail Transportation: The American Midwest as a Test Case, 1850-1860. National Bureau of Economic Research, Report No. w15520.
5. Banerjee, A., Duflo, E., Qian, N., 2012. On the road: Access to transportation infrastructure and economic growth in China. National Bureau of Economic Research, Report No. w17897.
6. Baum-Snow, N., 2007. Did highways cause suburbanization? The Quarterly Journal of Economics 122 (2), 775-805.

7. Berger, T., Enflo, K., 2014. Locomotives of Local Growth: The Short-and Long-Term Impact of Railroads in Sweden. Department of Economic History, Lund University, Repot No. 132.
8. Blonigen, B.A., Cristea, A.D., 2015. Air service and urban growth: Evidence from a quasi-natural policy experiment. Journal of Urban Economics 86, 128-146.
9. Brueckner, J.K., 2003. Airline traffic and urban economic development. Urban Studies 40 (8), 1455-1469.
10. Caltrans, 2015. Caltrans GIS Data Library. Available at http://www.dot.ca.gov/hq/tsip/gis/datalibrary/, Retrieved on December 29, 2015.
11. Chi, G., 2010. The impacts of highway expansion on population change: An integrated spatial approach. Rural Sociology 75 (1), 58-89.
12. Cidell, J., 2015. The role of major infrastructure in subregional economic development: an empirical study of airports and cities. Journal of Economic Geography 15 (6), 1125-1144.
13. CTA, 2003. CTA Railroad Network. Available at http://www-cta.ornl.gov/transnet/RailRoads.html, Retrieved on December 29, 2015.
14. Donaldson, D., Hornbeck, R., 2013. Railroads and American Economic Growth: A" Market Access" Approach. National Bureau of Economic Research, Report No. w19213.
15. Duranton, G., Morrow, P.M., Turner, M.A., 2014. Roads and Trade: Evidence from the US. The Review of Economic Studies 81 (2), 681-724.
16. Duranton, G., Turner, M.A., 2012. Urban growth and transportation. The Review of Economic Studies 79 (4), 1407-1440.
17. Elkind, E.N., Chan, M. Faber, T.V., 2015. Grading California's Rail Transit Station Areas: A Ranking of How Well They Accommodate Population Growth, Boost Economic Activity and Improve the Environment. Center for Law, Energy & the Environment, University of California, Berkeley.
18. Faber, B., 2014. Trade integration, market size, and industrialization: evidence from China's National Trunk Highway System. The Review of Economic Studies 81 (3), 1046-1070.
19. Franch, X., Morillas-Torné, M., Martí-Henneberg, J., 2013. Railways as a Factor of Change in the Distribution of Population in Spain, 1900–1970. Historical Methods: A Journal of Quantitative and Interdisciplinary History 46 (3), 144-156.
20. Garcia-Milà, T., Montalvo, J.G., 2007. The impact of new highways on business location: new evidence from Spain. Centre De Recerca En Economia Internacional, Universitat Pompeu Fabra, working paper.
21. Gibbons, S., Lyytikainen, T., Overman, H.G., Sanchis-Guarner, R., 2012. New road infrastructure: the effects on firms. SERC Discussion Papers, SERCDP117. Spatial Economics Research Centre (SERC), London School of Economics and Political Science, London, UK.
22. Green, R.K., 2007. Airports and economic development. Real Estate Economics 35 (1), 91-112.
23. Gregory, I.N. Henneberg, J.M., 2010. The Railways, Urbanization, and Local Demography in England and Wales. Social Science History 34 (2), 199-228.
24. Ho, D., Imai, K., King, G., Stuart, E., 2007. MatchIt: MatchIt: Nonparametric Preprocessing for Parametric Casual Inference. Political Analysis 15 (3), 199-236.
25. Honaker, J., King, G., 2010. What to do about missing values in time-series cross-section data. American Journal of Political Science 54 (2), 561-581.
26. Honaker, J., King, G., Blackwell, M., 2011. Amelia II: A program for missing data. Journal of Statistical Software 45 (7), 1-47.
27. Hornung, E., 2012. Railroads and Micro-Regional Growth in Prussia, ifo Institut – Leibniz-Institut für Wirtschaftsforschung an der Universität München, working paper.
28. Huettner, F., Sunder, M., 2015. [. rego] R-Squared decomposition. Available at http://www.uni-leipzig.de/~rego/, Retrieved on December 29, 2015.
29. Huettner, F., Sunder, M., 2012. Axiomatic arguments for decomposing goodness of fit according to Shapley and Owen values. Electronic Journal of Statistics 6, 1239-1250.

30. ICPSR, 2015a. County and City Data Book [United States] Series. Available at: http://www.icpsr.umich.edu/icpsrweb/ICPSR/series/23, Retrieved on December 29, 2015.
31. ICPSR, 2015b. County Characteristics, 2000-2007 [United States] (ICPSR 20660). Available at: http://www.icpsr.umich.edu/icpsrweb/DSDR/studies/20660, Retrieved on December 29, 2015.
32. IDOT, 2017. Passenger Rail Services. Available at http://www.idot.illinois.gov/travel-information/passenger-services/AMTRAK-services/index, Retrieved on October 23, 2017.
33. King, G., Honaker, J., Joseph, A., Scheve, K., 2001. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. American Political Science Association 95 (1), 49-69.
34. Koopmans, C., Rietveld, P., Huijg, A., 2012. An accessibility approach to railways and municipal population growth, 1840–1930. Journal of Transport Geography 25, 98-104.
35. Lee, K.J., Carlin, J.B., 2010. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. American Journal of Epidemiology 171 (5), 624-632.
36. Lieshout, R., 2012. Measuring the size of an airport's catchment area. Journal of Transport Geography 25, 27-34.
37. Marcucci, E. Gatta, V., 2011. Regional airport choice: consumer behaviour and policy implications. Journal of Transport Geography 19 (1), 70-84.
38. McGraw, M., 2014. Perhaps the Sky's the Limit? Airports and Employment in Local Economies. University of California, Berkeley, Job Market paper.
39. Metra, 2017. Maps and Schedules | Metra. Available at https://metrarail.com/maps-schedules, Retrieved on October 23, 2017.
40. Michaels, G., 2008. The effect of trade on the demand for skill: Evidence from the interstate highway system. The Review of Economics and Statistics 90 (4), 683-701.
41. MPC, 2011. National historical geographic information system: Version 2.0. Minnesota Population Center, University of Minnesota, Minneapolis, MN.
42. Mukkala, K., Tervo, H., 2013. Air transportation and regional growth: which way does the causality run? Environment and Planning A 45(6), 1508-1520.
43. Munnell, A.H., Cook, L.M., 1990. How does public infrastructure affect regional economic performance? New England Economic Review, 11-33.
44. Murakami, J., Cervero, R., 2010. California high-speed rail and economic development: Station-area market profiles and public policy responses. University of California Transportation Center.
45. Nice, D., 1998. The History and Politics of a National Railroad. Lynne Rienner Publishers, Boulder, Colorado.
46. NOAA, 2012. NOAA's List of Coastal Counties. Available at: https://www.census.gov/geo/landview/lv6help/coastal_cty.pdf, Retrieved on December 29, 2015.
47. NOAA, 2000. Ports List. Available at: http://www.ngs.noaa.gov/RSD/coastal/projects/coastal/ports_list.html, Retrieved on December 29, 2015.
48. Nordstrom, M., 2015. The Growth Effect: Commuter Rail in Southern California. California State University, Northridge, Doctoral dissertation.
49. NTAD, 2015. National Transportation Atlas Database | Bureau of Transportation Statistics. Available on: http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/national_transportation_atlas_database/2015/index.html, Retrieved on December 29, 2015.
50. Pan, J., 2010. The United States Outer Executive Departments and Independent Establishments & Government Corporations. Xlibris Corporation.
51. Percoco, M., 2010. Airport activity and local development: evidence from Italy. Urban Studies 47 (11), 2427-2443.

52. Pereira, R.M., Hausman, W.J., Pereira, A.M., 2015. Railroads and Economic Growth in the Antebellum United States. Department of Economics, College of William and Mary, Working paper No. 153.
53. Rosenbaum, P.R., 1989. Optimal matching for observational studies. Journal of the American Statistical Association 84 (408), 1024-1032.
54. Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66 (5), 688.
55. Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data. CRC press.
56. Schafer, J.L., Olsen, M.K., 1998. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavioral Research 33 (4), 545-571.
57. Sekhon, J.S., 2011. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. Journal of Statistical Software 42 (7).
58. Sheard, N., 2014. Airports and urban sectoral employment. Journal of Urban Economics 80, 133-152.
59. Sperry, B., Taylor, J., Roach, J., 2013. Economic Impacts of Amtrak Intercity Passenger Rail Service in Michigan: Community-Level Analysis. Transportation Research Record: Journal of the Transportation Research Board 2374, 17-25.
60. Stuart, E.A., 2010. Matching methods for causal inference: A review and a look forward. Statistical Science: A Review Journal of the Institute of Mathematical Statistics 25 (1), 1-21.
61. Thompson, G.L., 1993. The Passenger Train in the Motor Age: California's Rail and Bus Industries, 1910-1941. Ohio State University Press.
62. US Census Bureau, 2016. American Community Survey 1-year estimates. Retrieved from Census Reporter Profile page for Chicago-Naperville-Elgin, IL-IN-WI Metro Area. Available at: https://censusreporter.org/profiles/31000US16980-chicago-joliet-naperville-il-in-wi-metro-area/, Retrieved on October 23, 2017.
63. US Census Bureau, 2010. County and City Data Book: 2007. Available at: http://www.census.gov/library/publications/2010/compendia/databooks/ccdb07.html, Retrieved on December 29, 2015.
64. van den Heuvel, F., Rivera, L., van Donselaar, K.H., de Jong, A., Sheffi, Y., de Langen, P.W., Fransoo, J.C., 2014. Relationship between freight accessibility and logistics employment in US counties. Transportation Research Part A: Policy and Practice 59, 91-105.
65. Wang, Y., Wu, B., 2015. Railways and the Local Economy: Evidence from Qingzang Railway. Economic Development and Cultural Change 63 (3), 551-588.