# THÈSE

**En vue de l'obtention du**

**DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE**

**Délivré par l'Université Toulouse 1 Capitole**

**Présentée et soutenue par**

**Fernando STIPANICIC MÁRQUEZ**

Le 11 juillet 2022

## Essays on International Trade and Innovation

Ecole doctorale : **TSE - Toulouse Sciences Economiques**

Spécialité : **Sciences Economiques - Toulouse**

Unité de recherche :
**TSE-R - Toulouse School of Economics - Recherche**

Thèse dirigée par
**Christian HELLWIG**

Jury
**M. Mark SCHANKERMAN,** Rapporteur
**M. Marti MESTIERI,** Rapporteur
**M. Christian HELLWIG,** Directeur de thèse
**M. Saleh MOHAMED,** Président

# The Creation and Diffusion of Knowledge: Evidence from the Jet Age*

Stefan Pauly† Fernando Stipanicic‡

April 23, 2022

Click here for the latest version

This paper provides new causal evidence of the impact of air travel time on the creation and diffusion of knowledge. We exploit the beginning of the *Jet Age* as a quasi-natural experiment. We digitize airlines' historical flight schedules and construct a novel data set of the flight network in the United States. Between 1951 and 1966, travel time between locations more than 2,000 km apart decreased on average by 41%. The reduction in travel time explains 33% of the increase in knowledge diffusion as measured by patent citations. The increase in knowledge diffusion further caused an increase in the creation of new knowledge. The results provide evidence that jet airplanes led to innovation convergence across locations and contributed to the shift in innovation activity towards the South and the West of the United States.

JEL Classification: O31, O33, R41, N72

*First version: February 2021. A previous version of this paper was named *The diffusion of knowledge: Evidence from the Jet Age*. We are grateful for the continuous support of Christian Hellwig since the beginning of this project. We are indebted to Björn Larsson who has made this project possible by sharing with us his collection of historical flight schedules. We thank Taylor Jaworski and Carl Kitchens for sharing highway data with us. This paper has benefited from comments of Guillermo Alves, Tiziana Assenza, Antonin Bergeaud, Matteo Bobba, Christian Bontemps, Felipe Carozzi, Thomas Chaney, Fabrice Collard, Patrick Fève, Ruben Gaetani, Victor Gay, Ulrich Hege, Patrick Llerena, João Pereira dos Santos, Mohamed Saleh, Mark Schankerman, Andrea Smurra, Nicolas Werquin, Alex Whalley, Miguel Zerecero, and numerous participants in seminars. During 2021 this paper has been presented in: Geneva Local GTDW, Paris GSIE Seminar, INSEAD/Collège de France, European Meeting UEA, ACDD Strasbourg, UAB Lunch Webinar, African Meeting ES, IAAE Annual Conference, Annual Conference ITEA, IMF, Bavarian Young Economist Meeting, SED Annual Meeting, YES Princeton, European Economics Association, ifo Dresden Regional Economics, RIEF Paris, LSE Graduate Economic History Seminar, RIDGE Growth, European Winter Meeting ES.

†Sciences Po, Ph.D. Candidate. Email: *stefan.pauly@sciencespo.fr*

‡Toulouse School of Economics, Ph.D. Candidate. Email: *fernando.stipanicic@tse-fr.eu*

# 1. Introduction

*"If I have seen further it is by standing on the shoulders of Giants."*

– Isaac Newton (1675)[1]

*"(...) if one man starts a new idea, it is taken up by others and combined with suggestions of their own; and thus becomes the source of further new ideas."*

– Alfred Marshall (1890)[2]

In their famous quotations Isaac Newton and Alfred Marshall illustrate that access to knowledge is key for the creation of new knowledge. Understanding the process of creation of new knowledge is crucial as it has been characterized as one of the main causes of economic growth (Lucas (1993), Aghion and Howitt (1997) and Jones (2002)). Access to knowledge spurs the creation of new knowledge (Furman and Stern (2011), Acemoglu et al. (2016)). Physical proximity, by facilitating face to face interactions, is a key driver of the diffusion of knowledge and hence of access to knowledge (Storper and Venables (2004), Glaeser (2011)).

Providing evidence of the effect of access to knowledge on the creation of new knowledge is an empirical challenge. Agents who are highly productive in terms of creation of knowledge may endogenously sort towards locations with high access to knowledge, leading to reverse causality. Additionally, access to knowledge is correlated with other drivers of innovation as access to markets, resulting in a potential omitted variable bias due to confounding factors.

This paper provides new causal evidence on this question by exploiting as a quasi-natural experiment the beginning of the *Jet Age* in the United States. During the 1950s the introduction of jet engines into civil aviation led to a large reduction in travel time.

[1]Quoted from a letter of Isaac Newton to Robert Hooke, 1675. A digital copy of the letter can be found at: `https://digitallibrary.hsp.org/index.php/Detail/objects/9792`

[2]Quoted from Duranton and Puga (2004), page 2066.

We exploit changes in travel time to identify changes in knowledge diffusion, which are further translated into changes in access to knowledge. Then, we exploit changes in access to knowledge to study the impact on the creation of new knowledge. The results provide evidence that jet airplanes led to innovation convergence across locations and contributed to the shift in innovation activity towards the South and the West of the United States.

We start by constructing a new dataset of the flight network in the United States during the 1950s and 1960s. We digitize historical flight schedules of the major interstate airlines operating in the period and obtain the fastest route between every two airports in the network.[3] We document that between 1951 and 1966 travel time decreased on average by 29%, and the decrease is on average of 41% for airports located more than 2,000km apart.[4]

This nationwide shock was arguably exogenous as it happened in a strictly regulated environment. We decompose the change in travel time and find that 90% of the change is due to the improvement in aircrafts' speed, while 10% is due to a change in the flight routes. This is consistent with the fact that during this period the Civil Aeronautics Board (CAB) was imposing strong regulation in the interstate airline market. With the objective to promote a *stable* airline industry, the CAB determined ticket prices and restricted entry of airlines into new or existing routes.

Additionally, during the 1950s and 1960s airplanes were predominantly used to transport people and not goods. Hence, the change in travel time represented a shock to the mobility of people while not significantly affecting the shipment of goods.

To study knowledge creation and diffusion we use patent data. We follow Jaffe

[3]The 6 domestic airlines in our data accounted for 75% of total air passenger transport.

[4]New York and Boston are about 300km apart, while New York and San Francisco are located about 4,130 km apart. Between 1951 and 1966 we observe a reduction of travel time of 23% (13 minutes reduction) between New York and Boston, while the reduction is of 50% (5 hours 30 minutes reduction) between New York and San Francisco.

et al. (1993) and use patent citations as our observable measure of knowledge flow. We assemble one dataset with all corporate patents granted by the United States Patent and Trademark Office (USPTO) with filing year between 1949 and 1968, which includes for each patent: filing year, technology classification, location (Metropolitan Statistical Area, MSA) of the inventors when they applied for the patent, owner of the patent and citations to other patents which were granted by the USPTO.

We document three facts of patenting activity during our sample period. First, patent growth was stronger both in initially less innovative MSAs and in MSAs in the South and the West of the US. Second, over time multi-establishment firms expanded geographically and accounted for a larger share of patents. Third, the mass of citations shifted towards longer distances. Our results show that the decrease in travel time contributed to all three facts.

We do our analysis in three steps. In the first step, we estimate a gravity equation to obtain the elasticity of citations to travel time. We identify the elasticity exploiting only within establishment-pair across-time variation in citations and travel time. The estimated elasticity implies that citations increased on average 2.4% due to the decrease in travel time between 1951 and 1966. We find that the absolute value of the elasticity is increasing with the distance between the citing and cited establishments. At a distance of more than 2,000km, the change in travel time implies an increase in citations of 6.9%. This accounts for 32.7% of the observed increase in citations in this distance range.

In order to rule out the possibility that the opening of new routes or the timing of adoption of jets at the route level was driven by variables that also affected knowledge flows, we perform an instrumental variables estimation. We instrument the observed travel time with a fictitious travel time computed by fixing routes to the initial time period and assuming in each year all routes are operated with the year's average airplane. Hence, changes in fictitious travel time are only due to the nationwide roll out of jets and is thus independent of decisions at the route level. The results do not change significantly, reflecting the reduced scope for endogeneity of travel time. In addition, the results are robust to controlling for potential confounding factors such as changes in highway travel time, telephone connectivity and flight ticket prices. Finally, the results also remain after restricting the sample to contain only establishments that existed in the initial time period.

In the second step, using the estimated elasticity of diffusion of knowledge, we compute a measure of knowledge access that is specific to each location-technology. The measure captures changes in knowledge access that are only consequence of the change in travel time. We use knowledge access as an externality that affects the production of new patents and estimate the elasticity of new patents to knowledge access. We identify the elasticity at the establishment level comparing only across time variation in patents and knowledge access across establishments within a location, conditional on aggregate technological trends. Thus, the identification is independent of location specific changes in local population or R&D subsidies. The estimated elasticity implies that the amount of new patents filed increased at a yearly growth rate of 3.5% due to the increase in knowledge access, which accounts for 79.5% of the observed yearly growth rate.

Given the reduction in travel time was larger for longer distances, the increase in knowledge access was stronger in locations geographically far from the initial innovation centers located in the Midwest and the Northeast. Hence, by increasing access to knowledge, the reduction in travel time led to a shift in the distribution of innovative activity towards the South and the West of the US. The South and the West had an average yearly growth rate of patenting 2.1 percentage points higher than the Northeast and the Midwest during our sample period. The change in travel time explains 35% of the observed differential growth.

We find that the value of the elasticity of patents to knowledge access is bigger in magnitude for establishments located in initially less innovative locations. Within each technology class, we rank locations according to the amount of patents in the initial time period and split them into four quartiles. We find that the increase in knowledge access predicts a 4.5% yearly growth rate of patenting in locations in the lowest quartile of initial innovativeness, while it predicts a 3.4% yearly growth rate in the highest quartile. The difference in growth rates indicates that the increase in knowledge access acted as a convergence force between locations, and it can explain 21% of the convergence observed in the data. Results go in the same direction if we rank locations in terms of patents per capita.

Our results are robust to controlling for changes in market access by highway, changes in market access by airplanes and time changing telephone connectivity. Results do not change if we compute knowledge access using only knowledge located at long distances. Additionally, we present suggestive evidence that the results are not driven by a decrease in financial frictions.

In the third step, we uncover the sources of the increase in patenting. We find that most of the effect of knowledge access on new patents happens through two entry margins: entry of establishments of new firms and entry of subsidiaries of firms that expanded from other locations. The two entry margins are stronger in initially less innovative locations, meaning that convergence comes both from new firms and the geographic expansion of multi-establishment firms.

To more directly test the firm expansion channel, we study if a firm's subsidiary's location decision depends on travel time to headquarters. We estimate a probability model to analyze if the locations in which firms have inventors applying for patents depends on travel time to the firm's headquarters. We identify the change in the probability only from changes in travel time and locations in which the firm starts patenting or stops patenting. We find that the probability that a firm has inventors applying for patents in a certain location goes up when then travel time from that location to the firm's headquarters reduces. In addition, the change in the probability is stronger for potential recipient locations that were initially less innovative, again highlighting the importance of this channel for convergence.

This paper contributes to multiple branches of literature. First, it contributes to the literature on agglomeration and knowledge spillovers. Agglomeration forces are usually understood as happening in a geographically localized manner (Glaeser (2011), Arzaghi and Henderson (2008)). The literature on tech clusters also documents this fact (Duranton et al. (2009), Kerr and Robert-Nicoud (2020), Moretti (2021)). The seminal paper Jaffe et al. (1993) finds that patent citations decay rapidly with distance. Our results show that jet airplanes allowed long distance knowledge spillovers, facilitating the development of tech clusters in other regions. The literature that provides evidence of knowledge spillovers usually focuses on changes in the supply of knowledge (Bloom et al. (2013), Acemoglu et al. (2016)). In our case we fix the supply of knowledge and focus on changes in the degree of accessibility.

We contribute to the literature on transportation by studying a new quasi-natural experiment that isolates a shock to the mobility of people. To do so we construct a new dataset that could be used to answer many other questions.[5] Other papers have studied the impact of transportation improvements on innovation. Agrawal et al. (2017) study the impact on innovation of a region's stock of highways, while Perlman (2016) uses 19th century data on locations' density of railroads. Andersson et al. (2017) and Tsiachtsiras (2021) do so using the historical railroad expansion in Sweden and France. Relative to them, we contribute by exploiting a quasi-natural experiment that allows us to isolate a channel of face to face interaction, with little scope for a trade channel. In contemporaneous work Bai et al. (2021) estimate the elasticity of patent citations to air travel time using the introduction of new airline routes in a more recent period, post deregulation of the airline market. Relative to them, we contribute by exploiting a set up in which the risk for endogeneity of travel time is limited. Our work is related to

[5]Our dataset also includes international flights. We are currently digitizing more airlines to increase coverage both inside the US and internationally.

other literature which found that business travel affects innovation (Hovhannisyan and Keller (2015)), trade (Söderlund (2020)) and industrial activity (Coscia et al. (2020)). Also, air travel shapes collaboration between researchers (Catalini et al. (2020)).

The impact of transportation improvements in economic outcomes has long been a subject of study (Fogel (1963), Baum-Snow (2007), Michaels (2008), Donaldson and Hornbeck (2016), Jaworski and Kitchens (2019) and Herzog (2021)). Our convergence result contrasts with previous studies on improvements in other means of transport. Pascali (2017) finds that the introduction of steam engine vessels in the second half of the 19th century led to an increase in international trade which contributed to economic divergence between countries. Faber (2014) finds that the expansion of the highway system in China led to a reduction of GDP growth of peripheral counties, with evidence suggesting a trade channel. While both papers emphasize a trade channel, in our set up the trade channel would not be of first order. Hence, we uncover a new effect of improved connectivity.

Finally, we contribute to the literature on firm's location decision. Our result about firms deciding their establishments' locations based on travel time to headquarters is comparable to the one found by Giroud (2013), who finds that a reduction in air travel time to headquarters increases plant level investment and total factor productivity. Similarly, Campante and Yanagizawa-Drott (2017) finds that firms' cross country investment decision depends on connectivity to headquarters.

The paper is structured as follows. First, we present a simple theoretical framework which lays the foundations of how to think about the creation and diffusion of knowledge. The framework shows the two key parameters to estimate. Second, we describe the historical context in which jet airplanes were introduced. Third, we present the two datasets that we use: travel times and patents. Fourth, we perform the analysis to estimate the impact of travel time on the diffusion of knowledge, the creation of knowledge, and firm's location decision. Fifth, we conclude.

## 2. Conceptual framework

This section lays out a simple theoretical framework to think about the creation of knowledge. The framework clearly shows the two key parameters to estimate empirically: the elasticity of knowledge diffusion to travel time and the elasticity of knowledge creation to knowledge access.

Following Carlino and Kerr (2015) we consider a production function of knowledge which includes external returns in the form of knowledge spillovers. Knowledge output of a firm depends not only on firm's specific characteristics as its idiosyncratic productivity and input decisions, but also on an externality due to knowledge spillovers. We consider a production function of knowledge of the following form:

$$\text{New Knowledge}_{Fi} = f(z_{Fi}, inputs_{Fi}) \times \text{Knowledge Access}_i^{\rho} \tag{1}$$

where $\text{New Knowledge}_{Fi}$ is the knowledge created by firm $F$ located in $i$. The output of $Fi$ depends on an *internal* component and on an *external* component. The *internal* component is the firm's idiosyncratic productivity $z_{Fi}$ and choice of inputs $inputs_{Fi}$. The *external* component represents the externality to which all firms $F$ in location $i$ are exposed to: $\text{Knowledge Access}_i$. This externality, *Knowledge Access*, represents the total amount of knowledge spillovers that the firm is exposed to. The degree to which the externality affects the production of knowledge is governed by the parameter $\rho$. If $\rho$ is zero then knowledge spillovers have no effect on the creation of new knowledge. On the other hand, a positive $\rho$ implies that, keeping productivity and inputs constant, an increase in the level of knowledge spillovers leads to an increase in firm $F$'s creation of new knowledge.

A long standing literature studies the importance of knowledge spillovers for the creation of new knowledge.[6] The concept of knowledge spillovers goes back at least to

[6]The chapters of Audretsch and Feldman (2004) and Carlino and Kerr (2015) in the Handbook of Regional and Urban Economics provide an excellent review on the literature on knowledge spillovers,

Marshall (1890) who explains it as one of the agglomeration forces. Krugman (1991) refers to knowledge spillovers as one of the justifications for external increasing returns, and that the degree of spillovers are dependent on physical distance. The geographic decay of spillovers is grounded in the fact that not all knowledge is easy to codify, usually referred to as *tacit knowledge*, and geographic proximity increases the degree of knowledge spillovers by facilitating face to face interactions (Storper and Venables (2004), Glaeser (2011)). Hence, we consider the total amount of knowledge spillovers to which the firm $F$ in location $i$ is exposed to has the following functional form:

$$\text{Knowledge Access}_i = \sum_j \text{Knowledge stock}_j \times \text{distance}_{ij}^{\beta} \tag{2}$$

where $\text{Knowledge stock}_j$ is the total amount of knowledge in location $j$ (which is non-negative) that could potentially spill over to location $i$ and $\text{distance}_{ij}$ is a measure of distance from $j$ to $i$. The amount of knowledge that spills over from $j$ to $i$ depends on distance and the degree with which distance impedes spillovers, governed by the parameter $\beta$. If $\beta$ is zero, then distance does not affect knowledge spillovers from $j$ to $i$ and all locations perfectly share the same level of *Knowledge Access*. On the contrary, a negative $\beta$ implies a decay in knowledge spillovers when distance increases. In other words, a negative $\beta$ implies that if we reduce the distance from $j$ to $i$ while keeping every other distance constant, the amount of spillovers from $j$ to $i$ will weakly increase.

This theoretical framework bears resemblance to the concept of *Market Access* presented in Donaldson and Hornbeck (2016) and Redding and Venables (2004). If we interpret *Knowledge Access* as one of the inputs in the production function of knowledge, then $\text{Knowledge Access}_i$ could be interpreted as a measure of *Input Market Access*. This measure captures how cheaply firms in location $i$ can access pre-existing knowledge, where the cost of accessing knowledge depends on distance between $i$ and $j$. Also, *Knowledge Access* is similar to a measure of network centrality. The centrality of each location $i$ (node) is the weighted sum of distance (edges) to every location, where the

their geographic decay and how they affect the creation of knowledge.

weight of each location is given by its knowledge stock.

The theoretical framework highlights the two parameters to estimate: $\rho$ and $\beta$. Empirically, we use travel time as a measure of distance to first estimate $\beta$ and then conditional on $\beta$ we estimate $\rho$. Changes in travel time due to improvements in commercial aviation allow us to estimate both parameters. First, we use citations between patents as a proxy for the diffusion of knowledge. We estimate $\beta$ by relating changes in travel time between research establishments to changes in citations between them. Second, we use the stock of patents filed by inventors in each location as proxy for each location's stock of knowledge. We construct a measure of knowledge access using the patent stock, travel times and the value of $\beta$. New patents in each location proxy for new knowledge. Changes in travel time lead to changes in knowledge access which allow us to estimate $\rho$.

# 3. Historical context

## 3.1. Air transport: jet arrival

The jet aircraft was first invented in 1939 for military use, with the German Heinkel He 178 being the first jet aircraft to fly. The first commercial flight by a jet aircraft was in 1952 by the British Overseas Airways Corporation (BOAC) from London, UK to Johannesburg, South Africa with a Havilland Comet 1. Nonetheless, given the amount of accidents of the Havilland Comet 1 due to metal fatigue, jet commercial aviation did not truly take off until the Boeing 707 entered commercial service in late 1958. The 24th of January of 1959 represented a major milestone in the jet era: American Airlines Flight 2 flew with a Boeing 707 jet aircraft from Los Angeles to New York, the first non-stop transcontinental commercial jet flight.[7]

[7] The reader passionate of aviation history would enjoy reading the following New York Times article which tells the experience of the first transcontinental jet flight: `https://www.nytimes.com/2009/01/26/nyregion/26american.html`

In 1951 New York City and Los Angeles were connected with a one-stop flight in 10 hours and 20 minutes. The flight had a forced stop in Chicago and was operated with the propeller aircraft Douglas DC-6, which had a cruise of 500 kmh. By 1956, New York City and Los Angeles were connected with a non-stop flight in 8 hours and 30 minutes. This was accomplished due to the introduction of the propeller aircraft Douglas DC-7 which had a cruise speed of 550kmh, and a change in regulation which increased maximum flight time of a crew from 8 to 10 hours within a 24-hour window.[8] In 1961, the route was covered with the jet aircraft Boeing 707 in a non-stop flight in 5 hours 15 minutes, reaching 5 hours 10 minutes in 1966. The Boeing 707 had a cruise speed of 1000kmh, cutting travel time from New York City to Los Angeles in half between 1951 and 1966.

## 3.2. Air transport: moving people, not goods

During the 1950s and 1960s, air transportation served to transport people but not goods. Figures 1 and 2 are images (edited for better readability) from annual reports of the Interstate Commerce Commission of 1967 and 1965 respectively. Figure 1 displays the amount of passenger-miles[9] for Air, Motor and Rail transportation from 1949 to 1966. We observe that, while transport of people by rail decreased and by motor remained relatively constant, transport of people by air multiplied by 6 in a 16-year period, which translates to around 12% compound annual growth. In 1966, air transport accounted for more passenger-miles than both rail and motor transportation together, reflecting the growing importance of this mean of transport.

[8] AA and TWA had transcontinental non-stop propeller flights scheduled since at least 1954. These flights were scheduled to take 7 hours 55 minutes, just under the maximum flight time allowed by regulation in domestic flights: regulation impeded pilots from being on duty more than 8 hours within a 24 hours window. Nonetheless, the propeller aircrafts used in these flights, the Douglas DC-7 and the Lockheed Super Constellation, overheated their engines due to excessive demand to cover the route in less than 8 hours. AA fought intensely until the CAB approved a waiver that allowed non-stop transcontinental flights to take up to 10 hours to accomplish the non-stop transcontinental flight. See page 16 of the edition of the 21st of June 1954 of the Aviation Week magazine `https://archive.org/details/Aviation_Week_1954-06-21/page/n7/mode/2up`

[9] Passenger-miles is a standard unit of measurement in transport, where one passenger-mile accounts for one person traveling one mile. The reasoning is the same for ton-miles, with one ton of goods traveling one mile.

Figure 2 shows shipments in ton-miles for the period 1939 to 1964 by mean of transport: Airways, Pipelines, Inland Waterways, Motor, Railroads. Interestingly, we observe that air transport of goods, even if it increased, it accounted for less than 0.1% of transport of goods in 1964.[10]

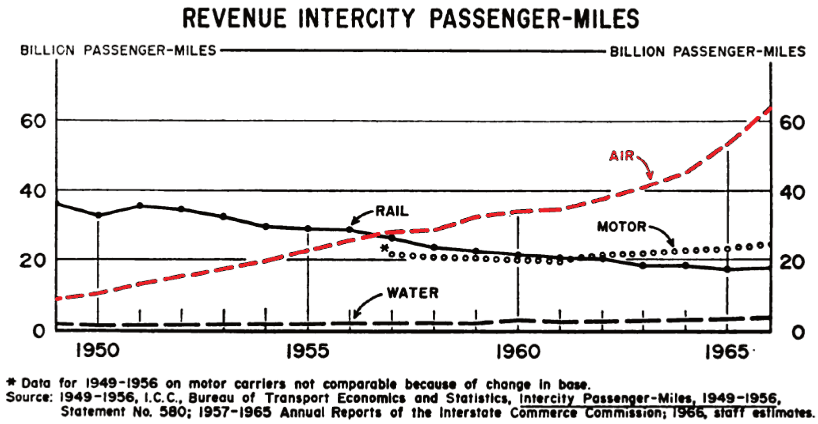



Figure 1: Passenger Miles
Source: Interstate Commerce Commission, Annual Report 1967
Edited by the authors

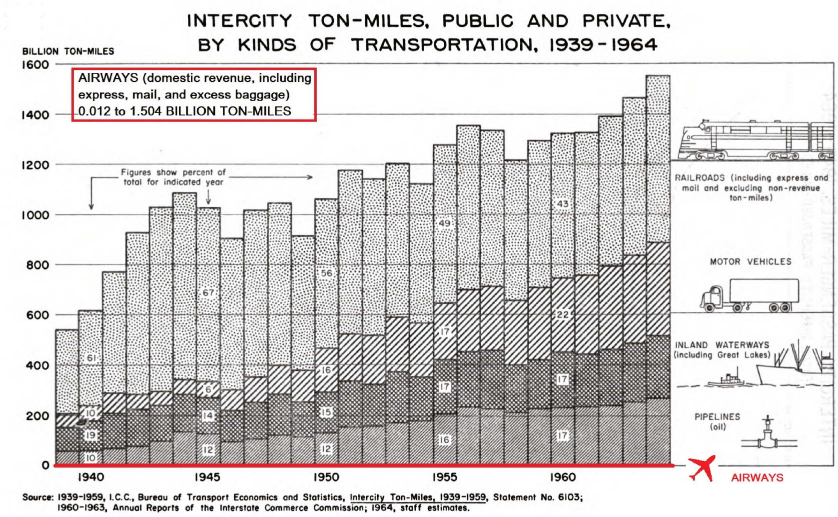



Figure 2: Ton Miles
Source: Interstate Commerce Commission, Annual Report 1965
Edited by the authors

[10]We have not found data about shipments by mean of transport measured in monetary values.

## 3.3. Regulation

As explained in Borenstein and Rose (2014), in the 1930s the airline industry was seen as suffering from coordination issues, destructive competition and entry. Additionally, the industry was developing in a context of financial instability and increasing military concerns post Great Depression. A strong domestic airline industry was perceived as an interest of national defense. As consequence, the Civil Aeronautics Board (CAB) was created in 1938 with the objective to promote, encourage and develop civil aeronautics.[11] It was empowered to control entry, fares, subsidies and mergers.[12] In other words, the CAB regulated the market by deciding which airlines could fly, in which routes they could operate, the price that they charged in each route, the structure of subsidies and merger decisions. The CAB regulated the airline industry in a barely unchanged manner until it ceased to exist in 1985.

When the CAB was created, it conceived special rights to the existing airlines over the connections they were operating. The CAB did not permit entry of new airlines on interstate routes and gradually allowed current airlines to expand their routes. The CAB controlled both the system and each airline's network. The network was designed to maintain industry stability and minimize subsidies, leading to a system where each route was mainly operated by one or two airlines.[13] Importantly, Borenstein and Rose (2014) in pages 68-69 explain that *"the regulatory route award process largely prevented airlines from reoptimizing their networks to reduce operation costs or improve service as technology and travel patterns changed."* As a consequence, any technological improvement such as increases in aircraft speed, capacity or range would not affect each airline's flight network in the short term.

[11]The CAB was a federal agency hence, in principle, would not have control over intrastate routes. Nonetheless, according to Borenstein and Rose (2014) the CAB managed to have some intrastate markets under its control using legal arguments.

[12]Safety regulation was under the control of the Federal Aviation Administration.

[13]Borenstein and Rose (2014) in page 68, based on Caves (1962), expose *"In 1958, for example, twenty-three of the hundred largest city-pair markets were effectively monopolies; another fifty-seven were effectively duopolies; and in only two did the three largest carriers have less than a 90 percent share."*

By regulating fares, the CAB explicitly encouraged airlines to adopt new aircraft. Airlines, when operating an older aircraft, would apply for a fare reduction arguing that it is needed in order to preserve demand for low quality service. The CAB would refuse this application, hence airlines would have to adopt new aircraft or risk losing consumers who would choose another airline which flies newer aircrafts.

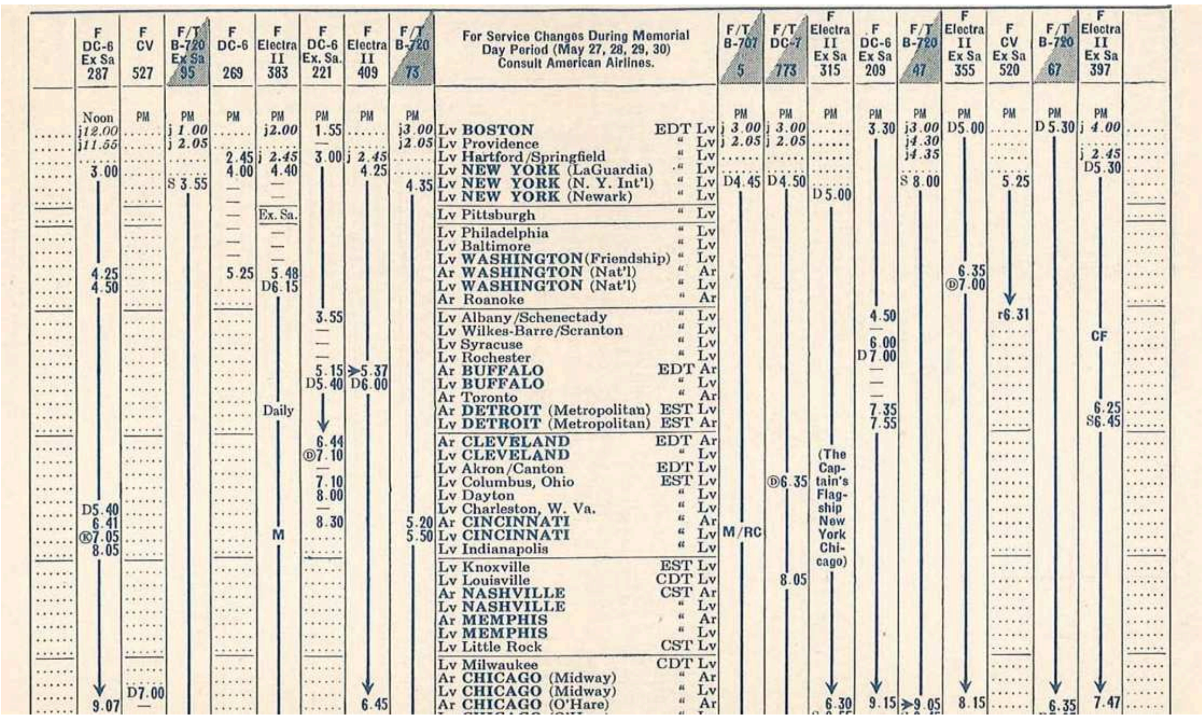
## 4. Air travel data

We construct a new data set of the flight network in the United States during the 1950s and 1960s. We collected and digitized information of all the flights operated by the main airlines and obtained the fastest route and travel time between every two airports in the network.

To construct the flight network we use historical flight schedules of the main airlines operating in 1950s and 1960s. Figure 3 is a fragment from an example page of the 1961 flight schedule of American Airlines. In the flight schedule we observe in the center column the name of departure and arrival cities (which we match to airports using airlines' historical ticket office geographical location), while the small columns on the sides depict flights. In the top of the small columns we observe the type of service provided (first class, coach or both), aircraft operated, days operated (daily if information is missing) and flight number. The content of the small columns displays the departure and arrival time (local time, bold numbers represent PM) at each city, including all intermediate stops.

We digitize flight schedules for the years 1951, 1956, 1961 and 1966 of six domestic airlines: American Airlines (AA), Eastern Airlines (EA), United Airlines (UA), Trans World Airlines (TWA), Braniff International Airways (BN), Northwest Airlines (NW),[14]

[14] These are six of the fifteen trunk (interstate) airlines operating in 1951. Many of the remaining trunk airlines would merge with another trunk airline over the years, and there would be zero entry of new airlines. We are currently digitizing the remaining trunk airlines and we plan to add them to the travel time dataset in the future. We have already digitized: Allegheny Airlines, Capital Airlines, Colonial Airlines, Continental Airlines and Delta Air Lines. We have also digitized the year 1970 for


For Service Changes During Memorial Day Period (May 27, 28, 29, 30) Consult American Airlines.

Lv BOSTON EDT Lv
Lv Providence " Lv
Lv Hartford/Springfield " Lv
Lv NEW YORK (LaGuardia) " Lv
Lv NEW YORK (N. Y. Int'l) " Lv
Lv NEW YORK (Newark) " Lv
Lv Pittsburgh " Lv
Lv Philadelphia " Lv
Lv Baltimore " Lv
Lv WASHINGTON (Friendship) " Lv
Ar WASHINGTON (Nat'l) " Ar
Lv WASHINGTON (Nat'l) " Lv
Ar Roanoke " Ar
Lv Albany/Schenectady " Lv
Lv Wilkes-Barre/Scranton " Lv
Lv Syracuse " Lv
Lv Rochester " Lv
Ar BUFFALO EDT Ar
Lv BUFFALO " Lv
Ar Toronto " Ar
Ar DETROIT (Metropolitan) EST Lv
Lv DETROIT (Metropolitan) EST Ar
Ar CLEVELAND EDT Ar
Lv CLEVELAND " Lv
Lv Akron/Canton EDT Lv
Lv Columbus, Ohio EST Lv
Lv Dayton " Lv
Lv Charleston, W. Va. " Lv
Ar CINCINNATI " Ar
Lv CINCINNATI " Lv
Lv Indianapolis " Lv
Lv Knoxville EST Lv
Lv Louisville CDT Lv
Ar NASHVILLE CST Ar
Lv NASHVILLE " Lv
Ar MEMPHIS " Ar
Lv MEMPHIS " Lv
Lv Little Rock CST Lv
Lv Milwaukee CDT Lv
Ar CHICAGO (Midway) " Ar
Lv CHICAGO (Midway) " Lv
Ar CHICAGO (O'Hare) " Ar

Figure 3: Fragment of flight schedule American Airlines 1961
The center column displays the name of departure and arrival cities. The small columns on the sides display flights with departure and arrival time (local time, bold numbers represent PM). The top of the small columns shows the type of service provided (first class, coach or both), aircraft operated, days operated (daily if information is missing) and flight number.

and one international airline: Pan American Airways (PA). This group of airlines includes the *Big 4*: AA, EA, UA and TWA, which accounted for between 69% and 74% of interstate air revenue passenger miles in the US in the years collected. BN and NW were digitized in order to have a wide geographical coverage, while PA provides international flights. Based on C.A.B. (1966), in the years collected, the six domestic airlines together accounted for between 77% and 81% of interstate air revenue passenger miles.

In total we have digitized 6,143 US flights (unique combinations of flight number-year, 7,007 worldwide). However, flights often have multiple stops. If we count each non-stop part (*leg*) of these flights separately, our sample contains 17,737 legs in the US

the six airlines used in this paper and Pan American. Due to a time constraint we have not included them in the current analysis. We plan to digitize BOAC to obtain more international flights, and to cover the time period 1930 to 2000 for all airlines that is possible.

and 21,210 worldwide. Our data connects 275 US airports (434 worldwide) creating 2,563 unique origin-destination (directional) airport links (3,466 worldwide). Figure 4 displays the flight network in continental United States pooling all years together. In Appendix A.2 we show the US flight network by year, around 80% of the non-stop flights remain year-on-year.

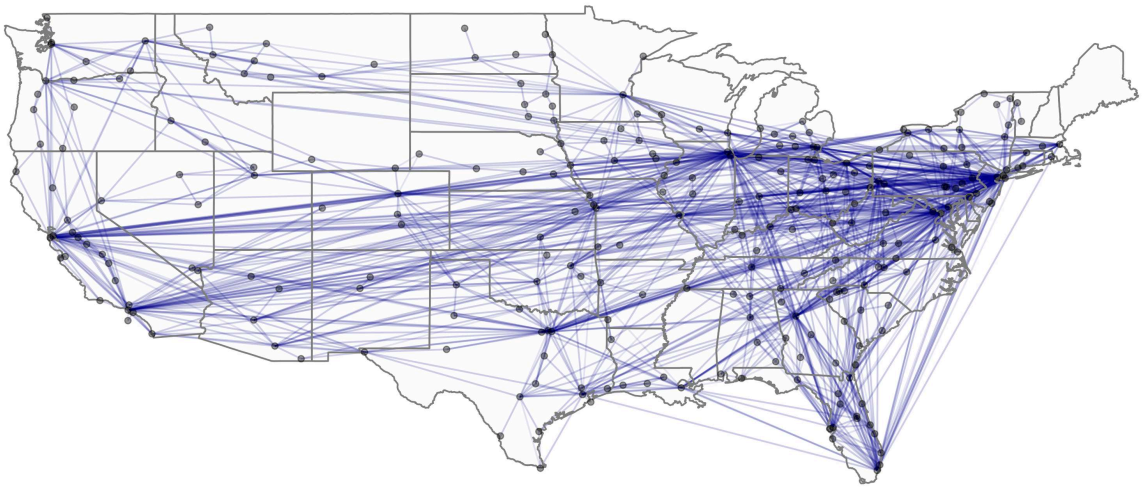


Figure 4: United States flight network 1951-1966

Using departure and arrival time of each flight at each airport, we obtain the fastest route and corresponding travel time between every two airports in our data. To obtain the fastest route and travel time we modify the Dijkstra algorithm to account for layover time in case the fastest route includes connecting flights.[15]

Once the fastest route between every two airports is computed, we match every airport to 1950 Metropolitan Statistical Areas (MSA) using the shape file from Manson et al. (2020). We consider only MSAs in contiguous United States. We use MSAs as the geographical unit of analysis because they are constructed taking into account

[15]We are currently working on setting a minimum waiting time for switching airplanes, such that the change is not permitted unless waiting time is more than the minimum. For the time being we have set the minimum waiting time to zero, meaning that in our calculation one passenger would be able to switch from one airplane to another if departure of the following flight is one minute later than arrival of the previous flight. This a rather implausible assumption and we are estimating the minimum waiting time in each airport depending on the airport's congestion.

commuting flows. We assume that people in an MSA would use, for each desired route, the most appropriate airport lying inside or nearby the MSA. We match each airport to all MSAs for which it lies inside the MSA or is at most 15km away from its boundary.[16] 176 out of 275 US airports are matched to at least one MSA. Meanwhile, 142 out of 168 MSAs are matched to one or more airports in at least one year, and 108 MSAs are matched to one or more airports in the four years. We use the sample of 108 MSAs that had an airport in the four years as our baseline travel time data.[17]

## 4.1. Descriptive statistics: Air travel

To understand the changes in travel time we will first study travel time of non-stop flights and then of all routes including connecting flights. Figure 5 displays the non-stop fastest flight within each MSA pair that was operating in each year. In 1951 the longest non-stop flight across MSAs was between Chicago and San Francisco using the Douglas DC-6, covering a distance of 2,960 km in 7 hours 40 minutes. This travel time was just under 8 hours, the maximum flight time allowed for a crew in a 24-hour period.[18] In 1956, new regulation allowed up to 10 hour flights for transcontinental flights, the longest non-stop flight between MSAs was New York to San Francisco with the Douglas DC-7, covering a distance of 4,151 km in 9 hours. Between 1951 and 1956, while we observe an increase in average flight speed that went up to 17%, the main change observed is that longer non-stop routes were possible.

In 1961, the first year in which we have jet aircrafts in the travel time data, there is a reduction in travel time between MSA-pairs, especially for those far apart from each

[16]The 15km distance was chosen after inspecting airports near the border that should arguably be matched, as for example, Atlanta ATL airport.

[17]In Appendix A.2 we include a table with the 168 MSAs, those connected at least once and those connected in the four years. Among the MSAs not connected is San Jose, California, which in our patent sample accounted for around 2% of patents. San Jose had an airport (SJC) during our time period but it was not served by any of our airlines, so it is not included in our analysis. In the future we plan to include the currently non-connected MSAs by matching them to airports that may have served them and accounting for the commuting travel time.

[18]Honolulu was not concerned by the regulation. Honolulu was connected with non-stop flights to San Francisco (9 hours 40 minutes), Los Angeles (11 hours) and Portland (12 hours 55 minutes).

other. In 1966, there is a further decrease in travel time due to a widespread adoption of jet aircrafts in shorter distances. In Appendix Figure 22 we show the jet adoption rate by distance for MSAs connected with a non-stop flight. All MSA-pairs more than 3,000km apart connected with a non-stop flight operated at least one jet flight in 1961, and this expanded to all those more than 2,000km apart in 1966. The speed gain of jets relative to propeller aircrafts is increasing with the amount of time that the jet can fly at its cruise speed, arguing in favor of an adoption that is increasing with the distance between origin and destination.[19]

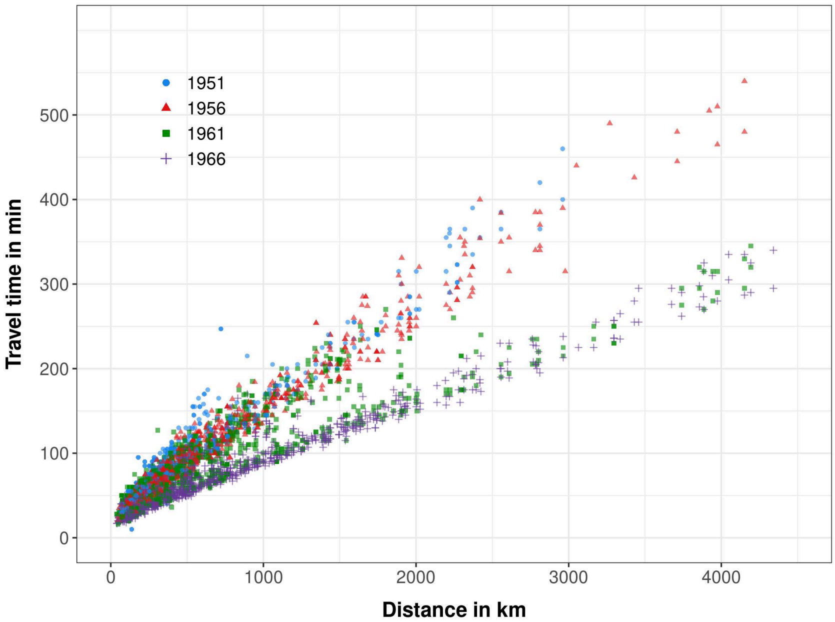



Figure 5: Non-stop fastest flights United States MSAs

The change in travel time in non-stop flights is also reflected in the travel time for connecting flights. Figure 6 shows, relative to 1951, the average and standard deviation change in travel time for all MSA-pairs, including non-stop and connecting flights.[20]

[19]We are currently exploring the differential timing of jet adoption across airlines. Differences in (pre-existing) route distance and past contractual relationships with aircraft suppliers potentially led to different adoption rates at each time period. For example, Eastern Airlines' routes were particularly shorter than for other airlines. Also, those committed to buy Douglas airplanes (the leader US commercial aircraft supplier pre-jet era) would have adopted jets later, as Douglas launched jet airplanes later than Boeing.

[20]The plot includes only MSA-pairs with travel time in all time periods. The standard deviation for MSA-pairs less than 250km apart is big relative to the ones at other distances. Hence we decided not

Between 1951 and 1956, there is an average reduction in travel time of 9.2% which is roughly constant for all distances over 500km. Between 1951 and 1961, there is a reduction in travel time that is increasing with distance. The average decrease in travel time is of 16.8%, while the reduction is of 29.4% for a distance of more than 2,000km and 39.2% for a distance of 4,250-4,500km. Between 1951 and 1966, there is an even stronger decrease in travel time at all distances. The average reduction in travel time is 28.7% across all distances, 40.8% for a distance of more than 2,000km and 48.4% for a distance of 4,250-4,500km. The increased adoption of jets for short distance flights implied that both non-stop flights at short distance and connecting flights at farther distance had a decrease in travel time.

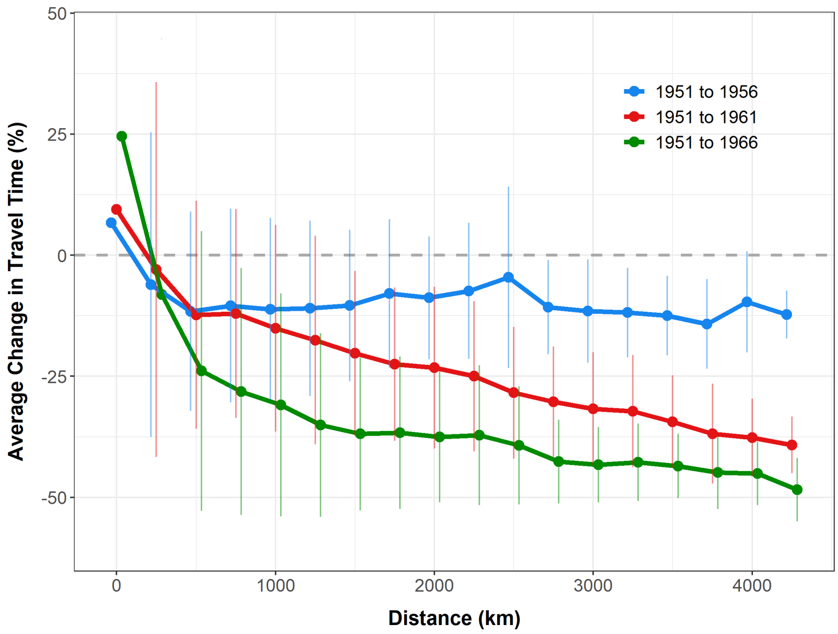



Figure 6: Change in MSAs travel time

Figure 25 in Appendix A.2 shows that the change in travel time is accompanied by a reduction of the amount of legs needed to connect two MSAs at every distance. This reduction is especially marked between 1951 and 1956, and 1961 and 1966. Between 1956 and 1961, we do not observe a big reduction in the amount of legs, implying that the decrease in travel time observed in Figure 6 between 1956 and 1961 comes

to include it because it distorts the visualization of the rest of the plot.

from a source other than the amount of legs. In Appendix Figure 26 we open up the change in travel time by the way an MSA pair was connected in 1951 and 1966: either directly (non-stop flight) or indirectly (connecting flight). We observe that much of the increase in travel time for MSA pairs less than 250km apart comes from routes that in 1951 were operated non-stop while in 1966 were operated with connecting flights.[21] Interestingly, for MSA-pairs more than 2,000km apart travel time reduced on average 42% for those pairs that were connected indirectly in both periods, and 51% for those that switched from indirect to direct. This fact shows the relevance of improvements in flight technology even for MSAs that were not directly connected.

It could be the case that a reduction in the amount of legs or an increase in frequency of flights reduces layover time, which then translates into a reduction of travel time. In Appendix Figure 28 we compare the change in travel time from 1951 to 1966 with a counterfactual change in travel time in which we eliminate layover time in both time periods. We observe that the average change in travel time is stronger at every distance in the counterfactual scenario without layover time. This implies that the relative importance of layover time to total travel time within a route increased between 1951 and 1966, so total travel time did not decrease proportionally to the change of in-flight travel time. In short, layover time attenuated the reduction in travel time.

## 4.2. Constructing an instrument

In this section we construct an instrumental travel time that is based on the pre-existing flight routes and the time-varying nationwide roll out of jets. In this way, the instrument abstracts from the endogenous decisions of two agents: First, regulator's decision on the opening/closure of routes. Second, airlines' decision about to which routes allocate

[21] Appendix Figure 27 repeats the exercise discarding layover time in all time periods. By comparing Figure 26 and Figure 27 we can disentangle the effect of layover time and the change in in-flight time. For MSA pairs less than 250km that changed from direct to indirect connection, 80% of the increase in travel time is due to the increase in layover time (which was previously zero as it was a non-stop flight), and 20% is due to the increase of in-flight time.

jet vs propeller airplanes and scheduling (frequency of flights and layover time). We first explain the idea and identifying assumptions of the instrument, and then we detail how it is constructed.

In Borenstein and Rose (2014) it is argued that, due to strict regulation, it was difficult for airlines to adapt their flight network when technology to fly changed. However, we may be concerned that the decision of the regulator to grant new routes could be targeted to specific pairs or correlated with unobservable variables that also affect the creation and diffusion of knowledge.[22] Hence, as the first step in the construction of our instrument, we *fix routes* to the ones we observe in 1951. In this way the instrumental travel time is computed only using non-stop flights present in 1951, and does not consider appearance or disappearance of non-stop flights in the data. The identifying assumption is that the network of flight routes in 1951 did not yet include the changes that would be optimal to operate with jet airplanes. In other words, we require that the regulator did not change routes already by 1951 in anticipation of the arrival of jet airplanes.[23,24]

Airlines could decide on two factors that affect travel time: the type of airplane (jet vs. propeller) operated in each route and scheduling, which consists on the frequency of flights and layover time in case of connecting flights.[25] We may be concerned that, as with the regulator, airlines' decisions could be correlated with unobservables that also affect the creation and diffusion of knowledge.[26] The second step in the construction of

[22] For example, the regulator could have targeted the opening of new routes between places in order to boost their economic activity.

[23] For example, in the instrument there are no non-stop transcontinental routes.

[24] In our estimations we exploit time variation for identification. Hence, if pre-existing routes affect the levels at the origin-destination level, this does not drive our identification. However, we may be concerned that pre-existing routes could affect future growth and not only levels. To address this concern, in robustness analysis we estimate the elasticity of citations to travel time using only MSA-pairs that are always indirectly connected. Results go in the same direction.

[25] In 1961, all non-stop flights of more than 3,000km had at least one jet operating within them, while in 1966 it was the case in all non-stop flights of more than 2,000km. Therefore the endogeneity of jet adoption is a smaller concern for long distance flights.

[26] For example, airlines may have decided to prioritize the allocation of jets to routes which had a higher share of business travel, which may be correlated with the diffusion of knowledge.

our instrument is to discard layover time (hence discarding all scheduling decisions) in all time periods, and assume that in each year all routes are operated with a *fictitious average airplane* of the year. Hence, the change in instrumental travel time in a route is independent of the type of airplane used in the route and it only depends on the nationwide roll out of jets. The identifying assumption is that no single route had the power to shift the average speed of the year.

To construct the instrumental travel time we first estimate, separately for each year, a linear regression of travel time on flight distance using only the fastest non-stop flight in each origin-destination airport pairs.[27] These yearly regressions provide us with the fictitious average airplane of each year: the intercept gives the take-off and landing time of the airplane while the slope provides the (inverse) speed. Second, we fit these regressions to obtain predicted travel time in each non-stop flight and year. Third, for each year, we compute the fastest travel time using the Dijkstra algorithm. The Dijkstra algorithm looks for the fastest path using only 1951 non-stop flights, while the travel time in each non-stop flight in each year is given by the predicted travel time from the previous step. Layover time is set to zero in all years.

Figure 7 shows the percentage change in observed and instrumental travel time relative to 1951. We compute the percentage change within each MSA-pair for each year and then take averages within 250km bins. We observe that the instrumental travel time follows pretty closely the observed change in travel time in each year. Especially, it replicates the pattern of a stronger decrease in travel time for MSAs located farther apart. It is only for MSAs less than 250-500km apart that the change in the instrumental travel time departs from the observed change.[28] This finding shows that most of the change in travel time that we observe is due to the change in speed of airplanes, and

[27]The use of a linear regression is motivated by the linearity between travel time and distance displayed in Figure 5. To estimate these regressions we use all routes appearing in each year.

[28]We observe an increase in travel time for short distances in 1961 relative to 1951. Given that non-stop routes are fixed and that for longer distances there is a decrease in travel time, the increase in travel time in short distances comes from an increase in the value of the intercept relative to the slope in 1961, relative to 1951.

that the endogeneity concern is limited for MSAs located far away from each other.

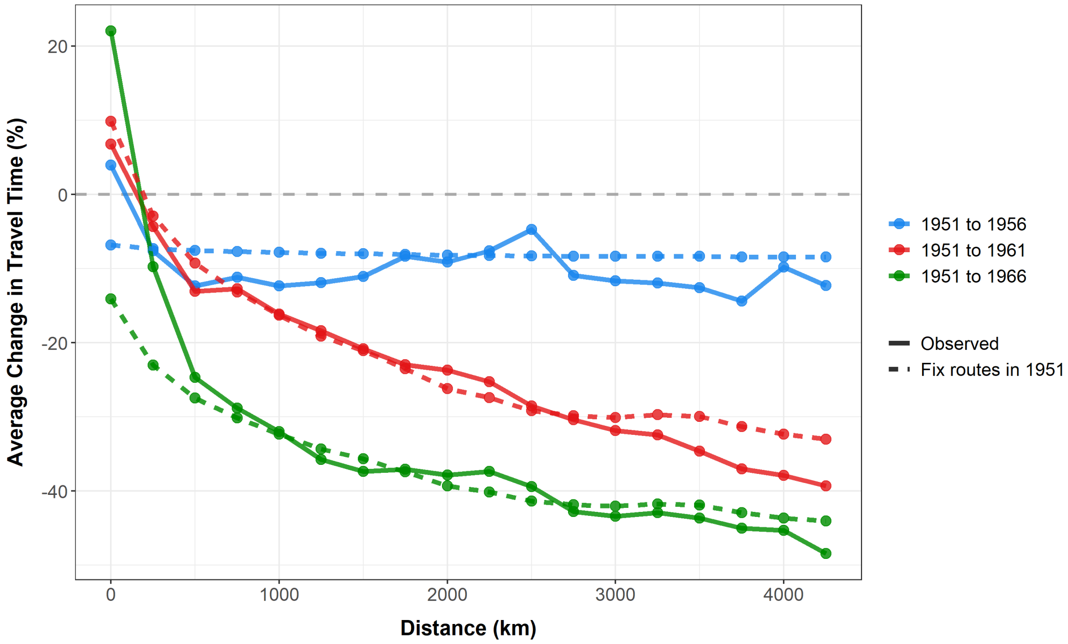



Figure 7: Instrumental Travel Time between US MSAs.

In Appendix A.2 we present other two counterfactual travel times: one in which we fix airplanes to be the average airplane of 1951 and allow routes to evolve, and another in which both the average airplane and routes are varying. These two counterfactuals together with the one presented in this section allow us to decompose the change in travel time by the change in routes and the change in speed of airplanes. We obtain that around 90% of the change in travel time is due to the change in speed of airplanes, while around 10% of the change is due to the change in the flight routes. Appendix Figure 30 shows that the share is roughly constant for all distances. This finding confirms that most of the observed changes in travel time are due to improvements in flight technology.

# 5. Patent data

We use patent data as our source of innovation information. We construct a dataset of all patents granted by the United States Patent and Trademark Office (USPTO) with filing year[29] between 1949 and 1968, which includes for each patent: filing year, technology classification, location of the inventors when they applied for the patent, owner of the patent and citations to other patents also granted in the United States. This dataset provides the distribution of patents and citations over the geographic space, allowing to take into account ownership structure.

To construct the patent dataset we downloaded from Google Patents all patents granted by the USPTO with filing year between 1949 and 1968. This dataset contains patent number, filing year and citations.[30,31] Based on the patent number we merge it with multiple datasets. First, we obtained technology class from the USPTO Master Classification File[32] and we aggregated them to the six technology categories of Hall et al. (2001). Second, we obtained geographic location of inventors from three datasets: HistPat (Petralia et al. (2016)) and HistPat International (Petralia (2019)) for patents published until 1975, Fung Institute (Balsmeier et al. (2018)) for patents published after 1975.[33] We match all inventors' locations to 1950 Metropolitan Statistical Areas (MSAs) in contiguous United States. To do the match we obtain geographical coordinates from

[29]Filing year, also called application year, is the closest date to the date of invention that is present in the data and it represents the date of the first administrative event in order to obtain a patent. In the other hand, the publishing (also called granting year) is a later year in which the patent is granted. The difference between filing and publishing year depends on diverse non-innovation related factors (as capacity of the patent office to revise applications) and changes over time. Hence filing year is the date in our data that approximates the best to the date of invention.

[30]Very few patents had missing information on filing year. We complemented both missing filing year and citations with the OCR USPTO dataset.

[31]We note that the patent citation record starts in 1947, year in which the USPTO made it compulsory to have front page citations of prior art. Gross (2019)

[32]`https://www.google.com/googlebooks/uspto-patents-class.html`

[33]Due to the gap between the filing year and publishing year we also do the matching to patents published after 1968. Our underlying patent data actually covers a longer time period of filing years, which we need for example to construct forward and backward citation lags. However, there are limitations to use the geographic data in filing years 1971-1972. In Appendix B we show that during filing years 1971-1972 the rate of unmatched patents to inventors' location increases. This is probably due to Histpat and Fung data not being a perfect continuation one of the other.

the GeoNames US Gazetteer file and Open Street Maps, and use the MSAs shape file from Manson et al. (2020). Third, we obtain ownership of patents from two sources: Kogan et al. (2017) for patents owned by firms listed in the US stock market and Patstat (Magerman et al. (2006)) for the remaining unmatched patents.[34]

For the descriptives presented below and the posterior analysis we truncate and aggregate the data in the following way. We drop patents that are owned by universities or government organizations. To count patents that are classified into multiple technology categories, we do a fractional count by assigning proportionally a part of the patent to each category. Citations are counted as the multiplication of the technology weight of the citing and cited patents. We drop patents (and their citations) that have inventors in multiple MSAs[35] and citations in which the citing owner is the same as the cited owner.[36]

We aggregate the patent data to 4 time periods of 5 years each, with the center of each period being the year of travel time data collected. The periods are: 1951 (which contains the years 1949-1953), 1956 (1954-1958), 1961 (1959-1963) and 1966 (1964-1968). We consider only patents in MSAs that are matched to an airport in the four periods.[37] The final dataset contains 108 MSAs with patents and travel time.

## 5.1. Descriptive statistics: Patents

This section presents three facts about US patents over our sample period: First, initially less innovative locations had a higher patenting growth rate. The average yearly

[34]Patent ownership in both datasets comes from the patent text, which is self declared by the patent applicant. Particularly, Kogan et al. (2017) does not explicitly state if it takes into account firm-ownership structure to determine the ultimate owner of a patent, neither does Patstat.

[35]Working with multi-MSA patents requires an assumption on how to compute distance and travel time between the citing and cited patents, as it is not a single origin-destination location pair. We hence prefer to abstract from multi-MSA patents. In the other hand, collaboration of inventors located in different MSAs is a interesting subject and it is part of our research agenda.

[36]Incentives to self-cite may be different than to cite patents of other owners.

[37]We drop around 9% of patents that are in MSAs which are not matched to an airport in the four time periods. Descriptive statistics including those patents are similar to the ones presented here.

growth rate of locations in the lowest quartile of initial innovativeness was 7.2% while it was 1.9% for those in the highest quartile. High growth locations were also primarily in the South and the West of the US. The South and the West grew three times as fast as the Midwest and the Northeast. Second, over time firms grew larger as measured by the amount of MSAs in which they had research establishments. At the same time, the share of patents filed by large multi-establishment firms increased. The amount of firms with research establishments in more than 10 MSAs almost tripled over the time period and their share of patents doubled. Third, the mass of citations shifted towards longer distances. While the first quartile of citation distance remained relative stable over the time period, the third quartile increased its distance by 39%. At the same time, the share of citations at more than 2,000km increased by 30%.

We compute descriptives by technology. In here we present descriptives of averages across technologies. Technology specific descriptives are included in Appendix B.

**Fact 1.a.: Initially less innovative locations had a higher patenting growth rate**

In the period 1951 to 1966 we observe that the highest growth of patenting takes place in locations that were initially less innovative. The differential growth rate implies a convergence rate of 5.3% per year.

Figure 8 shows the geographic distribution of patenting in 1951. Darker colors refer to a higher level of *initial innovativeness*, which is defined as the amount of patents filed by inventors in the MSA in 1951.[38] We observe that MSAs in the top quartile of patenting are concentrated in the Northeast (which includes New York) and the

[38] To compute the level of initial innovativeness we only use patents filed in 1951 (years 1949-1953). We aggregate patents to the MSA-technology level and then compute the quantile-position of each MSA in the technology. Lower values of quantile-position refers to lower amount of patents in the technology (relative to other MSAs). Each MSA has a different value of quantile-position in each of the 6 technology categories. To obtain the MSA level quantile we take the average quantile across technologies within the MSA. Finally we classify MSAs into quartiles depending on whether the average quantile is higher or lower than the thresholds 0.25, 0.50, 0.75.

Midwest (which includes Chicago), with few additional MSAs in the West.[39,40]

Figure 9 shows the geographic distribution of patenting growth in 1951-1966.[41] We observe a striking pattern relative to Figure 8: high growth MSAs were those that were initially less innovative. High growth happens in initially less innovative locations in the South and the West but also in the Northeast. We confirm this pattern in Figure 10, which shows the MSA's ranking of innovativeness in 1951 and its subsequent patenting growth rate in 1951-1966. Figure 10 shows that MSAs that were initially more innovative (lower values in the ranking) are those that saw lower values of subsequent patenting growth.[42,43] We estimate a linear regression with an intercept and a slope, and find that the slope is positive and statistically different from zero. At the mean, lowering initial innovativeness by 10 positions in the ranking was associated with a subsequent 0.42 percentage points higher yearly growth rate of patenting.

Figure 10 presents average growth rates across technologies within a MSA. We obtain a result that goes in the same direction if we compute the average growth rates across MSAs within a technology and quartile of initial innovativeness, and then take the average across technologies. The average yearly growth rate of MSA-technologies in the lowest quartile of initial innovativeness is 7.2% while it is 1.9% in the highest quartile.[44] The percentage point difference between the two growth rates implies that

[39] In Appendix B we show that the 1951 geographic distribution of patents looks similar across technology categories.

[40] The top 5 patenting MSAs in 1951 were: New York City (25% of all patents), Chicago (11%), Los Angeles (8%), Philadelphia (6%) and Boston (4%).

[41] We compute the growth rate of patenting in each technology within a MSA and then take the average across technologies within the MSA.

[42] Each dot in Figure 10 is an MSA. To compute the MSA ranking we need to double-rank MSAs. First we rank all MSAs in each technology. Second we take the across-technology average ranking of each MSA. Third we rank all MSA's averages. To compute the MSA's yearly growth rate we first take the 1951-1966 growth rate for each technology in the MSA. We then take the average across technology. Finally we obtain the MSA's yearly growth rate by computing: $yearly\_growth\_rate = (1 + 19\_year\_growth\_rate)^{(1/19)} - 1$ (the 1951 to 1966 period is a 20 year window, we take growth rates as being from the first year 1949 to the last one 1968, which is 19 year growth).

[43] In Appendix B we show replicate the plot differentiating geographic regions. MSAs that were initially less innovative and had high subsequent growth were located in all four regions, although they were primarily located in the South and the West.

[44] We first compute the 1951-1966 growth rate (19-year growth rate) for each MSA-technology. We then

locations in the lowest quartile converged towards locations in the highest quartile at a speed of 5.3% per year.[45] The convergence in patenting across MSAs is consistent with *The Postwar Decline in Concentration, 1945-1990* described in Andrews and Whalley (2021).

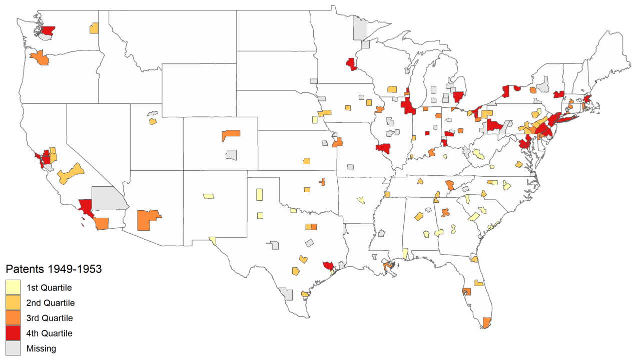



Figure 8: Geography of Patenting 1951

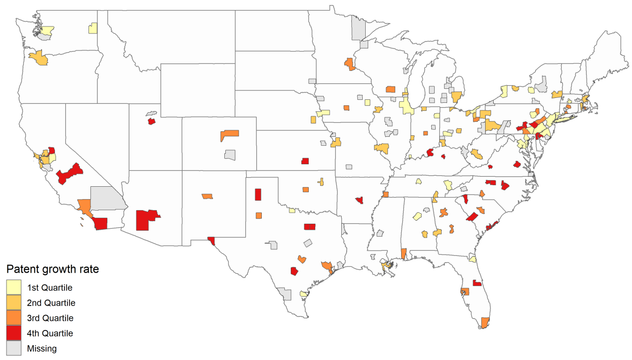



Figure 9: Patent growth 1951-1966

take averages across MSAs within a quartile-technology, and after take averages across technologies within a quartile. Finally, we convert the 19-year growth rate into an average yearly growth rate.

[45]We note that the aggregate growth of patents is much smaller than the across MSAs unweighted average, and this is exactly because initially less innovative MSAs grew faster. If we compute the growth rate in nationwide amount of patents in each of the technologies and then average across technologies we obtain a yearly growth rate of 1.5%.

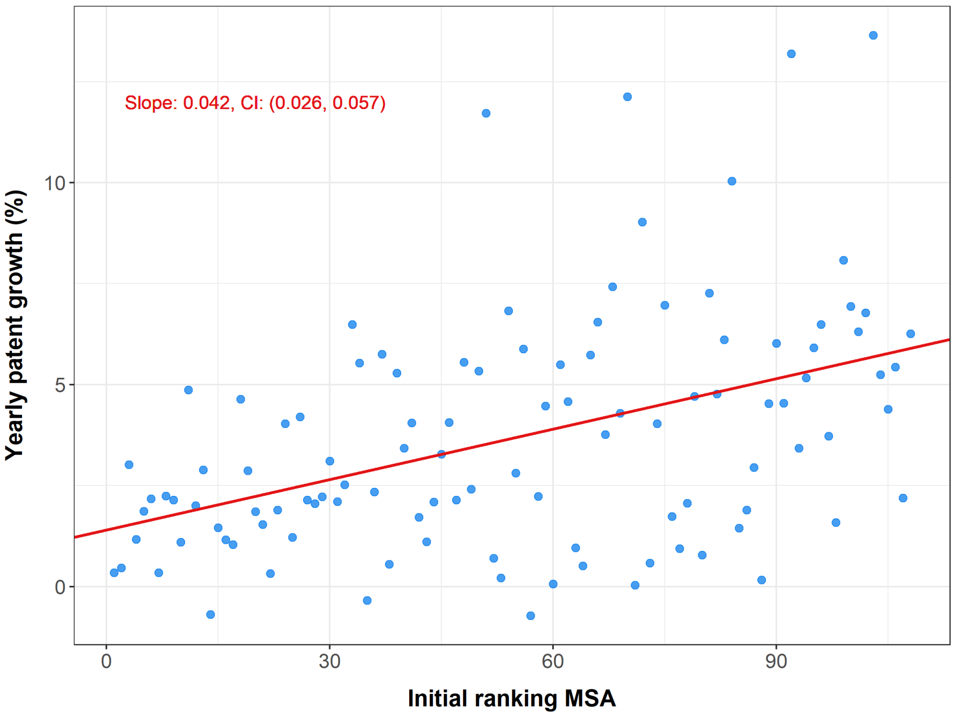



Figure 10: Patent growth rate by initial innovativeness ranking of MSA

**Fact 1.b.: The South and the West of the US had a higher patenting growth rate**

Figure 9 shows that MSAs located in the South and the West of the US had a higher patenting growth rate in 1951-1966. We classify MSAs using Census Regions of the US (Midwest, Northeast, South and West)[46] and aggregate patents within each region-technology-year. Figures 11 and 12 present averages across technologies within a region-year. Figure 11 shows that the share of patents filed by inventors located in the Midwest and the Northeast decreased from 75% in 1951 to 68% in 1966, while the share of patents filed in the South and the West increased from 25% to 32%. The opposite change in the shares implies that the South and the West had a higher growth rate of patenting relative to the Midwest and the Northeast.

Figure 12 shows that in the period 1951-1966 the South and the West increased their amount of patenting by 80%, while the Midwest and the Northeast had a 22% growth.[47]

[46] In Appendix C we present a map with the four Census Regions.
[47] Growth rates are computed by region-technology and then averaged across technologies within region.

Translated into yearly growth rates, the South and the West grew three times as fast as the Midwest and the Northeast (3.14% vs. 1.05% per year).[48]

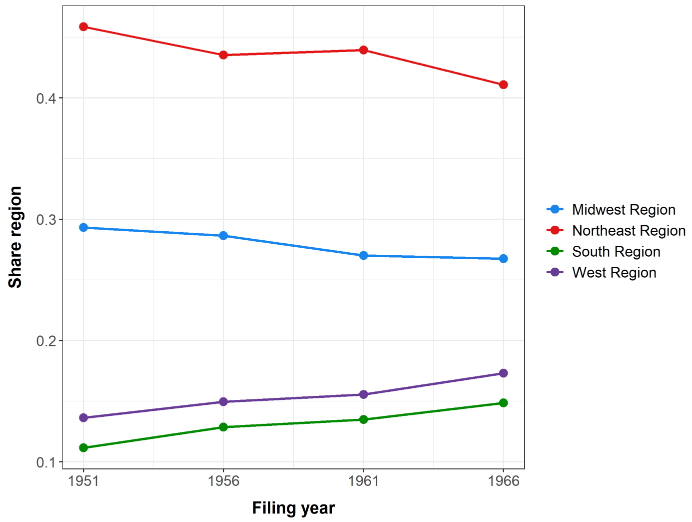



Figure 11: Share of patents by region

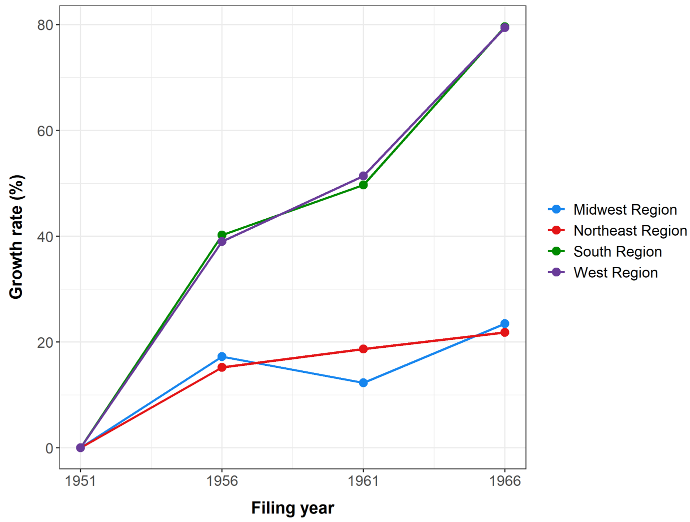



Figure 12: Patent growth by region

**Fact 2: Multi-establishment firms expanded geographically and accounted for a higher share of patents**

Using all the patents of the same owner we identify all locations in which a patent owner had inventors applying for patents. We label a patent owner a *firm* and assume that a firm has a *research establishment* in the MSAs in which it has inventors applying for patents. Combining all patents belonging to the same firm we know if a firm has research establishments in multiple MSAs, if a firm expands over time and where it locates its establishments.

In Table 1 we count the number of firms and compute their share of patents according to whether the firm had 1, 2 to 5, 6 to 10, 11 to 20, or more than 20 establishments in each respective year. As we can see, the vast majority of firms had one establishment (95.8% in 1951), while very few had 11 or more establishments (0.1% in 1951). In 1951, single-establishment firms accounted for 57% of all patents. At the same time, firms

[48] $3.14\% = 1.80^{(1/19)} \times 100$, $1.05\% = 1.22^{(1/19)} \times 100$

with 11 or more establishments (42 firms, 0.1% of all firms) accounted for 15% of all patents.

From 1951 to 1966, the amount of single establishment firms declined by 1% while the amount of firms with 11 or more establishments increased by 283%. In other words, the amount of firms with presence in 11 MSAs or more grew from 42 to 119 firms. At the same time, the share of patents accounted by firms with 11 or more establishments increased from 15% to 31%. Simultaneously, the share of patents of single-establishment firms decreased from 57% to 46%. Hence, Table 1 illustrates that both the amount of multi-establishment firms and their share of patents grew over time.[49] In Appendix B we show that multi-establishment firms increased their share of patents in all quartiles of MSAs' initial innovativeness, with a stronger increase in initially less innovative MSAs.

| | Number of firms | | | | | Share of patents | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Year \ N. estab. | 1 | 2 to 5 | 6 to 10 | 11 to 20 | +20 | 1 | 2 to 5 | 6 to 10 | 11 to 20 | +20 |
| 1951 | 41,133 | 1,684 | 75 | 34 | 8 | 0.57 | 0.19 | 0.08 | 0.07 | 0.08 |
| 1956 | 42,590 | 1,927 | 111 | 60 | 12 | 0.52 | 0.19 | 0.09 | 0.11 | 0.08 |
| 1961 | 37,366 | 2,112 | 131 | 80 | 18 | 0.48 | 0.19 | 0.09 | 0.13 | 0.12 |
| 1966 | 40,711 | 2,086 | 132 | 89 | 30 | 0.46 | 0.15 | 0.09 | 0.14 | 0.17 |

Table 1: Number of firms and share of patents by firm's geographic coverage
Geographic coverage is computed as the amount of Metropolitan Statistical Areas (MSAs) in which the firm has inventors applying for patents (*research establishments*) in a certain year. Bins of geographic coverage are 1 MSA, 2 to 5 MSAs, 6 to 10 MSAs, 11 to 20 MSAs, more than 20 MSAs. The maximum possible is 108 MSAs.

While we observe an increase in the number of multi-establishment firms, we also observe an increase in the distance between establishments of the same firm. Figure 13 shows that, for firms that have multiple establishments in the respective year, the

[49] Within each year and bin of firm size, we compute the share of patents by technology and then take the average across technologies. We have computed the across-firms Herfindahl index within technology (so it shows the level of across-firm concentration within a technology) and we do not observe a clear pattern of either concentration or deconcentration.

average distance across establishments within the firm increased over time.[50]

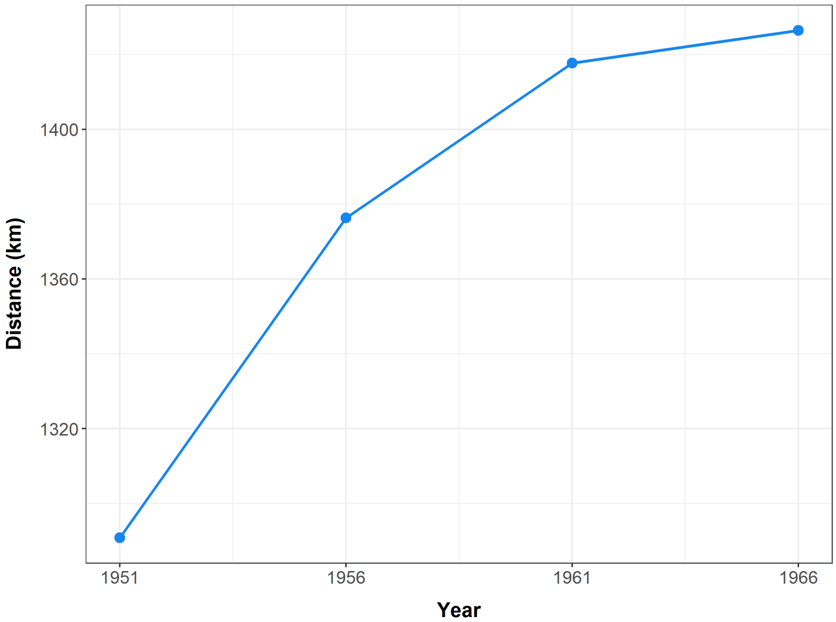



Figure 13: Average distance across establishments within the firm

#### Fact 3: Distance of citations increased

In our analysis we use citations as a proxy for knowledge diffusion. According to Jaffe et al. (1993) *"a citation of Patent X by Patent Y means that X represents a piece of previously existing knowledge upon which Y builds."* (page 580).[51] We compute the distance between the citing inventor and the cited inventor. Figure 14 shows the evolution over time of the first, second and third quartile of citation distance.[52] We observe that 25% of citations happened between inventors located less than 300km apart throughout our sample period. For the middle 50% of citations we observe that over time inventors cited other inventors located farther away. The third quartile of citation distance in-

[50]The increase in distance across establishments within firms could well be the result of firms that are growing and randomly producing new patents in different locations. However, in Section 8 we show that the process firms' geographic expansion was not random: firm's expansion was directed towards locations that got larger reductions in travel time to the firm's headquarters.

[51]Jaffe et al. (1993) discusses the reasons why to cite and why not to cite. Using a survey of inventors, Jaffe et al. (2000) find that there is communication among inventors and citations are a *"noisy signal of the presence of spillovers."*

[52]We compute distance between MSA centroids.

creased from 1,642km in 1951 to 2,284km in 1961, a 39% increase in the distance.[53] In other words, the mass of citations shifted towards longer distances.

In Figure 15 we present the share of citations by distance range between the citing and cited inventors.[54] The distance cutoffs where chosen in order to have a balanced share of citations in the initial time period, and considering the changes in travel time presented in Section 4.1. The share of citations that happen between inventors located more than 2,000km apart grew from 21.5% in 1951 to 27.9% in 1966. The 6.4 percentage points increase represents an increase of 30% of the share of citations at more than 2,000km.

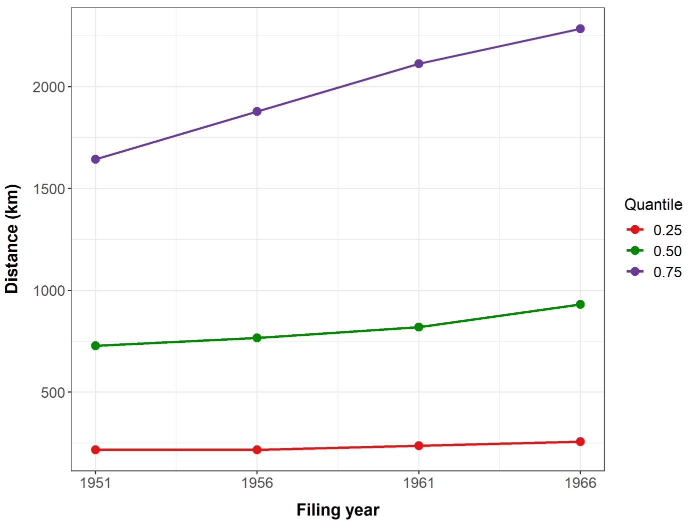



Figure 14: Quantiles of citation distance

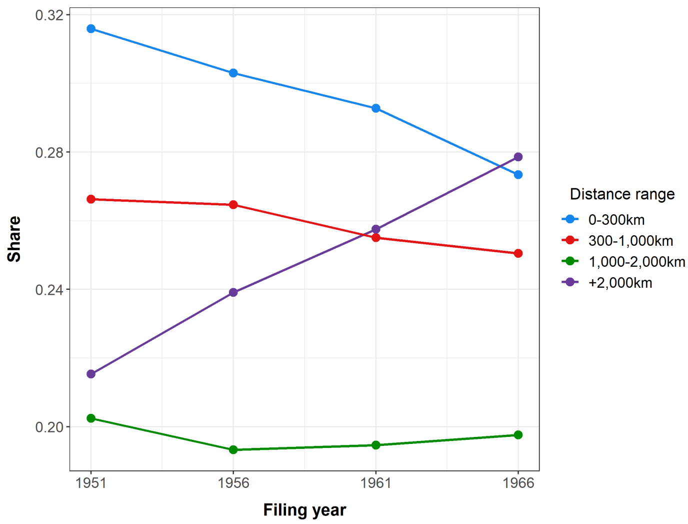



Figure 15: Share of citations by distance

[53]As a reference, the distance from New York City NY to other places is: Boston MA 300km, Chicago IL 1,140km, Dallas TX 2,200km, San Francisco CA 4,130km. The quantile 0.10 of was at 0km in every period, implying that 10% of citations took place within MSA. The quantile 0.90 was around 3,611km to 3,789km over the sample period.

[54]While Figure 14 shows how the distance of each quantile changes over time, Figure 15 shows the mass of citations (and hence the quantile to which belongs) in a certain distance cutoff. For example, in 1951 the share of citations in the 0-300km range was 31.6%, which is equal to saying that the quantile 0.316 in 1951 was 300km.

# 6. Diffusion of knowledge

In this section we show that the reduction in travel time led to an increase in knowledge diffusion, especially over long distances. In doing so we estimate the parameter $\beta$ highlighted in equation (2): the elasticity of knowledge diffusion to travel time.

To perform the analysis we merge the Air Travel and Patent datasets to obtain a final dataset that contains for each patent owner-location, the amount of patents filed in a certain 5-year period and technology class, the amount of citations to other patents with their respective owner identifier, location and technology class, and the travel time to every location. We aggregate citations to the citing-cited establishment-technology within each period. We assume that passengers take a return flight, hence we make travel times symmetric.[55]

We estimate a gravity equation which relates citations between two establishments-technologies with their pairwise travel time.[56] We estimate the following regression:

$$citations_{FiGjhkt} = \exp\left[\boldsymbol{\beta}\, log(\text{travel time}_{ijt}) + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt}\right] \times \varepsilon_{FiGjhkt} \quad (3)$$

where $citations_{FiGjhkt}$ is the amount of citations from patents filed by the establishment of firm $F$ in location $i$, technology $h$ and time period $t$, to patents filed by establishment of firm $G$ in location $j$ and technology $k$. We call $Fi$ the research establishment of firm $F$ in location $i$. travel time$_{ijt}$ is the air travel time (in minutes) between location $i$ and $j$ at time period $t$. The parameter of interest in the regression is $\boldsymbol{\beta}$, which represents the elasticity of citations to travel time.[57] If citations are affected negatively by travel time we would expect a negative value of $\beta$.

[55] $travel\ time_{ijt} = (travel\ time^{original}_{ijt} + travel\ time^{original}_{jit})/2$ where $travel\ time^{original}_{ijt}$ stands for the travel time between MSA $i$ and $j$ at time period $t$.

[56] For explanation and micro foundations of the gravity equation see Head and Mayer (2014) and references thereof.

[57] A 1 percent increase in travel time has an effect of $\beta$ percent increase (or decrease in the case of a negative $\beta$) in citations.

Given the panel structure of our data, we can include the fixed effect $FE_{FiGjhk}$ that absorbs any time invariant citation behavior within the *citing establishment-technology and cited establishment-technology*. This fixed effect flexibly controls for persistent relationships within an establishment pair that would lead to relatively more (or less) citations. That includes characteristics like physical distance, but also pre-existing commercial relationships between establishments. The fixed effects $FE_{Fiht}$ and $FE_{Gjkt}$ control for the time changing general level of citations specific to each establishment and technology. For example $FE_{Fiht}$ controls for the fact that if $Fih$ files more patents in a given period, it would mechanically make more citations to every establishment. On the other hand, $FE_{Gjkt}$ controls for $Gjk$ filing more patents or higher quality patents that would receive more citations from every establishment.[58]

The inclusion of $FE_{FiGjhk}$ implies that only variation across time within an establishment-pair is used for identification. By additionally including the fixed effect $FE_{Fiht}$, the across-time variation is compared only between citing-cited establishment-technology pairs $FiGjhk$ within a citing establishment-technology $Fih$ in period $t$. As we also include $FE_{Gjkt}$, the comparison is done while controlling for the size of the cited establishment-technology $Gjk$ in period $t$. Put differently and simplifying slightly, the identification of $\boldsymbol{\beta}$ relies on changes in citations and travel time within an establishment-pair, relative to another establishment-pair with the same citing establishment, conditional on the two cited establishments' sizes.

Following Silva and Tenreyro (2006), we estimate the gravity equation by Poisson Pseudo Maximum Likelihood (PPML).[59] This estimation methodology has two advantages over a multiplicative model that is then log-linearized to obtain a log-log specification. First, it only requires the conditional mean of the dependent variable to be correctly specified, while the OLS estimation of the log-linearized model would lead to

[58] In the International Trade literature, the parallel of the fixed effects (simplified for exposition) would be: $FE_{ij}$ country-pair fixed effect, $FE_{jt}$ origin-time fixed effect and $FE_{it}$ destination-time fixed effect.

[59] We use the package *fixest* (Bergé (2018)) in R to estimate high dimensional fixed effects generalized linear models *feglm* with Poisson link function.

biased estimates in the presence of heteroskedascity. Second, it allows to include zeros in the dependent variable, which is especially relevant when using disaggregated data. One downside of estimating PPML with the fixed effects that we include is that both coefficients and standard errors have to be corrected due to the incidental parameter problem (Weidner and Zylkin (2021)). We follow Weidner and Zylkin (2021) to use split-panel jackknife bias-correction on the coefficients and Dhaene and Jochmans (2015) to bootstrap standard errors which we also bias-correct with split-panel jackknife.[60]

Whenever $FiGjhk$ has positive citations in at least one period and missing value in another, we impute zero citations in the missing period.[61] Travel time is set to one minute whenever $i = j$.[62]

Column (1) in Table 2 presents the results of estimating equation (3). The value of the elasticity of citations to travel time is estimated to be $-0.083$, statistically significant at the 1% level. Given the average reduction in travel time of 31.4% in the full estimating sample, the elasticity implies that citations increased on average 2.6% as consequence of the reduction in travel time. If we consider the average decrease in travel time across all MSAs in the baseline travel time data, the implied increase is 2.4%.[63]

The importance of air transport relative to other means of transport potentially depends on the distance to travel. Also, we observed in section 4.1 that the improvements in air travel time depended on the distance to travel, with a difference in jet adoption

[60]Details on the bias correction and bootstrap procedures are provided in Appendix D.

[61]We do not impute zeros in $FiGjhk$ that are always zero, as those observations would be dropped due to not being able to identify $FE_{FiGjhk}$.

[62]We measure air travel time in minutes. In our sample 13% of citations happen within the same MSA. The inclusion of those citations in the estimation increases the amount of observations available to identify of $FE_{Fiht}$ and $FE_{Gjkt}$, and hence keeping them increases the amount of $FiGjhkt$ that remain in the effective sample to identify $\beta$. In order to include them we then need to impute a within-location travel time. We assume that within-location (air) travel time is not changing across time periods. Nonetheless, the identification of $\beta$ is not affected by the value chosen for the within-location (time invariant) travel time, as $\beta$ is identified by across time variation. In the appendix we show results using other values of (time invariant) within MSA travel time and the coefficients remain equal.

[63]These values come from the multiplication of the elasticity of citations to travel time 0.083 and the average decrease in travel time between 1951 and 1966: 31.4% in the full estimating sample and 28.7% in the raw data of travel time across MSAs.

| | PPML | | IV PPML | |
|---|---|---|---|---|
| Dep. variable: *citations* | *citatitions*$_{FiGjhkt}$ | | | |
| | (1) | (2) | (3) | (4) |
| log(travel time) | $-0.083^{***}$ (0.019) | | $-0.152^{***}$ (0.029) | |
| log(travel time) $\times$ 0-300km | | 0.019 (0.036) | | $-0.076$ (0.221) |
| log(travel time) $\times$ 300-1,000km | | $-0.089^{***}$ (0.023) | | $-0.134^{***}$ (0.044) |
| log(travel time) $\times$ 1,000-2,000km | | $-0.094^{***}$ (0.033) | | $-0.112^{**}$ (0.047) |
| log(travel time) $\times$ +2,000km | | $-0.169^{***}$ (0.039) | | $-0.203^{***}$ (0.043) |
| N obs. effective | 4,703,010 | 4,703,010 | 4,703,010 | 4,703,010 |
| R2 | 0.88 | 0.88 | 0.88 | 0.88 |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 2: Elasticity of citations to travel time

Column (1) shows the result of Poisson Pseudo Maximum Likelihood (PPML) estimation of $citations_{FiGjhkt} = \exp\left[\boldsymbol{\beta}\, log(\text{travel time}_{ijt}) + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt}\right] \times \varepsilon_{FiGjhkt}$, for citations of patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$, to patents filed by establishment of firm $G$ in location $j$ and technology $k$. travel time$_{ijt}$ is the travel time in minutes between location $i$ and $j$ at time period $t$, and it is set to 1 when $i = j$. When $FiGjhk$ has positive citations in at least one period and no citations in another, we attribute zero citations in the missing period. Column (2) includes the interaction of travel time$_{ijt}$ with a dummy for distance bin between the citing establishment $Fi$ and the cited establishment $Gj$. Column (3) and (4) show the result of two step instrumental variables estimation, where $log(\text{travel time}_{ijt})$ is instrumented with $log(\text{travel time}_{ijt}^{\text{fix routes}})$, the travel time that would have taken place if routes were fixed to the ones observed in 1951 and in each year routes were operated with the average airplane of the year. Bootstrap standard errors are presented in parentheses. The coefficients and standard errors in columns (1) and (2) are jackknife bias-corrected. R2 is computed as the squared correlation between observed and fitted values.

for travel distances under and over 2,000km. Taking these two characteristics into account, we estimate a variation of equation (3) in which we allow the elasticity of citations to travel time to vary by distance interval between the locations of citing and cited establishments.[64] Column (2) in Table 2 shows the result of this estimation.[65] The estimated value of the elasticity in absolute terms increases with distance, reaching $-0.169$ for distances of more than 2,000km. Between 1951 and 1966 the average change in travel time in the full estimating sample is 47.7% for a distance of more than 2,000km. The estimated elasticity implies that citations between establishments at more than 2,000km apart increased by 8.1% due to the decrease in travel time. In total citations at more than 2,000km increased by 21%, implying that the change in travel time can account for 38.2% of the observed increase. If instead we consider the 40.8% average reduction in travel time across MSAs in the raw data, the elasticity implies an increase in citations of 6.9%, accounting for 32.7% of the total citation increase.

In Appendix B we investigate different heterogeneous effects. We study how travel time affects the extensive margin of citations (whether an establishment cites another establishment or not) and the intensive margin (conditional on citing, how much it cites). We find the effect comes from both margins. We estimate an heterogeneous elasticity depending on the level of spatial concentration of the citing technology and the cited technology, we do not find a statistical difference. We also look at whether it is older patents or younger patents that get diffused, finding some slight evidence that it is technologies that take longer time to diffuse that increase more their diffusion with the reduction in travel time. We study citations to and from government patents, and self citations, on the whole we do not find a different pattern from the baseline. We also do not find a particular pattern of the elasticity depending on the citing *firm's size* as measured by the amount of patents filed in 1949-1953. Finally, we estimate the elasticity by citing and cited technology and most of the effect seems to come when the citing and cited technologies are the same.

[64] We compute distance between the geographical center of each MSA.

[65] The share of observations (citations) in each distance interval is: 0-300km 26.1% (28.5%), 300-1,000km 30.7% (28.5%), 1,000-2,000km 19.7% (23.4%), +2,000km 23.4% (19.6%).

There are two types of threats to identification in estimating equation (3): (i) the potentially targeted changes in travel time, which could be due to the opening of new routes, the allocation of jets across routes, or changes in scheduling, and (ii) time changes in other variables at the MSA-pair level which also drive the diffusion of knowledge and are correlated with the changes in travel time. In the remaining of this section we address the first type of threat by estimating the model by instrumental variables. In the following subsection we address the second type of threat by adding multiple controls. In both cases we show that results do not change.

As mentioned in Section 4.2, we may be concerned that the timing and allocation of jets to routes and that the opening/closure of routes were not random. In case there is an omitted variable that drives both the change in travel time at the MSA pair level and the change in citations across establishments within the same MSA, we would estimate biased coefficients. In order to tackle the endogeneity concern due to omitted variable we do an instrumental variables estimation using the instrument proposed in Section 4.2. To implement the instrumental variables estimation we follow a control function approach described in Wooldridge (2014). We proceed in two steps estimating the following two equations:

$$\begin{aligned} log(\text{travel time})_{FiGjhkt} \; &= \lambda_2 \, log(\text{travel time}^{\text{fix routes}}_{FiGjhkt}) \\ &\quad + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt} \; + \; u_{FiGjhkt} \end{aligned} \tag{4}$$

$$\begin{aligned} citations_{FiGjhkt} \; &= \exp\left[\beta \, log(\text{travel time}_{ijt}) \; + \; \lambda \, \hat{u}_{FiGjhkt} \right. \\ &\quad \left. + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt}\right] \; \times \; v_{FiGjhkt} \end{aligned} \tag{5}$$

In a first step we estimate equation (4) and obtain estimated residuals $\hat{u}_{FiGjhkt}$. In a second step we use the estimated residuals as a regressor in equation (5) which *controls* for the endogenous component of travel time. To perform inference we bootstrap standard errors.[66] According to Wooldridge (2014), there would be evidence of endogeneity

[66]Appendix D includes details on the bootstrap procedure.

if the parameter $\lambda$ in equation (5) is estimated to be statistically different from zero.

Columns (3) and (4) of Table 2 show the results of the instrumental variables estimation. If airlines were allocating jet airplanes to routes that would have witnessed a higher degree of exchange of knowledge even in the absence of jets, then we would expect the instrumental variables estimate to be smaller in absolute terms relative to the baseline coefficient. On the other hand, if the regulator targeted the opening of new routes between places that were in a lower trend of exchange of knowledge, we would expect the instrumented coefficient to be larger in absolute terms. Column (3) estimates the elasticity to be -0.152, bigger in absolute value compared to the non-instrumented estimate. The instrumental variables corrects for a downward bias in absolute terms, which represents evidence in favor of the regulator targeting the opening of new routes between places that had a lower degree of exchange of knowledge.[67,68]

In column (4) of Table 2 we see the coefficients of the instrumental variable estimation by distance between the citing and cited establishments. We observe the presence of a bias in the same direction as in column (3), however the magnitude of the bias is smaller except for the distance bin 0-300km, which is not precisely estimated. In particular, at more than 2,000km, the coefficient is relatively similar to the baseline estimation. In Appendix E we show the regression including coefficients on the residual *controls*. If the coefficients on controls are statistically significant, that is evidence of endogeneity. While the control is statistically significant when using only one coefficient for all distances, none of them is statistically significant when opening the coefficient

[67] The incidental parameter problem is potential present also in the instrumental variables estimation (IV PPML). However, there is currently no bias-correction procedure available for IV-PPML that we are aware off. Hence, columns (3) and (4) in Table 2 are not bias-corrected. In column (2) of Table 3 we present the PPML estimation not bias-corrected.

[68] The literature on weak instruments for non-linear instrumental variables is scarce. The rule of thumb of Staiger and Stock (1997) based on the F statistic is constructed using the bias that a *weak instrument* generates in a linear second stage (see Staiger and Stock (1997), Stock and Yogo (2005) and Sanderson and Windmeijer (2016) for testing for weak instruments in linear IV regression). For informative purposes, in the first stage of the model estimated in column (3) in Table 2 we obtain $\hat{\lambda}_2 = 0.95$ with a standard error 0.039 (clustered at the non-directional location pair level, *ij* is the same location pair as *ji*), and a within R2 of 0.38 (the share of residual variation explained by the instrument, after projecting out fixed effects).

by distance range. In other words, we do not find evidence of endogeneity at long distances, especially at +2,000km.

The instrument used in the instrumental variables estimation is constructed using the 1951 flight network. We may be concerned that the 1951 flight network is correlated with future changes of citations.[69] In order to address this concern in Appendix E we estimate equation (3) by restricting the sample to establishments in MSA-pairs that are always indirectly connected. Results go in the same direction.

## 6.1. Diffusion of knowledge: Robustness

We may be concerned that there are other variables that could drive the diffusion of knowledge and at the same time be correlated with the change in travel time. In order to bias the coefficients, such omitted variables should be time-changing at the origin-destination MSA pair and be systematically correlated with the change in MSA-pair air travel time.[70] We consider three potential variables that could bias our estimates: improvements in highways, improvements in telephone communication and changes in flight ticket prices. In Table 3 we show the results controlling for this variables separately, while in Appendix E we include them simultaneously. Estimates are robust to including these controls.

Columns (1) and (2) in Table 3 present the elasticity of citations to travel time by distance bin. In column (1) the elasticity is bias-corrected while in column (2) it is not.[71] We observe that not doing the bias correction does not qualitatively affect the results. Columns (3) to (6) include the additional controls and should be compared to column (2).

[69] We include a establishment pair fixed effect in the regressions, so a potential correlation between the 1951 flight network and the level of citations between research establishments does not affect our estimation.

[70] Variables that are not time changing or that are time changing at the MSA or establishment level do not represent a threat to identification, as they are flexibly controlled for with the fixed effects.

[71] The jackknife bias-correction due to the incidental parameter problem is computationally intensive. Due to the computational burden, we have still not bias-corrected all estimations. Columns (2) to (6) of Table 3 do not include bias-correction.

First, in 1947 the Congress published the official plan for the Interstate Highway System, a nationwide infrastructure plan to improve existing highways and build new ones (see Baum-Snow (2007), Michaels (2008), Jaworski and Kitchens (2019) and Herzog (2021)). In case the change in travel time by air is correlated with the change in travel time by highway, we would have an omitted variable bias if we include only one of them in the estimation. Taylor Jaworski and Carl Kitchens have graciously shared with us data on county-to-county highway travel time and travel costs for 1950, 1960 and 1970, which we converted to MSA-to-MSA and linearly interpolated to convert to the same years of our air travel data. Hence we have a MSA-to-MSA time-varying measure of travel time. In Appendix E we show the correlation of MSA-to-MSA change in air travel time and highway travel time.

Second, other means of communication like telephone lines may have expanded or changed their price during the period of analysis. Haines et al. (2010) contains information on the share of households within each city with telephone lines in 1960. We aggregate the variable to the MSA level. For each MSA-pair, we take the log of the mean share of households with telephone lines.[72] To include the variable as control we interact it with a time dummy to make the measure time variant. The assumption behind the interaction is that, if telephone lines expanded or changed their price over the time period, this time-change specific to each year was proportional to the 1960 log mean share of the MSA-pair.

Third, during the period of analysis ticket prices were set by the Civil Aeronautics Board, so airlines could not set prices of their own tickets. Some airlines included a sample of prices in the last page of their booklet of flight schedules, which we digitized. In Appendix E we document multiple facts about prices. The relevant fact for this section

[72]Data from the 1962 City Data Book which comes from the US Bureau of the Census. log(mean telephone share$_{ij}$ = log((telephone share$_i$+telephone share$_j$)/2). We take the log of the mean share because the share is a linear combination of origin MSA and destination MSA characteristics, hence perfectly explained by origin and destination fixed effects. Taking the log prevents this.

is that during 1962-1963 we observe a drop in prices of around 20% for routes of more than 1,000km distance. We may be concerned that the change in flow of knowledge is actually consequence of the change in prices, which happens to be correlated with the change in travel time. Given that we do not have ticket prices for each route and year, we use an estimated route price which is time varying. We obtain estimated prices by using the sample of prices that we digitized and fitting, for each year, price on a third degree polynomial of distance between origin and destination. We use log of estimated prices as control.[73]

Column (3) to (5) of Table 3 include the described controls. Assuming the covariance across coefficients is zero, none of the coefficients is statistically different from the baseline coefficients either in column (1) or (2).[74]

Fourth, we control for a time varying effect of distance on citations. We may believe that other variables may have an effect on the diffusion of knowledge, and those variables are related to the distance between the citing and cited establishments. In column (6) we include as control $log(distance)$ interacted with a time dummy. We observe that the coefficients reduce in magnitude, potentially due to the fact that the change in travel time is also correlated with distance, hence controlling for a time-varying effect of distance absorbs part of the effect. In spite of that, the coefficient for distance of more than 2,000km remains statistically significant at the 5%. This result shows that travel time and distance are not equivalent measures. Hence, it highlights the importance of the origin-destination time varying travel time data when studying the impact of face to face interactions. At the same time, this result differentiates our analysis from the one of Feyrer (2019) who uses two types of time-invariant distance (sea distance and geographical distance) interacted with time dummies.

[73]In order to perform inference we should adjust standard errors by the fact that we have a predicted regressor as control variable.

[74]To perform a test of statistical difference across coefficients of different regressions we need to estimate the covariance between them. We are currently doing a joint-bootstrap to obtain the covariance and perform the test.

| | PPML bias-corrected | PPML not bias-corrected | | | | |
|---|---|---|---|---|---|---|
| Dep. variable: *citations* | | $citations_{FiGjhkt}$ | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| log(travel time) $\times$ 0-300km | 0.019 | 0.021 | 0.023 | 0.0198 | 0.025 | 0.032 |
| | (0.036) | (0.039) | (0.039) | (0.039) | (0.038) | (0.040) |
| log(travel time) $\times$ 300-1,000km | $-0.089^{***}$ | $-0.099^{***}$ | $-0.096^{***}$ | $-0.094^{***}$ | $-0.102^{***}$ | $-0.075^{**}$ |
| | (0.023) | (0.027) | (0.028) | (0.027) | (0.027) | (0.030) |
| log(travel time) $\times$ 1,000-2,000km | $-0.094^{***}$ | $-0.093^{**}$ | $-0.089^{**}$ | $-0.071^{*}$ | $-0.104^{**}$ | $-0.040$ |
| | (0.033) | (0.042) | (0.044) | (0.042) | (0.042) | (0.052) |
| log(travel time) $\times$ +2,000km | $-0.169^{***}$ | $-0.185^{***}$ | $-0.180^{***}$ | $-0.172^{***}$ | $-0.196^{***}$ | $-0.124^{**}$ |
| | (0.039) | (0.049) | (0.050) | (0.050) | (0.049) | (0.059) |
| N obs. effective | 4,703,010 | 4,703,010 | 4,703,010 | 4,703,010 | 4,703,010 | 4,703,010 |
| R2 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 |
| Controls: | | | | | | |
| log(highway time) | - | - | Yes | - | - | - |
| log(telephone share) $\times$ time | - | - | - | Yes | - | - |
| log(price) | - | - | - | - | Yes | - |
| log(distance) $\times$ time | - | - | - | - | - | Yes |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 3: Robustness: Elasticity of citations to travel time
Part 1

Column (1) shows the result of Poisson Pseudo Maximum Likelihood (PPML) estimation of $citations_{FiGjhkt} = \exp\left[\sum_d \boldsymbol{\beta}_d \, \mathbb{1}\{distance_{ij} \in d\} \, log(\text{travel time}_{ijt}) + \sum_d \alpha_d \, \mathbb{1}\{distance_{ij} \in d\} \, \mathbb{1}\{X_{FiGjhkt}\} \, log(\text{travel time}_{ijt}) + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt}\right] \times \varepsilon_{FiGjhkt}$, for citations of patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$, to patents filed by establishment of firm $G$ in location $j$ and technology $k$. travel time$_{ijt}$ is the travel time in minutes between location $i$ and $j$ at time period $t$, and it is set to 1 when $i = j$. $d$ are distance intervals: $[0 - 300km]$, $(300km - 1000km]$, $(1000km - 2000km]$, $(2000km - max]$. Column (1) presents jackknife bias-corrected coefficients and bias-corrected bootstrap standard errors. Column (2) repeats column (1) without bias-correction. Relative to (2), columns (3) through (6) contain additional controls. Column (3) controls for log highway time between $i$ and $j$ at period $t$. Column (4) controls for the log of the mean share of households with telephone line in 1960 in $ij$ pair interacted with a time dummy. Column (5) controls for log flight ticket price between $i$ and $j$ at period $t$. Column (6) controls for log distance $ij$ interacted with a time dummy. When $FiGjhk$ has positive citations in at least one period and no citations in another, we attribute zero citations in the missing period. Columns (2) through (6) present standard errors clustered at the non-directional location in parentheses ($ij$ is the same non-directional location pair as $ji$). R2 is computed as the squared correlation between observed and fitted values.

Finally, as we will see in section 8.2, the entry and exit of research establishments was not uniform across locations during the sample period. We may then be concerned that the change in diffusion of knowledge is only consequence of the change in the geographic location of innovation. In Appendix E we re-estimate equation (3) with different samples: first, using only citing establishments that were present in 1949-1953, and second using only citing and cited establishments that were present in 1949-1953. We find the coefficient at more than 2,000km remains comparable to the one in the baseline regression, statistically significant at the 1%.

# 7. Creation of knowledge

In this section we show that the reduction in travel time to innovative locations led to an increase in knowledge creation. We show that the effect on knowledge creation was stronger in initially less innovative locations, leading to convergence across locations in terms of innovation. Additionally, the reduction in travel time contributed to a change in the geographic distribution of knowledge creation, increasing the relative importance of locations in the South and the West of the United States.

We construct a measure of *Knowledge Access* by adapting equation (2) to an empirical set up with multiple technology categories and time periods. The measure of *Knowledge Access* ($KA_{iht}$) shows how *easy* it is in time period $t$ for research establishments in location $i$ and technology $h$ to access knowledge created in other locations. We compute *Knowledge Access* as follows:

$$KA_{iht} = \sum_{k} \omega_{hk} \sum_{j, j \neq i} \text{Patent stock}_{jk,t=1953} \times \text{travel time}_{ijt}^{\beta} \tag{6}$$

where, from right to left, travel $\text{time}_{ijt}^{\beta}$ is the travel time between locations $i$ and $j$ at time period $t$, to the power of the elasticity of diffusion of knowledge to travel time. Patent $\text{stock}_{jk,t=1953}$ is the discounted sum of patents produced in location $j$ and technology $k$ between 1941 and 1953.[75] $\omega_{hk}$ is the share of citations of technology $h$ that go to technology $k$ at the aggregate level in 1949-1953, similar to an input-output weight.[76] Then, $KA_{iht}$ is a weighted sum of the patent stock in each other location and technology, where the weights are how easy it is to access that patent stock (travel time$^{\beta}_{ijt}$) multiplied by how relevant that knowledge is ($\omega_{hk}$).

In order to reduce concerns of potential endogeneity of accessing knowledge and creating knowledge, we exclude the patent stock in the location itself from the sum (we only use $j \neq i$).[77]

The measure of *Knowledge Access* contains across-time variation within a location-technology $ih$, and cross-sectional variation across technologies $h$ within a location $i$. The across-time variation is only due to the change in travel time between locations, every other component of the measure is fixed to its 1949-1953 level. The cross-sectional variation comes from a distribution of Patent stock$_{jk,t=1953}$ within $k$ that is not uniform across $j$, and from the input-output weights $\omega_{hk}$. The joint across-time and cross-sectional variation means that if travel time for $ij$ reduces, there will be a differential change in *Knowledge Access* across $h$ within $i$ which depends on the initial patent stock and input-output weights.

The degree with which changes in travel time are reflected in access to knowledge

[75] Patent stock$_{jk,t=1953} = \sum_{y \in [1941,1953]}$ Patents$_{jky}$ $\times$ $(1 - \text{depreciation rate})^{1953-y}$ . We use a depreciation rate of 5%, which is in the range of average depreciation rates of R&D found by De Rassenfosse and Jaffe (2017). We decided to fix the patent stock and not to allow it to change over time, as changes in travel time will potentially lead to changes in patent stock creating a dynamic reinforcing effect between knowledge access and new knowledge. In this sense, we abstract from *dynamic* externalities that could be at play.

[76] $\omega_{hk} = citations_{hk,t=[1949,1953]} / citations_{h,t=[1949,1953]}$ is included to weight each *source* technology category $k$ by how important it is for the *destination* technology category $h$.

[77] The theory makes no distinction on whether the knowledge stock is in $i$ or $j$, so in principle we would like to include the patent stock of $i$ in the knowledge access of $i$. However, this could lead to econometric problems. First, we do not have exogenous variation of travel time within $i$. Second, if knowledge creation in $i$ is a persistent process, by including the patent stock of $i$ we would introduce a mechanical relationship between knowledge access and knowledge creation. Hence, our baseline measure of knowledge access of $i$ does not consider the patent stock of $i$. In Appendix E we show that the inclusion of $i$'s patent stock does not affect the results.

depend on how *important* travel time is to get knowledge to diffuse, which is the elasticity of knowledge diffusion to travel time that we estimated in Section 6. As the baseline we use $\beta = 0.185$, which is the elasticity of citations to travel time at more than 2,000 km not bias corrected. In robustness we use distance-specific $\beta$ and in Appendix E we do sensitivity analysis of the results to changing the value of $\beta$.

The measure of *Knowledge Access* allows us to translate changes in travel time between pairs of MSAs into a single location-technology specific characteristic, and to represent it on the same scale as patent growth in Figure 9. Figure 16 depicts the time change in log *Knowledge Access* from 1951 to 1966, averaged across technologies within each MSA. Darker colors represent higher growth in *Knowledge Access*. As with patent growth, we observe that MSAs that had the strongest growth are generally located in the South and the West of the United States, far from the knowledge centers of New York and Chicago. The reduction in travel time was larger between locations far apart, implying that locations which happened to be far from knowledge centers increased relatively more their *Knowledge Access*.

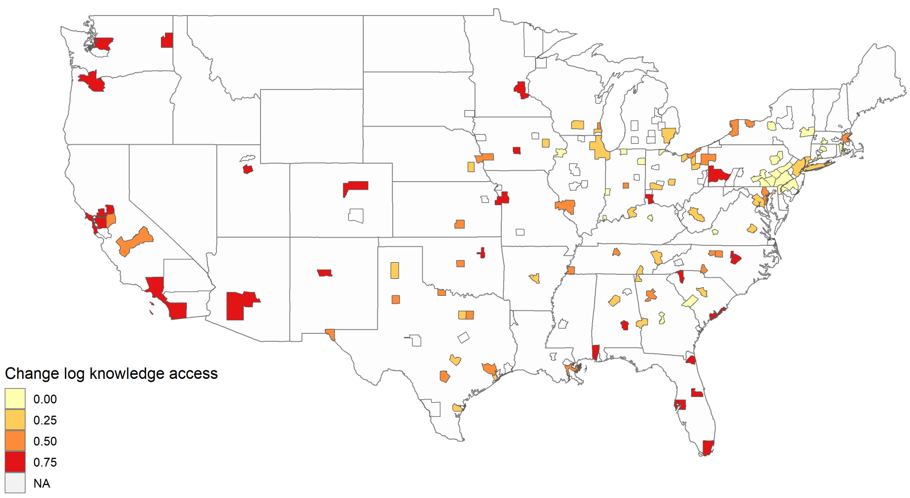



Figure 16: Change in log Knowledge Access 1951 - 1966

With the measure of *Knowledge Access* we then adapt equation (1) to estimate:

$$\text{Patents}_{Fiht} \;=\; \exp\left[\,\boldsymbol{\rho}\, log(KA_{iht}) + FE_{Fih} + FE_{it} + FE_{ht}\right]\,\times\,\xi_{Fiht} \tag{7}$$

where $\text{Patents}_{Fiht}$ are patents applied by establishment of firm $F$ in location $i$ and technology $h$ at time period $t$. The measure of knowledge access $KA_{iht}$ is at the $iht$ location-technology-time level, meaning that all establishments within an $iht$ share the same level of knowledge access. The parameter of interest $\rho$ is the elasticity of (the creation of new) patents to knowledge access. In the presence of knowledge spillovers as suggested in section 2, we would expect $\rho$ to be positive and statistically significant.

The fixed effect $FE_{Fih}$ absorbs time invariant characteristics at the firm-location-technology level, as for example the productivity of the establishment-technology. This fixed effect is more fine grained than just a location-technology, which would absorb the comparative advantage of a location in a certain technology. The fixed effect $FE_{it}$ absorbs characteristics that are time variant at the location level. For example, changes in R&D subsidies that are location specific and common across all technologies would be absorbed by this fixed effect. Also, better flight connectivity could spur economic activity as shown in Campante and Yanagizawa-Drott (2017), leading to an increase in patenting activity in the location. If that increase is general across technologies within the location, then $FE_{it}$ would absorb it. Finally, the fixed effect $FE_{ht}$ absorbs characteristics that are time variant at the technology level. If technologies had different time-trends at the national level, then the fixed effect would control for these trends in a flexible way.

The inclusion of $FE_{Fih}$ implies that only across-time variation within an establishment-technology is used to identify $\rho$. The inclusion of $FE_{it}$ implies that only variation across-technologies within a location-time is exploited, so across-time variation is compared across establishments within a location, and not across locations. The inclusion of $FE_{ht}$ implies that the identifying across-time variation is conditional on aggregate trends of the technology. In short, identification of $\rho$ relies on across-time changes in the amount of patents and knowledge access of an establishment, relative to other establishments in the same location, conditional on aggregate technological trends.

| | PPML | PPML q innovation | IV PPML | IV PPML q innovation |
|---|---|---|---|---|
| Dependent Variable: *Patents* | | $Patents_{Fiht}$ | | |
| | (1) | (2) | (3) | (4) |
| log(knowledge access) | 10.14*** (3.66) | 9.36** (3.69) | 11.24* (6.35) | 10.26 (6.38) |
| log(knowledge access) × 3rd quartile | | 2.05*** (0.58) | | 2.32*** (0.66) |
| log(knowledge access) × 2nd quartile | | 3.80*** (0.90) | | 4.21*** (0.84) |
| log(knowledge access) × 1st quartile | | 5.00*** (1.30) | | 5.77*** (1.11) |
| R2 | 0.85 | 0.85 | 0.85 | 0.85 |
| N obs. effective | 991,480 | 991,480 | 991,480 | 991,480 |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 4: Effect of knowledge access on patents, by MSA innovativeness quartile

Column (1) shows the result of Poisson Pseudo Maximum Likelihood (PPML) estimation of $\text{Patents}_{Fiht} = \exp\left[\boldsymbol{\rho}\, log(KA_{iht}) + FE_{Fih} + FE_{it} + FE_{ht}\right] \times \xi_{Fiht}$, for patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$. $KA_{iht}$ is knowledge access of establishments in location $i$ technology $h$ and time period $t$. Column (2) opens the coefficient $\rho$ by the quartile of innovativeness of location $i$ within technology $h$, computed using patents filed in 1949-1953. Higher quartile indicates higher initial level of innovativeness. The fourth quartile is used as reference category. Column (3) and (4) show the result of two step instrumental variables estimation, where $KA_{iht}$ is instrumented with $\widetilde{KA}_{iht}$, knowledge access computed using the counterfactual travel time that would have taken place if routes were fixed to the ones in 1951 and each year routes were operated at the average aggregate flying speed of the year. Standard errors are presented in parentheses. Column (1) and (2) present clustered at the location-technology $ih$. Column (3) and (4) present bootstrap standard errors. R2 is computed as the squared correlation between observed and fitted values.

Column (1) in Table 4 shows the result of estimating equation (7). The elasticity of patents to knowledge access is estimated to be 10.14, significant at the one percent level. The average change in knowledge access at the location-technology level[78] is 9%, implying that on average the change in travel time predicts a 3.5% average yearly

[78]Due to entry, we cannot compute the growth rate at the establishment-technology level for 70% of establishment-technology, given that they had 0 patents in the initial time period. In the case of location-technology, 5% did not have patents in the initial period. We the prefer to interpret coefficients using location-technology growth rates, which we compute using the remaining 95% of location-technologies that had positive patents in the initial time period.

growth rate of patents.[79] The observed average yearly growth rate of new patents at the location-technology is 4.4%.[80] Comparing the predicted and observed growth rates, the improvement in air travel time has the power to account for 79.5% of the observed average yearly patent growth rate.[81]

We aggregate predicted changes in patent growth at the Census Region level. The change in travel time predicts a yearly growth rate 0.74 percentage points higher in the South and the West relative to the Midwest and Northeast. In the data we observe 2.1 percentage points difference in the growth rate, implying that the change in travel time can account for 35% of the observed differential growth rate.[82]

Section 5.1 showed that in the data, initially less innovative MSAs had a larger growth rate of patenting. In column (2) in Table 4 we investigate if the increase in knowledge access had an heterogeneous effect on the amount of new patents created depending on the initial innovativeness of the location $i$ in technology $h$. We compute the quartile of innovativeness of location $i$ in technology $h$ in the time period 1949-1953 and interact it with $\log(KA_{iht})$.[83] We use as reference category the highest quartile of initial innovativeness, hence the coefficient on $\log(KA_{iht})$ without interaction is the elasticity for the highest quartile. Coefficients on other quartiles should be interpreted relative to the highest quartile.

[79] The elasticity of 10.14 predicts an increase of 91.3% over the time period of 19 years ($10.14 \times 0.09 = 0.913$), which translates into a 3.5% average yearly growth rate ($(1+0.913)^{1/19}-1 \approx 0.035$).

[80] From the first time period (1949-1953) to the last time period (1964-1968) we observe an average growth rate of new patents of 127%. We obtain $0.044 \approx ((1+1.27)^{1/19} - 1$

[81] $79.5 = 3.5/4.4 \times 100$

[82] Using the coefficient of column (1) in Table 4, we compute the MSA-technology predicted level of patents for 1966 and aggregate it at the Census region - technology level. Then, we compute yearly growth rates within each region-technology and take averages across technologies. Next, we take the average between S and W, and MW and NE, and finally compute the differential predicted growth. If we use the quartile-specific coefficients of column (2) in Table 4 we obtain a predicted differential growth rate of 0.86 percentage points, which implies that the change in travel time can account for 41% of the observed differential growth rate.

[83] We use the quartiles of innovativeness defined in section 5.1, computed using the amount of patents of location $i$ in technology $h$ filed in the time period 1949-1953. Each location $i$ has (potentially) a different value quartile in each technology $h$. The 1st quartile refers to the 25% initially least innovative MSAs in technology $h$.

We find that the coefficients on lower quartiles of initial innovativeness are positive and statistically different from the coefficient in the highest quartile. Thus, knowledge access had a greater effect on patenting for establishments that were located in initially less innovative locations.[84] Given the difference in the coefficients, the increase in knowledge access predicts an average yearly growth of new patents of 4.5% for the initially lowest quartile of innovativeness, while it predicts 3.4% for the highest quartile.[85] The change in knowledge access predicts differential growth rate of 1.1 percentage points. In the data we observe that the average yearly growth rate of patents in the lowest quartile is 5.3 percentage points higher than in the highest quartile. Comparing the predicted and observed differential growth rates, the improvement in knowledge access as consequence of the reduction in travel time explains 21% of the difference in growth rates of new patents between locations in the lowest and highest quartile of innovativeness.[86]

As in Section 6, we may be concerned that decisions of the regulator or airlines which affect travel time are endogenous to the diffusion of knowledge and consequently to knowledge access. Therefore, we construct an instrument for knowledge access in which instead of using observed travel time, we use the fictitious travel time presented in section 4.2 in which routes are fixed to the ones in 1951 and each route is operated with the average airplane of the year:

$$\widetilde{KA}_{iht} = \sum_{k} \omega_{hk} \sum_{j, j \neq i} \text{Patent stock}_{jk,t=1953} \times \left(\text{travel time}_{ijt}^{\text{fix routes}}\right)^{\beta} \tag{8}$$

We then implement the instrumental variables estimation by control function as in

[84]A given percentage change in knowledge access led to a stronger increase in patenting in initially less innovative locations.

[85]The change in knowledge access for the lowest quartile is on average 9.1%, which multiplied by the coefficient 14.36 (obtained by doing 9.36+5.00=14.36) gives a predicted growth of 131% over 19 years. Translated into average yearly growth it is 4.5% = $[(1+1.31)^{(1/19)} - 1] \times 100$. For the highest quartile, knowledge access changed on average 9.5%, which multiplied by the coefficient 9.36 predicts 89% growth rate, which is 3.4% yearly growth rate.

[86]$21\% \approx 1.2/5.1 \times 100$

Section 6. The results are presented in columns (3) and (4) in Table 4. The coefficients do not show an important change and the convergence prediction obtained using non-instrumented PPML remains valid.[87,88]

Figure 17 shows in the left panel the patent growth observed in the data (it replicates Figure 9), while in the right panel it is the predicted patent growth. We compute the prediction using the observed change in travel time and quartile specific elasticities of column (2) in Table 4. Similarly to what is observed in the data, the change in travel time predicts a larger patenting growth rate in the South and the West. At the same time, the change in travel time predicts smaller growth rates in New York, Chicago and their surroundings.

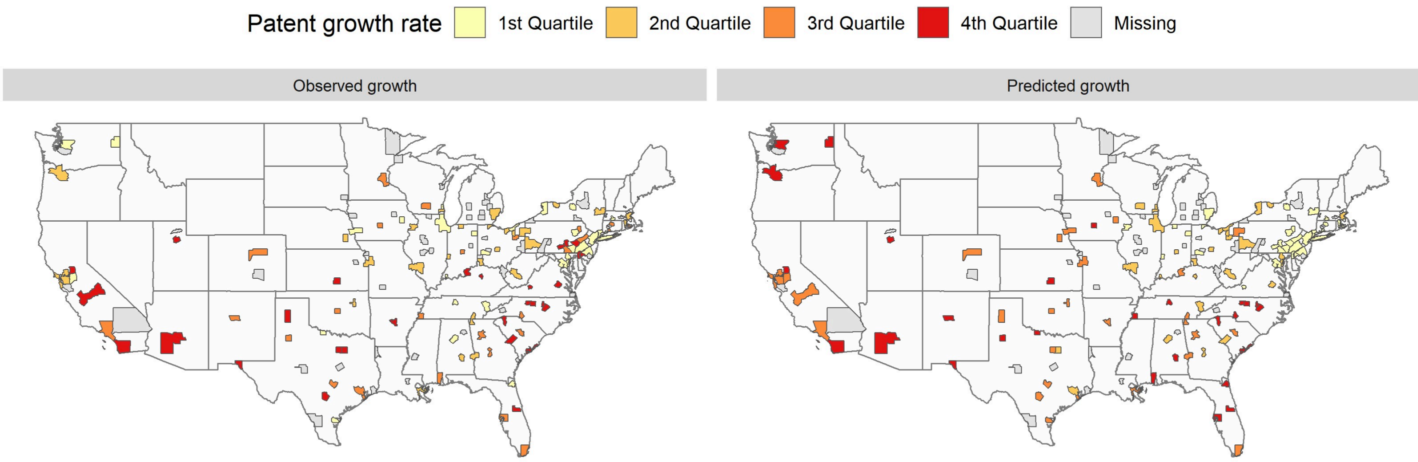



Figure 17: Observed vs. predicted patent growth 1951 - 1966

The result in column (2) implies that a given change in *Knowledge Access* had a stronger effect on patenting growth in less innovative locations. In other words, knowledge spillovers as an externality had a more predominant role in the production of

[87]The first stage of the model estimated in column (3) of Table 4 gives a $\hat{\lambda}_2 = 1.01$ with standard error 0.03 (clustered at the location-technology level *ih*), and a within R2 of 0.53.

[88]Using IV estimates, the predicted yearly patent growth rate in the lowest quartile is 4.9% while it is 3.7% in the highest quartile. The predicted differential growth rate is then 1.2 percentage points, meaning that the change in knowledge access can explain $(1.2/5.3) \times 100 \approx 23\%$ of the observed differential growth rate.

knowledge in locations that initially produced relatively fewer patents. Theoretically, this result implies that the parameter $\rho$ in equation (1) varies depending on the level of previous production of knowledge of location $i$. Empirically the implication is that a given increase in knowledge spillovers leads to innovation convergence across locations. As seen in section 5.1, during 1949-1968 we observe innovation-convergence across locations and that is exactly what the estimated coefficients predict following a reduction in travel time.

In order to understand the convergence result and compare it with other findings in the literature it is important to remember that commercial airplanes during 1950s and 1960s were a means of transportation mainly for people. On the other hand, other transportation improvements as those in water transport, railroads or highways also contain another ingredient: they were used to carry goods. Hence, other means of transportation have a simultaneous impact on face to face interactions and trade. Pascali (2017) finds that the introduction of the steam engine vessels in the second half of the 19th century had an impact on international trade that led to economic divergence between countries. Faber (2014) finds that the expansion of the highway system in China led to a reduction of GDP growth in peripheral counties, with evidence suggesting a trade channel due to reduction in trade costs. In our setup, the introduction of jet airplanes represented a big shock to the mobility of people while not affecting significantly the transport of merchandise. Therefore, studying the introduction of jet airplanes allows us to focus on improved face to face interactions, while the trade channel would be a second order effect.

## 7.1. Creation of knowledge: Robustness

In this section we show that the effect of *Knowledge Access* on the creation of new patents and the convergence effect remains after including different controls. Table 5 shows the results.

Jaworski and Kitchens (2019) show that improvements in the Interstate Highway System led to local increases in income through an increased market access. In our set up, if the effect of market access affects innovation in the same way across technologies, then it would be absorbed by the MSA-time fixed effect $FE_{it}$ in equation (7). However, if the effect of market access on innovation varies across technologies, then it would be a confounder. To control for this potential confounder, we compute market access by highway and interact it with a technology dummy. We compute market access as:

$$\text{Market Access}_{it} = \sum_j \text{Population}_{j,t=1950} \times \tau_{ijt}^{\theta} \tag{9}$$

where $\text{Population}_{j,t=1950}$ is population in MSA $j$ in 1950, $\tau_{ijt}$ are the shipping costs provided in the data of Taylor Jaworski and Carl Kitchens computed using each year's highway driving distance, highway travel time, petrol cost and truck driver's wage. $\theta$ is the elasticity of trade to trade costs which we set to -8.28, the preferred value of Eaton and Kortum (2002) and in the range of many other estimates in the literature (Head and Mayer (2014), Caliendo and Parro (2015), Donaldson and Hornbeck (2016)). Columns (3) and (4) of Table 5 show the results, we do not observe an important difference with the baseline estimates.

Campante and Yanagizawa-Drott (2017) shows that better connectivity by airplane leads to an increase in economic activity as measured by satellite-measured night light. Söderlund (2020) shows that an increase in business travel in the late 1980s and early 1990s led to an increase in trade between countries. In a similar way to market access, we could think that better connectivity by airplane could have led to an increase in market access due to a reduction in information frictions, with goods being shipped by land. Similarly to highway market access, if the effect of market access by airplane is common to all technology categories the effect would be absorbed by the MSA-time fixed effect $FE_{it}$. In order to account for a technology-specific effect, we construct a measure of airplane market access and interact it with a technology dummy. The measure of airplane market access is similar to equation (9) where $\tau$ is the travel time by airplane and $\theta$ is set to -1,22, the elasticity of trade to travel time from Söderlund (2020). The results are shown in columns (5) and (6) of Table 5. While the coefficients in all quartiles are reduced, the estimated value of $\rho$ is positive and significant and the result on convergence remains.

Potential contemporaneous improvements in other means of communication, like telephones, could have spurred the creation of new patents. In columns (7) and (8) we include the log of the MSA's share of households with telephones in 1960 and double-interact it with a technology dummy and a time dummy. The results remain invariant with respect to the baseline.

Another potential explanation for the increase of patenting could be that better connectivity decreased technology-specific financial frictions. The potential reduction in financial frictions, rather than a confounder, would be a mechanism through which airplanes increased innovation. However, according to Jayaratne and Strahan (1996) during 1950s and 1960s interstate lending and bank branching were limited. Prior to the 1970s, banks and holdings were restricted in their geographic expansion within and across state borders. Additionally, the Douglas Amendment to the Bank Holding Company Act prevented holding companies from acquiring banks in other states. Therefore, it is unlikely that interstate bank financing would be a driving force. Nonetheless, if other sector-specific modes of financing like venture capital were active, it could be confounding the results. In Appendix E we construct multiple measures of access to capital by using market capitalization of patenting firms listed in the stock market. The results present suggestive evidence that access to capital is not driving the results.

Finally, in Appendix E we include additional robustness checks. We compute different versions of *Knowledge Access*: we use distance-specific $\beta$ from section 6, we consider the patent stock only of locations $j$ far from $i$, we do sensitivity analysis using different values of $\beta$. Also, we re estimate the effects by quartile of initial innovativeness using patents per capita. Last, we re-do the baseline regression using OLS estimation. Re-

| | PPML | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Dependent Variable: *Patents* | $Patents_{Fiht}$ | | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| log(knowledge access) | 10.14*** (3.66) | 9.36** (3.69) | 9.28** (3.68) | 8.23** (3.69) | 6.22* (3.58) | 5.84 (3.60) | 10.34*** (3.44) | 9.25*** (3.43) |
| log(knowledge access) × 3rd quartile | | 2.05*** (0.58) | | 2.16*** (0.57) | | 2.06*** (0.59) | | 2.23*** (0.57) |
| log(knowledge access) × 2nd quartile | | 3.80*** (0.90) | | 3.89*** (0.89) | | 3.75*** (0.88) | | 3.93*** (0.91) |
| log(knowledge access) × 1st quartile | | 5.00*** (1.30) | | 5.13*** (1.30) | | 5.08*** (1.29) | | 5.18*** (1.32) |
| N obs. effective | 991,480 | 991,480 | 991,480 | 991,480 | 991,480 | 991,480 | 991,480 | 991,480 |
| R2 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 |
| Controls: | | | | | | | | |
| log(Highway market access) × technology | - | - | Yes | Yes | - | - | - | - |
| log(Airplane market access) × technology | - | - | - | - | Yes | Yes | - | - |
| log(Telephone share) × technology × time | - | - | - | - | - | - | Yes | Yes |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 5: Elasticity of new patents to knowledge access, by MSA innovativeness quartile

Column (1) shows the result of Poisson Pseudo Maximum Likelihood (PPML) estimation of $\text{Patents}_{Fiht} = \exp\left[\boldsymbol{\rho}\, log(KA_{iht}) + FE_{Fih} + FE_{it} + FE_{ht}\right] \times \xi_{Fiht}$, for patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$. $KA_{iht}$ is knowledge access of establishments in location $i$ technology $h$ and time period $t$. Column (2) opens the coefficient $\rho$ by the quartile of innovativeness of location $i$ within technology $h$, computed using patents in 1949-1953. Higher quartile indicates higher initial level of innovativeness. The fourth quartile is used as reference category. Relative to columns (1) and (2), columns (3) and (4) control for technology specific effect of log(highway market access), columns (5) and (6) control for technology specific effect of log(airplane market access), columns (7) and (8) control for technology and time specific effect of log(telephone share). Standard errors clustered at the location-technology *ih* are presented in parentheses. R2 is computed as the squared correlation between observed and fitted values.

sults go in the same direction: an increase in knowledge access leads to an increase in patenting and the effect is stronger in initially less innovative locations.

# 8. Firms' geographic expansion

In section 5.1 we showed that there was innovation-convergence across regions and this happened simultaneously with an increase in the amount of multi-establishment firms. In section 7 we showed that the reduction in travel time predicts innovation-convergence across locations. In this section we uncover one of the mechanisms that led to innovation-convergence: the geographic expansion of multi-establishment firms. We proceed in two steps. First, we show that the increase in patenting is driven by two types of entry: entry of establishments of new firms, and entry of establishments of pre-existing firms. The second type of entry is due to the geographic expansion of firms. Second, we show that the decrease in travel time led firms to expand geographically and this expansion was stronger towards initially less innovative locations.

## 8.1. Entry of new establishments

We use all patents of the same firm to identify all locations in which the firm had research establishments in each time period.[89] Using patents applied during the first time period (1949-1953), we classify all the research establishments that applied for patents in every subsequent period. We classify research establishments into three mutually exclusive categories: the establishment (and hence the firm) applied for patents in 1949-1953 (*existing firm and est*), the establishment did not apply for patents but the firm had establishments in other locations applying for patents in 1949-1953 (*existing firm new est*), neither the establishment nor the firm applied for patents in 1949-1953 (*new firm new est*).[90] The dummies *new firm new est* and *existing firm new est* capture two

[89] All our *firm* and *research establishment* information comes from the patent data. Hence, we only observe an establishment in a certain time period if it applies for patents in that time period.

[90] We define if an establishment exists or not if it applied for patents in any technology $h$ in 1949-1953. We define the establishment at the $Fi$ level (as opposed to $Fih$) as our object of interest a firm-location.

types of entry margin. *new firm new est* captures a new establishment of a new firm, while *existing firm new est* captures entry due to the geographic expansion of firms. The dummy *existing firm and est* captures jointly an intensive and exit margin.

We estimate a variation of equation (7) that includes interactions with dummies which indicate the status of the establishment in 1949-1953:

$$\text{Patents}_{Fiht} \;=\; \exp\left[\sum_{e} \boldsymbol{\rho_e}\, log(KA_{iht}) \;\times\; \mathbb{1}\{Fi \in e\} + FE_{Fih} + FE_{it} + FE_{ht}\right] \times\; \nu_{Fiht} \quad (10)$$

where $\text{Patents}_{Fiht}$ are patents applied by establishment of firm $F$ in location $i$ and technology $h$ at time period $t$. $KA_{iht}$ is the knowledge access at the location-technology-time level. $\mathbb{1}\{Fi \in e\}$ is an indicator variable that takes value 1 of $Fi$ is of the type $e = \{$*new firm new est*, *existing firm new est*, *existing firm and est*$\}$. The results are displayed in column (2) of Table 6. The results show that the effect of innovation access on the increase of patenting happened through the two entry margins: entry of new establishments of new firms and entry of new establishments of firms that previously existed in other locations.

In Table 7 we open up the effect by including a double interaction of $Fi$ establishment type and location-technology $ih$ quartile of initial innovativeness. We use the highest quartile as the reference category. The two margins of entry are active in all quartiles of initial innovativeness, with a stronger effect in lower quartiles. In the case of the entry of establishments that belong to firms that already existed in other locations, the pattern is more prominent. The intensive and exit margin does not appear active in any quartile of innovativeness except for the last one. The combined effect of entry and intensive/exit suggests that, in locations in the lowest quartile of initial innovativeness, the churn rate of patenting firms is increased as consequence of the increase in knowledge access.

An interesting avenue of research is to study within-establishment changes in the technological composition of patenting.

| Dependent Variable: *Patents* | $Patents_{Fiht}$ | |
|---|---|---|
| | (1) | (2) |
| log(knowledge access) | $10.14^{***}$ | |
| | (3.66) | |
| log(knowledge access) × new firm new est | | $23.71^{***}$ |
| | | (4.46) |
| log(knowledge access) × existing firm new est | | $23.79^{***}$ |
| | | (4.47) |
| log(knowledge access) × existing firm and est | | $-0.28$ |
| | | (4.70) |
| R2 | 0.85 | 0.81 |
| N obs. effective | 991,480 | 991,480 |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 6: Patents and knowledge access: Entry, exit and continuing firms

Column (1) shows the result of Poisson Pseudo Maximum Likelihood (PPML) estimation of $\text{Patents}_{Fiht} = \exp\left[\boldsymbol{\rho}\, log(KA_{iht}) + FE_{Fih} + FE_{it} + FE_{ht}\right] \times \xi_{Fiht}$, for patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$. $KA_{iht}$ is knowledge access of establishments in location $i$ technology $h$ and time period $t$. Column (2) adds an interaction of $log(KA_{iht})$ with $e$ the type of establishment $Fi$ in a classification on whether the establishment and/or the firm existed in 1949-1953. Standard errors clustered at the location-technology $ih$ are presented in parentheses. R2 is computed as the squared correlation between observed and fitted values.

The results of Table 6 and Table 7 indicate that one part of the increase in patenting is consequence of multi-establishment firms that expand across locations, and more so in initially less innovative locations. Hence, multi-establishment firms contributed to innovation-convergence across locations by expanding geographically.

| Quartile innovativeness \ Establishment type | New firm & New est | Existing firm & New est | Existing firm & Existing est |
|---|---|---|---|
| log(knowledge access) | 22.84*** (4.40) | 22.00*** (4.41) | −0.36 (4.67) |
| log(knowledge access) × 3rd quartile | 3.40*** (1.14) | 6.35*** (1.44) | −1.33 (1.19) |
| log(knowledge access) × 2nd quartile | 5.95*** (1.48) | 6.74*** (1.67) | −2.20 (2.33) |
| log(knowledge access) × 1st quartile | 4.88** (1.97) | 10.98*** (2.15) | −15.62*** (3.25) |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 7: Patents and knowledge access: entry, exit and continuing firms
The table shows the results of one Poisson Pseudo Maximum Likelihood (PPML) estimation of $\text{Patents}_{Fiht} = \exp\left[\sum_e \boldsymbol{\rho_e}\, log(KA_{iht}) \times \mathbb{1}\{Fi \in e\} + \sum_{q\in\{1,2,3\},e} \boldsymbol{\rho_{q,e}}\, log(KA_{iht}) \times \mathbb{1}\{ih \in q\} \times \mathbb{1}\{Fi \in e\} + FE_{Fih} + FE_{it} + FE_{ht}\right] \times \nu_{Fiht}$, for patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$. $KA_{ikt}$ is knowledge access of establishments in location $i$ technology $h$ and time period $t$. $q$ is the quartile of initial innovativeness of location $i$ within technology $h$, computed using patents filed in 1949-1953. Higher quartile indicates higher initial level of innovativeness. The fourth quartile is used as reference category. $e$ is the type of establishment $Fi$ in a classification on whether the establishment and/or the firm existed in 1949-1953. Standard errors clustered at the location-technology $ih$ are presented in parentheses. R2 is computed as the squared correlation between observed and fitted values. All columns and rows belong to the same regression. The number of observations is 991,480.

## 8.2. Geographic expansion of multi-establishment firms

In this subsection we show that the decrease in travel time gave rise to the geographic expansion of multi-establishment firms. We focus on all firms that patented in the initial time period and follow their subsequent opening and closure of establishments. We find that firms directed the opening (closure) of new establishments towards locations that got larger (smaller) reductions in travel time to the firm's headquarters.

We define the headquarters location $q$ of firm $F$ as the location in which the firm filed the largest amount of patents in the period 1945-1953. If firm $F$ did not file any patent in 1945-1953, or there is no unique location with the maximum amount of patents (e.g. two locations have the maximum amount of patents), then no headquarters is assigned.[91] Firms with no headquarters assigned are dropped from the estimations that required headquarters location.

[91] Using patents applied in the period 1949-1953 does not significantly affect the results. We use 1945-1953 instead as it allows us to identify headquarters location for 7% more firms.

We compute the travel time of every firm $F$'s headquarters's location $q$ to each other location $j$. We then estimate a linear probability model to study if the location decision of establishments of a firm depend on travel time to a firm's headquarters. We estimate the following regression:

$$\mathbb{1}\{establishment_{Fqjt}\} = \gamma\, log(\text{travel time}_{qjt}) + FE_{Fqj} + FE_{Fqt} + FE_{jt} + \zeta_{Fqjt} \tag{11}$$

where $\mathbb{1}\{establishment_{Fqjt}\}$ is a dummy variable that takes value 1 if firm $F$ with headquarters in location $q$ has a research establishment in location $j$ at time period $t$.[92] The coefficient $\gamma$ is a semi-elasticity: $\gamma/100$ is the change in percentage points of the probability that firm $F$ has an establishment in location $j$ when travel time increases by one percent. If travel time has a negative impact on the probability then we would expect $\gamma$ to be negative.

The inclusion of the fixed effect $FE_{Fqj}$ implies that $\gamma$ is identified only from changes in travel time and *opening* and *closure* of research establishments across time.[93] Fixed effects $FE_{Fqt}$ and $FE_{jt}$ control flexibly for changes in firm $F$ expanding and the opening establishments everywhere else, and $j$ becoming more attractive for every firm.

Table 8 presents the results jointly with predicted and observed growth rate of the probability. Column (1) presents the results of estimating equation (11). We find that the probability of firm $F$ having a subsidiary research establishment in location $j$ increases when the travel time between the firm's headquarters's location $q$ and $j$ decreases. The coefficient is -0.0364, which if we multiply it by the average change in travel time between headquarters' location and every other potential location (-34.7%), the decrease in travel time predicts an increase in the share of existing subsidiaries of 0.0126 percentage points. The result goes in the same direction as Giroud (2013) who finds

[92] $\mathbb{1}\{establishment_{Fqjt}\}$ takes value 0 if firm $F$ does not file patents in location $j$ at time period $t$. The headquarters location $q$ remains fixed for all time periods.

[93] *Opening* refers from $\mathbb{1}\{establishment_{Fqjt}\}$ switching from 0 to 1, while *closure* refers to the inverse.

that a reduction in travel time between a firm's subsidiary and its headquarters leads to an increase in investment in the subsidiary.

| | Baseline | Quartile receiving location | Initial probability | Change travel time | Predicted yearly growth rate | Observed yearly growth rate |
|---|---|---|---|---|---|---|
| Dependent Variable: | $\mathbb{1}\{establishment_{Fqjt}\}$ | | | | | |
| | (1) | (2) | | | | |
| log(travel time) | -0.0364*** (0.0088) | | 0.000810 | -34.7% | 15.94% | 1.50% |
| log(travel time) × 4th quartile | | -0.0749*** (0.0187) | 0.001895 | -36.0% | 15.41% | 0.98 % |
| log(travel time) × 3rd quartile | | -0.0150*** (0.0031) | 0.000364 | -33.4% | 15.22% | 3.03% |
| log(travel time) × 2nd quartile | | -0.0102*** (0.0028) | 0.000145 | -35.2% | 18.67% | 3.86% |
| log(travel time) × 1st quartile | | -0.0079*** (0.0025) | 0.000068 | -33.8% | 21.40% | 5.75% |
| R2 | 0.49 | 0.50 | | | | |
| N obs. effective | 19,755,792 | 19,755,792 | | | | |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 8: Subsidiaries' location and travel time to headquarters

The table shows the estimation of a linear probability model. The left panel of the table shows estimation results while the right panel shows observed and predicted growth rates of the probability. Column (1) presents the results of OLS estimation of $\mathbb{1}\{establishment_{Fqjt}\} = \gamma\, log(\text{travel time}_{qjt}) + FE_{Fqj} + FE_{Fqt} + FE_{jt} + \zeta_{Fqjt}$ or firm $F$ which has headquarters in location $q$where $\mathbb{1}\{establishment_{Fqjt}\}$ is a dummy that takes value one if firm $F$ which has headquarters in location $q$ has an establishment *open* in location $j$ at time period $t$. We define an establishment of firm $F$ in location $j$ at time period $t$ as *open* if $F$ has inventors located in $j$ that apply for patents at time period $t$. travel time$_{qjt}$ is the travel time in minutes between $F$'s headquarters location $q$ and location $j$ at time period $t$. Column (2) includes an interaction of $log(\text{travel time}_{qjt})$ with the across-technology average quartile of initial level of innovativeness of $j$. $j$'s quantile of initial innovativeness in technology $h$ is computed using the level of patents of $j$ in 1949-1953 in technology $h$. Standard errors at the non-directional location pair are presented in parentheses ($qj$ is the same non-directional location pair as $jq$). Predicted growth rates are obtained using the estimated coefficient and the change in travel time, relative to the initial probability. Yearly growth rates $g$ are obtained by computing $g = [(1 + \text{nineteen year growth rate})^{(1/19)} - 1] \times 100$, where 19 is the amount of years between 1949 and 1968.

Column (2) of Table 8 estimates the semi-elasticity of the probability of having an establishment to travel time by the quartile of innovativeness of location $j$ in 1949-1953. We compute the quartile of innovativeness at the location level by taking the average quantile across technologies within a location, only for those technologies in which the location has positive patents in 1949-1953. To be able to compare the relative impact of travel time on the growth rate of the probability, we need to relate the semi-elasticity to the baseline probability. The semi-elasticity in the lowest quartile of initial innovativeness is around 1/10th the one in the highest quartile. However, the initial probability in the lowest quartile is around 1/30th of the one in the highest quartile. Therefore, a given percentage change in travel time has an impact on the growth rate of the probability in the lowest quartile that is around 3 times the one in the highest quartile.[94] In other words, given the very low initial probability of locations in the lowest quartile of innovativeness to receive a subsidiary from a firm headquartered in another location, the small increase in percentage points represents a big relative increase in the probability.

The yearly growth rate of subsidiaries implied by the change in travel time is 21.4% for the lowest quartile while it is 15.4% for the highest quartile, implying a predicted difference of 6 percentage points in the yearly growth rate.[95] In the data we observe an average yearly growth rate which is 4.8 percentage points higher for the lowest quartile relative to the highest quartile.[96] Hence, the reduction in travel time not only predicts a geographic expansion of firms, but it also predicts that the geographic expansion is tilted towards initially less innovative locations. This pattern of geographic expansion is in line with the one observed in the data.

[94]These are approximate numbers. The precise computations: the ratio of coefficients is $0.106 = (-0.0079)/(-0.0749)$, the ratio of initial probability is $0.036 = 0.000068/0.001895$, the ratio of the growth rate is $2.94 = (-0.0079/0.000068)/(-0.0749/0.001895)$. The initial probabilities are computed as the amount of observed subsidiaries in 1949-1953 divided by the amount of (time invariant) potential subsidiaries. The amount of potential subsidiaries is the amount of firms for which we identify headquarters multiplied by the amount of locations other than headquarters location (we have 108 locations in the data, meaning that each firm has 107 potential locations for subsidiaries).

[95]For the lowest quartile, the model predicts a 3,869% increase in the probability over 19 years ($19 = 1968 - 1949$), which translates into an average yearly growth rate of 21.4%. For the highest quartile the predicted increase is 1,422%, an average yearly growth rate of 15.4%. Consistent with the computation of the relative growth rate presented in the main text: $1,422/3,869 = 0.36 \approx 0.34 \times (33.8/36.0)$, where 0.34 has to be adjusted by the fact that the average change in travel time is not the same across quartiles. The 19-year growth rates are obtained by multiplying the change in travel time (-33.8% vs -36.0%) by the coefficient (-0.0079 vs -0.0749) divided by 100, and finally dividing by the initial probability (0.000069 vs 0.001895) and multiplying by 100. For the lowest quartile: $3,869 = [(-33.8) \times (-0.0079/100)/0.000069] \times 100$, and for the highest quartile:$[1,422 = (-36.0) \times (-0.0749/100)/0.001895] \times 100$. The average yearly growth rates are computed as $21.4 \approx [(1+38.69)^{1/19} - 1] \times 100$ and $15.4 \approx [(1+14.22)^{1/19} - 1] \times 100$.

[96]The average yearly growth rate of the probability for the lowest quartile is 5.75% while it is 0.98% for the highest quartile.

# 9. Conclusion

This paper constructed a new dataset of the flight network in the United States during the beginning of the *Jet Age* and studied the impact of improvements of air travel on the creation and diffusion of knowledge. We found that the reduction in travel time led to an increase in knowledge diffusion, especially between research establishments located far apart. The reduction in travel time also led to an increase in the general access to knowledge, which had positive spillovers for the creation of new knowledge. The effect in the increase of creation of knowledge was stronger in locations initially less innovative, generating a convergence force which goes in the same direction as what is observed in the data. One of the drivers of the increase in the creation of knowledge and convergence was the geographical expansion of firms.

We provide causal evidence of *standing on the shoulders of giants*: new knowledge builds upon pre-existing knowledge. We do so by first estimating one new key parameter: the elasticity of diffusion of knowledge to travel time. Second, extending a production function of knowledge proposed in Carlino and Kerr (2015), we estimate the impact of knowledge spillovers on the creation of new knowledge. Conditional on the pre-existing distribution of knowledge, changes in travel time translate into changes in knowledge spillovers. The results show that knowledge spillovers are important for the creation of new knowledge and more so in locations which are initially less innovative.

Our novel dataset document a historical country wide event that dramatically changed the way we see time and space. Our results provide new evidence of how the introduction of jet airplanes changed the geography of innovation. Better connectivity to innovation centers in the Midwest and the Northeast led to an increase in innovation in the South and the West of the United States. In this way, jet airplanes were one of the contributing factors in the shift of innovative activity towards the South and the West of the United States.

We would like to point to the limitations of the current analysis. The results found in this paper are identified by exploiting differential time changes across establishments. As consequence, we are able to identify differential impacts and not aggregate ones. The results obtained could be consequence of general increase in the amount of diffusion and creation of knowledge, a relocation of previous diffusion and creation, or a mix of both. At the same time, the potential relocation of resources as consequence of the reduction in travel time may have increased the allocative efficiency and therefore increasing the amount of knowledge creation.

In order to separately identify the aggregate effects of travel time from relocation we plan to estimate a structural model. We consider two types of models that could potentially account for the increase in the diffusion of knowledge and the increase of innovation in the South and the West. The first option is to extend Donaldson and Hornbeck (2016) including an intermediate sector which produces knowledge, where knowledge access would enter the production function of knowledge. The second option is to modify Davis and Dingel (2019), who find that a system of cities is an equilibrium outcome in the presence of localized knowledge spillovers. We would extend the model to allow for knowledge spillovers across cities, where the degree of across-city spillovers depends on the across-city travel time. To include multi-establishment firms we would build upon Oberfield et al. (2020) who present a model of spatial equilibrium with multi-establishment firms. This model includes the location interdependency of establishments within a firm: the ideal location of an establishment of a firm depends on the location of every other establishment of the firm.

## References

Acemoglu, D., U. Akcigit, and W. R. Kerr (2016). Innovation network. *Proceedings of the National Academy of Sciences 113*(41), 11483–11488.

Aghion, P. and P. Howitt (1997). *Endogenous Growth Theory*. The MIT Press.

Agrawal, A., A. Galasso, and A. Oettl (2017). Roads and innovation. *The Review of Economics and Statistics 99*(3), 417–434.

Andersson, D., T. Berger, and E. Prawitz (2017). On the right track: Railroads, mobility and innovation during two centuries. In *On the Move: Essays on the Economic and Political Development of Sweden*, pp. 203–288. Institute for International Economic Studies, Stockholm University.

Andrews, M. J. and A. Whalley (2021). 150 years of the geography of innovation. *Regional Science and Urban Economics*, 103627.

Arzaghi, M. and J. V. Henderson (2008). Networking off madison avenue. *The Review of Economic Studies 75*(4), 1011–1038.

Audretsch, D. B. and M. P. Feldman (2004). Knowledge spillovers and the geography of innovation. In *Handbook of regional and urban economics*, Volume 4, pp. 2713–2739. Elsevier.

Bai, J. J., W. Jin, and S. Zhou (2021). Proximity and knowledge spillovers: Evidence from the introduction of new airline routes.

Balsmeier, B., M. Assaf, T. Chesebro, G. Fierro, K. Johnson, S. Johnson, G.-C. Li, S. Lck, D. O'Reagan, B. Yeh, G. Zang, and L. Fleming (2018). Machine learning and natural language processing on the patent corpus: Data, tools, and new measures. *Journal of Economics & Management Strategy 27*(3), 535–553.

Baum-Snow, N. (2007). Did highways cause suburbanization? *The quarterly journal of economics 122*(2), 775–805.

Bergé, L. (2018). Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm. *CREA Discussion Papers* (13).

Bloom, N., M. Schankerman, and J. Van Reenen (2013). Identifying technology spillovers and product market rivalry. *Econometrica 81*(4), 1347–1393.

Borenstein, S. and N. L. Rose (2014, June). *How Airline Markets Work... or Do They? Regulatory Reform in the Airline Industry*, pp. 63–135. University of Chicago Press.

C.A.B. (1951, 1956, 1961, 1966). *Air carrier traffic statistics.* Civil Aeronautics Board.

Caliendo, L. and F. Parro (2015). Estimates of the trade and welfare effects of nafta. *The Review of Economic Studies 82*(1), 1–44.

Campante, F. and D. Yanagizawa-Drott (2017, 12). Long-Range Growth: Economic Development in the Global Network of Air Links*. *The Quarterly Journal of Economics 133*(3), 1395–1458.

Carlino, G. and W. R. Kerr (2015). Agglomeration and innovation. *Handbook of regional and urban economics 5*, 349–404.

Catalini, C., C. Fons-Rosen, and P. Gaulé (2020). How do travel costs shape collaboration? *Management Science 66*(8), 3340–3360.

Caves, R. E. (1962). *Air transport and its regulators: an industry study*. Harvard economic studies v. 120. Cambridge: Harvard University Press.

Coscia, M., F. Neffke, and R. Hausmann (2020). Knowledge diffusion in the network of international business travel. *Nature Human Behaviour 4*(10).

Davis, D. R. and J. I. Dingel (2019). A spatial knowledge economy. *American Economic Review 109*(1), 153–70.

De Rassenfosse, G. and A. B. Jaffe (2017). Econometric evidence on the r&d depreciation rate. Technical report, National Bureau of Economic Research.

Dhaene, G. and K. Jochmans (2015). Split-panel jackknife estimation of fixed-effect models. *The Review of Economic Studies 82*(3), 991–1030.

Dijkstra, E. W. et al. (1959). A note on two problems in connexion with graphs. *Numerische mathematik 1*(1), 269–271.

Donaldson, D. and R. Hornbeck (2016). Railroads and american economic growth: A market access approach. *The Quarterly Journal of Economics 131*(2), 799–858.

Duranton, G., P. Martin, T. Mayer, and F. Mayneris (2009). The economics of clusters: evidence from france.

Duranton, G. and D. Puga (2004). Micro-foundations of urban agglomeration economies. In *Handbook of regional and urban economics*, Volume 4, pp. 2063–2117. Elsevier.

Eaton, J. and S. Kortum (2002). Technology, geography, and trade. *Econometrica 70*(5), 1741–1779.

Faber, B. (2014). Trade integration, market size, and industrialization: evidence from china's national trunk highway system. *Review of Economic Studies 81*(3), 1046–1070.

Feyrer, J. (2019). Trade and incomeexploiting time series in geography. *American Economic Journal: Applied Economics 11*(4), 1–35.

Fogel, R. W. (1963). *Railroads and American economic growth: essays in econometric history*. Ph. D. thesis, Johns Hopkins University.

Furman, J. L. and S. Stern (2011, August). Climbing atop the shoulders of giants: The impact of institutions on cumulative research. *American Economic Review 101*(5), 1933–63.

Giroud, X. (2013, 03). Proximity and Investment: Evidence from Plant-Level Data *. *The Quarterly Journal of Economics 128*(2), 861–915.

Glaeser, E. (2011). *Triumph of the City*. Macmillan.

Gross, D. P. (2019). The consequences of invention secrecy: Evidence from the uspto patent secrecy program in world war ii. Technical report, National Bureau of Economic Research.

Haines, M. R. et al. (2010). Historical, demographic, economic, and social data: the united states, 1790–2002. *Ann Arbor, MI: Inter-university Consortium for Political and Social Research*.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2001, October). The nber patent citation data file: Lessons, insights and methodological tools. Working Paper 8498, National Bureau of Economic Research.

Hansen, B. E. (2021). *Econometrics*.

Head, K. and T. Mayer (2014). Chapter 3 - gravity equations: Workhorse,toolkit, and cookbook. In G. Gopinath, E. Helpman, and K. Rogoff (Eds.), *Handbook of International Economics*, Volume 4 of *Handbook of International Economics*, pp. 131–195. Elsevier.

Herzog, I. (2021). National transportation networks, market access, and regional economic growth. *Journal of Urban Economics 122*, 103316.

Hovhannisyan, N. and W. Keller (2015, March). International business travel: an engine of innovation? *Journal of Economic Growth 20*(1), 75–104.

Jaffe, A. B., M. Trajtenberg, and M. S. Fogarty (2000, May). Knowledge spillovers and patent citations: Evidence from a survey of inventors. *American Economic Review 90*(2), 215–218.

Jaffe, A. B., M. Trajtenberg, and R. Henderson (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics 108*(3), 577–598.

Jaworski, T. and C. T. Kitchens (2019). National policy for regional development: Historical evidence from appalachian highways. *Review of Economics and Statistics 101*(5), 777–790.

Jayaratne, J. and P. E. Strahan (1996). The finance-growth nexus: Evidence from bank branch deregulation. *The Quarterly Journal of Economics 111*(3), 639–670.

Jones, C. I. (2002, March). Sources of u.s. economic growth in a world of ideas. *American Economic Review 92*(1), 220–239.

Kerr, W. R. and F. Robert-Nicoud (2020). Tech clusters. *Journal of Economic Perspectives 34*(3), 50–76.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman (2017, 03). Technological Innovation, Resource Allocation, and Growth*. *The Quarterly Journal of Economics 132*(2), 665–712.

Krugman, P. R. (1991). *Geography and trade*. MIT press.

Lucas, R. E. (1993). Making a miracle. *Econometrica 61*(2), 251–272.

Magerman, T., B. Van Looy, and X. Song (2006). Data production methods for harmonized patent statistics : patentee name harmonization.

Manson, S., J. Schroeder, D. Van Riper, T. Kugler, and S. Ruggles (2020). Ipums national historical geographic information system: Version 15.0.

Marshall, A. (1890). *Principles of economics*.

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.

Moretti, E. (2021, October). The effect of high-tech clusters on the productivity of top inventors. *American Economic Review 111*(10), 3328–75.

Oberfield, E., E. Rossi-Hansberg, P.-D. Sarte, and N. Trachter (2020). Plants in space. Technical report, National Bureau of Economic Research.

Pascali, L. (2017). The wind of change: Maritime technology, trade, and economic development. *American Economic Review 107*(9), 2821–54.

Perlman, E. R. (2016). Dense enough to be brilliant: patents, urbanization, and transportation in nineteenth century america. *Work. Pap., Boston Univ.*

Petralia, S. (2019). HistPat International Dataset.

Petralia, S., P.-A. Balland, and D. Rigby (2016). HistPat Dataset.

Redding, S. and A. J. Venables (2004). Economic geography and international inequality. *Journal of international Economics 62*(1), 53–82.

Sanderson, E. and F. Windmeijer (2016). A weak instrument f-test in linear iv models with multiple endogenous variables. *Journal of econometrics 190*(2), 212–221.

Silva, J. M. C. S. and S. Tenreyro (2006). The log of gravity. *The Review of Economics and Statistics 88*(4), 641–658.

Söderlund, B. (2020). The importance of business travel for trade: Evidence from the liberalization of the soviet airspace. Working Paper Series 1355, Research Institute of Industrial Economics.

Staiger, D. and J. H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica 65*(3), 557–586.

Stock, J. H. and M. Yogo (2005). *Testing for Weak Instruments in Linear IV Regression*, pp. 80108. Cambridge University Press.

Storper, M. and A. J. Venables (2004, 08). Buzz: face-to-face contact and the urban economy. *Journal of Economic Geography 4*(4), 351–370.

Tsiachtsiras, G. (2021). Transportation networks and the rise of the knowledge economy in 19th century france.

Weidner, M. and T. Zylkin (2021). Bias and consistency in three-way gravity models. *Journal of International Economics 132*, 103513.

Wooldridge, J. M. (2014). Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables. *Journal of Econometrics 182*(1), 226–234.

# A. Appendix: Travel Time Data

## A.1. Data Construction

We construct a dataset of travel times by plane between US MSAs for the years 1951, 1956, 1961, 1966. We get information of direct flights from airline flight schedules and feed this information into an algorithm to allow for indirect flights. For each MSA pair with airports served by at least one of the airlines in our dataset we compute the fastest travel time in each of the four years.

Using images of flight schedules, we digitized the flight network for six major airlines: American Airlines (AA), Eastern Air Lines (EA), Trans World Airlines (TWA), United Airlines (UA), Braniff International Airways (BN) and Northwest Airlines (NW). Note that the first four in this list were often referred to as the *Big Four*, highlighting their dominant position in the market. They alone accounted for 74% of domestic trunk revenue passenger-miles from February 1955 to January 1956. Together the six airlines accounted for 82% of revenue passenger-miles in that same period, 77% from February 1960 to January 1961 and 78% from February 1965 to January 1966 (C.A.B., 1966). Our sample of airlines thus covers a vast share of the domestic market for air transport. In addition, the airlines were chosen to maximize geographic coverage.

In total we obtain a sample of 5,910 flights. These flights often have multiple stops. If we count each origin-destination pair of these flights separately, our sample contains 17,469 legs.

Table 9 lists the exact dates of when flight schedules we digitized became effective. Due to limited data availability not all flight schedules are drawn from the same part of the year. As seasonality of the network seems limited and given the large market share of the airlines we consider, our data is a good approximation of the network in a given year.

Table 9: Date of Digitized Flight Schedules

| Airline | 1951 | 1956 | 1961 | 1966 |
|---|---|---|---|---|
| AA | September 30 | April 29 | April 30 | April 24 |
| EA | August 1 | October 28 | April 1 | April 24 |
| TWA | August 1 | September 1 | April 30 | May 23 |
| UA | April 29 | July 1 | June 1 | April 24 |
| BN | August | August 15 | April 30 | April 24 |
| NW | April 29 | April 29 | May 28 | March 1 |
| PA | June 1 | July 1 | August 1 | August 1 |

Figure 18 shows two pages of the flight schedule published by American Airlines in 1961. Each column corresponds to one flight. As can be seen, one flight often has multiple stops. Departure and arrival times in most flight schedules are indicated using the 12-hour system. PM times can be distinguished from AM times by their bold print. In the process of digitization we converted the flight schedules to the 24-hour system. Times in most tables are in local time. We thus recorded the time zones that are indicated next to the city name and converted them to Eastern Standard Time.

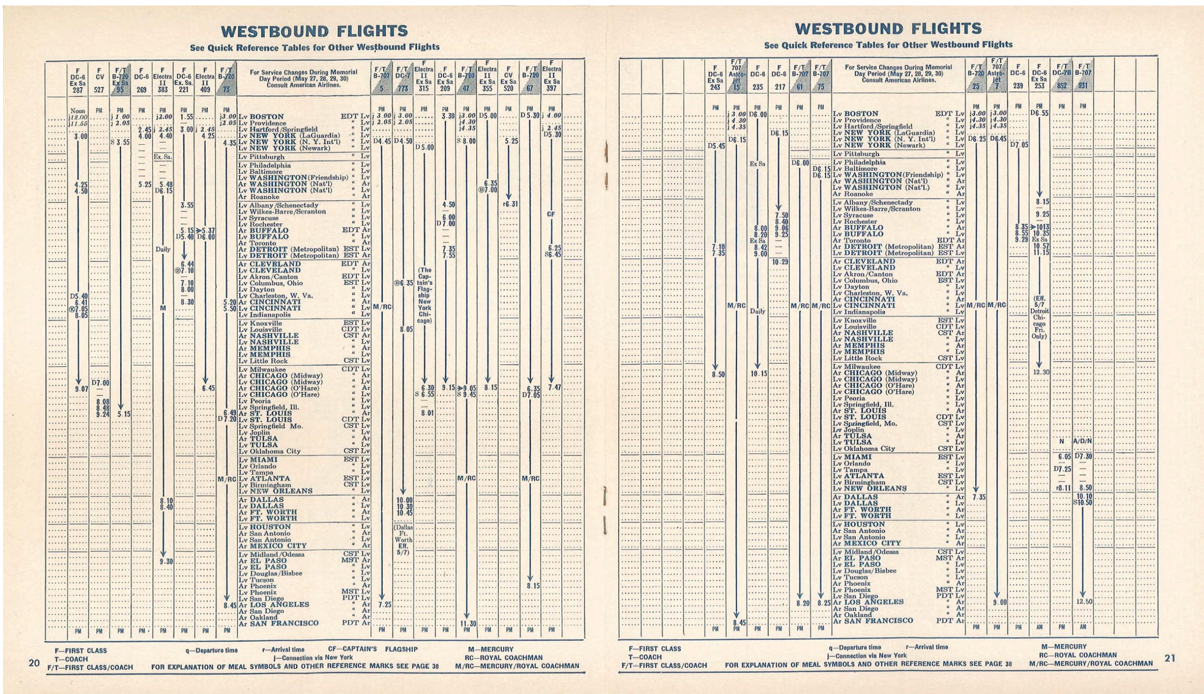



Figure 18: Flight Schedule American Airlines 1961.

To obtain exact geographical information on where airports are located, we match city names to their IATA airport codes. We use the addresses of ticket offices that are indicated on the last pages of the flight schedules. Most of the ticket offices were located directly at the airport, allowing to infer the airport the airline was serving in a given year. For some flight schedules we are missing these last pages and used information from adjacent years in order to identify airports. We also manually verified the airport match using various online sources. We then obtain geographical coordinates from a dataset provided by `https://ourairports.com/` (downloaded July 2020).

From the flight schedule we also collect information on the aircraft model, indicated next to the flight number. Using various online sources, we manually identified aircraft models that are powered by a jet engine. We thus know on which connections airlines were using jet aircraft.

Flight Schedules also contain information on connecting flights. For example, the second column in figure 18 indicates a departure from Boston leaving at 12.00 local time. A footnote is added to the departure time indicating that this departure is a connection via New York. It is thus not operated by flight 287 otherwise described in column 2, but it is just supplementary information for the passenger. As we are interested in the speed of aircraft and the actual travel time on a given link, this information on connecting flights would pollute our data and we thus delete this supplementary information.

As outlined above, the digitization requires human input. It is thus prone error-prone. The travel time calculation relies on each link in the network, and if one important connection has a miscoded flight, it might potentially distort the travel time between many MSA pairs. We thus implement an elaborate method to detect mistakes in the digitization process. In particular, after the initial transcription, we regress the observed duration of the flight on a set of explanatory variables: the full interaction of distance, a set of airline indicators, a set of year indicators and a dummy variable indicating whether the aircraft is powered by a jet engine or not. This linear model yields an

$R^2$ above 95%. We then compute the predicted duration of each flight and obtain the relative deviation from the observed duration. If the deviation is above 50%, we manually check whether the transcribed information is correct. If we find a mistake, we correct the raw data, rerun the regression and recompute relative deviations, until all the observations with more than 50% deviation have been manually verified.

For 15 connections, the information was correctly transcribed from the flight schedule, but the flight time differed a lot from other flights with similar distances that used the same aircraft. The implied aircraft speed for these cases is either unrealistically high or low, in one case the implied flight time is even negative. These cases seem to be typos introduced when the flight schedule was created (e.g. a "2" becomes a "3"). Instead of inferring what the true flight schedule was which is not always obvious, we drop these cases. Table 10 lists all 15 cases.

Table 10: Dropped Connections

| | Airline | Year | Origin | Destination | Departure Time | Arrival Time |
|---|---|---|---|---|---|---|
| 0 | UA | 66 | TYS | DCA | 1940 | 2036 |
| 1 | UA | 66 | LAX | BWI | 2150 | 1715 |
| 2 | UA | 66 | CHA | TYS | 1635 | 1909 |
| 3 | PA | 66 | SFO | LAX | 2105 | 1850 |
| 4 | PA | 66 | SEA | PDX | 705 | 935 |
| 5 | PA | 56 | PAP | SDQ | 830 | 835 |
| 6 | PA | 51 | HAV | MIA | 800 | 903 |
| 7 | PA | 51 | SJU | SDQ | 825 | 830 |
| 8 | NW | 66 | HND | OKA | 655 | 1135 |
| 9 | EA | 66 | ORD | MSP | 2340 | 2340 |
| 10 | EA | 56 | SDF | MDW | 1352 | 1418 |
| 11 | EA | 56 | GSO | RIC | 2207 | 2204 |
| 12 | AA | 56 | PHX | TUS | 1630 | 1655 |
| 13 | PA | 51 | STR | FRA | 1320 | 1540 |
| 14 | EA | 66 | TPA | JFK | 1330 | 1548 |

As our analysis is at the MSA level, we match airports to 1950 MSA boundaries. Each airport is matched to all MSAs for which it lies inside the MSA boundary or at most

15km away from the MSA boundary. If we focus only on airports contained within MSA boundaries, we would, for example, drop Atlanta's airport. Of 275 US airports, 156 airports are matched to at least one MSA. 18 of these are matched to two MSAs and Harrisburg International Airport is matched to three MSAs: Harrisburg, Lancaster and York. Out of 168 MSAs, 142 are at some point connected to the flight network in our dataset. In table 11 we present the 168 MSAs, the ones that are connected at least once, and the ones that are connected in the four years.

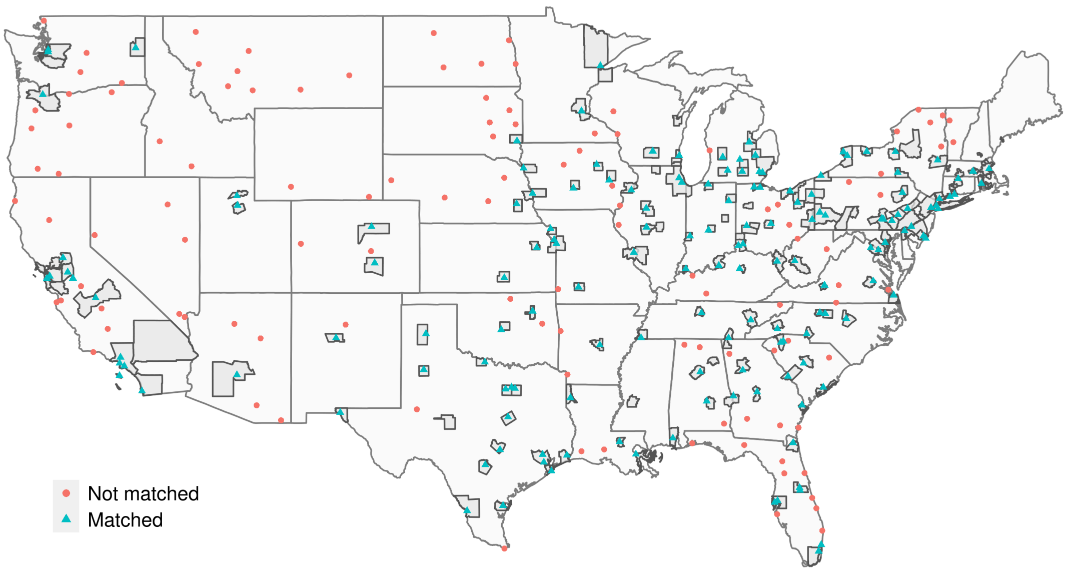



Figure 19: Airports matched to MSAs.

Next, we compute the shortest travel time for every airport pair, and then take the minimum to obtain shortest travel time at the MSA pair level. In particular, we apply Dijkstra's algorithm to compute shortest paths (Dijkstra et al., 1959). We adjust this algorithm to take into account the exact timing of the flight schedules. We consider a possible departure time $t$ from origin city $o$ and then compute the shortest path to destination city $d$ at this time of the day. If getting to $d$ requires switching flights, we account for the required time at the location of the layover. We repeat this procedure for every possible departure time $t$ at origin city $o$ and then take the minimum that gives us the fastest travel time from $o$ to $d$, $\tau_{od}$.

The flight schedule format requires us to make one assumption. In particular, the flight schedule for a multi-stop flight may either indicate the arrival time or the departure time for a particular stop. If the flight schedule only lists the departure time, we need to infer the arrival time and vice versa. We allow for five minutes between arrival and departure. This is relatively low, but still in the range of observed difference between departure and arrival for cases where we observe both. As correspondences may have been ensured by airlines in reality, i.e. one aircraft waiting with departure until other aircraft arrive, we opted for the lower end of the observed range of stopping times.

Finally, since the shortest travel time measure may not capture the benefits of a highly frequented hub, we also calculate the daily average of the shortest travel time. In particular, we compute the shortest travel time at every full hour of the day and take the average. This measure thus captures the benefits of being located near an airport where flights depart many times per day.

To conclude, we end up with a set of four origin-destination matrices indicating the fastest travel time (and another set with the average daily travel time) between US MSAs in 1951, 1956, 1961 and 1966.

## A.2. Descriptive Statistics

Table 12 shows the number of non-stop connections between MSAs by year and airline. It underlines the dominant position of the *Big Four* (AA, EA, TW, UA) which were much bigger than their competitors (BN and NW). The growth of the airline industry is also apparent. All airlines had the lowest number of connections in 1951 and subsequently extended their network. At the same time the average distance of the connections gradually increased over time. Part of this may have been due to jet technology allowing for longer aircraft range. We thus analyze a period where more and longer flights are introduced.

Table 12: Domestic Non-Stop Connections by Airline and Year

| Airline | Year | Number of connections | Jet Share (connections) | Jet Share (km) | Mean Distance (in km) |
|---|---|---|---|---|---|
| AA | 1951 | 258 | 0.00 | 0.00 | 515.32 |
| AA | 1956 | 367 | 0.00 | 0.00 | 889.66 |
| AA | 1961 | 325 | 22.15 | 50.50 | 768.24 |
| AA | 1966 | 282 | 73.40 | 89.52 | 1020.36 |
| BN | 1951 | 96 | 0.00 | 0.00 | 317.90 |
| BN | 1956 | 210 | 0.00 | 0.00 | 380.60 |
| BN | 1961 | 176 | 8.52 | 18.84 | 460.41 |
| BN | 1966 | 150 | 72.00 | 76.64 | 553.09 |
| EA | 1951 | 345 | 0.00 | 0.00 | 319.87 |
| EA | 1956 | 479 | 0.00 | 0.00 | 412.60 |
| EA | 1961 | 595 | 3.70 | 13.28 | 441.42 |
| EA | 1966 | 492 | 54.47 | 75.46 | 569.01 |
| NW | 1951 | 77 | 0.00 | 0.00 | 521.70 |
| NW | 1956 | 95 | 0.00 | 0.00 | 724.77 |
| NW | 1961 | 127 | 11.02 | 32.43 | 824.59 |
| NW | 1966 | 136 | 77.94 | 90.86 | 945.81 |
| TW | 1951 | 210 | 0.00 | 0.00 | 503.69 |
| TW | 1956 | 253 | 0.00 | 0.00 | 711.78 |
| TW | 1961 | 240 | 28.75 | 54.63 | 807.72 |
| TW | 1966 | 265 | 86.42 | 96.05 | 1143.30 |
| UA | 1951 | 291 | 0.00 | 0.00 | 492.88 |
| UA | 1956 | 361 | 0.00 | 0.00 | 714.39 |
| UA | 1961 | 323 | 31.89 | 65.32 | 803.49 |
| UA | 1966 | 533 | 49.91 | 79.54 | 781.38 |

While these changes in the network are remarkable, airlines were constrained by the regulator in opening new routes. Accordingly, table 13 shows that the network remains relatively stable over time with more than three quarters of connections remaining intact within a five-year window. Interestingly, during the beginning of the jet age (i.e. 1956 to 1961), the network appears to have been especially stable, with only 11% of connections either disappearing or newly being added. Thus, the rise of jet aircraft did not lead to a vast reshaping of the network. Given the very different technology, this may be surprising, but may partly be due to heavy regulation.

The table also shows that newly introduced routes were over long distances whereas those discontinued were operating on shorter distances. When changes in the network took place, they thus seemed to improve the network for places further apart.

Table 13: Network Changes (weighted by frequency)

| Period | Remain connected | Newly connected | Disconnected |
|---|---|---|---|
| Share of Non-stop Connections (%) | | | |
| 1951 to 1956 | 78.47 | 16.79 | 4.74 |
| 1956 to 1961 | 88.96 | 6.43 | 4.6 |
| 1961 to 1966 | 80.64 | 12.37 | 6.99 |
| Mean distance (km) | | | |
| 1951 to 1956 | 411 | 1075 | 337 |
| 1956 to 1961 | 524 | 914 | 972 |
| 1961 to 1966 | 568 | 769 | 450 |

Table 14: Network Changes

| Period | Remain connected | Newly connected | Disconnected |
|---|---|---|---|
| Connected MSAs | | | |
| 1951 to 1956 | 119 | 7 | 8 |
| 1956 to 1961 | 122 | 0 | 4 |
| 1961 to 1966 | 114 | 7 | 8 |
| Non-stop Connections | | | |
| 1951 to 1956 | 721 | 357 | 124 |
| 1956 to 1961 | 908 | 231 | 170 |
| 1961 to 1966 | 912 | 331 | 227 |

Changes in the number of connected MSAs and connections among them. A MSA is connected if in our data it appears as having at least one incoming and one outgoing flight. A non-stop connection refers to a pair of origin MSA-destination MSA between which a non-stop flight operates.

Figure 20 shows all non-stop connections in our data weighted by the (log) frequency. Initially, the network was concentrated in the Eastern states and transcontinental routes were not yet established, due to technological limitations. In contrast, in the 1960s, after the jet is introduced, intercontinental routes quickly emerge and are operated at a high frequency. Similarly, direct connections from the Northeast to Florida intensify. The figure echos the findings from table 14 which illustrates that the overall number of MSA pairs with a direct connection increases over time.

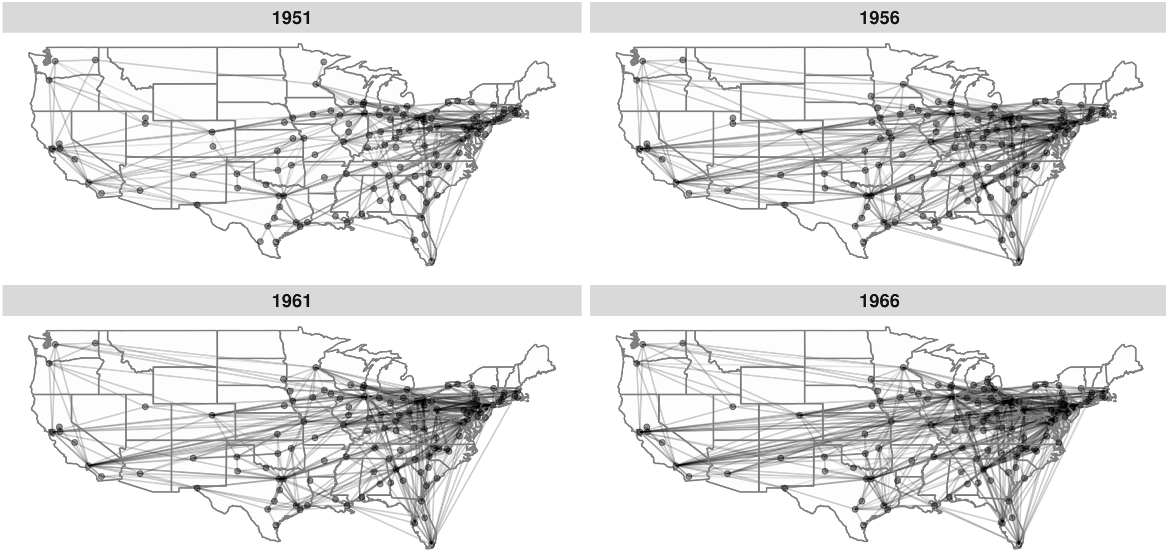



Figure 20: Flight Network by Year. Weighted by log weekly frequency.

Airlines differed in their speed of adoption of the newly arrived jet aircraft. Table 12 shows that, in 1961, 65% of UA's connections between MSAs were flown using a jet aircraft (weighted by distance), whereas this was only true for 13% of EA's connections. While adoption was heterogeneous across airlines, adoption was fast. By 1966, all airlines were operating 75% of their connections with jet aircraft (weighted by distance).

Figure 21 show the average speed of jet and propeller aircraft by distance. Generally, jet aircraft were substantially faster, but especially so on long-distance flights, where they could be up to twice as fast as propeller-driven aircraft. This particularly stark difference in speed for long-haul flights is also reflected by adoption. Figure 22 shows that jet aircraft were first introduced on long-haul flights. Only 50% of MSA pairs at around 1,500 km distance had at least one jet aircraft operating, whereas 100% of pairs above 3,000 km. Then, in the late 1960s, they were also gradually introduced on shorter distances. In fact, for all pairs above 2,000 km there was at least one jet engine-powered flight.

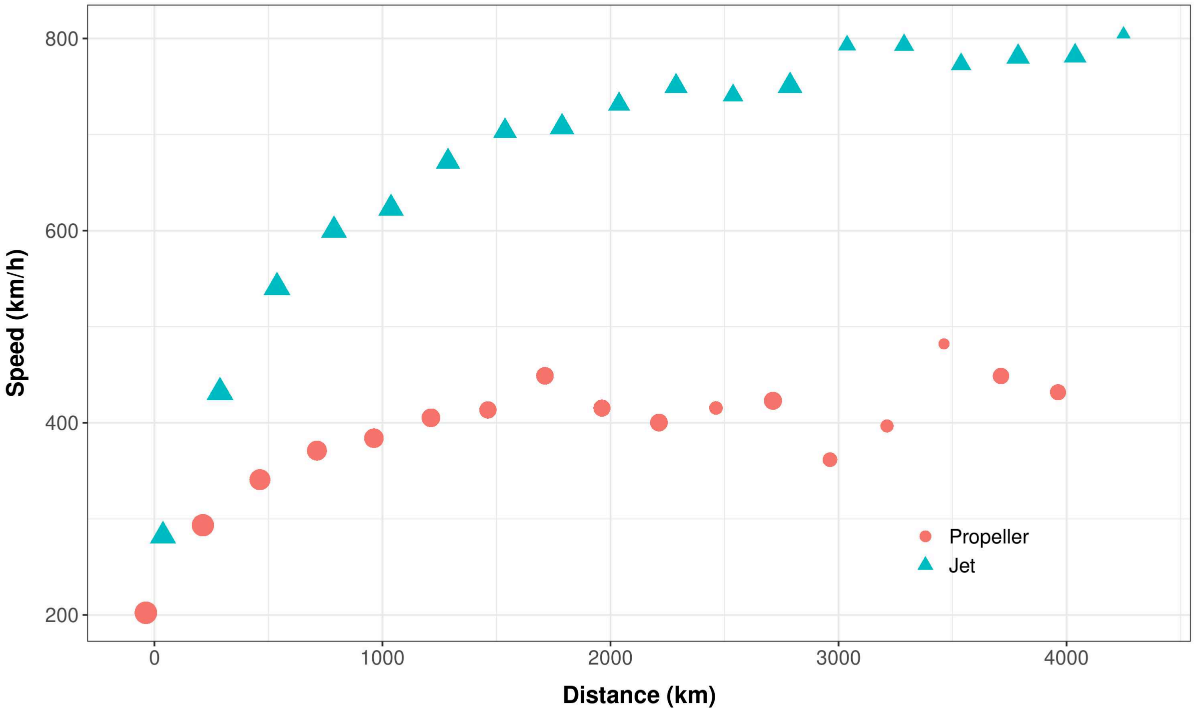



Figure 21: Speed by Aircraft Type. Pooling all Years.

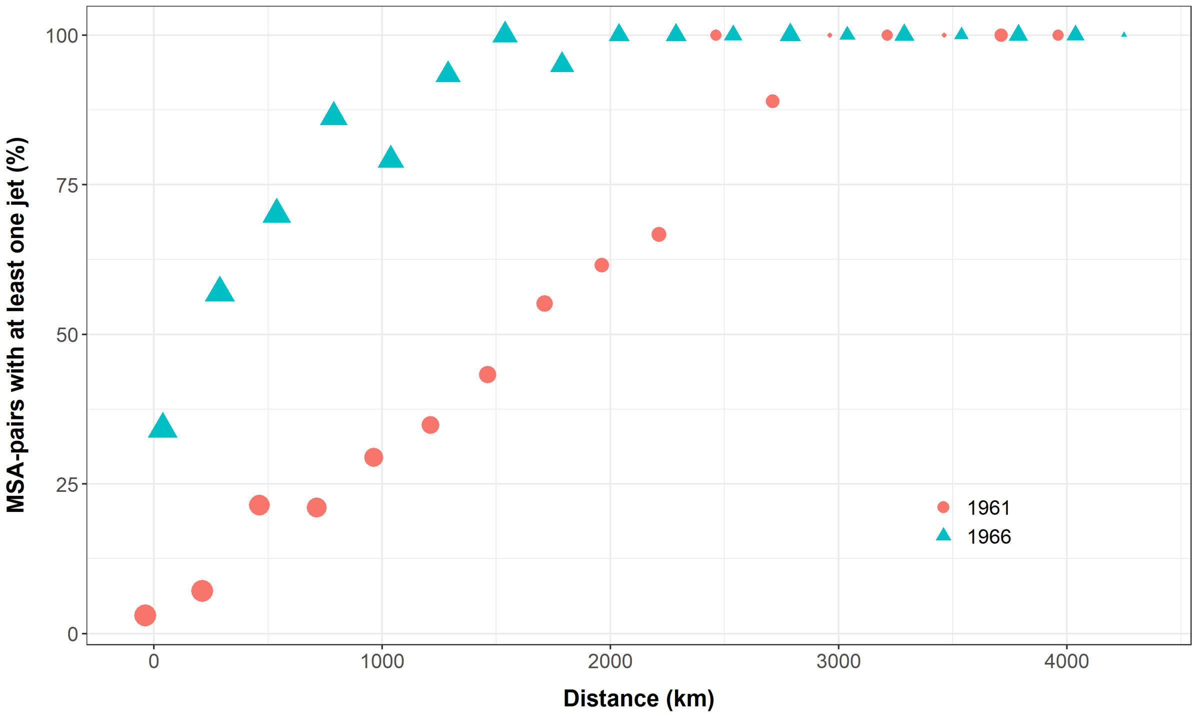



Figure 22: Jet Adoption

Figure 23 shows on which routes jets were operating. In the early days of the jet age it was mainly the transcontinental corridor between New York and California that benefited. In 1966 propeller aircraft were already being phased out and only operating in the dense Eastern part of the US where distances between cities are relatively small.

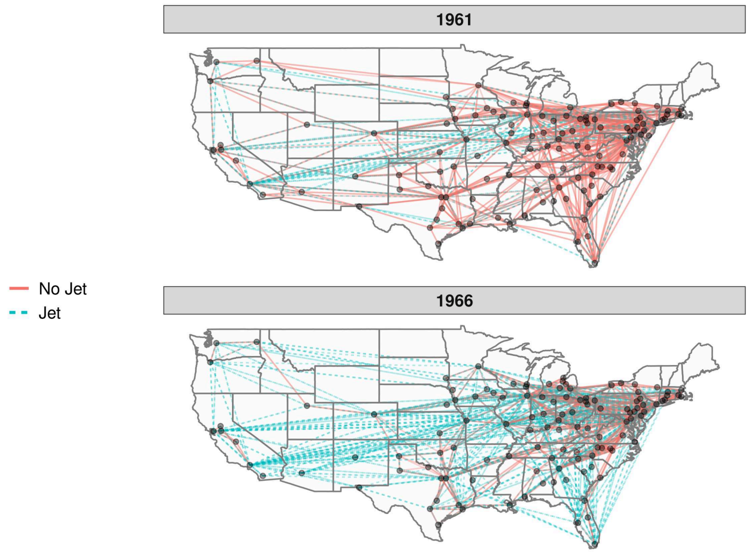



Figure 23: Jet Adoption by Year

The increase in speed due to jet aircraft caused a dramatic reduction in travel times between US cities. When looking at the full origin-destination matrix, i.e. including indirect flights, a network-wide reduction in travel time becomes apparent. Figure 24 shows travel times between US MSAs. While the figure shows a gradual decline in travel time from 1951 to 1966, it also illustrates that conditional on distance and year a large amount of variation in travel time remains, as only a small fraction of all MSA pairs were connected via a direct flight (around 8.5% in 1966).

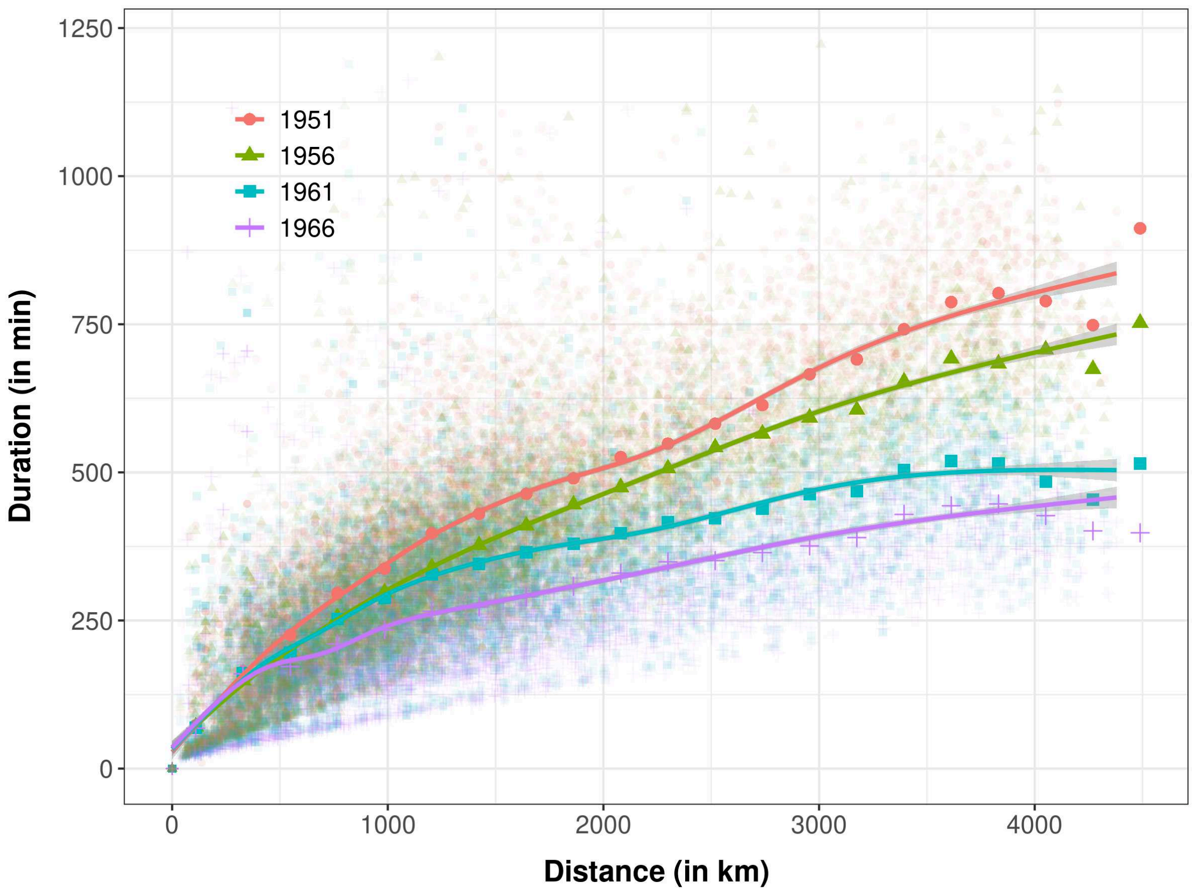



Figure 24: Travel Times between US MSAs.

Figure 25 that the change in travel time is accompanied by a reduction of the amount of legs needed to connect two MSAs at every distance. This reduction is specially marked between 1951 and 1956, and 1961 and 1966. In Figure 26 we open up the change in travel time by the way an MSA pair was connected in 1951 and 1966: either directly (non-stop flight) or indirectly (connecting flight). We observe that much of the increase in travel time for MSA pairs less than 250km apart comes from routes that were operated non-stop and then it needed a connecting flight. Interestingly, for MSA-pairs more than 2,000km apart travel time reduced on average 42% for those pairs that were connected indirectly in both periods, and 51% for those that switched from indirect to direct. This fact shows the relevance of improvements in flight technology even for MSAs not directly connected. It could be the case that a reduction in the amount of legs or an increase in frequency of flights reduces layover time. In Figure 28 we compare the change in travel time from 1951 to 1966 with a fictitious change in travel time in which we eliminate layover time in both time periods. We observe that the average change in travel time is stronger at every distance if we disregard layover time. This implies that the relative importance of layover time over total travel time increases between 1951 and 1966, preventing total travel time to decrease proportionally to the change of in-flight travel time.

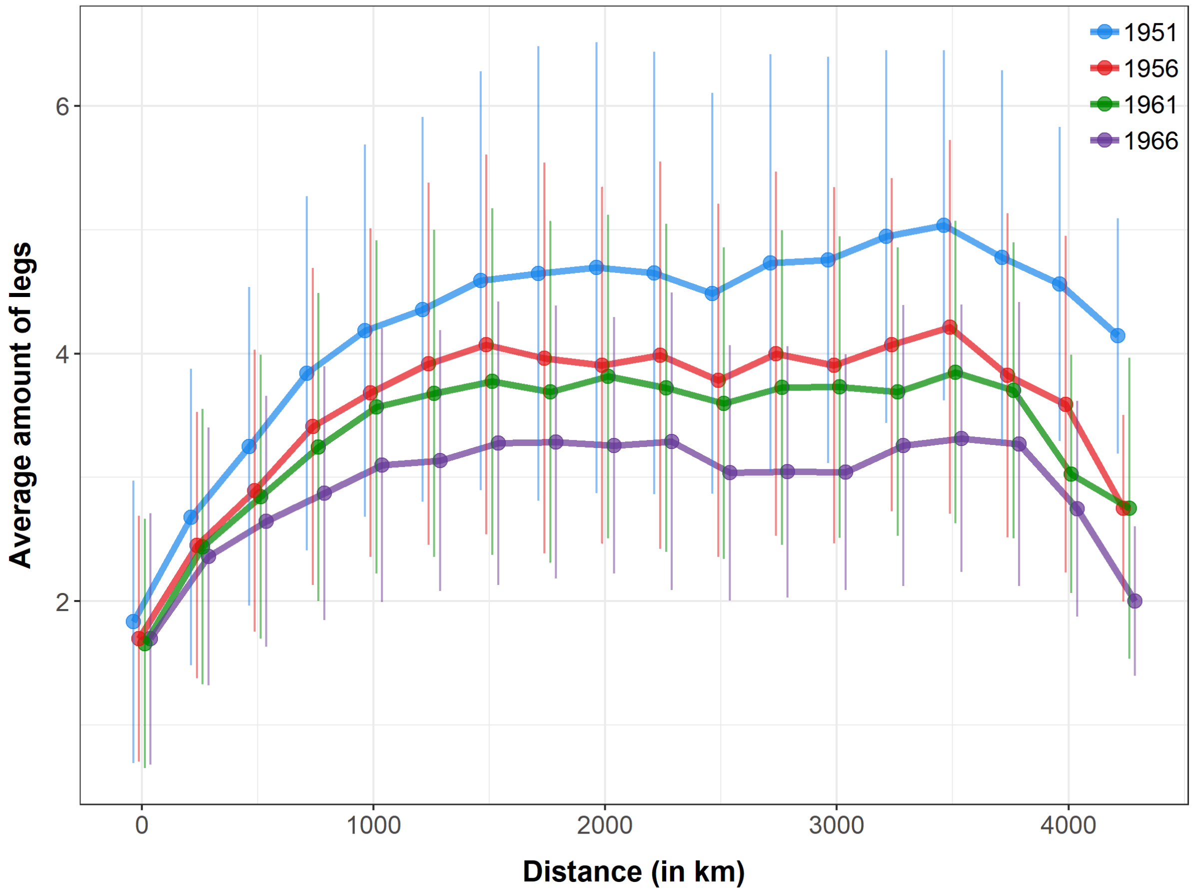



Figure 25: Average amount of legs per route

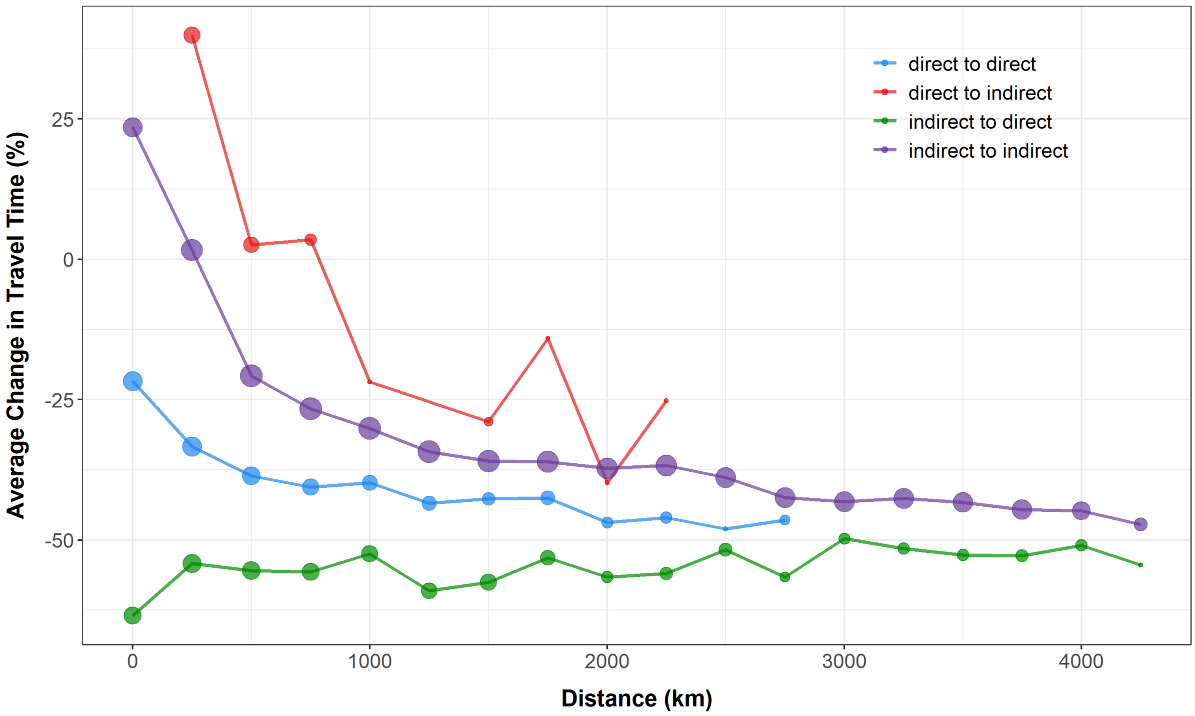



Figure 26: Change in US travel time 1951 to 1966: connections
97

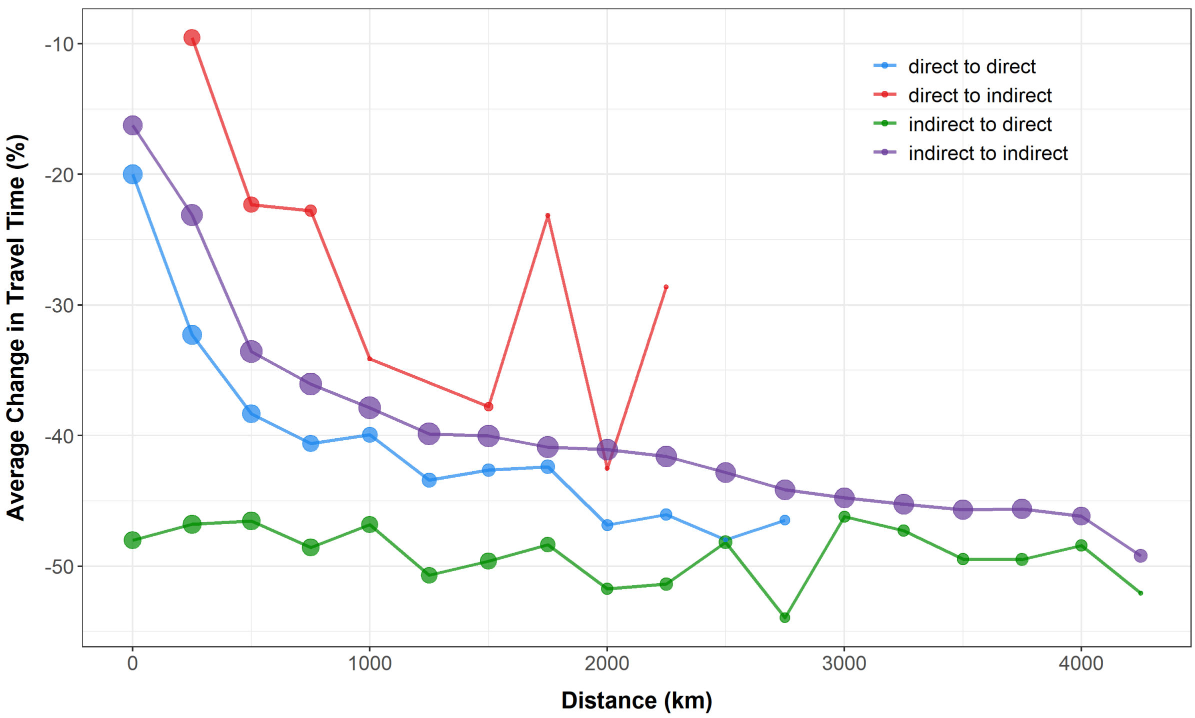



Figure 27: Change in US travel time 1951 to 1966: connections, discarding layover time
98

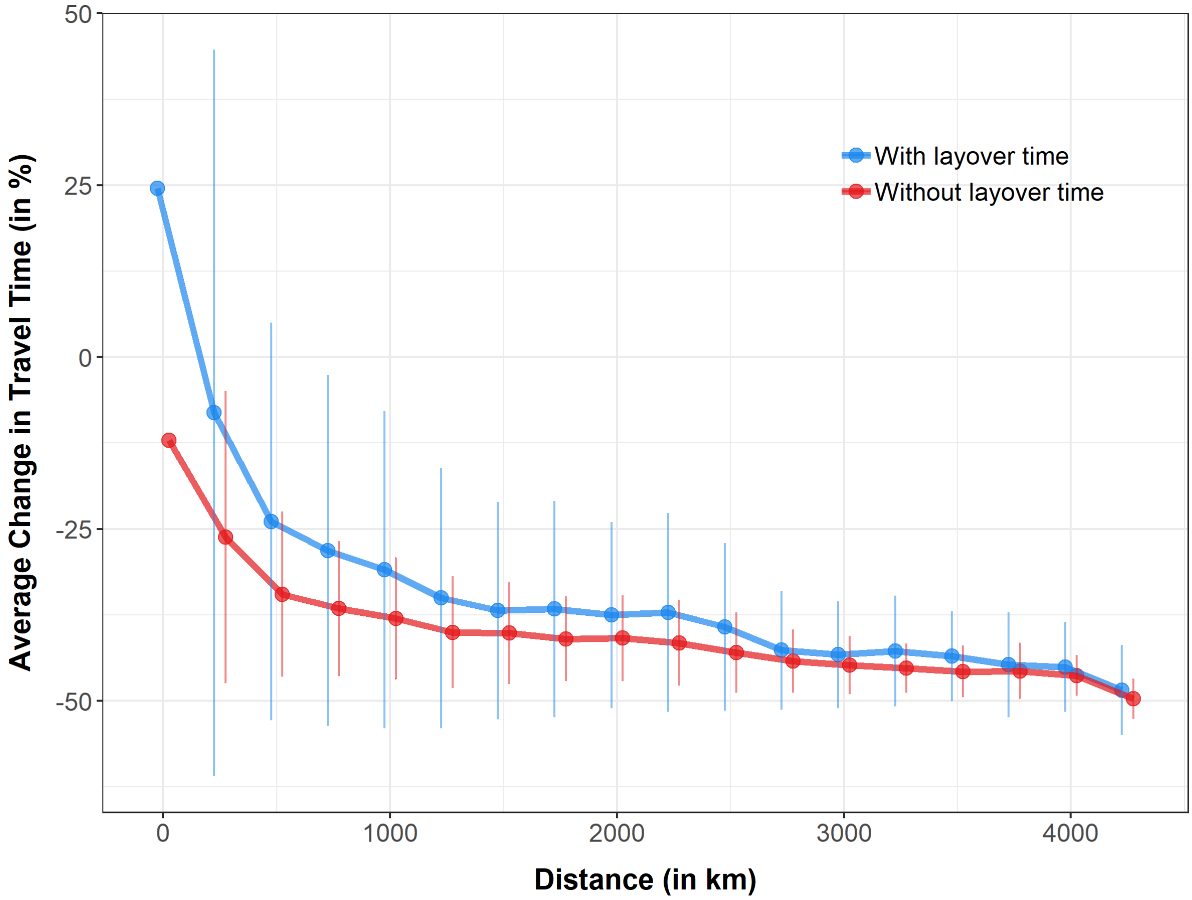



Figure 28: Change in US travel time 1951 to 1966: layover time

In figure 29 we show the average change in travel time in three counterfactual flight networks. The first counterfactual fixes the flight routes[99] and allows aircraft speed to evolve. The second counterfactual fixes aircraft speed and allows flight routes to evolve. The third counterfactual allows both flight routes and aircraft speed to evolve. We obtain that around 90% of the change in travel time is due to the change in speed of aircrafts, while around 10% of the change is due to the change in the flight routes. In the figure 30 in the appendix we show that the proportion is relatively constant for all distances. This confirms that most of the observed changes in the network are due to improvements in the flight technology.

[99]Fixes the origin-destination airports that are connected with a non-stop flight

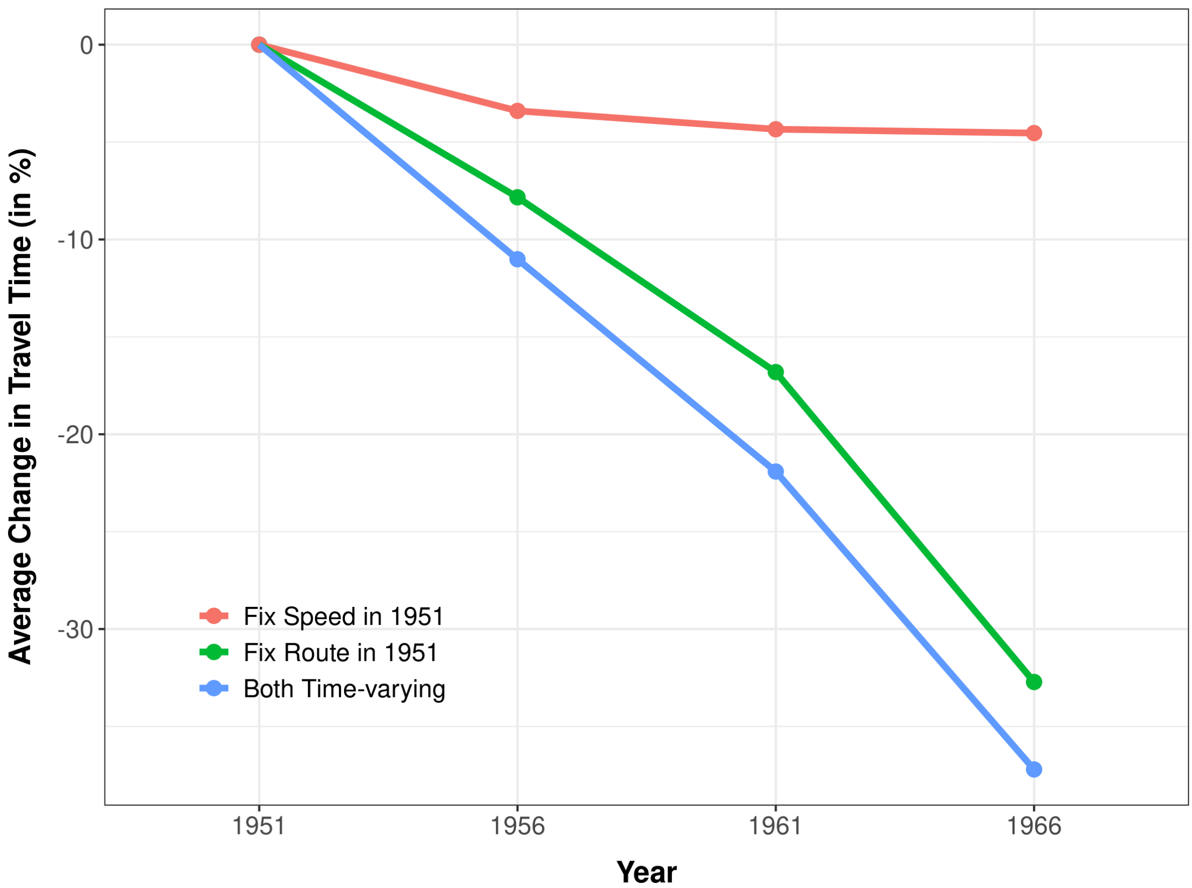



Figure 29: Counterfactual change in travel time

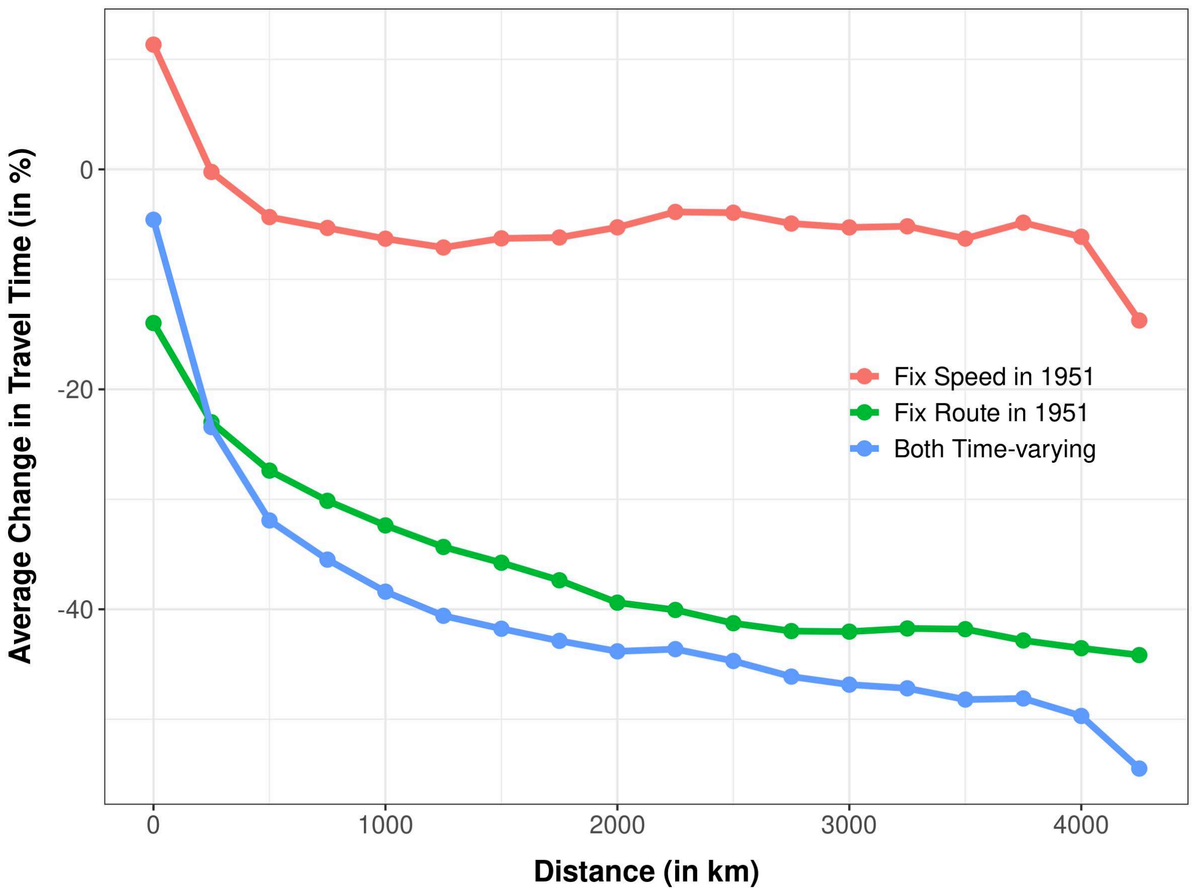



Figure 30: Counterfactual change in travel time 1951-1966

In addition to the changes over time in the network leading to faster travel times, another feature of the US airline industry becomes salient in the data: airlines' regional specialization. As figure 31 shows, while there was competition among the airlines in our dataset on the major routes (Lower West Coast to the Midwest and Upper East Coast to the Midwest), some airlines are very specialized and face no competition from any of the other five airlines on certain routes. In particular, NW controls the routes connecting Seattle to the Midwest and EA controls much of the connections from Florida to New York and surroundings.

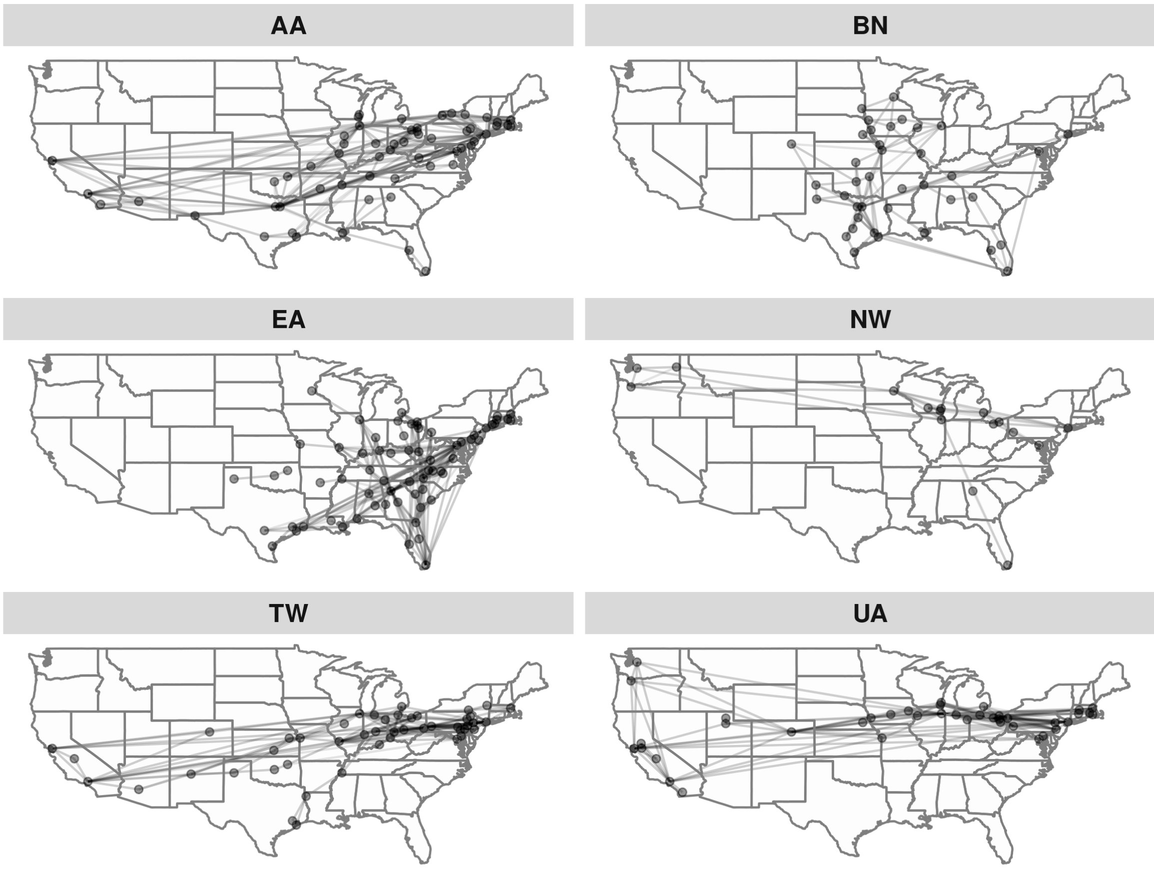



Figure 31: Flight Network in 1956 by Airline (weighted by log frequency).

# B. Appendix: Patent data

In this appendix we describe facts that we observe in the US patent data, for patents filed[100] between 1945 and 1975. US patents data containing citations and filing year

[100]Filing year, also called application year, is the closest date to the date of invention that is present in the data and it represent the date of the first administrative event in order to obtain a patent. In the other hand, publishing or also called granting year, is the later year in which the patent is granted. The difference between filing and granting year depends on diverse non-innovation related factors (as

have been downloaded from Google Patents. Then, it was merged with multiple datasets (see Appendix Patent Data Construction for more details):

- Technology classification: NBER patent database.
- Geographic location of inventors: Histpat and Histpat International for patents published until 1975, Fung Institute for patents published after 1975. Both matched to 1950s Metropolitan Statistical Areas (MSAs).
- Ownership: Kogan et al. (2017) for patents owned by firms listed in the US stock market, Patstat for the remaining patents not matched to Kogan et al. (2017).

We highlight two details from the matching process: 1. During filing years 1971-1972 the rate of non-geocoded patents increases, possibly due to Histpat and Fung data not being a perfect continuation one of the other. 2. Kogan et al. (2017) seems to use a matching method based on the patent owner declared in the patent text, as Patstat does. Specially, Kogan et al. (2017) does not explicitly say if it takes into account firm-ownership structure to determine patent ownership, neither does Patstat.

For the analysis presented in this appendix we will use the resulting dataset from the matching procedure, where unless evident or noticed, we will use only patents that have inventors within MSAs. We discard patents that have inventors in multiple MSAs and patents that belong to government organizations or universities. We assign patents to technology categories using fractional count: if a patent is listed in two technology categories, then we assign half a patent to each category. We discard self citations (citations in which the citing patent owner is the same as the cited patent owner) because self-citations may be due to different incentives.

## B.1. Matching patents to locations

In figure 32 we observe that the matching rate decreases from around 95% before 1970, to around 80% in 1971 and 1972, and then it stabilizes around 99% after 1975.

capacity of the patent office to revise applications) and changes over time. Hence filing year is the date in our data that approximates the best to the date of invention.

Hence, geogprahical results during years 1970-1975 will contain an increased amount of measurement error.

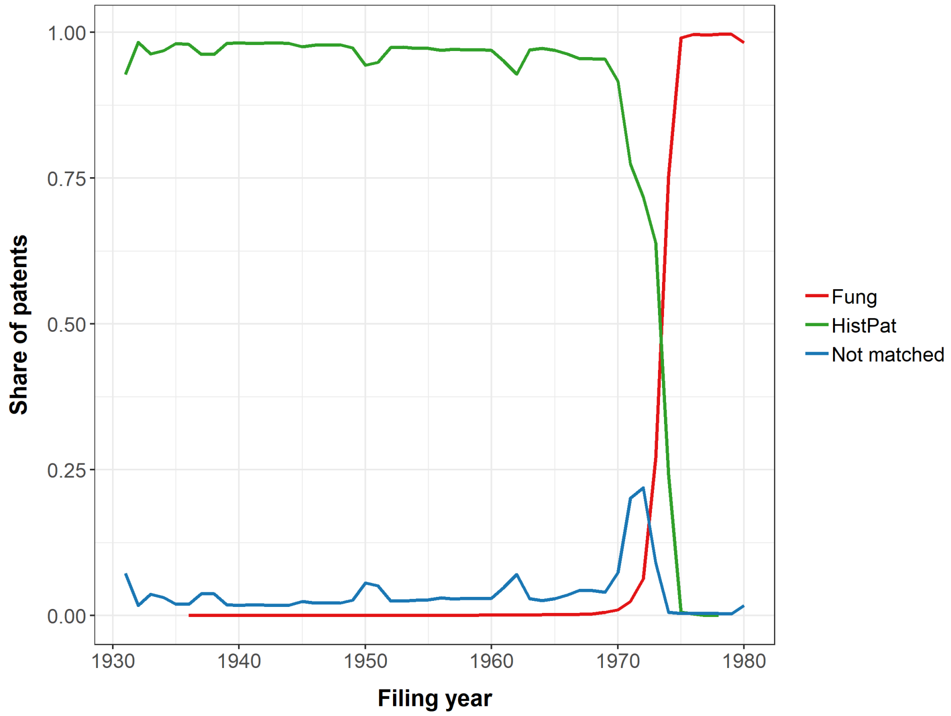



Figure 32: Non-matching rate HistPat, HistPat International and Fung

Figure 33 shows the share of patents that have inventors inside MSAs, and figure 34 displays the same by technology category.[101]

[101] Technologies are aggregated to six big groups, as explained in HJT 2002

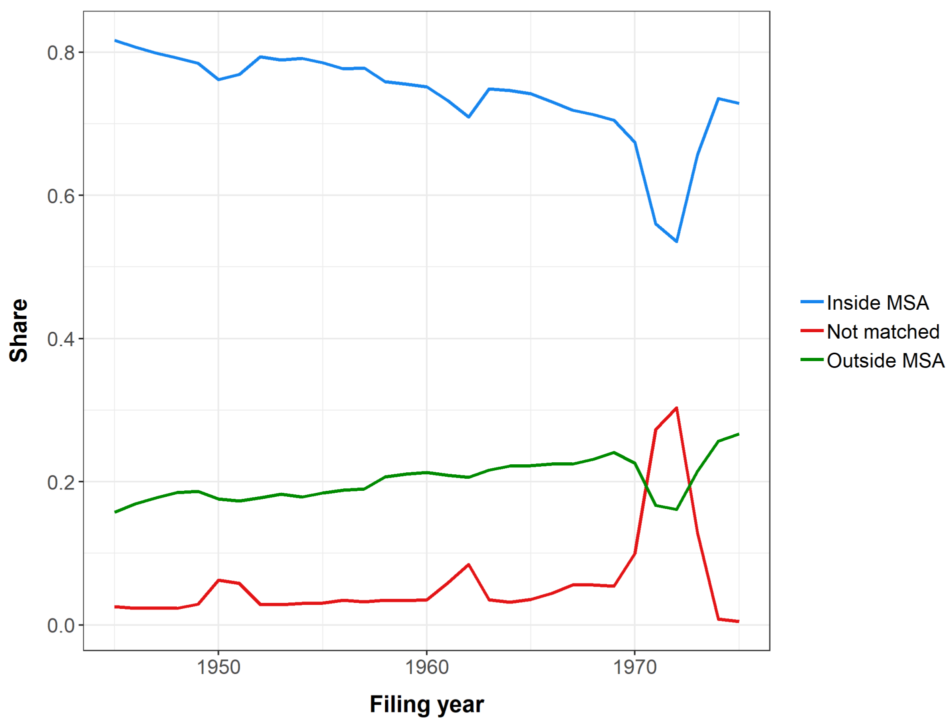



Figure 33: Share patents in Metropolitan Statistical Areas

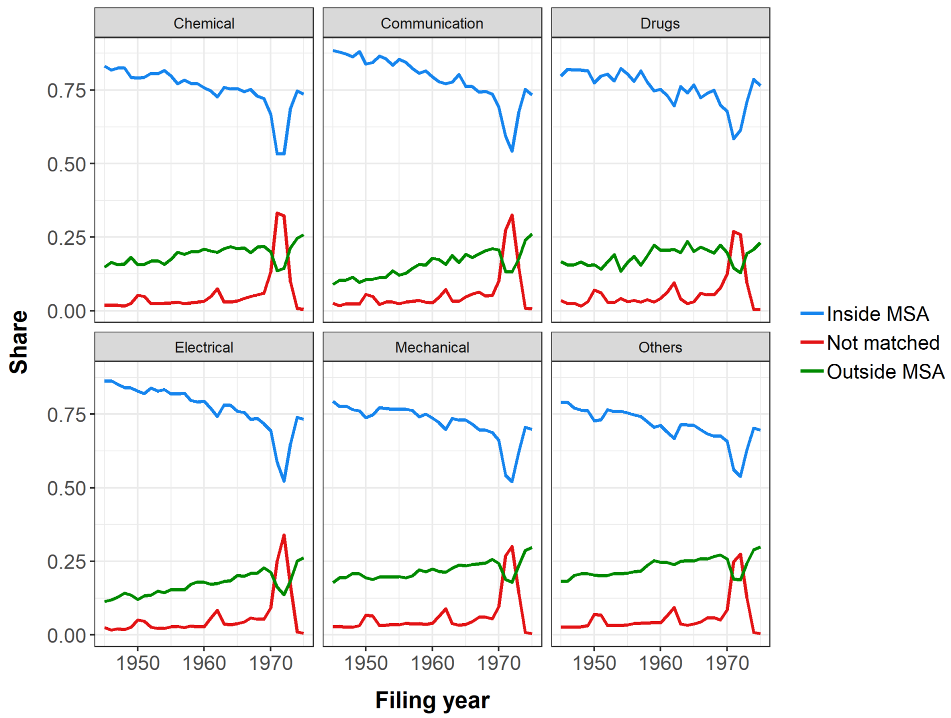



Figure 34: Share patents in Metropolitan Statistical Areas

## B.2. Input-Output of patents

In the same spirit as how Input-Output tables of industries are constructed, we can use citations as a reflection of sourced (input) knowledge. In this case, we interpret the cited patent as being a source of knowledge, and the citing patent as being a destination. In Figure 35 we aggregate citations by citing-cited technology category in the years 1949-1953. Rows represent the source technology and columns the destination technology. Columns should sum to 1 (round errors may exist). We highlight in bold those IO coefficients that are higher than 0.1. We observe that the diagonal has coefficients greater than 0.5, implying that technologies rely on themselves to create new knowledge. At the same time, we observe the importance of Electrical to create Communication technologies, and the small relevance of Drugs for every other technology.

| Source/Destination | Chemical | Communication | Drugs | Electrical | Mechanical | Others |
|---|---|---|---|---|---|---|
| Chemical | **0.74** | 0.01 | **0.13** | 0.03 | 0.05 | 0.05 |
| Communication | 0 | **0.6** | 0 | 0.07 | 0.01 | 0.01 |
| Drugs | 0.01 | 0 | **0.6** | 0 | 0 | 0.01 |
| Electrical | 0.03 | **0.28** | 0.03 | **0.7** | 0.05 | 0.04 |
| Mechanical | **0.11** | 0.07 | 0.07 | 0.1 | **0.72** | **0.15** |
| Others | **0.11** | 0.05 | **0.16** | 0.09 | **0.16** | **0.75** |
| Total | 1 | 1 | 1 | 1 | 1 | 1 |

Figure 35: Input-Output of technologies 1949-1953

## B.3. General Electric research establishments

Using the patent owner identifier we can display the geographical distribution of research establishments for a selected firm. Figure 36 shows the research establishments of General Electric in the period 1945-1953. We say that a firm $F$ had a research establishment in location $i$ in time period $t$ if firm $F$ filed at least one patent in time period $t$ with inventors located in location $i$. The headquarters location $q$ of firm $F$ is defined as the location in which the firm filed the largest amount of patents in the period 1945-1953. General Electric had research establishments in 62 MSAs in the period 1945-1953, and the MSA with the largest amount of patents was Schenectady, New York. Figure 37 shows the location of patents cited by patents filed by General Electric with inventors in Fort Wayne, Indiana, in the period 1949-1953. Figure 38 shows the research establishments of General Electric during periods 1949-1953 and 1964-1968. General Electric had research establishments in 51 MSAs in 1949-1953 and in 76 MSAs in 1964-1968. 42 out of them appear in both time periods.

## Table 11: Connected MSAs

| MSA fips | MSA name | <=3 periods | 4 periods | MSA fips | MSA name | <=3 periods | 4 periods |
|---|---|---|---|---|---|---|---|
| 80 | Akron, OH SMA | X | X | 4680 | Macon, GA SMA | X | X |
| 160 | Albany-Schenectady-Troy, NY SMA | X | X | 4720 | Madison, WI SMA | X | X |
| 200 | Albuquerque, NM SMA | X | X | 4760 | Manchester, NH SMA | | |
| 240 | Allentown-Bethlehem-Easton, PA-NJ SMA | X | X | 4920 | Memphis, TN SMA | X | X |
| 280 | Altoona, PA SMA | | | 5000 | Miami, FL SMA | X | X |
| 320 | Amarillo, TX SMA | X | X | 5080 | Milwaukee, WI SMA | X | X |
| 480 | Asheville, NC SMA | X | | 5120 | Minneapolis-St. Paul, MN SMA | X | X |
| 520 | Atlanta, GA SMA | X | X | 5160 | Mobile, AL SMA | X | X |
| 560 | Atlantic City, NJ SMA | X | | 5240 | Montgomery, AL SMA | X | X |
| 600 | Augusta, GA-SC SMA | X | X | 5280 | Muncie, IN SMA | | |
| 640 | Austin, TX SMA | X | X | 5360 | Nashville, TN SMA | X | X |
| 720 | Baltimore, MD SMA | X | X | 5400 | New Bedford, MA SMA | | |
| 760 | Baton Rouge, LA SMA | X | | 5440 | New Britain-Bristol, CT SMA | | |
| 800 | Bay City, MI SMA | X | | 5480 | New Haven, CT SMA | X | X |
| 840 | Beaumont-Port Arthur, TX SMA | X | | 5560 | New Orleans, LA SMA | X | X |
| 960 | Binghamton, NY SMA | X | | 5600 | New York-Northeastern NJ, NY-NJ SMA | X | X |
| 1000 | Birmingham, AL SMA | X | X | 5720 | Norfolk-Portsmouth, VA SMA | X | |
| 1120 | Boston, MA SMA | X | X | 5840 | Ogden, UT SMA | X | |
| 1160 | Bridgeport, CT SMA | X | X | 5880 | Oklahoma City, OK SMA | X | X |
| 1200 | Brockton, MA SMA | | | 5920 | Omaha, NE-IA SMA | X | X |
| 1280 | Buffalo, NY SMA | X | X | 5960 | Orlando, FL SMA | X | X |
| 1320 | Canton, OH SMA | X | X | 6120 | Peoria, IL SMA | X | |
| 1360 | Cedar Rapids, IA SMA | X | X | 6160 | Philadelphia, PA-NJ SMA | X | X |
| 1440 | Charleston, SC SMA | X | X | 6200 | Phoenix, AZ SMA | X | X |
| 1480 | Charleston, WV SMA | X | X | 6280 | Pittsburgh, PA SMA | X | X |
| 1520 | Charlotte, NC SMA | X | X | 6320 | Pittsfield, MA SMA | | |
| 1560 | Chattanooga, TN-GA SMA | X | X | 6400 | Portland, ME SMA | | |
| 1600 | Chicago, IL-IN SMA | X | X | 6440 | Portland, OR-WA SMA | X | X |
| 1640 | Cincinnati, OH-KY SMA | X | X | 6480 | Providence, RI SMA | X | X |
| 1680 | Cleveland, OH SMA | X | X | 6560 | Pueblo, CO SMA | X | |
| 1760 | Columbia, SC SMA | X | X | 6600 | Racine, WI SMA | X | X |
| 1800 | Columbus, GA-AL SMA | X | X | 6640 | Raleigh, NC SMA | X | X |
| 1840 | Columbus, OH SMA | X | X | 6680 | Reading, PA SMA | X | X |
| 1880 | Corpus Christi, TX SMA | X | X | 6760 | Richmond, VA SMA | X | X |
| 1920 | Dallas, TX SMA | X | X | 6800 | Roanoke, VA SMA | X | X |
| 1960 | Davenport-Rock Island-Moline, IA-IL SMA | X | X | 6840 | Rochester, NY SMA | X | X |
| 2000 | Dayton, OH SMA | X | X | 6880 | Rockford, IL SMA | | |
| 2040 | Decatur, IL SMA | | | 6920 | Sacramento, CA SMA | X | X |
| 2080 | Denver, CO SMA | X | X | 6960 | Saginaw, MI SMA | X | |
| 2120 | Des Moines, IA SMA | X | X | 7000 | St. Joseph, MO SMA | X | |
| 2160 | Detroit, MI SMA | X | X | 7040 | St. Louis, MO-IL SMA | X | X |
| 2240 | Duluth-Superior, MN-WI SMA | X | | 7160 | Salt Lake City, UT SMA | X | X |
| 2280 | Durham, NC SMA | X | X | 7200 | San Angelo, TX SMA | | |
| 2320 | El Paso, TX SMA | X | X | 7240 | San Antonio, TX SMA | X | X |
| 2360 | Erie, PA SMA | X | | 7280 | San Bernardino, CA SMA | | |
| 2440 | Evansville, IN SMA | X | X | 7320 | San Diego, CA SMA | X | X |
| 2480 | Fall River, MA-RI SMA | X | X | 7360 | San Francisco-Oakland, CA SMA | X | X |
| 2640 | Flint, MI SMA | X | | 7400 | San Jose, CA SMA | | |
| 2760 | Fort Wayne, IN SMA | X | X | 7520 | Savannah, GA SMA | X | |
| 2800 | Fort Worth, TX SMA | X | X | 7560 | Scranton, PA SMA | X | X |
| 2840 | Fresno, CA SMA | X | X | 7600 | Seattle, WA SMA | X | X |
| 2880 | Gadsden, AL SMA | | | 7680 | Shreveport, LA SMA | X | |
| 2920 | Galveston, TX SMA | X | X | 7720 | Sioux City, IA SMA | X | |
| 3000 | Grand Rapids, MI SMA | X | | 7760 | Sioux Falls, SD SMA | X | |
| 3080 | Green Bay, WI SMA | | | 7800 | South Bend, IN SMA | X | X |
| 3120 | Greensboro-High Point, NC SMA | X | X | 7840 | Spokane, WA SMA | X | X |
| 3160 | Greenville, SC SMA | X | X | 7880 | Springfield, IL SMA | X | |
| 3200 | Hamilton-Middletown, OH SMA | | | 7920 | Springfield, MO SMA | X | |
| 3240 | Harrisburg, PA SMA | X | X | 7960 | Springfield, OH SMA | | |
| 3280 | Hartford, CT SMA | X | X | 8000 | Springfield-Holyoke, MA-CT SMA | X | X |
| 3360 | Houston, TX SMA | X | X | 8040 | Stamford-Norwalk, CT SMA | X | |
| 3400 | Huntington-Ashland, WV-KY-OH SMA | X | | 8120 | Stockton, CA SMA | X | X |
| 3480 | Indianapolis, IN SMA | X | X | 8160 | Syracuse, NY SMA | X | X |
| 3520 | Jackson, MI SMA | X | | 8200 | Tacoma, WA SMA | | |
| 3560 | Jackson, MS SMA | | | 8280 | Tampa-St. Petersburg, FL SMA | X | X |
| 3600 | Jacksonville, FL SMA | X | X | 8320 | Terre Haute, IN SMA | X | X |
| 3680 | Johnstown, PA SMA | | | 8400 | Toledo, OH-MI SMA | X | X |
| 3720 | Kalamazoo, MI SMA | X | | 8440 | Topeka, KS SMA | X | |
| 3760 | Kansas City, MO-KS SMA | X | X | 8480 | Trenton, NJ SMA | | |
| 3800 | Kenosha, WI SMA | | | 8560 | Tulsa, OK SMA | X | X |
| 3840 | Knoxville, TN SMA | X | X | 8680 | Utica-Rome, NY SMA | | |
| 4000 | Lancaster, PA SMA | X | X | 8800 | Waco, TX SMA | X | |
| 4040 | Lansing, MI SMA | X | | 8840 | Washington, DC-MD-VA SMA | X | X |
| 4080 | Laredo, TX SMA | X | | 8880 | Waterbury, CT SMA | | |
| 4160 | Lawrence, MA SMA | | | 8920 | Waterloo, IA SMA | X | |
| 4280 | Lexington, KY SMA | X | X | 9000 | Wheeling-Steubenville, WV-OH SMA | X | |
| 4320 | Lima, OH SMA | | | 9040 | Wichita, KS SMA | X | X |
| 4360 | Lincoln, NE SMA | X | X | 9080 | Wichita Falls, TX SMA | X | X |
| 4400 | Little Rock-North Little Rock, AR SMA | X | X | 9120 | Wilkes-Barre–Hazleton, PA SMA | X | X |
| 4440 | Lorain-Elyria, OH SMA | X | X | 9160 | Wilmington, DE-NJ SMA | X | X |
| 4480 | Los Angeles, CA SMA | X | X | 9220 | Winston-Salem, NC | X | X |
| 4520 | Louisville, KY-IN SMA | X | X | 9240 | Worcester, MA SMA | X | |
| 4560 | Lowell, MA SMA | | | 9280 | York, PA SMA | X | X |
| 4600 | Lubbock, TX SMA | X | X | 9320 | Youngstown, OH-PA SMA | X | X |

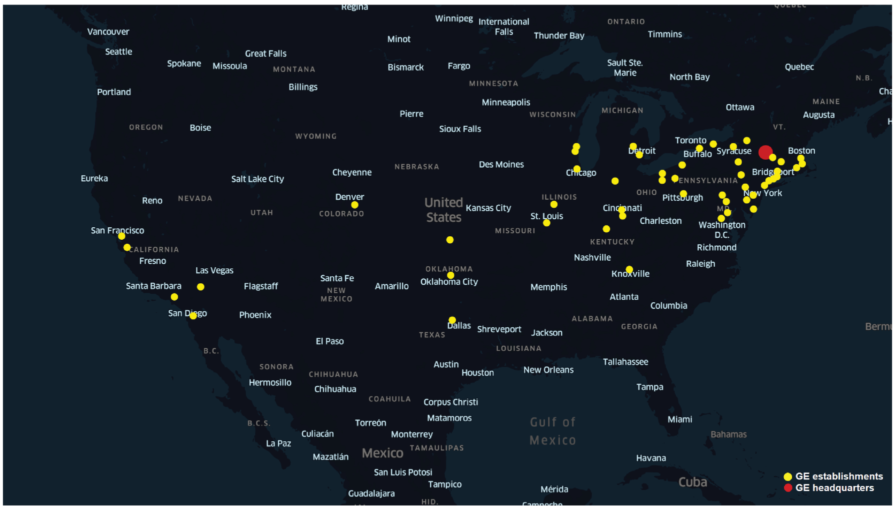



Figure 36: Research establishments of General Electric 1949-1953


Figure 37: Citations General Electric at Fort Wayne IN 1949-1953


Figure 38: Change location research establishments of General Electric between 1949-1953 and 1964-1968

## B.4. Descriptive statistics

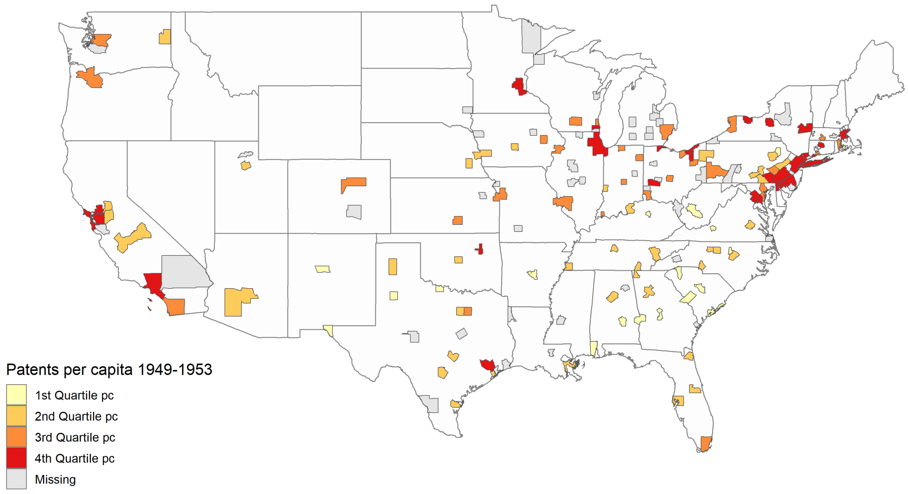



Figure 39: Patents per capita in 1951
Quantiles of patents per capita are computed in each technology and then averaged across technologies. Population is from 1950 Census.

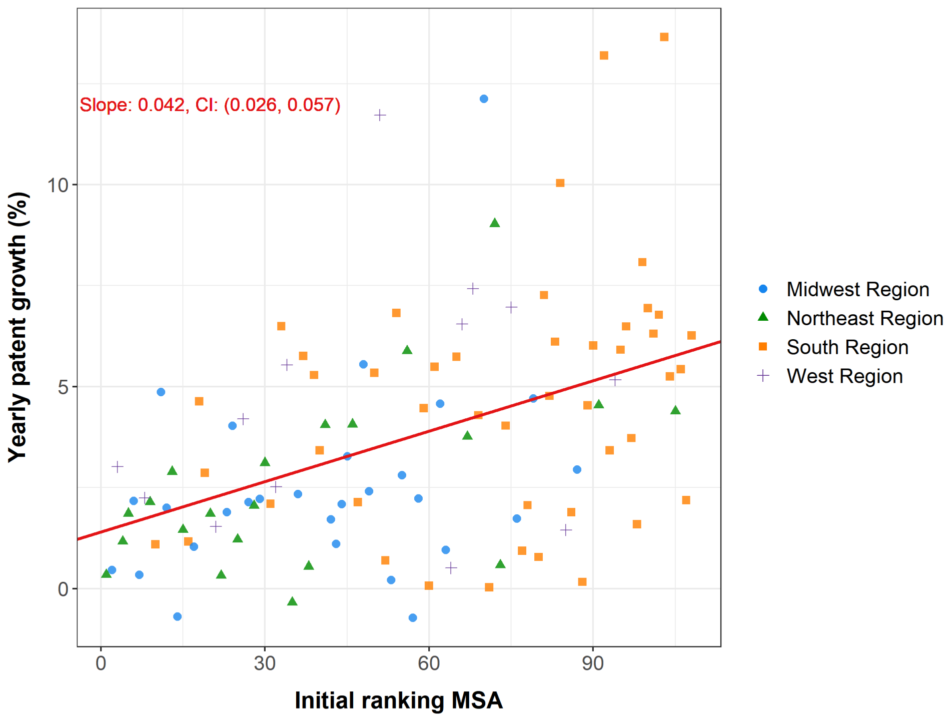



Figure 40: Patent growth by initial innovativeness ranking of MSA

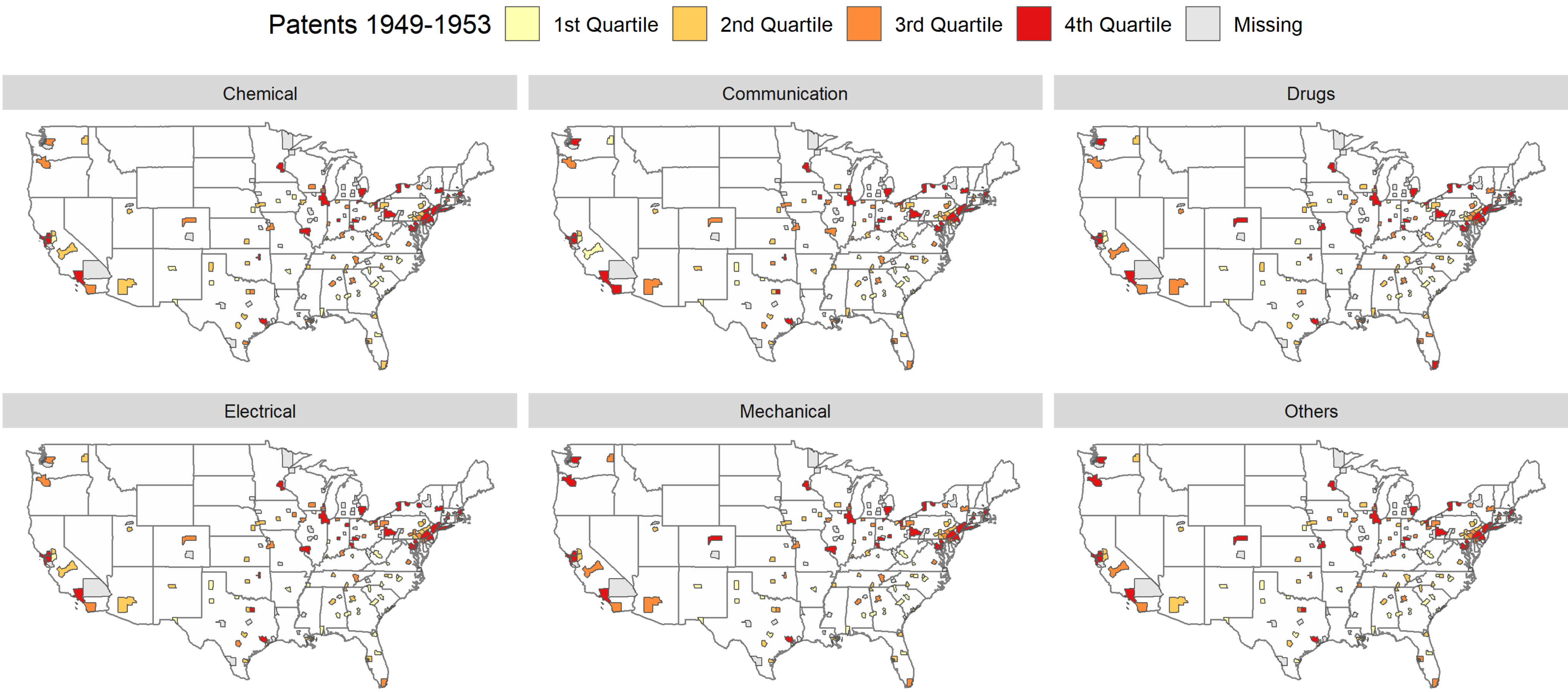



Figure 42: Geography of patenting 1951

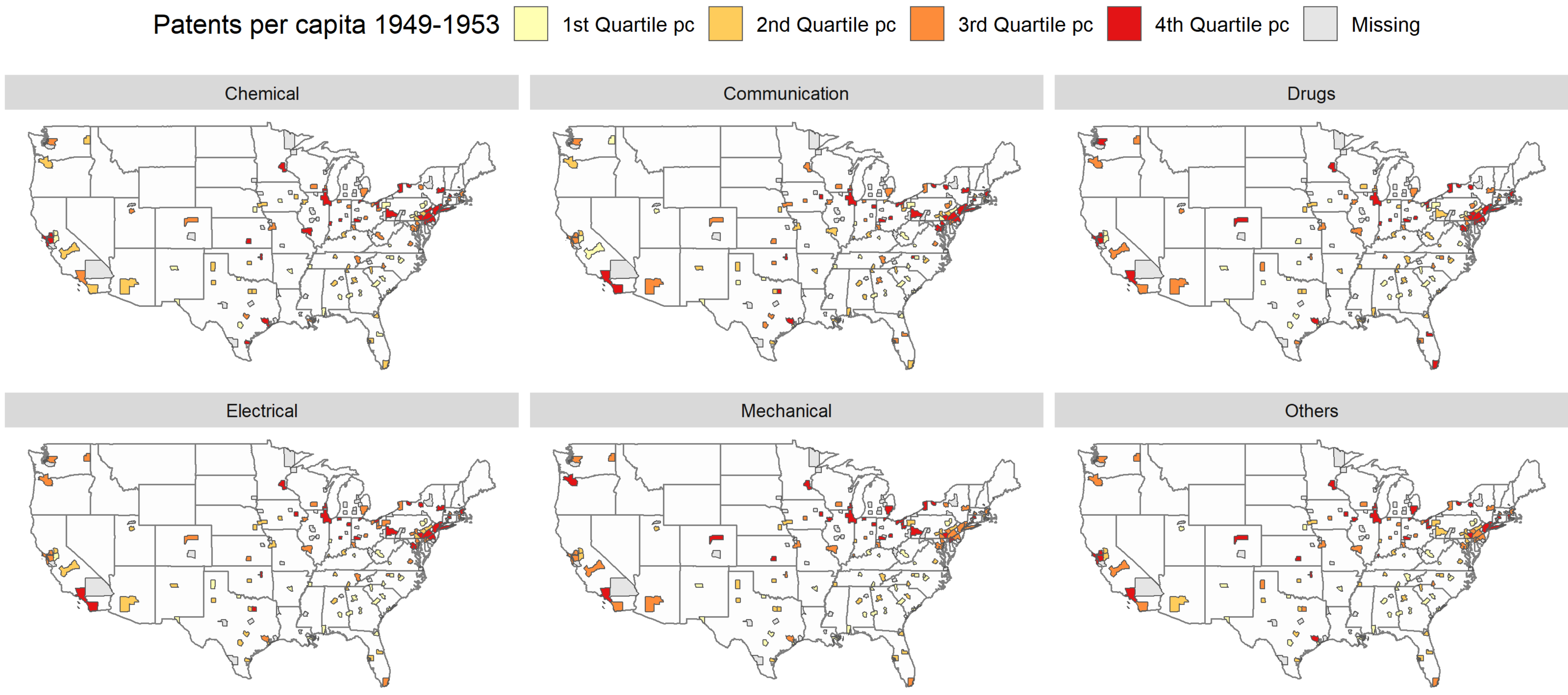



Figure 43: Patents per capita in 1951

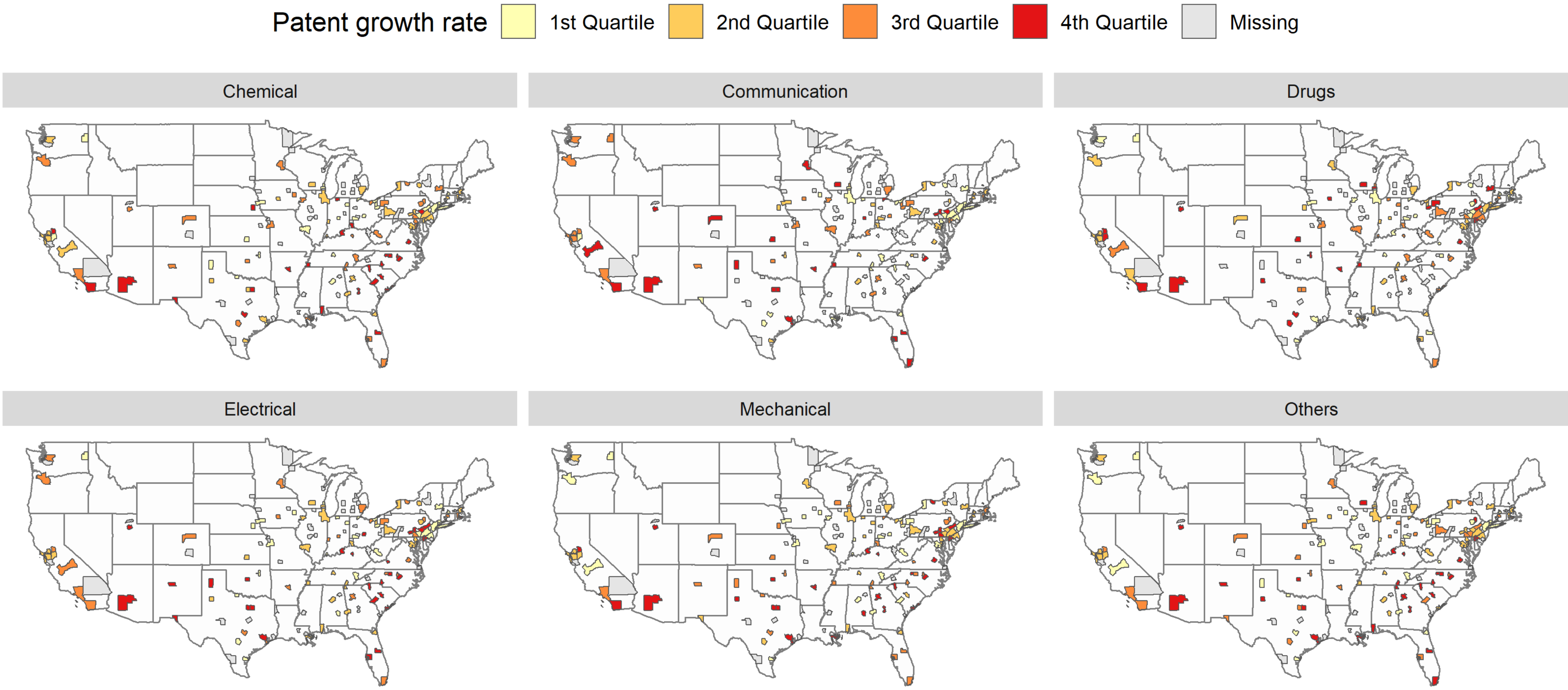



Figure 44: Patent growth rate 1951-1966

#### B.4.1. Descriptive statistics by technology

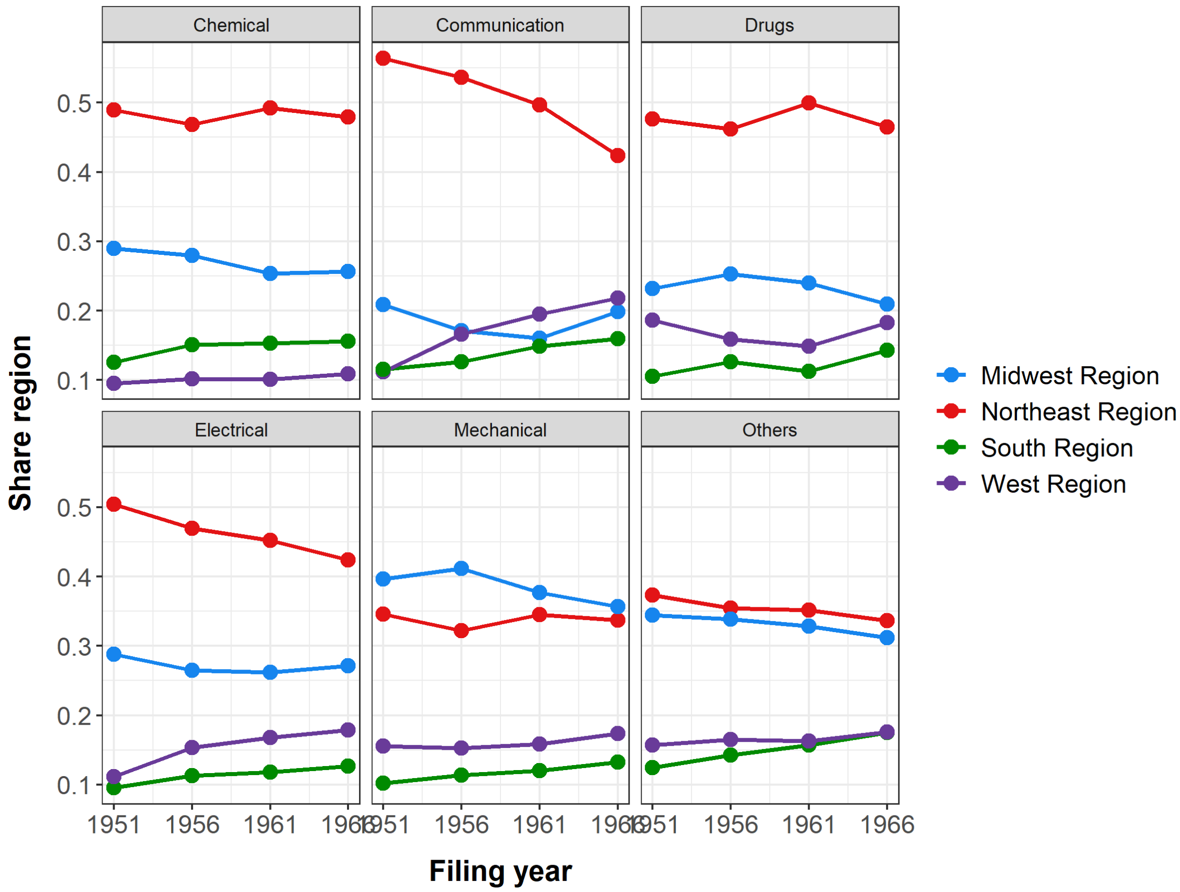



Figure 41: Share of patents by region

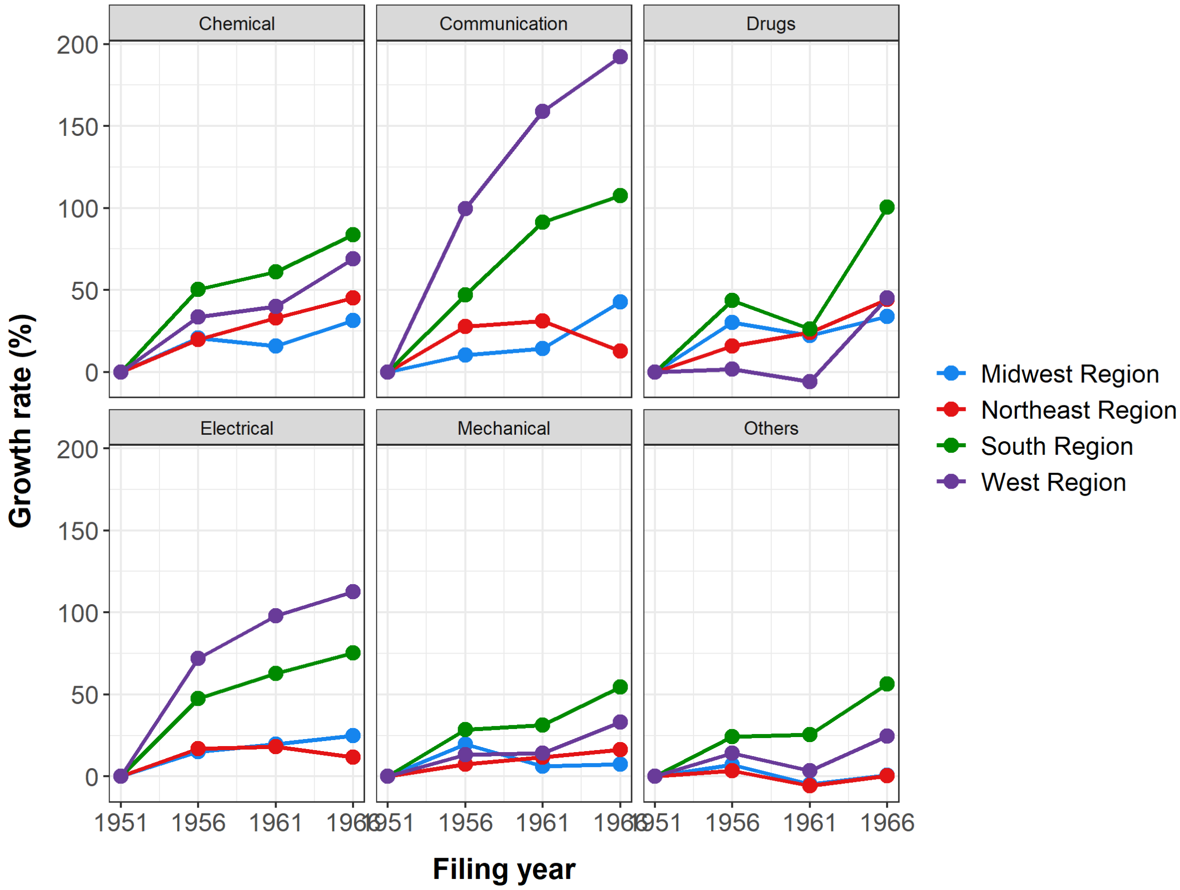



Figure 45: Patent growth rate by region

| Technology | N. estab. / Year | Number of firms: 1 | 2 to 5 | 6 to 10 | 11 to 20 | +20 | Share of patents: 1 | 2 to 5 | 6 to 10 | 11 to 20 | +20 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Chemical | 1951 | 6,773 | 892 | 72 | 34 | 8 | 0.39 | 0.25 | 0.13 | 0.18 | 0.06 |
| | 1956 | 7,196 | 953 | 108 | 60 | 12 | 0.38 | 0.22 | 0.11 | 0.19 | 0.10 |
| | 1961 | 6,728 | 1067 | 125 | 80 | 18 | 0.32 | 0.21 | 0.15 | 0.15 | 0.17 |
| | 1966 | 7,092 | 1120 | 125 | 89 | 30 | 0.29 | 0.17 | 0.13 | 0.19 | 0.22 |
| Communication | 1951 | 1,956 | 270 | 43 | 24 | 8 | 0.42 | 0.18 | 0.18 | 0.04 | 0.18 |
| | 1956 | 2,292 | 337 | 56 | 43 | 11 | 0.36 | 0.19 | 0.15 | 0.18 | 0.12 |
| | 1961 | 2,413 | 441 | 62 | 63 | 15 | 0.34 | 0.16 | 0.09 | 0.20 | 0.20 |
| | 1966 | 2,320 | 414 | 75 | 66 | 29 | 0.32 | 0.14 | 0.08 | 0.17 | 0.29 |
| Drugs | 1,951 | 1675 | 163 | 20 | 21 | 5 | 0.76 | 0.19 | 0.02 | 0.04 | 0.00 |
| | 1956 | 1,706 | 198 | 40 | 35 | 9 | 0.66 | 0.18 | 0.07 | 0.06 | 0.02 |
| | 1961 | 1,705 | 247 | 57 | 45 | 16 | 0.62 | 0.19 | 0.09 | 0.07 | 0.04 |
| | 1966 | 2,115 | 251 | 49 | 53 | 24 | 0.62 | 0.13 | 0.08 | 0.11 | 0.06 |
| Electrical | 1,951 | 7394 | 789 | 73 | 33 | 8 | 0.47 | 0.20 | 0.08 | 0.08 | 0.18 |
| | 1956 | 8,182 | 962 | 97 | 59 | 12 | 0.44 | 0.19 | 0.10 | 0.12 | 0.15 |
| | 1961 | 8,077 | 1,092 | 123 | 80 | 18 | 0.40 | 0.19 | 0.07 | 0.17 | 0.17 |
| | 1966 | 7,885 | 1,006 | 126 | 87 | 30 | 0.37 | 0.16 | 0.09 | 0.16 | 0.23 |
| Mechanical | 1951 | 18,509 | 1,348 | 75 | 34 | 8 | 0.64 | 0.20 | 0.06 | 0.05 | 0.04 |
| | 1956 | 18,735 | 1,498 | 109 | 60 | 12 | 0.59 | 0.20 | 0.06 | 0.08 | 0.06 |
| | 1961 | 16,873 | 1,703 | 130 | 80 | 18 | 0.54 | 0.21 | 0.07 | 0.10 | 0.08 |
| | 1966 | 17,856 | 1,669 | 132 | 89 | 30 | 0.52 | 0.17 | 0.09 | 0.11 | 0.11 |
| Others | 1951 | 24,994 | 1,343 | 75 | 34 | 8 | 0.76 | 0.15 | 0.04 | 0.03 | 0.02 |
| | 1956 | 24,650 | 1,527 | 110 | 60 | 12 | 0.71 | 0.16 | 0.05 | 0.05 | 0.03 |
| | 1961 | 20,914 | 1,683 | 131 | 80 | 18 | 0.65 | 0.16 | 0.06 | 0.08 | 0.05 |
| | 1966 | 22,982 | 1,625 | 132 | 89 | 30 | 0.63 | 0.15 | 0.05 | 0.08 | 0.08 |

Table 15: Number of firms and share of patents by firm's geographic coverage

Geographic coverage is computed as the amount of Metropolitan Statistical Areas (MSAs) in which the firm has inventors applying for patents (*research establishments*) in a certain year. Research establishments are defined irrespective of the technology in which technology it patents. Hence a firm applying for patents in technology $h$ in one establishment and technology $k$ in another establishment is defined as having two establishments, and counts as a two-establishment firm both in technology $h$ and technology $k$. Bins of geographic coverage are 1 MSA, 2 to 5 MSAs, 6 to 10 MSAs, 11 to 20 MSAs, more than 20 MSAs. The maximum possible is 108 MSAs.

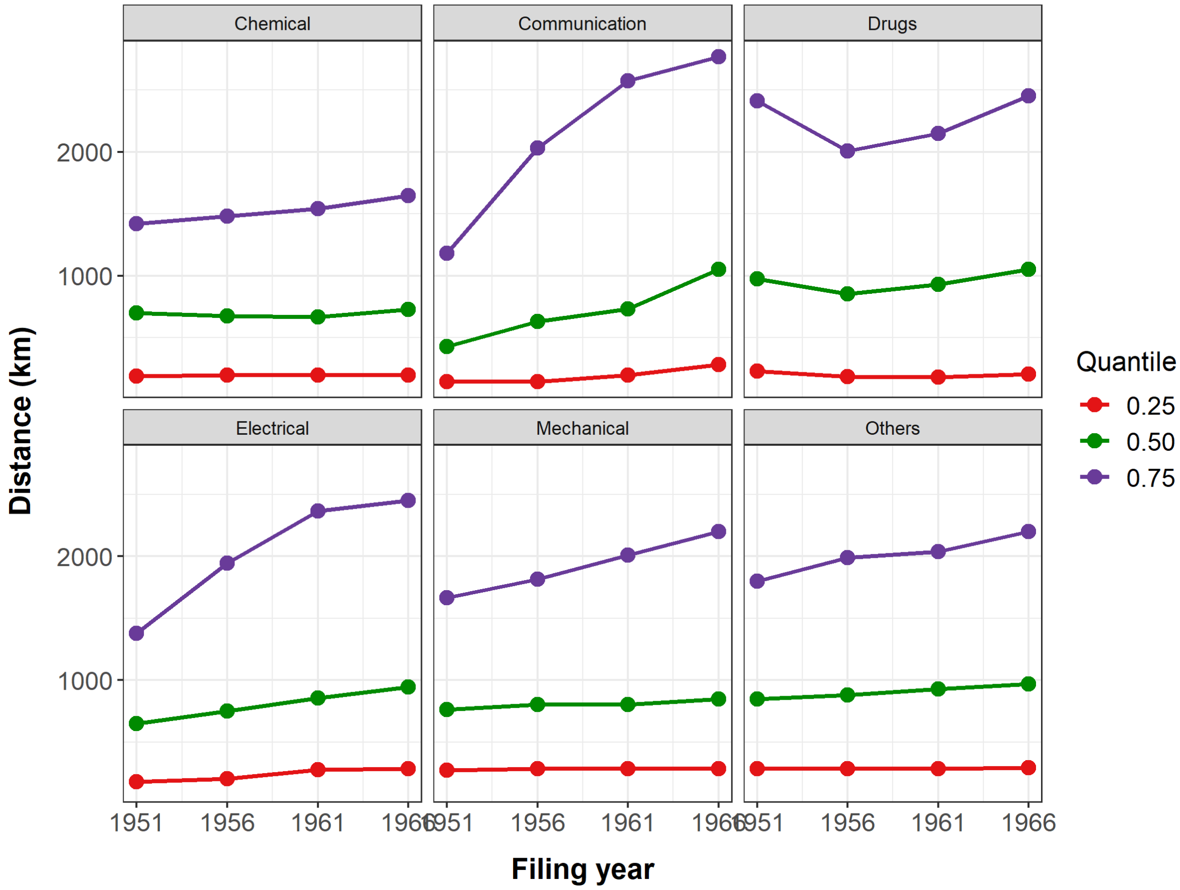



Figure 46: Quantiles of citation distance

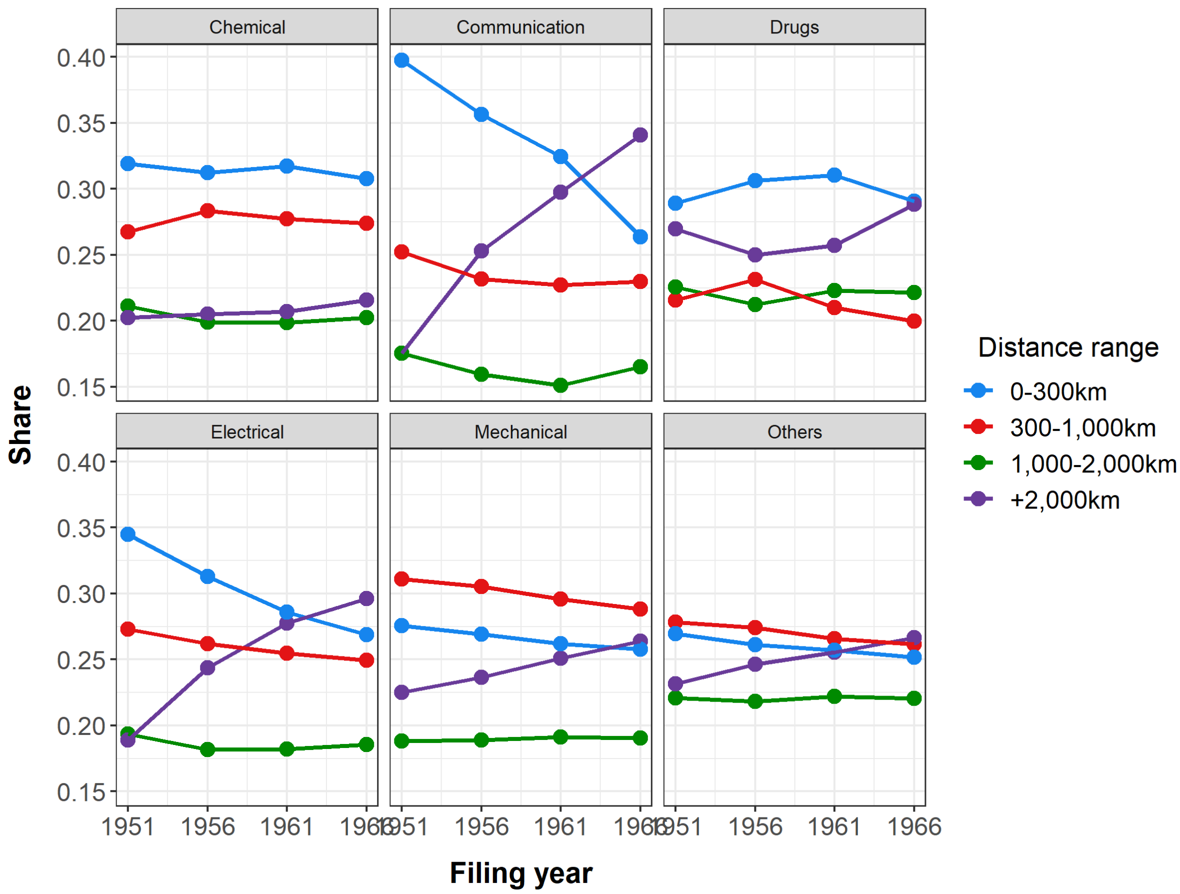



Figure 47: Share of citations by distance

# C. Appendix: US Census Regions

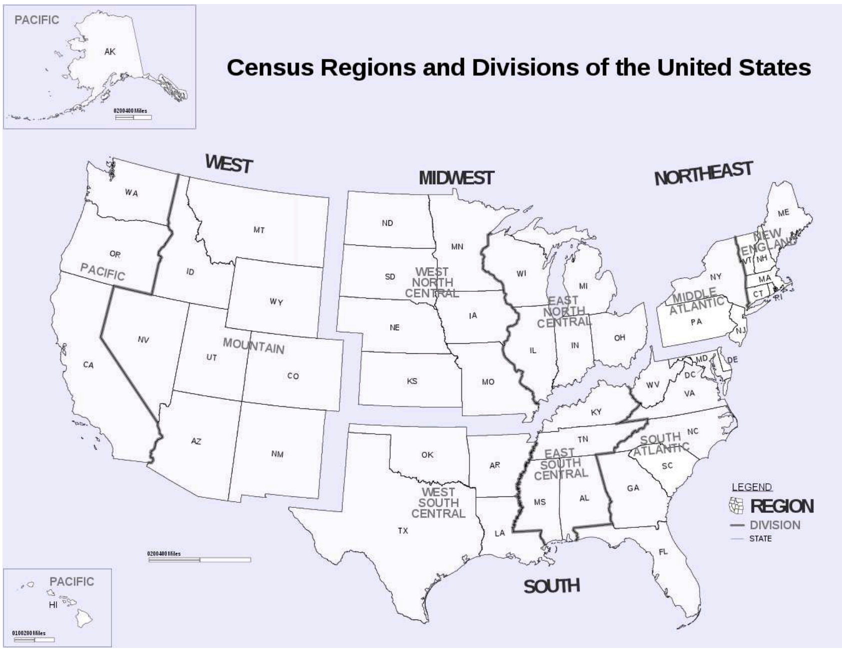



Figure 48: US Census Regions
Source: US Census Bureau

# D. Appendix: Bias Correction and IV estimation

## D.1. Split-panel jackknife bias correction

Weidner and Zylkin (2021) show that PPML estimation of gravity equations with three-way fixed effects (origin-time, destination-time, origin-destination) is consistent but asymptotically biased. In their words: "*the asymptotic distribution of the estimates is not centered at the truth as* $N \rightarrow \infty$" (page 2). The asymptotic bias concerns both point estimates and standard errors. In order to correct the bias we apply their suggested split-panel jackknife bias correction of section 3.4.1 to both point estimates and bootstrap standard errors. The idea of the jackknife bias correction is to estimate the model in many subsamples and then subtract the average coefficients of the subsamples from (twice) the original coefficient.

As suggested in Weidner and Zylkin (2021) when using real world data (as opposite to simulated data), we estimate the bias correction repeatedly. We modify equation (14) in Weidner and Zylkin (2021) to define the bias corrected coefficient as:

$$\tilde{\beta}_N^J := 2 \times \hat{\beta} - \frac{1}{Z} \sum_z \sum_p \frac{\hat{\beta}_{(p,z)}}{4} \tag{12}$$

where $p$ is a random subsample of size 1/4th of the original sample, and $Z$ is the amount of times to subsample.

The *procedure to estimate bias corrected point estimate* $\tilde{\beta}_N^J$ is as follows:

1. Estimate $\hat{\beta}$: the not-bias-corrected estimate of equation (3)
2. Randomly allocate all citing establishment-technology $Fih$ into two equally sized groups (groups are time-invariant). Call them citing groups $a$ and $b$.
3. Randomly allocate all cited establishment-technology $Gjk$ into two equally sized groups (groups are time-invariant). Call them cited groups $a$ and $b$.
4. Create four $p$ subsamples of the original data: (a,a), (a,b), (b,a), (b,b). Subsamples keep the same granularity as the original data $FiGjhkt$.
5. Estimate equation (3) (gravity equation of the main text) in each of the subsamples from the previous step to obtain $\hat{\beta}_{(p,z)}$.[102] Store the four estimated coefficients.
6. Repeat $Z$ times steps 2 to 5.
7. Compute equation 12

To compute bias-corrected bootstrap standard errors we need to bias-correct the point estimate $\tilde{\beta}_m^J$ of each bootstrap iteration $m$. The *procedure to estimate bias corrected standard errors* is as follows:

[102]Given that we require to identify the fixed effects, the *effective subsample* in all four $p$ estimations does not have the same amount of observations. However, in our estimations the *effective subsample* size across $p$ subsamples does not differ by more than 5%.

1. Sample establishment-technology-pairs $FiGjhk$ with replacement such that we obtain a re-sampled data of the same size as the original data (hence, some $FiGjhk$ will be repeated in the re-sampled data). Sampled $FiGjhk$ are kept for all time periods in order to keep the source of identification of $\beta$: across time variation within a establishment pair. Label this new dataset $data_m$.

2. Using $data_m$, estimate equation (3) to obtain $\hat{\beta_m}$ (this is a point estimate of the specific $data_m$)

3. Using $data_m$, repeat $Z_M$ times steps 2 to 5 of the *procedure to estimate bias corrected point estimate*. This step provides $Z_M \times 4$ point estimates $\hat{\beta}_{(p,m,z_M)}$

4. Compute the bias corrected point estimate of bootstrap $m$ $\tilde{\beta}^J_m = 2 \times \hat{\beta}_m - \frac{1}{Z_M} \sum_{z_M} \sum_p \frac{\hat{\beta}_{(p,m,z_M)}}{4}$.

5. Store the bias corrected point estimate of bootstrap $m$

6. Repeat steps 1 to 5 $M$ times to obtain $M$ bias corrected bootstrap point estimates $\tilde{\beta}^J_m$

7. Compute the variance-covariance matrix of bias corrected bootstrap coefficients $\tilde{\beta}^J_m$ and use it to compute standard errors of $\tilde{\beta}^J_N$

The bias correction of point estimates and bias correction of bootstrap standard errors implies estimating $Z \times 4 + Z_M \times M \times 4$ models. This is a computationally demanding task. To estimate columns (1) and (2) of Table 2 we set $Z = 100$, $Z_M = 5$ and $M = 200$, adding up to $1,100$ models to estimate for each column.

As recommended in Hansen (2021), in the Table 16 we repeat Table 2 but reporting 0.025 and 0.975 quantile values of bootstrap estimates (bias corrected for columns (1) and (2)) instead of standard errors:

| | PPML | | IV PPML | |
|---|---|---|---|---|
| Dep. variable: *citations* | $cit_{FiGjhkt}$ | | $cit_{FiGjhkt}$ | |
| | (1) | (2) | (3) | (4) |
| log(travel time) | $-0.083$ <br> $(-0.129; -0.056)$ | | $-0.152$ <br> $(-0.210; -0.097)$ | |
| log(travel time) × 0-300km | | 0.019 <br> $(-0.054; 0.082)$ | | $-0.076$ <br> $(-0.542; 0.384)$ |
| log(travel time) × 300-1,000km | | $-0.089$ <br> $(-0.141; -0.052)$ | | $-0.134$ <br> $(-0.246; -0.066)$ |
| log(travel time) × 1,000-2,000km | | $-0.094$ <br> $(-0.156; -0.022)$ | | $-0.112$ <br> $(-0.192; -0.022)$ |
| log(travel time) × +2,000km | | $-0.169$ <br> $(-0.277; -0.105)$ | | $-0.203$ <br> $(-0.311; -0.136)$ |
| N obs. effective | 4,703,010 | 4,703,010 | 4,703,010 | 4,703,010 |
| R2 | 0.88 | 0.88 | 0.88 | 0.88 |

Table 16: Elasticity of citations to travel time

Column (1) shows the result of Poisson Pseudo Maximum Likelihood (PPML) estimation of $citations_{FiGjhkt} = \exp\left[\boldsymbol{\beta}\, log(\text{travel time}_{ijt}) + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt}\right] \times \varepsilon_{FiGjhkt}$, for citations of patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$, to patents filed by establishment of firm $G$ in location $j$ and technology $k$. travel time$_{ijt}$ is the travel time in minutes between location $i$ and $j$ at time period $t$, and it is set to 1 when $i = j$. When $FiGjhk$ has positive citations in at least one period and no citations in another, we attribute zero citations in the missing period. Column (2) includes the interaction of travel time$_{ijt}$ with a dummy for distance bin between the citing establishment $Fi$ and the cited establishment $Gj$. Column (3) and (4) show the result of two step instrumental variables estimation, where $log(\text{travel time}_{ijt})$ is instrumented with $log(\text{travel time}_{ijt}^{\text{fix routes}})$, the travel time that would have taken place if routes were fixed to the ones observed in 1951 and in each year routes were operated with the average airplane of the year. 0.025 and 0.975 quantile bootstrap estimates are presented in parentheses. The coefficients and bootstrap estimates in columns (1) and (2) are jackknife bias-corrected. R2 is computed as the squared correlation between observed and fitted values.

## D.2. Instrumental variables PPML

To implement the instrumental variables of Poisson estimation we follow the control function approach described in Wooldridge (2014). We explain the procedure using the estimation of the elasticity of citations to travel time. The procedure is similar for the elasticity of (new) patents to knowledge access. We proceed in two steps estimating the following two equations:

$$\begin{aligned} log(\text{travel time})_{FiGjhkt} =& \lambda_2 \, log(\text{travel time}^{\text{fix routes}}_{FiGjhkt}) \\ &+ FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt} + u_{FiGjhkt} \end{aligned} \tag{13}$$

$$\begin{aligned} citations_{FiGjhkt} =& \exp\left[\beta \, log(\text{travel time}_{ijt}) + \lambda \, \hat{u}_{FiGjhkt} \right. \\ &\left. + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt}\right] \times v_{FiGjhkt} \end{aligned} \tag{14}$$

In a first step we estimate equation (13) and obtain estimated residuals $\hat{u}_{FiGjhkt}$. In a second step we use the estimated residuals as a regressor in equation (14) which *controls* for the endogenous component of travel time.

To perform inference we bootstrap standard errors in the following way:

1. Sample establishment-technology-pairs $FiGjhk$ with replacement such that we obtain a re-sampled data of the same size as the original data (hence, some $FiGjhk$ will be repeated in the re-sampled data). Sampled $FiGjhk$ are kept for all time periods in order to keep the source of identification of $\beta$: across time variation within a establishment pair. Label this new dataset $data_m$

2. Using $data_m$, estimate equations (13) and (14) to obtain the bootstrap estimate $\hat{\beta}_m$. Store $\hat{\beta}_m$.

3. Repeat $M$ times steps 1 and 2.

4. Compute the variance-covariance matrix of $\hat{\beta}_m$ and use it to compute standard errors of $\hat{\beta}$

For columns (3) and (4) of Table 2, and columns (3) and (4) of Table 4 we set $M = 200$.

# E. Appendix: Additional results

## E.1. Diffusion of knowledge

### E.1.1. Heterogeneous effects

First, we perform an intensive margin/extensive margin decomposition of the effect of travel time on citations. We find that the effect is coming from both margins. In the instrumental variables approach, the intensive margin is only statistically different from zero for distance greater than 2,000km, while for the extensive margin it is for distance greater than 300km. Results for the baseline analysis are shown in Table 17 and for the IV estimation in Table 18.

Second, we investigate if the elasticity varies by the degree of concentration of patents across establishments in the citing technology or cited technology, we find no statistically significant heterogeneous effect. Results are shown in columns (1) and (2) of Table 20.

Third, we check if the elasticity varies by the median forward and backward citation lags of the cited and citing technologies. We find that the elasticity of citations to travel time is *more negative* both for technologies that accumulate citations during a longer time period and for technologies that cite older patents. To be able to precisely show if it is *newer* or *older* technologies that diffuse better as consequence of the jet requires an analysis with the citation level forward and backward lag, and not using the median lag in the technology. Nonetheless, the results seem to suggest that jets improved the diffusion of *older* technologies. Results are shown in columns (3) and (4) of Table 20.

Fourth, we extend the sample of patents to include patents with a patent owner identified as a government organization or university. Column (5) of Table 20 opens the elasticity of citations to travel time by whether the citing patent belongs to a government organization of university. Column (6) includes a dummy for whether the cited patent belongs to a government organization or university. We do not observe a particular change in the pattern of the elasticity of citations to travel time.

Sixth, we extend the sample to include self citations (citations in which the citing and cited patents belong to the same patent owner $F$). Column (7) of Table 20 shows that the elasticity is not statistically different for self citations.

Seventh, we check if the elasticity varies with the level of innovativeness of the citing firm. It may be the case that those firms that actually have the -time and monetary- budget to take a plane are only the most innovative ones. We rank firms $F$ in technology $h$ according to the amount of patents filed by $F$ in technology $h$ at the initial time period 1949-1953. We define quantile 0.00 as all those firms that did not file patents in 1949-1953, while quantile 0.01 is assigned to those that filed patents but not as many as to be in the quantile 0.25 or higher. Results are shown in Table 19. We do not find a particular pattern related to the initial innovativeness.

Eighth, we check if the elasticity varies with the citing technology, cited technology and citing-cited technology pair. Results are shown in Table 21 and Table 24. We find that the elasticity is negative and significant mainly when the citing and cited technology are the same. In Appendix B we show that most citations happen within a technology, so most identification power would be when citing and cited technologies are the same.

| | PPML | | log-log | | linear probability | |
|---|---|---|---|---|---|---|
| Dep. variable: *citations* | $cit_{FiGjhkt}$ | | $log(cit_{FiGjhkt})$ | | $cit_{FiGjhkt} > 0$ | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| log(travel time) | $-0.083^{***}$ (0.019) | | $-0.071$ (0.098) | | -0.013$^{***}$ (0.003) | |
| log(travel time):0-300km | | 0.019 (0.036) | | $0.318^{**}$ (0.152) | | $-0.0045$ (0.005) |
| log(travel time):300-1000km | | $-0.089^{***}$ (0.023) | | $-0.265^{*}$ (0.145) | | $-0.008^{***}$ (0.003) |
| log(travel time):1000-2000km | | $-0.094^{***}$ (0.032) | | $-0.231$ (0.209) | | $-0.013^{***}$ (0.003) |
| log(travel time):+2000km | | $-0.169^{***}$ (0.039) | | $-0.424^{**}$ (0.192) | | $-0.024^{***}$ (0.005) |
| N obs. effective | 4,703,010 | 4,703,010 | 16,412 | 16,412 | 10,106,940 | 10,106,940 |
| R2 | 0.88 | 0.88 | 0.86 | 0.86 | 0.70 | 0.70 |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 17: Elasticity of citations to travel time: intensive and extensive margin

Column (1) shows the result of Poisson Pseudo Maximum Likelihood (PPML) estimation of $citations_{FiGjhkt} = \exp\left[\beta\, log(\text{travel time}_{ijt}) + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt}\right] \times \varepsilon_{FiGjhkt}$, for citations of patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$, to patents filed by establishment of firm $G$ in location $j$ and technology $k$. travel time$_{ijt}$ is the travel time in minutes between location $i$ and $j$ at time period $t$, and it is set to 1 when $i = j$. When $FiGjhk$ has positive citations in at least one period and no citations in another, we attribute zero citations in the missing period. Column (3) shows the result of an OLS estimation of $log(citations_{FiGjhkt}) = \alpha\, log(\text{travel time}_{ijt}) + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt} + \varepsilon_{FiGjhkt}$, with a sample of establishment-technology pairs ($FiGjhk$) that have positive citations in all periods. Column (5) shows the result of an OLS estimation of $\mathbb{1}\{citations_{FiGjhkt} > 0\} = \gamma\, log(\text{travel time}_{ijt}) + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt} + \varepsilon_{FiGjhkt}$, with the same sample as (1). Column (2), (4) and (6) open, respectively, the coefficients $\beta$, $\alpha$, $\gamma$ by distance between the citing establishment $Fi$ and the cited establishment $Gj$. Standard errors are presented in parentheses. Columns (1) and (2) present coefficients and bootstrap standard errors jackknife bias corrected. Columns (3) through (6) present standard errors clustered at the non-directional location pair ($ij$ is the same non-directional location pair as $ji$). R2 is computed as the squared correlation between observed and fitted values.

| | IV PPML | | IV log-log | | IV linear probability | |
|---|---|---|---|---|---|---|
| Dep. variable: *citations* | $cit_{FiGjhkt}$ | | $log(cit_{FiGjhkt})$ | | $cit_{FiGjhkt} > 0$ | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| log(travel time) | $-0.152^{***}$ (0.029) | | -0.396$^{**}$ (0.175) | | -0.027$^{***}$ (0.004) | |
| log(travel time):0-300km | | -0.076 (0.221) | | 1.324 (1.680) | | -0.028 (0.036) |
| log(travel time):300-1000km | | $-0.134^{***}$ (0.044) | | -0.148 (0.378) | | -0.022$^{***}$ (0.007) |
| log(travel time):1000-2000km | | $-0.112^{**}$ (0.047) | | -0.314 (0.200) | | -0.021$^{***}$ (0.005) |
| log(travel time):+2000km | | $-0.203^{***}$ (0.043) | | -0.388$^{**}$ (0.185) | | -0.032$^{***}$ (0.005) |
| N obs. effective | 4,703,010 | 4,703,010 | 16,412 | 16,412 | 10,106,940 | 10,106,940 |
| R2 | 0.88 | 0.88 | 0.86 | 0.86 | 0.70 | 0.70 |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 18: Elasticity of citations to travel time: IV estimation intensive and extensive margin

Column (1) shows the result of Instrumental Variables Poisson estimation of $citations_{FiGjhkt} = \exp\left[\beta\, log(\text{travel time}_{ijt}) + \lambda\, \hat{u}_{FiGjhkt} + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt}\right] \times \varepsilon_{FiGjhkt}$, for citations of patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$, to patents filed by establishment of firm $G$ in location $j$ and technology $k$. $\text{travel time}_{ijt}$ is the travel time in minutes between location $i$ and $j$ at time period $t$, and it is set to 1 when $i = j$. The variable $\hat{u}_{FiGjhkt}$ is constructed as $\hat{u}_{FiGjhkt} = \text{travel time}_{FiGjhkt} - \hat{\lambda}_2\, \text{travel time}^{\text{fix network}}_{FiGjhkt}$. When $FiGjhk$ has positive citations in at least one period and no citations in another, we attribute zero citations in the missing period. Column (3) shows the result of an IV-2SLS estimation of $log(citations_{FiGjhkt}) = \alpha\, log(\text{travel time}_{ijt}) + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt} + \varepsilon_{FiGjhkt}$, with a sample of establishment-technology pairs ($FiGjhk$) that have positive citations in all periods. Column (5) shows the result of an IV-2SLS estimation of $\mathbb{1}\{citations_{FiGjhkt} > 0\} = \gamma\, log(\text{travel time}_{ijt}) + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt} + \varepsilon_{FiGjhkt}$, with the same sample as (1). Columns (3) and (5) use $\text{travel time}^{\text{fix network}}_{ijt}$ as an instrument for $\text{travel time}_{ijt}$. Column (2), (4) and (6) open, respectively, the coefficients $\beta$, $\alpha$, $\gamma$ by distance between the citing establishment $Fi$ and the cited establishment $Gj$. Standard errors are presented in parenthesis. In Columns (1) and (2) standard errors are bootstrapped. In Columns (3) to (6) standard errors clustered at the non-directional location pair ($ij$ is the same non-directional location pair as $ji$). R2 is computed as the squared correlation between observed and fitted values.

| | Concentration citing | Concentration cited | Cited lag forward | Citing lag backward | Citing govnt & uni | Cited govnt & univ | Self citation |
|---|---|---|---|---|---|---|---|
| Dep. variable: *citations* | | | | $cit_{FiGjhkt}$ | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| log(travel time):0-300km | 0.103 | 0.160 | −0.045 | 0.1907 | 0.021 | 0.018 | 0.002 |
| | (0.121) | (0.114) | (0.472) | (0.538) | (0.038) | (0.038) | (0.039) |
| log(travel time):300-1000km | −0.105 | −0.039 | −0.546 | −0.145 | −0.102*** | −0.099*** | −0.077*** |
| | (0.084) | (0.095) | (0.364) | (0.366) | (0.027) | (0.027) | (0.029) |
| log(travel time):1000-2000km | −0.138 | −0.117 | 0.086 | 0.101 | −0.094** | −0.093** | −0.094** |
| | (0.105) | (0.116) | (0.480) | (0.498) | (0.042) | (0.041) | (0.040) |
| log(travel time):+2000km | −0.287*** | −0.268*** | 0.720** | 0.560 | −0.185*** | −0.188*** | −0.153*** |
| | (0.105) | (0.090) | (0.344) | (0.472) | (0.049) | (0.048) | (0.040) |
| log(travel time):0-300km × $X$ | −1.180 | −2.013 | 0.028 | −0.066 | −0.125 | 0.481 | 0.038 |
| | (1.843) | (1.712) | (0.185) | (0.211) | (0.367) | (0.543) | (0.252) |
| log(travel time):300-1000km × $X$ | 0.079 | −0.880 | 0.178 | 0.018 | −0.088 | −0.609* | 0.077 |
| | (1.188) | (1.366) | (0.144) | (0.145) | (0.265) | (0.330) | (0.127) |
| log(travel time):1000-2000km × $X$ | 0.634 | 0.341 | −0.073 | −0.078 | −0.282 | −0.370 | 0.082 |
| | (1.412) | (1.606) | (0.191) | (0.197) | (0.366) | (0.385) | (0.210) |
| log(travel time):+2000km × $X$ | 1.436 | 1.157 | −0.366*** | −0.299 | −0.328 | 0.015 | −0.073 |
| | (1.456) | (1.136) | (0.137) | (0.188) | (0.410) | (0.295) | (0.170) |
| N obs. effective | 4,703,010 | 4,703,010 | 4,703,010 | 4,703,010 | 4,800,144 | 4,800,144 | 4,835,001 |
| R2 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 0.94 |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 20: Elasticity of citations to travel time: Heterogeneity (part 1)

Result of Poisson Pseudo Maximum Likelihood (PPML) estimation of $citations_{FiGjhkt} = \exp\left[\sum_d \beta_d \mathbb{1}\{distance_{ij} \in d\}\, log(\text{travel time}_{ijt}) + \sum_d \alpha_d \mathbb{1}\{distance_{ij} \in d\} \mathbb{1}\{X_{FiGjhkt}\}\, log(\text{travel time}_{ijt}) + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt}\right] \times \varepsilon_{FiGjhkt}$, for citations of patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$, to patents filed by establishment of firm $G$ in location $j$ and technology $k$. travel time$_{ijt}$ is the travel time in minutes between location $i$ and $j$ at time period $t$, and it is set to 1 when $i = j$. $d$ are distance intervals: $[0-300km]$, $(300km-1000km]$, $(1000km-2000km]$, $(2000km-max]$. The variable $X$ takes different value depending on the column: in column (1) it is the across-MSA Herfindahl index of the citing technology, in column (2) it is the across-MSA Herfindahl index of the cited technology, in column (3) it is median forward citation lag of the cited technology, in column (4) it is median backward citation lag of the citing technology. In column (5) and (6) the sample includes government and university patents, in column (5) $X$ is a dummy that takes value one if the citing patent belongs to a university or government organisation, in column (6) it is a dummy that takes value one if the cited patent belongs to a university or government organisation. In column (7) the sample includes self citations, the variable $X$ is a dummy that takes value one if the citing firm $F$ cited firm $G$ are the same. When $FiGjhk$ has positive citations in at least one period and no citations in another, we attribute zero citations in the missing period. Standard errors clustered at the non-directional location pair are presented in parenthesis ($ij$ is the same non-directional location pair as $ji$). R2 is computed as the squared correlation between observed and fitted values.

| | Citing quantile | Cited quantile |
|---|---|---|
| Dep. variable: *citations* | $cit_{FiGjhkt}$ | |
| | (1) | (2) |
| log(travel time) × quantile 0.00 | -0.151*** (0.058) | -0.111*** (0.039) |
| log(travel time) × quantile 0.01 | -0.078 (0.114) | -0.084 (0.101) |
| log(travel time) × quantile 0.25 | -0.081 (0.103) | -0.159* (0.093) |
| log(travel time) × quantile 0.50 | -0.139 (0.091) | -0.063 (0.083) |
| log(travel time) × quantile 0.75 | -0.262*** (0.079) | -0.033 (0.068) |
| log(travel time) × quantile 0.90 | -0.029 (0.066) | -0.127** (0.057) |
| log(travel time) × quantile 0.95 | -0.001 (0.037) | -0.123*** (0.038) |
| log(travel time) × quantile 0.99 | -0.130*** (0.035) | -0.066* (0.039) |
| log(travel time) × quantile 0.999 | -0.070 (0.045) | -0.070 (0.045) |
| N obs. effective | 4,703,010 | 4,703,010 |
| R2 | 0.88 | 0.88 |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 19: Elasticity of citations to travel time: Heterogeneity (part 2)

Column (1) shows the result of Poisson Pseudo Maximum Likelihood (PPML) estimation of $citations_{FiGjhkt} = \exp\left[\sum_q \boldsymbol{\beta}_q\, log(\text{travel time}_{ijt})\, \mathbb{1}\{quantile_{Fh} \in q\} + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt}\right] \times \varepsilon_{FiGjhkt}$, for citations of patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$, to patents filed by establishment of firm $G$ in location $j$ and technology $k$. travel time$_{ijt}$ is the travel time in minutes between location $i$ and $j$ at time period $t$, and it is set to 1 when $i = j$. $quantile_{Fh}$ is the quantile of firm $F$ in the distribution of firms within technology $h$, using patents applied by $F$ in $h$ in the time period 1949-1953. Column (2) repeats the analysis using the quantile of the cited firm $G$ in technology $k$. When $FiGjhk$ has positive citations in at least one period and no citations in another, we attribute zero citations in the missing period. When $FiGjhk$ has positive citations in at least one period and no citations in another, we attribute zero citations in the missing period. Standard errors clustered at the non-directional location in parentheses ($ij$ is the same non-directional location pair as $ji$). R2 is computed as the squared correlation between observed and fitted values.

| | PPML | |
|---|---|---|
| | Citing technology | Cited technology |
| Dep. variable: *citations* | $cit_{FiGjhkt}$ | |
| | (1) | (2) |
| log(travel time) × Chemical | $-0.066$ (0.045) | $-0.093^{**}$ (0.045) |
| log(travel time) × Computers & Communications | $-0.100$ (0.079) | $-0.140^{*}$ (0.077) |
| log(travel time) × Drugs & Medical | $-0.053$ (0.162) | $-0.005$ (0.181) |
| log(travel time) × Electrical & Electronic | $-0.070$ (0.048) | $-0.054$ (0.046) |
| log(travel time) × Mechanical | $-0.080^{**}$ (0.031) | $-0.087^{***}$ (0.032) |
| log(travel time) × Others | $-0.147^{***}$ (0.045) | $-0.113^{**}$ (0.044) |
| N obs. effective | 4,703,010 | 4,703,010 |
| R2 | 0.88 | 0.88 |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 21: Elasticity of citations to travel time by citing and cited technology
Part 1

Column (1) shows the result of Poisson Pseudo Maximum Likelihood (PPML) estimation of $citations_{FiGjhkt} = \exp\left[\sum_{tech} \beta_h \mathbb{1}\{tech = h\} \times log(\text{travel time}_{ijt}) + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt}\right] \times \varepsilon_{FiGjhkt}$, for citations of patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$, to patents filed by establishment of firm $G$ located in $j$, in technology $k$. $\mathbb{1}\{tech = h\}$ is a dummy variable that takes value 1 when the citing technology $h$ is equal to technology $tech$. In column (2) the dummy is modified to $\mathbb{1}\{tech = k\}$ such that it takes value 1 when the cited technology $k$ is equal to technology $tech$. travel time$_{ijt}$ is the travel time in minutes between location $i$ and $j$ at time period $t$, and it is set to 1 when $i = j$. When $FiGjhk$ has positive citations in at least one period and no citations in another, we attribute zero citations in the missing period. Standard errors clustered at the non-directional location pair are presented in parenthesis ($ij$ is the same non-directional location pair as $ji$). R2 is computed as the squared correlation between observed and fitted values.

### E.1.2. IV PPML: first and second stage estimation

### E.1.3. Robustness

#### Sample of establishments

During the time period there was entry and exit of research establishments that was not uniform across locations. We may then think that the change in diffusion of knowledge is only consequence of the change in the geographical location of innovation. To test this possibility, in Table 25 we estimate the baseline regression 3 with different samples.

| | First stage OLS | Second stage PPML |
|---|---|---|
| Dep. variable: | log(travel time) | $cit_{FiGjhkt}$ |
| | (1) | (2) |
| log(travel time fix routes) | $0.951^{***}$ (0.039) | |
| log(travel time) | | $-0.152^{***}$ (0.029) |
| residual | | $0.094^{***}$ (0.035) |
| N obs. effective | 10,106,940 | 4,703,010 |
| R2 | 0.99 | 0.88 |
| Within R2 | 0.38 | |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 22: Elasticity of citations to travel time: first and second stage IV PPML

| | OLS First stage 0-300km | OLS First stage 300-1,000km | OLS First stage 1,000-2,000km | OLS First stage +2,000km | Second stage PPML |
|---|---|---|---|---|---|
| Dep. variable: | log(travel time) | | | | $cit_{FiGjhkt}$ |
| | (1) | (2) | (3) | (4) | (5) |
| log(travel time fix routes) × 0-300km | $0.278^{**}$ (0.122) | 0.073 (0.057) | 0.024 (0.026) | $0.040^{*}$ (0.022) | |
| log(travel time fix routes) × 300-1,000km | $-0.103^{***}$ (0.032) | $1.113^{***}$ (0.041) | $-0.013$ (0.011) | 0.010 (0.011) | |
| log(travel time fix routes) × 1,000-2,000km | $-0.064^{***}$ (0.024) | $-0.052^{***}$ (0.020) | $1.059^{***}$ (0.044) | $0.017^{*}$ (0.009) | |
| log(travel time fix routes) × +2,000km | $-0.058^{***}$ (0.022) | $-0.046^{***}$ (0.017) | $-0.020^{**}$ (0.010) | $1.097^{***}$ (0.018) | |
| log(travel time) × 0-300km | | | | | $-0.076$ (0.221) |
| log(travel time) × 300-1,000km | | | | | $-0.134^{***}$ (0.044) |
| log(travel time) × 1,000-2,000km | | | | | $-0.112^{**}$ (0.047) |
| log(travel time) × +2,000km | | | | | $-0.203^{***}$ (0.043) |
| residual × 0-300km | | | | | 0.100 (0.196) |
| residual × 300-1,000km | | | | | 0.045 (0.053) |
| residual × 1,000-2,000km | | | | | 0.026 (0.069) |
| residual × +2,000km | | | | | 0.043 (0.078) |
| N obs. effective | 10,106,940 | 10,106,940 | 10,106,940 | 10,106,940 | 4,703,010 |
| R2 | 0.99 | 0.99 | 0.99 | 0.99 | 0.88 |
| Within R2 | 0.04 | 0.46 | 0.80 | 0.88 | |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 23: Elasticity of citations to travel time: first and second stage IV PPML

| Cited \ Citing | Chemical | Computers & Communications | Drugs & Medical | Electrical & Electronic | Mechanical | Others |
|---|---|---|---|---|---|---|
| Chemical | $-0.092^{**}$ (0.052) | 0.219 (0.262) | 0.113 (0.199) | $-0.299^{***}$ (0.094) | $-0.025$ (0.071) | $-0.070$ (0.068) |
| Computers & Communications | $-0.089$ (0.259) | $-0.306^{***}$ (0.095) | $-0.657$ (0.976) | 0.107 (0.090) | 0.122 (0.149) | 0.095 (0.169) |
| Drugs & Medical | 0.224 (0.239) | 0.567 (1.205) | $-0.278$ (0.268) | $-0.230$ (0.561) | $-0.334$ (0.362) | 0.358 (0.323) |
| Electrical & Electronic | $0.233^{**}$ (0.093) | $0.171^{*}$ (0.096) | $-0.224$ (0.634) | $-0.102^{**}$ (0.056) | 0.087 (0.070) | $-0.063$ (0.079) |
| Mechanical | $-0.060$ (0.076) | 0.151 (0.145) | $-0.152$ (0.402) | 0.106 (0.082) | $-0.129^{***}$ (0.035) | $-0.032$ (0.056) |
| Others | 0.042 (0.074) | 0.173 (0.169) | 0.204 (0.274) | 0.052 (0.072) | 0.019 (0.053) | $-0.209^{***}$ (0.054) |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 24: Elasticity of citations to travel time by citing and cited technology
Part 2

Column (1) shows the result of one single Poisson Pseudo Maximum Likelihood (PPML) estimation of $citations_{FiGjhkt} = \exp\left[\sum_{\text{tech pair}} \beta_{hk}\, \mathbb{1}\{\text{tech pair} = hk\} \times log(\text{travel time}_{ijt}) + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt}\right] \times \varepsilon_{FiGjhkt}$, for citations of patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$, to patents filed by establishment of firm $G$ located in $j$, in technology $k$. $\mathbb{1}\{\text{tech pair} = hk\}$ is a dummy variable that takes value 1 when the citing technology $h$ is equal to technology $tech$. In column (2) the dummy is modified to $\mathbb{1}\{tech = k\}$ such that it takes value 1 when the cited technology $k$ is equal to technology $tech$. travel time$_{ijt}$ is the travel time in minutes between location $i$ and $j$ at time period $t$, and it is set to 1 when $i = j$. When $FiGjhk$ has positive citations in at least one period and no citations in another, we attribute zero citations in the missing period. Standard errors clustered at the non-directional location pair are presented in parenthesis ($ij$ is the same non-directional location pair as $ji$). R2 is computed as the squared correlation between observed and fitted values. The amount of observation in the effective sample is 4,703,010.

In column (1) we include the baseline results.[103] In column (2) we use only citing establishments $Fi$ that filed patents during the initial time period 1949-1953. In column (3) we further restrict the sample to both citing establishments $Fi$ and cited establishments $Gj$ that filed patents in 1949-1953.[104] We find that the coefficient at more than 2,000km remains comparable to the one in the baseline regression, statistically significant at the 1%.

[103]Coefficients are not bias corrected.

[104]We require $Fi$ and $Gj$ to have positive amount of patents applied during 1949-1953. However, those establishments need not to have cited each other.

| | All | Citing establishment | Citing & Cited establishment |
|---|---|---|---|
| Dep. variable: *citations* | | $cit_{FiGjhkt}$ | |
| | (1) | (2) | (3) |
| log(travel time) × 0-300km | 0.021 | 0.020 | 0.028 |
| | (0.039) | (0.043) | (0.043) |
| log(travel time) × 300-1,000km | $-0.099^{***}$ | $-0.095^{***}$ | $-0.095^{***}$ |
| | (0.027) | (0.029) | (0.030) |
| log(travel time) × 1,000-2,000km | $-0.093^{**}$ | $-0.092^{**}$ | $-0.062$ |
| | (0.042) | (0.047) | (0.050) |
| log(travel time) × +2,000km | $-0.185^{***}$ | $-0.155^{***}$ | $-0.179^{***}$ |
| | (0.049) | (0.052) | (0.052) |
| N obs. effective | 4,703,010 | 3,109,285 | 1,960,851 |
| R2 | 0.88 | 0.88 | 0.89 |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 25: Elasticity of citations to travel time: Fix sample of establishments

Column (1) shows the result of Poisson Pseudo Maximum Likelihood (PPML) estimation of $citations_{FiGjhkt} = \exp\left[\sum_d \beta_d \times \mathbb{1}\{distance_{ij} \in d\} \times log(\text{travel time}_{ijt}) + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt}\right] \times \varepsilon_{FiGjhkt}$, for citations of patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$, to patents filed by establishment of firm $G$ in location $j$ and technology $k$. travel time$_{ijt}$ is the travel time in minutes between location $i$ and $j$ at time period $t$, and it is set to 1 when $i = j$. $d$ are distance intervals: $[0 - 300km]$, $(300km - 1000km]$, $(1000km - 2000km]$, $(2000km - max]$. Column (2) truncates the sample keeping only citing establishments $Fi$ that where present in the initial time period $1949 - 1953$. Column (3) truncates the sample keeping only citing establishments $Fi$ and cited establishments $Gj$ that where present in the initial time period. When $FiGjhk$ has positive citations in at least one period and no citations in another, we attribute zero citations in the missing period. Standard errors clustered at the non-directional location pair are presented in parenthesis ($ij$ is the same non-directional location pair as $ji$). R2 is computed as the squared correlation between observed and fitted values.

#### Ticket prices

During the period of analysis ticket prices were set by the Civil Aeronautics Board, so airlines could not set prices of their own tickets. Some airlines included a sample of prices in the last page of their booklet of flight schedules a sample of prices, which we digitized. We have digitized American Airlines 1951, 1961, 1966; TWA 1951 and United Airlines 1956 and 1961.[105]. The sample includes prices for 11,590 directional airport pair years. We document multiple facts about prices.

First, prices were set in the form of an intercept plus a variable increment depending on distance between origin and destination (until 1962-1963). A linear regression with an intercept and a slope estimated separately for each year (including 1966), service class (first class or coach service), and aircraft type (propeller or jet) gives a R2 of 0.98

[105]The sample of prices digitized was limited due to data availability.

or higher in each regression, with an average R2 of 0.993.

Second, all airlines operating within the same route charged exactly the same price. In 1951, in our digitized price data we have 432 airport pairs in which both American Airlines and TWA were operating and reported the price for first class service. 94% of those airport pairs had exactly the same price in both airlines.

Third, ticket prices of flights operated by jet airplanes had a surcharge of around 6% on top of the one operated by propeller airplanes.

Fourth, the change in prices over time had a similar pattern until 1961: a stronger increase in short distances (probably due to an increase in fixed costs of take-off and landing, although not reflected in the intercept of the linear regressions), and a relatively constant increase for flights between airports more than 1,000 km apart. In the period 1961 to 1966 we observe a drop in prices of around 20% for routes of more than 1,000km distance, breaking the linearity of prices in distance previously observed. We had visually inspected price tables and detected that the drop in prices happened in 1962-1963.

Figure 49 shows prices for first class service by year and aircraft type, deflated by the consumer price index to 1951 values. Figure 50 presents the percentage change in deflated prices of first class service. Both figures show the previous facts: prices are generally linear in distance until 1966 in which we observe a break after 1,000 km.

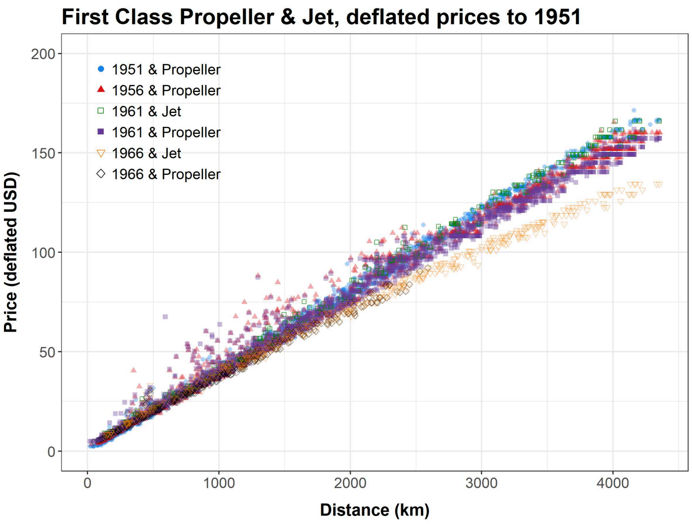



Figure 49: Flight ticket prices, deflated by CPI

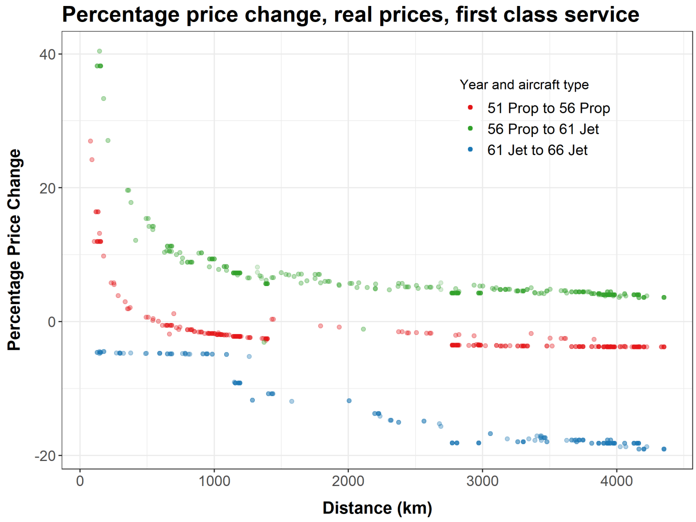



Figure 50: Change flight ticket prices, deflated by CPI

We convert our sample of prices at the airport-pair level to prices of the population of MSA-pairs as follows: first, we obtain a pricing function that can flexibly approximate prices by regressing deflated prices on a cubic polynomial of distance separately for each year. We use prices of first class service for all years, propeller aircraft for 1951 and 1956 and jet aircraft for 1961 and 1966. Second, we predict prices for each MSA-pair and year using the MSA-pair distance and the year's estimated regression.

### Highway travel time

Taylor Jaworski and Carl Kitchens have graciously shared with us data on county-to-county highway travel time and nominal travel costs for 1950, 1960 and 1970. Travel time is constructed using maximum speed limit in each highway segment and year. Travel costs uses, for each year, travel time, highway distance, truck driver's wage and petrol costs. See Jaworski and Kitchens (2019) for details. The dataset is constructed using 2010 county boundaries and contains county centroids. We converted it to MSA-to-MSA by matching counties' centroids to 1950 MSAs using the shape file from Manson et al. (2020). We take the minimum travel time and minimum travel costs among all county pairs that belong to the same MSA pair. We convert nominal travel costs to

1950 real travel costs deflating by the consumer price index. We convert 1950, 1960 and 1970 travel times and travel costs to 1951, 1956, 1961 and 1966 by linearly interpolating (e.g. travel $\text{time}_{ij,1951} =$ travel $\text{time}_{ij,1950} \times \frac{1960-1951}{10} +$ travel $\text{time}_{ij,1960} \times \frac{1951-1950}{10}$).

The within MSA-pair correlation of the 1951-1966 change in travel time by highway and airplane is 0.068 for all MSA-pairs, and -0.011 for MSA-pairs more than 2,000 km apart. Figure 51 presents the MSA-pair 1951-1966 change in travel time by highway and airplane, where for exposition we only present MSA-pairs that had a reduction in travel time by both means of transport. Estimating a linear regression of change in air travel time on the change in highway travel time gives a slope of -0.02 not statistically different from zero, with a R2 of 0.00005.[106] Figure 52 repeats the exercise where MSA-pairs are weighted by the amount of establishment-technology pairs used to estimate the elasticity of citations to travel time (equation (3)). In this case the estimated regression has a slope of 0.73 statistically significant at the 1% level and a R2 of 0.09.[107]

In Tables 5 and 26 we present the results of adding highway travel time as control. The low correlation between the change in travel time by highway and airplane implies that the estimated elasticity of citations to air travel time remains almost unchanged, relative to the baseline estimation.[108]

[106] 8.7% of MSA-pairs had an increase in travel time either by highway or by airplane. The regression with all MSA-pairs has a slope of 0.60 significant at the 1% level. However, the R2 of the regression remains very low: 0.0046.

[107] With all MSA-pairs the slope is 1.01 statistically significant at the 1% level and the R2 is 0.04.

[108] In order to perform a test of statistical difference of coefficients we would need to compute the covariance between the two regressions. Assuming the covariance is zero, in columns (1) and (2) 26 the coefficients of air travel time at +2,000km are not significantly different.

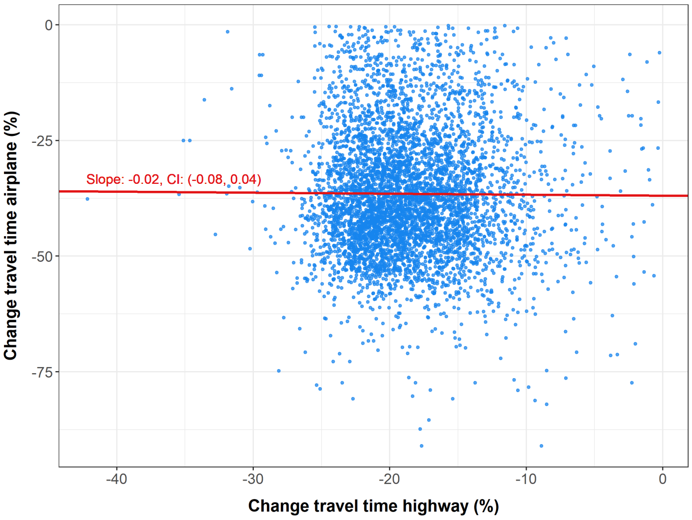



Figure 51: Change travel time by airplane and highway 1951-1966

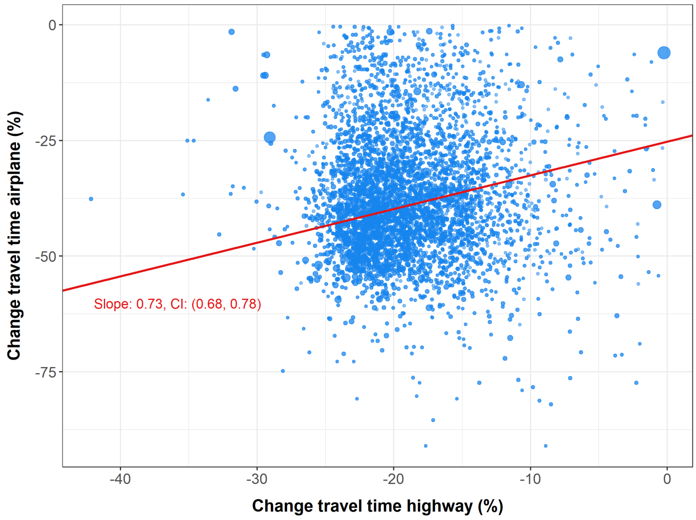



Figure 52: Change travel time by airplane and highway 1951-1966, weighted

| | PPML | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Dep. variable: *citations* | $cit_{FiGjhkt}$ | | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| log(travel time) × 0-300km | 0.0213 | 0.0276 | 0.0198 | 0.0318 | 0.0252 | 0.0349 | 0.0283 | 0.0313 |
| | (0.0388) | (0.0385) | (0.0391) | (0.0393) | (0.0389) | (0.0391) | (0.0396) | (0.0393) |
| log(travel time) × 300-1,000km | -0.0990*** | -0.1040*** | -0.0935*** | -0.0745** | -0.1014*** | -0.0857*** | -0.0748** | -0.0861*** |
| | (0.0269) | (0.0292) | (0.0265) | (0.0303) | (0.0290) | (0.0312) | (0.0303) | (0.0312) |
| log(travel time) × 1000-2,000km | -0.0928** | -0.1155** | -0.0710* | -0.0395 | -0.0948* | -0.0498 | -0.0318 | -0.0435 |
| | (0.0418) | (0.0485) | (0.0423) | (0.0523) | (0.0502) | (0.0573) | (0.0520) | (0.0576) |
| log(travel time) × +2,000km | -0.1848*** | -0.1761*** | -0.1724*** | -0.1238** | -0.1658*** | -0.1052* | -0.1236** | -0.1041* |
| | (0.0492) | (0.0531) | (0.0498) | (0.0587) | (0.0542) | (0.0607) | (0.0590) | (0.0609) |
| N obs. effective | 4,703,010 | 4,703,010 | 4,703,010 | 4,703,010 | 4,703,010 | 4,703,010 | 4,703,010 | 4,703,010 |
| R2 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 |
| Controls: | | | | | | | | |
| log(highway time) | - | Yes | - | - | Yes | Yes | - | Yes |
| log(telephone share) × time | - | - | Yes | - | Yes | - | Yes | Yes |
| log(distance) × time | - | - | - | Yes | - | Yes | Yes | Yes |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 26: Elasticity of citations to travel time: additional controls

Column (1) shows the result of Poisson Pseudo Maximum Likelihood (PPML) estimation of $citations_{FiGjhkt} = \exp\left[\sum_d \beta_d \mathbb{1}\{distance_{ij} \in d\}\, log(\text{travel time}_{ijt}) + \sum_d \alpha_d \mathbb{1}\{distance_{ij} \in d\}\, \mathbb{1}\{X_{FiGjhkt}\}\, log(\text{travel time}_{ijt}) + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt}\right] \times \varepsilon_{FiGjhkt}$, for citations of patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$, to patents filed by establishment of firm $G$ in location $j$ and technology $k$. travel time$_{ijt}$ is the travel time in minutes between location $i$ and $j$ at time period $t$, and it is set to 1 when $i = j$. $d$ are distance intervals: $[0 - 300km], (300km - 1000km], (1000km - 2000km], (2000km - max]$. Relative to (1), columns (2) to (8) contain additional controls. Log highway time between $i$ and $j$ changes in every time period $t$. The log mean share of households with telephone line in $ij$ pair interacted in 1960 is interacted with a time dummy. Log distance $ij$ is interacted with a time dummy. When $FiGjhk$ has positive citations in at least one period and no citations in another, we attribute zero citations in the missing period. Standard errors clustered at the non-directional location in parentheses ($ij$ is the same non-directional location pair as $ji$). R2 is computed as the squared correlation between observed and fitted values.

**Frequency adjusted travel time**

The frequency of flights may have changed simultaneously with the introduction of jet airplanes. The change in travel time could then be consequence of higher frequency rather than changes in airplanes' speed. Given that some MSA pairs are connected indirectly (with connecting flights), accounting for frequency is not straight forward: the frequency of each leg of the flight route matters (actually, it is not only frequency of each leg but also the synchronization among all potential legs). In order to take into account potential changes in the frequency of flights we computed the daily average travel time. This travel time is the average across all fastest travel times if the passenger was to depart at each full hour (1am, 2am, ..., 1pm, 2pm, etc.). The computation of this travel time includes the waiting time that is affected by frequency: the time until first departure and layover time of each connecting flight. Hence, the daily average travel time is a frequency-adjusted travel time: changes in the daily average travel time that are larger than in the fastest travel time denote that frequency of flights increased and therefore there is less waiting time. If we observe the reverse that means that frequency did not improve as much as the speed of airplanes.

Figure 53 shows the within MSA-pair decrease in the fastest travel time and the daily average travel time.[109] Both measures of travel time follow a similar pattern: slight decrease in 1956, a stronger decrease in 1961 especially for long distance routes, and a further decline in 1966. However, we observe that the decrease of the fastest travel time is on average larger than the one of the daily average travel time: the frequency of flights, if any, attenuated the potential decrease in travel time from the improvements in airplanes' speed. This observation is also in line with a comparison of the fastest travel time with and without layover time (Figure 28 in the Appendix of the paper): layover time attenuated the change in travel time.

In table 27 we estimated the elasticity of citations to travel time using first the fastest

[109]The within MSA-pair correlation of the (1951-1966) change in fastest travel time and the change daily average travel time is 0.60.

travel time (baseline, columns 1 and 2) and the daily average travel time (columns 3 and 4). The estimated elasticity is similar using both measures, which gives confidence that our results are not driven by changes in the frequency of flights.

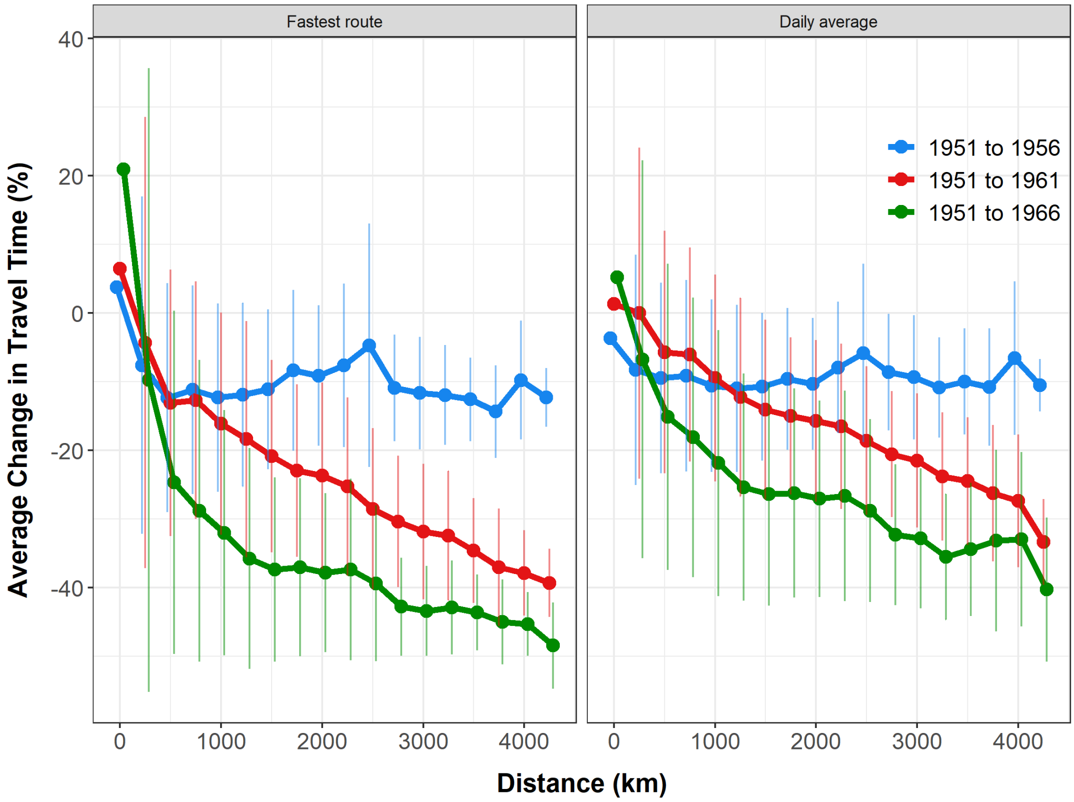



Figure 53: Change in MSAs travel time: fastest travel time and daily average travel time

| | PPML not bias-corrected | | | |
|---|---|---|---|---|
| Dep. variable: *citations* | $cit_{FiGjhkt}$ | | | |
| | (1) | (2) | (3) | (4) |
| log(travel time) | $-0.088^{***}$ (0.024) | | | |
| log(travel time) × 0-300km | | 0.021 (0.039) | | |
| log(travel time) × 300-1,000km | | $-0.099^{**}$ (0.027) | | |
| log(travel time) × 1000-2,000km | | $-0.093^{**}$ (0.042) | | |
| log(travel time) × +2,000km | | $-0.185^{***}$ (0.049) | | |
| log(travel time daily avg) | | | $-0.100^{***}$ (0.039) | |
| log(travel time daily avg) × 0-300km | | | | 0.034 (0.037) |
| log(travel time daily avg) × 300-1,000km | | | | $-0.142^{***}$ (0.047) |
| log(travel time daily avg) × 1000-2,000km | | | | $-0.170^{***}$ (0.072) |
| log(travel time daily avg) × +2,000km | | | | $-0.236^{***}$ (0.064) |
| N obs. effective | 4,703,010 | 4,703,010 | 4,703,010 | 4,703,010 |
| R2 | 0.88 | 0.88 | 0.88 | 0.88 |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 27: Elasticity of citations to travel time: daily average travel time

Column (1) shows the result of Poisson Pseudo Maximum Likelihood (PPML) estimation of $citations_{FiGjhkt} = \exp\left[\boldsymbol{\beta}\, log(\text{travel time}_{ijt}) + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt}\right] \times \varepsilon_{FiGjhkt}$, for citations of patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$, to patents filed by establishment of firm $G$ in location $j$ and technology $k$. travel time$_{ijt}$ is the travel time in minutes between location $i$ and $j$ at time period $t$, and it is set to 1 when $i = j$. When $FiGjhk$ has positive citations in at least one period and no citations in another, we attribute zero citations in the missing period. Column (2) includes the interaction of travel time$_{ijt}$ with a dummy for distance bin between the citing establishment $Fi$ and the cited establishment $Gj$. Column (3) and (4) use the daily average travel time, which is computed as the average of the fastest travel time departing at every full hour (the average across all 24 potential departing times). Standard errors clustered at the non-directional location are presented between parentheses ($ij$ is the same non-directional location pair as $ji$). R2 is computed as the squared correlation between observed and fitted values.

| | First stage OLS | Second stage PPML |
|---|---|---|
| Dep. variable: | log(knowledge access) | $Patentscit_{Fiht}$ |
| | (1) | (2) |
| log(knowledge access fix routes) | 1.01*** (0.032) | |
| log(knowledge access) | | 11.24* (6.35) |
| residual | | −2.31 (7.20) |
| N obs. effective | 991,480 | 91,480 |
| R2 | 0.99 | 0.85 |
| Within R2 | 0.53 | |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 28: Elasticity of patents to knowledge access: first and second stage IV PPML

| | OLS First stage reference quartile | OLS First stage 3rd quartile | OLS First stage 2nd quartile | OLS First stage 3rd quartile | Second stage PPML |
|---|---|---|---|---|---|
| Dep. variable: | log(knowledge access) | | | | $Patents_{Fiht}$ |
| | (1) | (2) | (3) | (4) | (5) |
| log(knowledge access time fix routes) | 1.00*** (0.03) | 0.01 (0.06) | 0.03 (0.03) | 0.00 (0.01) | |
| log(knowledge access fix routes) × 3rd quartile | 0.01* (0.004) | 1.11*** (0.03) | −0.00 (0.00) | −0.00 (0.00) | |
| log(knowledge access fix routes) × 2nd quartile | 0.00 (0.01) | −0.01 (0.04) | 1.11*** (0.03) | −0.00 (0.00) | |
| log(knowledge access fix routes) × 1st quartile | 0.01 (0.01) | −0.00 (0.04) | −0.04 (0.04) | 1.15*** (0.04) | |
| log(knowledge access) | | | | | 10.26 (6.38) |
| log(knowledge access) × 3rd quartile | | | | | 2.32*** (0.66) |
| log(knowledge access) × 2nd quartile | | | | | 4.21*** (0.84) |
| log(knowledge access) × 1st quartile | | | | | 5.77*** (1.11) |
| residual | | | | | −2.25 (7.27) |
| residual × 3rd quartile | | | | | −2.55 (1.59) |
| residual × 2nd quartile | | | | | −4.32** (1.97) |
| residual × 1st quartile | | | | | −8.27** (3.28) |
| N obs. effective | 991,480 | 991,480 | 991,480 | 991,480 | 991,480 |
| R2 | 1.00 | 1.00 | 1.00 | 1.00 | 0.85 |
| Within R2 | 0.53 | 0.89 | 0.90 | 0.90 | |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 29: Elasticity of patents to knowledge access: first and second stage IV PPML

| | Baseline | Quartile absolute | Quartile per capita |
|---|---|---|---|
| Dependent Variable: *Patents* | | $Patents_{Fiht}$ | |
| | (1) | (2) | (3) |
| log(knowledge access) | 10.14*** (3.66) | 9.36** (3.69) | 7.77** (3.70) |
| log(knowledge access) × quartile 0.50 | | 2.05*** (0.58) | 0.75** (0.34) |
| log(knowledge access) × quartile 0.25 | | 3.80*** (0.90) | 1.58*** (0.50) |
| log(knowledge access) × quartile 0.00 | | 5.00*** (1.30) | 4.03*** (0.77) |
| N obs. effective | 991,480 | 991,480 | 991,480 |
| R2 | 0.85 | 0.85 | 0.85 |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 30: Elasticity of new patents to knowledge access: absolute and per capita MSA innovativeness

Column (1) shows the result of Poisson Pseudo Maximum Likelihood (PPML) estimation of $\text{Patents}_{Fiht} = \exp\left[\boldsymbol{\rho}\, log(KA_{iht}) + FE_{Fih} + FE_{it} + FE_{ht}\right] \times \xi_{Fiht}$, for patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$. $KA_{iht}$ is knowledge access of establishments in location $i$ technology $h$ and time period $t$. Column (2) opens the coefficient $\rho$ by the quartile of innovativeness of location $i$ within technology $h$, computed within technology using the absolute level of patents in the MSA-technology in 1949-1953. Column (3) computes the quartile of innovativeness using patents per capita in the MSA-technology in 1949-1953 using 1950 population. Higher quartile indicates higher initial level of innovativeness. The fourth quartile is used as reference category. Standard errors clustered at the location-technology $ih$ are presented in parentheses. R2 is computed as the squared correlation between observed and fitted values.

## E.2. Creation of knowledge

### E.2.1. Heterogeneous effects

### E.2.2. IV PPML: first and second stage estimation

### E.2.3. Robustness

| | PPML | | $\beta$ by distance | | +300km | | +1,000km | | +2,000km | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dependent Variable: *Patents* | $Patents_{Fiht}$ | | | | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| log(knowledge access) | 10.14*** | 9.36** | 18.17*** | 16.50** | 10.09** | 8.70* | 18.82*** | 19.08*** | 12.70 | 10.26 |
| | (3.66) | (3.69) | (4.63) | (4.76) | (4.66) | (4.67) | (5.82) | (5.74) | (8.18) | (7.92) |
| log(knowledge access) × 3rd quartile | | 2.05*** | | 2.70*** | | 2.12*** | | 2.08*** | | 1.94*** |
| | | (0.58) | | (0.84) | | (0.58) | | (0.53) | | (0.49) |
| log(knowledge access) × 2nd quartile | | 3.80*** | | 5.96*** | | 4.19*** | | 3.97*** | | 3.64*** |
| | | (0.90) | | (1.42) | | (0.88) | | (0.81) | | (0.73) |
| log(knowledge access) × 1st quartile | | 5.00*** | | 8.94*** | | 5.49*** | | 5.28*** | | 4.68*** |
| | | (1.30) | | (1.97) | | (1.25) | | (1.23) | | (1.07) |
| N obs. effective | 991,480 | 991,480 | 991,480 | 991,480 | 991,480 | 991,480 | 991,480 | 991,480 | 991,480 | 991,480 |
| R2 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 31: Elasticity of new patents to knowledge access, varying beta or distance.

Column (1) shows the result of Poisson Pseudo Maximum Likelihood (PPML) estimation of $\text{Patents}_{Fiht} = \exp\left[\boldsymbol{\rho}\, log(KA_{iht}) + FE_{Fih} + FE_{it} + FE_{ht}\right] \times \xi_{Fiht}$, for patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$. $KA_{iht}$ is knowledge access of establishments in location $i$ technology $h$ and time period $t$. Column (2) opens the coefficient $\rho$ by the quartile of innovativeness of location $i$ within technology $h$, computed using patents in 1949-1953. Higher quartile indicates higher initial level of innovativeness. The fourth quartile is used as reference category. Relative to columns (1) and (2), columns (3) and (4) compute Knowledge Access using four distance-specific $\beta$ parameter according to distance bins between $i$ and $j$. The bins are [0km, 300km], (300km, 1000km], (1000km, 2000km], +2,000km. Columns (5) to (10) use the same $\beta$ as column (1) and (2), but computing Knowledge Access with a truncated sample of $j$ that are further than a certain distance threshold from $i$. Standard errors clustered at the location-technology $ih$ are presented in parentheses. R2 is computed as the squared correlation between observed and fitted values.

| | PPML | | OLS | |
|---|---|---|---|---|
| Dependent Variable: *Patents* | $Patents_{Fiht}$ | | $log(Patents_{Fiht})$ | |
| | (1) | (2) | (3) | (4) |
| log(knowledge access) | 10.14*** (3.66) | 9.36** (3.69) | 6.83* (3.19) | 6.27* (3.20) |
| log(knowledge access) × 3rd quartile | | 2.05*** (0.58) | | 0.92* (0.51) |
| log(knowledge access) × 2nd quartile | | 3.80*** (0.90) | | 2.64** (1.03) |
| log(knowledge access) × 1st quartile | | 5.00*** (1.30) | | 3.82** (1.79) |
| N obs. effective | 991,480 | 991,480 | 300,539 | 300,539 |
| R2 | 0.85 | 0.85 | 0.87 | 0.87 |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 32: Elasticity of new patents to knowledge access: PPML and OLS

Column (1) shows the result of Poisson Pseudo Maximum Likelihood (PPML) estimation of $\text{Patents}_{Fiht} = \exp\left[\boldsymbol{\rho}\, log(KA_{iht}) + FE_{Fih} + FE_{it} + FE_{ht}\right] \times \xi_{Fiht}$, for patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$. $KA_{iht}$ is knowledge access of establishments in location $i$ technology $h$ and time period $t$. Column (3) estimates $log(\text{Patents})_{Fiht} = \boldsymbol{\rho}\, log(KA_{iht}) + FE_{Fih} + FE_{it} + FE_{ht} + \xi_{Fiht}$. Columns (2) and (4) open the coefficient $\rho$ by the quartile of innovativeness of location $i$ within technology $h$, computed within technology using the absolute level of patents in the MSA-technology in 1949-1953. Higher quartile indicates higher initial level of innovativeness. The fourth quartile is used as reference category. Difference in amount of observations is due to dropping zeros in columns (3) and (4). Standard errors clustered at the location-technology $ih$ are presented in parentheses. R2 is computed as the squared correlation between observed and fitted values.

**Access to capital**

We construct four measures of access to capital using 1949-1953 market capitalization of firms listed in the stock market. The four measures are similar in their essence but differ in the computation of a firm's technology and the firm's location. The measure is computed as follows:

$$capital\ access_{iht} = \sum_{k} \psi_{hk} \sum_{j, j \neq i} \text{Capital stock}_{jk,t=1951} \times \text{travel time}_{ijt}^{\xi} \tag{15}$$

where $\text{Capital stock}_{jk,t=1951}$ is a proxy for the capital which is specific to technology $k$ located in $j$ at the initial time period 1951. $\psi_{hk}$ is an input-output weight of capital flows and $\xi$ is the elasticity of capital flows between to travel time. As a proxy for capital we use market capitalization of firms.

We construct four measures of $capital\ access_{iht}$ which differ on: (i) the way we define the allocation of the firm's capital to each location (either using all inventors' locations or only the assigned headquarters), and (ii) the way we allocate a firm's capital across technologies (using the share of a technology within the firm, or relative to the national share of that technology). We use COMPUSTAT as our source of data for market capitalization.

We proceed as follows:

1. Use share's market price at closure calendar year multiplied by the number shares outstanding. We use the variables *prcc_c* and *csho* to maximize coverage of firms given that other variables have missing value for many firms.

2. Take the yearly average market capitalization to maximize coverage (many firms have missing in a certain year). This step potentially introduces measurement error due to changes in total stock market capitalization but allows us to increase the amount of firms included in the sample.

3. Determine a firm's MSA using patent inventor location. Two ways to determine the location, 1. only HQ location, 2. all locations where the firm had inventors applying for patents in 1949-1953

4. Determine the share of each technology firm's technology using patent technology. Two ways to determine the share oftechnology: 1. the share of each tech within firm + share within firm relative to national share

5. In the absence of data on a capital input-output weight, assume it is the same as the technology input-output weight, i.e. $\psi_{hk} = \omega_{hk}$

6. In the absence of data on the elasticity of capital flows to travel time assume $\xi = -1$

The four measures of access to capital are as follows:

1. Attribute all capital to headquarters and use the absolute share of each technology in the firm

2. Attribute all capital to headquarters and use the share of each technology in the firm relative to the national share
3. Attribute capital to establishments using their pat share and use the absolute share of each technology in the firm
4. Attribute capital to establishments using their pat share and use the share of each technology in the firm relative to the national share

Table 33 shows the results of estimating the elasticity of new patents to knowledge access while at the same time controlling for capital access.

| Dependent Variable: *Patents* | $Patents_{Fiht}$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| log(knowledge access) | 10.14*** (3.66) | | | | | 9.96** (4.50) | 11.29*** (4.32) | 10.67** (4.70) | 12.90*** (4.43) |
| log(finance access hq) | | 0.54** (0.26) | | | | 0.02 (0.30) | | | |
| log(finance access hq rel) | | | 0.40 (0.25) | | | | -0.14 (0.28) | | |
| log(finance access est) | | | | 0.56* (0.31) | | | | -0.07 (0.39) | |
| log(finance access est rel) | | | | | 0.31 (0.30) | | | | -0.39 (0.38) |
| N obs. effective | 991,480 | 991,480 | 991,480 | 991,480 | 991,480 | 991,480 | 991,480 | 991,480 | 991,480 |
| R2 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 |

***$p < 0.01$; **$p < 0.05$; *$p < 0.10$

Table 33: Elasticity of new patents to knowledge access and finance access

Column (1) shows the result of Poisson Pseudo Maximum Likelihood (PPML) estimation of $\text{Patents}_{Fiht} = \exp\left[\boldsymbol{\rho}\, log(KA_{iht}) + FE_{Fih} + FE_{it} + FE_{ht}\right] \times \xi_{Fiht}$, for patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$. $KA_{iht}$ is knowledge access of establishments in location $i$ technology $h$ and time period $t$. Column (2) to (5) use as regressor the finance access of establishments in location $i$ technology $h$ and time period $t$, where the measure of finance access changes across columns. Columns (6) to (9) estimate the regression using both knowledge access and finance access. Standard errors clustered at the location-technology $ih$ are presented in parentheses. R2 is computed as the squared correlation between observed and fitted values.

### Sensitivity to $\beta$

### Indirectly connected MSAs

If the 1951 flight network was constructed in order to connect city pairs that would see future growth in citations, we can alleviate this endogeneity concern by focusing only

| $\beta$ | $\rho$ | $\beta \times \rho$ | Predicted yearly growth p.p. | Share yearly growth explained | Predicted yearly growth differential p.p. | Share yearly growth differential explained |
|---|---|---|---|---|---|---|
| -0.186 | 10.14 | -1.89 | 3.47 | 0.78 | 1.1 | 0.21 |
| -0.1 | 19.35 | -1.94 | 3.5 | 0.78 | 1.07 | 0.2 |
| -0.2 | 9.4 | -1.88 | 3.47 | 0.78 | 1.1 | 0.21 |
| -0.3 | 6.1 | -1.83 | 3.45 | 0.77 | 1.14 | 0.22 |
| -0.4 | 4.48 | -1.79 | 3.44 | 0.77 | 1.16 | 0.22 |
| -0.5 | 3.52 | -1.76 | 3.44 | 0.77 | 1.19 | 0.23 |
| -0.6 | 2.91 | -1.74 | 3.45 | 0.77 | 1.2 | 0.23 |
| -0.7 | 2.48 | -1.73 | 3.47 | 0.78 | 1.22 | 0.23 |
| -0.8 | 2.17 | -1.73 | 3.5 | 0.78 | 1.22 | 0.23 |
| -0.9 | 1.93 | -1.73 | 3.52 | 0.79 | 1.24 | 0.24 |
| -1 | 1.72 | -1.72 | 3.51 | 0.79 | 1.28 | 0.24 |
| -2 | 0.58 | -1.16 | 2.8 | 0.63 | 1.55 | 0.3 |
| -5 | 0.04 | -0.19 | 1.19 | 0.27 | 3.65 | 0.7 |
| -8 | 0.09 | -0.76 | 8.22 | 1.84 | 6.96 | 1.33 |
| -10 | 0.11 | -1.08 | 15.16 | 3.4 | 8.19 | 1.56 |
| -20 | 0.13 | -2.63 | 69.8 | 15.65 | 21.66 | 4.14 |
| -50 | 0.16 | -8.22 | 531.34 | 119.16 | 219.49 | 41.94 |
| -100 | 0.12 | -12.33 | 5428.85 | 1217.49 | 2971.74 | 567.91 |

Table 34: Effect of knowledge access on new patents: varying the value of elasticity of knowledge diffusion

on indirectly connected pairs.

Table 35 presents PPML regressions not bias-corrected. Columns (1) and (2) are the baseline regressions (all MSA-pairs), columns (3) and (4) drop MSA-pairs that are ever connected with one leg (a non-stop flight), and columns (5) and (6) drop MSA-pairs that are ever connected with one flight number. The difference between non-stop and one flight number is that one flight number could serve multiple MSAs by making intermediate stops.[110] The estimated coefficients are in the ballpark of the initial estimates, especially for +2,000km, providing evidence that it is reasonable to use the pre-existing network as the baseline to construct the instrument.

[110]For example, in 1951 NYC-LA was connected with one flight number that included one stop in Chicago, that is two legs but only one flight number: passengers did not have to change airplanes).

| | PPML not bias-corrected | | | | | |
|---|---|---|---|---|---|---|
| Dep. variable: *citations* | $cit_{FiGjhkt}$ | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| log(travel time) | $-0.088^{***}$ (0.024) | | $-0.202^{***}$ (0.051) | | $-0.241^{***}$ (0.061) | |
| log(travel time) × 0-300km | | 0.021 (0.039) | | $-0.237^{***}$ (0.116) | | $-0.410^{**}$ (0.165) |
| log(travel time) × 300-1,000km | | $-0.099^{**}$ (0.027) | | $-0.147^{*}$ (0.081) | | $-0.210^{**}$ (0.095) |
| log(travel time) × 1000-2,000km | | $-0.093^{**}$ (0.042) | | $-0.157^{*}$ (0.092) | | $-0.216^{**}$ (0.109) |
| log(travel time) × +2,000km | | $-0.185^{***}$ (0.049) | | $-0.297^{***}$ (0.085) | | $-0.242^{***}$ (0.090) |
| N obs. effective | 4,703,010 | 4,703,010 | 1,735,427 | 1,735,427 | 1,396,393 | 1,396,393 |
| R2 | 0.88 | 0.88 | 0.94 | 0.94 | 0.94 | 0.94 |
| *Observation selection:* | | | | | | |
| All | X | X | | | | |
| Discard one leg | | | X | X | | |
| Discard one flight number | | | | | X | X |

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Table 35: Elasticity of citations to travel time: dropping directly connected MSA pairs

Column (1) shows the result of Poisson Pseudo Maximum Likelihood (PPML) estimation of $citations_{FiGjhkt} = \exp\left[\boldsymbol{\beta}\, log(\text{travel time}_{ijt}) + FE_{FiGjhk} + FE_{Fiht} + FE_{Gjkt}\right] \times \varepsilon_{FiGjhkt}$, for citations of patents filed by establishment of firm $F$ in location $i$, technology $h$ and time period $t$, to patents filed by establishment of firm $G$ in location $j$ and technology $k$. travel time$_{ijt}$ is the travel time in minutes between location $i$ and $j$ at time period $t$, and it is set to 1 when $i = j$. When $FiGjhk$ has positive citations in at least one period and no citations in another, we attribute zero citations in the missing period. Column (2) includes the interaction of travel time$_{ijt}$ with a dummy for distance bin between the citing establishment $Fi$ and the cited establishment $Gj$. Column (3) and (4) discards all $ij$ that are ever connected with one leg (non-stop flight), while columns (5) and (6) discard all $ij$ that are ever connected with one flight number. The difference between non-stop and one flight number is that one flight number could serve multiple MSAs by making intermediate stops. Standard errors clustered at the non-directional location are presented between parentheses ($ij$ is the same non-directional location pair as $ji$). R2 is computed as the squared correlation between observed and fitted values.

## E.3. Firms' geographic expansion

To be completed.

# Historical Air Travel Times: United States

Stefan Pauly* Fernando Stipanicic†

May 5, 2022

In this paper we present a novel dataset of air travel times in the United States. We have digitized and web-scrapped information on airlines' flight schedules for 13 years in approximately five year intervals covering the period 1951 to 1999. The dataset contains 755 airports and 11,058 directed airport-pair links. We document that between 1951 and 1999, the big drop in travel time of non-stop flights occurred in the period 1956-1961 with the introduction of jet airplanes. We also document the appearance of long-range flights in 1970. Using airline information, we show the expansion of American Airlines and the development of its hubs. We propose to study the Airline Deregulation Act of 1978 as a policy change that may have driven the change from a point-to-point to a hub-and-spoke network.

*Sciences Po, Ph.D. Candidate. Email: *stefan.pauly@sciencespo.fr*
†Toulouse School of Economics, Ph.D. Candidate. Email: *fernando.stipanicic@tse-fr.eu*

# 1 Introduction

In this paper we present a novel dataset of air travel times in the United States. We have digitized and web-scrapped information on airlines' flight schedules for 13 years in approximately five year intervals covering the period 1951 to 1999. For each scheduled commercial flight we know the airport of origin and destination, scheduled departure and arrival time, airline, aircraft model, frequency and type of food service provided. To our knowledge, this is the first dataset on air travel times covering the second half of the 20th century.[1] In the future we plan to extend the dataset to cover the period 1930-1999. If we take as an observation a combination of origin airport - destination airport - origin time - destination time - airline - flight number - year, the dataset contains 404,536 observations.

In the remaining of the paper we first present how the data was constructed. Second, we present basic descriptive statistics of the data. Third, we present descriptive statistics on the evolution of travel time for airports-pairs connected by non-stop flights and including connecting flights. Fourth, we present descriptive statistics on the network of flights. Fifth, we sketch a research proposal. For additional details on data construction and descriptives on the period 1951-1966 we refer to Pauly and Stipanicic (2022).

# 2 Data construction

The construction of the dataset can be split into two periods: 1951-1970 and 1975-1999. The first period of the dataset has been constructed by collecting historical flight schedules of individual airlines and digitizing them. The second period has been constructed by web-scrapping flight schedules that were initially published in hard-copy by an air travel company.

[1] The most similar dataset that we are aware of is the one of Giroud (2013) which covers the period 1977 to 2005.

For the period 1951-1970 we digitized 6 domestic airlines in this time period: American Airlines (AA), Eastern Airlines (EA), United Airlines (UA), Trans World Airlines (TW), Northwest (NW), Braniff (BN) and Delta (DL).[2] We chose these airlines to maximize both the amount passenger transport covered (AA, EA, UA and TW constituted the group of the *Big 4* and accounted for more than 70% of passenger transport in each year of the period), and the geographic coverage (NW provided service from the East to Oregon and Washington, BN covered the Midwest to Texas, and DL grew quickly connecting Atlanta with other regions). We have digitized the years 1951, 1956, 1961, 1966 and 1970. These years were selected depending on data availability and with the objective to have observations that near-equally spaced over time. Figure 1 is an example of the raw data that was digitized.

For the period 1975-1999 we changed the method of data collection due to the availability of digital resources. We have scrapped the website of a third party who has made public historical records of flight schedules. The data is published by airport of destination. Figure 2 is an example of the raw data that has been scraped. We have collected the years 1975, 1979, 1981, 1985, 1989, 1991, 1995 and 1999.[3]

While the company which has compiled the hard-copy data would in principle provide information on all commercial flights scheduled by all airlines, we are uncertain on whether the website has published all of them. Across years we observe substantial variation in the amount of airports of destination. However, our educated guess is that for all destinations that have been published, all scheduled flights of all airlines are included. To get an approximate solution to this missing-destination problem, in the descriptive statistics we will work with a dataset that we have done symmetric: for all origin-destination flights we have created a fictitious destination-origin flight with the same travel time.

[2]We have also digitized the airline Panamerican which in this time period operated international flights only.

[3]We are in the process of collecting international flight data.

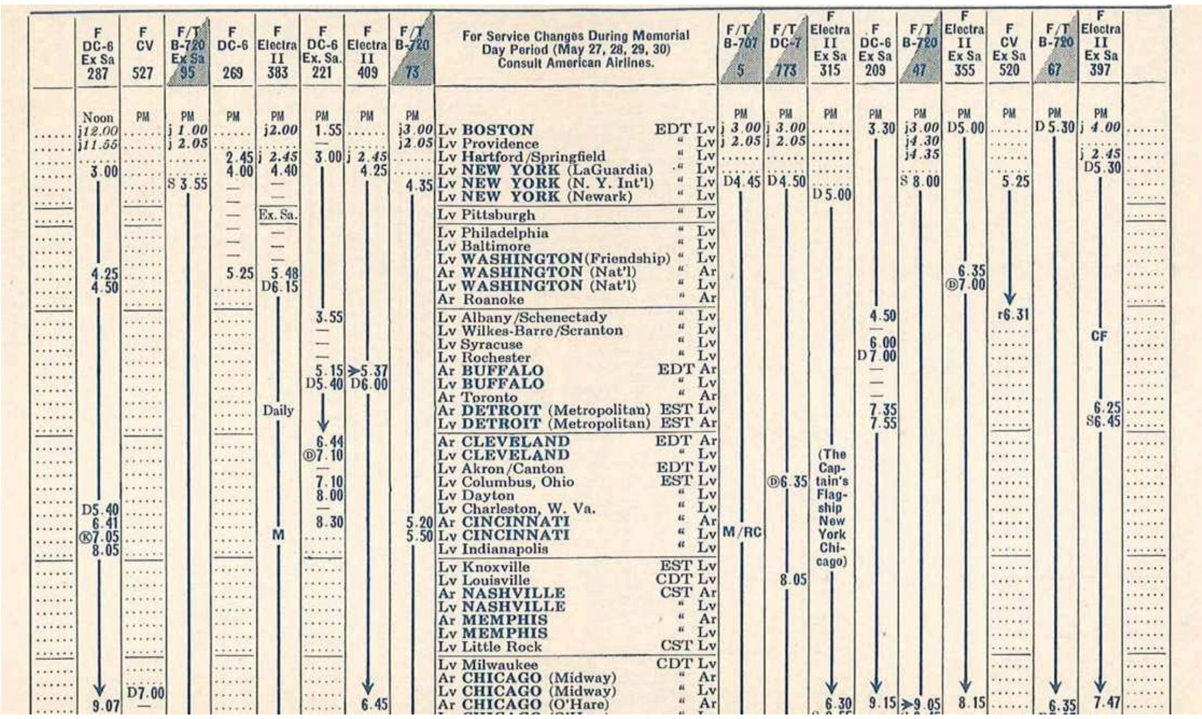



Figure 1: Fragment of flight schedule American Airlines 1961

The center column displays the name of departure and arrival cities. The small columns on the sides display flights with departure and arrival time (local time, bold numbers represent PM). The top of the small columns shows the type of service provided (first class, coach or both), aircraft operated, days operated (daily if information is missing) and flight number.

# 3 Basic descriptive statistics

The dataset contains a multitude of variables for each scheduled flight: flight number, departure time and airport, arrival time and airport, airline, aircraft model, frequency (e.g. daily, Mondays only, etc), food service provided. Currently we have done a proper cleaning of the variables: departure time and airport, arrival time and airport. Hence, information on airlines or flight numbers should be taken with caution.

Table 1 presents simple descriptives. For the period 1951-1970 by construction the dataset contains 6 airlines. For the period 1975-1999 the amount of airlines fluctuates between 117 and 332. However, this number should be taken as an upper bound. As an example, American Airlines appears written in 7 different ways: "America", "American", "aa", "AAA', "AAA Airlines", "Americam", "American'". A proper

**To New York LaGuardia Airport (Page 1 of 4)**
**Effective April 15, 1975**

| Arriving From | Airline | Flight | Departs | Arrives | Meals | Equipment | Frequency |
|---|---|---|---|---|---|---|---|
| Akron/Canton, OH | United | UA 544 | 3:45pm | 4:53pm | S/ | 737 | ExSat |
| Atlanta, GA | Eastern | EA 80 | 5:40am | 7:34am | B | L10 | Daily |
| Atlanta, GA | Delta | DL 188 | 6:11am | 7:59am | B | 72S | Daily |
| Atlanta, GA | Delta | DL 212 | 12:15p | 2:03pm | L | 72S | Daily |
| Atlanta, GA | Eastern | EA 918 | 3:00pm | 5:05pm | S | 727 | MonOnly |
| Atlanta, GA | Delta | DL 120 | 3:02pm | 4:50pm | S | 72S | Daily |
| Atlanta, GA | Eastern | EA 544 | 3:15pm | 5:10pm | S | 72S | Daily |
| Atlanta, GA | United | UA 368 | 4:00pm | 5:59pm | D | 727 | ExSat |
| Atlanta, GA | Delta | DL 200 | 6:00pm | 7:48pm | D | 72S | Daily |
| Atlanta, GA | United | UA 424 | 6:00pm | 8:00pm | D | 727 | ExSat |
| Atlanta, GA | Eastern | EA 100 | 6:32pm | 8:32pm | D | D9S | Daily |
| Atlanta, GA | Delta | DL 128 | 8:22pm | 10:10pm | | 72S | Daily |
| Atlanta, GA | Eastern | EA 432 | 10:15pm | 12:07am | | 727 | Daily |

Figure 2: Example website flight schedule

string cleaning is part of our future tasks.

The dataset contains 755 airports and 11,058 unique directional airport pairs (links) in the whole period 1951-1999. However, this amount of airports and links is not constant across years. This could be due both to the opening of new airports and routes, but could also be due to measurement error, e.g. not having a constant sample of airlines. We balance the dataset in two ways: fixing airports, fixing links and fixing airlines. The dataset with a balanced set of airports has 676 airports and 6,014 links present in all years. If we fix the links, the balanced dataset has 83 airports and 268 links. It could additionally be possible to balance the dataset based on airlines (once we have cleaned the variable), however this is not trivial as airlines may have merged, appeared or disappeared over time.

If we count the amount of flight numbers (which is a combination of airline name and the number of the flight e.g. "AA 1"), we have 76,784 unique flight numbers (if a flight number appears in two years we just count it once). Many of these flight numbers have multiple stops, which leads to a total of 352,808 unique legs (non-stop segments by flight number). If we count the amount of flight numbers and legs in each year and then sum across years, we have 132,281 flight numbers and 404,536 legs. Fig-

| Year | N airlines | N airports | N links | N flights | N legs |
|---|---|---|---|---|---|
| 1951 | 6 | 220 | 1,240 | 953 | 6,596 |
| 1956 | 6 | 215 | 1,654 | 1,493 | 9,640 |
| 1961 | 6 | 194 | 1,566 | 1,443 | 8,696 |
| 1966 | 6 | 173 | 1,570 | 2,021 | 9,980 |
| 1970 | 6 | 164 | 1,984 | 2,948 | 11,790 |
| 1975 | 163 | 541 | 4,316 | 7,275 | 27,032 |
| 1979 | 332 | 593 | 4,914 | 11,226 | 34,064 |
| 1981 | 250 | 601 | 5,064 | 11,644 | 38,046 |
| 1985 | 281 | 575 | 5,208 | 15,345 | 48,584 |
| 1989 | 165 | 534 | 4,992 | 16,579 | 45,628 |
| 1991 | 155 | 511 | 5,296 | 18,718 | 53,732 |
| 1995 | 157 | 517 | 5,436 | 21,390 | 57,408 |
| 1999 | 117 | 432 | 4,858 | 21,246 | 53,340 |
| All years unbalanced | 941 | 755 | 11,058 | 76,784 | 352,808 |
| All years balanced airports | 858 | 676 | 6,014 | 71,663 | 255,781 |
| All years balanced links | 235 | 83 | 268 | 18,899 | 51,832 |

Table 1: Simple count air travel data per year

The table presents the amount of unique airlines, airports, links (directional airport pairs served with a non-stop flight), flight numbers and legs (link served by a flight number). Rows on "All years (...)" presents the amount of unique values of each variable: across all years (unbalanced), across all years with a constant set of airports (balanced airports), across all years with a constant set of links (balanced links).

ure 3 shows each year's geographic distribution of airports in our data. We observe that we have a wide geographical coverage in all years, although since 1975 the increase in amount of airports in our data is noticeable.

## 4 Descriptive statistics: evolution of travel time

In this section we describe the evolution in travel time. We first focus on the fastest flight of every two airports that are connected with a non-stop flight. Second, we present descriptives on travel time between 1950 Metropolitan Statistical Areas (MSAs) allowing for connecting flights.

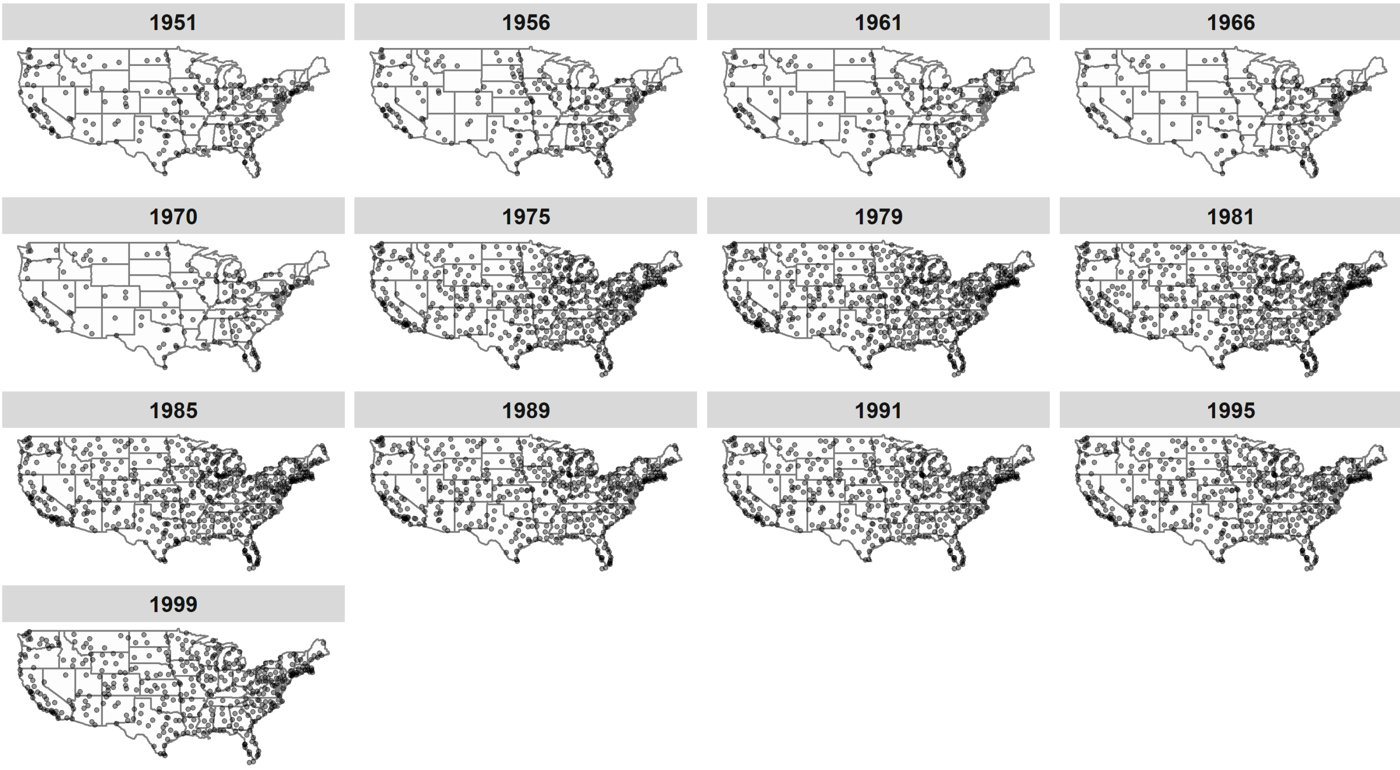



Figure 3: Airports

## 4.1 Non-stop flights

Figure 4 shows, for a selection of years, the fastest flight for every pair of airports that are connected with a non-stop flight. We keep a constant set of airports both for origin and destination, but we allow the set of links (directional airport pair) to evolve. A clear pattern emerges: for the period 1951 to 1999, the big changes in travel time happened between 1951 and 1966. Since 1966 travel times of non-stop flights have been almost constant.[4] The big drop in travel time observed between 1956 and 1961 is the period in which jet airplanes were introduced. Jet airplanes were introduced in late 1958 and had a cruise speed that was around 1,000kmh, almost twice the speed of propeller airplanes used during the 1950s. Between 1961 and 1966, the increased adoption rate of jet airplanes further pushed down travel times. Pauly and Stipanicic (2022) describes in detail the regulation environment and changes in flight technology in the period 1951 to 1966.

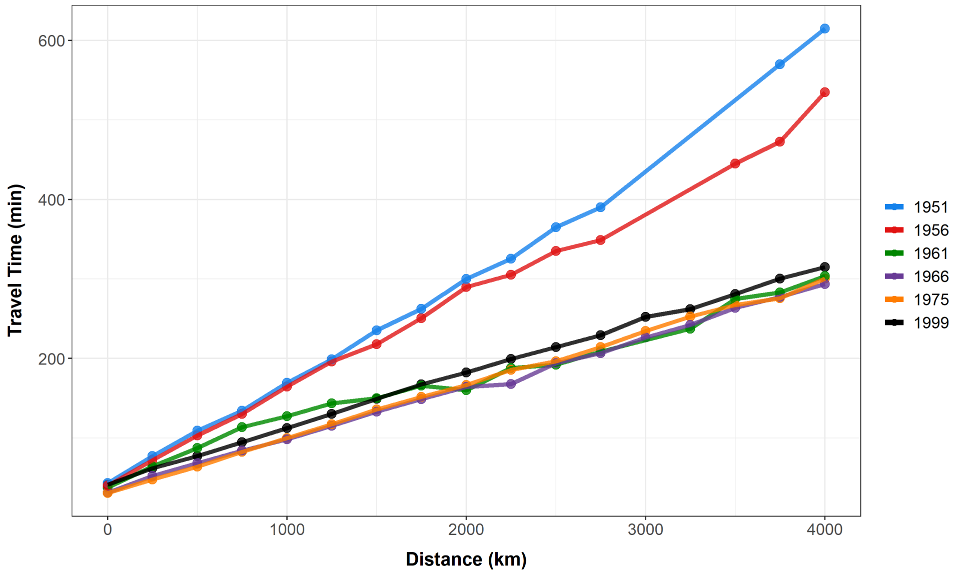



Figure 4: Travel time in non stop flights, balanced set of airports

Table 2 presents the fastest and the longest non-stop flight for each year.[5] In 1951

[4]The small increase in travel time for all distances between 1975 to 1999 may be a signal of a potential measurement or data cleaning error. We plan to review in detail the data of this year to check what is the source of this increase.

[5]We compute the haversine distance between the origin and destinations, which is different from the

the fastest flight was between Ancorage (Alaska) and Seattle, covering 2,326km in 5 hours with an implied average speed of 465kmh. In 1956 the fastest flight had an average speed of 567kmh, in 1961 it went up to 864kmh and in 1966 it further increased to 1,116kmh.[6] Since 1966 the fastest flight had an average speed that fluctuated around 950-1,000kmh.[7]

In 1951 the longest flight was between Honolulu and Portland, covering 4,188km in 13 hours and 5 minutes, with an implied average speed of 320kmh. In 1961 the longest flight was between Ancorage and New York covering 5,434km, while in 1966 it was between Anchorage and Chicago covering 4,593km.[8] In 1970 we observe a big increase in traveled distance of the longest flight: there was non-stop service between Honolulu and New York. This was possible due to the introduction of wide-body long-range airplanes as the Boeing 747. Since 1970 we observe that the longest flight was between Honolulu and another city in either the Midwest or Northeast of continental United States.[9]

## 4.2 Travel times

In order to compute the travel time between two airports we run the Dijkstra algorithm, using as edges the fastest non-stop flight for every airport pair. In this way, we compute the in-flight travel time between two airports of the fastest route. How-

actual flight distance. Airplanes may deviate from the shortest route, for example, to exploit air streams that allow them to go faster and save petrol.

[6] The air stream in Northern US makes it possible to cover routes from West to East much faster than without an air stream.

[7] However, this may not be the case for international flights. In the international data that we are currently collecting we expect to capture commercial flights operated by the supersonic airplane Concord.

[8] Although it is possible that this flight did take place by exploiting the air stream, it may be possible that there was actually a stop in between Ancorage and New York. We have detected multiple errors in 1961 data that we are still in the process of correcting. In the flight schedules airlines provided "full tables" and "quick tables". In the quick tables airlines sometimes did not report all the stops that flights made.

[9] In the international data we expect to capture the introduction of ultra-long-range airplanes in late 1980s and beginning of 1990s. These airplanes were able to cover up to 10,000km non-stop. **??** exploits the discontinuity around 10,000km created by the introduction of such airplanes.

| Year | Longest flight | Minutes | Km | Kmh | Fastest flight | Minutes | Km | Kmh |
|---|---|---|---|---|---|---|---|---|
| 1951 | HNL-PDX | 785 | 4,188 | 320 | ANC-SEA | 300 | 2,326 | 465 |
| 1956 | HNL-PDX | 600 | 4,188 | 419 | SFO-MDW | 315 | 2,977 | 567 |
| 1961 | ANC-JFK | 405 | 5,434 | 805 | SEA-JFK | 270 | 3,887 | 864 |
| 1966 | ANC-MDW | 335 | 4,593 | 823 | ANC-SEA | 125 | 2,326 | 1,116 |
| 1970 | HNL-JFK | 562 | 8,007 | 855 | BWI-OMA | 100 | 1,645 | 987 |
| 1975 | HNL-JFK | 555 | 8,007 | 866 | ACV-MFD | 212 | 3,462 | 980 |
| 1979 | HNL-ORD | 540 | 6,819 | 758 | SEA-CDB | 180 | 2,892 | 964 |
| 1981 | HNL-ORD | 540 | 6,819 | 758 | MKE-MIA | 120 | 2,029 | 1,014 |
| 1985 | HNL-ORD | 470 | 6,819 | 871 | SFO-STK | 100 | 1,678 | 1,007 |
| 1989 | ATL-HNL | 555 | 7,234 | 782 | LMT-MFD | 205 | 3,241 | 949 |
| 1991 | ATL-HNL | 550 | 7,234 | 789 | HNL-STL | 405 | 6,635 | 983 |
| 1995 | ATL-HNL | 490 | 7,234 | 886 | LMT-MFD | 210 | 3,241 | 926 |
| 1999 | EWR-HNL | 550 | 7,973 | 870 | SFO-ATL | 217 | 3,435 | 950 |

Table 2: Longest and fastest flight in each year

ever, this computation does not account for layover time in case of connecting flights. We could potentially account for layover time, as it is done in Pauly and Stipanicic (2022), using the arrival time of one flight and the departure time of the following one. However, for the period 1975-1999, the cross year variation in amount of destinations suggest that the data is incomplete and we have missing destinations in some years. To partially solve the problem we have decided to create a fictitious return flight for each origin-destination that we observe in the data, with a travel time equal to the one observed for the origin-destination. This approach has the advantage that we have a more complete dataset, which probably approximates better the reality as usually origin-destination non-stop flights also have their corresponding non-stop return flight destination-origin. The limitation of this approach is that we do not know the time of departure and arrival of fictitious flights, which means that we cannot compute the layover time when connecting with a fictitious flight. Hence, our fastest route calculation only includes in-flight travel time. We obtain the fastest route and corresponding in-flight travel time by using the extended dataset (observed flights + fictitious flights) and running the standard *static* Dijkstra algorithm.[10]

[10]We refer to the Dijkstra algorithm as *static* Dijkstra algorithm because it does not exploits information on departure and arrival time of each flight, it just uses duration of an origin-destination non-stop flight. We have developed a *dynamic* version of the Dijkstra algorithm that uses actual departure

We match airports to 1950 Metropolitan Statistical Areas (MSAs) by whether they lay inside the MSA or less than 15km away from its border. By using travel time across MSAs instead of airports we can account for within-city across-airport shifting of flight activity (e.g. Chicago Midway to Chicago O'Hare). Figure 5 shows the travel time between a constant set of 1950 Metropolitan Staistical Areas. The figure replicates what was observed in non-stop flights: there was a big drop in travel time between 1956 and 1961, and then travel times remained roughly constant from 1966 onwards. However, the changes in travel time were not uniform across all distances. Figure 6 is an example of the change in travel time for four origin-destination MSAs. The change in travel time is computed as the percentage difference relative to the travel time of the same MSA pair in 1951. All three pairs of New York to Boston, Chicago and Los Angeles were non-stop flights from 1956 onwards (only New York to Los Angeles had one stop in 1951). New York to Boston are 300km apart, New York to Chicago are around 1,150km apart and New York to Los Angeles are 3,970km apart. We observe that for that longer routes had a larger reduction in travel time. The drop is observed very clearly between 1956 and 1961. The pair Tampa-Portland is presented as an example of other MSA-pairs which required multiple flights to connect each other also observed a big drop in travel time. Tampa-Portland are located 4,010km apart. In 1951 the fastest route included 3 stops (Atlanta, Chicago and Salt Lake City) and took 10 hours 58 minutes. In 1999 the fastest route included 1 stop (Denver) and took 5 hours 31 minutes.

# 5 Descriptive statistics: flight network

Figure 7 shows the flight network of American Airlines (including all its 7 different labellings). We observe an expansion of airports in which American Airlines operates. Note for example the increase of American Airlines operations in Dallas and Miami

and arrival time of each flight number, hence better approximating total travel time that would have taken to passengers by including layover time. The *dynamic* Dijkstra algorithm is used in Pauly and Stipanicic (2022)

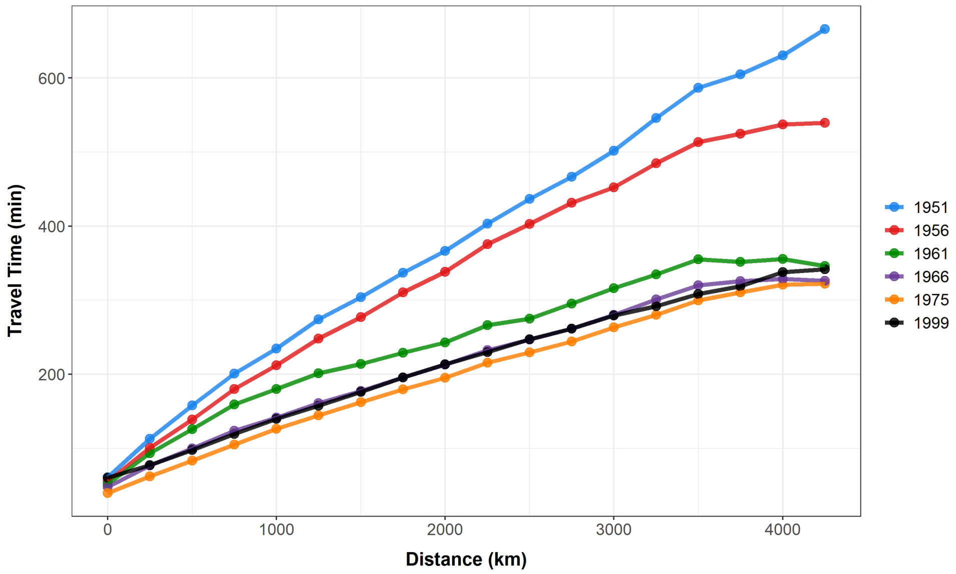



Figure 5: Travel time between 1950s MSAs, balanced set of MSAs

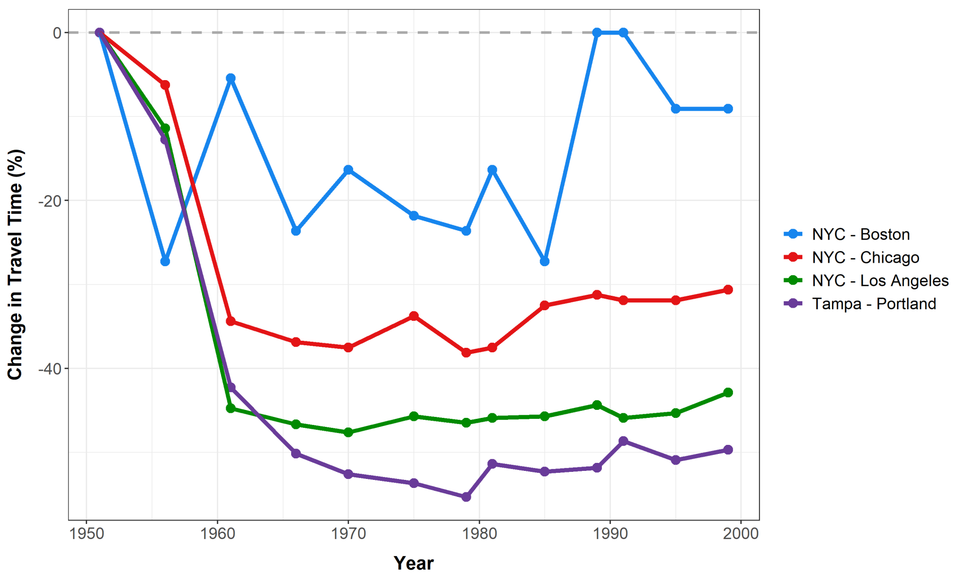



Figure 6: Example change in travel time between MSAs

during the 1980s, and an expansion of the service to Portland and Seattle during the 1990s.

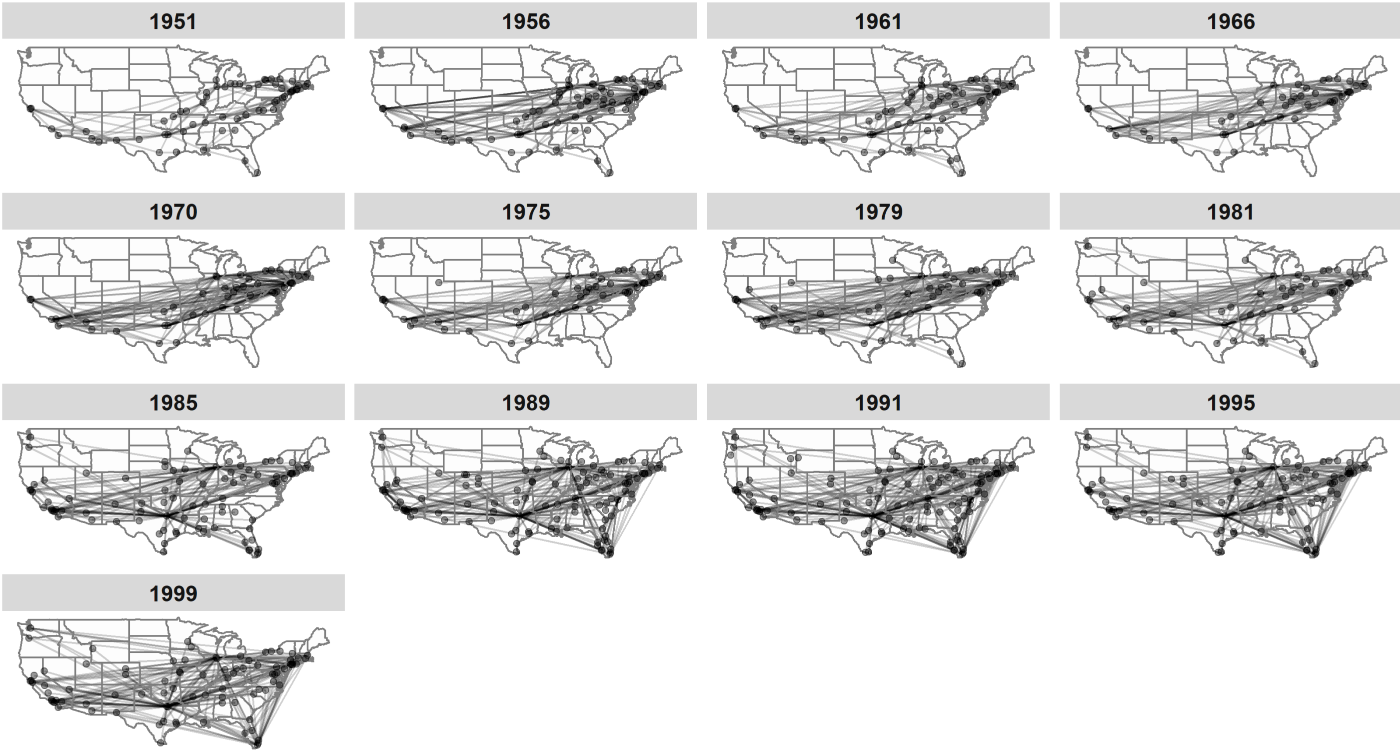



Figure 7: American Airlines flight network

Table 3 shows in the second column the share of flight numbers of the top 10 airports ranked by amount of flight numbers.[11] We observe that in 1951 the top 10 airports were either the origin or destination of 36% of non-stop flights. In 1999 the share was 27%. However, we note that the data collection procedure is likely to be one of the reasons for the decrease in the share: in 1970 the share was 39% while in 1975 it was 29%. The third column shows the share of links (unique origin-destination airport pairs) of the top 10 airports ranked by amount of links. Top 10 airports accounted by

[11] As the data has been done symmetric, this is the share of flight numbers in which the airport appears as origin or destination airport.

20% of links in 1951 while they accounted for 22% of links in 1999. The data collection procedure may also affect the share in this case: in 1975 and later we have smaller airlines in the data which potentially serve smaller and remote airports which would drive down the share of the top 10 airports. The increase in amount of links shown in Table 1 is consistent with this explanation.

| Year | Share flights | Share links |
|---|---|---|
| 1951 | 0.36 | 0.20 |
| 1956 | 0.37 | 0.22 |
| 1961 | 0.34 | 0.23 |
| 1966 | 0.39 | 0.26 |
| 1970 | 0.39 | 0.27 |
| 1975 | 0.29 | 0.18 |
| 1979 | 0.26 | 0.18 |
| 1981 | 0.27 | 0.18 |
| 1985 | 0.26 | 0.18 |
| 1989 | 0.26 | 0.19 |
| 1991 | 0.24 | 0.19 |
| 1995 | 0.25 | 0.20 |
| 1999 | 0.27 | 0.22 |

Table 3: Share flights and links top 10 airports

Table 4 shows the 3-digit IATA code of the top 10 airports by amount of links in each year (these airports correspond to the third column in 3). Measured by amount of links, Chicago continuedly had the most connected airport in our data: it was Chicago Midway (MDW) until 1956 and Chicago O'Hare (ORD) since 1961 onwards. We also observe the transition from Dallas Love Field (DAL) to Dallas Fort Worth (DFW). The table displays the decay of Washington Dulles (DCA) and all three New York airports (Kennedy JFK, La Guardia LGA and Newark EWR), while we observe the raise of Atlanta (ATL), Denver (DEN), Minneapolis Saint Paul (MSP), Cincinnati/Northern Kentucky (CLT) and Charlotte North Caroline (CVG). While shifts as the one of Chicago O'Hare, Dallas Fort Worth or Atlanta may be a reflection of flight activity, we are cautious about other trends that may be correlated with the change in the data collection procedure.

| Airport rank | 1951 | 1956 | 1961 | 1966 | 1970 | 1975 | 1979 | 1981 | 1985 | 1989 | 1991 | 1995 | 1999 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | MDW | MDW | ORD | ORD | ORD | ORD | ORD | ORD | ORD | ORD | ORD | ORD | ORD |
| 2 | LGA | DCA | DCA | JFK | ATL | ATL | ATL | ATL | ATL | ATL | DFW | ATL | ATL |
| 3 | DCA | LGA | ATL | DCA | JFK | DFW | DEN | DEN | DFW | DFW | ATL | DFW | DFW |
| 4 | DAL | ATL | JFK | ATL | LAX | DEN | DFW | PIT | DEN | DEN | PIT | DEN | MSP |
| 5 | EWR | LAX | MDW | EWR | EWR | DCA | PIT | DFW | STL | PIT | DEN | PIT | PIT |
| 6 | CLE | EWR | EWR | CLE | LGA | LGA | LAX | LGA | PIT | MSP | CLT | MSP | DEN |
| 7 | LAX | JFK | DAL | DAL | DCA | LAX | STL | STL | LAX | CLT | STL | CLT | CVG |
| 8 | SFO | DAL | CLE | LAX | CLE | STL | BOS | DCA | LGA | DTW | MSP | STL | DTW |
| 9 | AGC | SFO | LAX | PIT | STL | BOS | LGA | LAX | EWR | BWI | DTW | DTW | STL |
| 10 | ATL | CLE | PIT | STL | PHL | JFK | DCA | BOS | BOS | EWR | PHL | CVG | CLT |

Table 4: Top 10 airports by amount of links

Table 5 shows the top 10 links (airport pairs connected by a non-stop flight) ranked by the amount of flight numbers. In 1951, the link that had the biggest amount of flights operating between them was New York's La Guardia airport (LGA) to Washington DC's Dulles airport (DCA). The link San Francisco (SFO) to Los Angeles (LAX) appears in the top 10 links in all years. Table 6 shows the share of flight numbers of each of the links. We observe that the share of the top 10 links was 12.9% in 1951, 7.8% in 1970, 4.2% in 1975 and 4.0% in 1999. While in Table 3 we observe that the share of flights by airport of origin/destination increased between 1951 and 1970, Table 6 shows that the share of flights within top links decreased.

| Link rank | 1951 | 1956 | 1961 | 1966 | 1970 | 1975 | 1979 | 1981 | 1985 | 1989 | 1991 | 1995 | 1999 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | LGA-DCA | LGA-DCA | SFO-LAX | SFO-LAX | ORD-LGA | SFO-LAX | ONT-LAX | SAN-LAX | SFO-LAX | SEA-PDX | SAN-LAX | OGG-HNL | SAN-LAX |
| 2 | LGA-BOS | LGA-BOS | ORD-JFK | SEA-PDX | ORD-MSP | SJC-OAK | HOU-DAL | SFO-LAX | OGG-HNL | SFO-LAX | SEA-PDX | SAN-LAX | SEA-PDX |
| 3 | MDW-LGA | MDW-DTW | SEA-PDX | STL-MKC | SEA-PDX | SEA-PDX | ORD-LGA | ONT-LAX | SEA-PDX | OGG-HNL | SFO-LAX | SEA-PDX | LAX-LAS |
| 4 | SFO-OAK | DAY-CMH | EWR-DCA | ROC-BUF | SFO-LAX | ORD-LGA | IAH-DFW | IAH-DFW | SAN-LAX | SAN-LAX | HOU-DAL | LAX-LAS | OGG-HNL |
| 5 | YIP-MDW | SEA-PDX | SAN-LAX | ORD-LAX | LAX-JFK | LAX-LAS | SNA-LAX | LAS-GCN | HOU-DAL | LIH-HNL | OGG-HNL | SFO-LAX | SFO-LAX |
| 6 | DCA-BWI | MDW-LGA | DAY-CMH | SFO-ORD | ORD-CLE | ORD-MSP | SLC-DEN | HOU-DAL | HYA-ACK | PHX-LAX | PHX-LAX | LIH-HNL | MIA-MCO |
| 7 | PHL-EWR | SFO-LAX | DCA-BWI | HOU-DAL | ORD-LAX | IAH-DFW | ORD-MSP | ORD-MSP | STL-ORD | HOU-DAL | PHX-LAS | PHX-LAX | JFK-BOS |
| 8 | SFO-LAX | SFO-OAK | TPA-MIA | ORD-JFK | ORD-EWR | ORD-DCA | SFO-LAX | SEA-PDX | EWR-BOS | SEA-GEG | LIH-HNL | ORD-MSP | PHX-LAX |
| 9 | DAY-CMH | STL-MKC | BOS-BDL | LAX-JFK | LGA-BOS | ORD-MKE | MIA-JFK | ORD-LGA | PHX-LAX | ORD-LGA | SAN-PHX | IAH-DFW | TPA-MIA |
| 10 | MDW-LAX | PHL-LGA | STL-MKC | SAN-LAX | HOU-DFW | PHL-DCA | ORD-MKE | OGG-HNL | DEN-COS | ORD-MSP | ORD-MSP | PHX-LAS | ORD-MSP |

Table 5: Top 10 airport pairs by flight numbers

| Link rank | 1951 | 1956 | 1961 | 1966 | 1970 | 1975 | 1979 | 1981 | 1985 | 1989 | 1991 | 1995 | 1999 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.023 | 0.016 | 0.015 | 0.014 | 0.013 | 0.006 | 0.004 | 0.007 | 0.006 | 0.007 | 0.006 | 0.006 | 0.006 |
| 2 | 0.019 | 0.015 | 0.011 | 0.011 | 0.009 | 0.005 | 0.004 | 0.006 | 0.005 | 0.006 | 0.005 | 0.005 | 0.005 |
| 3 | 0.017 | 0.010 | 0.011 | 0.009 | 0.009 | 0.005 | 0.004 | 0.004 | 0.004 | 0.005 | 0.005 | 0.004 | 0.004 |
| 4 | 0.015 | 0.010 | 0.008 | 0.008 | 0.007 | 0.005 | 0.004 | 0.004 | 0.004 | 0.005 | 0.005 | 0.004 | 0.004 |
| 5 | 0.010 | 0.010 | 0.008 | 0.008 | 0.007 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.003 | 0.004 |
| 6 | 0.009 | 0.009 | 0.008 | 0.008 | 0.007 | 0.004 | 0.003 | 0.003 | 0.003 | 0.004 | 0.004 | 0.003 | 0.004 |
| 7 | 0.009 | 0.009 | 0.008 | 0.008 | 0.007 | 0.004 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 |
| 8 | 0.009 | 0.009 | 0.007 | 0.008 | 0.006 | 0.004 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 |
| 9 | 0.008 | 0.008 | 0.006 | 0.008 | 0.006 | 0.004 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 |
| 10 | 0.008 | 0.008 | 0.006 | 0.007 | 0.006 | 0.004 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 |
| Total | 0.129 | 0.105 | 0.088 | 0.088 | 0.078 | 0.042 | 0.035 | 0.040 | 0.039 | 0.042 | 0.040 | 0.037 | 0.040 |

Table 6: Share of flights top 10 airport pairs

Figure 8 displays the log rank of an airport by amount of links against the airport's log amount of links. The airport in rank 1 is the airport with the largest amount of links. The change in the data collection procedure, by increasing the amount of airlines, airports and links, would mechanically shift the curves to the upper-right part of the plot. However, in the figure we observe that over time there is a change in the slope of the log rank-log size distribution. For example, relative to 1975, in 1999 we observe a sharper decrease in the amount of links for the less connected airport. At the same time, most connected airports remain equally or more connected airports. The same pattern is repeated in 1970 relative to 1951.

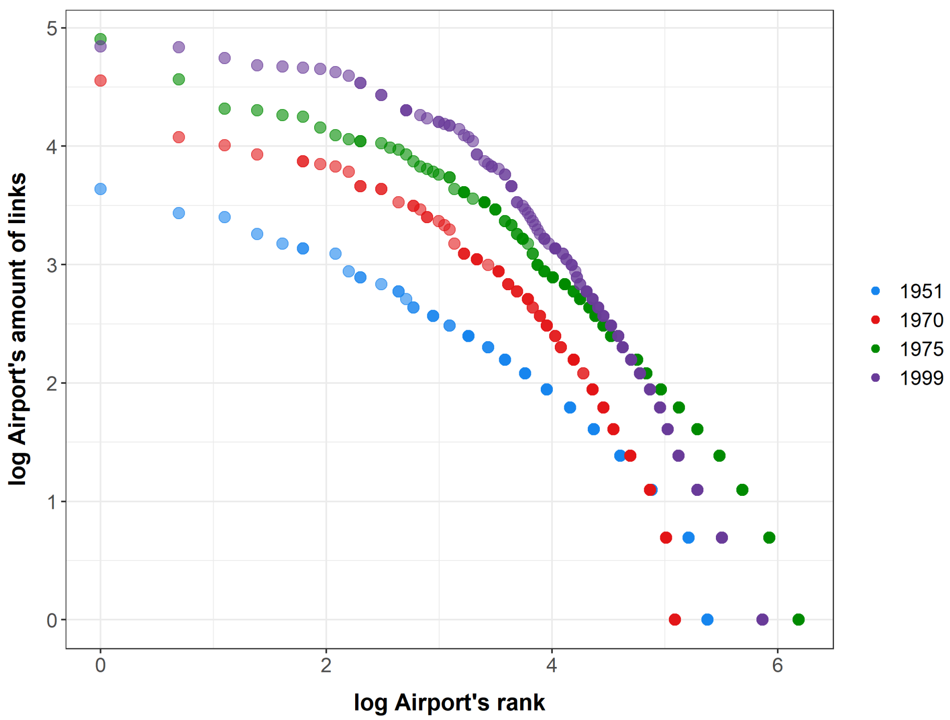



Figure 8: Log rank - log size of airports by amount of links

Figures 9 and 10 present the quantiles of airport's outdegree centrality. Outdegree centrality measures the amount of destination airports that an origin airport has, relative to the maximum amount of destinations it could have. The first figure contains the period 1951-1970, while the second one contains 1975-1999. In both figures we observe that the 90th quantile of outdegree centrality increases, while the lower quantiles do not increase as much. This implies that most connected airports became more connected, and even more relative to less connected airports.

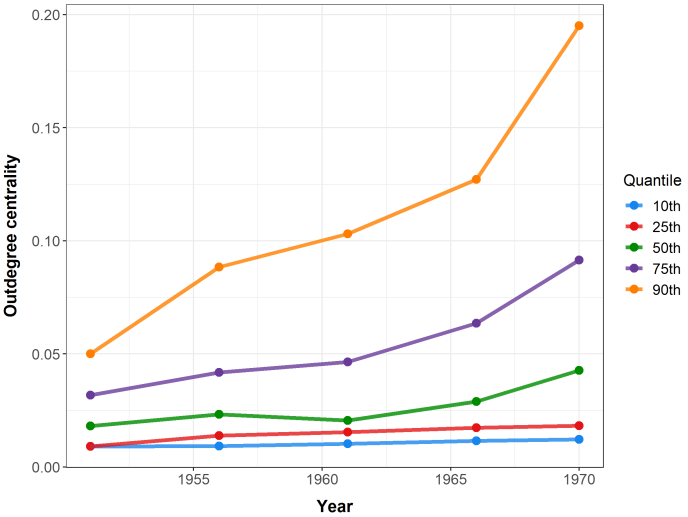



Figure 9: Quantiles of outdegree centrality until 1970

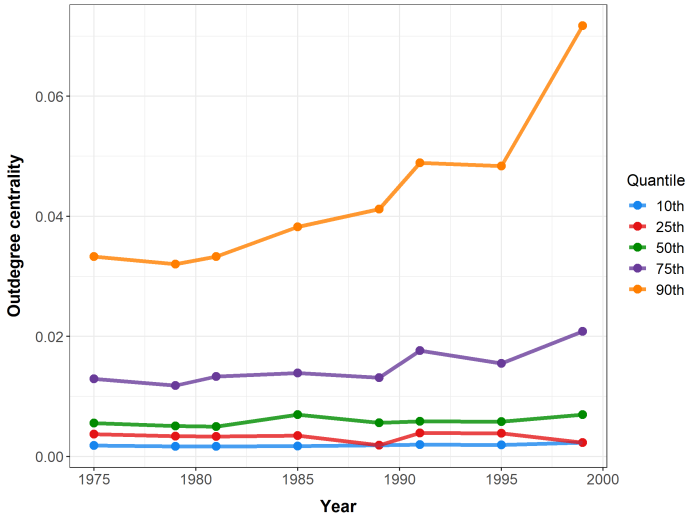



Figure 10: Quantiles of outdegree centrality since 1975

Figure11 is constructed using the fastest route between every two airports. The figure shows the cumulative distribution of the amount of legs needed to get from one airport to another in the fastest route. We observe a shift to the upper left part of the figure, meaning that over time less legs were required to get from one airport to another.

Figures 12 to 19 present the distribution of airports' outdegree centrality for varying set of airports and links used to compute the centrality value. Figures 12 and 13 show the outdegree centrality for all airports present in each year. In 1970 we observe that there is a larger mass of more central airports relative to 1951. Although less salient, the same is true in 1999 relative to 1975. Figures 14 and 15 repeat the computations but using only airports and links that are served by American Airlines.[12] While in 1970 relative to 1951 we observe the same pattern as using all airports and flights, we do not observe the same in 1999 relative to 1975. In 1999 we observe an increase in the mass

[12]Hence, we compute centrality only for airports that are operated by American Airlines only using links that are served by American Airlines. In other words, we compute an airport's centrality within American Airlines flight network.

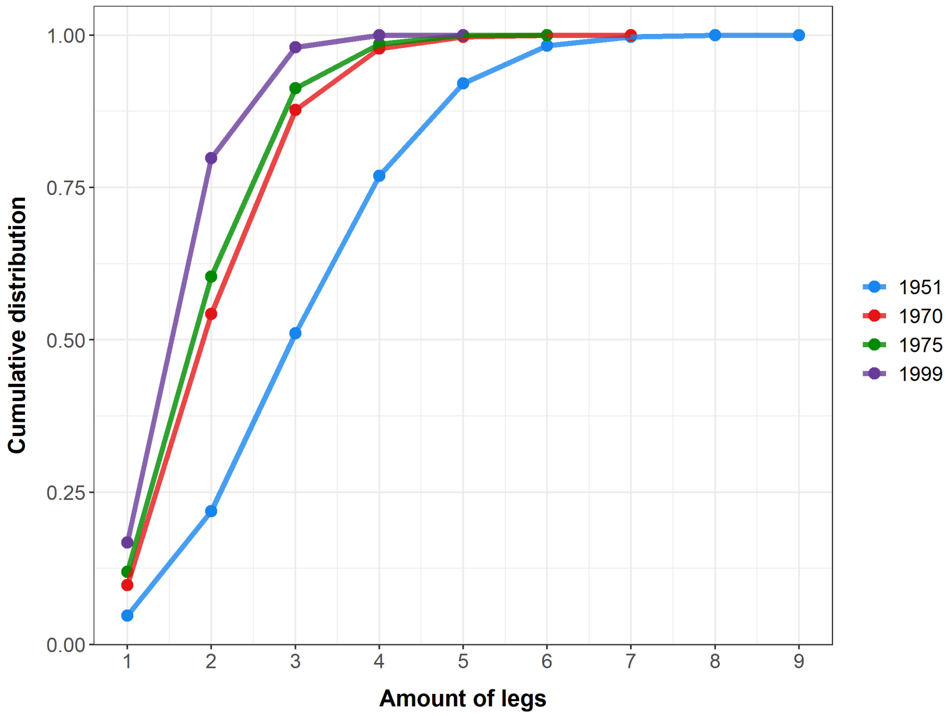



Figure 11: Cumulative distribution amount of legs fastest route within airport pairs

of low centrality airports and the appearance of super-central airports in American Airlines flight network: Dallas Fort Worth (centrality = 0.85), Chicago O'Hare (0.55) and Miami International (0.25).

We may be concerned that the change in the data collection procedure highly drives the airport's centrality. Hence, we could repeat the analysis fixing the set of airports and allowing the links to evolve. Figures 16 and 17 are the counterpart to 12 and 13. We observe the same patterns in 1970-1951 and 1999-1975 relative to the the centrality values using all airports. Figures 18 and 19 present a pattern that goes in the same direction as the one observed in Figures 14 and 15.

# 6 Research proposal: deregulation of the airline market

Between 1938 and 1978 the US airline market was under strict regulation by the Civil Aeronautics Board (CAB). Among other mandates, the CAB determined the routes in

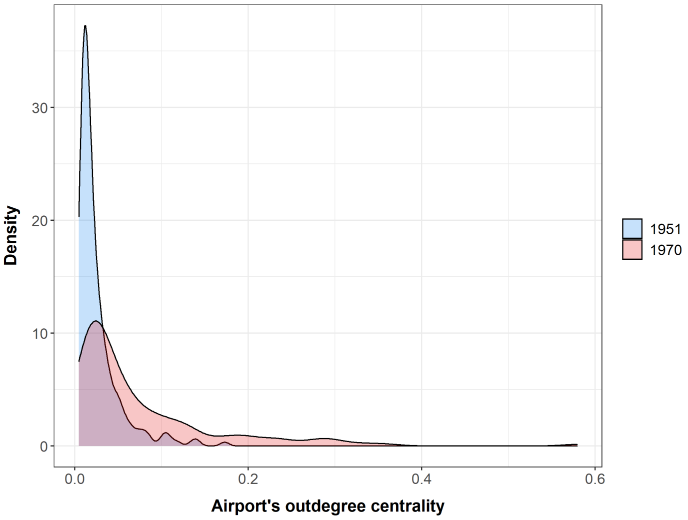



Figure 12: Airports' outdegree centrality allowing airports to change

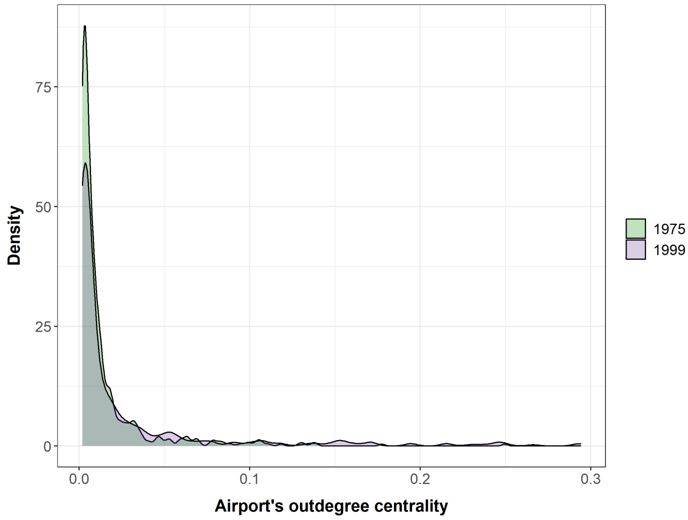



Figure 13: Airports' outdegree centrality allowing airports to change

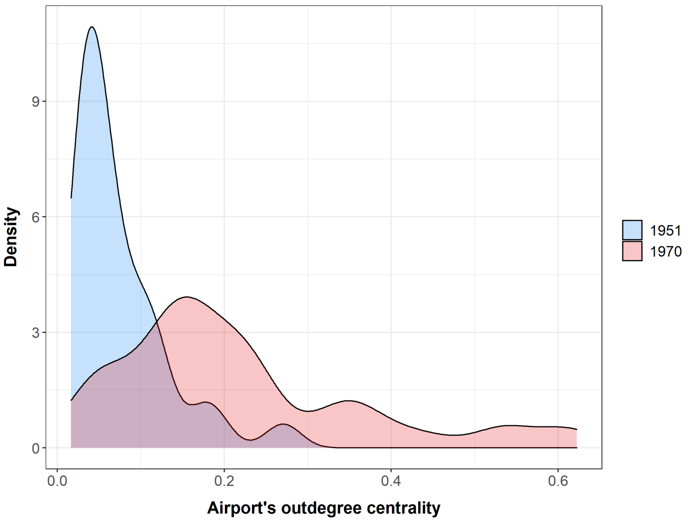



Figure 14: Airports' outdegree centrality allowing airports to change, only American Airlines flights

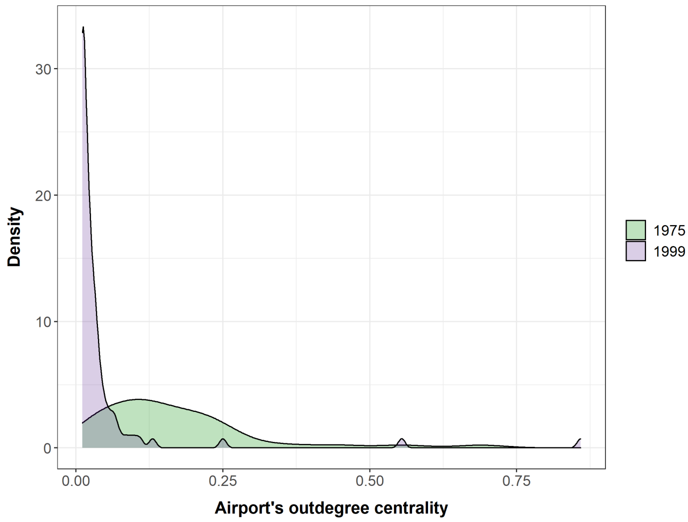



Figure 15: Airports' outdegree centrality allowing airports to change, only American Airlines flights

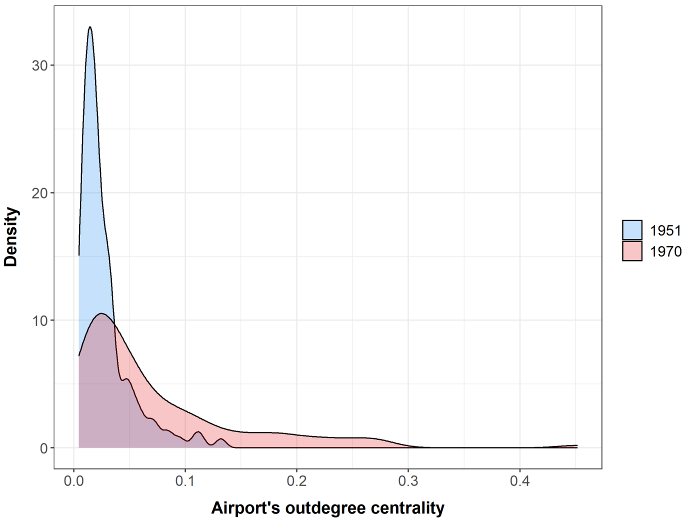



Figure 16: Airports' outdegree centrality fixing airports

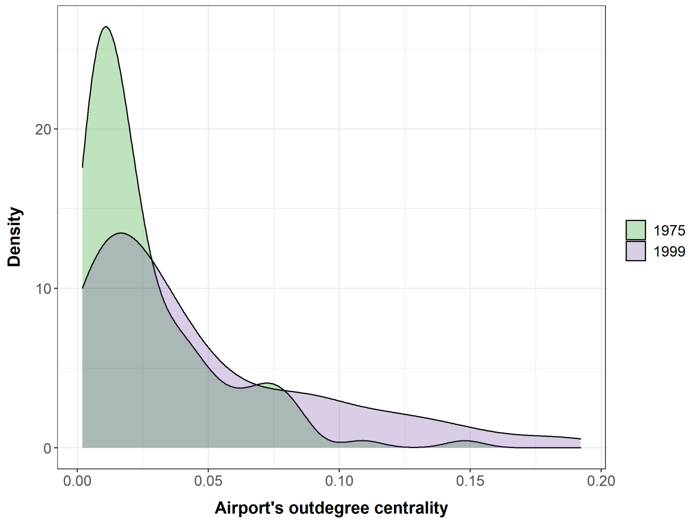



Figure 17: Airports' outdegree centrality fixing airports

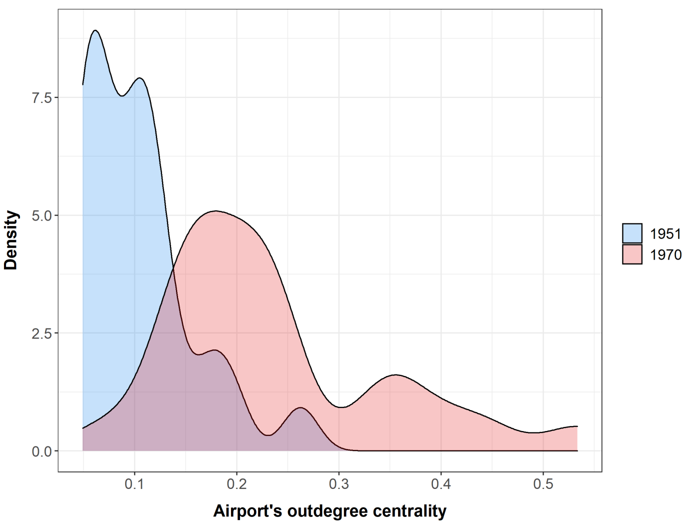



Figure 18: Airports' outdegree centrality fixing airports, American Airlines

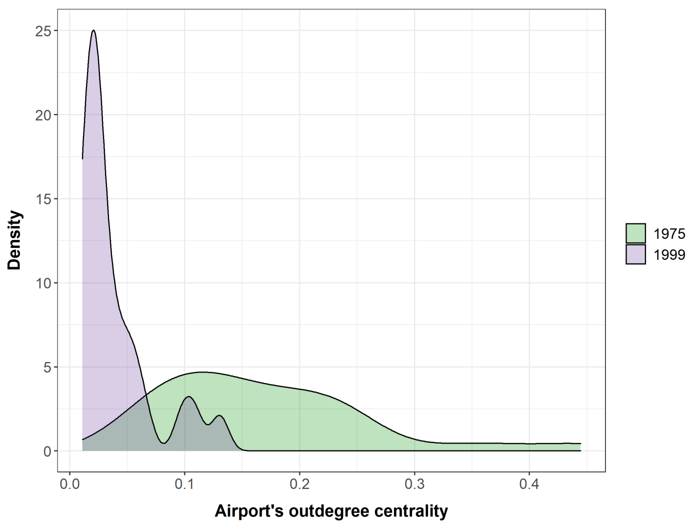



Figure 19: Airports' outdegree centrality fixing airports, American Airlines

which each airline operated and the prices that they charged. In this manner, the CAB controlled both an airline's network and the system network. In 1978, US president James Carter signed the Airline Deregulation Act which gradually removed restrictions to airlines' over the following four years. By 1981 restrictions of entry into domestic routes were eliminated, and by 1983 regulation on domestic fares (ticket prices) eliminated. In 1985 the CAB ceased to exist.

We plan to study how the deregulation changed the flight network and its welfare gains or losses. The deregulation of the airline market implied the switch from an optimization problem from a centralized to a decentralized perspective. While the objective of the CAB was to keep a *stable airline industry*, we could also think of it as optimizing the whole US flight network. On the other hand, post deregulation each airline was optimizing its own flight network.

Previous literature shows that post deregulation the US airline market transformed from a point-to-point network into a hub-and-spoke network (McShan and Windle (1989)). This shift is argued to be consequence of the existence of economies of scale at the airline level (Brueckner et al. (1992)). Although in our descriptive analysis we do not find strong evidence at the US-wide network level of a shift towards hub-and-spoke, Figure 15 suggests that this may have been the case for American Airlines.

One of the argued benefits of hub-and-spoke networks versus point-to-point is that it reduces the amount of legs required to travel between two points. The comparison between years 1975 and 1999 in Figure 11 is suggestive evidence that this was indeed what happened. However, Figure 5 shows that the decrease in legs did not translate into a reduction of in-flight travel time.[13]

[13]Nonetheless, the decrease in legs may have led to a decrease in total travel time through a reduction of layover time. As noted in Section 2, our source of data does not contain all destinations for each origin. We circumvented this issue by making the dataset symmetric: all origin-destination have an equal destination-origin with the same flight duration. The disadvantage of this approach is that we do not have information on departure and arrival times needed to compute layover time.

The constructed dataset allows us to study the gains or losses in terms of travel time. We plan to merge this dataset with information on origin-destination passenger transport which includes the route taken by each passenger. This dataset would allow us to observe market size and shifts of flows through different routes for the same origin-destination.

By further matching the dataset with production data by industry, location and year, as for example with the *County Business Patterns* data, we would be able to study the impact of changes of the flight network on local economic activity. Airlines, by shifting towards hub-and-spoke networks, may have led to an increase in the concentration of economic activity in hub cities relative to spokes.

# 7 Conclusion

In this paper we presented a novel dataset of air travel times in the United States for the period 1951 to 1999. The dataset contains information for each scheduled commercial flight on the airport of origin and destination, scheduled departure and arrival time, airline, aircraft model, frequency and type of food service provided. While the dataset still needs improvements, we believe it will allow to answer a broad set of questions ranging from flight network analysis, the industrial organization of airlines, competition and innovation, migration, macroeconomic impacts of connectivity, and more.

# References

Brueckner, J. K., N. J. Dyer, and P. T. Spiller (1992). Fare determination in airline hub-and-spoke networks. *The RAND Journal of Economics*, 309–333.

Giroud, X. (2013, 03). Proximity and Investment: Evidence from Plant-Level Data *. *The Quarterly Journal of Economics 128*(2), 861–915.

McShan, S. and R. Windle (1989). The implications of hub-and-spoke routing for airline costs. *Logistics and transportation review 25*(3), 209.

Pauly, S. and F. Stipanicic (2022). The Creation and Diffusion of Knowledge: Evidence from the Jet Age.

# The Geography of Innovation: France 1978-2010

Gabrielle Gambuli* Fernando Stipanicic†

May 5, 2022

In this paper we present a series of facts on the geographic distribution of patenting activity in France in the period 1978 to 2010. During this period we observe that the share of multi-inventor patents that have inventors in multiple departments increased from 38% to 47%. At the same time, the share of those patents with at least one inventor in Paris decreased from 35% to 30%. Therefore, across-department patent collaboration increased at a national level but diverted away from Paris. We also observe that the share of inventors that year-on-year migrate to or from Paris increases over time. These phenomena happened simultaneously with the expansion of high speed railways. We construct a new dataset of yearly train travel time across departments' capitals. Between 1980 and 2018, the introduction of high speed railways led to an average decrease in travel time of more than 20% for capital-pairs located more than 400km apart. We plan to study how changes in travel time affect inventors' collaborations, inventors' teams characteristics and inventors' mobility.

*Ph.D. Candidate CY Cergy Paris Universitè
†Ph.D. Candidate Toulouse School of Economics

# 1 Introduction

In this paper we present a series of facts on the geographic distribution of patenting activity in France in the period 1978 to 2010. We focus on inventors' collaboration networks, inventors' mobility and the technological composition of inventors' teams and cities. We then present a new dataset on travel times by train which we are constructing. This dataset includes the introduction of high speed railways which led to large reduction in travel time for connected cities.

Currently the analysis is in early stages and we present just a preview of it. We additionally include a summary of research questions on which these datasets together would be able to provide new insights. The descriptive analysis will be used to guide future research projects, exploiting the introduction of high speed railways as a tool to provide causal evidence. In the future we plan to include French administrative data of firms which we would match to patent assignees.[1]

We exploit the patent dataset of Morrison et al. (2017) who have geo-localized inventors based on their residential address and constructed unique inventor identifiers. Using this dataset we identify patents that have multiple inventors and the department of each inventor.

Between 1978 and 2010 the share of patents with more than one inventor, which we label collaborative patents, went from 38% to 61%. At the same time, we observe that the share of collaborative patents that have inventors in multiple departments increased from 38% to 47%. However, the changes in across-department patent collaboration were not homogeneous for all departments. The share of across-departments collaborative patents which involve at least one inventor in Paris decreased from 35% to 30%, while the counterpart share of across-department collaborative patents which involved only inventors located in other departments increased from 65% do 70%.[2]

[1] We have already been granted access to the confidential administrative data of France.

[2] This behavior is not explained by migration of inventors from Paris to the departments bordering

This decrease happens simultaneously with a decrease of the share of patents in Paris in all patents in France from 26% to 15%.[3] Hence, we observe a shift in both patenting activity and across-department collaboration away from Paris.

Our analysis will aim to study if there are potential network effects of collaborations, leading to a potential collaboration-diversion, similar to the potential trade-diversion effect of trade agreements. We will also study if the development of the high speed railway (HSR) may have contributed to this phenomena. Preliminary descriptive analysis shows that following the HSR connection of a department to Paris, the share of the department's collaboration with Paris increases. However, once the department gets connected with other departments, the share of the department's collaboration with Paris decreases.

The inventor identifier of Morrison et al. (2017) accounts for inventor mobility. This is a distinct characteristic of this dataset which allows us to follow the inventor's location across different years in which she patents. For inventors that patent in at least two years, the share of inventors who change department (*move*) at least once is 14.4%. Of those inventors who move at least once, 20% move at least one more time. Of those who move two times or more, 50.8% of them return to the department declared in the address of their first patent. In each year, 15% inventors who patent are not natives from the department in which they patent.

In the future we will seek for potential heterogeneity in the movement of inventors across departments and the relation with reductions in travel time. Paris appears as having both a high share of immigrant inventors and inventors who emigrated, with a yearly shares of inventors who immigrate and emigrate which are around 10% and

Paris. If we create a fictitious department called "Paris extended" which is constituted by Paris and its bordering departments (Hauts de Seine, Seine Saint Denis and Val de Marne), the share of across-departments collaborative patents which involve at least one inventor in "Paris extended" decreased from 60% to 46%. Note that to obtain this number we have used the fictitious departments to compute the amount of across-department collaborative patents.

[3]The share of all collaborative patents -both within department and across departments- with at least one inventor in Paris went from 33% to 19%.

15% respectively. We also plan to describe the distinct characteristics of those inventors who are highly mobile, characteristics of departments that have a high immigration, emigration and overall turnover rate.

Another angle of future research is to study the dynamics of technological composition of collaborative teams and departments. Changes in the costs to collaborate and migrate may trigger compositional changes, leading to more diverse or specialized teams and departments. It may also affect the joint dynamics: a pair of departments with large improvements in connectivity may become more similar to each other. Preliminary descriptive analysis shows little correlation between the technological similarity of two departments and their share of joint collaborations.

We plan to use the opening of high-speed railways (which in this paper we refer indistinctevely as *Train à Grande Vitesse* (TGV) or *Ligne à Grande Vitesse* (LGV)) as a shifter in the communication and migration costs between inventors in different departments.[4] In 1981 the first TGV line opened to connect Lyon and Saint Florentin, to later reach Paris in 1983. Since then many other important cities of France were connected with Paris: Lille, Marseille, Bordeaux, Montpellier.

In this paper we present a new dataset of travel times by train from 1980 to 2018. According to our estimations, the average operating speed of TGV is 220kmh, 83% faster than the operating speed of Intercités trains. Using the opening dates of TGV lines and the existing train line network in 2021 we create a dataset of train travel times between all prefecture-pairs for each year between 1980 and 2018.[5] Our dataset, which is still in a preliminary stage, suggests that the TGV network laid out until 2018 led to an average decrease in travel time of between 5% and 33%, with a larger reduction for prefectures located further apart. We validate our constructed dataset by comparing

[4]In this paper whenever we refer to high speed trains (*Train à Grande Vitesse*) we are actually referring to high speed trains operating in high speed railways (*Ligne à Grande Vitesse*). High speed trains also operate on non-high speed railways and in such cases they do not run at high speed.

[5]Prefectures are the capital city of each department. In our patent data 87% of inventors reside in the urban area of Prefectures.

with an external dataset that contains yearly travel times for a sub-sample of city pairs. Our constructed dataset can explain 83% of the within city-pair change in travel time of this sub-sample.

The rest of the paper is organized as follows. First we present the source of patent data. Second we present descriptives on the geography of patenting activity: patents and patenting growth, inter-regional collaboration, technological specialization, inventors' teams and inventors' mobility. Third, we present the train travel time data set. Last, we conclude.

## 2 Patent data: sample selection

The source of our patent data is Morrison et al. (2017). From their data we select patents filed at the European Patent Office (EPO). We first subset patents with assignees located in France using the file `LinkedAssigneeNameLocData.txt`. We then obtain each patent's application year from the file `all_disambiguated_patents_withLocal.txt` and the patent's inventors' location from `LinkedInventorNameLocData.txt`. We drop all inventors which Morrison et al. (2017) recognize as being having geo-coded with low quality.[6] We then match inventors' geo-coordinates to French departments (NUTS3 classification) by intersecting geo-points with a shape file.

Figure 1 shows the amount of patents per year according to each of the steps of the data slicing. The dataset from 1978 to 2010 has 176,268 EPO patents assignees located in France. Among these patents, we drop 17,757 patents which only have inventors located abroad. We also drop patents for which all inventors are unlocalized or are being localized with low quality. We observe a big drop in the amount of patents for patents

[6]We prefer to discard inventor-patent observations that are low quality geo-coded given that they are highly likely to include measurement error. Inventors who are geo-coded with low quality for a certain patent frequently appear in multiple locations within a patent. The inventor's address is the one declared to the EPO by the inventor in order for the EPO to send official documentation in a physical form. This is supposed to be the residential address of the inventor and it is unlikely -or maybe not even allowed- that the inventor declares more than one address.

with application year after 2010. This is probably due to the fact that Morrison et al. (2017) uses EPO data from 2014. Our analysis is based on a total of 145,914 patents from 1978 to 2010, with inventors localized at the NUTS3 level (red line in Figure 1).

Figure 2 shows the evolution of the amount of patents from the selected dataset over time. We label *co-patents* to patents that have multiple inventors. *intra-regional co-patents* and *inter-regional co-patents* are respectively patents with all inventors within the same department and with inventors in multiple departments. We observe an increase in the amount and the share of co-patents on all patents, going from around 50% in 1980 to around 70% in 2010.

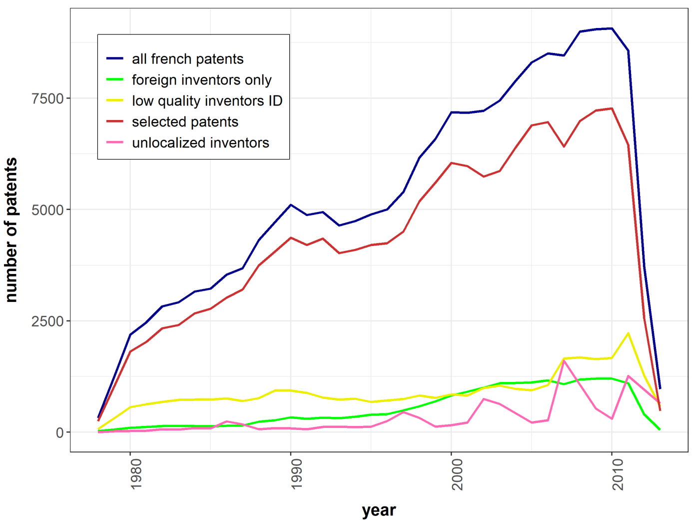



Figure 1: Patent selection

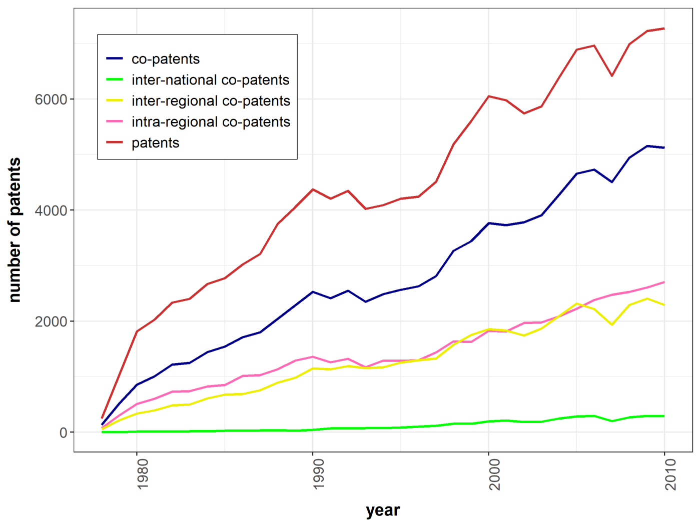



Figure 2: Amount of collaboration patents

# 3 The Geography of Patenting Activity

Figure 3 shows the location of inventors from 1978 to 2010. Blue dots represent inventors' locations while red dots represent departments' capitals (called *préfecture*).[7] 44% of all the inventors are located within préfecture boundaries, while 87% of all inventors are located in the urban area of their region's prefecture. Hence, innovative

[7]Each inventor is geo-coded independently in each patent. Hence, if the inventor files multiple patents with the same location, the observed blue dot in the map is darker.

activity takes place either in the préfecture and its surroundings. Figure 4 shows the location of inventors by sub-periods.[8]

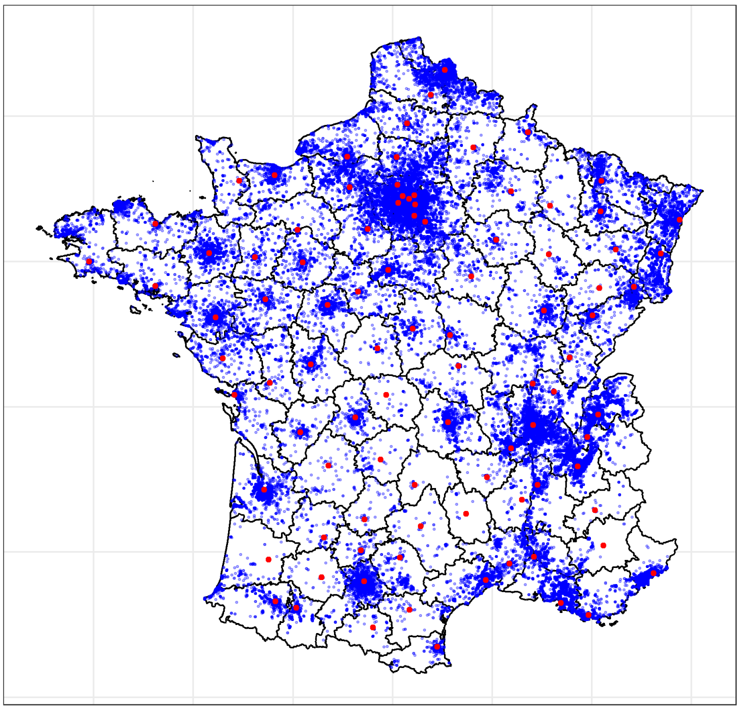
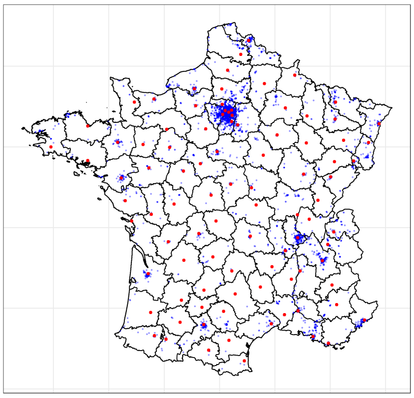


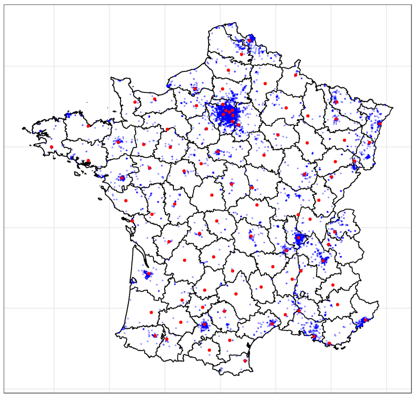
Figure 3: Inventors' location (1978-2010)

For a clear comparison in the intensity of patenting activity in each department, we aggregate patents within department. Figure 5 shows the map of France with departments colored according to the log-amount of patents in 1978-2010. We have combined into a single region Paris and its three bordering departments, i.e. Hauts-de-Seine, Seine-Saint-Denis and Val-de-Marne.[9] We have done this with the objective of having spatial units of analysis that approximate commuting zones. In the future we plan to use commuting zones as defined by the *Institut National de la Statistique et des Études Économiques* (INSEE). The amount of patents in Paris' combined area gathers a total of 50,266 over the whole sample period. It is far beyond Yvelines (Versailles), Essone (vry), Rhne (Lyon) and Isre (Grenoble) which, in descending order, each account for between 10,000 and 12,000. The department with the lowest amount of patents is Lozre, accounting for 17 patents in the whole sample period.

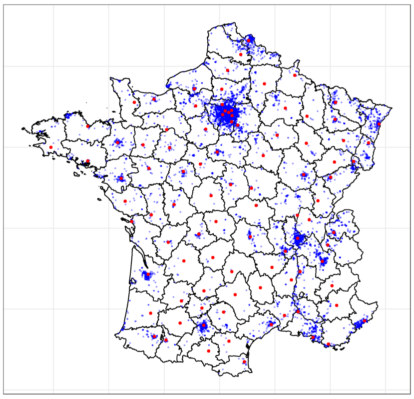
[8] All sub-periods are 4 years long, except for the last one that is 5 years long.
[9] This region is commonly referred as *Paris' little crown*.



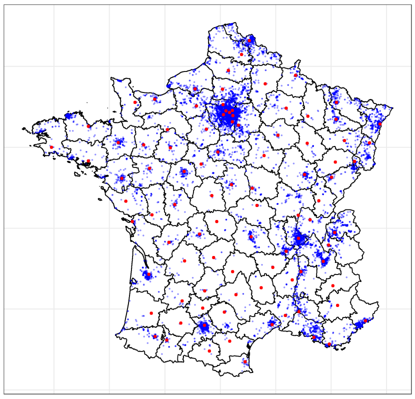
(a) 1978-1981



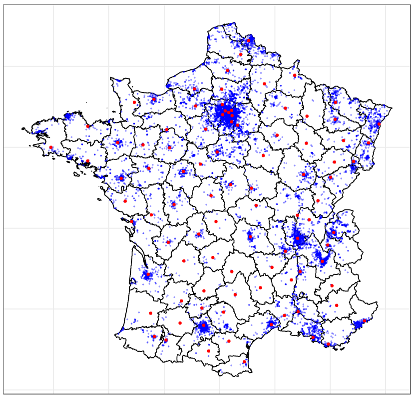
(b) 1982-1985



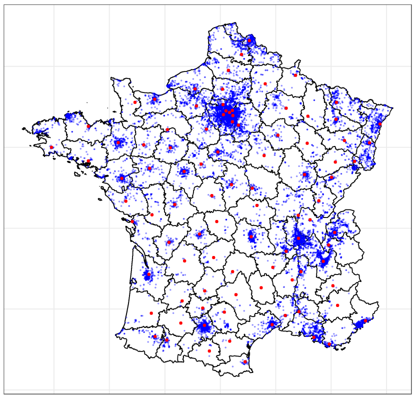
(c) 1986-1989



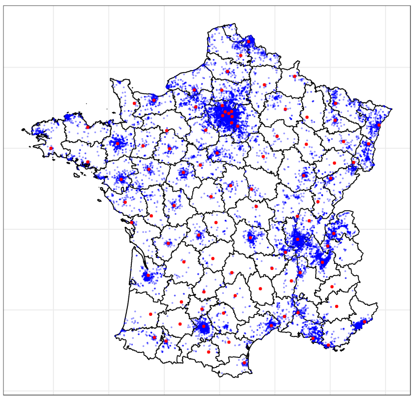
(d) 1990-1993



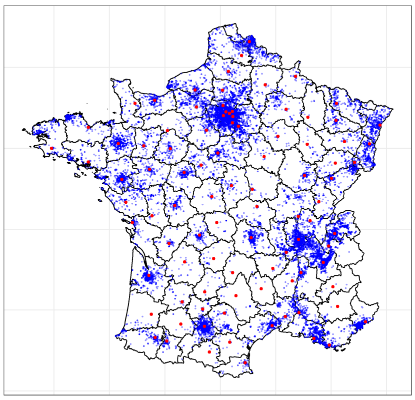
(e) 1994-1997



(f) 1998-2001



(g) 2002-2005



(h) 2006-2010



Figure 4: Inventors' location over time (1978-2010)

Figure 6 colors departments according to the department's average growth rate sub period-on-sub period, where periods are the ones displayed in figure 4.[10] The departments in darker colors are Haute-Marne, Aveyron and Lozre which experienced an average growth between 125% and 143%. However, even if they had a large growth rate, each of these departments accounted for between 6 and 35 patents in 2006 to 2010. Paris' combined area experienced an average growth of around 28%. It is one of the lowest growth among all the regions. The correlation between a department's total amount of patents and the average growth rate is $-0.27$.

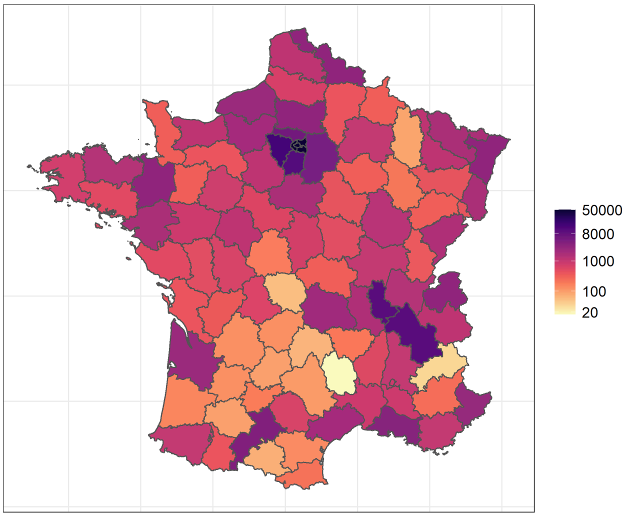



Figure 5: Amount of patents by department (1978-2010)

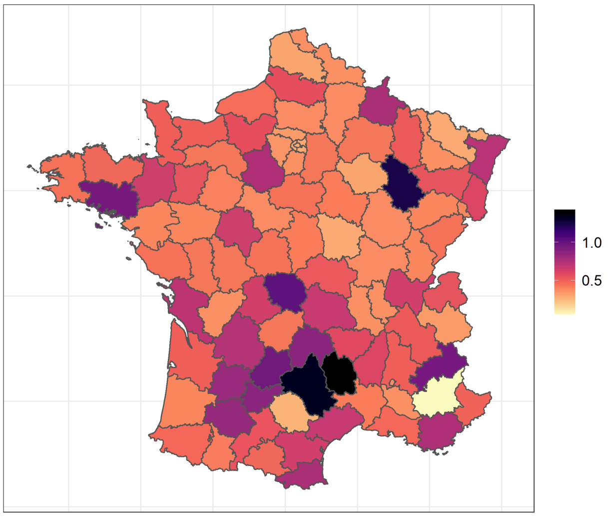



Figure 6: Average period-to-period patent growth rate

### 3.1 Openness to inter-regional collaboration

Figure 7 presents the share of collaboration patents in all patents, and the share of inter-regional collaboration patents in both collaboration patents and all patents. We observe that the share of collaboration patents on all patents increased over time. At the same time, the share of collaboration patents that contain inventors in multiple departments increased from 40% in 1980 to 50% in 2000, and then decreased to 46% in

[10]The across-year within-department aggregation is necessary in order to reduce the amount of zero patents in certain departments which would impede the computation of growth rates.

2010.

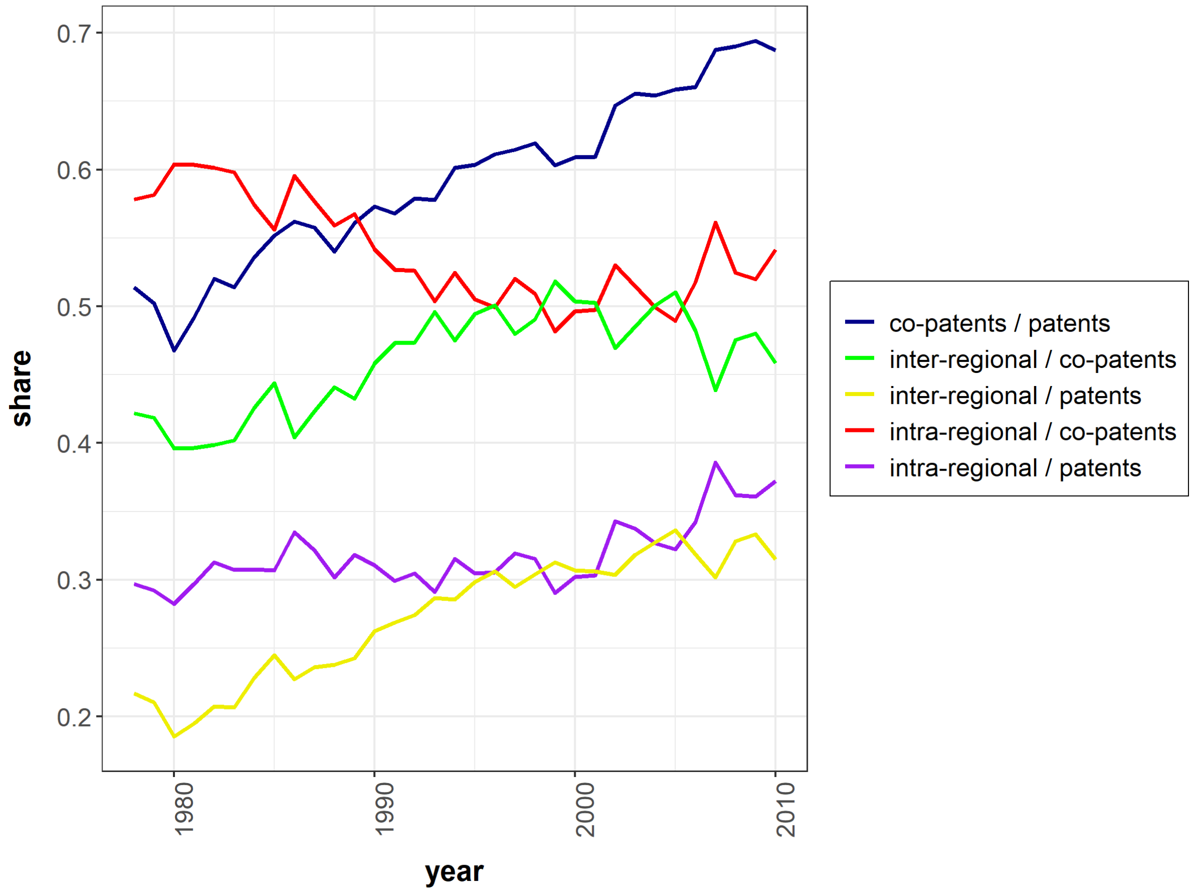



Figure 7: Shares of patents by type (1978-2010)

Figure 8 shows the ratio of inter-regional co-patents over all co-patents.[11] This ratio is between 35% for Haute-Savoie and 100% for Hautes-Alpes. Departments surrounding the Paris extended region also have a high ratio of inter-regional collaboration. Less than one quarter of regions have a share of inter-regional co-patents in all co-patents of less than 50%. The across departments mean and median share of inter-regional co-patents is about 63%, and the third quartile is about 73%.

Figure 9 presents the average growth in the ratio of inter-regional co-patents over the amount of all co-patents. The departments with the higher average growth rate are

[11] Co-patents are defined here as the sum of the amount of patents developed under inter-regional collaboration and under intra-regional collaboration.

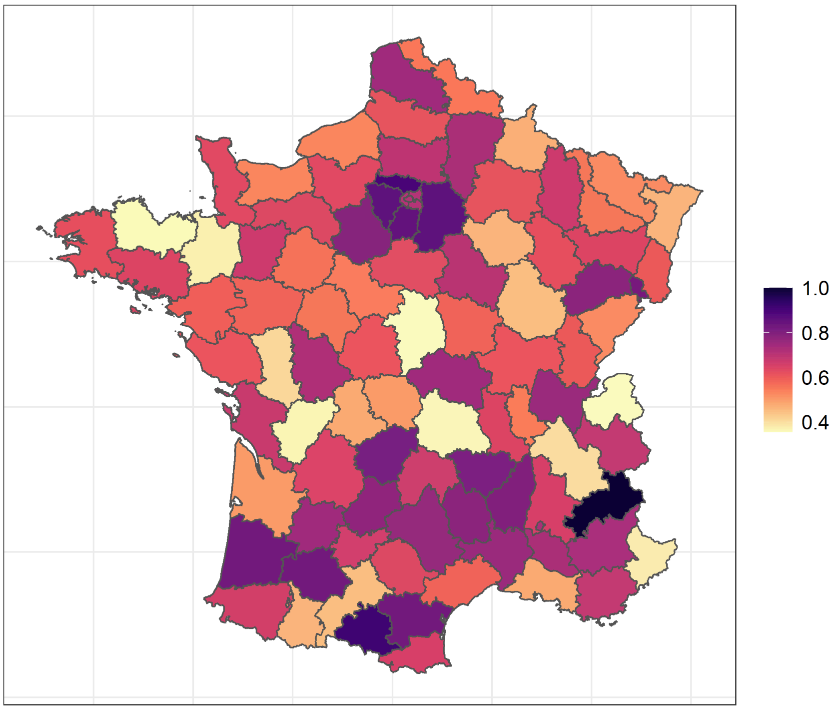



Figure 8: Share of inter-regional collaboration patents in all collaboration patents

Haute-Marne, Gers, Tarn, Dordogne, which display a growth rate higher than 89%. No region has a negative average growth rate, which would denote a redirection of collaboration towards within the department. Hence, we observe that over time inventors increase their degree of collaboration with inventor is in other departments. We also observe a negative correlation between a department's average growth rate of inter-regional collaboration and its growth rate of patenting.

Figure 10 presents the average distance of inter-regional collaborations.[12] We observe that Paris extended area has an average distance that is lower than in other departments, probably due to the fact that patenting is concentrated in that area. A future computation will compare observed collaborations relative to the potential collaborations by doing a matching method similar to Jaffe et al. (1993).

[12]We compute average distance of collaborations of department $i$ as $distance_i = \frac{1}{\sum_{j,j\neq i} \text{collaboration patents}_{ij}} \sum_{j,j\neq i} distance_{ij} \times \text{collaboration patents}_{ij}$, where collaboration patents$_{ij}$ is the amount of patents in collaboration between departments $i$ and $j$

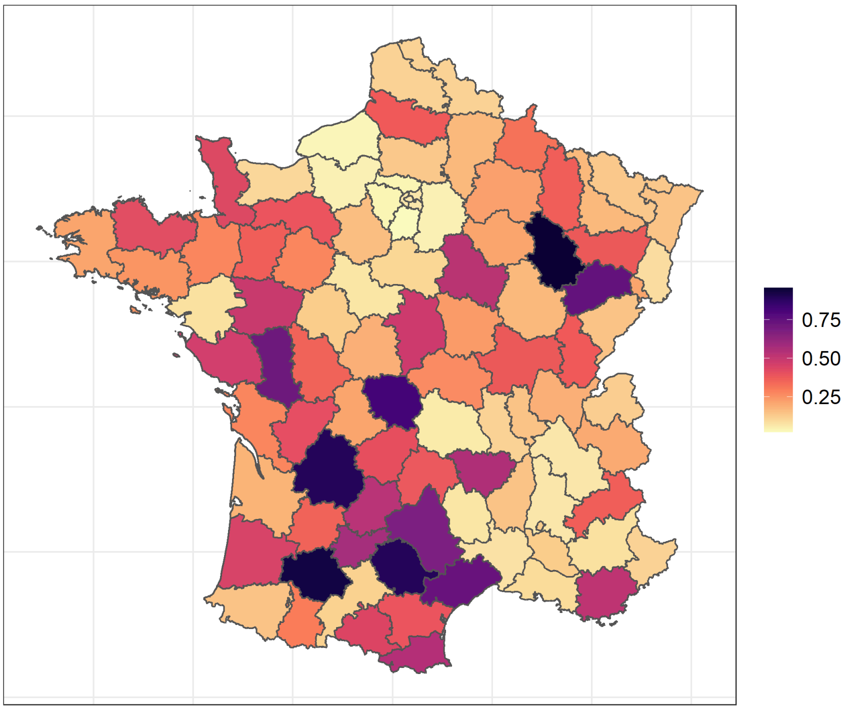



Figure 9: Average growth rate of intensity of inter-regional collaboration patenting
A department's intensity of inter-regional collaboration patenting is defined as the department's amount of inter-regional collaboration patents divided by the department's amount of collaboration patents. Average growth rate is computed period-on-period.

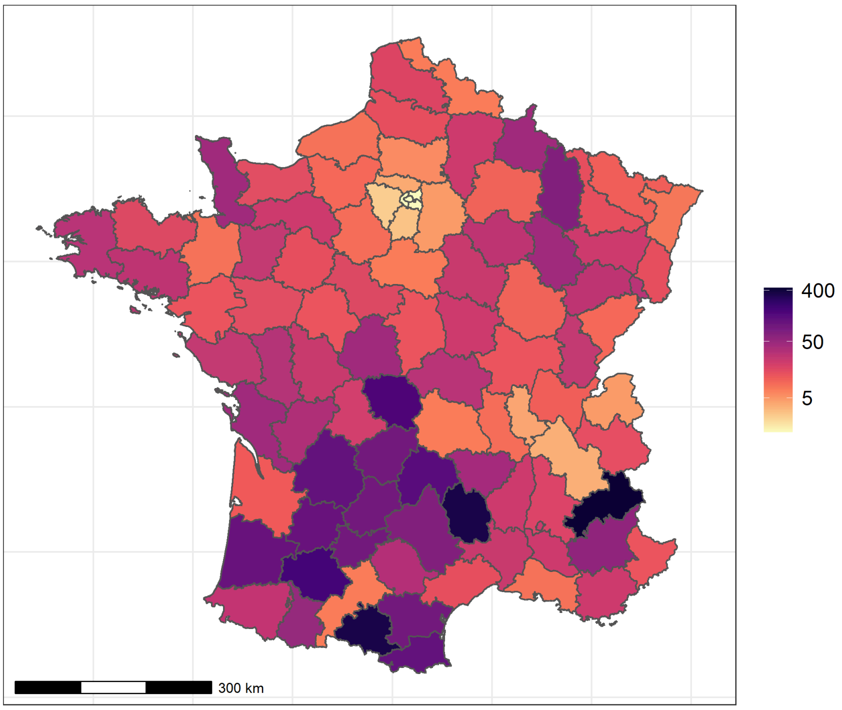



Figure 10: Average distance of inter-regional collaborations (1978-2010)

# 4 Patent collaboration across departments

In this section we show the evolution of the amount of co-patents by pair of departments, where we refer to departments by their capital city (prefecture).

## 4.1 Paris' innovation partners

We select the patents developed in collaboration with Paris extended region. From 1978 to 2010, Paris counts 74% of its co-patents developed in collaboration with other regions. Figure 11 shows the co-patenting partners of Paris colored by their share of co-patents with Paris in all Paris' co-patents. Figure 12 colors departments by their share of co-patents with Paris in the department's co-patents.

From the point of view of Paris, its bigger collaborating departments are those surrounding Paris, plus Lyon and Grenoble. In each time period Paris collaborates on average with 50 departments. From the perspective of other departments' co-patents, Paris is a major collaborating partner. The across-department mean and median share of collaboration with Paris are 12% and 15% respectively, with a maximum value of 59%.

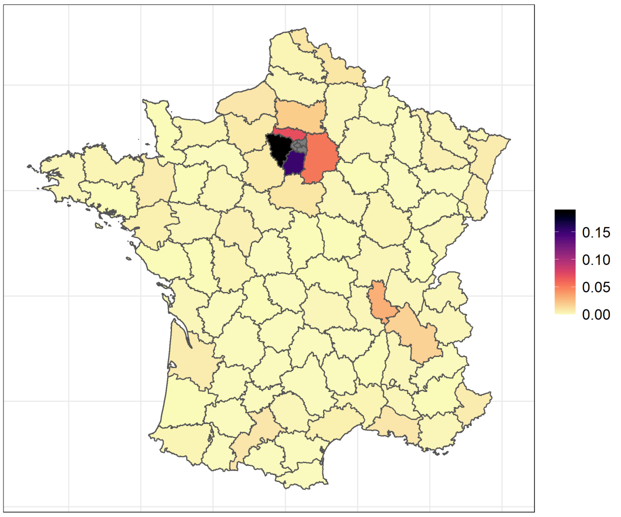



Figure 11: Paris and partners' co-patents as shares in Paris' total co-patents

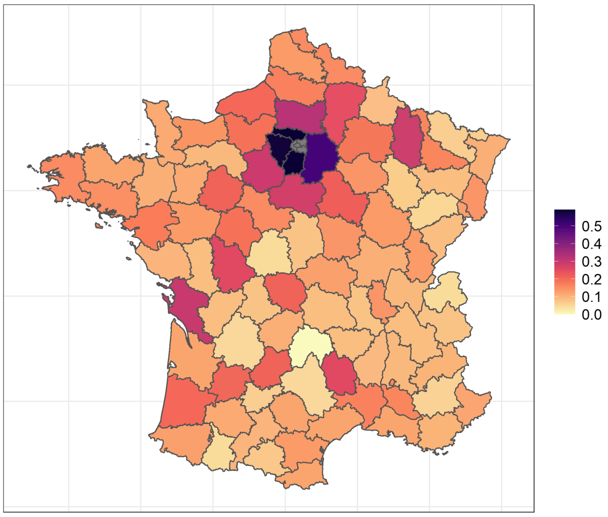



Figure 12: Paris and partners' co-patents as shares in other regions' total co-patents

## 4.2 Secondary cities' innovation partners

We focus now in pairs involving secondary regions that have been impacted by the HSR network during the years of analysis. We select the regions with the following main cities: Lille (connected to Paris in 1993), Le Mans and Tours (connected to Paris in 1989 and 1990), Strasbourg (connected to Paris in 2007), Rennes (impacted by the HSR through the connection to Le Mans), Lyon (connected to Paris in 1981-1983), Grenoble (impacted by the HSR through the connection to Lyon), Marseille, Montpellier and Nice (connected to Lyon by HSR in 2001).

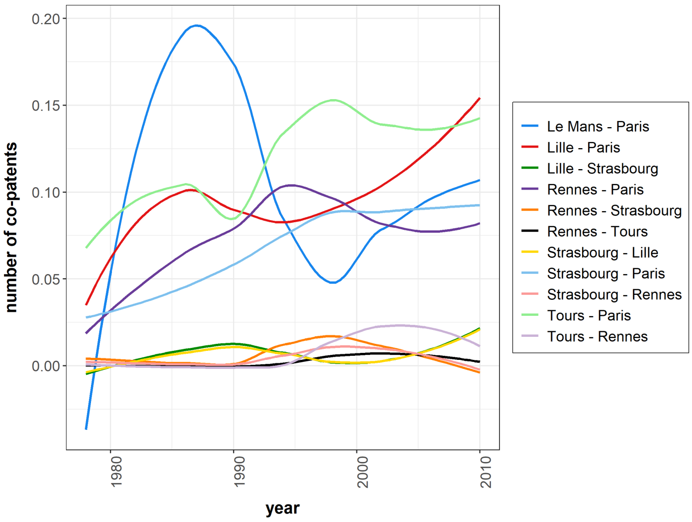



Figure 13: Share of co-patents in first region's total patents: Northern pairs

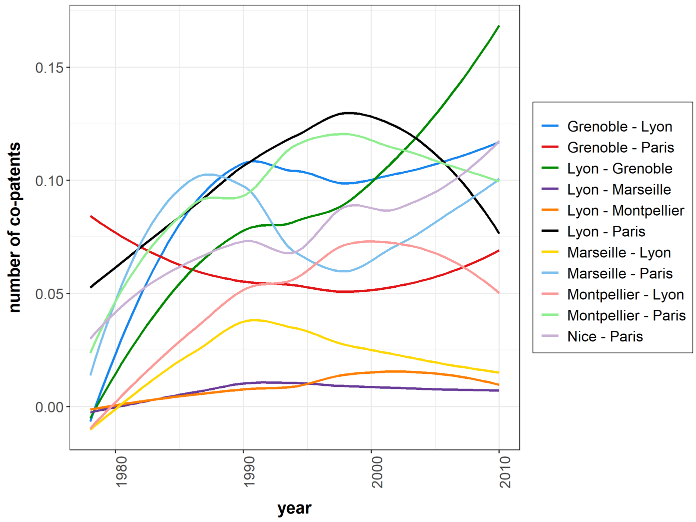



Figure 14: Share of co-patents in first region's total patents: Southern pairs

We show the evolution of the share of co-patents developed in the total amount of patents of the first region named in the pair. Figure 13 and 14 gather pairs that developed the biggest amount of patents overall the years of the study, separating the sample among Northern and Southern regions. Plotted curves are smoothed using a polynomial transformation.

Among the Norther pairs we observe Le Mans - Paris. Before the HSR roll-out in 1989, Le Mans witnessed an increasing share of patents in collaboration with Paris, reaching almost 20% in its peak. After the HSR roll-out, we see that the share of co-patents with Paris has decreased sharply. Not shown in the plots, during the period 1990-2000 we observe in the data that Le Mans has increased the number of regions it collaborates with, which naturally decrease the shares. Another pair which experienced a similar trend is Rennes - Paris. The share of patents in collaboration with Paris in the total amount of patents developed in Rennes has increased before the HSR roll-out, and then decreased.

On the other hand, we find pairs who experienced an increase in their co-patents share after the shock in travel time induced by the HSR expansion. The pair Lille - Paris experienced an increase in the amount of co-patents relative to the total amount of patents developed in Lille, with a boost after 1993 when the pair has been connected by an HSR. A similar observation can be done for the pair Tours - Paris, connected by an HSR in 1990.

In Figure 14 we would like to highlight a pattern that we suspect may be consequence of a combination of local shocks and network effects. After Lyon has been connected by HSR to cities in the south in 2001, the shares of co-patents developed between Lyon and Paris decreased. At the same time, we observe an increasing share of patents co-developed between Grenoble and Paris, and between Grenoble and Lyon as well. Lyon seems to shift its collaborative activity from Paris to Lyon. At the same time, Grenoble increases its share of collaborations with Lyon and Paris. It remains to be studied if this shift in collaborative behavior is related to the *trade diversion effect* common in the trade literature.

# 5 Technological specialization

## 5.1 Patent collaboration by technology field

Patents applications contain information on patent claims, which give the scope of the protection conferred by a patent. It gives information on the technology field each part of the invention to be protected belongs to. Technological field are classified following the International Patent Classification (IPC).[13] We work with eight primary fields presented in Table 1.

| Primary field | Description |
|---|---|
| A | Human Necessities |
| B | Performaing Operations and Transporting |
| C | Chemistry and Metallurgy |
| D | Textiles and Paper |
| E | Fixed Constructions |
| F | Mechanical Engineering, Lighting, Heating, Weapons and Blasting |
| G | Physics |
| H | Electricity |

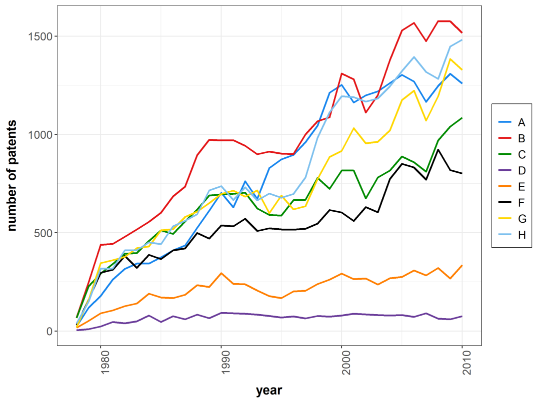
Table 1: Primary technology fields

In order to observe the main technology field of each patent, we aggregate each patent's claims by primary field presented above, from sector A to sector H. For each patent, we count the amount of claims belonging to each field, and then select the main sector a patent belongs to according to the sector that counts the highest amount of claims. Then, we are able to observe the diversity of technological knowledge for each region, computing their specialization portfolio by counting the relative amount of patents it counts by field.

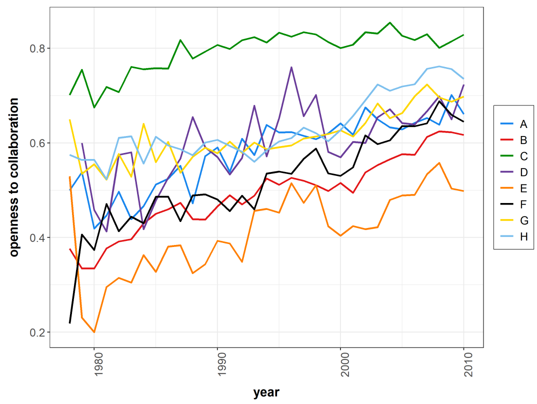
First we plot the evolution of the amount of patents developed by sector, illustrated in Figure 15a. The amount of patents in each field follows an increasing trend, such as the total amount of patents over time showed in Figure 5, except for fields of Textiles and Paper (D) and Fixed Constructions (E) which are stagnating at the same level.

Figure 15b shows the openness to collaboration of each field. The variable of interest is computed as the total amount of co-patents (both inter-regional and intra-regional co-patents) over the total amount of patents by field. It shows the propensity of each

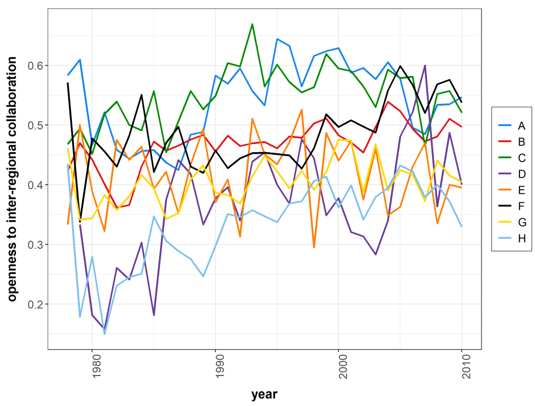
[13] To find more information on the International Patent Classification system, visit the WIPO's IPC webpage



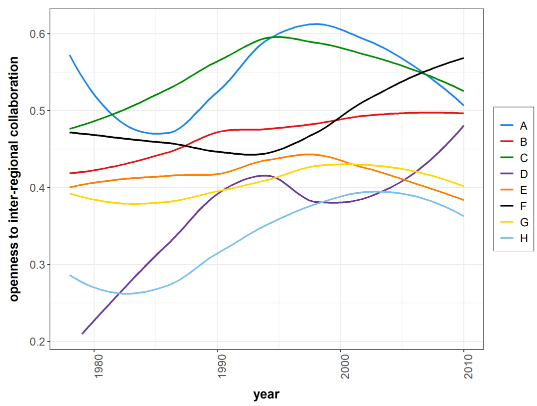

(a) Amount of patents




(b) Co-patents as share in total amount of patents




(c) Inter-regional co-patents as share in the total amount of co-patents




(d) Inter-regional co-patents as share in the total amount of co-patents (trends)

Figure 15: Patenting activity by sector over time (1978-2010)

field to be relying on collaborative work. The field that relies the most on collaboration is the one of Chemistry and Metallurgy (C) with an openness ratio between 0.7 and 0.9. We observe over time all fields increase their share of collaborative patents.

Finally, Figure 15c and 15d show the ratio of openness to cross-regional collaboration of each field, in level and under smoothing transformation respectively. It computes the amount of inter-regional co-patents on the total amount of co-patents over time. Again, the field of of Chemistry and Metallurgy (C) appear to be the one more open to collaboration with other regions, as well as the field of Human Necessities (A), which gathers patents in the subfields of Agriculture, Food, Domestic articles, Health and Amusement. Both fields reach level of openness to inter-regional collaboration about 70% over all co-patents. We find variation in their propensity to innovate with other regions. The trend is generally increasing first, but then decreases for some of the fields, i.e. fields A, C, E, G and H.

## 5.2 Technology field by region

To learn more about the regional knowledge portfolios, we compute the amount of patents developed in each field relative to the total amount of patents in each region for the whole period 1978-2010. We plot results in Figure 16.

As we observe, Paris' little crown region is particularly diversified in its patent activity across fields. It appears in light colors for each field, and display a maximal proportion of patents in Electricity field (H), which represent about 20% of its overall amount of patents.

Few regions are strongly specialized in one technology field. For example, Cher (which capital city is Bourges) has the highest ratio of patents in one field, which is this case is Mechanical engineering (F) accounting for 65% of Cher's patents. The second most specialized region is Loir-et-Cher (which capital city is Blois) with a share of 45% in the same field.

The technology fields Human Necessities (A) and Performing Operations and Transporting (B) have a relatively high presence in all regions. The average amount of patents in these fields among all regions is about 20% and 27% respectively. On the contrary, few regions are particularly specialized in the field of Textiles and Paper (D) and Fixed Constructions (E), with ratios of patents about 2% and 7% respectively. We would like to point that the propensity to invent and to patent conditional on inventing may vary across technology fields and hence this would be directly reflected in the amount of patents across technologies. However, by looking at the patenting intensity within the same technology across regions we can still learn which regions are more specialized in a field than others.

In Figure 17, we plot the sector that arrives first in terms of specialization for each region. We observe that specialization in a field is quite localized. Regions and their neighbours often share the same technological knowledge specialization. When looking at fields of technological specialization across periods, we find that in average, regions display around 3 different fields in which they specialize. Some regions are quite diversified and count 6 different main fields across periods.

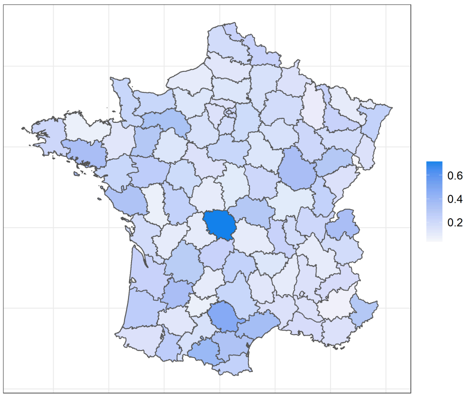
## 5.3 Technological proximity

We compute a measure of technological proximity across regions by using patents' claims' technology class at the IPC 35, a set of 35 technology subclasses. We compute the measure following Jaffe (1989), applying it for regions rather than firms.

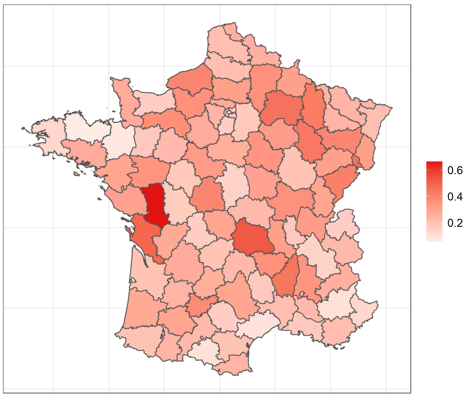
We first define $P_{ikt} = \frac{C_{ikpt}}{C_{ipt}}$ as the amount of occurrences of each technology field $k$ by claim $C_{ikpt}$ over the total amount of claims $C_{ipt}$ for each patent. Then, we compute $F_{ikt} = \frac{P_{ikt}}{P_{it}}$ as the frequency of the amount of patents developed in region $i$ at time $t$ in each technology field $k$ over the total amount of patents $P_{it}$. It measures the degree of a specialization in each field from 0 to 1 with respect to other fields. Finally, we compute vectors $F_{it} = (F_{i1t}, ..., F_{iKt})$ for each region and year, which represents the vector of technological portfolios of region $i$ at time $t$, with $k$ the technology field, with $K = 35$. These variables being defined, we can express a contemporaneous technological proximity measure, computed as a cosine similarity index between regions $i$ and $j$:

$$\text{technological proximity}_{ijt} = \frac{F_{it} \cdot F_{jt}}{\sqrt{(F_{it} \cdot F_{it})(F_{jt} \cdot F_{jt})}} \tag{1}$$

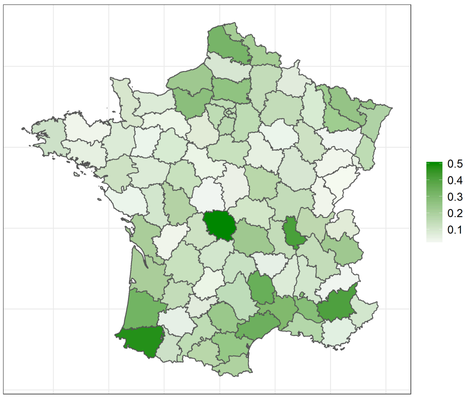
The correlation coefficient between contemporaneous share of inter-regional co-patents in all co-patents and the technological proximity within a department-pair is 0.03. If we aggregate collaborations within a department-pair across all years, and compute



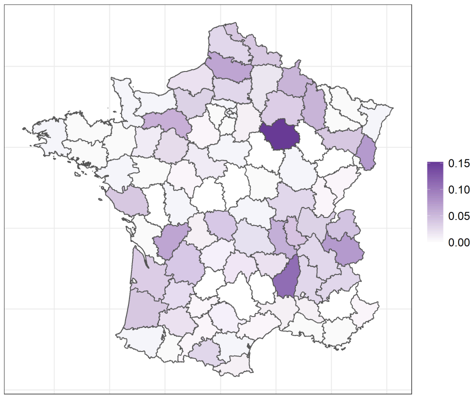

(a) Human Necessities



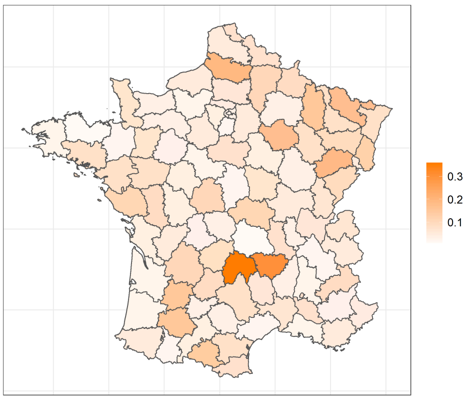

(b) Performing Operations and Transporting



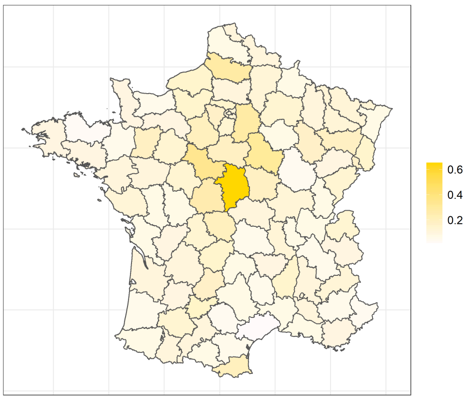

(c) Chemistry and Metallurgy



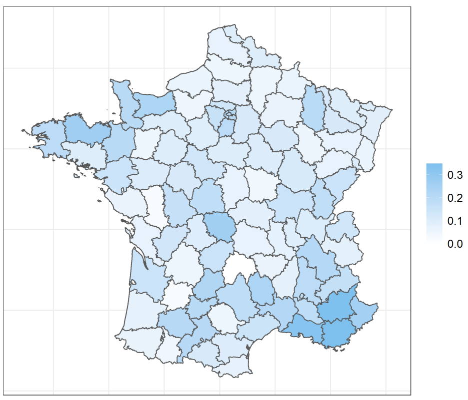

(d) Textiles and Paper



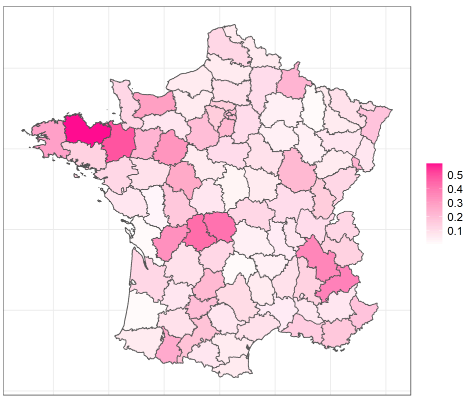

(e) Fixed Constructions




(f) Mechanical Engineering, Lighting, etc.




(g) Physics




(h) Electricity

Figure 16: Technological composition of departments

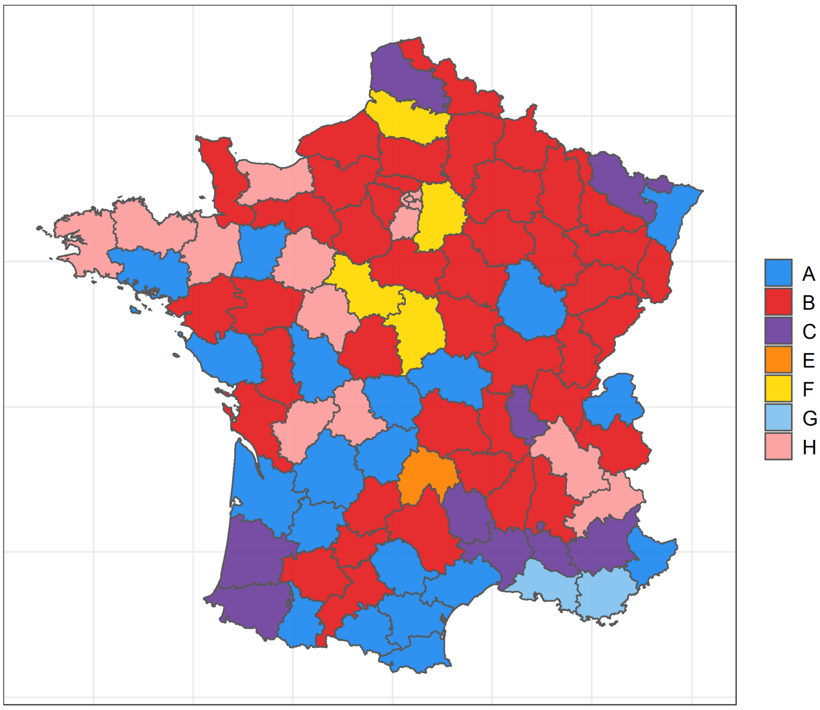



Figure 17: Main technology field by department

the proximity index using claims of all years, we find a correlation coefficient equal to 0.09. We interpret this value as being low: regions that collaborate more with each other are not necessarily similar.[14] However, the department-pair may not be the appropriate level to measure proximity, as collaboration may be a more micro-level phenomena. In the future we plan to compute proximity between firms and inventors that collaborate with each other.

# 6 Inventors' teams

This section describes the Morrison et al. (2017) dataset with respect to inventors' teams. In our sample, 37% of the patents are developed by a unique inventor. The remaining 64% are developed by at least two inventors, which compose a research team. In the next subsections we study the composition of research teams.

[14]We plan to recompute the proximity index dropping inter-regional collaboration patents, as they would mechanically increase the similarity between regions within the collaboration.

## 6.1 Amount of inventors within a team

The average amount of inventors involved in a patent is showed by the blue curve in Figure 18. The amount of inventors involved in a patent has increased over time from 1.8 to 2.5 inventors in average. The amount of inventors localized in French departments involved in a patent increased from 1.6 to 2.1 in average (purple curve). The difference between the two curves is due to poorly located and non-located inventors. on average, we count 0.14 unlocalized inventors per patent, and 0.04 inventors with an identifier of poor quality. Collaboration with foreign inventors increased overtime from an average of 0.01 foreign inventors per patent to 0.13. In the subsequent analysis we only keep inventors located within French departments with high quality.

The average amount of inventors in inter-regional patents increased from around 2.6 inventors in 1990 to 3 inventors in 2010. At the same time, the amount of inventors in intra-regional patents (patents in which all inventors reside in the same department) went from around 1.9 inventors to around 2.1 inventors. Hence, inter-regional patents have a larger team on average.

The map in Figure 19 shows the overall average amount of inventors per team involved in a inter-regional co-patent by region. The darker the region, the bigger the average team size the region is involved in. Figure 20 shows the average proportion of domestic inventors involved in a inter-regional co-patent with respect to the total number of inventors localized in French regions involved in the patent. We see that in average, teams count around the half of the inventors that compose the team in one region. Indeed, regions count in average 33% to 54% of a research team located domestically. The share is so high in average since 50% of the patents in the sample count 2 regions involved in a cross-regional patent. In average, a patent involves 2.4 regions, and at most, it counts inventors from 34 regions in a same team.

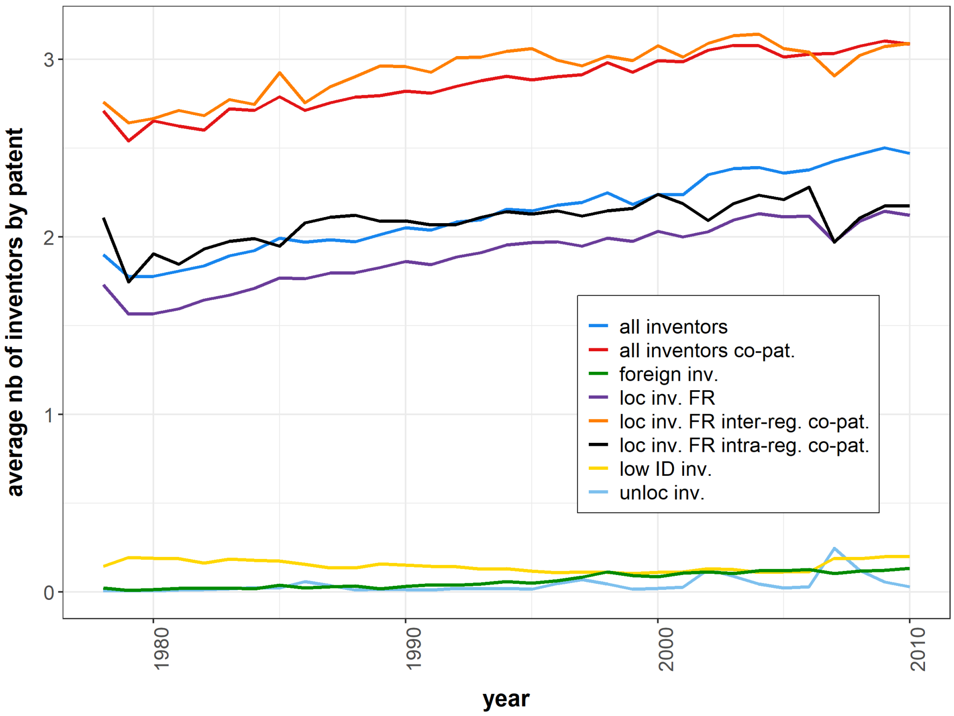


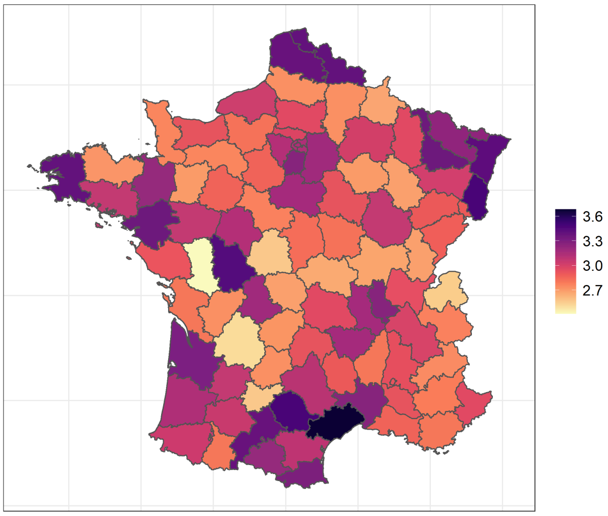

Figure 18: Average amount of inventors involved in a patent team

## 6.2 Average distance between inventors by team

Figure 21 shows the distance between inventors computed based on their geographical coordinates. The average distance between inventors involved in a same patent project has increased from around 50-60 kilometers to 90 kilometers. The average distance between inventors within the same team and region, we see that the figure increased from 5 to 12 kilometers. Finally, the average distance between inventors of a same team has increased from 100 kilometers to almost 150 kilometers at the end of the period.

## 6.3 Inventors' network and regional network proximity

The literature on research team collaboration has been increasingly adapting concepts from the social network literature. This literature explains that the creation of a collaboration between two individuals can be explained by an existing relationship between



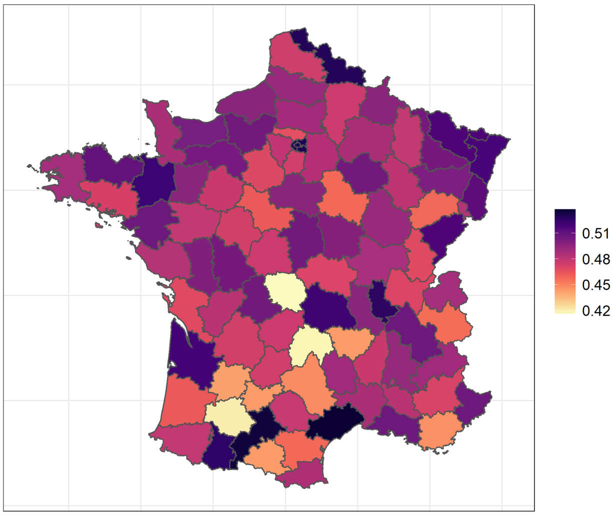




Figure 19: Average amount of inventors involved in an inter-regional co-patent team

Figure 20: Average proportion of domestic inventors - inter-regional co-patent team

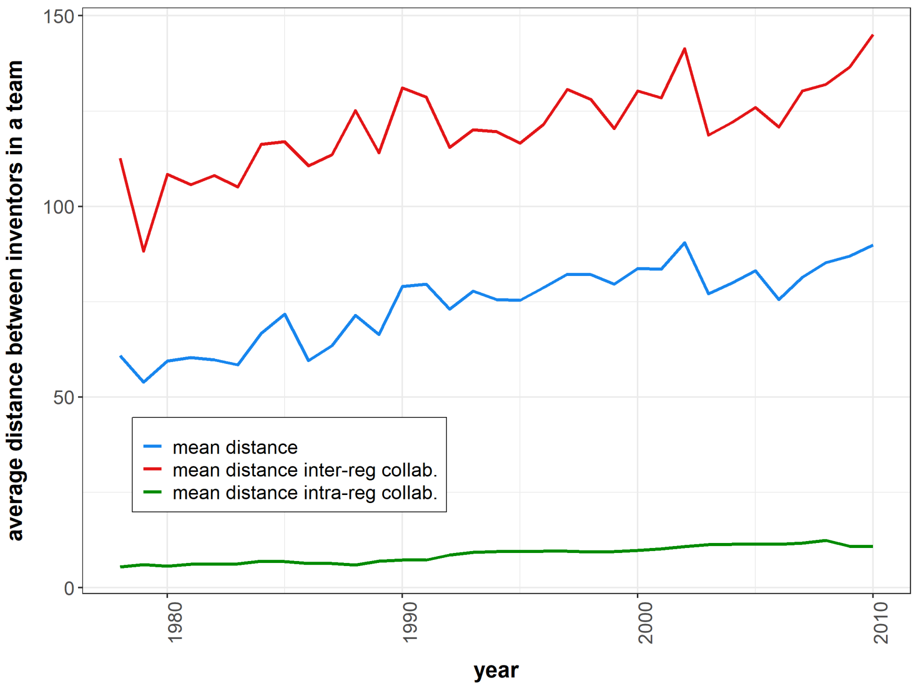



Figure 21: Average distance between inventors within team

both, or an existing relationship with a common third individual. If individuals A and C collaborate together, and individual C is also part of a research project with individual B, there are chances for A and B to hear about each other from their common collaborator and finally begin to collaborate with each other, or that their respective acquaintances meet and collaborate. Hence, indirect connections between individuals can be at the source of collaboration creation.

Bergé (2015) adapts the concept of bridging paths to regional network proximity. Figure 22 illustrates the method. Two regions are considered to be close, in terms of their inventors network, if they count a numerous amount of indirect connections between their inventors. We see from the figure that region $i$ and $j$ have two pairs of direct connections as their individuals collaborate with each other. But they also count an indirect connection, also called bridging path, as one inventor in each region collaborate with a same third inventor in region $k$. On the other hand, regions $i$ and $k$ count two bridging paths, while regions $j$ and $k$ do not count any. Note that to have a bridging path between regions $i$ and $j$ it is necessary that the same inventor in $r$ collaborates with inventors in both $i$ and $j$.

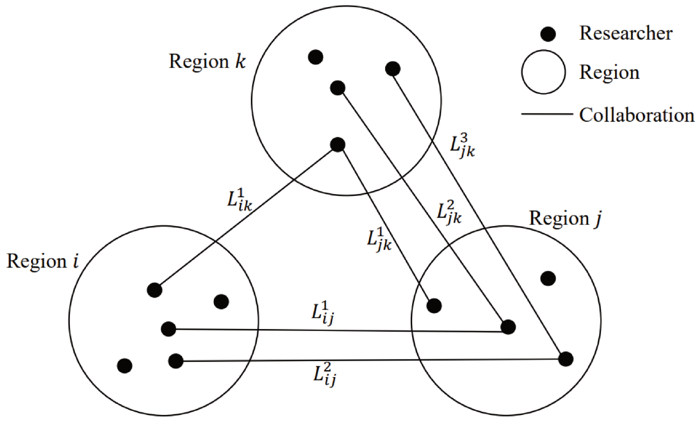



Figure 22: Bridging path - Bergé (2015)

We compute the amount of bridging paths between all regions $i$ and $j$ by summing the amount of inventors in all other regions $r$ that collaborate with $i$ and $j$ in year $t$.[15] In doing so we we only consider collaboration patents that include the inventors in $i$ and $r$, but not inventor in $j$. Similarly with patents with the same inventor in $r$ and another inventor in $j$, but no inventor in $i$.

The amount of bridging paths between each pair of regions overall the years of the same ranges from 0 to 84, with 0 being the median as well as the 3rd quartile. 35% of the region-pairs have at least one bridging path in all the sample period. Among those that have at least one bridging path in one year, the average amount of bridging paths per year is about 0.4. Using all region-pair-years, the correlation coefficient between the amount of co-patents with a region pair and the amount of bridging paths is about 0.44.[16]

# 7 Inventors' mobility

## 7.1 Data characteristics

In this section we focus on inventors and their mobility across French departments. We base our analysis on the patent data provided by Morrison et al. (2017). In their inventor disambiguation procedure, Morrison et al. (2017) first create location-specific inventor identifiers and then use information of those inventors to check if they have shared co-inventor, shared assignee, shared triadic family, or shared citations. Among inventors who have shared characteristics they run a name-matching procedure to identify those inventors who are actually the same person.[17] This last step in their

[15]This is different from the original method proposed by Bergé (2015). The computation in Bergé (2015) is: network proximity$_{ijt} = \sum_{r \neq Q_i \cap Q_j} \frac{\text{copat}_{irt} \times \text{copat}_{jrt}}{\text{pat}_{rt}}$ where $Q_i$ (resp. $Q_j$) the set of all patents of region $i$ (resp. $j$), copat$_{irt}$ the number of patents under collaboration between regions $i$ and $r$ at time $t$, and pat$_{rt}$ the number of patents developed by inventors in regions $r$ at time $t$, which proxy for the number of researchers in the region.

[16]We plan to compute bridging paths in different variations: within time periods instead of years, allowing for the connecting inventor in $r$ to be two different inventors that are located closer than X kilometers, using firms instead of inventors.

[17]Morrison et al. (2017) does not use technology similarity to do the matching of inventors in different locations: "because they are less personal (working on a research topic is far less informative than

procedure creates an identifier for mobile inventors: inventors who file at least two patents with addresses in different locations.[18]

## 7.2 Descriptive statistics

In Table 2 we present simple statistics of inventor's mobility. There are 102,082 inventors in the sample of inventors who were identified with high quality. Of those inventors, 35.1% patent in at least two years. Restricting to the subsample of inventors who patent in multiple years we are able to follow inventors over time and space. We label the department of the first patent of the inventor as the inventor's home/native department. We measure moves of inventors across departments by using the address's department that the inventor declared in the patent. For inventors that patent in at least two years, the share of inventors who move at least once is 14.4%. Of those inventors who move at least once, 20% move at least one more time. Of those who move two times or more, 50.8% of them return to the department declared in the address of their first patent. Among inventors who patent in at least two years, each inventor moves on average 0.19 times. Among inventors who move at least once, each inventor moves on average 1.33 times.

For the rest of the descriptives we restrict the sample to inventors that we are able to follow over time, meaning inventors that patent in at least two years. Figure 23 shows the map of France, where each of the departments is colored by the log-amount of unique inventors that ever patent in the department in the period 1980-2010.[19] The department with the highest number of inventors is Paris (5,237), followed by Hauts-

collaborating with an individual) and the correct level of aggregation is unclear."

[18]This section has been compiled separately from previous sections. Differently to previous sections, in this section we have not merged Paris with its neighboring departments to create an extended Paris region. In the future we will replicate the analysis using the extended Paris region instead. Hence, inventors' moves across departments within the extended Paris region would not be counted as moves.

[19]Each inventor is counted in the departments that it patents. If an inventor patents twice in the same department she is counted only once for such department. However, if an inventor patents in two departments she is counted once for each department.

| | Inventors | Moves |
|---|---|---|
| N inventors | 102,082 | |
| N inventors +2 years | 35,878 | |
| N inventors move | 5,176 | 6,888 |
| N inventors +2 move | 1,039 | 1,563 |
| N inventors return home | 528 | 579 |
| N moves per inventor | 0.19 | |
| N moves per mobile inventor | 1.33 | |

Table 2: Amount of inventors and cross-department moves in 1980-2010

de-Seine (3,974, its prefecture is Nanterre), Yvelines (2,947, Marseille), Rhône (2,421, Lyon) and Isère (2,368, Grenoble).

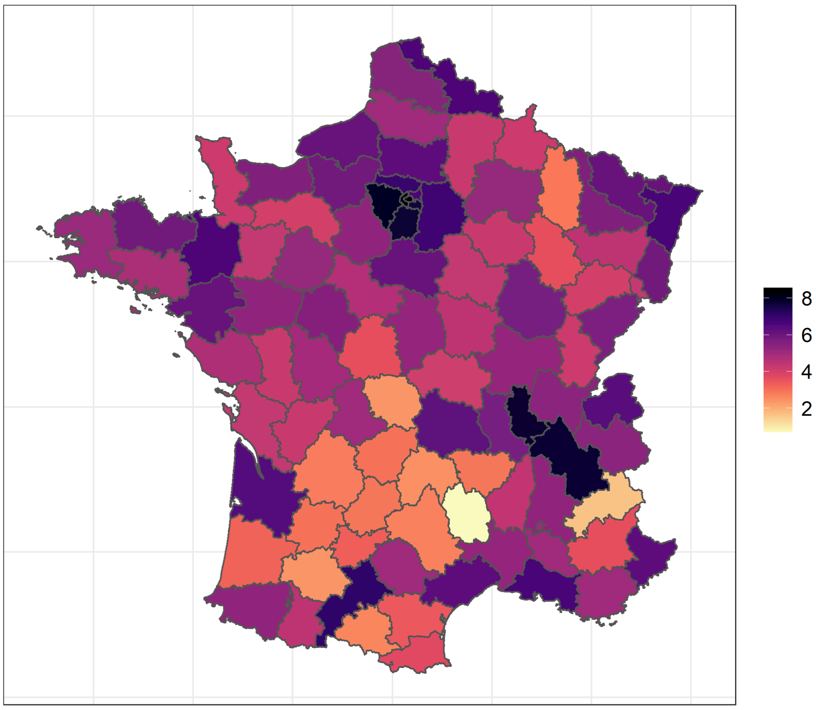



Figure 23: Log amount of inventors per department

Figure 24 presents, for each year, the share of inventors who: (1) patent in a department different to the department of the inventor's initial application (Migrants), (2) patent in a department different from the previous-patent department (Movers), (3) is the first time that they patent in a department different from the previous-patent department (First movers). The difference between (1) and (2) is that the inventor could have moved in the past and continues to patent in the *foreign* department, while in (2) it is required that the inventor changed department with respect to her last patent observed. The difference between (2) and (3) is that (2) includes all moves, while (3) only includes first-time moves (hence, the difference between (2) and (3) are 2nd and later time moves). Since 1990 we observe that around 7.5% of inventors who patent are movers, out of which 5 percentage points are first time movers. We also observe that since the year 2001, around 15% of inventors who patent in a certain year are immigrants in the location that they patent.

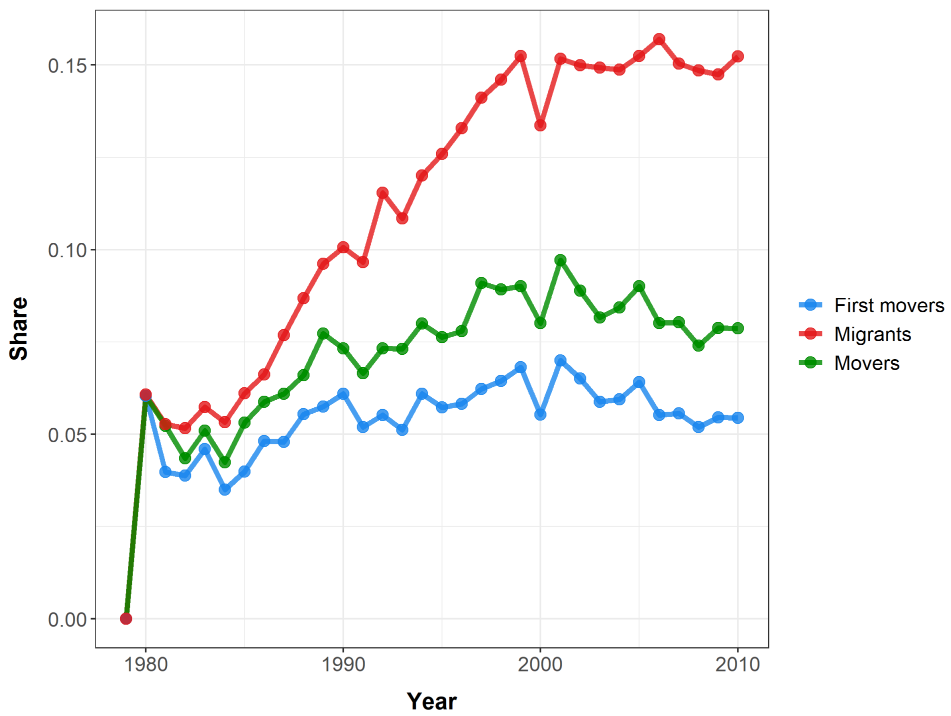



Figure 24: Share of inventors that are first movers, movers and migrants

Figure 25 shows in blue the share of inventors that patent in Paris who's first patent's department was not Paris (which we call inflows or immigration). In red it is represented, for year t, the amount of inventors whose last patent's department is Paris and in year $t$ patent in another department, divided by the amount of inventors applying for patents in Paris in year $t$ (outflows or emigration).[20] We observe that since 1990 at least 5% of inventors in Paris just moved from another department to Paris, and the figure displays an upward trend, with the share consistently being more than 10% after 2005. We also observe that since 1990 around 15% of inventors who patented in Paris, their following patent is filed with an address in another department. Overall, the figure shows that there is a significant amount of turnover in the composition of inventors in Paris.

Figure 26 replicates Figure 25 but for a different selection of departments: Bouches-du-Rhône (Marseille), Gironde (Bordeaux), Haute-Garonne (Toulouse), Isère (Grenoble), Nord (Lille) and Rhône (Lyon). The figure shows that the turnover in Bouches-du-Rhône, even if volatile, it is notably higher than in other departments. Also, the figure shows an upward trend in both inflows and outflows in Rhône.

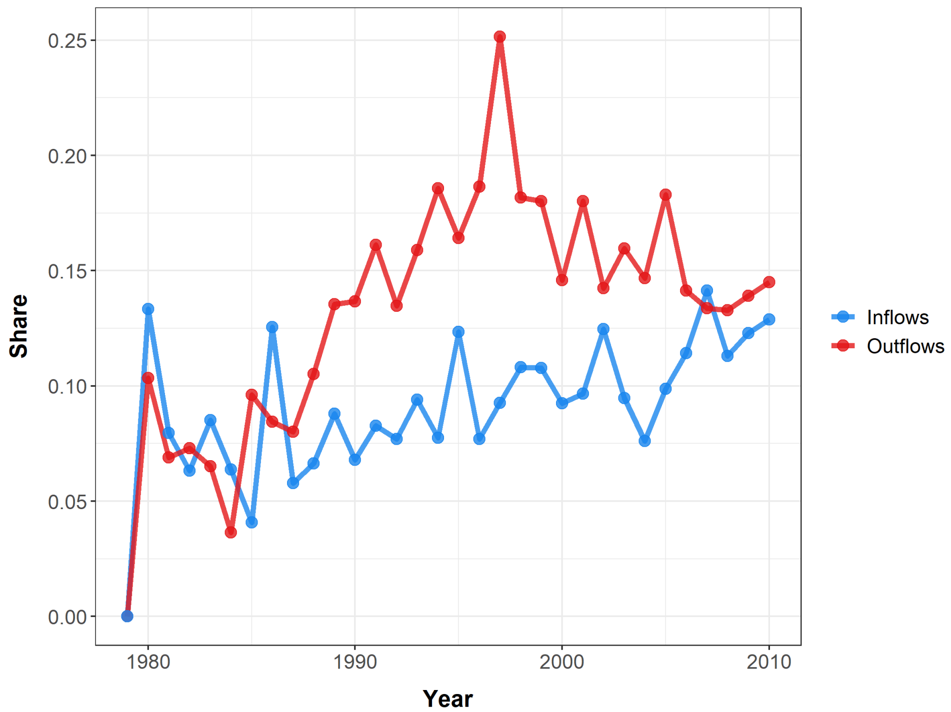



Figure 25: Share of inventors that emigrate/immigrate from/to Paris

[20] We choose to use the year $t$ inventors in Paris as denominator to have a common denominator between outflows and inflows. Additionally, taking year $t$ inventors makes it easily interpretable: taking year $t - x$ would be cumbersome to both compute and interpret as $x$ is be inventor specific.

Figures 27 and 28 show a department's emigration and immigration of inventors. For the whole sample period, Figure 27 shows, for the native department, the share of native inventors who later on patented in another department (emigrants).[21] Figure 28 shows for each department the share of inventors that ever patented in that department who are not native from that department (immigrants). Paris and its surrounding departments have a high share of both emigrants and immigrants. The share of immigrants is also high for departments in the South and South-West of France. The correlation between a department's emigration and immigration share is 0.22 and it is 0.30 for its ranking.

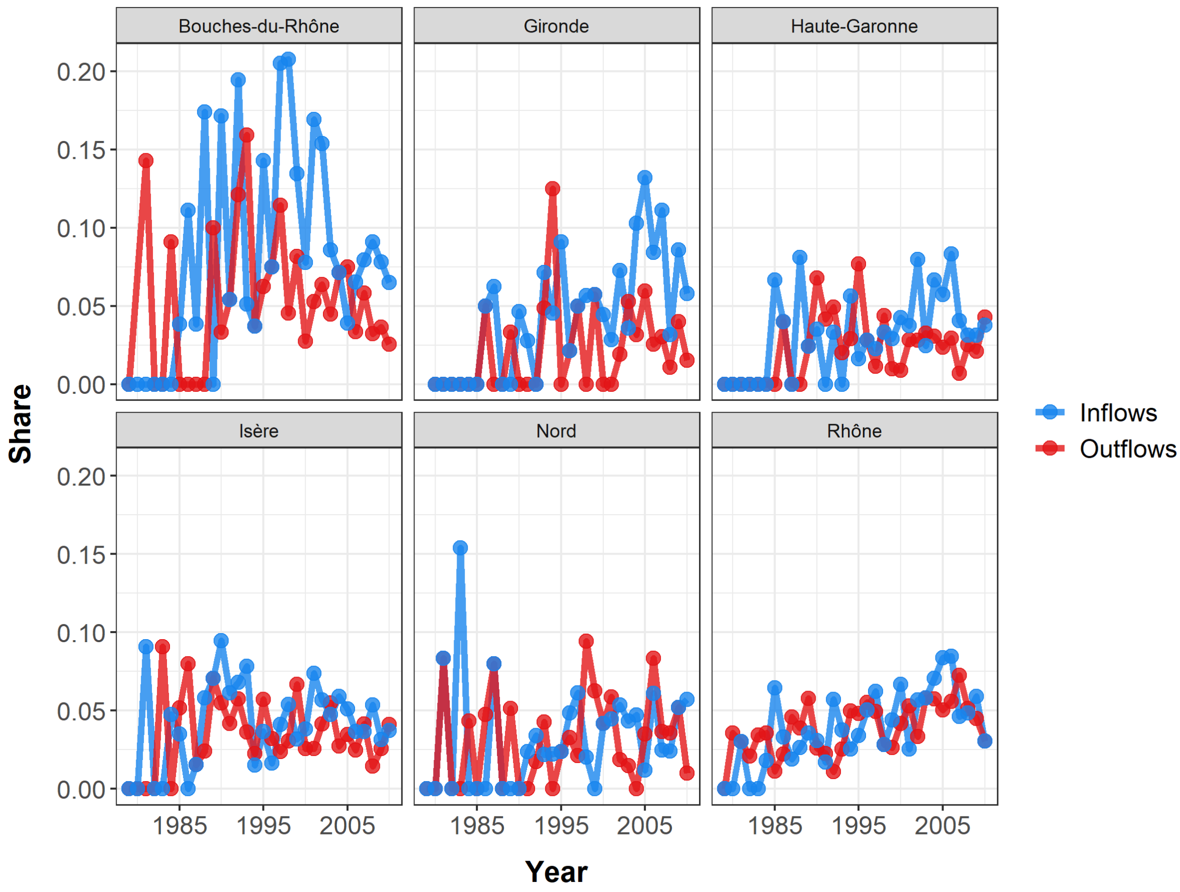



Figure 26: Share of inventors that emigrate/immigrate from/to selected departments

[21]The department Lozère has a share of 0.5 emigrant inventors. We decided not to display the share of Lozère in the map as it distorts the scale.

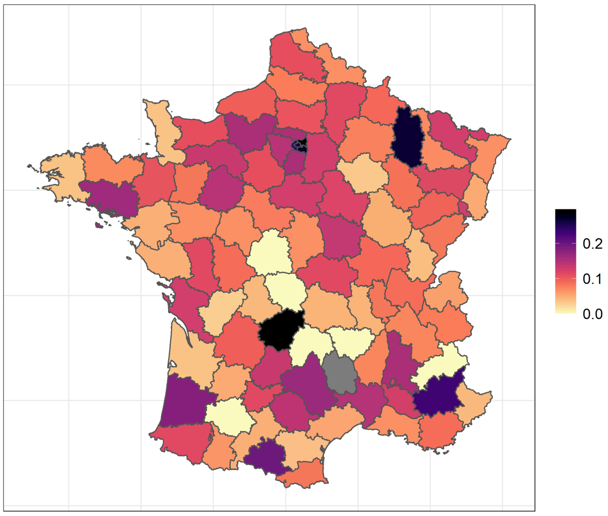



Figure 27: Share of native inventors that emigrate

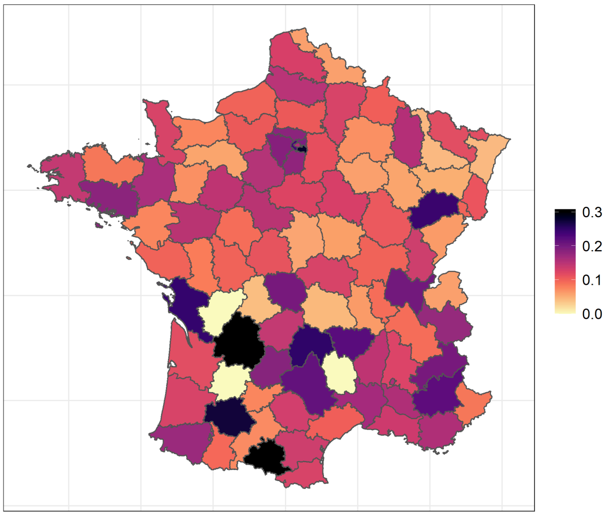



Figure 28: Share of local inventors that are immigrants

# 8 Train travel time

From early 1980's until 2018 the expansion of the *Train à Grande Vitesse* (TGV) in France led to a big reduction in travel times by train. In this paper we construct a dataset that approximates the travel time between the *Prefecture* of all French departments for each year from 1980 to 2018. Our dataset is constructed using the train lines and stations existing in 2021, and an approximation of the speed for each train type for each year. We validate our dataset by using a restricted set of city pairs for which we observe the yearly travel time, and compare with the approximated travel times. We find that approximated travel times can account for 83% of the time-variation in travel time within these city pairs. Future version of the data set would include improvements to better replicate changes of travel time for non-TGV lines.

The objective of the data construction is to obtain an approximation of the levels and changes in travel time of the train network in France. This dataset would allow us to study the effect of changes in travel time on different outcomes. In a related study, Charnoz et al. (2018) has constructed a dataset of train travel times in France by digitizing historical record of French train company Société Nationale des Chemins de

Fer Français (SNCF). We expect to replicate our analysis with the dataset of Charnoz et al. (2018) once it becomes available.

## 8.1 History of the Train à Grande Vitesse

In 1981 France inaugurated its first TGV service operating between Saint Florentin and Lyon, a high speed line which was expanded in 1983 to connect Paris and Lyon. The TGV operates at a maximum speed of 320kmh, compared to the less than 160kmh of Intercités and TER. The following is a list of the date in which different high speed train lines (Ligne à Grande Vitesse - LGV) opened:

- 1981: First part of the line LGV Sud-Est aiming at connecting Lyon to Paris. The line opened on May 22nd between Saint Florentin and Lyon. Saint Florentin is around halfway between Paris and Lyon.
- 1983: Second part of the line LGV Sud-Est reaching Paris opens on April 25th
- 1989: First branch of the line LGV Atlantique, Paris - Courtalain - Connerré - Le Mans
- 1990: Second branch of the line LGV Atlantique, Courtalain - Saint-Pierre-des-Corps - Monts
- 1991: Opening of Massy station
- 1992: East bypass of Lyon from Montanay to Saint-Quentin-Fallavier (in order to create the line LGV Rhône Alpes, which would connect LGV Sud-Est to the extreme south of France) in December 1992
- 1993: LGV Nord from Paris Gare du Nord to Lilles-Flandres, opened on May 18th
- 1993: LGV Nord from Lille-Europe to Calais-Fréthun (towards the Channel Tunnel which connects France to England), opened on September 26th
- 1994: LGV Rhône-Alpes, from Lyon-Saint-Exupéry to Valence

- 1994: East bypass of Paris Interconnexion Est at Vmars-Couvert-Crisenoy (connects Aroport Charles de Gaulle and Marne-la-Vallée Chessy to the LGV).
- 1996: East bypass of Paris Interconnexion Est at Valenton-Coubert
- 1997: Lille-Flandres to Belgium border towards Bruxelles, opened on December 10th
- 2001: LGV Méditerranée connecting Marseille to Paris via LGV Sud-Est (Paris-Lyon), LGV Rhône-Alpes (Lyon-Valence) and LGV Méditerranée (Valence-Marseille). Fork towards Avignon and Nîmes.
- 2007: First part of the line LGV Est from Paris-Est to Baudrecourt, opened on June 10th.
- 2010: Line towards Spain, Perpignan-Figuières, opened in December 19th.
- 2011: First part of LGV Rhin-Rhône from Villiers-les-Pots to Petit-Croix, opened on December 11th.
- 2016: Second part of the line LGV Est between Baudrecourt and Vendenheim, opened on July 3rd.
- 2017: LGV Sud Atlantique between Tours and Bordeaux, opened on July 2nd (extension of the second branch of the line LGV Atlantique which connects Paris to Tours).
- 2017: LGV Bretagne-Pays de la Loire, which connects Le Mans to Rennes, opened on July 2nd (extension of the first branch of the line LGV Atlantique which connects Paris to Le Mans).
- 2017: Bypass of Nîmes and Montpellier, LGV Méditerranée, opened on December 10th.

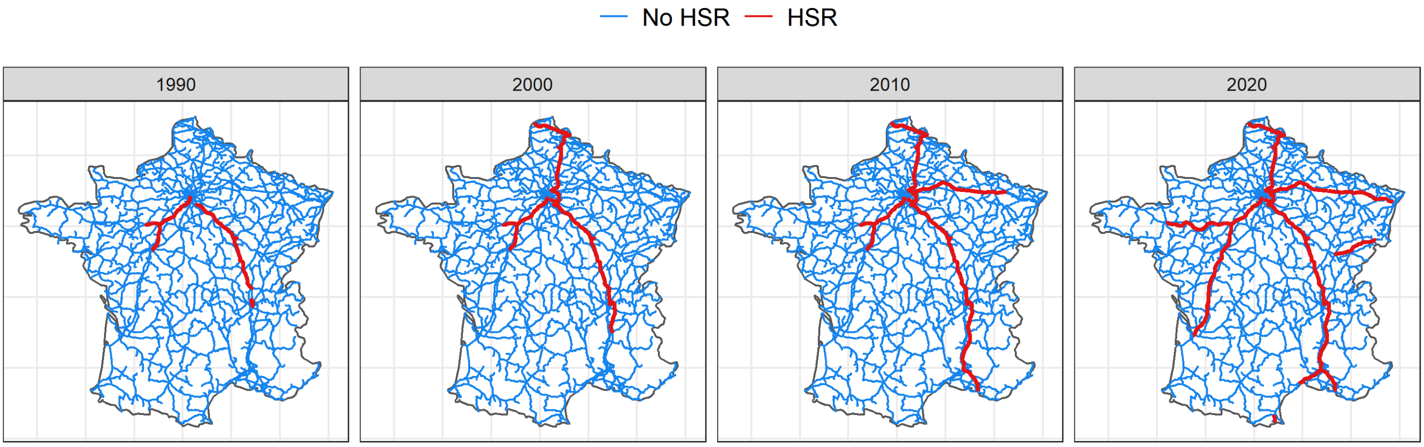



Figure 29: High Speed Railway expansion

## 8.2 Data construction

We construct travel times by computing a the yearly counterfactual travel time for each pair of stations that are directly connected. Then, for each year, we run the dijkstra algorithm to find the fastest route in between stations that belong to prefectures. In the following we list the datasets that we have used to construct the prefecture-pairs travel time. In parentheses we include the source.

1. List of train lines and their geographical coordinates in 2021 (SNCF)

2. List of stations active in 2021 (SNCF)

3. List of departure and arrival time of all (non-stop) train services realized between the 8th and the 16th of December 2021. Train services are classified by train type: TGV, Intercites and TER. (SNCF)[22]

4. List of train lines and dates in which they became high speed railways (done by the authors)

5. List of all departments' prefectures and their geographic urban area

[22]We use TGV, Intercites and TER services. In the future we plan to include Transilien services also that operate in the Ile-de-France region.

We face two main challenges using these datasets. First, we do not have the travel distance between each pair of stations, only the travel time in the week for which we have departure and arrival times. Hence, we need to infer the travel distance by computing the shortest path between every pair of stations that had operated a non-stop service. Second, train stations are not overlapping with train lines. However, we know which lines pass through the station. Therefore, we need to attribute a location of the station within the train line. We construct the dataset in the following steps:

**Step 1**: We identify train line intersections and locate stations within train lines. First, We compute the distance between each beginning and end of train line to every other train line. If the beginning or end of a train line is less than 100 meters of another, we interpret that they intersect (or connect to) each other. This is necessary because coordinates of train lines do not exactly overlap and hence appear to be disconnected. Second, we identify the location of the station in each of the train lines that pass through it by using the line's geographical point that is closest to the station. We split train lines into segments using the geographical coordinates of where train lines intersect and where there are stations. Then we compute the train line distance within each segment.

**Step 2**: We compute the train line distance between stations that are connected with a non-stop service in December 2021. We do so by running Dijkstra algorithm to find the shortest path, using train line segments as edges, and segments beginning-end points as nodes. This implies that the shortest path is allowed to change direction in every intersection and station. We compare the train line distance of the shortest path against the station-to-station straight line distance. We select pair of stations for which the train line distance is 15km or 100% greater than the straight line distance, these are paths that are potentially wrong due to missing segments (e.g. two train lines that have a *raccordement* in between but the *raccordement* was not properly intersected with both lines).[23]

[23]The threshold was set by using a sample of station pairs and doing a visual inspection of the segments

For each selected pair of stations, we select the beginning and end points of all existing segments that lay in the rectangle area which is 0.1 degrees larger in each cardinal direction North/South/East/West with respect to the rectangle formed by the two stations (sides of the rectangle are parallel to latitude and longitude lines). We compute the straight distance between all end points of a segment with the beginning point of every other segment within that square. If the distance between the end and beginning of two segments is less than 5km, we create a new fictitious segment connecting both pre-existing segments.[24] We then re-start with Step 2 and iterate a maximum of 5 times adding fictitious segments.[25]

Figure 30 shows the train line distance between two stations connected with a non-stop service in December 2021, and the observed fastest travel time between those two stations by train type. Hence, values in the x-axis are computed, while those in the y-axis are observed. As expected, for a given distance, travel time between two stations is generally lower in TGV trains compared to Intercites and TER. However, we also observe that for a given distance some TGV services have a substantially higher travel travel time. This could be due to a miscoding of train types or due to the fact that TGV trains also operate on non-high speed railways, leading to a limited speed comparable to the one of Intercites or TER trains. In the next step we aim to estimate the speed by train type and hence there will be a downward bias in the estimated (high-speed) TGV speed.[26]

**Step 3**: We compute each year's travel time for all pair of stations connected with a non-stop service in December 2021. We have computed it following two different ap-

followed in the shortest path and the path that Google Maps provides as train service in April 2022.

[24]Distance threshold was chosen by doing visual inspection of connections that exist according to Google Maps but are not included in our dataset.

[25]The maximum amount of iterations was chosen discretionally. A further analysis should be done to choose the threshold by jointly minimizing false negatives and false positives.

[26]In order to correct this measurement error, in the future we plan to distinguish those station pairs which are operated fully at high-speed TGV from those that are not.

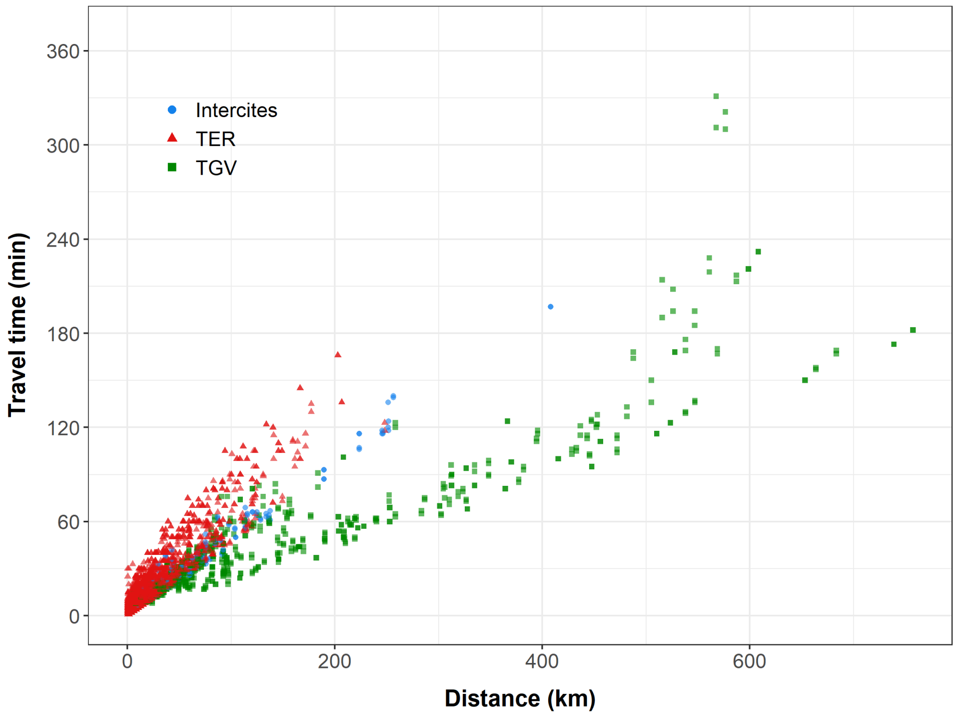



Figure 30: Observed Travel Time and Computed Distance

proaches. Both approaches use predicted travel time based on train line distance and contemporaneous train speed by train type (TGV, Intercites and TER), and replaces TGV speed by Intercites speed in the years prior to introducing TGV in that line (or segment). We first estimate one linear model for each train type in which we regress travel time on train line distance using station-pairs connected with a non-stop service in December 2021. In Table 3 we present the results. The coefficient on distance is the inverse of implied speed of each train type.[27] The implied speeds are: TER 90kmh, Intercites 128kmh, TGV 220kmh. We note that in the case some train services are coded as TGV while they only partially operated at high speed, then the estimated implied TGV speed would be biased downwards. We fit the models to obtain predicted travel time in each station-pair or train line segment, for each train type. We then use the predicted travel times to compute travel time in the two approaches as follows:

The first approach (A) is computed at the station-pair level and only uses predicted

[27] Implied speed in kmh = $\frac{1}{\text{coefficient in minutes}/60}$

| Dependent Variable: | Travel Time (min) | | |
|---|---|---|---|
| | Intercites | TER | TGV |
| Model: | (1) | (2) | (3) |
| *Variables* | | | |
| Intercept | 5.30*** | 1.45*** | 9.21*** |
| | (0.51) | (0.07) | (0.94) |
| Distance (km) | 0.47*** | 0.67*** | 0.27*** |
| | (0.01) | (0.00) | (0.00) |
| *Fit statistics* | | | |
| Observations | 214 | 8,624 | 696 |
| $R^2$ | 0.97 | 0.87 | 0.85 |
| Implied speed (km/h) | 128.11 | 89.88 | 220.75 |

*IID standard-errors in parentheses*
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 3: Travel time and train line distance
The table presents the results of a linear regression of observed travel time between two stations in 2021 on the computed train line distance. The coefficient on distance represents the inverse of speed measured in kilometers per minute. E.g. estimated TGV speed is 1/0.27 = 3.70 km/minute, which is 220 km/h.

travel time for train lines that in December 2021 were high speed railways, using as counterfactual travel time one computed using contemporaneous Intercites speed and train line distance. For all other station-pairs it uses the contemporaneous travel time.

The second approach (B) is computed by working at the segment level, using predicted travel time in each segment according to the fastest train type that is operated in the segment and year. For segments that were TGV in December 2021, it replaces TGV speed by Intercites speed in the years prior to the segment becoming TGV. Using the predicted travel time for each segment and year, we run the Dijkstra algorithm to find the fastest path (in terms of travel time) between each pair of stations connected with a non-stop service in December 2021.

In approach (A) the whole station-pair operates either TGV, Intercites or TER, and follows the same path in each year. In approach (B) a station-pair could be partially

TGV, Intercites and TER, and allows the fastest path to change in each year. However, while approach (A) uses only predicted travel time for a small set of station-pairs, approach (B) uses predicted travel time for all segments and hence all station-pairs. In both approaches the time variation comes only from the opening of TGV lines, the travel time in every other line (segment) remains unchanged.

**Step 4**: We compute each year's travel time between all stations that are contained within the urban area of all departments' prefectures. For each year we use the Dijkstra algorithm to find the fastest path between all prefectures' stations, using as edges the travel time between stations connected with a non-stop service. Then, for each year and prefecture-pair, we take the minimum travel time between all stations contained in the prefecture-pair. For each year we obtain two measures of travel time between prefectures, one using approach (A) and another using approach (B).

## 8.3 Validation exercise

SNCF provides the yearly travel time since 1920 for a selected set of city pairs, most of them being travel time to Paris, Marseille and Lyon. In order to do a validation exercise of our constructed dataset, we restrict it to the same city pairs that is contained in the SNCF dataset.[28] With those city pairs and years we estimate the following regression:

$$\textit{log(observed travel time)}_{odt} = \beta \times \textit{log(predicted travel time)}_{odt} + FE_{od} + FE_{ot} + FE_{dt} + \epsilon_{odt}$$

Where *observed travel time*$_{odt}$ is the travel time that comes from the SNCF data set for origin $o$, destination $d$ and year $t$, *predicted travel time*$_{odt}$ is the predicted travel time (either through approach (A) or (B)), $FE_{od}$ is an origin-destination pair fixed effect, $FE_{ot}$ is an origin-time fixed effect and $FE_{dt}$ is a destination-time fixed effect.

[28] In the future we plan to use the SNCF data set to estimate the Intercites and TER train speed in each year, and then use that estimated speed in our constructed data set.

Table 4 presents the results. Columns (2) and (6) include only the origin-destination pair fixed effect, while columns (4) and (8) include three-way fixed effects. In the four columns, given that the origin-destination pair fixed effect is included, the within-pair across-time variation is used to estimate $\beta$. In columns (2) and (6) $\beta$ is estimated by comparing time variation in one origin-destination pair with the time variation in another origin-destination pair. In columns (4) and (8), the same variation is used but conditioning on overall time changes at the origin and destination. In these four columns, the within R2 represents the variation in the observed travel time which is explained by the predicted travel time, after projecting out the fixed effects. Hence, in the case of columns (4) and (8), 49% and 51% of the within-pair change in travel time is captured by our predicted travel time. The values go up to 86% and 83% if we do not project out the time variation at origin and destination.

| Dependent Variable: | *log(observed travel time)* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Model: | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| *Variables* | | | | | | | | |
| *log(predicted travel time A)* | 1.01*** | 1.03*** | 0.91*** | 0.84*** | | | | |
| | (0.00) | (0.04) | (0.31) | (0.19) | | | | |
| *log(predicted travel time B)* | | | | | 0.97*** | 1.33*** | 0.79 | 1.16*** |
| | | | | | (0.01) | (0.06) | (0.61) | (0.34) |
| *Fixed-effects* | | | | | | | | |
| origin-destination | | Yes | | Yes | | Yes | | Yes |
| origin-year | | | Yes | Yes | | | Yes | Yes |
| destination-year | | | Yes | Yes | | | Yes | Yes |
| *Fit statistics* | | | | | | | | |
| Observations | 1,508 | 1,508 | 1,508 | 1,508 | 1,508 | 1,508 | 1,508 | 1,508 |
| $R^2$ | 0.92 | 0.99 | 0.99 | 1.0 | 0.78 | 0.99 | 0.98 | 1.0 |
| Within $R^2$ | | 0.86 | 0.78 | 0.49 | | 0.83 | 0.44 | 0.51 |

*Clustered (origin-destination) standard-errors in parentheses*
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 4: Observed vs. Predicted Travel Time
The table shows the result of a linear regression of log observed travel time for a pair of cities on the log predicted travel time in the same year. The observations include a selected sample of city pairs for which we observe travel time for the period 1980 to 2018.

Figure 31 shows the evolution of observed and predicted travel time for three city pairs. We observe that both predicted travel times replicate the drops in observed travel time to Paris, although in an attenuated manner. Both connections to Paris opened TGV lines during our period of analysis. However, Bordeaux - Marseille observed travel time of Bordeaux - Marseille is not well tracked by the predicted travel time. The observed travel time drops in 1990, probably because it is routing through segments that use the newly opened TGV segments that are part of the service between Bordeaux and Paris. In this case, our predicted travel time is not capturing the drop. The plot shows that, although the predicted travel time represents to some extent the travel time reductions as consequence of TGV openings, an improvement of the dataset is required, especially to better replicate changes in non-TGV travel times.

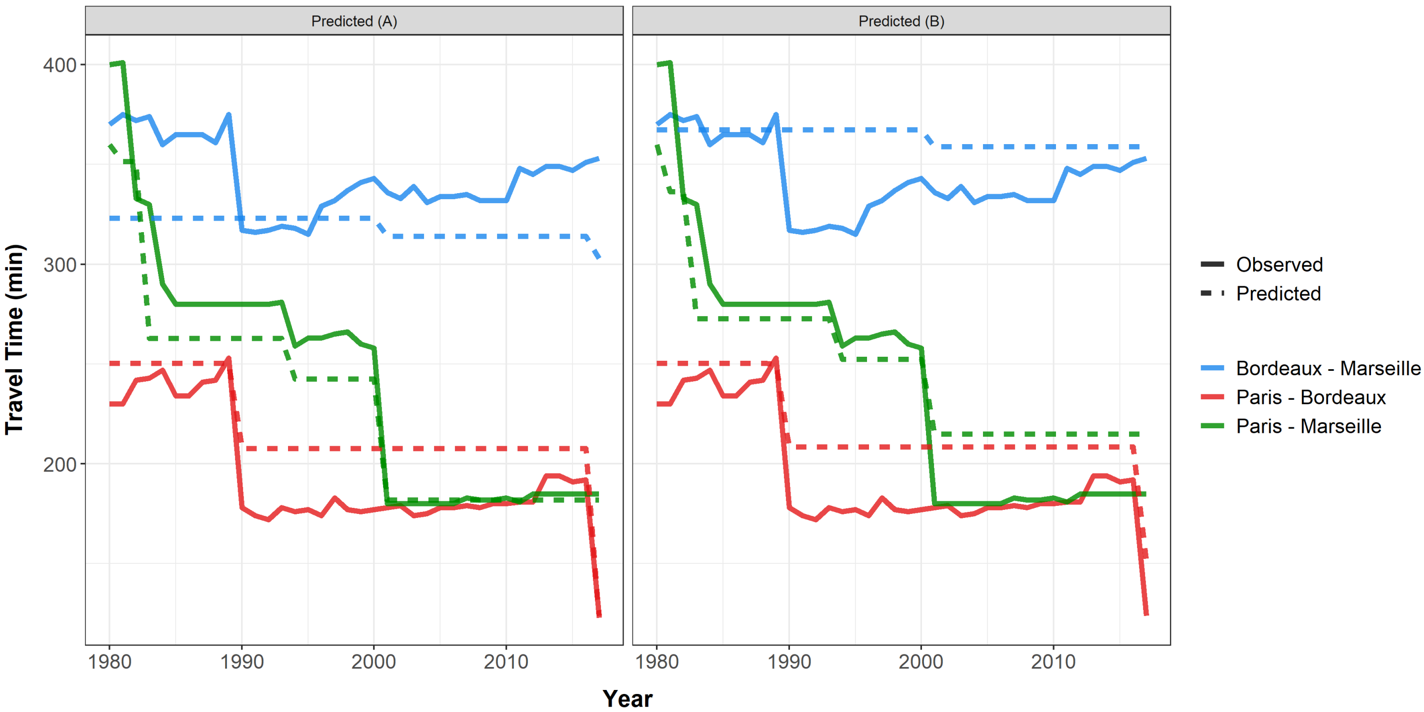



Figure 31: Observed and Predicted Travel Time

## 8.4 Descriptives travel time

Figure 32 presents the within prefecture-pair change in travel time relative to 1980 travel time. We observe that the decrease in travel time is stronger for longer distances. In 2018, travel time for prefectures less than 100km apart was on average 5%

less than in 1980. However, for prefectures located more than 900km apart the average reduction in travel time was around 33%.

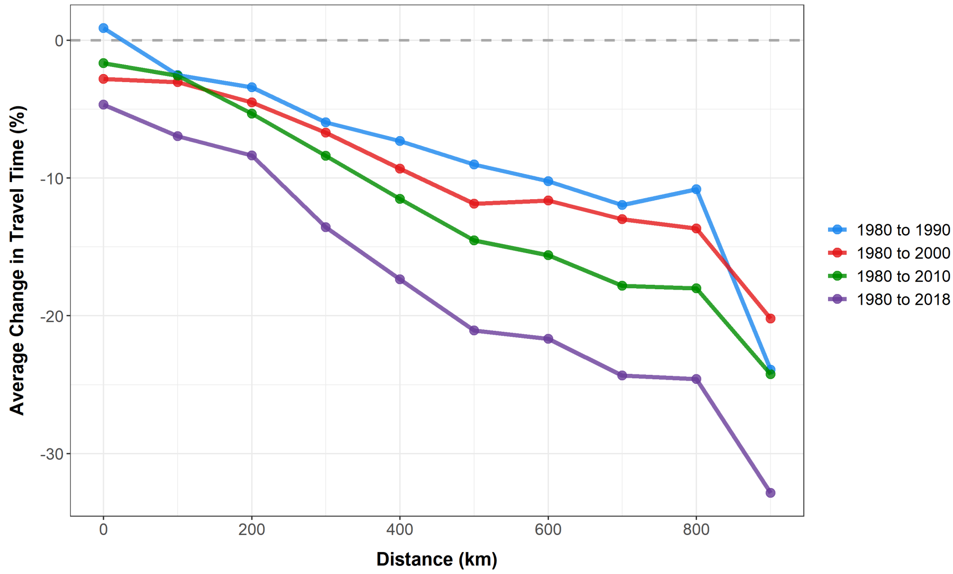



Figure 32: Predicted Travel Time

## 8.5 Next steps to improve the dataset

The first step to improve the dataset will be to identify station pairs directly connected in December 2021 that are fully connected with a high speed railway, as opposite to partially connected. This will reduce our measurement error that leads to a downward bias in the estimation of TGV speed when operating in high speed railway. The second step is to add the Transilien train network. This network would be relevant for locations surrounding Paris. The third step is to improve the quality of the match for lines intersecting or overlapping each other, and identifying train stations within train lines. To do so we plan to use GIS software to create a buffer around train lines, in a similar manner as in Donaldson and Hornbeck (2016). The fourth step will be to obtain a time-varying estimation of train speed for Intercités and TER trains using the observed historical travel times provided by SNCF. This will allow us to better replicate travel time changes in non-TGV lines.

# 9 Conclusion

In this paper we have presented descriptives of the geography of patenting activity in France. We have also introduced a new dataset of train travel time that includes a large reduction in travel time. In the future we plan to exploit changes in travel time as a shifter in communication costs that may have changed the geography of patenting activity. We plan to focus on how changes in travel time affected inter-regional collaboration, the relation of collaboration with technological proximity and its dynamics, inventors' mobility and team formation. Confidential administrative data of balance sheet of firms and matched employer-employee information would complement our analysis. While this research project is still in early stages, we think of it as a potentially fruitful research agenda.

# References

Bergé, L. R. (2015, March). Network proximity in the geography of research collaboration. Papers in Evolutionary Economic Geography (PEEG) 1507, Utrecht University, Department of Human Geography and Spatial Planning, Group Economic Geography.

Charnoz, P., C. Lelarge, and C. Trevien (2018, 05). Communication Costs and the Internal Organisation of Multi-Plant Businesses: Evidence From the Impact of the French High-Speed Rail. *The Economic Journal 128*(610), 949–994.

Donaldson, D. and R. Hornbeck (2016). Railroads and american economic growth: A market access approach. *The Quarterly Journal of Economics 131*(2), 799–858.

Jaffe, A. B. (1989). Characterizing the technological position of firms, with application to quantifying technological opportunity and research spillovers. *Research Policy 18*(2), 87–97.

Jaffe, A. B., M. Trajtenberg, and R. Henderson (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics 108*(3), 577–598.

Morrison, G., M. Riccaboni, and F. Pammolli (2017). Disambiguation of patent inventors and assignees using high-resolution geolocation data. *Scientific data 4*(1), 1–21.