Economies of Scale and Scope in Railroading*

Pedro G. Degiovanni† and Ron Yang‡

November 11, 2023

Latest version.

Abstract

To what extent do transportation costs depend on the amount shipped, and how does infrastructure investment shape these costs? We model railroads as multiproduct firms and estimate the link between capacity utilization and costs using firm choices, the network structure of production, and publicly available routing data. We find a U-shaped relationship between marginal costs and rail utilization: As utilization increases, costs decrease by 30% to a low point at 55% utilization, before increasing by another 30%. Increased congestion in the rail network can explain a third of the 50% increase in real rail prices observed since 2004. We use our framework to study two normative and one positive policy questions: First, we estimate the network externalities of rail infrastructure investment, finding that investment in Arizona provides the highest returns, but only 3% are captured by Arizona itself. Next, we evaluate the cost efficiencies that would arise from a merger between Union Pacific and Burlington Northern Santa Fe. We find that such a merger would reduce costs by 17.1% due to reduced misallocation and process innovation. Lastly, we study the effect of the China shock on freight costs. We show that the reallocation of imports

*We thank Manuel Amador, Pol Antras, Yanyou Chen, Dagny Dukach, Ed Glaeser, Mark Fagan, Jose Gomez-Ibanez, Doireann Fitzgerald, Jeff Gortmaker, Myrto Kalouptsidi, Gabriel Kreindler, Robin Lee, Marc Melitz, Ariel Pakes, Chris Walker, Leisy Zhang, and numerous colleagues and workshop attendees for their suggestions and comments. We also gratefully acknowledge support from the Chae Family Economics Research Fund which helped finance the project. All errors that remain are our own.

†Ph.D. Candidate, Harvard University.

‡Assistant Professor in the Strategy and Business Economics Division at the University of British Columbia's Sauder School of Business.

toward the West Coast led to a 3% increase in shipment costs in Los Angeles and Chicago, with heterogeneous effects across space and firms.

Keywords: Railroads, cost function estimation, routing problem, multiproduct firms.

JEL Classification: D24, L92, R41

1 Introduction

Transportation costs are key in shaping the spatial distribution of economic activity and trade.[1] These are not just technological parameters: they are complex endogenous objects arising from the optimizing behavior of firms across multiple modes, and as such they respond to changes in demand, market structure, and policy.[2] A recent literature has studied the role of public action in shaping these costs, driven by renewed interest in infrastructure investment at the federal level.[3] Rail features prominently in this discussion, as it is a growing sector that presents a low-emissions alternative to trucking. In this paper, we study the structure and determination of transportation costs in US railroads.

Railroads are multiproduct firms, jointly producing shipping services between various city pairs. They do this by utilizing a rail network that they own and operate. The production of these services features economies of scale, both positive—a full train is more efficient than a half-empty one—and negative—the more traffic there

[1] For a comprehensive review of the literature, see Redding and Turner [2015].

[2] These modes range from atomistics trucks and freight ships operating in a competitive (yet frictional) market (see Yang [2023], Brancaccio et al. [2020])), to large oligopolies controlling a large rail network.

[3] This literature, spearheaded by Allen and Arkolakis [2022], Fajgelbaum and Schaal [2020] and, more recently, Fuchs and Wong [2022], models transport costs as arising from the structure and utilization of the transportation network, with a focus on identifying the specific links where additional investment would be the most productive. Their models feature either a completely decentralized transportation system with infinitesimal shippers which take transport costs as given, or a social planner with complete knowledge and control over the flows.

is, the harder it is to efficiently operate it.[4] Joint production of shipments gives raise to economies of scope: shipments share the same tracks and rail yards, generating cost spillovers on each other. These economies of scale and scope will shape how costs and conduct react to shocks, as well as the effectiveness of regulation and infrastructure investment.

We estimate costs in rail, using rich data and a flexible model that accounts for these economies of scale and scope. We use our estimates to conduct three policy exercises: First, we show how the network structure of production implies that infrastructure investment engender substantial network spillovers, which local and state governments are unlikely to internalize. Next, we study the effect of a merger between the two largest American railroads: the merged firm is better able to exploit economies of scale and scope, leading to cost efficiencies of 17%. Finally, we examine how a localized demand shock – the accession of China to the World Trade Organization (WTO) – led to heterogeneous changes in rail costs across firms and space.

We study the US railroad system using shipment-level data from 1998 to 2018 to infer the routes that railroads use to ship products through their networks. From these inferred routes, we document three empirical patterns. First, railroads do not minimize distance: 40 percent of shipments deviate from the shortest possible path, and the modal route of these deviating shipments is 20 percent longer than the shortest possible path.[5] Second, deviations from the shortest path are correlated with capacity constraints.[6] For every given city pair, railroads have a preferred route which they use almost exclusively when utilization is low. But when capacity utilization along this main route exceeds 50%, railroads become increasingly likely to use multiple routes. Third, spot-market prices for rail shipments exhibit a U-shaped relationship with capacity utilization. These three findings challenge the

[4]Since the advent of rail regulation in the late 19th century, economists have been interested in the existence and importance of economies of scale in rail transportation. A wide-ranging literature has followed, analyzing the topic from both theoretical (see, for example, Wellington [1893], Ripley [1927], Jones [1931], Daniels [1932]) and empirical (see, among others, Lorenz [1916], Borts [1952], Borts [1954], Meyer et al. [1961], Borts [1960], Griliches [1972], Keeler [1974], Braeutigam et al. [1982], Jara-Díaz and Cortés [1996]) perspectives to explore whether costs are constant or decreasing in quantity.

[5]In contrast, Fuchs and Wong [2022] assume that all shipments follow the shortest feasible path between origin and destination.

[6]For each segment of the rail network, we define capacity as the highest observed monthly flow through that segment plus 10%.

traditional view of railroads as an industry characterized by increasing returns to scale, and provide evidence of congestion at high utilization levels.

Based on these three findings, we build a model of freight railroads as multiproduct firms selling the service of transporting carloads between city pairs. Each rail firm owns a rail network, represented as a directed graph in which track segments are edges and rail yards are nodes. Railroads choose routes to minimize the total cost of shipping through their network. We decompose the marginal cost of shipping an additional unit on a given route into three components: First, it increases the flow in all edges belonging to the route, incurring an edge marginal cost. Second, it increases the flows going in the opposite direction for all edges belonging to the reverse route. Finally, it increases the amount of traffic through all the nodes belonging to that route, incurring a node marginal cost. Each of these marginal costs is in turn a flexible function of both capacity utilization on that segment and other track characteristics. Because each firm owns and operates their own infrastructure, they will internalize any spillovers they impose on their own shipments. Their choices will then inform us on the size of these spillovers. We focus on cases for which we observe a firm using multiple routes to fulfill shipments with the same origin and destination and show how cost minimization implies that the cost of shipping a marginal unit through each route must be the same.

We use this indifference condition to construct a set of moment conditions that we can take to the data. We use the Generalized Method of Moments to estimate the cost function of the two largest American railroads: Burlington Northern Santa Fe (BNSF) and Union Pacific (UP). These railroads collectively account for 50% of industry revenue and handle most traffic west of the Mississippi River. To control for endogeneity in firms' routing choices, we use the 2010s post-shale-boom decline in coal production as an exogenous demand shock to construct a shift-share instrument.

We find evidence of substantial returns to scale and congestion in rail: Marginal costs are U-shaped, with a minimum around 55% utilization, and imply significant economies of scale. Traversing an edge with 0% utilization is $2.00 more expensive than a shipment at 55% utilization. As utilization increases past this efficient level and approaches 100%, costs increase by $1.25. These differences represents 40% and 25% of the average price for a carload-mile during this period. We also find that infrastructure investment is effective in reducing costs, with an additional set of

parallel tracks leading to a decrease in costs of 40%.

Finally, we show how our framework can inform answers to both positive and normative policy questions. We start by analyzing the impact of public investment in rail infrastructure. Many state and local governments fund rail infrastructure investments, aiming to reduce road traffic congestion for motorsits and reduce the cost of shipping to and from their locations.[7] The network structure of production implies that these infrastructure investments engender network externalities, which local and state governments are unlikely to internalize. We measure the cost savings of upgrading rail networks for each US state and find substantial heterogeneity across states: The return on investment for the most productive state (Arizona) is 25 times larger than that of the least productive state (South Dakota). In addition, we estimate that only 3% of the total benefits accrue to Arizona, with most of the gains being captured by California, Texas and Illinois. This suggests that decision-makers may fail to internalize the benefits from infrastructure investment, highlighting the potential role of the federal government in coordinating rail investment.

Next, we study the effect of merging UP and BSNF. Such a merger could potentially reduce costs in three ways: First, the merged firm could better coordinate the flows between each firm's network, redirecting traffic to the least congested areas. Second, the joint firm would have access to new, otherwise inaccessible routes. Finally, the merger would reduce contractual frictions that currently arise when using trackage rights in each other's networks. We find that such a merger would reduce costs by 17.1% compared to the decentralized equilibrium.

Finally, we study the effect of China's accession to the WTO ("the China shock") on domestic U.S. transportation costs. We find that the increase in Chinese goods arriving in West Coast ports led to a doubling in rail flows originating in Los Angeles. The aggregate effect of the shock is small, but there is substantial heterogeneity across both space and firms. We find that Los Angeles experienced the largest increase in costs (around 4%), but regions located between Los Angeles and Chicago also experienced increased costs. Conversely, we find that UP was previously operating below its optimal capacity, and as a result, the cost of their shipments decreased after the shock.

Our work contributes to the literature on transportation costs across networks.

[7] Two large city-level investments are the Alameda Corridor, built in Los Angeles in the early 2000s, and CREATE, an ongoing project in Chicago. Both of these resulted from a partnership between railroads, city governments and, in the case of CREATE, the state and federal governments.

Allen and Arkolakis [2022] model transportation costs as endogenous and dependent on the traffic flowing through each edge in the network. They then use US data to compute the welfare benefits from improving the efficiency of individual edges in the network using hat algebra, with the goal of directing infrastructure investment towards its most efficient applications. Fajgelbaum and Schaal [2020], on the other hand, make stronger assumptions to obtain an expression for the optimal infrastructure network as a function of the data. Fuchs and Wong [2022] build on this foundation by adding a mode-choice decision, explicitly modeling US road, rail, and ports with a focus on the intermodal facilities that allow for switching between networks. This literature has primarily featured infinitesimal agents that take transportation costs as given, while we study large oligopolistic firms that internalize their effect on costs.

This paper also contributes to the well-established literature on economies of scope within multiproduct firms. Despite the prevalence and importance of multiproduct firms in both developing and developed countries,[8] multiproduct firms have been challenging to study due to the absence of a production function, compounded by the scant data on the products produced by firms or the allocation of inputs across them. The theoretical underpinnings have been shaped by contributions from Panzar and Willig [1975], Panzar and Willig [1981], and Teece [1980]. Empirically, the estimation techniques have involved utilizing firm-level cost data,[9], or leveraging information on demand See Ding [2022], Argente et al. [2020], and Khmelnitskaya et al. [2023]. Unlike most of the prior literature, we observe a fine mapping from inputs (tracks) to outputs (shipments), as well as detailed information on the various ways a product can be produced (routes). This allows us to use a more flexible, revealed preference approach to estimation.

Another body of literature has taken a structural approach to studying individual segments of the transportation network, including ports (Bailey [2021] and Ducruet et al. [2020]), trucking (Yang [2023] and Allen et al. [2023]), ocean shipping (Brancaccio et al. [2020] and Ganapati et al. [2021]), river shipping (Caris et al. [2014]), and air freight (Feng et al. [2015]). The most similar paper is Chen [2023], which also studies oligopolistic competition by railroads. In his model, Chen assumes a constant marginal cost per mile, which can only be altered by investment in

[8] See Bernard et al. [2010] and Goldberg et al. [2010].

[9] As evidenced by Hall [1973], Kohli [1981], Brock [1983], Johnes [2012], directly estimating the cost functions using quantity and input data[10] and Zhang and Malikov [2022].

maintenance. The efficiency of this investment is then used to measure economies of scope, since it reflects the benefits of consolidating flows. In contrast, we allow the cost per mile to depend on the flows and the observed rail infrastructure. Although we do not explicitly feature investment in our model, our results allow us to recover the effects of altering specific rail characteristics, such as the slope of the terrain or the number of parallel tracks.

Section 2 presents a brief description of the American railroad industry. Section 3 describes our data and details how we obtain routing choices from shipment level characteristics. In Section 4, we describe a series of reduced form findings, and in Section 5, we construct a model of railroad choices. Section 6 presents our estimation strategy, Section 7 documents our results, Section 8 reviews three counterfactual exercises, and finally, Section 9 concludes.

2 Industry Background

The United States has the largest freight rail network in the world, with over 140,000 miles of track. This network was constructed in the 19th and early 20th centuries, and it reached its peak size in 1917. Following World War II, the widespread adoption of cars, together with the development of commercial flight and trucking, reduced rail's market share and led to half the network being abandoned. This decline lasted until railroad deregulation in the 1980s, after which traffic started growing once more.[11] Since then, the industry has boomed, but starting in 2004, real prices have surged, climbing 50% by 2018 (see Figure A.1 in the Appendix).

Today the US freight railroad industry is composed of for-profit firms which own and operate rail infrastructure,[12] with six major firms operating the vast majority of all rail shipping.[13] In this paper, we focus on the two largest firms: Burlington Northern Santa Fe (BNSF) and Union Pacific (UP). These railroads have a combined

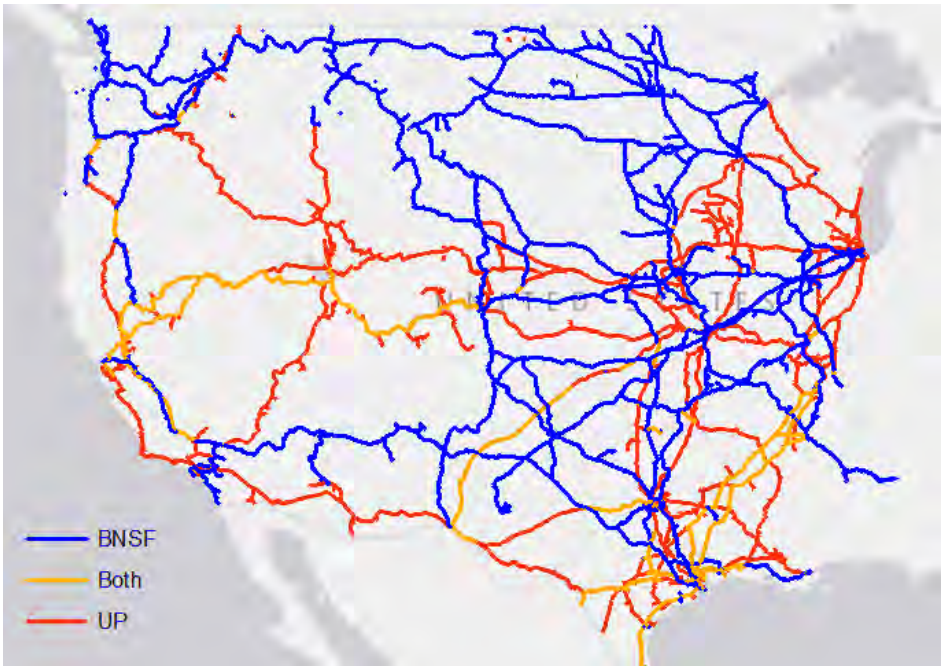
[11]The network has been stable and slowly shrinking over the past 20 years, having decreased by 4 percent between 1999 and 2014. In our analysis, we will consider the rail network as fixed, although railroads have the option of "abandoning" a line by setting the quantities transported through it to zero.

[12]In contrast, most other developed economies feature either partial or complete nationalization of the sector.

[13]The six "Class I railroads" are Union Pacific (UP), Burlington Northern Santa Fe (BNSF), CSX Transportation (CSX), Norfolk Southern (NS), Canadian Pacific Kansas City (CPKC), and the Canadian National Railway (CN). We offer a brief history of the industry and regulations in Appendix A.

market share of 50%, and they handle most of the traffic west of the Mississippi River (we depict their networks in Figure 1).

Figure 1: *Networks for BNSF and UP.*



"Both" depicts tracks that are owned by one of the two firms, but that the other one can access via trackage rights. Source: NTAD.

Railroads are multiproduct firms, jointly producing shipping services between various city pairs using several types of inputs: labor, fuel, capital goods (locomotives and cars), and rail infrastructure. In this paper, we will focus on this last input: the utilization of rail infrastructure, including tracks and yards. In Section 3, we show how to infer how these inputs are used to produce individual shipments.

The basic unit of analysis for this paper is a carload: a railroad car containing some commodity that needs to be shipped from some origin to some destination. Firms sell this service to customers through a mix of spot-market transactions and long-term contracts.

Railroads primarily ship low-cost, long-distance commodities. Modern locomotives can achieve fuel efficiencies of around 400 miles per gallon of diesel, making them significantly more fuel-efficient than trucks. However, their maximum speed is lower than that of trucks, and their movement is limited to existing rail infras-

tructure. Commodities with the highest rail market share include coal, agricultural products, motor vehicles, and, increasingly, containers.[14]

Economies of scale and scope. In this paper, we consider a railroad to exhibit economies of scale if the marginal cost of moving a shipment through a track or yard depends on the quantity shipped through that track or yard. We will refer to these economies of scale as "returns to scale" when marginal costs decrease with quantity, and as "congestion" when marginal costs increase with quantity. We consider a railroad to feature economies of scope if the marginal cost of producing a shipment between one origin-destination pair depends on the quantity shipped between other origin-destination pairs.

Returns to scale can be generated by lumpy or fixed costs. For example, a locomotive and a full crew can pull anywhere between one and one hundred freight cars, thus leading to decreasing marginal costs for the first hundred units shipped. Unit trains can also lead to returns to scale: If a client ships a whole train's worth of cargo from i to j, the train can avoid stopping at yards in the way. Finally, Lai and Barkan [2009] document that the more similar trains are to each other in terms of their length and power-to-weight ratios, the more efficiently traffic operates—this is consistent with the law of large numbers, which suggests that scale should lead to more efficient traffic management.

As the quantity of shipments increases, tracks can also become congested. Whenever a faster train encounters a slower train, the fast train must reduce its speed and wait until it reaches an "overtaking siding," where the slower train can move aside and allow the faster train to pass. Similarly, most tracks support traffic in both directions, so if two trains are about to cross paths, one of them must stop on a "passing siding" until the way is cleared. Outages also play a role: If a locomotive breaks down, all trains on the same track must wait until it's repaired or removed.

Congestion is also present at the yard level. Yards play the role that connecting airports play in passenger flights: Upon arrival at a yard, trains (locomotives plus their cargo) are broken up and reassembled, so that the cargo can reach its ultimate destination. The more traffic, the more cumbersome this process becomes.[15]

[14] In 2012, containers (i.e., any manufactured good transported in a container) surpassed coal as the most common commodity on rail.

[15] Famously, Union Pacific faced a massive yard jam after its merger with Southern Pacific in 1997. The sudden increase in the scale of its operations caused its yards to be gridlocked for months, with losses estimated at $100 million.

3 Data

3.1 Data Sources

Our main dataset is the confidential Carload Waybill Sample (CWS), provided by the US Surface Transportation Board. This dataset contains a 2.7% stratified sample of all waybills from Class I railroads. A waybill is a receipt created by a railroad for every carload of freight that travels on its rails. It provides, among other information, the carload's origin, destination, and trip distance; the commodity transported, its quantity, and price charged; and a list of all the railroads involved in the trip. The CWS contains around 600,000 waybills per year from 1996 to 2018,[16] but we focus on just 2015 to 2018 in our analysis to take advantage of improved routing data (see Appendix E for more details).

Throughout the paper, we use a variety of additional sources. Most notably, we use the North American Rail Lines shapefile provided by the National Transportation Atlas Database (NTAD) to measure infrastructure characteristics, such as the average number of tracks, the frequency of rail sidings, and the slope of the terrain. In addition, we leverage monthly coal production data from the Energy Information Administration, as well as imports data at the country level (provided by the World Bank's World Integrated Trade Solution) and at the port level (provided by the US Census Bureau).

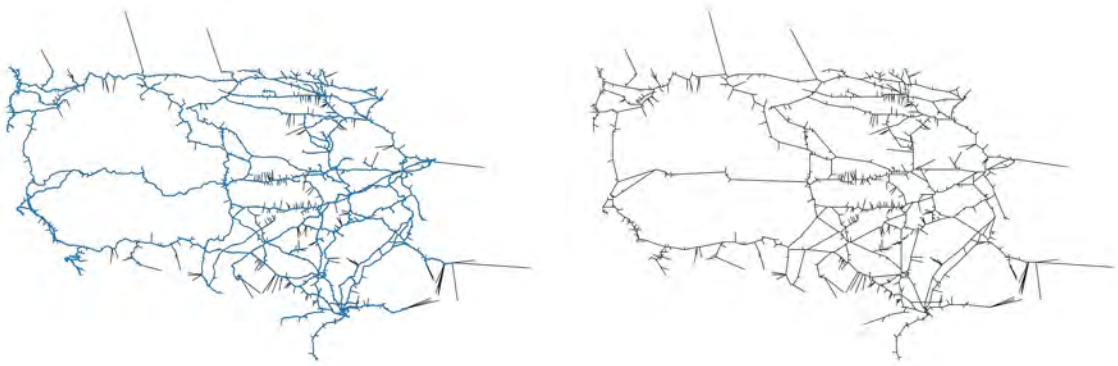
3.2 Inferring routes

The CWS does not directly provide railroad's routing choices. Instead, it tells us the endpoints of a given trip, the distance traveled, and all the states visited during the trip. We use the fact that the rail network is relatively sparse to invert this information and recover the routes taken.

We start by constructing a directed network that reflects the rail infrastructure available to each firm. We detail the steps we take in Appendix B. Figure 2 depicts

[16] Around 90% of waybills are fulfilled as part of contracts, as opposed to spot-market transactions. Other than a contract flag, our data does not provide any additional information on these contracts, which can often be quite complex: They tend to be long-term and include investment requirements by the railroad and/or the shipper, minimum quantity requirements for specific time periods, rebates, etc. A growing literature (Government Accountability Office [2006]) has documented that the CWS reported price is a poor indicator of the actual prices charged by the firm. As such, in our analysis, we refrain from using pricing data for estimation.

Figure 2: *BNSF Original (left) and simplified (right) networks*



the original network on the left, and our graphical representation on the right.

For each origin-destination pair, we observe multiple combinations of distance traveled and states visited, with each combination representing a different route.[17] For each of these origin-destination-route tuples, we define a subnetwork containing the tracks belonging to the visited states. We plot one such subnetwork in Figure 3. Next, we compute the k-shortest loopless routes between origin and destination which are fully contained within that subnetwork. We progressively increase k until we find all routes within 20 miles of the target distance.

If there is only one such route, we conclude that it is the route implied by the origin-destination-route tuple. In most cases, there will be multiple routes within the target distance; if that happens, we assume that all such routes are equally likely to be the actual route. We then compute the probability that an edge or node belongs to a given route by looking at the proportion of potential routes within the target distance that includes that edge or node.

As an illustration, Figure 4 plots the inferred route for a 2500-mile shipment originating in Los Angeles, terminating in Chicago, and visiting California, Arizona, New Mexico, Texas, Oklahoma, Missouri, and Illinois. The width and color of an edge indicate the probability that a given edge belongs to the actual route. We repeat this process for all routes in the sample to build a dataset that is—to the best of our knowledge—the first publicly available dataset of US railroad route choices.[18]

[17] In order to simplify our network, we combine all rail within each 0.2 decimal degrees square. Each of these squares has a diameter of approximately 20 miles. As a consequence, we round all distances to the closest multiple of 20 when computing routes.

[18] The dataset is available upon request, and it will be available in our websites in the future.

Figure 3: *Subnetwork containing the states of California, Arizona, New Mexico, Texas, Oklahoma, Missouri, and Illinois.*



Routing choices. Our analysis relies on observing the origin, destination, states visited, and distance of all routes taken by the firm. Since this is such an important variable, we will briefly discuss how it is created and how it informs our approach to analysis: First, Railinc receives the origin, destination, departure date, and railroads involved in the shipment. This information is then matched to the Railinc Event Data dataset, which contains the locations of individual cars over time. Around 60% of all trips can be directly matched to this second dataset, and when the two datasets cannot be matched, Railinc matches trips to comparable waybills (i.e., waybills that share the same origin, destination, commodity, and interchanges), and then computes the median distance of matching trips. Finally, if there are no comparable trips, Railinc uses its own cost-minimizing model to compute a route for the waybill.

Figure 4: *Inferred route*



Inferred route for a 2500-mile shipment originating in Los Angeles, terminating in Chicago, and visiting California, Arizona, New Mexico, Texas, Oklahoma, Missouri, and Illinois. The width and color of an edge indicate the probability that a given edge belongs to the actual route.

4 Evidence for decreasing returns to scale

In this Section, we use the data described in Section 3 to document three facts about railroads' routing choices which will inform our subsequent analysis.

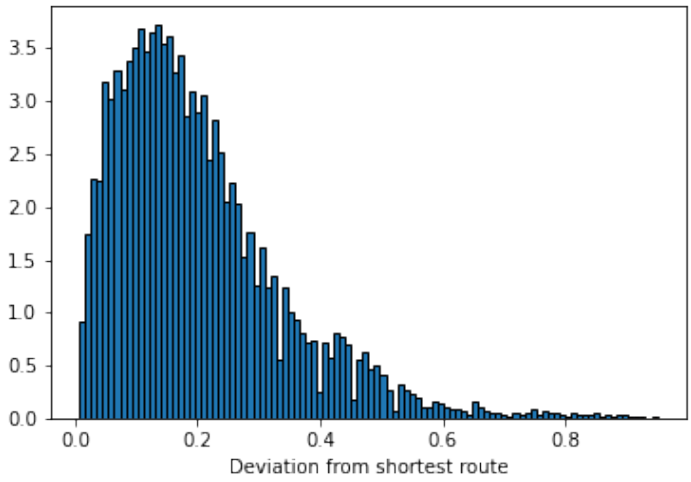
4.1 Railroads do not minimize distance

We start by comparing the route taken by a shipment to the shortest path between its origin and destination. We measure the extent of the deviation from the shortest route by dividing the "excess" miles traveled by the length of the shortest path.[19] Figure 5 plots a histogram of this deviation measure, conditional on it being strictly positive, showing that 40 percent of shipments take a route longer than the shortest path. Among these shipments, the modal shipment takes a route 14% longer than the shortest path. There is also a long right tail, with 10% of shipments taking a

[19]For example, if i and j are 100 miles apart by rail, and the railroad chooses a 150-mile trip, that trip has a deviation of $(150 - 100)/100 = 50\%$.

route 35% longer than shortest path. This indicates that railroads trade off distance with other factors when making their routing choices. Next, we examine what these other factors may be.

Figure 5: *Deviation from shortest path. Observations correspond to shipments, and are weighted by their sampling weights.*



4.2 Capacity utilization predicts route choices

We first explore whether capacity constraints are relevant for routing choices. We do not directly observe the capacity of tracks and yards. Instead, we use the largest observed monthly flow through each track or yard to estimate capacity. To allow for the possibility of measurement error, we assume that actual capacity is 10% greater than this highest observed monthly flow. Letting $F_{uv,t}$ be the amount flowing through track going from u to v at time t, we define its capacity K_{uv} as:

$$K_{uv} = 1.1 \max_t F_{uv,t}$$

(and likewise for yards). Around 80% of tracks and yards (representing 95% of flows) reached their maximum observed flows before 2006, but very few do so during our period of analysis (2015-2018).

We define a track's "capacity utilization," $f_{uv,t}$ as the observed monthly flow

divided by its capacity:

$$f_{uv,t} = \frac{F_{uv,t}}{K_{uv}}$$

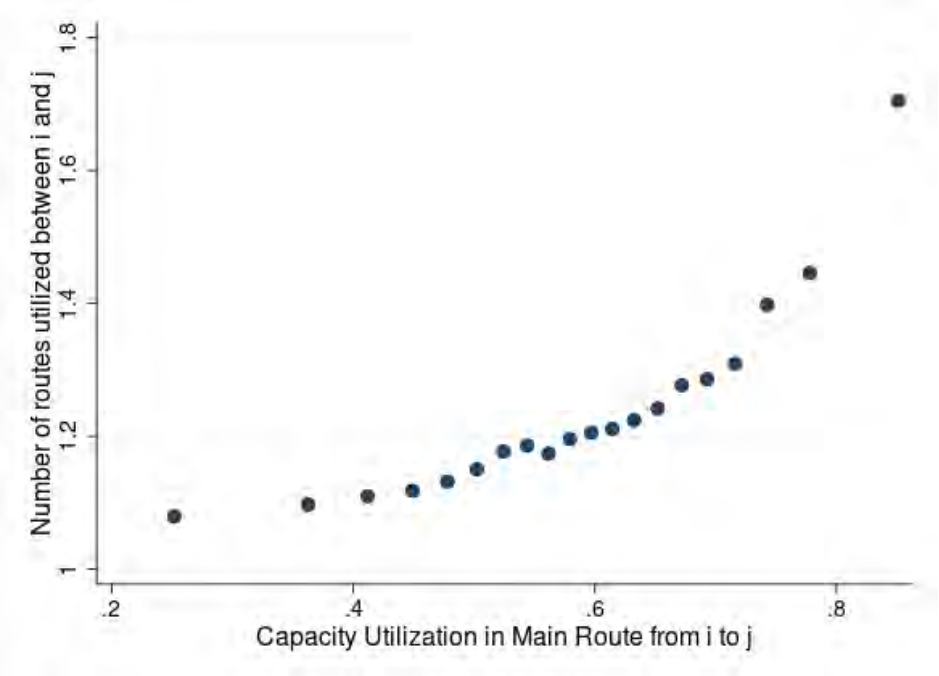
Likewise, we define a route's capacity utilization as the average track utilization for edges belonging to that route:

$$f_t^i = \sum_{uv \in i} \frac{1}{\sum_{uv \in i} 1} f_{uv,t}$$

Finally, for each origin-destination pair, we define its "main route" as the route most commonly used to service that pair.

In Figure 6, we plot the number of unique routes used to service a city pair against the capacity utilization of its main route. We find that firms use their preferred route almost exclusively when its utilization is below 40%. However, as utilization for the main route increases, they become increasingly likely to use multiple routes to service the same city pair. This behavior is consistent with increased capacity utilization leading to congestion. As such, we model firm costs as directly dependent on capacity utilization of their infrastructure.

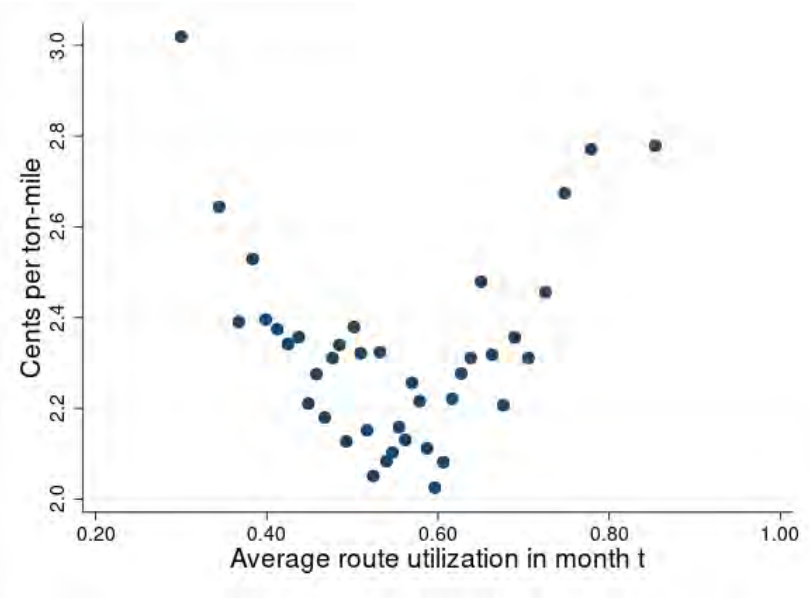
Figure 6: *Number of routes utilized versus utilization of main route.*



4.3 Capacity utilization is correlated with prices

Our final fact relates to the relationship between capacity utilization and prices. We look at the spot market for shipments—that is, the approximately 10% of shipments which are not subject to long-term contracts. Figure 7 plots a binned scatter plot of these prices against the capacity utilization of the route used to fulfill each shipment.

Figure 7: *Spot market prices and capacity utilization.*



Spot market prices and capacity utilization for BNSF between 1996 and 2018. We trim the top 1% of observations in terms of price, and the bottom 1% in terms of quantity. Each observation is weighted by the inverse of its sampling probability.

We find a U-shape relationship between spot market prices and capacity utilization: Prices decrease with capacity utilization until 50% utilization, and increase thereafter. In particular, minimum prices are 33% lower than prices at lower utilization.[20] This relationship need not be causal, and it likely reflects a combination of varying markups and costs across space. In order to let our model inform us on this point, we will let marginal costs be a convex function of capacity utilization.

[20]Figure 7 includes no controls. If we add commodity and time fixed effects, the relationship is attenuated but remains significant. Origin and destination fixed effects leave an upwards sloping curve, without the declining part of the U-shape.

5 Model

We develop a model in which railroads choose routes to minimize costs, taking shipping demand and the behavior of other railroads as givens. Motivated by our descriptive results, we allow the cost function to depend flexibly on each rail segment's traffic, with either increasing or decreasing returns to scale at different levels of capacity utilization.

We begin by describing the rail network and the railroad's production function. Next, we present the railroad's cost minimization problem. We then describe additional assumptions related to the railroad's cost function before deriving an indifference condition implied by cost minimization, which will be useful for estimation.

5.1 Preliminaries

Network. The railroad owns and controls a rail network, represented by a symmetric directed graph $G = (V, E)$, with $|V| = N_N$ nodes and $|E| = N_E$ edges.

Routes. To deliver a shipment from origin o to destination d, the railroad uses sequences of edges, which we will call routes. Each route i is defined by two indicator vectors, R^{iE} and R^{iN}, where $R^{iE}_{uv} = 1$ if edge uv belongs to the route, and $R^{iN}_{uv} = 1$ if node n belongs to the route. There are N_{od} distinct, non-looping routes between o and d.

5.2 Firm problem

In each month t, the railroad chooses a matrix of route choices, $\mathbf{X}$, to minimize costs given a quantity vector Q_t where $X_{od,it}$ is the quantity (in carloads) the firm decides to ship between o and d via route i at time t, and $Q_{od,t}$ is the quantity that the firm must ship from o to d. The firm's problem is:

$$\min_{\mathbf{X}} C(\mathbf{X})$$

subject to

$$0 \leq X_{od,it} \ \forall i, od \tag{1}$$

$$\sum_{i=1}^{N_{od}} X_{od,it} = Q_{od,t} \ \forall od \tag{2}$$

Constraint (1) requires the firm to ship a nonnegative quantity along each route. Constraint (2) ensures that the quantities shipped across all routes are sufficient to fulfill demand.

This formulation represents the problem faced by the head of the Operations Department in a Class I railroad: She knows how much needs to be shipped where, she must determine the cheapest way to do so. This assumes that the firm takes demand as given when making routing choices, and is consistent with the behavior documented by Williams [2022] in the airline industry and Chen [2023] in the railroad industry. In Appendix C, we show how with modest assumptions regarding demand, we can extend the problem to allow for joint determination of routing choices and quantities. In particular, we can allow consumers to prefer certain routes to others, as long as their inverse demand is separable in quantities and route characteristics. This condition is satisfied by logit demand, which is the most common specification used in empirical analyses of the industry (see, for example, Chen [2023]).

5.3 Cost Function

We start by imposing structure on the cost function. We assume that $C(\mathbf{X})$ equals the sum of the cost of traversing each edge and node in the network. We model the cost of traversing an edge as a function of the flows through that edge in both directions:

$$c_{uv,t} = c_{uv}(F_{uv,t}, F_{vu,t})$$

where c_{uv} is an edge-specific and time-invariant function. Edge traffic $F_{uv,t}$ sums routing choices over all routes which use edge uv,

$$F_{uv,t} = \sum_{od} \sum_{i=1}^{N_{od}} X_{od,it} R_{uv}^{iE}$$

Similarly, we let the cost of traversing a node to be a function of the traffic through that node:

$$c_{n,t} = c_n(F_{n,t})$$

where node traffic $F_{n,t}$ sums over all routes which use node n,

$$F_{n,t} = \sum_{od} \sum_{i=1}^{N_{od}} X_{od,it} R_n^{iN}$$

Note that by allowing both c_{uv} and c_n to be nonlinear functions of quantity, we are baking in economies of scale at the edge and node level.

Thus, we rewrite the cost function as:

$$C(F(\mathbf{X})) = \sum_{uv} c_{uv}(F_{uv,t}(\mathbf{X}), F_{vu,t}(\mathbf{X})) + \sum_{n} c_n(F_{n,t}(\mathbf{X})) \tag{3}$$

5.4 Indifference Condition

In this subsection, we show how to use the cost function to derive an indifference condition that will allow us to estimate the model using routing choices.

We start by rewriting (3) as a Lagrangian incorporating the restrictions stated in equations (1) and (2). Dropping the time subindex for simplicity, we have:

$$\min_{\mathbf{X},\lambda} \mathcal{L} = \tilde{C}(F(\mathbf{X})) + \sum_{od,i} \lambda_{1,od,i} X_{od,i} + \sum_{od} \lambda_{2,od} [Q_{od} - \sum_{i} X_{od,i}]$$

The first order condition for a generic choice $X_{od,i}$ is given by:

$$\frac{\partial C(F(\mathbf{X}))}{\partial X_{od,i}} + \lambda_{1,od,i} + \lambda_{2,od} = 0$$

Consider an origin-destination pair od for which more than one route is used—say, routes i and j. In that case, constraint (1) is not binding, and its associated Lagrangian multipliers equal zero. The FOCs for i and j reduce to:

$$\frac{\partial C(F(\mathbf{X}))}{\partial X_{od,i}} + \lambda_{2,od} = 0 \tag{4}$$

$$\frac{\partial C(F(\mathbf{X}))}{\partial X_{od,j}} + \lambda_{2,od} = 0 \tag{5}$$

Subtracting (5) from (4) we obtain the following result, showing that the marginal cost of both routes must equal each other.

Lemma 1. If an origin-destination pair od features two routes with positive flows i and j, their first order conditions satisfy:

$$\frac{\partial C(F(\mathbf{X}))}{\partial X_{od,i}} = \frac{\partial C(F(\mathbf{X}))}{\partial X_{od,j}} \tag{6}$$

Our cost function assumptions imply a more intuitive form of this equation. For each route i, let $\tilde{R}^{iE}$ be the indicator vector for the reverse route of i, containing the same nodes but reversing the direction of each edge. The partial derivative of the cost function with respect to shipping along a single route can be decomposed into

$$\frac{\partial C(F(\mathbf{X}))}{\partial X_{od,it}} = \underbrace{\sum_{uv \in i} \frac{\partial c_{uv}}{\partial F_{uv}}}_{\text{Edge flows}} + \underbrace{\sum_{vu \in i} \frac{\partial c_{uv}}{\partial F_{vu}}}_{\text{Reverse-edge flows}} + \underbrace{\sum_{n \in i} \frac{\partial c_n}{\partial F_n}}_{\text{Node flows}} \tag{7}$$

Equation (7) highlights the three ways in which shipping an extra unit via route i affects costs: First, it increases the flow in all edges belonging to route i, incurring an edge marginal cost. Second, it increases the flows going in the opposite direction for all edges belonging to the reverse route to i, $\tilde{R}^i$. Finally, it increases the amount of traffic through all nodes belonging to route i, incurring a node marginal cost.

Let $\frac{\partial \vec{c}_{uv}}{\partial F_{uv}}$, $\frac{\partial \vec{c}_{uv}}{\partial F_{vu}}$, and $\frac{\partial \vec{c}_n}{\partial F_n}$ be vectors stacking the derivatives of the cost of traversing edges and nodes with respect to edge-flows in the same direction, edge-flows in the opposite direction, and node flows, respectively. Equation (7) can be compactly rewritten as follows:

$$\frac{\partial C(F(\mathbf{X}))}{\partial X_{od,it}} = R'^{iE} \frac{\partial \vec{c}_{uv}}{\partial F_{uv}} + \tilde{R}'^{iE} \frac{\partial \vec{c}_{uv}}{\partial F_{vu}} + R'^{iN} \frac{\partial \vec{c}_n}{\partial F_n} \tag{8}$$

Plugging equation (8) into equation (6), we obtain a matrix version of the first order condition:

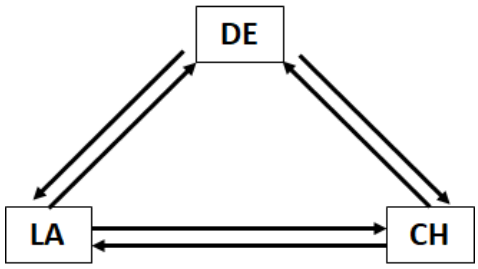
$$\left[R^{i,E} - R^{j,E}\right]' \frac{\partial \vec{c}_{uv}}{\partial F_{uv}} + \left[\tilde{R}^{i,E} - \tilde{R}^{j,E}\right]' \frac{\partial \vec{c}_{uv}}{\partial F_{vu}} + \left[R^{i,N} - R^{j,N}\right]' \frac{\partial \vec{c}_n}{\partial F_n} = 0 \tag{9}$$

In the next section, we develop an empirical analogue to this equation for estimation.

5.5 Example

To build intuition, we present a simple 3-node version of the model, depicted in Figure 8. There are $N_N = 3$ nodes (LA, DE, CH) and $N_E = 6$ edges.

Figure 8: *A sample network*



There are two possible non-looping routes between LA and CH: Route 1 is $\{(LA, DE), (DE, CH)\}$, while route 2 is $\{(LA, CH)\}$. We can represent this mapping of routes to specific edges and nodes using a pair of matrices R^E and R^N, where:

$$R^E = \begin{bmatrix} R^{1E} \\ R^{2E} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$
$$R^N = \begin{bmatrix} R^{1N} \\ R^{2N} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

The rows of R^E and R^N correspond to routes 1 and 2, while the columns correspond to the six edges for R^E and the three nodes (LA, DE, CH) for R^N.[21]

Suppose the firm must deliver a single shipment from LA to CH, that is, $Q_{LA,CH} = 1$. The firm chooses how much to ship through each of the two routes between LA and CH, $X_{LA,CH} = (x_1, x_2)$, where x_i is the amount shipped on route i. Given some choice of shipments $\mathbf{X}$, the flows through each edge are given by:

$$F^E(\mathbf{X})' = \mathbf{X}R^E = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} x_1 & x_1 & x_2 & 0 & 0 & 0 \end{bmatrix}$$

Similarly, the flows through each node are:

[21] The edges are, in order, (LA,DE), (DE,CH), (LA,CH), (DE,LA), (CH,DE), (CH,LA)

$$F^N(\mathbf{X})' = \mathbf{X}R^N = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} x_1 + x_2 & x_1 & x_1 + x_2 \end{bmatrix}$$

The cost of traversing edge (LA-DE) equals

$$c_{LA,DE}(F_{LA,DE}, F_{DE,LA}) = c_{LA,DE}(x_1, 0)$$

while the cost of traversing node LA equals

$$c_{LA}(F_{LA}) = c_{LA}(x_1 + x_2)$$

5.6 Comments on the model

In our model, all relevant costs are a function of flows expressed at the edge-month and node-month levels. c_{uv} represents the cost of shipping, for example, 10,000 carloads between u and v in a given month. This abstracts away the details involved in the implementation of this shipment, such as how many individual trains are required, how their schedules will be determined, how locomotives, labor, and fuel will be allocated, etc.

Our model aims to replicate (with limited data) the algorithm used by rail firms to choose their routes. One such algorithm, the Computer-Aided Routing and Scheduling (CARS) tool, is described in Huntley et al. [1995]. CARS also models the rail network as a directed graph, and aims to minimize the cost of shipping a given quantity of cargo. However, it differs from our model in two important ways: First, railroads observe the true costs of shipping, while we can only recover a projection of these costs based on the edge and node characteristics to which we have access. And second, the true algorithm is subject to an additional constraint beyond implementing demand: It must output feasible train schedules, while we abstract away shipment times.

6 Estimation Strategy

Our estimation relies on a revealed preferences argument based on cost minimization. In short, if we observe that a railroad is using two routes to fulfill shipments between a given city pair, cost minimization implies that the marginal costs of each

route must be equal. Otherwise, the firm could decrease its costs by reallocating shipments from the route with higher marginal costs to the route with lower ones.

Below, we detail the assumptions necessary to go from this observation to a set of moment conditions we can take to the data.

6.1 Moment conditions

We assume that the marginal costs of traversing edges/nodes are a parametric function of flows through that node/edge, observable edge/node characteristics (Y), and an unobservable error. That is:

$$\partial c_{uv}/\partial F_{uv} = mc_{1,uv}(F_{uv}, F_{vu}; Y_{uv}, \theta) + \varepsilon_{1,uv,t} \tag{10}$$

$$\partial c_{uv}/\partial F_{vu} = mc_{2,uv}(F_{uv}, F_{vu}; Y_{uv}, \theta) + \varepsilon_{2,uv,t} \tag{11}$$

$$\partial c_{n}/\partial F_{n} = mc_{n}(F_{n}; Y_{n}, \theta) + \varepsilon_{n,t} \tag{12}$$

The error terms ε account for measurement error in the observables (flows and edge characteristics), random variation in costs (due to, for example, weather patterns), and edge and node characteristics unobservable to us. We assume the firm observed ε before making its routing choices, making flows an endogenous variable.

Let an arrow superscript denote the vector stacking the corresponding function for all edges or nodes. Replacing equations (10) through (12) into the first order condition (9) and dropping the Y argument, we have:

$$\left[R^{i,E} - R^{j,E}\right]' \vec{mc}_{1,uv}(F_{uv}, F_{vu}; \theta) + \left[\tilde{R}^{i,E} - \tilde{R}^{j,E}\right]' \vec{mc}_{2,uv}(F_{uv}, F_{vu}; \theta) + \left[R^{i,N} - R^{j,N}\right]' \vec{mc}_{n}(F_{n}; \theta)$$
$$+ \left[R^{i,E} - R^{j,E}\right]' \vec{\varepsilon}_{1,uv,t} + \left[\tilde{R}^{i,E} - \tilde{R}^{j,E}\right]' \vec{\varepsilon}_{2,uv,t} + \left[R^{i,N} - R^{j,N}\right]' \vec{\varepsilon}_{n,t} = 0$$

Given a guess for θ and a functional form for the mc functions, we can compute the first three terms of this sum as a function of observable edge and node flows. For each guess of θ, we can recover a linear combination of the errors. This linear combination, together with an instrument, will give us a moment condition that will allow us to estimate the model parameters.

6.2 Parametrization

6.2.1 Edge marginal costs: mc_1 and mc_2

We start by parameterizing mc_1, the same-direction marginal cost. As mentioned in Section 2, firms seem to be considering capacity utilization f_{uv} when making route choices. Hence, we allow costs to depend on the capacity utilization of an edge non-monotonically.

In addition, since our definition of what constitutes an edge depends on an ad hoc resolution parameter, we would like a functional form that is robust to our definition of an edge.[22] In particular, if we were to split an edge in half, we would like the cost of traversing the original edge to be equal to the sum of the two new edges. As such, we choose costs to be linear in the edge length:

$$mc_{1,uv} = d_{uv}(\alpha_d + \alpha_Y Y_{uv} + g(f_{uv,t}, f_{vu,t}))$$

where d_{uv} is the length (in miles) of the edge, Y_{uv} is a vector of track characteristics, including the number of parallel tracks, the frequency of passing sidings, the slope of the terrain, whether the firm owns or rents the infrastructure, and a state fixed effect.

We need g to be any convex function that has an analytical integral. For simplicity, we choose a quadratic polynomial on f_{uv} and f_{vu}.[23]

$$g(f_{uv,t}, f_{vu,t}) = \beta_1 f_{uv} + \beta_2 f_{vu} + \gamma_1 f_{uv}^2 + \gamma_2 f_{vu}^2 + \gamma_3 f_{uv} f_{vu}$$

We normalize the coefficient of distance α_d to be one. This implies our costs are expressed in "effective miles," and so we can interpret g as reflecting the effect of traffic flows in mile-equivalent units.

Our choice of mc_1 has implications on the shape of c_{uv}, which in turn restricts the shape of mc_2. In particular, by the definition of $mc_{1,uv}$, the cost of traversing an edge, c_{uv}, must satisfy:

$$c_{uv} = \int mc_{1,uv} dF_{uv}$$

[22] As mentioned in Section 3, we merge all edges within a 0.2 decimal degrees square.

[23] Adding higher order polynomial terms does not significantly change the results.

Hence, the implied restriction on mc_2 is given by the following expression:

$$mc_{2,uv} = \frac{\partial c_{uv}}{\partial F_{vu}} = \frac{\partial}{\partial F_{uv}} \left(\int mc_{1,uv} dF_{uv} \right)$$

After integrating, this becomes:

$$mc_{2,uv} = d_{uv} \left(\beta_2 \frac{F_{uv}}{K_{vu}} + \gamma_2 \frac{F_{uv} F_{vu}}{K_{vu}^2} + \gamma_3 \frac{F_{uv}^2}{K_{uv} K_{vu}} \right)$$

6.2.2 Node marginal cost: mc_n

We only consider the cost of traversing nodes that correspond to rail yards owned by the firm. Traversing every other node will have a cost of zero.[24] Conditional on a node corresponding to a yard, we let node marginal costs, mc_n, be a quadratic function of node capacity utilization.

$$mc_n = 1\{\text{node } n \text{ is a rail yard}\} \left(\alpha_n + \beta_n f_{n,t} + \gamma_n f_{n,t}^2 \right)$$

6.3 Instrument

The firm makes its routing choices after observing the errors ε. As a result, edge- and node-flows are likely correlated with the error term. For example, if ε_{uv} is high due to a snowstorm increasing the cost of traversing edge uv, the firm is likely to reduce flows going through that edge, creating a negative correlation between the error and the observed flows.[25]

We can control for endogeneity using a source of exogenous variation to construct an instrument. Since we are estimating a cost function, the natural source of variation would be a demand shock that could help us trace the cost function. Ideally, the magnitude of the demand shock would differ from edge to edge in a way that is uncorrelated with the errors, allowing us to instrument for flows F using some set

[24] Besides rail yards, there are two additional types of nodes in the rail network: origins and destinations of shipments, and points where two tracks cross each other. All shipments must go through their origin and destination nodes regardless of their route; our current estimation approach does not allow us to identify these costs. We are working on incorporating rail crossing quality data into our analysis, but the current model does not include them.

[25] Since flows enter both positively and negatively in our estimating equation, we cannot conclude from this correlation that the OLS estimates will be downward biased.

of variables Z that satisfy:

$$\left[R^{i,E} - R^{j,E}\right]' \vec{Z}_{1,uv,t}\vec{\varepsilon}_{1,uv,t} + \left[\tilde{R}^{i,E} - \tilde{R}^{j,E}\right]' \vec{Z}_{2,uv,t}\vec{\varepsilon}_{2,uv,t} + \left[R^{i,N} - R^{j,N}\right]' \vec{Z}_{n,t}\vec{\varepsilon}_{n,t} = 0$$

We construct such an instrument using the phase-out of coal that followed the emergence of shale gas starting in 2009. We provide details on the validity and construction of this instrument in the next subsection.

6.3.1 Coal phase-out shift-share instrument

We instrument for flows at the edge- and node-level with a shift-share instrument interacting monthly coal production east of the Mississippi River with the exposure of each edge and node to coal production.

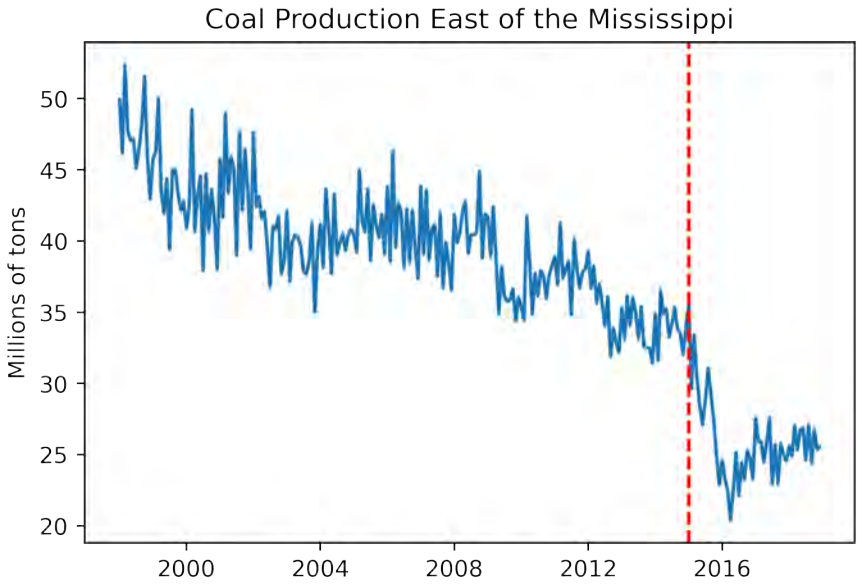
Until the mid-2010s, coal was the largest source of electricity in the United States, accounting for more than a third of the total electricity production (US Energy Information Administration). But in the 2010s, coal production in the US started to decline, as the emergence of shale gas reduced the price of natural gas and out-priced many coal energy plants. This decline further accelerated in 2015, when the Obama administration introduced the Clean Power Plan, increasing environmental regulation of coal plants. Between these two factors, monthly coal production declined by 35% from 2009 to 2018.

Rail plays a crucial role in transporting coal from mines to power plants across the US. In 2021, 69% of coal shipments were transported by rail, and as US coal production has decreased, the amount of coal transported by rail has decreased as well, falling by 61% from 2008 to 2021. Since coal is the top commodity transported by rail in terms of volume (accounting for 27% of all US freight in 2021), the decline in coal production has represented a substantial demand shock for the rail industry.[26]

We use this demand shock to construct a shift-share instrument for rail flows at the edge and node level. Since both railroads studied in this paper (BNSF and UP) service coal mines in the Mountain West, a potential threat to our instrument is the possibility of higher freight prices contributing to the decline in coal production. We address this by using as a shifter the total monthly US coal production for states East of the Mississippi River (depicted in Figure 9), outside of these railroads' area

[26] All the figures in this paragraph come from AAR Fact Sheets.

Figure 9: *US Monthly Coal Production East of the Mississippi River.*



Monthly US Coal Production East of the Mississipi River. The red line denotes January 2015, the first month used in our analysis. Source: EIA

of operations.[27]

We next construct a "share" reflecting the exposure of each edge and node to the decline in coal production. We do this by computing the total carloads of coal shipped between each pair of nodes o and d in the network for the period 1998-2003 (the earliest data to which we have access). We construct a shift-share variable at the city-pair level by multiplying the demand shifter and the share of total carloads shipped between a given city pair as a fraction of total coal shipments:

$$\text{Shift-Share}_{od,t} = \frac{\text{Carloads of Coal}_{od,98-03}}{\sum_{od} \text{Carloads of Coal}_{od,98-03}} \cdot \text{Coal Production}_t$$

In order to use this measure as an instrument, we need to convert it into a measure of exposure to coal for each edge uv and node n in the network. To do this,

[27] An additional concern is that the availability of cheaper shale oil led to both lower diesel costs for rail and lower coal production. To the extent that the decline in coal correlated with a decline in oil prices, this should affect all edges equally and cancel out (see Section 6.4), since diesel prices are highly correlated across space. Indeed, diesel prices across US regions had a correlation of at least 97.8% during 2015-2018 (EIA).

we compute the shortest route between o and d, and we allocate each city pair's exposure to all the nodes and edges belonging to that shortest route. Note that we use the shortest rather than the observed route since the observed route depends on the structural errors ε, and to the extent that these are persistent over time, could introduce bias in the results.

$$\text{Shift-Share}_{uv,t} = \sum_{od} 1\{uv \text{ belongs to shortest route between } o \text{ and } d\} \cdot \text{Shift-Share}_{od,t}$$

$$\text{Shift-Share}_{n,t} = \sum_{od} 1\{n \text{ belongs to shortest route between } o \text{ and } d\} \cdot \text{Shift-Share}_{od,t}$$

6.3.2 Nonparametric First Stage

We use these edge- and node-level shift-shares to construct an instrument for the endogenously determined flows at the edge- and node-level. Rather than choosing a specific functional form to go from these shift-shares to an instrument, we follow the approach proposed by Chen et al. [2020]to perform a nonparametric first stage with linear controls.

To summarize their approach, suppose we want to run the following regression:

$$Y_i = \alpha D_i + \beta X_i + \varepsilon_i$$

where D_i is endogenous, ε is unknown, and X_i is exogenous.[28] If we have a valid instrument Z_i, Chen et al. [2020] show how we can consistently estimate the coefficient of interest, α, together with the coefficient for the linear controls, β, with the following instrument:

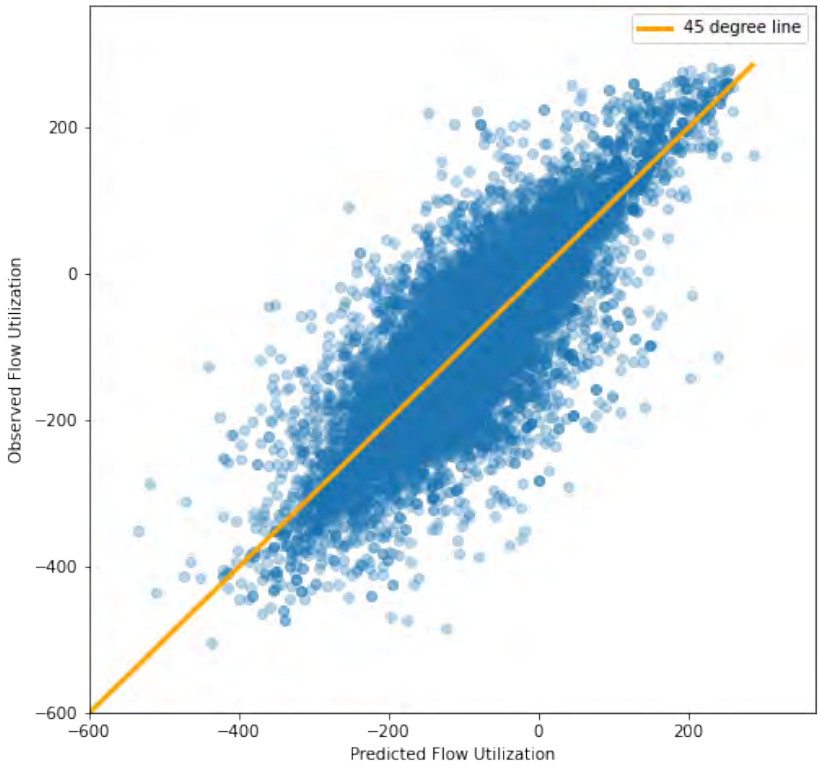
$$E[D_i|Z_i] + E^*\left[D_i|X_i - E[X_i|Z_i]\right]$$

where E^* is the Best Linear Predictor. In our application, we estimate the conditional expectation via a random forest. We refer the reader to Chen et al. [2020] for further details and a discussion on the econometric properties of this estimator.

In Figure 10, we plot a scatter plot of our instrument ("Predicted Flow Utilization") and the observed flows. The instrument tracks the endogenous variable closely, and this "first-stage" has an R^2 of 78%.

[28] In our application, the flows f_i play the role of D_i, and the edge characteristics are equivalent to X_i.

Figure 10: *Observed monthly flows vs. instrument.*



6.4 Comments on the estimation

Our estimation relies on deriving a condition under which the firm equalizes marginal costs across two routes. This approach has two drawbacks: First, any costs that do not depend on the route chosen will cancel out. Second, this condition can be trivially satisfied by setting all costs to zero (we implement a normalization to avoid this issue).

Levels vs changes. Our estimation approach cannot estimate any constant coefficient in the marginal cost function. This constant coefficient includes all costs that do not depend on the route taken, such as the wages of the employees operating the trains. Our cost function, then, relates exclusively to inputs whose utilization

is directly tied to the route used. These include the services provided by the rail infrastructure, as well as any route-dependent component of wages and capital costs.[29]

An additional implication is that we cannot recover the level of costs. Our approach, however, does allow us to detect changes in costs over time and across space. For example, it enables us to make statements such as "route i is 100 miles more expensive than route j."

Unit of measurement. Given our parametrization, all our moment conditions would be trivially satisfied if we set all coefficients equal to zero. To avoid this problem, we set the coefficient of distance on costs to one. We then interpret all costs in terms of equivalent miles: Saying that the cost of route i is 100 miles higher than the cost of route j implies that choosing i is equivalent to choosing j and shipping a carload an additional 100 miles.

In some cases, it is useful to have a dollar measure of costs. To obtain this measure, we start with the fuel efficiency of a locomotive, which stands is 400 ton-miles per gallon. A typical carload weights 115 tons, and the average price of diesel fuel during this period was \$2.71 (US Energy Information Administration).[30] Altogether, this suggests that the fuel costs of shipping one car-mile are $\frac{115}{400}\$2.71 = \$0.78$. Based on the Association of American Railroads (AAR), fuel represents around 20% of operating costs, so the aggregate cost of shipping one car-mile is \$3.89. For reference, the average price of shipping one car-mile during this period was \$5.00.

Selection and other caveats. In our analysis, we consider cases where the railroad uses more than one route to fulfill an origin-destination pair. However, in so doing, we are not using the potentially useful information that if a railroad chooses to use only one route to fulfill its demand, that choice must lead to a lower cost than if it used two routes. We could incorporate this information using moment inequalities for our estimation, allowing us to recover any fixed component of marginal costs. We have not implemented this approach yet, but we plan to explore this in future iterations of this paper.

An additional caveat related to the interpretation of the costs recovered from this approach—especially when they interact with the time to ship—is that we cannot

[29] For example, a longer route may take more time, and thus require more worker hours.

[30] Locomotives use a diesel fuel known as Diesel #2 or ULSD (Ultra-Low Sulfur Diesel).

directly account for the duration of any trip, since we do not observe this in our data. However, it is possible that trips using longer routes also take longer to reach their destinations. When this is the case, the firm needs to increase the labor and capital used for a single trip (i.e., it needs to assign a locomotive to the trip for a longer period of time). In such a case, our cost coefficients would also include capital and labor costs.

A trickier problem arises when consumers care about the duration of the trip. Under certain assumptions regarding demand, described in Appendix C, the model is isomorphic to one in which the firm offers a rebate to consumers proportional to the duration of the shipment. When that occurs, the marginal cost coefficients will capture a combination of the technological costs to the firm and the foregone profits due to the rebate.

7 Results

Estimation results. We use Generalized Method of Moments to estimate the coefficients of the marginal cost functions described in Section 6.2. We estimate separate coefficients for both BNSF and UP, allowing them to have different technologies. We present the results in Table 1. Columns (1) and (3) present OLS parameter estimates, columns (2) and (4) show our IV estimates, while columns (3) and (5) add passing sidings as an edge characteristic.[31] For the remaining sections, we use specifications (2) and (4).[32] These estimates inform us regarding the shape of the cost function and how the firms trade off distance, infrastructure characteristics, and capacity utilization. Note that the OLS and IV coefficients are similar in magnitude, with the main difference being that our IV estimation has tighter confidence intervals.[33]

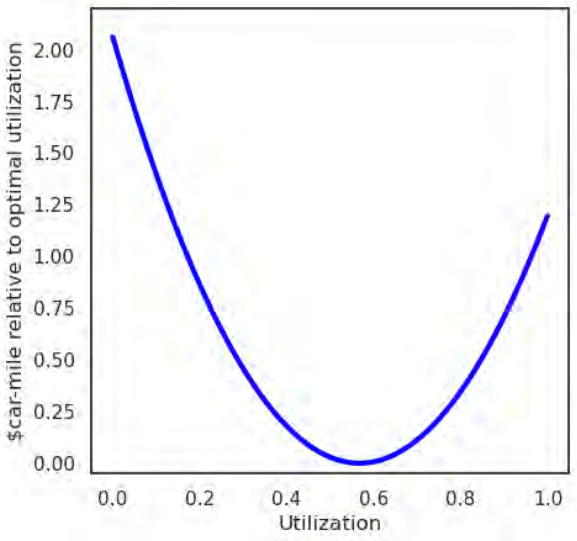
For both firms, better infrastructure reduces the cost of traversing an edge. An additional parallel track shortens the effective length of an edge by 18% for BNSF and 61% for UP, reducing the slope by 1% reduces edge-costs by 30%, and increasing the frequency of passing sidings reduces costs by 48-65%. The one notable difference between the two firms is the coefficient of the trackage dummy, which indicates

[31] At the time of writing this draft we have not yet finished running the route inversion for Union Pacific, thus leading to fewer observations. For more details, see Appendix B.

[32] We computed these estimates first, and have not yet rerun our analysis incorporating the new coefficients.

[33] This is consistent with our previous discussion of the lack of a clear bias in OLS.

Figure 11: *Estimated marginal costs at the edge level - BNSF.*



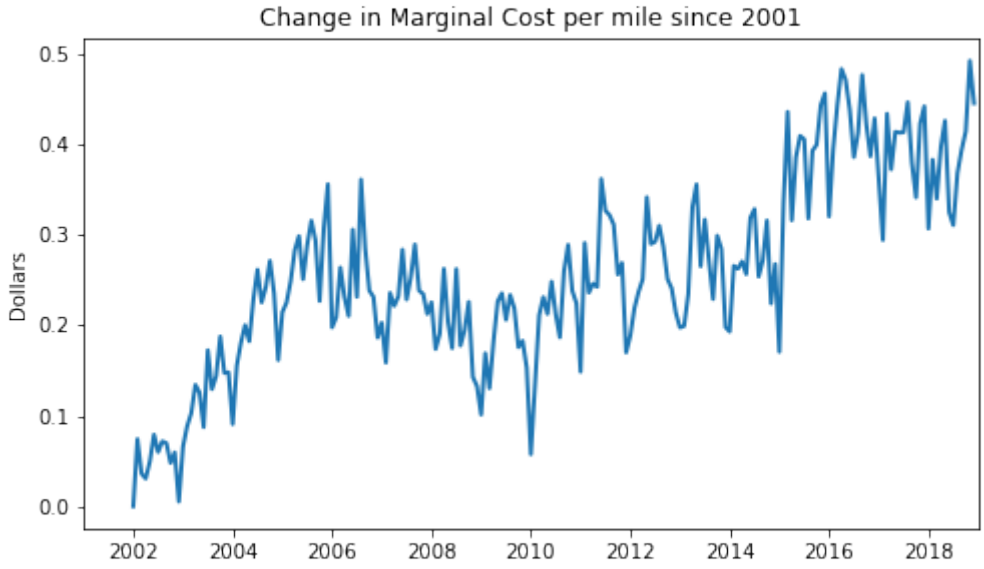
Edge-level marginal costs for BNSF. In this figure, we assume that the capacity utilization of an edge is the same in both directions.

whether the firm is using another firm's infrastructure. While UP is indifferent between using its own tracks and those of BNSF, BNSF systematically prefers using UP's tracks. This is consistent with an inefficiency arising from trackage rights, such that the price UP charges BNSF is too low. The node-costs coefficients imply that node-level marginal costs are hump shaped on capacity utilization, achieving a maximum at around 60% capacity utilization for BNSF and 40% for UP.

To interpret the estimated edge costs, we compute BNSF's edge-level marginal costs as a function of capacity utilization, depicted in Figure 11.[34] We convert the units from effective miles to dollars per carload-mile using the calculation described in Section 6.4. Since we do not identify levels, we normalize the cost of the most efficient unit to be zero. Marginal costs are U-shaped, with a minimum around 55% utilization, and imply significant economies of scale: Traversing an edge with 0% utilization is $2.00 more expensive than a shipment at 55% utilization; this difference represents 40% of the average price for a carload-mile during this period. These results are consistent with the behavior of prices in the spot market shown in Figure 7, even though pricing data was not included in the estimation.

[34]We show the same plot for Union Pacific in Appendix F.

Figure 12: *Estimated marginal cost per mile 1998-2018*



Reverse edge utilization has different effects depending on the specification. In specifications (2) and (4), with fewer track characteristics, increasing reverse utilization lowers both the optimal capacity and total costs. In specifications (3) and (6), with more controls, we find economically insignificant effects.[35]

Marginal costs over time. We estimate our model using data from 2015 to 2018. Under the assumption that the cost function is constant over time, we can estimate marginal costs by route for the rest of our data. Figure 12 plots the estimated change in marginal costs per mile from 2001 to 2018. We find that changes in capacity utilization increased the average cost per mile by around 50 cents during the period of analysis, with most of the increase occurring during 2004-2006 and 2015-2016. This pattern is consistent with the behavior of prices documented in Section 2, and explains around 30% of the total increase. Appendix F shows how costs varied for each of the two firms individually.

Figure 13 decomposes the time series presented in Figure 12 into edge-level and node-level marginal costs. Almost all the variation comes from changes in edge-level marginal costs, with node costs oscillating around zero for the entire period.[36]

[35]We are currently working on allowing reverse edge to affect costs differently depending on the rail infrastructure.

[36]Starting in the early 2000s, railroads have tried to increase the efficiency of their shipments by avoiding using yards, focusing increasingly on putting together "unit trains", in which all carloads

Figure 13: *Decomposition of estimated marginal costs per mile*

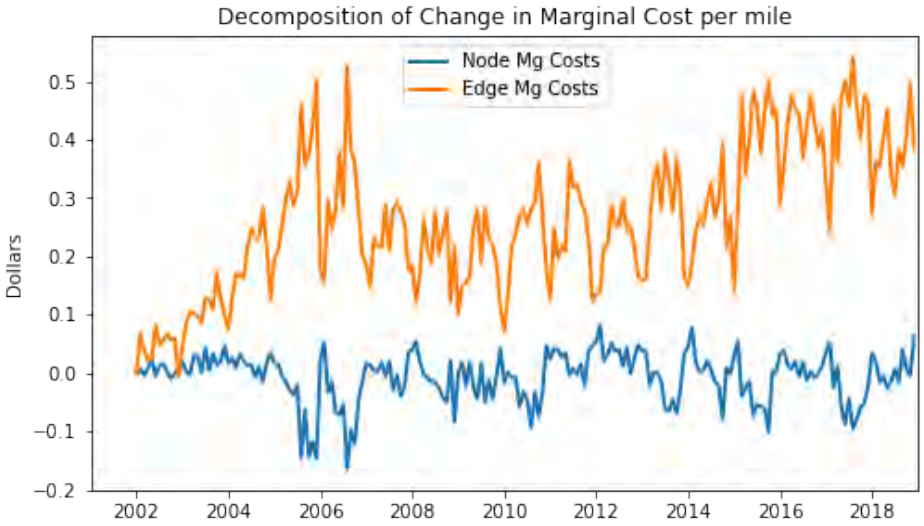


Table 1: *Estimation Results*

	BNSF			UP		
	(1)	(2)	(3)	(4)	(5)	(6)
Infrastructure Characteristics						
Additional Parallel Track	-0.233***	-0.183***	-0.151***	-0.710***	-0.612***	-0.588***
	[-16.64]	[-16.45]	[-18.27]	[-52.43]	[-21.25]	[-38.64]
Trackage Dummy	-0.329***	-0.309***	-0.400***	0.233***	0.182***	0.014
	[-57.61]	[-66.35]	[-41.43]	[29.83]	[11.78]	[1.40]
Slope of 1%	0.482***	0.418***	0.348***	0.434***	0.402***	0.293***
	[24.85]	[21.17]	[21.17]	[20.11]	[15.62]	[20.73]
Passing Lane Frequency			-0.651***			-0.480***
			[-24.85]			[-8.15]
Track Marginal Costs						
Utilization	-1.348***	-1.533***	-1.120***	-0.768***	-1.129***	-0.287***
	[-39.77]	[-37.61]	[-24.32]	[-19.34]	[-7.63]	[-3.96]
Util. Squared	1.139***	1.256***	0.918***	0.726***	1.117***	0.188**
	[29.12]	[27.52]	[22.51]	[17.17]	[7.34]	[2.45]
Util. opp. dir.	-0.198***	-0.334***	-0.179***	-0.106***	-0.132***	-0.139***
	[-6.81]	[-14.05]	[-7.17]	[-10.72]	[-8.25]	[-7.15]
Util. opp. dir. squared	-0.051***	-0.089***	0.012	-0.010	-0.034	-0.076***
	[-2.83]	[-5.20]	[0.85]	[-0.55]	[-1.41]	[-2.75]
Util. times utilization opp. dir.	0.260***	0.479***	0.197***	0.161***	0.236***	0.0262***
	[5.75]	[12.55]	[4.96]	[6.13]	[6.30]	[5.45]
Yard Marginal Costs						
Node-costs constant	-31.256***	-43.929***	-17.974***	-9.908***	-14.664*	8.322***
	[-16.62]	[-13.20]	[-7.08]	[-7.03]	[-1.89]	[5.24]
Node-costs slope	96.712***	157.709***	61.634***	52.762***	83.964***	-14.322**
	[12.25]	[12.05]	[7.28]	[10.59]	[2.79]	[-2.24]
Node-costs quadratic	-79.365***	-136.796***	-51.974***	-54.149***	-89.550***	16.676***
	[-10.96]	[-12.04]	[-7.66]	[-11.64]	[-3.30]	[2.67]
State FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	80561	80561	80561	20095	20095	20162
R^2	0.981			0.987		

Parameters represent costs in effective-miles. t statistics in brackets.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

8 Counterfactual Analyses

Thus far, we have shown how to model railroad shipping choices while accounting for the network structure of production, as well as how to obtain moment conditions to estimate the model. Next, we show that accounting for this network structure is useful to answer many policy questions. We highlight three features of the railroad industry and explore their implications for two normative policy questions and one positive policy question.

First, the network structure of production implies that infrastructure investments at an edge or node level have network-wide cost implications. Since a significant portion of rail investment is done by firms in cooperation with local and state governments, the presence of network externalities is likely to lead to underinvestment. We evaluate how returns to investment vary across states, and compute how much of these returns are captured by the states making the investment.

Next, since railroads own their rail networks, they internalize any economies of scope they impose on their own flows, potentially improving efficiency over a decentralized system. To highlight this feature, we study the effect of merging the two largest railroads, Union Pacific and Burlington Northern Santa Fe. Such a merger could potentially reduce costs in three ways: First, the merged firm could better coordinate the flows between each firm's network, redirecting traffic to the least congested areas. Second, the joint firm would have access to new, previously inaccessible routes. And finally, the merger would reduce the contractual frictions that are currently present when using trackage rights in each anoother firm's network. We find that such a merger would reduce costs by 17.1% compared to the decentralized equilibrium.

Third, railroads produce a vector of products, with each product representing a shipment between two different city pairs. To the extent that shocks are not uniform across space, they will have different effects depending on the set of products they affect. We study this phenomenon in the case of a particularly important shock: the accession of China to the World Trade Organization (WTO). We show how the reallocation of imports from ports in the Gulf Coast to the port of Los Angeles interacted with the capacity utilization of BNSF and UP to produce heterogeneous effects across markets.

share the same origin and destination. Thus, we interpret our results as reflecting the reduced importance of yards.

8.1 Infrastructure investment

Freight railroads own, build, and maintain most of the rail infrastructure in the US. From 2017 to 2022, railroad firms invested around $23 billion per year, equivalent to 39 percent of their revenue. Nevertheless, the federal government[37] as well as state and local government, often enter into partnerships with one or multiple railroads to jointly finance specific investments. Two salient examples are the Alameda Corridor and the Chicago Region Environmental and Transportation Efficiency Program (CREATE).

The Alameda Corridor, built in the late 1990s and early 2000s, was a joint undertaking by BNSF and UP and the cities of Los Angeles and Long Beach. The project aimed to improve rail access to the Port of Los Angeles by building a below-ground "trench," allowing rail traffic from the port to avoid grade crossings and cross urban areas at high speeds. Similarly, CREATE is a series of 70 projects financed by the Department of Transportation, the Illinois state government, the City of Chicago, and multiple private railroads.

Both of these initiatives aimed to reduce congestion in their respective cities by separating rail flows from road networks. In addition, these flows were expected to reduce rail congestion and increase rail access to cities. As long as part of the reduction in costs is passed on to consumers in the form of lower prices, cities will see a reduction in prices for goods shipped by rail, as well as increased market access for firms located in these cities.[38]

One natural question that emerges is whether state and local governments face the right incentives when deciding on whether to finance such infrastructure investments. This is particularly salient because these investments are likely to generate spillovers throughout the entire network. In particular, we study two questions: First, where do infrastructure investments have the highest returns? And second, how large are the network spillovers of these investment?s

[37]The Infrastructure Investment and Jobs Act of 2021 allocated $22 billion dollars to finance investment in rail and intermodal facilities.

[38]These projects highlight a major reason why railroad firms are unlikely to undertake these projects on their own: Improving rail infrastructure often involves changing road networks, making these projects impossible without governmental cooperation.

8.1.1 Returns to investment differ across states.

We start by evaluating how returns on infrastructure investments vary across states.[39] To do so, we separately improve the rail infrastructure of each state in our sample and assess how costs change in response. Specifically, we add one additional set of parallel tracks to every edge in the state. Since states differ in the amount of rail they have, we will report how cost changes for each mile of parallel tracks.

The additional set of tracks will reduce the cost for all shipments with a route along that edge.[40] In addition, the firm will be aware of its improved infrastructure, and will reallocate shipments to take advantage of the reduced costs. In order to recover this counterfactual change in flows, we solve the routing problem for each firm using the new edge-level costs. For more details on this computation, please see Appendix D.

We use June of 2018 as our reference month, meaning that firms will have to fulfill the Q_t observed in that month. We use our cost function to solve for firms' routing choices before and after the infrastructure investment in each state s, $\mathbf{X}_s^B$ and $\mathbf{X}_s^{CF}$.[41] Next, we compute how total costs change in response to the changing infrastructure:

$$\Delta C_s = C_s^{UP,CF}(\mathbf{X}_s^{UP,CF}) + C_s^{BNSF,CF}(\mathbf{X}_s^{BNSF,CF}) - C_s^{UP,B}(\mathbf{X}_s^{UP,B}) - C_s^{BNSF,B}(\mathbf{X}_s^{BNSF,B})$$

We do this for all 28 states where either UP or BNSF have rail infrastructure. For each state, we compute its Return on Investment:

$$\text{Return on Investment}_s = \frac{\Delta C_s}{\text{Miles of track}_s}$$

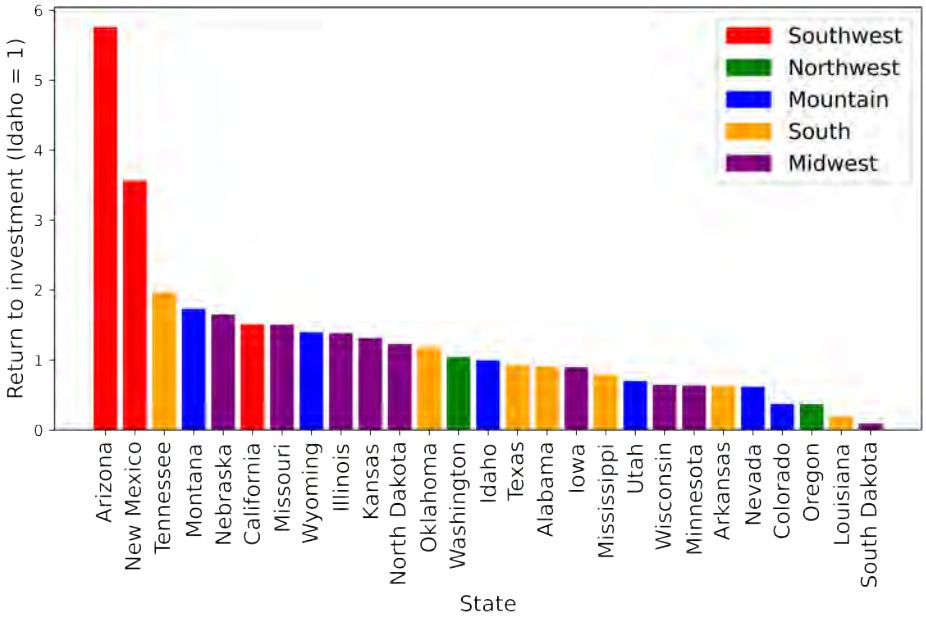
We normalize the returns to one for the median state in our sample (Idaho), and plot this measure for all states in Figure 14. As is evidenced by the Figure, returns on investment vary significantly across states. The state with the most productive investment (Arizona) has a return six times higher than that of the median state (Idaho), which in turn is five times as productive as the least productive state (South

[39] In principle, our model would allow us to detect the edges or nodes where investment is most productive. However, minimizing the cost function after every change takes about 3 hours, and with around 6,000 nodes and edges, this is unfeasible.

[40] Based on our parameter estimates, this is equivalent to reducing the cost of going through each edge by 40%.

[41] Note that these are $\mathbf{X}_s = \{\mathbf{X}_s^{UP}, \mathbf{X}_s^{BNSF}\}$

Figure 14: *Return to Investment in rail infrastructure.*



Change in costs in response to adding one additional set of parallel tracks to all edges in a state. Idaho = 1

Dakota). Returns on investment are higher in more central regions, and the states with the highest returns (AZ, NM, NE, MO, and KS) are all part of the largest rail corridor, which connects Los Angeles and Chicago.[42]

Going back to the two infrastructure projects that opened this subsection, we find that investment in California and Illinois is only 25% as productive as investment in Arizona. This large difference is partly explained by two shortcomings of this approach: First, our approach cannot recover the costs incurred at the start or end of a shipment. All shipments originating in Los Angeles, for example, must incur the cost of traversing the Los Angeles node regardless of the route taken. As such, any fixed costs originating at that point would cancel out. Similarly, to the extent that these investments focus on costs incurred during the last few miles of the trip, we would be underestimating their benefits.[43] Second, our counterfactual studies the

[42]Our current formulation does not allow for decreasing returns on investment: Investing one dollar in Arizona is as effective as investing one billion. We are working in incorporating decreasing returns in a future draft.

[43]We can use moment inequalities in estimation to recover these costs. We will incorporate that approach in future work.

returns of improving the infrastructure for all rail in California and Illinois, while more-targeted investments would likely yield higher returns.

8.1.2 Infrastructure investments produce significant network spillovers.

Next, we focus on the case of investment in Arizona, evaluating its effects on the costs of shipping to and from different regions of the United States. We will assume that changes in marginal costs translate one-to-one to changes in prices, although the qualitative results would hold as long as markups are constant across space.

We start by computing how the marginal cost of shipping a ton-mile between each city pair in our sample changes in response to the infrastructure investment. We denote this change as Δc_{ij}. Next, we evaluate how these changes translate into lower cost of shipping by constructing a Laspeyres price index for each location:

$$\text{Change in relative costs}_i = \frac{\sum_j Q_{ij}\Delta c_{ij} + \sum_j Q_{ji}\Delta c_{ji}}{\sum_j Q_{ij} + \sum_j Q_{ji}}$$

This index measures how the cost of shipping products has changed in response to the infrastructure investment. We plot these changes in Figure 15, which demonstrates that costs decreased the most in Arizona—up to 17% in the area around Flagstaff. This is to be expected, since all shipments originating or terminating in Arizona use the newly improved rail. In addition, there are substantial spillovers, as most of California and areas of Texas saw cost reductions of over 5%.

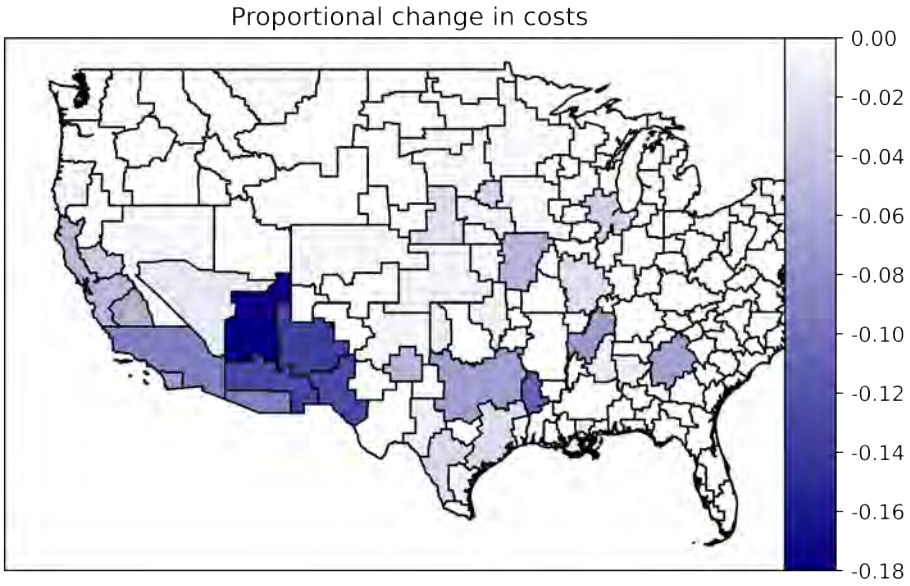
The outcome changes, however, when we look at the total change in costs. Although Arizona benefits the most from investment in Arizona rail, the state is not a major origin or destination for rail flows. We compute the total change in costs to be:

$$\text{Change in total costs}_i = \sum_j Q_{ij}\Delta c_{ij} + \sum_j Q_{ji}\Delta c_{ji}$$

Figure 16 shows which states benefit the most once we consider total costs. Arizona accrues only 3% of the total benefits, with 47% of the gains going to California, and 15% each to Texas and Illinois.[44] These numbers highlight how returns to investment need not align with incentives to invest.

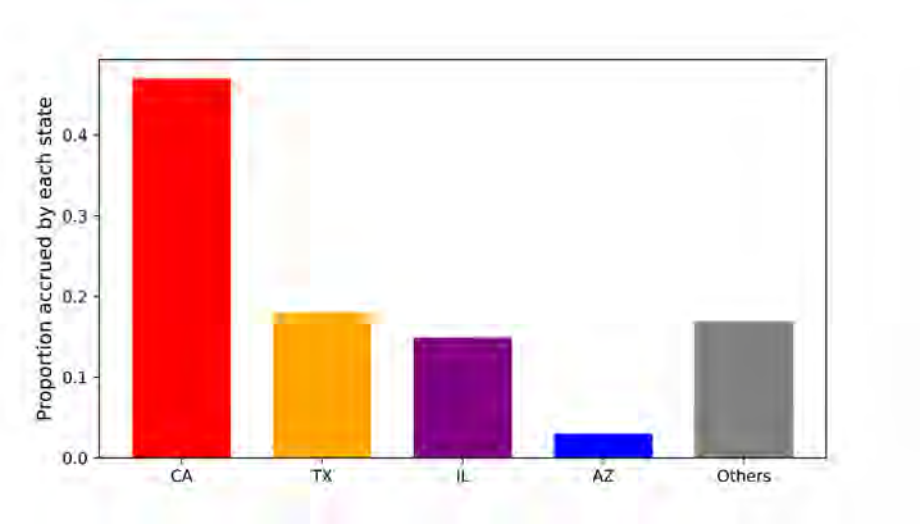
[44]Note that these numbers imply California would get a higher return from improving Arizona's infrastructure than from investing on its own.

Figure 15: *Changes in shipping costs per unit*



Weighted proportional change in shipping costs in response to investment in Arizona.

Figure 16: *Proportion of total benefits accruing to each state*



Proportion of total benefits accruing to each state following investment in Arizona.

8.2 Merger Analysis

The freight railroad industry underwent significant consolidation after its deregulation in 1980, contracting from over 40 Class I railroads in the late seventies to just seven by 1998. A 25-year pause followed, until early 2023, when Canadian Pacific acquired Kansas City Southern, bringing the total number of firms in the industry down to six. The Surface Transportation Board has been more lenient when evaluating anticompetitive concerns for end-to-end mergers, in which a railroad shipping between A and B acquires a firm shipping from B to C. These mergers are analogous to vertical integration, and have thus traditionally sparked fewer anticompetitive concerns among regulators. In contrast, horizontal mergers, in which a railroad acquires another railroad with similar routes, have been less common, and have required divestments or the concession of trackage rights to competitors.

Horizontal mergers are, however, more likely to allow the merged firm to better exploit the economies of scale and scope documented in this paper. Thus, for our second application, we evaluate the efficiency gains from a horizontal merger. We do this by simulating a merger between BNSF and UP.[45]

We start by discussing why cost synergies may arise from merging these two firms, and identifying three factors that may lead to cost reductions. Next, we explain how we use our model to simulate this merger, and delineate how much of the cost reductions would come from each factor.

8.2.1 Costs efficiencies from a merger

Before discussing the implementation of this counterfactual, it is useful to consider why a merger would lead to reductions in total costs. We identify three factors that would lead to lower costs: cooperation, process innovation, and reduced contractual frictions.

Cooperation. Suppose we allow UP and BNSF to merge, but we force them to keep their networks separate, with each individual shipment reqiored tp travel entirely within BNSF's or UP's original rail network. This arrangement would still lead to lower costs, as the merged firm would be able to optimally allocate flows to the subnetwork experiencing the lowest costs at any given moment. This could

[45]This merger is not under consideration and is unlikely to meet the STB's antitrust standards; still, we consider it an interesting thought experiment.

be because one of the firms is better at serving a given city pair than the other, or because one of them is congested and thus facing higher costs.

We call any reduction in costs that stems from improving the allocation across firms "cooperation." Note that this outcome cannot be replicated with the price system due to the presence of long-term contracts, which prevent the consumer from shifting to the firm with the lowest shipping costs. Absent switching costs between firms (not modelled in this paper), the firms could, in principle, replicate this outcome by hiring each other to deliver shipments.

Process Innovation. Beyond cooperation, suppose we let the merged firm integrate its networks and use any path when serving a given origin-destination pair. Because the firm can now use track belonging to either of the two original firms, it will have access to new routes that were previously unavailable. These new routes represent new ways to produce a shipment between two points and so, we will call the cost reduction they generate "process innovation." Regulation could also allow the firms to reach this outcome without a merger: The STB could require the extension of more trackage rights (which we will discuss more shortly) to make these paths available to the existing firms.

Contractual Frictions. Trackage rights allow a firm to use another firm's tracks for a fee. The STB has often demanded the extension of trackage rights to competitors in order to allow a merger, making these arrangements common in the industry.[46] The fees paid are adjusted annually in response to changes in aggregate costs, but do not generally reflect immediate traffic conditions. As a result, the firm exercising its trackage rights might not internalize the costs it imposes on its competitor.

We see evidence of this in our estimation. In particular, we find that BNSF has a clear preference for using UP's tracks, while UP is indifferent. This is consistent with BNSF paying a fee that does not fully internalize the cost of using UP's tracks.[47] These contractual frictions would be solved by a merger, since the merged firm would fully internalize the costs when choosing its routes.

[46] For example, UP can access around 15% of BNSF's tracks via trackage rights.

[47] If, instead, UP's infrastructure were unobservably better than that of BNSF, we would expect to see UP preferring its own rail to that of BNSF.

8.2.2 Implementation and Results

We simulate the merger using 2018 as the reference year. To isolate the effect of costs efficiencies, we assume that the merged firm does not change its prices, therefore shipping the same amount as the two individual firms.[48] We start by using our cost function parameters to compute a pre-merger baseline, in which each firm separately minimizes costs.[49]

$$\begin{aligned} &\min_X C^{BN}(X|X^{UP}) && \min_X C^{UP}(X|X^{BN}) \\ &\text{s.t. } X \in \mathbf{X}^{BN} && \text{s.t. } X \in \mathbf{X}^{UP} \\ &\sum X = Q^{BN} && \sum X = Q^{UP} \end{aligned}$$

We then compute the merger outcome in three stages, corresponding to the three factors described in the previous subsection. We start by simulating a cooperation equilibrium, in which the merged firm has access to the union of all routes available to BNSF and UP, $\mathbf{X}^M = \mathbf{X}^{UP} \cup \mathbf{X}^{BN}$. The firm minimizes a new cost function, C^M: This function has the same form as that described in equation (3), but can now include edges belonging to different firms.

$$\begin{aligned} &\min_X C^M(X) \\ &\text{s.t. } X \in \left\{\mathbf{X}^{BN} \cup \mathbf{X}^{UP}\right\} \\ &\sum X = Q^{BN} + Q^{UP} \end{aligned}$$

Next, we compute the process innovation equilibrium. To do so, we expand the choice set for the firm: It can now use routes that are only available by combining the two subnetworks, $\mathbf{X}^{NEW}$. The problem becomes:

[48] We could then ask, how much would markups have to increase to undo the cost efficiencies of the merger?

[49] Since our model does not fit the data perfectly, we compare model predictions to model predictions to isolate the changes coming from the merger. For details, see Appendix D.

$$\begin{aligned}
&\min_{X} C^{M}(X) \\
&\text{s.t. } X \in \left\{ \mathbf{X}^{BN} \cup \mathbf{X}^{UP} \cup \mathbf{X}^{NEW} \right\} \\
&\sum X = Q^{BN} + Q^{UP}
\end{aligned}$$

Finally, we compute the role of contractual frictions. We do this by assuming that, for the merged firm, whether an edge involves using trackage rights has no effect on costs. That is: $\alpha_{trackage} = 0$. We compute the firm optimal choices with the new set of parameters, and evaluate the previous choices using this new frictionless cost function.

Results. Our simulations deliver significant synergies. For the year 2018, we find that cooperation reduces total costs by 17% compared to the decentralized equilibrium.[50] We are currently working on implementing the other two scenarios.

8.3 China Shock

The accession of China to the WTO in 2001—one of the most studied shocks of the last 20 years—led to a boom in international trade in the United States. Total imports into the US, plotted in Figure A.7, increased by 115% during this period, or 15% faster than GDP. This shock was accompanied by a corresponding shift in the sources of imports: China's share of total US imports increased almost threefold, from 8.2% in 2000 to 21.6% in 2018 (Figure A.8). China's rising share was accommodated by a corresponding decline in imports from Canada and Japan. A vast literature has reported how this shock led to geographically diverse labor market impacts, as domestic firms were increasingly exposed to Chinese import competition.[51]

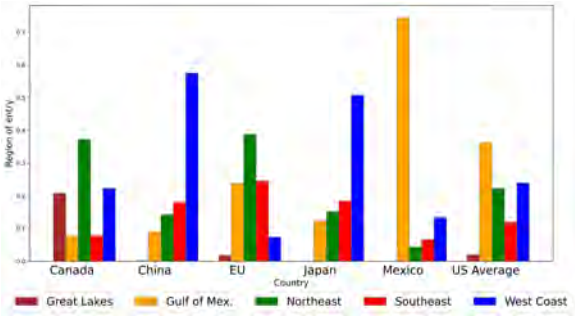
This change interacted with the geography of trade, as exporters typically opt for the nearest port to their home country as their port of entry. Figure 17a shows the five largest exporter's chosen ports of entry into the US in 2018. Unsurprisingly,

[50] Calculating this number is not straightforward, since we do not observe the levels of costs. Instead, we evaluate the costs before and after the merger, and compute the total reduction in costs in effective carload-miles. We then divide this value by the total carload-miles shipped, to obtain the 17% figure.

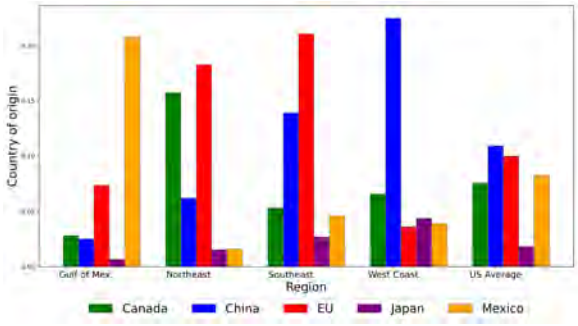
[51] See, among others, Autor et al. [2013], Pierce and Schott [2014], Dorn et al. [2021], Gerritse and Caragliu [2022].

Figure 17: *Import region of entry and country of origin, 2018*

(a) *Region of entry by country of origin*



(b) *Origin by port region*



Values weighted by tonnage. Only imports by ship are considered. Source: US Census Bureau. Ports included in each region: Great Lakes - Ogdensburg, Buffalo, Milwaukee, Chicago, Cleveland, Detroit, Duluth; Gulf of Mexico - Mobile, Port Arthur, New Orleans, Houston-Galveston; Northeast - Washington, Philadelphia, Portland, New York, Boston, Baltimore, Providence; Southeast - Charleston, Tampa, Norfolk, Miami, Savannah, Wilmington; West Coast - Seattle, San Francisco, San Diego, Los Angeles, Columbia-Snake.

imports from China and Japan enter the US predominantly through West Coast ports, Mexican imports through the Gulf of Mexico, European imports through ports in the East Coast, while Canadian imports are spread out evenly between the East Coast, West Coast, and the Great Lakes.

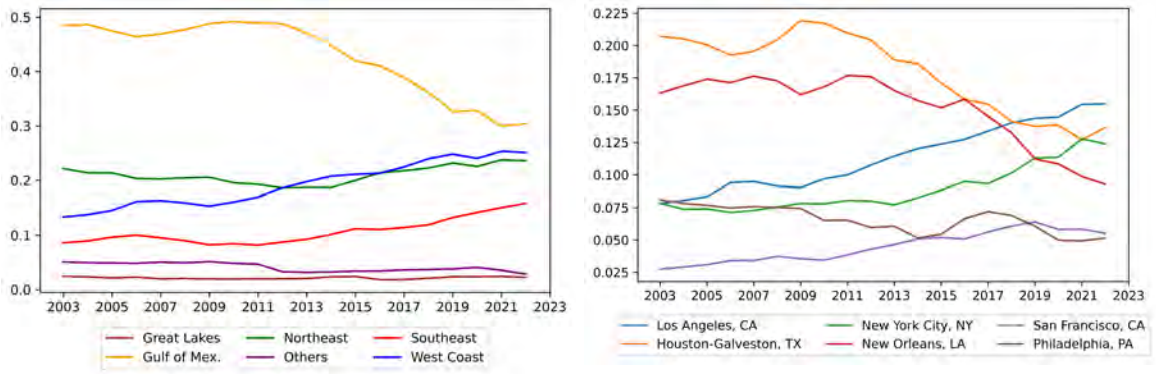
As such, each port experiences a different distribution of countries of origin, which will in general be different from that of the country as a whole. For example, Chinese imports represent more than 20% of all imports for ports in the West Coast, while they constitute only 2% of imports in ports in the Gulf of Mexico. Figure 17b shows this distribution for five US regions. We can interpret these graphs as showing how exposed different areas of the country are to shocks from different trading partners.

The combination of these two features (China's increased import share, coupled with Chinese exports going predominantly to the West Coast) led to a shift in the port and region of entry of US imports. At a regional level (Fig 18a), the share of imports entering through ports in the West Coast rose from 13% to 26%, while that of the Gulf of Mexico fell by 20 percentage points.[52] At the port level (Fig 18b), the

[52]This fall in part reflects the fact that the United States went from being a net importer of oil to a net exporter during this period.

Figure 18: *Region and port of entry for US imports. 2003-2023*

(a) *Region of Entry* **(b)** *Port of Entry*



Data through February 2023. Source: US Census Bureau

port of Los Angeles rose from fourth to first in terms of volume, surpassing New Orleans, Houston and New York.

Rail flows were similarly affected by this shift in international trade. Figure 19 shows how the share of rail trips originating in the Los Angeles area rose from less that 10% in 1998 to over 20% in 2018.

In this counterfactual analysis, we examine the extent to which this shift led to changes in rail costs across the United States.

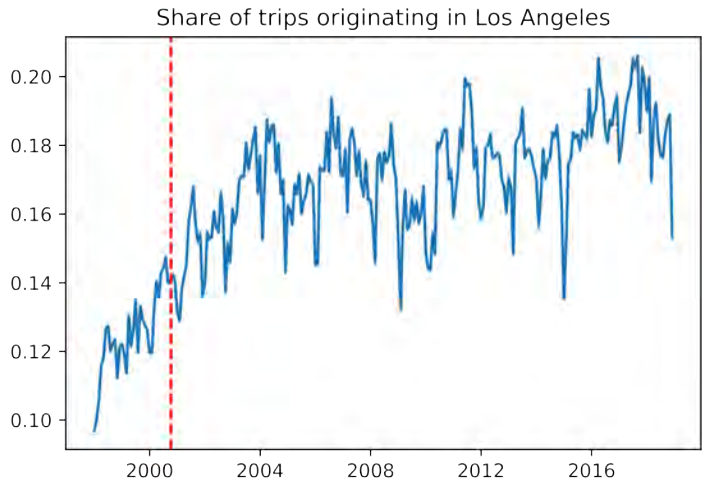
8.3.1 Counterfactual Flows

Our aim is to construct a set of counterfactual rail flows for a world in which Chinese exporters are forced to use ports according to their average size in the year 2000. To do so, we construct a measure of how much each port of entry was affected by the China shock, which we then subtract from the observed flows.

First, we use 2018 data to construct s_{op}, representing the proportion of imports from country o that entered the US through port p. This measure, similar to Figure 17b but at a more granular level, reflects the exposure of each port to trade from each country of origin. Next, for each country, we compute the value of imports in 2000 and 2018, $m_{o,00}$, $m_{o,18}$.

We can now define the counterfactual change in port-level imports, ΔQ_p^{CF}, by

Figure 19: *Proportion of trips originating in Los Angeles for two major US railroads, 1998-2018*



Trips weighted by size. The red line marks the passage of the United States-China Relations Act of 2000, in which Congress granted China Permanent Normal Trade Relations status and agreed to support its accession to the WTO.

keeping the aggregate import level (M_{18}) at its 2018 value, but changing the import composition to that of 2000.[53] This approach allows us to isolate the effect of the China shock on import composition while keeping the aggregate effect constant.[54]

$$\Delta Q_p^{CF} = \left(\sum_o s_{op} \left(m_{o,00} - m_{o,18} \right) \right) M_{18} \tag{13}$$

We match these ports p to nodes i in our rail network. We then convert units from tons of imports to carloads of freight by comparing the total imports in the port of Los Angeles to the import rail shipments originating in that port in our data.[55] We construct our counterfactual rail flows by scaling 2018 rail flows up and

[53] Due to data limitations, here we mix two measures of trade flows: s_{op} uses trade in tons, while m_{ot} uses trade in dollars.

[54] We do this to avoid modeling changes in domestic production in response to the change in imports. If we allowed aggregate imports to decrease, some of them would be replaced by local producers, which might then ship their products via rail.

[55] The CWS provides limited data on international trade. This variable is missing for almost all observations other than those for the port of Los Angeles.

down according to each port's exposure to the China shock.

$$Q_{ij}^{CF} = Q_{ij}^{obs} \frac{\max\left\{Q_i^{obs} + \Delta Q_i^{CF}, 0\right\}}{Q_i^{obs}} \tag{14}$$

8.3.2 Results

We start by examining how our counterfactual affects railroads' routing choices. To do so, we obtain each firm's routing choices by minimizing their estimated cost functions for both the observed demand matrix, and the counterfactual demand matrix defined in equation (14).[56] We estimate these choices separately for each month in 2018, and then average the results to obtain a yearly measure.

We plot the estimated change in flows for BNSF and UP induced by the China shock in Figures 20 and A.9. For each firm, we plot how flows changed for each edge, going both east and west. Both firms see an increase in eastbound flows originating in Los Angeles, and decreases in eastbound flows originating in San Francisco and westbound flows originating in New Orleans.

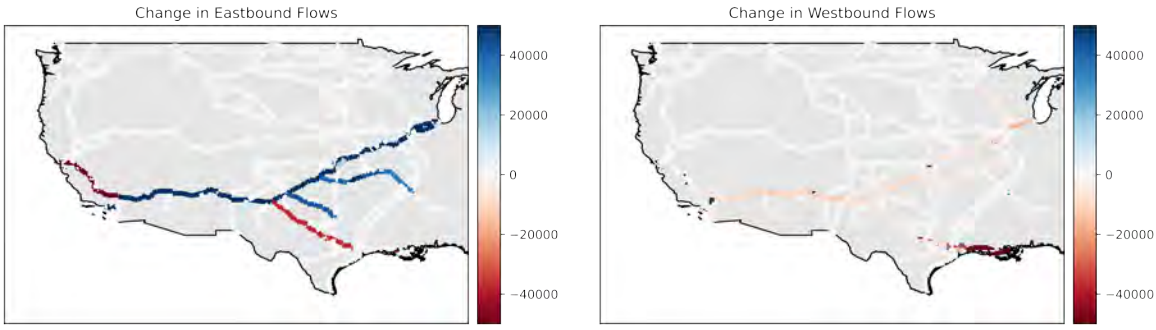
Despite this common shock, the effect on the network depends on routing choices. Consider the case of shipments originating in San Francisco: BNSF routes these shipments southeast through its main corridor in the Sun Belt. UP, on the other hand, sends them northeast through its tracks across the Mountain West. As a result, demand shocks in San Francisco propagate to Southern California if shipped by BNSF, and to Nevada and Colorado if shipped by UP.

Figures 21 and 22 present how our estimated marginal costs changed in response to the China shock for each firm. Whether an increase in flows translates into higher or lower marginal costs depends on the preexisting capacity utilization of the rail infrastructure. Section 7 shows that marginal costs at the edge-level have a U-shaped relationship with capacity utilization. Higher flows may lead to congestion if utilization falls to the right of the nadir of the U-shape, or they may lead to returns to scale if they fall to the left.

Consider, for example, a shipment traveling from Los Angeles to Chicago. Figures 20 and A.9 show that the traffic between those cities increased for both firms relative to the counterfactual. Despite this, Figures 21 and 22 show how edge-level costs increased for BNSF in almost all edges in this corridor, while they remained

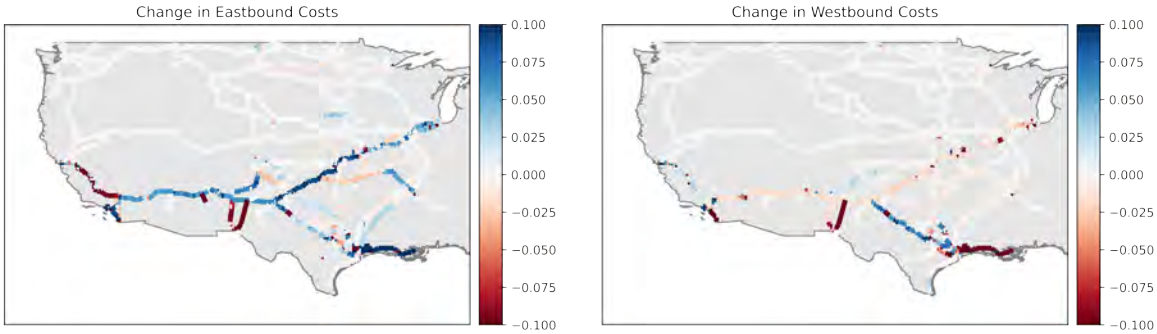
[56] For more details on how we implement this cost minimization, see Appendix D.

Figure 20: *Counterfactual Change in Flows, BNSF*



We plot the difference between the observed 2018 flows and the counterfactual flows. Values capped at $\pm 50,000$.

Figure 21: *Counterfactual Change in edge Marginal Costs, BNSF*

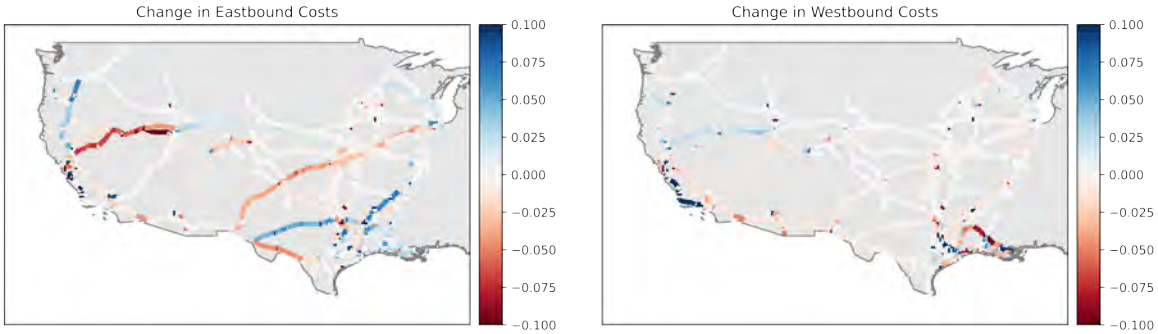


Values capped at ± 0.1.

constant or decreased for UP. Indeed, we observe that the same shock increased the cost of a shipment between these two cities by 7% for BNSF, while it decreased it by 1% for UP. This finding highlights the importance of considering the market structure of rail when conducting counterfactuals: The impact of a shock may differ depending on which firm is most affected by it.

Finally, we evaluate how these changes in edge-level marginal costs affected different regions of the country. We use regions constructed by the Bureau of Economic Analysis as our unit of observation. For each region, we construct a Laspeyres cost-index reflecting how shipping costs changed in response to the China shock. We take our estimated counterfactual change in Marginal Costs and

Figure 22: *Counterfactual Change in edge Marginal Costs, UP*



Values capped at ±0.1.

weight them using 2018 rail shipments for each firm-city-pair, Q_{odf}.

$$Outbound_o = \sum_d \left(\frac{Q_{odf}}{Q_o} \sum_f \sum_i x_{odi,f} \left(\sum_{uv} R_{uv}^{iE} \Delta mc_{uv} + \sum_n R_n^{iN} \Delta mc_n \right) \right) \tag{15}$$

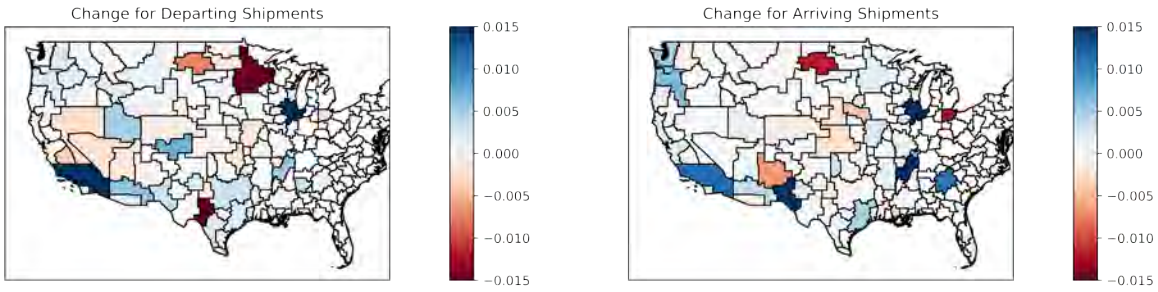
$$Inbound_d = \sum_o \left(\frac{Q_{odf}}{Q_d} \sum_f \sum_i x_{odi,f} \left(\sum_{uv} R_{uv}^{iE} \Delta mc_{uv} + \sum_n R_n^{iN} \Delta mc_n \right) \right) \tag{16}$$

where $x_{ijr,f}$ is the share of shipments between i and j for firm f taking route r. *Outbound* measures the change in costs for the average shipment originating in location o, which mostly affects the costs of firms producing in o, while *Inbound* measures the change in costs for the average shipment arriving to location d, which affects input costs and consumer prices for firms producing in d. We present these measures in Figure 23.

For departing and arriving shipments, the most affected regions are Los Angeles and Chicago, respectively, reflecting the increase in BNSF costs for that corridor.[57] We can see how the increase in costs in the Los Angeles area extends to Phoenix and, partially, to New Mexico, as the increased costs in the Los Angeles-Chicago

[57]The scale stops at 1.5% for legibility, but the actual measure for Los Angeles and Chicago is over 3%. This is to be expected, since the shock mostly affected shipments between these two cities. This is also consistent with Gerritse and Caragliu [2022], which found that most of the effect of the China shock on employment occurred in areas around ports, with increased freight prices partially shielding the rest of the country from competition.

Figure 23: *Counterfactual Change in region Shipping Costs, UP*



corridor affect these areas.

9 Conclusion and future work

In this paper, we developed a new framework to study railroad costs that explicitly considers the network structure of production. We documented how the firm's choices are consistent with nonmonotonic returns to scale and showed how to use a model of cost minimization, together with novel railroad choice data, to estimate railroads' cost functions. Finally, we used this cost function to study three policy questions: We showed how returns on investment vary more than 20x across states, and have significant network spillovers. We highlighted how market structure interacts with costs by simulating a merger that decreased total costs by 17%. And lastly, we showed how demand shocks have heterogeneous impacts across firms and space.

We remain aware of the limitations of this paper, and we plan to address them both in future work and future versions of this draft. We are currently working to expand our analysis to cover all six Class I railroads, allowing us to examine rail costs throughout the continental US. In addition, to examine the entire industry, it will be crucial to obtain an estimate of switching costs between firms. This will require us to move from GMM to moment inequalities, allowing us to recover fixed costs.

Another natural extension is to acquire additional data on rail contracts, to jointly estimate demand and costs. This would be particularly useful in our ongoing project, which integrates this structural model of rail with structural model of trucking described in Yang [2023], to be able to evaluate the freight system as a

whole.

References

T. Allen and C. Arkolakis. The Welfare Effects of Transportation Infrastructure Improvements [General Equilibrium Effects in Space: Theory and Measurement]. *Review of Economic Studies*, 89(6):2911–2957, 2022. URL https://ideas.repec.org/a/oup/restud/v89y2022i6p2911-2957..html.

T. Allen, D. Atkin, S. Cantillo, and C. Hernandez. Trucks. Working paper, 2023.

D. Argente, S. Moreira, E. Oberfield, and V. Venkateswaran. Scalable expertise. 2020.

J. Armstrong. *The Railroad: What it Is, what it Does*. Simmons-Boardman Books, 2008. ISBN 9780911382587. URL https://books.google.com/books?id=1ldhPgAACAAJ.

D. H. Autor, D. Dorn, and G. H. Hanson. The China Syndrome: Local Labor Market Effects of Import Competition in the United States. *American Economic Review*, 103(6):2121–2168, October 2013. URL https://ideas.repec.org/a/aea/aecrev/v103y2013i6p2121-68.html.

S. Bailey. Competition and coordination in infrastructure: Port authorities' response to the panama canal expansion. Working paper, 2021.

A. B. Bernard, S. J. Redding, and P. K. Schott. Multiple-product firms and product switching. *American Economic Review*, 100(1):70–97, March 2010. doi: 10.1257/aer.100.1.70. URL https://www.aeaweb.org/articles?id=10.1257/aer.100.1.70.

G. H. Borts. Production relations in the railway industry. *Econometrica*, 20:71–79, 1952.

G. H. Borts. Increasing Returns in the Railway Industry. *Journal of Political Economy*, 62:316–316, 1954. doi: 10.1086/257537. URL https://ideas.repec.org/a/ucp/jpolec/v62y1954p316.html.

G. H. Borts. The estimation of rail cost functions. *Econometrica*, 28:108, 1960. URL https://api.semanticscholar.org/CorpusID:154474402.

R. R. Braeutigam, A. F. Daughety, and M. A. Turnquist. The Estimation of a Hybrid Cost Function for a Railroad Firm. *The Review of Economics and Statistics*, 64 (3):394–404, August 1982. URL https://ideas.repec.org/a/tpr/restat/v64y1982i3p394-404.html.

G. Brancaccio, M. Kalouptsidi, and T. Papageorgiou. Geography, Transportation, and Endogenous Trade Costs. *Econometrica*, 88(2):657–691, March 2020. doi: 10.3982/ECTA15455. URL https://ideas.repec.org/a/wly/emetrp/v88y2020i2p657-691.html.

W. A. Brock. Contestable markets and the theory of industry structure: A review article. *Journal of Political Economy*, 91(6):1055–1066, 1983. ISSN 00223808, 1537534X. URL http://www.jstor.org/stable/1831204.

A. Caris, S. Limbourg, C. Macharis, T. van Lier, and M. Cools. Integration of inland waterway transport in the intermodal supply chain: a taxonomy of research challenges. *Journal of Transport Geography*, 41(C):126–136, 2014.

J. Chen, D. L. Chen, and G. Lewis. Mostly Harmless Machine Learning: Learning Optimal Instruments in Linear IV Models. Papers 2011.06158, arXiv.org, Nov. 2020. URL https://ideas.repec.org/p/arx/papers/2011.06158.html.

Y. Chen. Network structure and efficiency gains from mergers: Evidence from the u.s. freight railroads. Working paper, June 2023.

W. M. Daniels. *The Price of Transportation Service*. Harper & Bros., New York, 1932.

E. Dhyne, A. Petrin, V. Smeets, and F. Warzynski. Theory for Extending Single-Product Production Function Estimation to Multi-Product Settings. NBER Working Papers 30784, National Bureau of Economic Research, Inc, Dec. 2022. URL https://ideas.repec.org/p/nbr/nberwo/30784.html.

X. Ding. Industry linkages from joint production. Working Paper, 2022.

D. Dorn, D. Autor, and G. Hanson. On the Persistence of the China Shock. CEPR Discussion Papers 16688, C.E.P.R. Discussion Papers, Nov. 2021. URL https://ideas.repec.org/p/cpr/ceprdp/16688.html.

C. Ducruet, R. Juhasz, D. Krisztián, and C. Steinwender. All aboard: the effects of port development. LSE Research Online Documents on Economics 108496, London School of Economics and Political Science, LSE Library, Dec. 2020. URL https://ideas.repec.org/p/ehl/lserod/108496.html.

P. D. Fajgelbaum and E. Schaal. Optimal Transport Networks in Spatial Equilibrium. *Econometrica*, 88(4):1411–1452, July 2020. doi: 10.3982/ECTA15213. URL https://ideas.repec.org/a/wly/emetrp/v88y2020i4p1411-1452.html.

B. Feng, Y. Li, and Z.-J. M. Shen. Air cargo operations: Literature review and comparison with practices. *Transportation Research Part C: Emerging Technologies*, 56:263–280, July 2015.

S. Fuchs and W. F. Wong. Multimodal Transport Networks. FRB Atlanta Working Paper 2022-13, Federal Reserve Bank of Atlanta, Oct. 2022. URL https://ideas.repec.org/p/fip/fedawp/95074.html.

R. E. Gallamore and J. R. Meyer. *American Railroads: Decline and Renaissance in the Twentieth Century*. Harvard University Press, Cambridge, MA, 2014. ISBN 978-0-674-72564-5.

S. Ganapati, W. F. Wong, and O. Ziv. Entrepôt: Hubs, Scale, and Trade Costs. NBER Working Papers 29015, National Bureau of Economic Research, Inc, July 2021. URL https://ideas.repec.org/p/nbr/nberwo/29015.html.

M. Gerritse and A. Caragliu. Import competition and domestic transport costs. Tinbergen Institute Discussion Papers 22-071/VIII, Tinbergen Institute, Sept. 2022. URL https://ideas.repec.org/p/tin/wpaper/202200071.html.

P. K. Goldberg, A. K. Khandelwal, N. Pavcnik, and P. Topalova. Multiproduct Firms and Product Turnover in the Developing World: Evidence from India. *The Review of Economics and Statistics*, 92(4):1042–1049, November 2010. URL https://ideas.repec.org/a/tpr/restat/v92y2010i4p1042-1049.html.

Government Accountability Office. Freight railroads industry health has improved, but concerns about competition and capacity should be addressed. Report to Congressional Requesters GAO-07-94, Government Accountability Office, October 2006.

Z. Griliches. Cost Allocation in Railroad Regulation. *Bell Journal of Economics*, 3 (1):26–41, Spring 1972. URL https://ideas.repec.org/a/rje/bellje/v3y1972ispringp26-41.html.

R. E. Hall. The specification of technology with several kinds of output. *Journal of Political Economy*, 81(4):878–892, 1973. ISSN 00223808, 1537534X. URL http://www.jstor.org/stable/1831132.

C. L. Huntley, D. E. Brown, D. E. Sappington, and B. P. Markowicz. Freight Routing and Scheduling at CSX Transportation. *Interfaces*, 25(3):58–71, June 1995. doi: 10.1287/inte.25.3.58. URL https://ideas.repec.org/a/inm/orinte/v25y1995i3p58-71.html.

S. R. Jara-Díaz and C. E. Cortés. On the calculation of scale economies from transport cost functions. *Journal of Transport Economics and Policy*, 30(2):157–170, 1996. ISSN 00225258. URL http://www.jstor.org/stable/20053107.

G. Johnes. Costs and Industrial Structure in Contemporary British Higher Education. *The Economic Journal*, 107(442):727–737, 01 2012. ISSN 0013-0133. doi: 10.1111/j.1468-0297.1997.tb00038.x. URL https://doi.org/10.1111/j.1468-0297.1997.tb00038.x.

E. Jones. *Principles of Railway Transportation*. Macmillan Co., New York, 1931.

T. E. Keeler. Railroad Costs, Returns to Scale, and Excess Capacity. *The Review of Economics and Statistics*, 56(2):201–208, May 1974. URL https://ideas.repec.org/a/tpr/restat/v56y1974i2p201-08.html.

E. Khmelnitskaya, G. Marshall, and S. Orr. Identifying scope economies using demand-side data. 2023.

U. R. Kohli. Nonjointness and Factor Intensity in U.S. Production. *International Economic Review*, 22(1):3–18, February 1981. URL https://ideas.repec.org/a/ier/iecrev/v22y1981i1p3-18.html.

Y.-C. R. Lai and C. P. L. Barkan. Enhanced parametric railway capacity evaluation tool. *Transportation Research Record*, 2117(1):33–40, 2009. doi: 10.3141/2117-05.

M. O. Lorenz. Cost and Value of Service in Railroad Rate-Making. *The Quarterly Journal of Economics*, 30(2):205–232, 1916. URL https://ideas.repec.org/a/oup/qjecon/v30y1916i2p205-232..html.

B. Lu, H. Sun, M. Xu, P. Harris, and M. Charlton. Shp2graph: Tools to convert a spatial network into an igraph graph in r. *ISPRS International Journal of Geo-Information*, 7(8):293, 2018. URL https://doi.org/10.3390/ijgi7080293.

F. Maican and M. Orth. Determinants of economies of scope in retail. *International Journal of Industrial Organization*, 75(C), 2021. doi: 10.1016/j.ijindorg.2021.1. URL https://ideas.repec.org/a/eee/indorg/v75y2021ics0167718721000035.html.

J. R. Meyer, H. D. Mohring, M. J. Peck, J. Stenason, and C. J. Zwick. The economics of competition in the transportation industries. *Journal of the American Statistical Association*, 56:458, 1961. URL https://api.semanticscholar.org/CorpusID:123880247.

J. C. Panzar and R. D. Willig. Economies of scale and economies of scope in multi-output production. *Bell Laboratories economic discussion paper*, (33), 1975.

J. C. Panzar and R. D. Willig. Economies of scope. *The American Economic Review*, 71 (2):268–272, 1981.

J. R. Pierce and P. K. Schott. The Surprisingly Swift Decline of U.S. Manufacturing Employment. Technical report, 2014.

S. J. Redding and M. A. Turner. Chapter 20 - transportation costs and the spatial organization of economic activity. In G. Duranton, J. V. Henderson, and W. C. Strange, editors, *Handbook of Regional and Urban Economics*, volume 5 of *Handbook of Regional and Urban Economics*, pages 1339–1398. Elsevier, 2015.

doi: https://doi.org/10.1016/B978-0-444-59531-7.00020-X. URL https://www.sciencedirect.com/science/article/pii/B978044459531700020X.

W. Ripley. *Railroads, Rates, and Regulation*. Longmans, Green & Co., New York, 1927.

Surface Transportation Board. Confidential carload waybill sample, 1996–2018.

D. J. Teece. Economies of scope and the scope of the enterprise. *Journal of Economic Behavior & Organization*, 1(3):223–247, 1980. ISSN 0167-2681. doi: https://doi.org/10.1016/0167-2681(80)90002-5. URL https://www.sciencedirect.com/science/article/pii/0167268180900025.

US Energy Information Administration.

A. M. Wellington. *On the Economic Theory of the Location of Railways*. John Wiley & Sons, New York, 1893.

K. R. Williams. The Welfare Effects of Dynamic Pricing: Evidence From Airline Markets. *Econometrica*, 90(2):831–858, March 2022. doi: 10.3982/ECTA16180. URL https://ideas.repec.org/a/wly/emetrp/v90y2022i2p831-858.html.

R. Yang. (don't) take me home: Home preference and the effect of self-driving trucks on interstate trade. Working paper, 2023.

J. Zhang and E. Malikov. Off-balance sheet activities and scope economies in u.s. banking. *Journal of Banking & Finance*, 141:106534, 2022. ISSN 0378-4266. doi: https://doi.org/10.1016/j.jbankfin.2022.106534. URL https://www.sciencedirect.com/science/article/pii/S0378426622001285.

A Industry History

The US freight railroad industry is composed of private for-profit firms that both own and operate the rail infrastructure. In this respect, it differs from most other countries' rail industries, which feature either partial or complete nationalization of the sector.

The US rail network was mostly laid out in the late 19th and early 20th centuries, when rail was the main source of internal transportation for both commodities and passengers. The network peaked in size in 1916, at around 300,000 miles of track. After World War II, the widespread adoption of cars, along with the development of commercial flight and trucking, challenged the primacy of rail, which began a slow decline. Half of the original network was either abandoned or sold off to smaller, local firms.[58]

The industry can boast being one of the first industries to be regulated by the federal government: In 1887, in response to concerns over monopolistic practices, Congress created the Interstate Commerce Commission (ICC) to regulate railroads and promote competition. The ICC determined which markets firms were allowed to serve and its approval was required for any price changes, mergers, acquisitions, or line abandonments. Railroads were subject to the "common carrier restriction," a requirement to charge similar prices for all comparable services, which effectively prevented them from engaging in price discrimination.

This regulation remained practically unchanged until the 1970s. During this period, the development of automobiles and commercial planes turned rail passenger services, once a key source of profits, into a financial burden for firms. In addition, the ICC attempted to foster competition by protecting other transit firms (Gallamore and Meyer [2014]), which often stifled innovation.[59] As a result, by the 1970s railroads were in a dire financial situation, with many of them close to filing for bankruptcy.

Following the successful deregulation of the airline and trucking industries,

[58]The Minuteman Trail connecting Cambridge to Lexington is an example of a repurposed rail line.

[59]A famous case is that of Southern Railway's 'Big Johns'. These new cars featured a better load-to-weight ratio and allowed for more efficient shipping. Southern Railways thus applied for a 60% rate reduction in August of 1961, which was blocked by the ICC on grounds that it would "drive truckers and other railroads out of business." SR took the matter to the Supreme Court twice before a ruling in its favor in 1965. Source: http://www.railgoat.railfan.net/railwhales/a-axles.htm

Congress passed a series of acts during the Carter administration, culminating in the Staggers Act of 1980, which substantially deregulated the rail industry. The Staggers Act allowed firms to abandon unused lines, granted them greater freedom in setting prices, and permitted long-term contracts with clients.

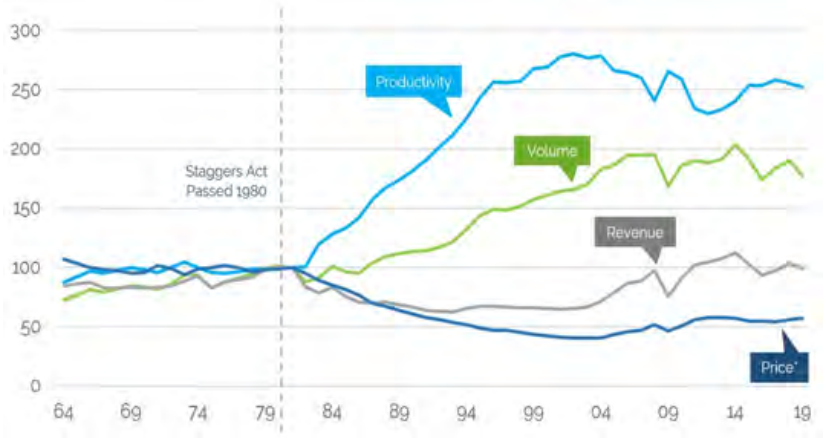
This successfully revitalized the industry. Despite shedding thousands of miles of unused track, the volume transported by Class I railroads doubled over the next 25 years, prices fell by 50% in real terms, and profits skyrocketed (Figure A.1). Key to this transformation was the railroads' newly-gained ability to sign long-term contracts, as these contracts were instrumental in solving the hold-up problem inherent in undertaking rail investments that will service other firms: Because each railroad is in charge of building its own infrastructure, firms may be reluctant to build tracks to service another firm without a guarantee that it will not ask for a discount once the investment is made. Long-term contracts allowed railroads to lock rates for long enough to recoup the investment costs.

The new regulations also gave railroads the freedom to set rates without the need for government approval. They did include a provision allowing firms without any other transportation alternatives to challenge rates if they exceeded 180% of the railroad's variable costs. But due to the difficulty of estimating these variable costs in a network industry, only a handful of rates have been challenged, with mixed results in the courts. Nevertheless, despite the lax government oversight, prices fell steadily as shipping volumes increased and firms abandoned unused track.

Regulators were also substantially more lenient in allowing mergers and acquisitions, especially between firms connected end-to-end. A flurry of mergers followed this deregulation, during which the 20-odd mid-sized US railroads consolidated into seven by the turn of the millennium. The remaining firms in the market—Union Pacific (UP), Burlington Northern Santa Fe (BNSF), CSX Transportation (CSX), Norfolk Southern (NS), Kansas City Southern (KCS), Canadian Pacific (CP), and the Canadian National Railway (CN)—own a network spanning 140,000 miles of track, making it the largest freight rail network in the world (see Figure A.2).[60] The pace of mergers slowed down in 2000 when the Surface Transportation Board (STB, the successor of the ICC) turned down a proposed merger between BNSF and CN; no new mergers were attempted for the next twenty years, until CP acquired KCS in 2023.

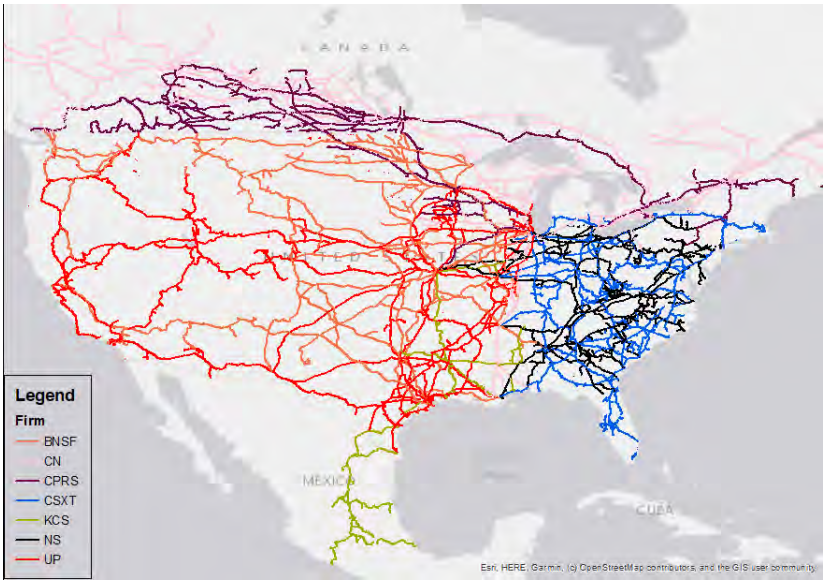
[60] Association of American Railroads (2023)

Figure A.1: *Industry performance after deregulation.*



Source: Association of American Railroads

Figure A.2: *US Rail Network in 2023*



B Dataset construction

In this subsection, we explain in more detail how we construct our data. We can answer additional questions or share our code upon request. **Confidential Waybill Sample.** We start by reading the raw Carload Waybill Sample .txt files into Stata. We use the 'Waybill Date' (the date at which the originating railroad prepares the waybill) as the date of each observation. We convert distance into miles (from tenths of a mile). We create a commodity categorical variable using 2-digit STCC codes. We drop all observations with zero revenue. We treat the American and the Canadian branches of Canadian Pacific as a single firm, and we label all Class II and Class III railroads as "Other."

Next, for each origin and destination, we match SPLC (Standard Point Location Code) locations to Zip codes using data scraped from Railinc. We then use the coordinates of the centroid of each Zip code (or BEA, if we could not match or SPLC was missing) as the origin or destination of each trip.

We then take all trips involving more than one railroad (about 17% of the total) and split them into multiple one-railroad trips. Each of these new trips will have as its origin the location where it received the cargo from the previous firm, and as its destination the place where it hands it over.

Network Construction. We start this process by constructing a shapefile containing all the tracks that a firm either owns or has trackage rights to by modifying the North American Rail Lines shapefile created by the NTAD. We transform this shapefile into a network using the Shp2Graph R package (Lu et al. [2018]). We then simplify this network to a resolution of 0.2 decimal degrees (roughly 14 by 11 miles, depending on the latitude). This process combines all nodes lying within the same 0.2 by 0.2 square into a single node. Two of these nodes are then connected by an edge if the original network featured an edge connecting points belonging to each of the two new nodes. When computing an edge's distance, we compute the shortest route between the two nodes closest to each node's geographical center. Finally, we delete all nodes that have only two edges and are never the origin or the destination of a trip.[61] We are left with roughly 1,700 nodes and edges.

Route definition. We define origins and destinations by rounding the coordinates

[61] This leaves us with three types of nodes: nodes that are the origin of some shipments, nodes that are the destination of some shipments, and nodes where the network branches into multiple directions.

of each location to the closest 0.2 degrees. We drop observations for which we do not observe the origin or the destination (1% of the total). For each origin-destination pair, we define a route as a unique combination of distance (rounded to the closest 20 miles) and set of states visited. We identify and drop a few routes (accounting for 0.04% of shipments) that have a data input error: These are observations where the distance traveled is over 10 times the shortest distance. We drop these observations. Next, for each railroad, we obtain the set of coordinates that are either an origin or a destination of at least one of their trips.

For the density of passing lanes, we use the classification used by NTAD, where sidings are classified with a 0 in the "number of tracks" variable. We distinguish between sidings (loops that connect back to the network) and spurs (tracks that branch out and connect to a destination). For each node, we compute the total mileage of tracks and of sidings. Then, for each edge connecting two nodes, we compute the ratio of the sum of siding mileage and track mileage.

We show preliminary results for UP, since we have not finished computing the routes for some city pairs. We allocate these missing observations to the closest route we have computed for that city pair. This accounts for 18% of flows at the moment, but that number is steadily falling as more routes are computed.

Variable construction. We construct "average tracks" using the "number of tracks" variable in the NTAD shapefile. The original variable applies to individual tracks. We fit it into our network by computing the average number of tracks in the 20 miles around each node; then, we convert this measure into an edge-level one by averaging this variable for the two nodes belonging to each edge. We compute the trackage measure in a similar way. We assign nodes according to whether they are owned by the firm. We then define an edge as involving "trackage" if one of its nodes is owned by another firm.

We compute the "grade" variable by computing the change in elevation between two neighboring nodes (before simplifying the network). If a locomotive travels with constant energy, all the energy lost in climbing up a slope is regained when going downhill, unless the locomotive needs to brake, in which case part of the energy is lost as heat (Armstrong [2008]). We account for this by capping the energy recovered when going downhill: If the slope is less than -1%, we assume the locomotive must brake, losing all energy beyond that number.

C Profit maximization

C.1 Expansion to Profit Maximization

Due to lack of good data on railroad pricing, we focus our analysis on the cost minimization problem. In this section, we show the assumptions required to embed our analysis in a profit maximization framework.

The firm faces a demand vector $P(Q_t, \mathbf{X})$, which depends on the quantities shipped, Q_t, and the routes chosen to implement them, $\mathbf{X}$. P is a $N_E^2 \times 1$ vector indicating the price the firm can obtain for each shipment as a function of the quantity shipped and its route choices.

Below, we commit a slight abuse of notation by letting Q_t refer to both the matrix of quantities to be shipped and its vectorized form, stacking all columns into a single vector. With that caveat in mind, standard profit maximization implies that the firm solves:

$$\max_{Q_t, \mathbf{X}} \{P(Q_t, \mathbf{X})Q_t - C(F(\mathbf{X}))\} \tag{17}$$

subject to equations (1) and (2).

To highlight the assumptions required to study cost minimization in isolation, let $P(Q_t, X) = P(Q_t) + h(X)$, where h is a function capturing how different routes affect the price the firm can charge for a given shipment. This decomposition works for a variety of demands, described later in this appendix. The problem then becomes:

$$\max_{Q_t} \left\{ P(Q_t)Q_t - \min_{\mathbf{X}} \underbrace{\{C(F(\mathbf{X})) - h(\mathbf{X})Q_t\}}_{\tilde{C}(\mathbf{X})} \right\} \tag{18}$$

Any cost function we recover will equal $\tilde{C}(\mathbf{X})$, which combines the true cost function plus the effect of the routes chosen on demand.

As a test of the likely size of h, we evaluate whether routing choices $\mathbf{X}$ predict prices P. We first use a random forest to estimate the conditional expectation of prices given all our observables excluding routing choices. Next, we compare that expectation to a second random forest where we include the routing choice as an explanatory variable. We find that including routing choices in the estimation does not significantly decrease the prediction error. Thus, we conclude that route choices

do not affect pricing behavior. This is consistent with pricing and routing being chosen independently by different departments within a firm. A similar behavior was documented by Williams [2022] in the airline industry.

C.2 Demands for which the separation works

Consider the firm problem as stated in equation (18). Under which conditions will the inverse demand function, $P(Q, \mathbf{X})$ be separable in Q and $\mathbf{X}$? That is, when will $P(Q, \mathbf{X}) = P(Q) + h(\mathbf{X})$? We list below some common demands, and note whether this separation is possible.

Linear Demand Suppose that demand is linearly separable in price and route characteristics:

$$D(P, \mathbf{X}) = \alpha P + \beta t(\mathbf{X})$$

This implies inverse demand which is separable in quantity and characteristics

$$P(Q, \mathbf{X}) = \alpha^{-1} Q - \beta \alpha^{-1} t(\mathbf{X})$$

As a result, let $h(\mathbf{X}) = -\beta \alpha^{-1} t(\mathbf{X})$. Under this formulation, we can interpret $t(\mathbf{X})$ as the time it takes to ship all products as a function of route choices; while $h(\mathbf{X})$ represents the effect this time has on the willingness to pay for the shipment.

Log-Linear Demand

$$D(P, X) = P^{\alpha} \mathbf{X}^{\beta} \implies \log Q = \alpha \log P + \beta \log \mathbf{X}$$

The inverse demand curve is

$$\log P = \alpha^{-1} \log Q - \alpha^{-1} \beta \log \mathbf{X}$$

This is linearly separable in logs, but not in levels.

CES Demand Suppose we have CES utility/profits over goods j, where $\xi(\mathbf{X})$ is a function of good j's characteristic (Hortacsu-Joo).

$$u(x,y) = \left(\sum_j \xi(X_j)^{1/\sigma} Q_j^{(\sigma-1)/\sigma} \right)^{\sigma/(\sigma-1)}$$

The implied demand function is, assuming income $Y = 1$,

$$Q = \frac{\xi(X_j) P_j^{-\sigma}}{\xi(X_j) P_j^{-\sigma} + \sum_{k \neq j} \xi(X_k) P_k^{-\sigma}}$$

Hold the other goods constant and denote them by κ.

$$Q = \frac{1}{1 + \kappa \xi(X_j)^{-1} P_j^{\sigma}}$$

$$P_j = (\kappa^{-1}(Q_j^{-1} - 1)\xi(X_j))^{1/\sigma}$$

In contrast, a CES utility like the following would generating linearly separable inverse demand curves.

$$u(x,y) = \left(\sum_j (\alpha \xi(X_j) + (1-\alpha) Q_j)^{(\sigma-1)/\sigma} \right)^{\sigma/(\sigma-1)}$$

Logit Demand Suppose that there is a continuum of consumers who receive i.i.d. logit shocks, and they care about characteristic $\xi(\mathbf{X})$ which is a function of routing $\mathbf{X}$. Suppose their mean value is $\pi = \alpha P + \beta X$ and there is an outside option with value 0. Demand for product i is

$$D(P,X) = \frac{\exp(\alpha P + \beta X)}{1 + \exp(\alpha P + \beta X)} = \frac{1}{\exp(-\alpha P - \beta X) + 1}$$

Inverse demand is

$$P = -\alpha^{-1} \log(Q^{-1} - 1) - \alpha^{-1} \beta X$$

So $h(X) = -\alpha^{-1}\beta$.

D Minimizing the cost function

The optimal routing problem is a version of the traveling salesman problem, which makes it an NP-hard problem. In order to solve it in finite time, we must theremore make some assumptions to reduce its complexity.

To that end, we restrict the firm to choose its routes from the set of routes it has used in our data going back to 1998. This reduces the number of potential routes from $N_E!$ to around 120,000 per firm. In some specifications where we need to further speed up computation time, we focus on routes that have carried at least 2.5% of total shipments. When we do this, we are left with around 6,000 routes for 1,500 city pairs.

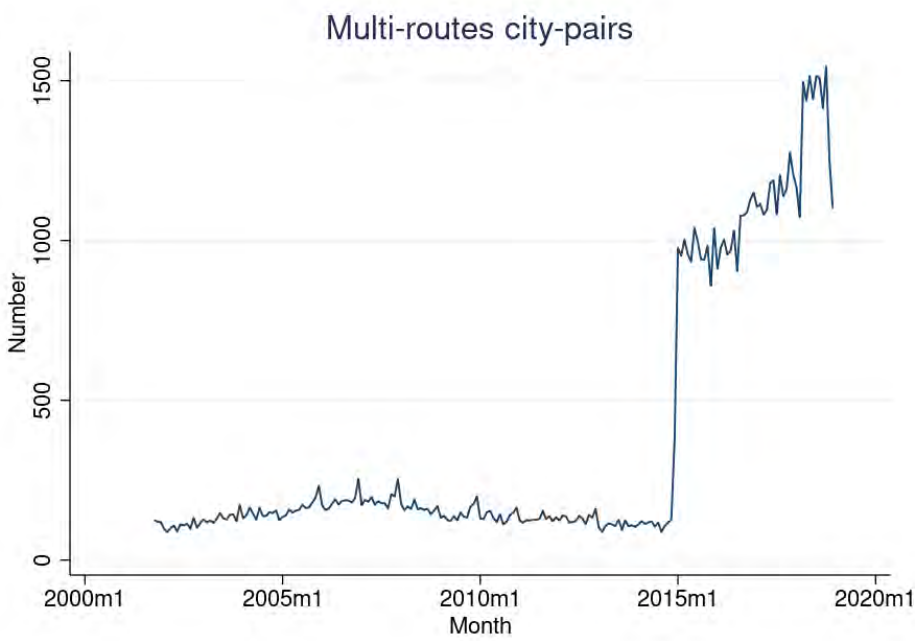
Because of the presence of trackage rights, firms' problems are interconnected: UP can choose to route part of its shipments using BNSF's networ,k and vice versa. We model this by having firm i take firm j's choices as given and computing its best response. We start with a guess for each firm's choices, and then iterate best responses until they converge to a Nash Equilibrium.

We solve the minimization problem using SQP implemented by KNITRO 13.2.0. We try multiple initial values and use the solution with the lowest costs. We find a Nash Equilibrium after about 4 hours.

E Routing data sample break

In Figure A.3, we plot the number of city pairs serviced that utilize more than one route by one of the largest four railroad companies. This number is fairly consistent from 2001 to 2014, but it has a clear break in 2015. To keep our analysis consistent, we restrict our attention to the 2015-2018 period.[62] In addition, to attenuate concerns about measurement error, we only consider pairs of routes that differ by at least 50 miles from each other. This leaves us with a sample of around 80,000 route-pair-month observations per railroad.

Figure A.3: *Number of city pairs utilizing more than one route*



F Additional Figures

[62]There is an additional, smaller jump between 2017 and 2018. It is small enough that we ignore it in our analysis.

Figure A.4: *Estimated marginal costs at the edge level - UP.*

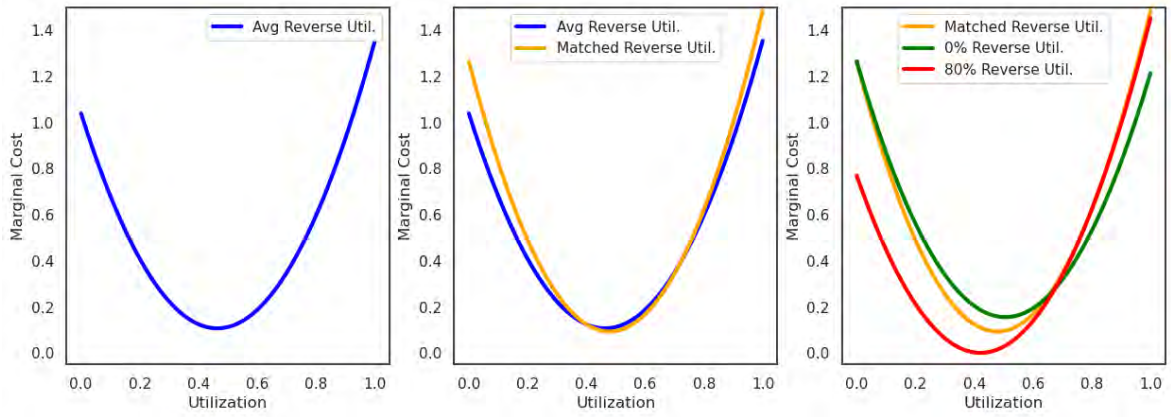


Figure A.5: *Estimated marginal cost per mile 1998-2018 - BNSF*

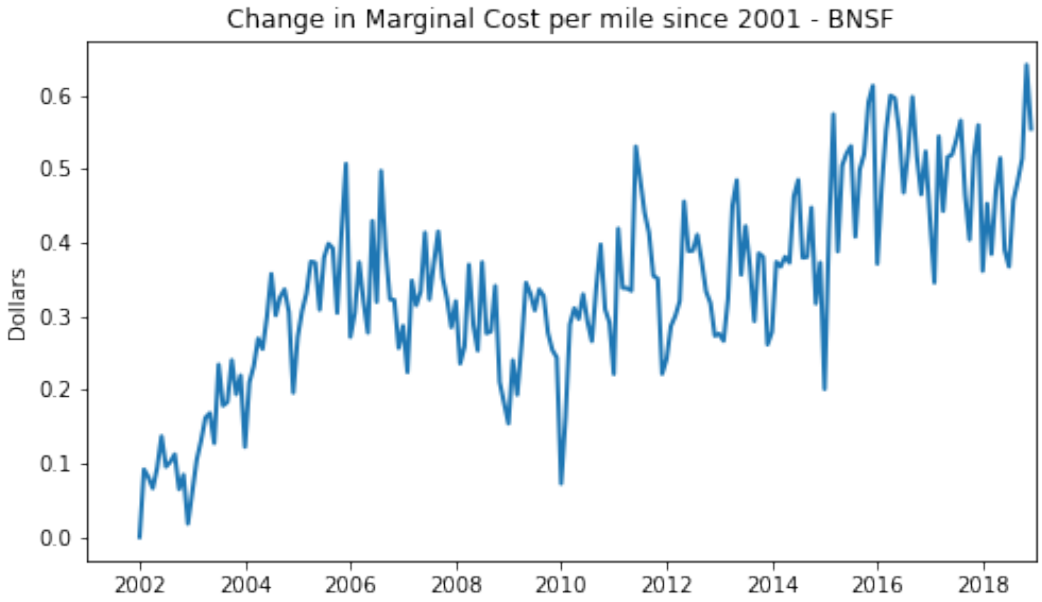


Figure A.6: *Estimated marginal cost per mile 1998-2018 - UP*

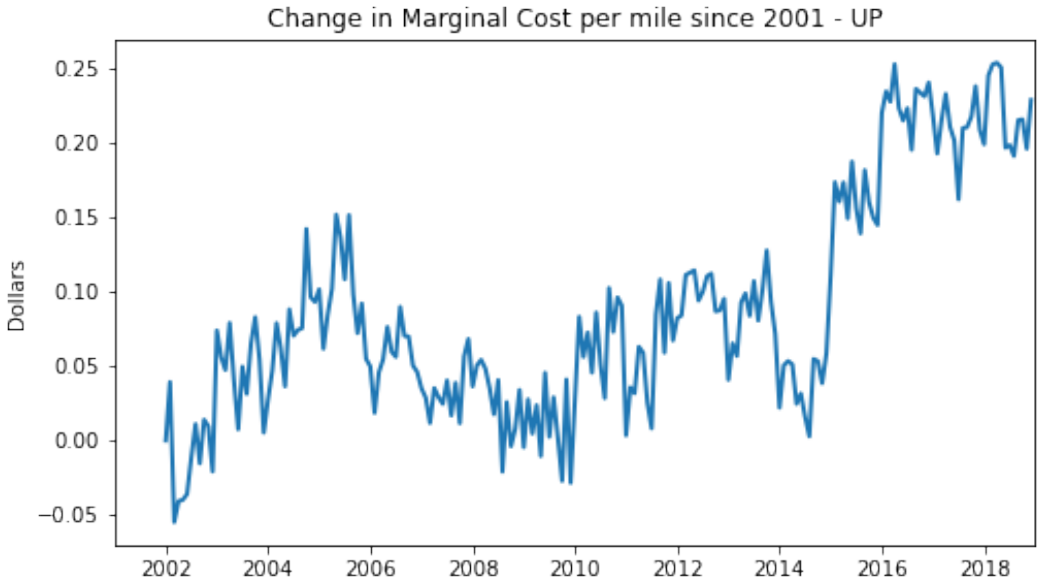
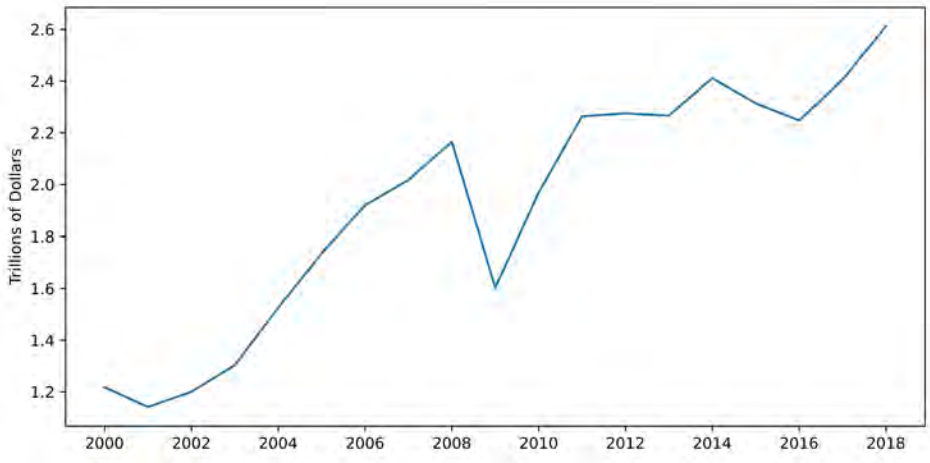
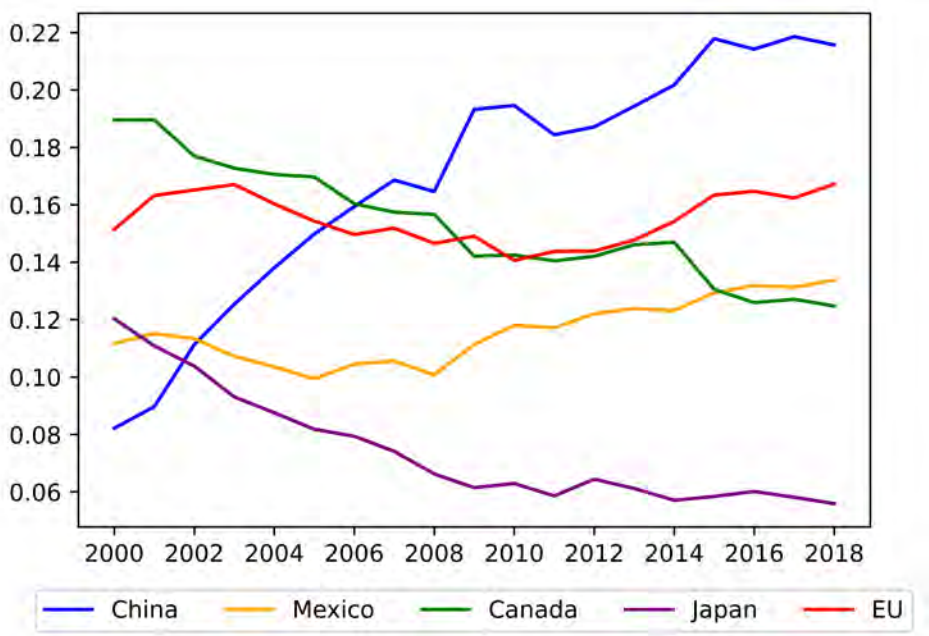


Figure A.7: *Total US imports, 2000-2018*



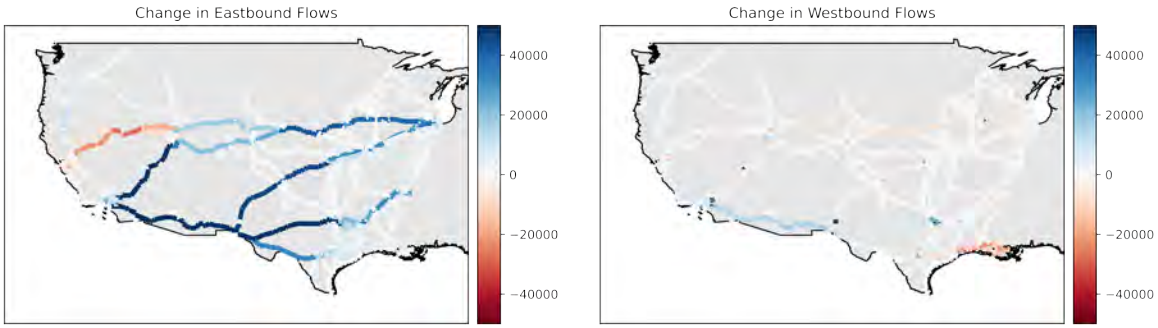
Total US imports in trillions of dollars. Source: WITS.

Figure A.8: *US imports by country, 2000-2018*



Share in US imports of 5 largest trading partners. Source: WITS. Note: EU refers to countries in the EU in 2022.

Figure A.9: *Counterfactual Change in Flows, UP*



Values capped at $\pm 50,000$.