Economies of Scale and Scope in Railroading*

Pedro G. Degiovanni† Ron Yang‡

October 17, 2024

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 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.

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

JEL Classification: D24, L92, R41

*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, and numerous colleagues and workshop attendees for their suggestions and comments. Leisy Zhang provided helpful research assistance. We acknowledge financial support from the Chae Family Economics Research Fund and UBC Sauder Exploratory Grant.

†Charles River Associates. Contact: pdegiovanni@crai.com

‡University of British Columbia. Contact: ron.yang@sauder.ubc.ca

Transportation costs are key in shaping the spatial distribution of economic activity and trade.[1] These costs are not technological parameters, but arise endogeneously 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 both national and local levels.[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 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 find that track-level costs exhibit both increasing and decreasing returns to scale at different levels of utilization, and

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

[2] These modes range from atomistic ships operating in a competitive (yet frictional) market (Brancaccio et al. (2020)), to large oligopolies controlling a large rail network (Chen (2023)).

[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.

[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.

physical infrastructure can significantly reduce costs. In counterfactual simulations, the network structure of production implies that infrastructure investment engender substantial network spillovers, which local governments are unlikely to internalize.

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: 51 percent of shipments deviate from the shortest possible path by 5 percent or more, and 29 percent deviate by 10 percent or more.[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 percent, 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 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

[5] In contrast, previous work studying the rail industry using the same dataset has assumed that all shipments follow the shortest feasible path between origin and destination, conditional on interchanges (Fuchs and Wong (2022)).

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

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 percent 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 percent utilization, and imply significant economies of scale. Traversing an edge with 0 percent utilization is $2.00 more expensive than a shipment at 55 percent utilization. As utilization increases past this efficient level and approaches 100 percent, costs increase by $1.25. These differences represents 40 percent and 25 percent 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 percent.

Finally, we use our framework to analyze the impact of public investment in rail infrastructure. Many state and local governments fund rail infrastructure investments, aiming to reduce road traffic congestion for motorists and reduce the cost of shipping to and from their locations. 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), but Arizona only captures 3 percent of these cost savings. More generally, high-centrality states are very productive

but capture a low share of cost savings. These results suggest that decision-makers may fail to internalize the benefits from infrastructure investment, highlighting the potential role of the federal government in coordinating rail investment.

Our work contributes to the literature on transportation costs across networks. 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 literature on economies of scope within multiproduct firms. Despite the prevalence and importance of multiproduct firms in both developing and developed countries, 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.[7] 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,[8] directly estimating the cost functions using quantity and input data,[9] or leveraging information on demand (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.

[7] See Bernard et al. (2010) and Goldberg et al. (2010).

[8] As evidenced by Hall (1973), Kohli (1981), Brock (1983), Johnes (2012).

[9] This approach is often referred as estimating a transformation function, see Dhyne et al. (2022) and Maican and Orth (2021), and Zhang and Malikov (2022).

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)), 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 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, and 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 presents out counterfactual exercise, and Section 9 concludes.

2 Industry Background

Railroad operations 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.[10] This paper focuses on the final input: the utilization of rail infrastructure, including tracks and yards.

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.

[10]Railroads primarily ship heavy, long-distance commodities, such as containers, coal, agricultural products, and motor vehicles. In 2012, containers surpassed coal as the most common commodity on rail.

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.

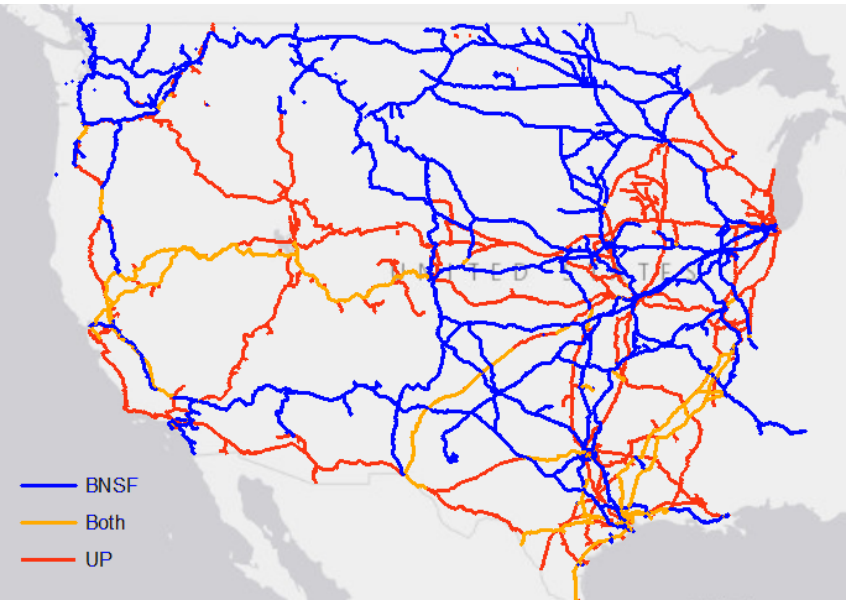
As the quantity of shipments increases, tracks can also become congested. Whenever a faster train encounters a slower train, the fast train must reduce its speed and wait until it reaches an "overtaking siding," where the slower train can move aside and allow the faster train to pass. Similarly, most 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.

Market Structure The US freight railroad industry is composed of for-profit firms which own and operate rail infrastructure, with six major firms operating the vast

majority of all rail shipping.[11] In this paper, we focus on the two largest firms: Burlington Northern Santa Fe (BNSF) and Union Pacific (UP).[12] These railroads have a combined 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: *Rail Networks*



Note: This figure plots the rail networks owned by either Union Pacific (UP), in red, or Burlington Northern (BNSF), in blue. Yellow networks are owned by one of the two firms, but the other firm can also access via trackage rights. Source: National Transportation Atlas Database.

[11]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. 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]The other Class I railroads include CSX Transportation (CSX), Norfolk Southern (NS), Canadian Pacific Kansas City (CPKC), and the Canadian National Railway (CN).

3 Data

3.1 Sources

Our main dataset is the U.S. Surface Transportation Board's Carload Waybill Sample, which contains a 2.7 percent stratified sample of all shipments on Class I railroads.[13] Each observation in the Waybill Sample provides route information (origin, destination, states visited, and distance), the commodity transported (as measured by STCC codes), its quantity, the price charged, and a list of all railroads involved in the trip.[14] The Waybill Sample contains around 600,000 waybills per year from 1996 to 2018.[15]

We use the National Transportation Atlas Database's North American Rail Lines dataset to define the rail network and construct physical characteristics. This dataset contains the networks of all Class I railroads, including infrastructure characteristics such as the average number of tracks, the frequency of rail sidings. We obtain elevations from the U.S. Geographic Names Information System (GNIS). To construct a demand shifter instrument, we leverage monthly coal production data from the Energy Information Administration.

3.2 Inferring Routes

We re-construct the paths taken by each shipment in the Waybill Sample using the contained route information: origin, destination, states visited, and distance. 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. Our approach exploits the

[13] A waybill is the receipt generated when a railraod delivers a carload of freight.

[14] The distance variable in the Waybill Sample is produced by Railinc, which matches waybills to an internal events database containing the movements of individual cars in the network. 60 percent of waybills are matched exactly. The remainder are matched to comparable waybills (i.e., waybills that share the same origin, destination, commodity, and interchanges). Finally, if there are no comparable trips, Railinc uses a proprietary cost-minimizing model to compute a route for the waybill.

[15] The waybill reported prices are known to be poor indicators of actual prices paid (Government Accountability Office (2006)), so we will not use reported prices in estimation. The dataset contains a binary indicator for contract shipments status, but does not provide any other information about contract structure.

sparsity of the rail network to identify the route or set of routes consistent with each waybill observation.

Simplified Network We begin by creating a simplified representation of the rail network.[16] For each railroad, we use Shp2Graph to transform the shapefile into the base network depicted in Panel A of Figure 2 (Lu et al. (2018)). We project this base network onto a grid of 0.2 decimal degree squares, and we create a simplified network where each node represents a grid square and each edge represents a connection between grid squares in the base network.[17] We set the length of the edge between two grid squares by finding the most central nodes in each square in the base network, and computing the distance in the base network between these two central nodes. Finally, some nodes only serve as intermediate points along longer routes: they have two edges and are never the origin or destination of a shipment. We replace these nodes and their adjacent edges with a single edge with length equal to the sum of the two edges, leaving us with three types of nodes: origins, destinations, and points where the network branches into multiple directions. Panel B presents our final simplified network.

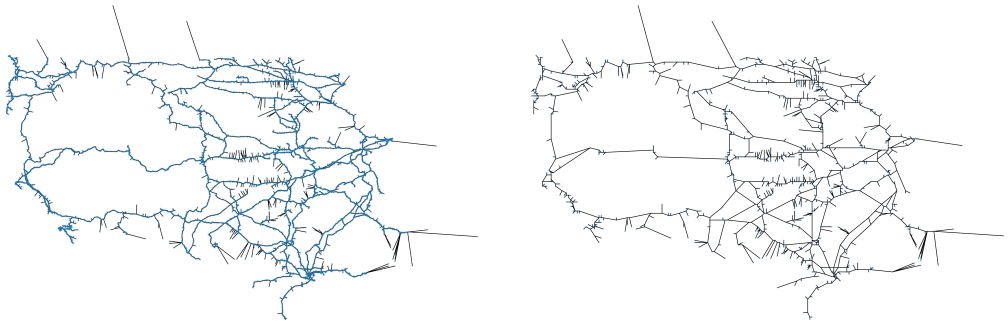
Characteristics We construct node and edge characteristics for the simplified network. We define the average number of tracks by computing the average for each grid square and taking an average of the grid squares visited for each edge.[18] For the density of passing lanes, we compute the total mileage of tracks and of sidings in each node. Then, for each edge connecting two nodes, we compute the ratio of the sum of siding mileage and track mileage. We compute the "grade" variable by

[16] Prior to creating the graph representation, we drop observations with zero reported revenues, observations corresponding to Class II and Class III railroads, and observations with missing origins and destinations. We round distances to miles, and drop observations if the reported distance is more than 10 times the shortest distance for the origin-destination pair. Shipments which involved more than one railroad are split into multiple single-railroad shipments. Origin and destination locations are coded using Standard Point Location Codes (SPLC); we use data from Railinc to match these to zip code centroids.

[17] At the average latitude of our data, a 0.2 decimal degree square translates into a 14 by 11 mile rectangle.

[18] For edges formed by dropping intermediate nodes, we take a distance-weighted average over all grid squares.

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



Note: Panel A plots BNSF's network in its raw shapefile format. Connections outside the rail network are represented by straight lines at the edge of the network. Panel B plots a simplified graph representation of the BNSF network, after rounding to 0.2 degrees and removing intermediate nodes. Source: National Transportation Atlas Database.

computing the change in elevation between two neighboring nodes, with a lower bound at -1 percent to account for energy lost to breaking.[19]

Route Construction For each origin-destination pair, we observe multiple combinations of distance traveled and states visited, with each combination representing a different route. 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 Panel A of Figure 3. Next, we compute the k-th shortest loopless route between origin and destination within that subnetwork. We progressively increase k until we find all routes within 20 miles of the target distance.[20]

If there is only one such route, we conclude that it is the route implied by the origin-destination-route tuple. If there are multiple routes within the target distance, 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

[19] 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 assume that a locomotive must break if the slope is less than 1 percent, limiting energy recovered when going downhill.

[20] The diameter of a grid square in our simplified network is approximately 20 miles.

Figure 3: *Subnetwork (left) and inferred route (right)*



Note: This figure illustrates the two steps of inferring the route for a 2,500-mile shipment originating in Los Angeles, terminating in Chicago, and visiting California, Arizona, New Mexico, Texas, Oklahoma, Missouri, and Illinois. Panel A presents the subnetwork containing only the transited states. Panel B presents the inferred routes which match the 2,500-mile distance. The width and color of an edge indicate the probability that a given edge belongs to the actual route. Source: Confidential Waybill Dataset, National Transportation Atlas Database.

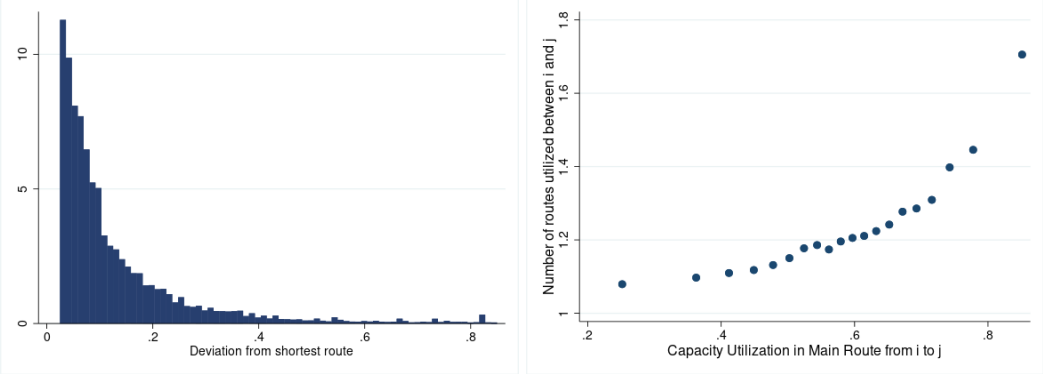
at the proportion of potential routes within the target distance that includes that edge or node. As an illustration, Panel B of Figure 3 plots the inferred route for a shipment from Los Angeles to Chicago, where the edge width indicates our inferred probability of belonging to the actual route. We repeat this process for all routes in the sample to build a dataset of U.S. railroad route choices.[21]

4 Reduced Form Evidence

In this section, we document three facts about railroads' routing choices which will inform our subsequent analysis: routes deviate from shortest path, capacity utilization predicts route choices, and prices have a U-shaped relationship with capacity utilization.

[21] We are able to compute route for 82 percent of UP shipments. For routes which we cannot compute, we allocate their traffic to the route with the closest distance.

Figure 4: *Deviations from shortest path*



Note: Panel A plots a histogram of the additional distance travelled by routes which deviate from shortest path. Deviations are measured in percentages of the shortest path distance. Observations correspond to shipments, and are weighted by their sampling weights. Panel B plots the number of routes used between a given origin-destination pair against the capacity utilization of the main route between that pair. Source: Confidential Waybill Dataset

4.1 Railroads do not minimize distance

Railroads often make shipments along routes that deviate from the shortest distance route. Using the routes constructed in Section , we find that 84 percent of shipments take a route which is at least 1 percent longer than the shortest route. Panel A of Figure 4 plots a histogram of the proportional distance these shipments travel above the shortest distance. Among shipments that deviate from the shortest route by at least 5 percent, the median shipment takes a route 11.6 percent longer than the shortest path, and 10 percent of shipments taking a route 40 percent longer than the shortest route. These patterns indicate that railroads may trade off distance with other factors when making routing choices.

4.2 Capacity utilization predicts route choices

Next, we document the relationship between capacity utilization and routing choices. We do not directly observe the capacity of tracks and yards, so we use the largest

observed monthly flow through each track or yard as a measure of capacity.[22] 80 percent of tracks and yards, representing representing 95% of total traffic, reached their maximum observed flows before 2006.

We define a track or yard's capacity utilization as the ratio of observed monthly traffic to capacity, and we define a route's capacity utilization as the average track utilization for tracks and yards belonging to that route. For each origin-destination pair, we define its "main route" as the route most commonly used to service that pair.

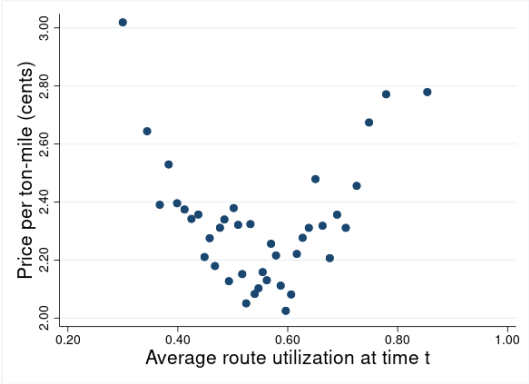
In Panel B of Figure 4, we plot the number of unique routes used to service a city pair against the capacity utilization of its main route. We find that firms use their preferred route almost exclusively when its utilization is below 40%. However, as utilization for the main route increases, they become 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 potentially depending on the capacity utilization of their infrastructure.

4.3 Capacity utilization is correlated with prices

Our final fact relates to the relationship between capacity utilization and prices. We focus on the prices for spot market shipments, which may be more likely to reflect contemporaneous market conditions than contract market shipments. Figure 5 plots a binned scatter plot of these prices against the capacity utilization of the route used to fulfill each shipment. We find a U-shape relationship between spot market prices and capacity utilization: Prices decrease with capacity utilization until 50 percent utilization, and increase thereafter. In terms of magnitude, minimum prices are 33 percent lower than prices at lower utilization. While this empirical relationship may reflect a combination of varying markups and costs across space, it motivates allowing a flexible cost function that can capture both increasing and decreasing regions.

[22] Formally, we define the capacity of a track segment from u to v as $K_{uv} = 1.1 \times \max_t F_{uv,t}$ and the capacity of a yard as $K_n = 1.1 \times \max_t F_{n,t}$, where $F_{uv,t}$ and $F_{n,t}$ are traffic through a track segment or yard in month t.

Figure 5: *Spot market prices and capacity utilization.*



Note: This figure plots spot market prices against the monthly average 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. Source: Confidential Waybill Dataset

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 the basis of our estimation approach.

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.[23] The firm's problem is:

$$\min_{\mathbf{X}} C(\mathbf{X}) \tag{1}$$

$$s.t. \quad 0 \leq X_{od,it} \; \forall i, od \tag{2}$$

$$\sum_{i=1}^{N_{od}} X_{od,it} = Q_{od,t} \; \forall od \tag{3}$$

Constraint (2) requires the firm to ship a nonnegative quantity along each route, and constraint (3) ensures that the quantities shipped across all routes are sufficient to fulfill demand.

For a given routing choice, edge traffic $F_{uv,t}$ sums routing choices over all routes which use edge uv,

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

[23] 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 A, 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. For example, this condition is satisfied by logit demand as used in Chen (2023).

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

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

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

$$\frac{\partial C(F(\mathbf{X}))}{\partial X_{od,i}} = \frac{\partial C(F(\mathbf{X}))}{\partial X_{od,j}} \tag{4}$$

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,

$$C(F(\mathbf{X})) = \sum_{uv} c_{uv}(F_{uv,t}(\mathbf{X}), F_{vu,t}(\mathbf{X})) + \sum_{n} c_n(F_{n,t}(\mathbf{X})) \tag{5}$$

where the edge cost c_{uv} is a function of flows through edge uv in both directions in month t, and the node cost c_n is a function of the flows through node n in month t.[24]

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

[24] 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.

cost function with respect to shipping along a single route can be decomposed into

$$\frac{\partial C(F(\mathbf{X}))}{\partial X_{od,it}} = \underbrace{\sum_{uv \in i} \frac{\partial c_{uv}}{\partial F_{uv}}}_{\text{Edge flows}} + \underbrace{\sum_{vu \in i} \frac{\partial c_{uv}}{\partial F_{vu}}}_{\text{Reverse-edge flows}} + \underbrace{\sum_{n \in i} \frac{\partial c_n}{\partial F_n}}_{\text{Node flows}} \tag{6}$$

Equation (6) 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 (6) can be compactly rewritten as follows:

$$\frac{\partial C(F(\mathbf{X}))}{\partial X_{od,it}} = R'^{iE} \frac{\partial \vec{c}_{uv}}{\partial F_{uv}} + \tilde{R}'^{iE} \frac{\partial \vec{c}_{uv}}{\partial F_{vu}} + R'^{iN} \frac{\partial \vec{c}_n}{\partial F_n} \tag{7}$$

Combining equations (7) and (4), we obtain a new first order condition:

$$\left[R^{i,E} - R^{j,E}\right]' \frac{\partial \vec{c}_{uv}}{\partial F_{uv}} + \left[\tilde{R}^{i,E} - \tilde{R}^{j,E}\right]' \frac{\partial \vec{c}_{uv}}{\partial F_{vu}} + \left[R^{i,N} - R^{j,N}\right]' \frac{\partial \vec{c}_n}{\partial F_n} = 0 \tag{8}$$

The next section develops an empirical analogue of equation 8 for estimation.

6 Estimation

We estimate the cost function using a revealed preferences argument based on cost minimization. Intuitively, 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.

In this section, we detail the assumptions necessary to go from this observation

to a set of moment conditions we can take to the data.

6.1 Econometric Model

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

$$\partial c_{uv}/\partial F_{uv} = mc_{1,uv}(F_{uv}, F_{vu}; Y_{uv}, \theta) + \varepsilon_{1,uv,t} \tag{9}$$

$$\partial c_{uv}/\partial F_{vu} = mc_{2,uv}(F_{uv}, F_{vu}; Y_{uv}, \theta) + \varepsilon_{2,uv,t} \tag{10}$$

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

The edge- and node-level error terms ε account for measurement error in the observable characteristics (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.

We can construct a route-pair-level equality condition based on the firm's first order condition. Let an arrow superscript denote the vector stacking the corresponding function for all edges or nodes. Replacing equations (9) through (11) into the first order condition (8) and dropping the Y argument, we have:

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

6.2 Parametrization

To take our moment condition to the data, we need to make a parametric assumption which approximates the unknown true function *mc*. Our parametric assumptions about *mc* have four objectives. First, we want to be flexible, to allow for either increasing or decreasing returns to scale at different levels of capacity utilization. Second, we want to be internally consistent: since mc_1 and mc_2 are derivatives of

the same cost function, assumptions about one imply restrictions about the other. Third, we want costs to be robust to our definition of an edge. 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. Fourth, we want our approximation to have an analytical integral, for ease of computing counterfactual simulations.

6.2.1 Edge marginal costs: mc_1 and mc_2

Same Direction Costs We start by parameterizing mc_1, the same-direction marginal cost, as an expression which is linear in edge length:

$$mc_{1,uv} = d_{uv}(\alpha_d + \alpha_Y Y_{uv} + g(F_{uv,t}, F_{vu,t} \mid K)) \tag{12}$$

where d_{uv} is the length (in miles) of the edge, Y_{uv} is a vector of track characteristics[25], and g is a function of edge flows, reverse edge flows, and capacity K. Under this expression, if characteristics are constant along an edge, then any division of that edge into two edges will have the same total marginal cost.

We approximate g using a quadratic polynomial on F_{uv} and F_{vu},[26],

$$\begin{aligned} g(F_{uv,t}, F_{vu,t} \mid K) &= \beta_1 \frac{F_{uv,t}}{K_{uv}} + \beta_2 \frac{F_{vu,t}}{K_{vu}} + \gamma_1 (\frac{F_{uv,t}}{K_{uv}})^2 + \gamma_2 (\frac{F_{vu,t}}{K_{vu}})^2 + \gamma_3 \frac{F_{uv,t}}{K_{uv}} \frac{F_{vu,t}}{K_{vu}} \\ &= [\beta_1 \ \beta_2 \ \gamma_1 \ \gamma_2 \ \gamma_3]' X_{uv,t} \end{aligned} \tag{13}$$

where $X_{uv,t}$ is a vector of polynomials in F_{uv}/K_{uv} and F_{vu}/K_{vu}.

In our model, the cost coefficients are identified up to a scale parameter. We normalize the coefficient of distance α_d to be one, so all costs are expressed in effective-mile units.

[25] Characteristics include the number of parallel tracks, the frequency of passing sidings, the slope of the terrain, whether the firm owns or rents the infrastructure. We also include state fixed effects.

[26] Continuous differentiability significantly increases the tractability of computing counterfactuals. Our estimation results are robust to approximating g with a step function or adding higher order polynomial terms; see Appendix F for details.

Opposite Direction Costs Assuming that F_{vu} only affects c_{uv} through mc_1, Equation 13 implies that

$$mc_{2,uv} = d_{uv}\left(\beta_2 \frac{F_{uv,t}}{K_{vu}} + \gamma_2 \frac{F_{uv,t}F_{vu,t}}{K_{vu}^2} + \gamma_3 \frac{F_{uv,t}^2}{K_{uv}K_{vu}}\right) = d_{uv}[\beta_2 \ \gamma_2 \ \gamma_3]' X_{2,uv,t}$$

where $X_{2,uv,t}$ is a vector of polynomials.

6.2.2 Node marginal costs

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.[27] Conditional on a node corresponding to a yard, we let node marginal costs, mc_n, be a quadratic function of node capacity utilization.

$$mc_n = \left(\alpha_n + \beta_n \frac{F_{n,t}}{K_n} + \gamma_n (\frac{F_{n,t}}{K_n})^2\right) = \alpha_n + [\beta_n \ \gamma_n]' X_{n,t} \tag{14}$$

6.3 Estimating Equation

Given our parametric assumptions, we want to write down a linear estimating equation. First, we define three operators which take the difference of a edge and node characteristics between routes i and j:

$$\begin{aligned} \Delta_{ij} x &= \sum_{uv} (R_{uv}^i - R_{uv}^j) x_{uv} \\ \tilde{\Delta}_{ij} x &= \sum_{uv} (\tilde{R}_{uv}^i - \tilde{R}_{uv}^j) x_{uv} \\ \Delta_n^N x &= \sum_{n} (R_n^i - R_n^j) x_n \end{aligned} \tag{15}$$

[27] 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.

Given these operators, our route-pair-level equality condition becomes:

$$\begin{aligned} \underbrace{-\Delta_{ij} d}_{\text{Net Distance}} &= \underbrace{\alpha_Y \times \Delta_{ij}(d \cdot Y)}_{\text{Net Track Char.}} + \underbrace{\theta \times \Delta_{ij}(d \cdot X) + \tilde{\theta} \times \tilde{\Delta}_{ij}(d \cdot X_2)}_{\text{Net Edge Util}} \\ &+ \underbrace{\theta^N \times \Delta^N_{ij} X_n}_{\text{Net Yard Util}} + \underbrace{\Delta_{ij}\varepsilon_1 + \tilde{\Delta}_{ij}\varepsilon_2 + \Delta^N_{ij}\varepsilon_n}_{\text{Error}} \end{aligned} \tag{16}$$

where $\theta = [\beta_1\ \beta_2\ \gamma_1\ \gamma_2\ \gamma_3]$, $\tilde{\theta} = [\beta_2\ \gamma_2\ \gamma_3]$, and $\theta^N = [\beta_n\ \gamma_n]$.

This equation states, for two routes which are both used, differences in distance (left-hand side) must be compensated by differences in (distance-weighted) route characteristics, (distance-weighted) edge utilization, or node utilization.

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. As a result, an OLS estimate may be biased.

To address the potential endogeneity, we use a route-pair-level instruments $Z = (Z_{ij}, \tilde{Z}_{ij}, Z^N_{ij})$ which affects net utilization independently of the error. We discuss our instrument construction in detail below in Section 6.4.We estimate θ using 2SLS with a nonparametric first stage, following Chen et al. (2020). First, we use a random forest to predict our endogenous variables ($D = (\Delta_{ij} dX, \tilde{\Delta}_{ij} dX_2, \Delta^N_{ij} X_n)$) using Z. Second, we construct the best partially linear predictor

$$\hat{D}^*(Z) = E[D_{ij}|Z_{ij}] + E^*\left[D_{ij}|\Delta_{ij} dY - E[\Delta_{ij} d|Z_{ij}]\right] \tag{17}$$

where E^* is the Best Linear Predictor, and use it to form our estimator

$$\hat{\theta} = \left(\frac{1}{NT}\sum_{ijt} \hat{D}^*(Z_{ijt}) T'_{ijt}\right)^{-1} \frac{1}{NT}\sum_{ijt} \hat{D}^*(Z_{ijt}) \Delta_{ij} dY \tag{18}$$

6.4 Instrument

Since we are estimating a cost function, the natural source of exogenous variation would be a demand shock that traces the cost function. The magnitude of this

demand shock should differ from edge to edge in a way that is uncorrelated with the errors, allowing us to instrument for flows F using some set 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 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.

Rail plays a crucial role in transporting coal from mines to power plants across the US; in 2021, 69 percent of coal shipments were transported by rail, and coal occupied the most volume of any commodity transported by rail. Further, monthly coal production declined by 35 percent from 2009 to 2018.[28] The decline in coal production therefore represents a substantial demand shock for the rail industry.

We use this demand shock to construct a shift-share instrument for rail flows at the edge and node level. As our "shifter", we use total monthly U.S. coal production for states east of the Mississippi River.[29] We construct a "share" by computing the share of all coal shipments shipped between each node pair from 1998 to 2003 (before our estimation sample period). We construct the shift-share variable at the

[28] 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 (2023)). 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.

[29] Appendix E presents a time series figure of the shifter variable. Using national U.S. coal production could be problematic since both railroads studied in this paper (BNSF and UP) service coal mines in the West, potentially generating reverse causality as higher freight prices to contribute to the decline in coal production. Another concern might be that the availability of cheaper shale oil led to both lower diesel costs for rail and lower coal production. However, diesel prices are highly spatially correlated, with a 97.8 percent correlation across U.S. regions in the 2015 to 2018 period. Any correlation between diesel costs and shale expansion therefore should not have a differential effect across edges. Since we normalize the overall level of costs, such a correlation would not affect our estimates.

month-city-pair level by multiplying the shifter and the shares:

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

Finally, we convert this variable into an edge- and node-level exposure measure by computing the shortest route between o and d and allocating each city pair's exposure to all the nodes and edges belonging to that shortest route.[30]

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

The instrument tracks the endogenous variable closely, and our first-stage has an R^2 of 78 percent.[31]

6.5 Comments on the estimation

Identification Our estimation of the cost function relies on the firm equalizing marginal costs across pairs of routes. This approach can identify the shape of the cost function up to an affine transformation, but cannot identify fixed costs, or any costs which are invariant to the choice of route (e.g., the non-variable component of wages). To fix scale, we normalize the coefficient of distance on costs to one and interpret all costs in terms of mile equivalent units.

For exposition, we convert miles into dollar units through an accounting-based approach. We start with the fuel efficiency of a locomotive, which is around 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 (2023)). Altogether, this suggests that fuel costs are \$0.78 per car-mile. The Association of American Railroads reports that fuel represents around 20 percent of operating costs, so we convert mile equivalent units into dollars at a rate of \$3.89 per mile.

[30] 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.

[31] Appendix E contains additional details on the first-stage nonparametric regression.

Routes Not Taken In our analysis, we focus on cases where a 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.

Duration We cannot directly account for the duration of any trip, since we do not observe this in our data. We may be concerned about this if shippers have preferences over duration, leading our cost minimization assumptions to fail. In Appendix A, we derive conditions under which our model can be nested within a broader model that allows for shipper preferences over route characteristics. We show that in these cases, the marginal costs which we estimate are a mix of physical costs and the effect of duration on demand.

A significant portion of variation in shipper preferences over duration may be explained by commodity type. For example, shippers of high-value commodities may prefer faster routes than shippers of low-value commodities, violating our assumption that shippers are indifferent over routes. In Appendix A.2, we estimate our baseline model on the set of route-pairs which we observed carried the same commodity in the same month, and we find that our results are robust to this restriction.

6.6 Results

Estimation Results We use Generalized Method of Moments to estimate the coefficients of the marginal cost functions described in Section 6.2, using as our sample all pairs of routes with positive traffic between 2015 and 2018.[32] We estimate separate coefficients for both BNSF and UP, allowing them to have different technologies. We present 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. For the remaining sections, we use specifications (2) and (4). These estimates inform us regarding the shape of the cost function

[32] As documented in Appendix C, the data quality of routing information is significantly higher quality beginning in 2015.

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 percent for BNSF and 61 percent for UP, reducing the slope by 1 percent reduces edge-costs by 30 percent, and increasing the frequency of passing sidings reduces costs by 48-65 percent. The one notable difference between the two firms is the coefficient of the trackage dummy, which indicates 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 percent capacity utilization for BNSF and 40 percent 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 6.[34] We convert the units from effective miles to dollars per carload-mile using the calculation described in Section 6.5. 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 percent utilization, and imply significant economies of scale: Traversing an edge with 0 percent utilization is $2.00 more expensive than a shipment at 55 percent utilization; this difference represents 40 percent of the average price for a carload-mile during this period. These results are consistent with the behavior of prices in Figure 5, even though pricing data was not used in estimation.

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.

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

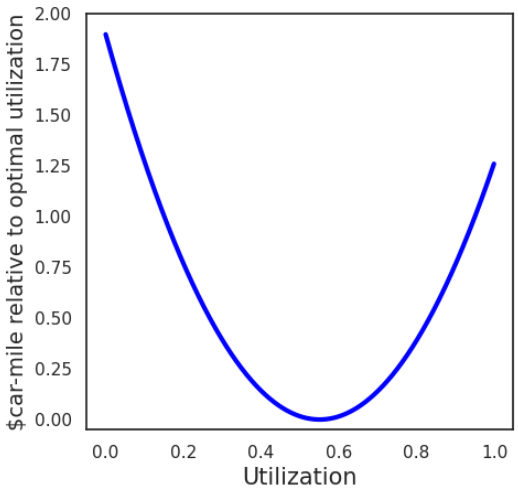
[34] We show the same plot for Union Pacific in Appendix D.

Table 1: *Estimation Results*

	BNSF			UP		
	(1) OLS	(2) IV	(3) IV	(4) OLS	(5) IV	(6) IV
Characteristics						
Additional 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 (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			-0.651			-0.480
			[-24.85]			[-8.15]
Track Costs						
Utilization_{ij}	-1.348	-1.533	-1.120	-0.768	-1.129	-0.287
	[-39.77]	[-37.61]	[-24.32]	[-19.34]	[-7.63]	[-3.96]
$\text{Utilization}_{ij}^2$	1.139	1.256	0.918	0.726	1.117	0.188
	[29.12]	[27.52]	[22.51]	[17.17]	[7.34]	[2.45]
Utilization_{ji}	-0.198	-0.334	-0.179	-0.106	-0.132	-0.139
	[-6.81]	[-14.05]	[-7.17]	[-10.72]	[-8.25]	[-7.15]
$\text{Utilization}_{ji}^2$	-0.051	-0.089	0.012	-0.010	-0.034	-0.076
	[-2.83]	[-5.20]	[0.85]	[-0.55]	[-1.41]	[-2.75]
$\text{Utilization}_{ij} \times \text{Utilization}_{ji}$	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 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]
Utilization_n	96.712	157.709	61.634	52.762	83.964	-14.322
	[12.25]	[12.05]	[7.28]	[10.59]	[2.79]	[-2.24]
Utilization_n^2	-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.

Figure 6: *Estimated marginal costs at the edge level (BNSF)*



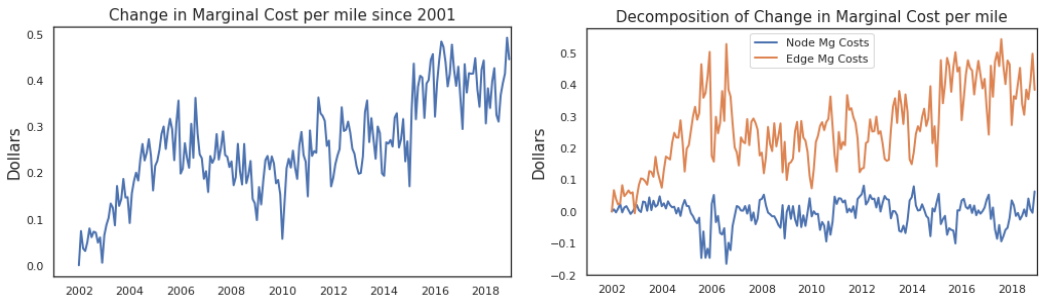
Note: This figure plots the implied edge-level marginal costs for BNSF as a function of same-direction utilization. Since the cost also depends on reverse-direction utilization, we assume that utilization is symmetric when producing this figure. Source: Confidential Waybill Dataset

Marginal costs over time We estimate our model using data from 2015 to 2018. Under the assumption that the cost function does not vary over time, we can estimate marginal costs by route for the rest of our data. Since we estimate separate cost functions for BNSF and UP, we construct an industry average marginal cost by averaging over routes, weighting each route by its observed traffic. Panel A of Figure 7 plots the estimated change in average marginal costs 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 percent of the total increase. Appendix D shows how costs varied for each of the two firms individually.

Panel B of Figure 7 decomposes the time series presented in Panel A 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

Figure 7: *Panel A: Estimated marginal cost per mile 2002-2018. Panel B: Decomposition of estimated marginal costs per mile*



Note: Panel A plots the implied change in average marginal costs due to changes in track and yard utilization. The cost in 2002 is normalized to 0. Panel B decomposes the change in average marginal costs into changes in edge and node marginal costs. Source: Confidential Waybill Dataset

period. [35]

7 Counterfactual Analyses

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.

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 as well as state and local government, often enter into partnerships with one or multiple railroads to jointly finance specific investments.[36]

[35] One interpretation of the time trend in costs could reflect increased use of unit trains. Starting in the early 2000s, railroads have tried to increase the efficiency of their shipments by avoiding using yards and instead assembling trains where all carloads share the same origin and destination. Such a technological transition could reduce the importance of yards as sites of congestion.

[36] At the federal level, the Infrastructure Investment and Jobs Act of 2021 allocated $22 billion

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 investments?

Returns to investment differ across states We start by evaluating how returns on infrastructure investments vary across states.[37] 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.[38] 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

dollars to finance investment in rail and intermodal facilities. At the state and local levels, two 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. 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.

[37] 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.

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

firm using the new edge-level costs.[39]

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 $\mathbf{X}_s^{CF}$.[40] Next, we compute how total costs change in response to the changing infrastructure:

$$\Delta C_s = C_s^{UP,CF}(\mathbf{X}_s^{UP,CF}) + C_s^{BNSF,CF}(\mathbf{X}_s^{BNSF,CF}) - C_s^{UP,B}(\mathbf{X}_s^{UP,B}) - C_s^{BNSF,B}(\mathbf{X}_s^{BNSF,B}) \tag{19}$$

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

$$\text{Return on Investment}_s = \frac{\Delta C_s}{\text{Miles of track}_s} \tag{20}$$

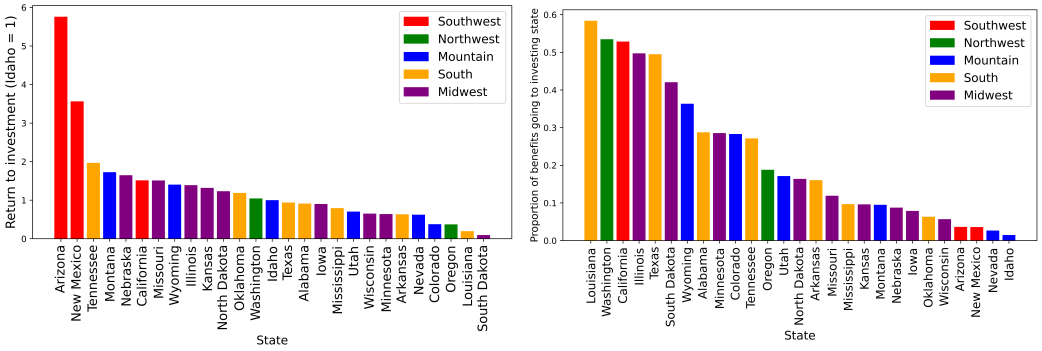
We normalize the returns to one for the median state in our sample (Idaho), and plot this measure for all states in Figure 8. As is evidenced by the Figure, returns on investment vary significantly across states. The state with the most productive investment (Arizona) has a return six times higher than that of the median state (Idaho), which in turn is five times as productive as the least productive state (South Dakota). Returns on investment are higher in more central regions, and the states with the highest returns are all part of the largest rail corridor, which connects Los Angeles and Chicago.

Investment Externalities Since improved infrastructure benefits both traffic that originates in a state, as well as traffic that passes through a state, state governments may not internalize all of the cost savings generated by local investments. In Panel B of Figure 8, we plot the share of the return to investment for each state which is captured by local shippers (i.e., the investing state is the origin or destination of the shipment). The median state captures around 20 percent of the returns to investment, but as in the previous figure, there is wide dispersion across states. In particular, the Southwestern states which generate the most returns to investment capture a

[39] Appendix B provides computational details.

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

Figure 8: *Return to Investment in rail infrastructure.*



Note: Panel A plots the change in costs in response to adding one additional set of parallel tracks to all edges in a state. The change in costs for the median state, Idaho, is normalized to 1. Panel B plots share of benefits (cost savings) captured by shipments which originate or terminate in the treated state, weighted by the number of shipments. Source: Confidential Waybill Dataset.

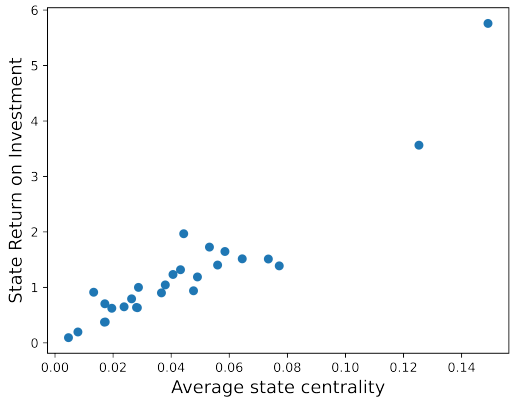
very small share of those benefits. The large size of these externalities suggests that state-level investments are not efficient, and national coordination can be beneficial.

Network Structure The negative correlation between the returns to investment and the share of returns captured can be explained by economies of scope on the rail network. Intuitively, central pieces of infrastructure have high traffic volumes, which increases the return to investment. However, a large share of that traffic tends to be through-traffic, connecting other origins and destinations, which lowers the share of returns captured. More formally, let Q denote the total flows through the network, $Q_{n\cdot}$ denote the total flows originating at node n, and $Q_{\cdot n}$ the total number of flows terminating at node n. The betweenness centrality of a node can be expressed as $(F_n - Q_{\cdot n} - Q_{n\cdot})/(Q - Q_{\cdot n} - Q_{n\cdot})$.

Panel A of Figure 9 presents a scatter plot of the return on investment on state centrality, indicating that more central states have higher returns to investment. Overall, centrality explains variation in returns on investment well.

Panel B presents results from a regression of the proportion of benefits captured on centrality and total state flows. A state is more likely to capture more of the benefits coming from infrastructure investment the higher the amount of flows that

Figure 9: *Centrality and Return on Investment*



	Proportion captured
Centrality Measure	-1.583
	(0.845)
State Flows	0.387
	(0.095)
Constant	0.176
	(0.052)
R-squared	0.438
R-squared Adj.	0.391
N	27
R2	0.438

Note: Panel A presents a scatter plot of the share of benefits captured by the investing state against the state's centrality measure as defined in the text. Panel B presents coefficients from a state-level regression of the proportion of benefits captured on the centrality measure and total inbound and outbound shipments. Source: Confidential Waybill Dataset.

originate or terminate in that state. This is intuitive, since our model assigns the benefits from lower infrastructure costs to the areas that originate or terminate flows. A higher centrality measure, on the other hand, leads to a lower ability to capture the benefits from higher investment. A central state is one which many others use for shipping products; as a consequence, investments there will benefit other states, leading to a reduced ability to capture the returns on investment.

8 Conclusion

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 non-monotonic 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 how returns on investment vary more across states and the effect of economies of scope on incentives for public investment.

References

Allen, Treb and Costas Arkolakis, "The Welfare Effects of Transportation Infrastructure Improvements [General Equilibrium Effects in Space: Theory and Measurement]," *Review of Economic Studies*, 2022, *89* (6), 2911–2957.

Argente, David, Sara Moreira, Ezra Oberfield, and Venky Venkateswaran, "Scalable Expertise," 2020.

Armstrong, J.H., *The Railroad: What it Is, what it Does*, Simmons-Boardman Books, 2008.

Bailey, Samuel, "Competition and Coordination in Infrastructure: Port Authorities' Response to the Panama Canal Expansion," 2021. Working paper.

Bernard, Andrew B., Stephen J. Redding, and Peter K. Schott, "Multiple-Product Firms and Product Switching," *American Economic Review*, March 2010, *100* (1), 70–97.

Borts, G. H., "Production Relations in the Railway Industry," *Econometrica*, 1952, *20*, 71–79.

Borts, George H., "Increasing Returns in the Railway Industry," *Journal of Political Economy*, 1954, *62*, 316–316.

_ , "The Estimation of Rail Cost Functions," *Econometrica*, 1960, *28*, 108.

Braeutigam, Ronald R, Andrew F Daughety, and Mark A Turnquist, "The Estimation of a Hybrid Cost Function for a Railroad Firm," *The Review of Economics and Statistics*, August 1982, *64* (3), 394–404.

Brancaccio, Giulia, Myrto Kalouptsidi, and Theodore Papageorgiou, "Geography, Transportation, and Endogenous Trade Costs," *Econometrica*, March 2020, *88* (2), 657–691.

Brock, William A., "Contestable Markets and the Theory of Industry Structure: A Review Article," *Journal of Political Economy*, 1983, *91* (6), 1055–1066.

Caris, An, Sabine Limbourg, Cathy Macharis, Tom van Lier, and Mario Cools, "Integration of inland waterway transport in the intermodal supply chain: a taxonomy of research challenges," *Journal of Transport Geography*, 2014, *41* (C), 126–136.

Chen, Jiafeng, Daniel L. Chen, and Greg Lewis, "Mostly Harmless Machine Learning: Learning Optimal Instruments in Linear IV Models," Papers 2011.06158, arXiv.org November 2020.

Chen, Yanyou, "Network Structure and Efficiency Gains from Mergers: Evidence from the U.S. Freight Railroads," June 2023. Working paper.

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

Dhyne, Emmanuel, Amil Petrin, Valerie Smeets, and Frederic Warzynski, "Theory for Extending Single-Product Production Function Estimation to Multi-Product Settings," NBER Working Papers 30784, National Bureau of Economic Research, Inc December 2022.

Ding, Xiang, "Industry linkages from joint production," 2022. Working Paper.

Ducruet, César, Reka Juhasz, Dávid Krisztián, and Claudia Steinwender, "All aboard: the effects of port development," LSE Research Online Documents on Economics 108496, London School of Economics and Political Science, LSE Library December 2020.

Fajgelbaum, Pablo D. and Edouard Schaal, "Optimal Transport Networks in Spatial Equilibrium," *Econometrica*, July 2020, *88* (4), 1411–1452.

Feng, Bo, Yanzhi Li, and Zou-Jun Max Shen, "Air cargo operations: Literature review and comparison with practices," *Transportation Research Part C: Emerging Technologies*, July 2015, *56*, 263–280.

Fuchs, Simon and Woan Foong Wong, "Multimodal Transport Networks," FRB Atlanta Working Paper 2022-13, Federal Reserve Bank of Atlanta October 2022.

Ganapati, Sharat, Woan Foong Wong, and Oren Ziv, "Entrepôt: Hubs, Scale, and Trade Costs," NBER Working Papers 29015, National Bureau of Economic Research, Inc July 2021.

Goldberg, Pinelopi K, Amit K Khandelwal, Nina Pavcnik, and Petia Topalova, "Multiproduct Firms and Product Turnover in the Developing World: Evidence from India," *The Review of Economics and Statistics*, November 2010, *92* (4), 1042–1049.

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.

Griliches, Zvi, "Cost Allocation in Railroad Regulation," *Bell Journal of Economics*, Spring 1972, *3* (1), 26–41.

Hall, Robert E., "The Specification of Technology with Several Kinds of Output," *Journal of Political Economy*, 1973, *81* (4), 878–892.

Huntley, Christopher L., Donald E. Brown, David E. Sappington, and Bernard P. Markowicz, "Freight Routing and Scheduling at CSX Transportation," *Interfaces*, June 1995, *25* (3), 58–71.

Jara-Díaz, Sergio R. and Cristian E. Cortés, "On the Calculation of Scale Economies from Transport Cost Functions," *Journal of Transport Economics and Policy*, 1996, *30* (2), 157–170.

Johnes, Geraint, "Costs and Industrial Structure in Contemporary British Higher Education," *The Economic Journal*, 01 2012, *107* (442), 727–737.

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

Keeler, Theodore E, "Railroad Costs, Returns to Scale, and Excess Capacity," *The Review of Economics and Statistics*, May 1974, *56* (2), 201–208.

Khmelnitskaya, Ekaterina, Guillermo Marshall, and Scott Orr, "Identifying Scope Economies Using Demand-Side Data," 2023.

Kohli, Ulrich R, "Nonjointness and Factor Intensity in U.S. Production," *International Economic Review*, February 1981, *22* (1), 3–18.

Lai, Yung-Cheng (Rex) and Christopher P. L. Barkan, "Enhanced Parametric Railway Capacity Evaluation Tool," *Transportation Research Record*, 2009, *2117* (1), 33–40.

Lorenz, M. O., "Cost and Value of Service in Railroad Rate-Making," *The Quarterly Journal of Economics*, 1916, *30* (2), 205–232.

Lu, Binbin, Huabo Sun, Miaozhong Xu, Paul Harris, and Martin Charlton, "Shp2graph: Tools to Convert a Spatial Network into an Igraph Graph in R," *ISPRS International Journal of Geo-Information*, 2018, *7* (8), 293.

Maican, Florin and Matilda Orth, "Determinants of economies of scope in retail," *International Journal of Industrial Organization*, 2021, *75* (C).

Meyer, John Robert, Herbert D Mohring, Merton J. Peck, John Stenason, and Charles J. Zwick, "The Economics of Competition in the Transportation Industries.," *Journal of the American Statistical Association*, 1961, *56*, 458.

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

_ and _ , "Economies of scope," *The American Economic Review*, 1981, *71* (2), 268–272.

Redding, Stephen J. and Matthew A. Turner, "Chapter 20 - Transportation Costs and the Spatial Organization of Economic Activity," in Gilles Duranton, J. Vernon Henderson, and William C. Strange, eds., *Handbook of Regional and Urban Economics*, Vol. 5 of *Handbook of Regional and Urban Economics*, Elsevier, 2015, pp. 1339–1398.

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

Teece, David J., "Economies of scope and the scope of the enterprise," *Journal of Economic Behavior & Organization*, 1980, *1* (3), 223–247.

US Energy Information Administration, "Today in Energy," `https://www.eia.gov/todayinenergy/detail.php?id=48896` 2023. Accessed: 07-27-2023.

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

Williams, Kevin R., "The Welfare Effects of Dynamic Pricing: Evidence From Airline Markets," *Econometrica*, March 2022, *90* (2), 831–858.

Zhang, Jingfang and Emir Malikov, "Off-balance sheet activities and scope economies in U.S. banking," *Journal of Banking & Finance*, 2022, *141*, 106534.

Online Appendix

A Shipper Preferences over Routes

Our analysis focuses on the firm's cost minimization problem given a fixed quantity of shipments to deliver. This baseline approach implicitly assumes that the routes chosen do not affect prices or quantity, which may be problematic if shippers have preferences over route characteristics (e.g., travel time). This section discusses several robustness checks exploring the potential role of route characteristics.

A.1 Routes and Prices

If shippers have preferences over routes, then routing choices should have predictive power over prices. For example, if shippers prefer shorter routes that are faster, then we might expect shorter routes to predict higher prices than longer routes.

We formalize this intuition by training a random forest to predict prices given all our observables excluding routing choices (distance and states visited) and comparing the predictions to those of a second random forest where we include the routing choices as explanatory variables.[41] Evaluated on a sample split into training and test sets, including routing choices in the random forest does not significantly improve the in-sample prediction accuracy, and lowers the out-of-sample R^2 from 90.35 percent to 89.57 percent.

A.2 Same-Commodity Pairs

If shippers of different commodities have heterogeneous preferences over routes, then using pairs of routes that serve different commodities could bias our estimation. For example, if container shippers prefer shorter routes more than coal shippers, then our indifference condition would not hold for comparing a longer route used to ship coal and a shorter route used to ship containers.

[41] As baseline observables, we include origin and destination fixed effects, commodity fixed effects, time dummies, weight, and the shortest feasible distance between origin and destination. We include states visited as a set of dummies for each unique combination of states visited.

We therefore perform a robustness check where we condition our estimation on only using pairs of routes that shipped the same commodity in the same month.[42] These pairwise comparisons would be valid even if shippers of different commodities have different preferences over routes (e.g., over route duration). Restricting the sample in this way drops about 10 percent of our sample.

Table 2 presents estimation results from using our baseline model on the same-commodity sample. The coefficients are similar to those from our baseline specification. We conclude that shipper preferences over duration or other unobserved route characteristics are driving our results.

A.3 Profit Maximization

In this section, we nest our model within a profit maximization framework and derive sufficient conditions for our analysis to be robust to shippers' route preferences.

Suppose that a firm faces an inverse demand function, $P(Q_t, \mathbf{X})$, which depends on the quantities shipped, Q_t, and the routes chosen to implement them, $\mathbf{X}$. The firm chooses quantities and routes to maximize profits:

$$\max_{Q_t, \mathbf{X}} \{P(Q_t, \mathbf{X})Q_t - C(\mathbf{X})\} \tag{21}$$

subject to equations (2) and (3).

If the inverse demand function P is separable in Q_t and X, $P(Q_t, X) = f(Q_t) + g(X)$, the profit maximization problem can be rewritten as

$$\max_{Q_t} \left\{ f(Q_t)Q_t - \min_{\mathbf{X}} \underbrace{\{C(\mathbf{X}) - g(\mathbf{X})Q_t\}}_{\tilde{C}(\mathbf{X})} \right\} \tag{22}$$

implying that the chosen routes minimize an augmented cost function,

$$X^* = \arg\max_X C(X) - G(X)Q_t^* = \arg\max_X \tilde{C}(X) \tag{23}$$

[42] We measure commodity type using 2-digit Standard Transportation Commodity Codes (STCC), a standard commodity classification system used in the rail industry. The full codes are seven digits, where the first two digits denote one of 36 Major Industry Groups.

Table 2: *Same Commodity Estimation Results*

	(1)
Utilization	-1.119
	(0.046)
Util. Squared	0.901
	(0.050)
Util. times utilization opp. dir.	0.219
	(0.044)
Util. opp. dir.	-0.195
	(0.032)
Util. opp. dir. squared	0.010
	(0.016)
Node-costs constant	-19.538
	(2.880)
Node-costs slope	66.826
	(10.224)
Node-costs quadratic	-55.858
	(8.501)
Trackage Dummy	-0.395
	(0.010)
Slope	36.092
	(1.225)
Number of Tracks	-0.142
	(0.009)
Passing lanes	-0.642
	(0.047)
State FE	Yes
Month FE	Yes
Observations	71246

Standard errors in parentheses

Our estimates can therefore be interpreted as recovering the augmented cost function, $\tilde{C}(\mathbf{X})$, which combines the true cost function and the effect of the routes chosen on demand.

We develop two examples of demand functions which have inverse demand functions that satisfy separability in prices and routes.

Linear Demand If demand is a linear function of prices and route characteristics, then inverse demand is separable in prices and characteristics.

$$D(P,\mathbf{X}) = \alpha P + g(\mathbf{X}) \implies P(Q,\mathbf{X}) = \alpha^{-1}Q - \alpha^{-1}g(\mathbf{X}) \tag{24}$$

Logit Demand Suppose that there is a continuum of shippers for each origin-destination pair who choose between rail shipping and an outside option. Each shipper has utility $\alpha P + g(X) + \varepsilon$ from choosing rail, and utility ε_{oo} from choosing the outside option, where ε and ε_{oo} are distributed i.i.d. from a Type I extreme value distribution.

Demand for rail takes on the form

$$D(P,X) = \frac{\exp(\alpha P + g(X))}{1 + \exp(\alpha P + g(X))} \tag{25}$$

which implies that inverse demand is separable in prices and characteristics.

$$P = -\alpha^{-1}\log(Q^{-1} - 1) - \alpha^{-1}g(X) \tag{26}$$

B Computational Details

Solving the firm's minimization problem is computationally challenging, so we make additional assumptions for tractability when computing counterfactual routes. First, we restrict each firm to choose routes from the set of historical routes with non-negligible traffic. Specifically, we exclude routes which carried less than 2.5 percent of shipments for their origin-destination pair over our full sample period (1996 to 2018). This filter leaves around 6,000 routes to optimize over for each firm.

Second, the optimization problem of the two firms are interconnected due to trackage rights: UP can choose to route part of its shipments using BNSF's network and vice versa. We model this by solving for a Nash Equilibrium in route choices, where firm i's route choices are a best response taking firm j's choices as given. We start with a guess for each firm's choices, and then iterate best responses until convergence. We solve the minimization problem using SQP implemented by KNITRO 13.2.0, trying multiple initial values and choosing the solution with the lowest costs.[43]

C Routing data sample break

Our approach relies on reconstructing routes using reported distances and states visited in waybills. In Figure A.1, we plot the number of origin-destination pairs in each year where a single firm reported two or more waybills with different distances or lists of states visited.[44] This number is fairly consistent from 2001 to 2014, but it has a clear break in 2015. We believe this break reflects changes in Railinc's methodology which improved the imputed distances for each waybill. To keep our analysis consistent, we restrict our attention to the 2015-2018 period.

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.

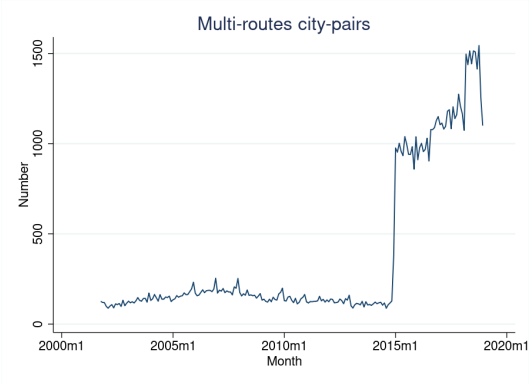
D Additional Estimation Results

In this section, we present additional figures that summarize our estimation results. In our estimation section, we reported the variation in marginal costs as capacity utilization rises from 0 percent to 100 percent for BNSF. In Figure A.2, we present the parallel cost curve for UP. The estimated cost function for UP has a similar shape to BNSF's cost function: marginal costs are convex in utilization with a minimum

[43] We used two classes of starting values: historical route allocations and equal allocations across all routes.

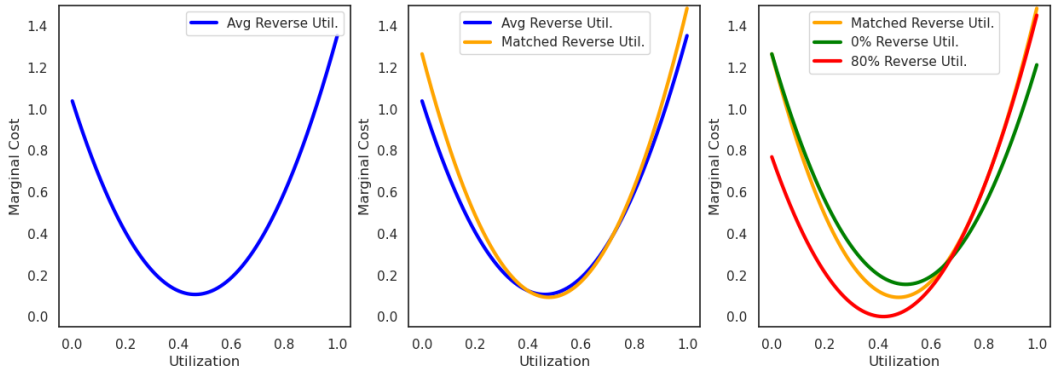
[44] These waybills are restricted to the largest four railroad companies.

Figure A.1: *Number of city pairs utilizing more than one route*



around 50 percent. In terms of magnitudes, we find that economies of scale are weaker for UP compared to BNSF.

Figure A.2: *Estimated marginal costs at the edge level - UP.*



In the estimation section, we reported a single cost measure that aggregated over BNSF and UP when we explored the evolution of marginal costs over time. Figure A.3 presents the corresponding figures for the marginal costs of BNSF and UP separately. Both firms exhibit a similar time trend of increasing marginal costs. Consistent with BSNF having stronger forces of economies of scale, the magnitude of BNSF's increase in marginal costs is larger than that of UP.

Figure A.3: *Estimated marginal cost per mile 1998-2018*

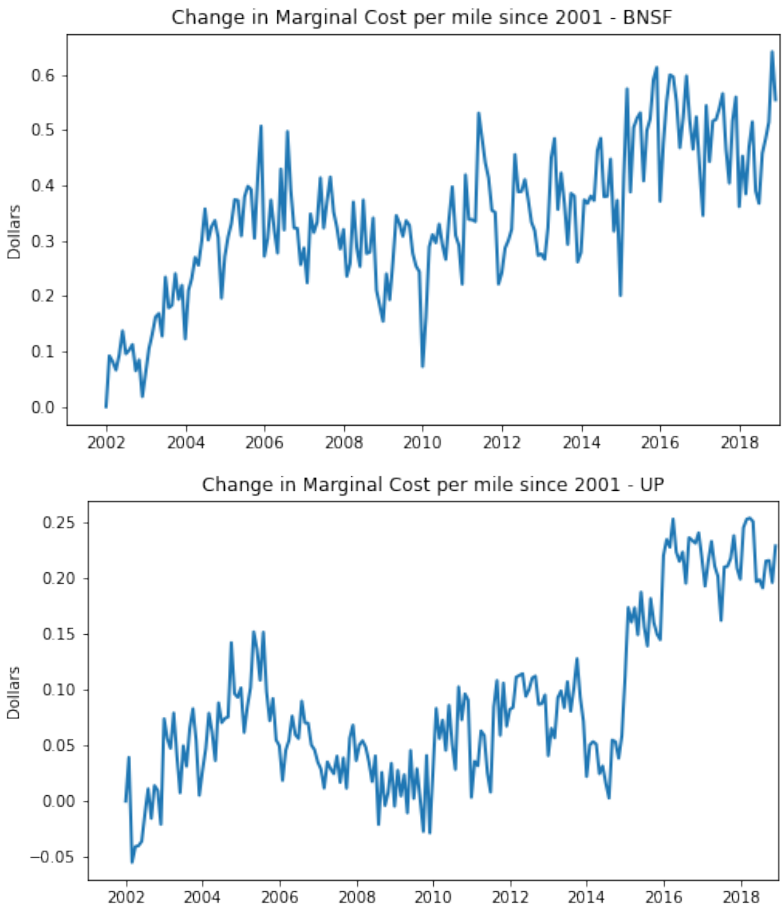


Figure A.4: *US Monthly Coal Production East of the Mississippi River.*

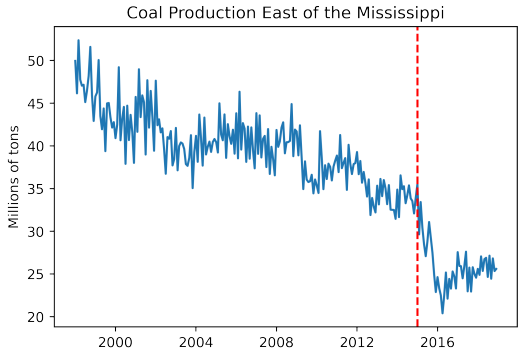


Figure A.4 plots total monthly U.S. Coal Production east of the Mississipi River. The red line denotes January 2015, the beginning of the sample period used in our main estimation. Source: EIA

E Instrument Details

This section presents additional figures related to our choice of instrument for estimation. In Figure A.4, we plot the time-series of U.S. coal production east of the Mississippi. there is an overall decline in production which is present both before and during our sample period. Since the two railroads we study predominantly serve markets west of the Mississippi, we believe that using Eastern coal production provides a stronger case for exogenous variation.

In Figure A.5, we plot a scatter plot of our constructed instrument (Predicted Flow Utilization) against the endogenous variable (Observed Flow Utilization). The first-stage has an R^2 of 78 percent.

F Functional Form

In our baseline analysis, we assume that g, the function capturing how traffic (capacity utilization) enters track marginal costs, is quadratic. In this section, we discuss alternative functional form specifications and show that our estimation results are robust.

First, we can consider letting g be a step function in capacity utilization, which

Figure A.5: *Observed monthly flows vs. instrument.*

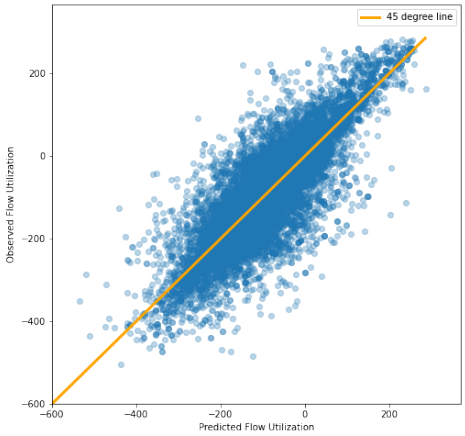


Figure A.5 presents a scatter plot of observed flows against predicted flows. Predicted flows are constructed by training a random forest to predict flows using the shift-share variable constructed in Section 6.4.

Figure A.6: *Robustness to Functional Form*

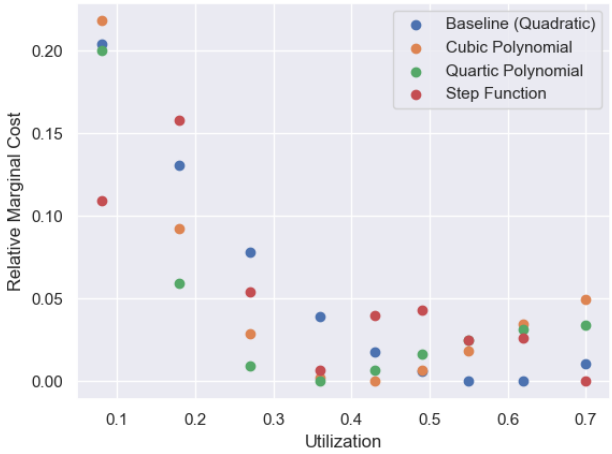


Figure A.6 presents a scatter plot of marginal cost estimates under four alternative specifications of same-direction capacity utilization affecting marginal costs: quadratic (our baseline), cubic, quartic, and a step function. The estimates are plotted at deciles of the capacity utilization distribution. The reference value of the step function is the first decile. This figure does not account for the role of backward flows.

implies a cost function which is piecewise linear. Specifically, we allow for 10 steps from 0 percent utilization to 100 percent utilization in g, where the steps are defined by deciles of the capacity utilization distribution.

Second, we extend the polynomial specification of g by adding third-order (F^3_{uv}) and fourth-order (F^4_{uv}) terms. We re-estimate the cost function with each of these alternate functional form specifications and present the cost curves in Figure A.6. Since the level of costs is not identified, we normalize all cost functions relative to their minimum.

All functions show a similar trend of rapid decreases in marginal cost at low levels of utilization, followed by a flat or modestly increasing region. The utilization level which minimizes marginal costs appears to be sensitive to the specification, with the cubic and quartic polynomials locating the minimal level around 40 percent compared to 55 percent for the quadratic.