A Look at Railroad Costs, Scale Economies, and Differential Pricing

USDA Report

Authors:

John Bitzan, Ph.D.
Director of the Sheila and Robert Challey Institute for Global Innovation and Growth at North Dakota State University

Fecri Karanki, Ph.D.

Doctoral student in the Department of Transportation, Logistics, and Finance at North Dakota State University

February 2021

The Sheila and Robert Challey Institute for Global Innovation and Growth aims to advance understanding in the areas of innovation, trade and institutions to identify policies and solutions that enhance economic growth and opportunity.

This research was supported by Cooperative Agreement Number 18-TMTSD-ND-0006, with the Agricultural Marketing Service (AMS) of the U.S. Department of Agriculture (USDA).

NDSU CHALLEY INSTITUTE FOR GLOBAL INNOVATION AND GROWTH

Contents

Introduction

It is well known that the railroad industry is characterized by a large amount of fixed and common costs.[1] As a result of these fixed and common costs—and of railroads operating longer trains as their traffic levels increase (a point made by Keeler, 1983)—railroads realize economies of scale. Given large scale economies, marginal cost pricing (setting price equal to the cost of an additional unit of service) does not allow railroad firms to recover their full cost of operations. Moreover, any method the railroad uses to assign common costs to individual shipments is arbitrary because they are unattributable to particular shipments. Thus, in order to remain in business, railroads have special authority in the law to charge different customers different prices, depending on their willingness to pay. This practice, known as differential pricing, enables railroads to recover more of their high fixed and common costs from traffic most dependent on their service (i.e. those with relatively inelastic demand).

However, while allowing for differential pricing, Congress also enacted protections to shippers by limiting the maximum price railroads can charge to shippers with few transportation options.[2] Recently, concerns over the high costs of pursuing a rate case for smaller shippers, concerns related to rates and service, and an increasingly profitable railroad industry have led to proposals to change the way maximum rates are regulated. An important consideration in the way maximum rates are regulated—and consequently in the degree of differential pricing that railroads practice—is the extent that railroads realize scale economies. The larger the degree of **economies of scale** realized[3] by railroads, the further marginal costs are from average costs, and the greater the degree of differential pricing railroads will use to remain in business.

This study examines the current cost structure of the U.S. railroad industry, including the extent of economies of scale in the industry and implications for differential pricing. Previous studies have found large economies of scale for the industry, as a result of spreading fixed way and structures costs[4] among more traffic and as a result of better utilization of labor and equipment due to longer and more frequent trains with more traffic. However, recent traffic increases and increasing congestion on some routes may suggest that a large portion of available economies of scale have been exhausted.[5] On the other hand, changes in technology such as electronic controls and railroads operating longer trains on higher density routes (using locomotives in the middle or at the end of the

[1] Fixed costs are those that do not vary with output, such as the cost of the right-of-way, and common costs are those that cannot be attributed to a particular product or service.

[2] The law does not specify a dollar threshold but says a railroad with market dominance cannot charge an unreasonable rate.

[3] Economies of scale refer to reductions in average cost resulting from producing more output. The degree of economies of scale refers to how much average costs fall as output increases.

[4] Way and structures costs are costs associated with the right-of-way, improvements in the right-of-way, and any facilities on or along the right of way. Examples include tunnels, bridges, buildings, terminals, ties, rail, ballast, and communications systems.

[5] It is important to note, however, that congestion on many rail routes is being alleviated by double tracking, additional sidings, and centralized traffic control. This may suggest that the congestion may be a short-run effect, and that long run returns to scale may still be increasing.

train) offer opportunities for railroads to accommodate more traffic at a lower cost.[6] Given the current debates regarding the appropriate regulatory procedures for railroad rates, and shipper concerns over rail rates and service, it is a good time to reevaluate the cost structure of the industry.

The following section provides a non-technical explanation of cost concepts. A discussion of the role of cost in a firm's pricing decisions follows. After an explanation of how cost concepts are applied to the railroad industry, the study briefly reviews the literature on railroad cost analysis. Finally, the empirical methods and data are described, followed by empirical results, and a brief summary of findings.

An Explanation of Cost Concepts

Studying costs provides important insights into product pricing. An understanding of the markups being charged by firms above costs, comparisons of markups of different products, and the overall profitability of various pricing alternatives require an understanding of costs.

In addition, the study of costs sheds light on the nature of production technology faced by firms in an industry, addressing questions with important policy implications. For example, a basic question about the nature of production in an industry is whether there is an advantage to firms producing on a large scale. This can be answered by examining whether proportional increases in the quantity or output produced by the firm lead to less than proportional increases in cost—that is, economies of scale. One could also examine whether there is an advantage to producing more than one product in the same firm. This can be answered by examining whether there are **economies of scope**, which refer to cost savings associated with producing more than one type of output.

Similarly, one might wonder whether inputs are easily substitutable for each other in production. For example, how easy is it for a motor carrier to substitute maintenance for fuel in providing trucking services? Shephard (1953) has shown that questions about the production technology faced by firms in an industry can be answered by examining costs. This section examines the basic theory used by economists to examine production technology using costs and to identify meaningful cost concepts, such as the extent to which costs will change as more output is produced (**marginal cost**) or the costs per unit (**average costs**).

In order to examine the relationships between costs, outputs, and system configurations, economists assume that firms choose inputs to minimize the cost of producing any output, given the prices of inputs and the technology available.[7] This assumption, implied by profit maximization, allows the

---

6 It is important to note that contrary to popular belief, economies of scale in the railroad industry are not only from savings in fixed capital costs. As Keeler (1983) and Miller (1973) pointed out, a big source of scale economies in the railroad industry is in line-haul operations through the introduction of longer and more frequent trains with higher traffic densities. Thus, even when lines become congested, economies of density can be realized through the use of longer trains. In simulating the costs of open access, Bitzan (2003) found significant scale economies in line-haul only operations.

7 It is important to note that much of the explanation of cost concepts and the use of costs in railroad pricing is in the context of a single-product firm. Railroads are multiple product firms, providing services from a variety of origins to a variety of destinations and handling a wide variety of products. The single-product explanation is provided for simplicity, as similar concepts apply in a multiple-product setting.

analyst to examine a wide variety of issues such as the extent of scale and scope economies, the substitutability of inputs, and the impacts of technological characteristics on costs.

Production Relationships

We start by examining the technology available to the firm. The technology available to the firm is defined by production possibilities; that is, the input/output combinations which are technologically feasible. **Inputs** are the factors of production (e.g. labor, fuel, materials) needed to produce a given amount of the firm's product or service (**output**). For example, an automobile repair center may need one hour of a mechanic's time and an hour's worth of tools (inputs) to perform one auto repair (output). One auto repair with an hour of mechanic time and tool time would be included in the firm's production possibilities, but two repairs with an hour of mechanic time and tool time would not.

For a firm that produces only one output (one product or service), the technology available to the firm can be represented by a production function. The **production function** shows the maximum amount of output that can be produced with different quantities of inputs. In cases where the firm produces more than one output (more than one type of product or service), the technology available to the firm is represented by a **transformation function**. The transformation function is similar to the production function in that it shows technologically feasible production. It shows the maximum possible set of outputs that can be produced with various quantities of inputs. (See the sidebar for a mathematical presentation of each.)

Production Function

Mathematically, the production function can be represented as follows:

$$Q = f(\boldsymbol{x}) \qquad (1)$$

where Q is the maximum output that can be produced with a vector of inputs, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, where x_1 is the first input type (e.g., labor), x_2 is the second input type (e.g., materials), etc.

Vectors are often used in mathematics (and mathematical economics) to simplify the notation. At a basic level, a vector encompasses a set of variables or attributes. For example, one's location in space consists of three coordinates: an x (longitude), y (latitude), and z (height or elevation). As the magnitude on one or more variables changes, a person's location changes.

When there is an improvement in technology, this is represented by a new production function.

Transformation Function

The transformation function can be represented as follows:

$$T(\mathbf{Q}, \mathbf{x}) = 0 \qquad (2)$$

where $\mathbf{Q}$ is a vector of outputs, $\mathbf{Q} = (Q_1, Q_2, \ldots, Q_m)$. The transformation function is only equal to zero when the maximum quantities of outputs are produced with given quantities of inputs.

The technology represented by the production or transformation function is translated into an output-cost relationship by solving the cost minimizing problem for the firm. The cost minimizing problem for the firm amounts to choosing inputs to produce the desired level of output, such that the desired level of output is produced at the lowest cost possible.

Long-Run and Short-Run Cost Minimization

Before solving the cost minimizing problem for the firm, it is important to distinguish between the short run and the long run. The **short run** is defined as a period of time when at least one of the inputs of the firm is fixed. For example, if the firm being considered is a railroad, there is some period of time where the amount of track in place and the quality of track in place is fixed. These are called **fixed inputs** because they cannot be adjusted.

If the railroad experiences an increase in the demand for shipments, the lowest cost method of increasing output to meet that demand increase might be to install additional track or to increase the quality of track in place. However, the railroad cannot instantaneously increase the amount of track or increase its quality. Moreover, if such an increase in demand is temporary, the firm is unlikely to want to increase its track investment knowing it will be difficult to reverse that investment. Thus, in this short-run period, where the amount of track in place and its quality cannot be adjusted (it is fixed), the firm will adjust to an increase in demand for shipments by increasing those inputs that can be instantaneously or readily increased, such as labor or fuel. These are called **variable inputs**.

In this example, if the increase in demand for shipments is a permanent increase in demand, the firm may eventually want to increase the amount of track or the quality of track in place. This period of time—where all inputs can be adjusted—is known as the **long run**.

To construct cost-output relationships, we start with the short-run cost minimizing problem of the firm. In the short run, the firm chooses variable inputs to minimize the cost of producing any output, given input prices and available technology (see the sidebar for a mathematical presentation).

Multiproduct Firm Cost Minimization

For the multiproduct firm, this cost minimization problem is represented mathematically as:

$$\min_{x_i \neq k} (\sum_i w_i x_i + w_k x_k) s.t. T(\boldsymbol{Q}, \boldsymbol{x}) = 0 \qquad (3)$$

where w_i are input prices for variable inputs, x_i are the quantities of variable inputs employed, w_k represents the input price of the fixed input, and x_k represents the quantity of the fixed input employed. This constrained cost minimization problem is solved using classical optimization techniques (calculus).

The solution to the problem gives quantities of inputs as functions of input prices, output levels, and the quantity of the fixed factor employed. These functions that show the cost minimizing quantities of inputs to use (for any level of output, input prices, and the quantity of the fixed factor employed) are called **short-run conditional input demand functions**.

The **short-run cost function** shows the minimum cost of producing any output level, given input prices and the quantity of the fixed factor employed. The short-run cost function is obtained when the conditional input demand functions are substituted into an expression showing total expenditures incurred in producing output—input prices times input quantities (see the sidebar for a mathematical presentation).

Short-Run Cost Function

The following walks through the mathematical derivation of the previously described steps to create the short-run cost function. For simplicity, it is presented for one output (q).[1] Here, cost concepts are explored in the context of a single-product firm. Analogous cost concepts are available in a multiple-product context, as well.

First, conditional input demands are obtained by solving the problem in (3):

$$x_i^* = x_i^*(\boldsymbol{w}, q, x_k) \qquad (4)$$

where x_i^* is the conditional input demand for input i, **w** is a vector of variable input prices, q is output, and x_k is the amount of the fixed input employed. Substituting conditional input demands into the expression for total expenditures incurred in producing output (**Equation 5**):

$$C = \sum_i w_i \cdot x_i^*(\boldsymbol{w}, q, x_k) + w_k \cdot x_k \qquad (5)$$

results in the firm's short run cost function (**Equation 6**):

$$C_{SR} = \emptyset(\boldsymbol{w}, q, x_k) + b(k) \qquad (6)$$

where $\emptyset(\boldsymbol{w}, q, x_k)$ are variable costs and $b(k) = w_k x_k$ are fixed costs. This short-run cost function shows the minimum cost of producing any output level, given input prices and the amount of the fixed input employed.

The short-run cost function is useful for a variety of reasons, including identifying short-run relationships between costs and output, and whether firms are operating at full capacity. However, to identify important cost characteristics such as economies of scale and scope, we need a long-run cost function.

As mentioned above, the long run is a period of time where the firm can freely adjust all of its inputs. Thus, in terms of the cost minimization problem faced by the firm, the difference between the short-run and the long-run is that the firm is not able to optimally adjust its fixed factor in the short run, but it is able to in the long run. In fact, the long-run cost function can be obtained from the short-run cost function by minimizing total short-run costs for any output level and input prices with respect to the amount of the fixed factor employed (see the sidebar for a mathematical presentation).

Minimizing Short-Run Costs to Obtain Long-Run Costs

Mathematically, classical optimization techniques (calculus) are used to minimize short run costs with respect to the amount of the fixed factor employed, as follows:

$$\frac{\partial C_{SR}}{\partial x_k} = 0 \qquad (7)$$

It is often assumed that the fixed factor is capital or facility size. Solving this results in an optimal amount of the fixed factor (capital) for any input prices and output level:

$$x_k^* = x_k^*(\boldsymbol{w}, w_k, q) \qquad (8)$$

This is substituted for x_k in the short-run cost function (**Equation 6**) to obtain the long run cost function.

Equation 6, $C_{SR} = \emptyset(\mathbf{w}, q, x_k^{\square}) + w_k x_k$, turns into:

$$\emptyset(\mathbf{w}, q, x_k^*) + w_k x_k^* = C_{LR}(\boldsymbol{w}, w_k, q) \qquad (9)$$

The **long-run cost function** shows the minimum cost of producing any output level, given input prices (depicted mathematically in the sidebar). The difference between the short-run cost function and the long-run cost function is that the long-run cost function shows the minimum cost of producing any output while capital is at its cost minimizing level for that output, while the short-run cost function shows the minimum cost of producing any output while capital is at some specified (fixed) level. This implies that the cost of producing any output on the long-run cost function is lower than or equal to the cost of producing any output on any short-run cost function.

Minimizing Long-Run Costs

The long-run cost function can also be obtained directly from the production function or transformation function if the short-run cost function is not known. In order to obtain the long-run cost function directly from the production function or the transformation function, the following cost minimization problem is solved, where the firm chooses all inputs to minimize costs:

$$\min_{x_i}(\textstyle\sum_i w_i x_i) s.t. T(\boldsymbol{Q}, \boldsymbol{x}) = 0 \qquad (10)$$

where w_i are all input prices and x_i are all inputs. Conditional input demands are obtained from solving this problem through classical optimization techniques. These are then substituted into the expression for expenditures to get the long-run cost function.

The relationship between short-run cost functions and the long-run cost function can be seen for a single output firm in **Figure 1**. **Figure 1** shows a variety of short run cost functions; each representing the minimum cost of producing any output level for a given level of capital (the fixed factor). For example, C_{SR} (K=1) represents a smaller plant size in comparison to C_{SR} (K=2). As the figure shows, each addition to capital leads to higher fixed costs (the vertical distance from the origin when Q is zero)[8] and consequently higher costs of producing small levels of output. However, larger amounts of capital lead to lower costs at higher amounts of output produced. The long-run cost function, which

[8] This distance from the origin to the curve represents fixed costs because fixed costs are incurred regardless of whether there is production or not (Q = 0).

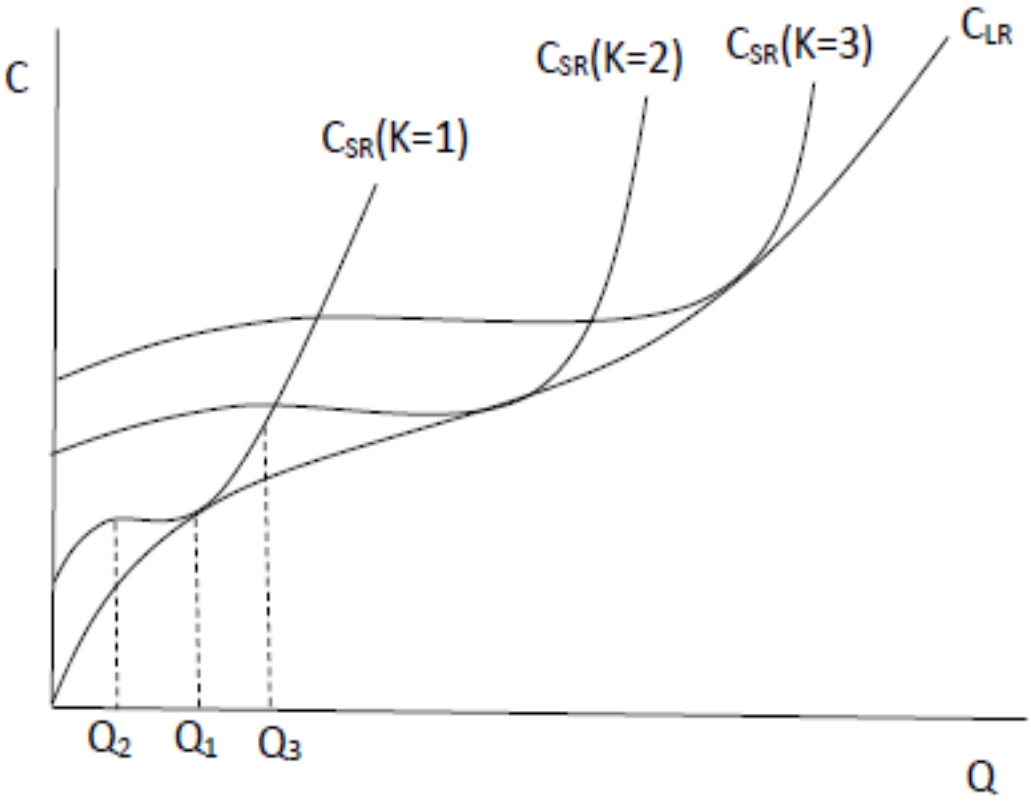

Figure 1: Short-Run and Long-Run Cost Functions

shows minimum cost when using the optimal amount of capital for any output, is tangent to an individual short-run cost function at every point. The point where the long-run cost function is tangent to a short-run cost function is at the output level where the level of capital represented by that particular short-run cost function is the cost minimizing level of capital. For example, the optimal amount of capital to produce output level Q_1 is K=1. Thus, the long-run cost function is the envelope of all short-run cost functions.

The long-run and short-run cost functions illustrated in **Figure 1** are usually estimated with cross-sectional, time-series, or panel data sets that have information on firm costs, outputs, input prices, and technological characteristics. Once these cost functions are estimated, they can be used to identify important characteristics about production technology in the industry. Moreover, they can be used to measure cost concepts that have important ties to pricing.

Economic Costs versus Accounting Costs

An important point to make about the measurement of costs used to estimate cost functions is that **economic costs** are used rather than **accounting costs**. The economic notion of costs is based on the principle of opportunity costs—that is, the value of a resource in its best alternative use. Market prices of products reflect opportunity costs. For example, land in New York City has a higher price than land in Fargo, ND because there are more and better alternative uses for land in New York.

However, because economic costs are based on opportunity costs, many economic costs differ from accounting costs. One reason is that in determining the costs of resources used to produce a product, accounting costs consider the historical cost of acquisition of those resources, while economic costs consider the value of those resources now in their best alternative use (e.g. if you could sell raw materials for a higher price now than the price you paid when acquiring them, your economic costs are higher than your accounting costs). Another reason for the difference is that economic costs consider the implicit opportunity costs realized when no market transaction takes place. For example, if a company uses its own money to purchase durable equipment, the money tied up in that equipment has an opportunity cost even if the company does not have to pay interest on it. Thus, even when companies show an accounting profit, they still may be generating an economic loss. In such cases the company is not generating enough profits to attract the capital that is necessary for continued investment in the industry. In essence, the rate of return being generated on investment in the industry is not as high as investors could earn in equally risky alternatives.

In defining short-run and long-run cost concepts, and in measuring costs, all subsequent discussion refers to economic costs rather than accounting costs. The following paragraphs identify important short-run and long-run concepts that will be used later in the context of pricing.

Fixed versus Sunk Costs

As highlighted previously, in the long run all inputs can be varied by the firm, while in the short run at least one input is fixed. Thus, as shown in **Equation 6**, the short run cost function includes a variable cost component and a fixed cost component.

Fixed costs are defined as those costs that do not vary with output. An example of a cost that does not vary with output in the short run for a railroad would be the opportunity cost of investment in right-of way (way) and structures. There is some period of time where the amount of track in place and the quality of track in place is fixed; that is, the railroad must produce with the given quality and quantity of track in place. If the railroad wants to increase the amount of track it has or make investments to improve the quality of track, it will take some time. Thus, in the short run, this opportunity cost of investment in way and structures is fixed. In the long run, the firm can freely vary the amount and quality of track; thus, the way and structures cost is variable in the long run.

On the other hand, **sunk costs** are costs that are incurred and cannot be recovered. Although sunk costs resemble fixed costs, in that they do not vary with output, they are different from normal fixed costs in that they cannot be reversed. For a railroad, an example of a sunk cost would be the cost of obtaining the right-of-way. If the right-of-way cannot be sold for an alternative use if the route is abandoned, then the cost of obtaining it is sunk. Thus, in calculating economic costs (opportunity costs), sunk costs are not included, since there are no opportunity costs associated with them. With a normal (reversible) investment there is some cost associated with having money tied up in the investment (the investment could be sold and the money used in an alternative endeavor), which is not the case for sunk costs.[9]

[9] As will be highlighted later, sunk costs are also important in that their existence precludes the existence of a contestable market.

Variable costs are defined as costs that vary directly with output. Examples of costs that vary directly with output in the short run would be fuel expenses, labor expenses, and material expenses. The firm is able to immediately purchase additional fuel, labor, and material to increase the amount of railroad services it provides. Thus, these expenses vary with output, even in the short run. Because the firm can freely vary all inputs in the long run, there are no fixed costs in the long run. In the railroad example, the railroad can use additional fuel, labor, materials, and track investment to provide additional railroad services. Thus, all of these expenses are variable.

Unit Costs

It is often useful to examine costs on a per-unit basis. Analysts, firms, and regulators are interested in the average costs of providing a unit of output and the change in total costs resulting from producing more units of output. As a result, the concepts of average, marginal, and incremental costs are often used.

Average costs, as the name suggests, are total costs divided by the amount of output produced.[10] There are three types of average costs that can be calculated: **average total costs** are total costs divided by the units of output produced:

$$ATC = \frac{TC}{q} \qquad (11)$$

where TC are total costs, and q is total output. **Average fixed costs** are total fixed costs divided by amount of output produced:

$$AFC = \frac{TFC}{q} \qquad (12)$$

where TFC are total fixed costs. **Average variable costs** are total variable costs divided by the amount of output produced:

$$AVC = \frac{TVC}{q} \qquad (13)$$

where TVC are total variable costs. Average total costs are also equal to the sum of average fixed costs and average variable costs.

When assessing how changes in output affect costs, we use the concept of marginal cost. **Marginal cost** is defined as the change in total cost resulting from a one-unit change in output. For a railroad, when output is measured in ton-miles, marginal cost would be defined as the change in total cost resulting from producing one-more ton mile of service. Algebraically, marginal cost can be defined as follows:

$$MC = \frac{\partial TC}{\partial q} = \frac{\partial TVC}{\partial q} \qquad (14)$$

where TC is total cost, TVC is total variable cost, and q is output. The reason the change in total variable cost resulting from a change in output can be used is that fixed costs do not vary with

[10] In a multiple output context, average costs cannot be used, since there is no common output that can be used as a denominator in calculating average costs. In such cases, a multiproduct equivalent, called ray average cost is used.

output. Graphically, marginal cost can be seen in **Figure 1** by looking at the slope of the respective cost function.

In examining marginal cost by looking at the slopes of the short-run cost functions and long-run cost function at various points in **Figure 1**, it can be seen that short-run marginal cost and long-run marginal cost often differ from each other. At any point where firms are producing on their long-run cost curve (i.e. they are using the optimal or cost-minimizing amount of the fixed factor to produce output), short-run and long-run costs coincide; short-run and long-run marginal costs also coincide at these points.

On the other hand, to the left of the points of tangency, the firm has excess capacity; that is, the amount of the fixed factor is larger than it should be to minimize the cost of producing the given level of output. To the right of those points, the firm is over-utilizing capacity; it could produce the output at a lower cost by increasing capacity. At both of these points—to the left and to the right of the point of tangency between short-run and long-run costs— short-run costs are higher than long run costs.

In terms of marginal costs, at points to the left of the point of tangency between short-run and long-run costs (where excess capacity exists), short-run marginal costs are below long-run marginal costs. This can be seen graphically in **Figure 1**. Suppose the firm is employing an amount of capital equal to K=1. If the firm is producing Q_2 units of output, its short-run marginal cost is below its long-run marginal cost (i.e., the slope of the short-run cost function is less than the slope of the long run cost function at this point). The intuition behind this is can be illustrated with a railroad example. Suppose that the fixed factor is the quality of track in place. If the railroad has a very high-quality track with very little traffic, the railroad can provide additional ton-miles very cheaply, since traffic will be able to move very rapidly meaning lower marginal labor, equipment, and fuel costs.

In contrast, at points to the right of the point of tangency between short-run and long-run costs, short-run marginal costs are above long-run marginal costs. The reason for this is that the cheapest way to produce more ton-miles involves increasing all inputs—including the fixed input; but, by definition, in the short-run the firm is unable to do so. This can be seen graphically by looking at output Q_3 (still assuming the firm has a capital stock of K=1). In the context of the railroad example, again assume that the quality of track is the fixed factor. If the quality of track is below where it should be for the level of traffic, the railroad may realize congestion and slower speeds when attempting to accommodate more traffic. This would mean that it costs more in additional labor, equipment, and fuel from increasing traffic in comparison to the combined costs of the additional roadway investment, labor, equipment, and fuel if the railroad were able to adjust roadway quality.

Marginal Costs versus Incremental Costs

Although the ideal from society's point of view is for firms to price at marginal cost (as highlighted in the next section), the actual units of output consumed for any product/service may not coincide with the way units of output are measured. In some cases, it is not possible for consumers to consume one unit of output. The railroad industry is an example. In the railroad industry, while units of output are often measured in ton-miles, where a ton-mile is one ton hauled for one mile, railroad consumers do not ever consume just one ton-mile. Instead, railroad consumers ship many carloads full of commodities for hundreds or thousands of miles. A North Dakota wheat shipper might ship 100

railcars of wheat to Portland, where each railcar holds 111 tons of wheat and the distance is 1,400 miles. This shipment would amount to more than 15 million ton-miles.
As a result, in such industries, the concept of incremental cost plays a very important role in pricing. Incremental cost is analogous to marginal costs, except it is for large changes in output. Instead of measuring the change in total cost resulting from a one-unit change in output, **incremental cost** measures the change in total cost resulting from some larger change in output. For the North Dakota shipment highlighted in the paragraph above, the incremental cost of the shipment is the change in total cost resulting from making that shipment. Incremental cost of output i is defined mathematically by **Equation 15**:

$$IC(q_i) = C(q) - C(q - q_i) \qquad (15)$$

where IC(q_i) is the incremental cost of output i, q is total output, and q_i is output i. For example, suppose a railroad was providing 800 shipments and considering the incremental cost of 200 additional shipments. In this case, q = 1,000 and q_i = 200. **Equation 15** produces incremental costs by comparing the cost of providing 1,000 units and the cost of providing 800 units. Thus, if the cost of providing 1,000 shipments is $100,000 and the cost of providing 800 shipments is $84,000, then the incremental cost of providing 200 shipments is $16,000. Using these cost concepts as building blocks, the following section examines the role that costs play in railroad pricing.

The Role of Cost in Pricing

It should be obvious that costs play an important role in the pricing of any product or service. A firm will maximize its profits by charging a price so that its **marginal revenues** (the change in revenue from a one-unit change in quantity sold) are equal to its marginal costs. The intuition behind this rule is straightforward; as long as the firm generates more additional revenue than additional costs from selling another unit, it should keep selling. However, once an additional unit sold adds more to costs than to revenues, it should stop.

This pricing rule and the profit motive tend to promote the interests of society in most cases. When society increases the value it places upon on a particular product or service, this generates opportunities for firms to increase their production of that good or service, since doing so will create additions to revenues that exceed their additions to costs. On the other hand, when society reduces the value it places on a particular good or service, firms are motivated to decrease their production of that good or service. Moreover, to the extent that markets are competitive, firms are signaled to produce an amount that maximizes the difference between the benefits received by society from the good or service and the costs of producing the good or service.

While the free market system and the profit motive align the interests of firms with those of society in most cases, there are also cases where the market fails and some type of regulation may be necessary. An example where regulation may be necessary is when an industry is a **natural monopoly**, where the product or service can be provided at a lower cost by one firm than by more than one firm. In industries characterized by natural monopoly, additional firms—which would inject competition—would also waste resources. However, without competition, the firm may pursue pricing and output policies that are detrimental to society. Thus, regulation may be desired to prevent wasteful competition, while preventing the firm from pursuing pricing and output policies that harm society. The next section presents a simple, non-technical framework that can be used to

evaluate the cost to society from monopoly, and to show how an assessment of cost by regulators can be used in evaluating the appropriate prices charged for products/services.

Social Welfare and the Cost of Monopoly

In order to understand the cost to society from monopoly and the regulatory interest in using cost to determine whether the prices charged by a particular firm are appropriate, it is useful to understand social welfare. **Social welfare** is defined as the value placed on goods and services by society in excess of the costs of resources used to produce those goods and services. Ideally, the goal of regulation is to ensure that social welfare is maximized in cases where there is some kind of market failure in the absence of such regulation. The following paragraphs explain social welfare in more detail.

Total social welfare in an individual market consists of two components: consumer's surplus and producer's surplus. **Consumer's surplus** is defined as the value placed on the good or service by all consumers in excess of the price that they have to pay for it. **Producer's surplus** is defined as the total revenues received by producers in selling the good or service in excess of the costs of producing the good or service. Producer's surplus is the same as economic profits for the firm.

Because social welfare is maximized in competitive markets, it is useful to illustrate social welfare in the context of a competitive market. Although a perfectly competitive market is an ideal that is not entirely achieved in the real world, local markets for agricultural commodities have many of the characteristics of perfect competition, and therefore, are often used as illustrations of competitive markets.

Figure 2 represents the interaction of supply and demand in the market for wheat. In the figure, the supply curve (S) represents the horizontal summation of individual wheat producers' marginal cost curves, and the demand curve (D) shows the horizontal summation of individual consumers' demands for wheat (the amounts they are willing to purchase at various prices.) The equilibrium in this market is where the supply (MC) curve intersects with the demand curve (Point B), resulting in an equilibrium price of P_E and an equilibrium quantity of Q_E. This is also the point where price is equal to marginal cost, since the supply curve represents the horizontal summation of individual producers' marginal cost curves.

In **Figure 2**, the area P_1 B P_E is called consumer's surplus. The demand curve shows the prices that consumers are willing to pay to consume each unit of wheat; therefore, it is the value placed on wheat by consumers. For all quantities of wheat less than Q_E, consumers are willing to pay a price higher than P_E, yet they only have to pay P_E to acquire those quantities. The lower price that consumers actually pay in comparison to the price they are willing to pay makes them better off, which is a benefit to society. Similarly, the producers of wheat would be willing to sell all quantities of wheat below Q_E at prices lower than P_E, as the extra cost from producing each unit (marginal cost) is below P_E at those quantities. This benefit to producers (known as producer's surplus), and hence to society, is shown as the area P_E B A in **Figure 2**.

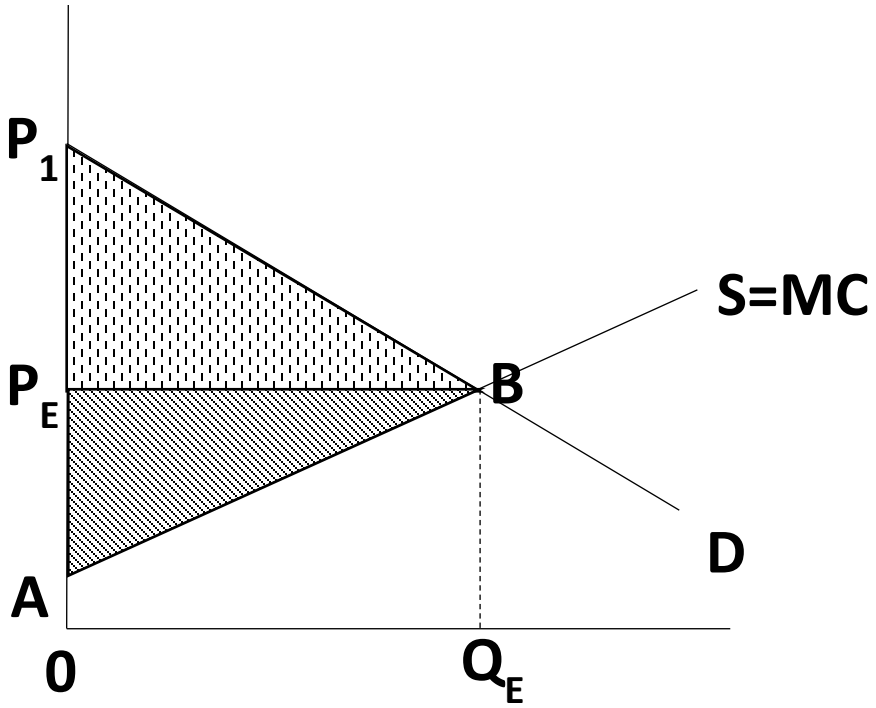

Figure 2: Illustration of Social Welfare in the Market for Wheat

The combination of consumer's and producer's surplus shows the value placed on the good by society in excess of the costs of the resources needed to produce it. This is known as social welfare.

In examining **Figure 2**, it should be apparent that the price and quantity combination of P_E Q_E is the one that maximizes social welfare. At any quantities less than Q_E, the value placed on additional units of the good or service is higher than the costs of resources needed to produce those additional units; thus, there is underproduction of the good or service. At any quantities above Q_E, the cost of resources needed to produce additional units of the good or service exceeds the value placed on those additional units; thus, there is overproduction of the good or service. Thus, price equal to marginal cost is society's ideal or **(first) best** pricing approach; society values the resources used to produce another unit of the good as much as the good itself.

Just as the figure above shows that social welfare is maximized in a competitive market and where price is equal to marginal cost, this same type of framework can be used to show why an unregulated monopoly can harm social welfare. **Figure 3** shows the same competitive equilibrium as **Figure 2**, but also adds the monopoly equilibrium.

In **Figure 3**, the competitive equilibrium is still at the intersection of supply and demand (Point B), or where price is P_E and quantity is Q_E. Since perfect competition assumes that individual firms are a small part of the entire market, individual firms are unable to influence price through their own actions. Thus, economists say these individual firms are "price takers"; that is, they can sell as much as they want at the going price. This means that the extra revenue generated by selling an additional unit of the good (marginal revenue) is equal to the price. Firms maximize profits by continuing to sell as long as the extra revenue from selling another unit (the price) exceeds the extra cost from selling another unit (the marginal cost); thus, they produce where price is equal to marginal cost.

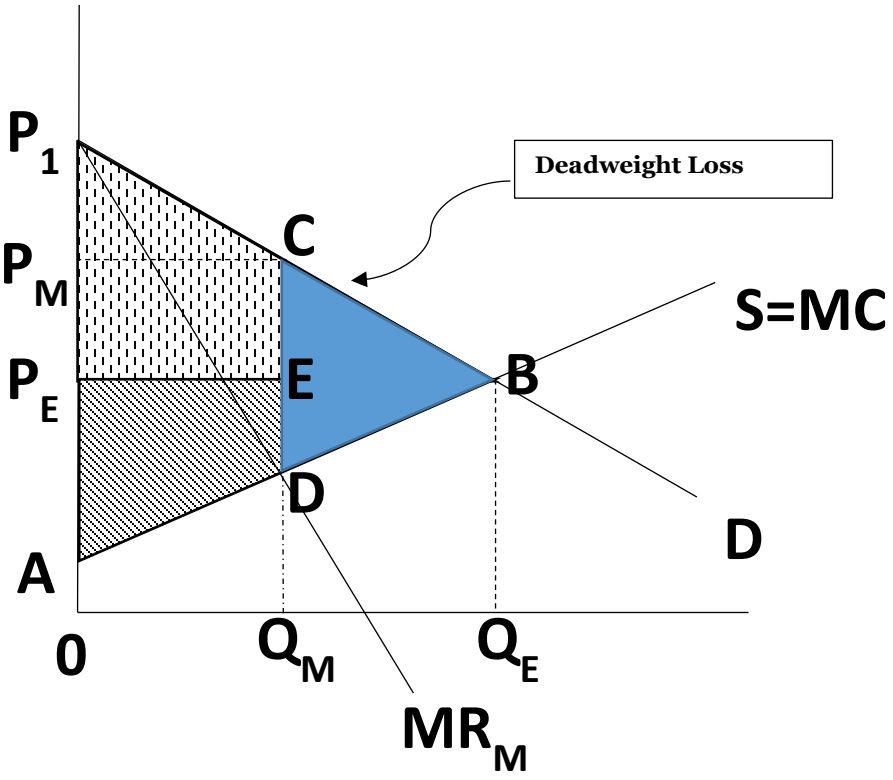

Figure 3: Illustration of Social Welfare Loss from Monopoly

On the other hand, the monopoly firm can only sell more by lowering the price they charge. This is the case because they are the only firm in the market, and therefore, face a downward sloping demand curve. Because the monopoly firm would have to reduce price (on all sold units) to sell a higher quantity, the extra revenue from selling another unit (marginal revenue) is less than the price. The marginal revenue curve for the monopolist is shown as MR_M in **Figure 3**.

As with any firm, the monopolist maximizes profits by producing more as long as the extra revenue from selling another unit (marginal revenue) exceeds the extra cost from selling another unit (marginal cost). As **Figure 3** shows, the maximum profits are obtained at a quantity of Q_M, where marginal revenue equals marginal cost. This results in a price of P_M, obtained from the demand curve.

From the viewpoint of economists, the problem with monopoly is that it leads to a loss in social welfare through an underutilization of resources. As shown in **Figure 3**, the monopoly produces only Q_M, which is much lower than the social welfare maximizing quantity of Q_E. All quantities of the good or service between Q_M and Q_E are valued more by society than the costs of the resources needed to produce those quantities, yet they are not produced. This loss to society is the triangle C B D and is called a **deadweight loss**. Society loses due to this misallocation of resources, as not enough of the product is produced.

Another effect of monopoly is that there is a transfer of wealth (surplus) from consumers to the producer. This transfer, which neither creates nor destroys surplus, does not harm social welfare but is often of interest to society. The transfer of wealth is the area P_M C E P_E. This area, which would be consumer's surplus with a perfectly competitive equilibrium, becomes producer's surplus in the monopoly equilibrium. Although this transfer is of interest to many, the true loss to society is the net loss of consumer's and producer's surplus due to the underutilization of resources (the deadweight loss triangle C B D). In this situation, the regulator wishing to maximize social welfare would force the monopoly firm to charge a price equal to marginal cost.

Regulation with Increasing Returns to Scale

While the regulatory task seems simple based on the above diagrams, there are many things in the real world that make this task more difficult. One consideration that complicates the above analysis is that some industries are characterized by **economies of scale**. Economies of scale occur when average costs are declining with increases in output.[11]

Figure 4 shows average and marginal costs for a firm in a single-product industry that is characterized by economies of scale. As the figure shows, because marginal cost is the change in cost resulting from producing another unit of the good and average cost is total cost divided by output, average cost is always declining when it is above marginal cost, and it is always increasing when it is below marginal cost.[12]

If average cost is declining at the point where marginal cost intersects demand, then forcing the firm to price at marginal cost would force the firm to lose money. Because of this dilemma, the **second best** solution (that which minimizes the loss in social welfare without forcing the firm to lose money) is for price to be set equal to average cost. This results in a loss in social welfare (a deadweight loss triangle), but one that is smaller than if the monopoly sets the profit maximizing price. Thus, even with economies of scale, the regulatory task is still relatively straightforward, at least theoretically.

[11] This becomes more complex in a multiproduct context.

[12] For the non-economist an intuitive example might help to understand this. Suppose that 10 people are in a room and their average age is 50. If an eleventh person, who is 40, is brought into the room the average age drops to approximately 49. The age of 40, by the eleventh person could be thought of as the marginal age. On the other hand, if an eleventh person who is 60, is brought into the room, the average age rises to about 51.

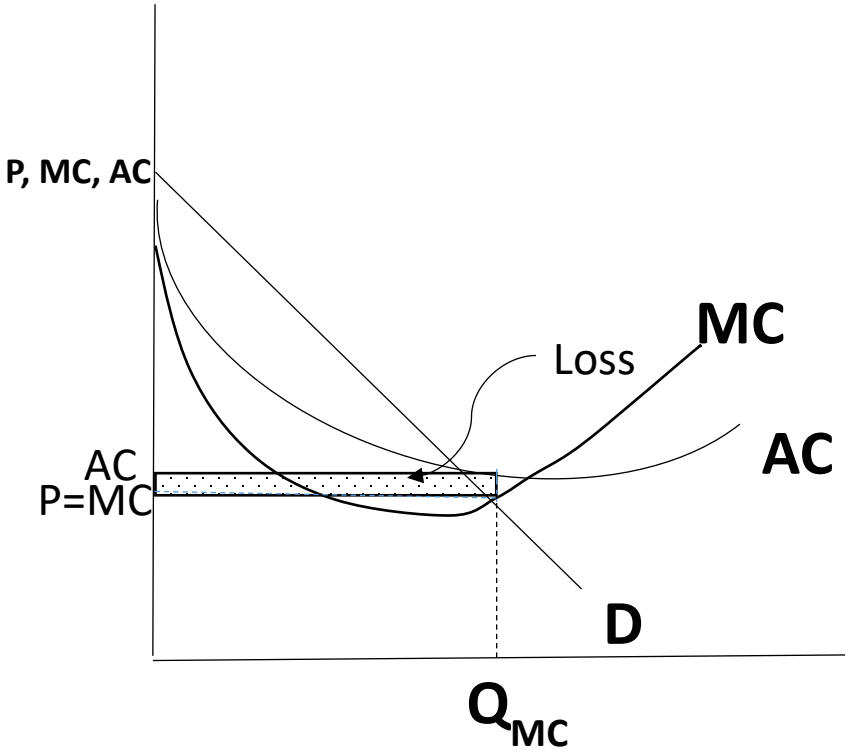

Figure 4: Illustration of Pricing with Scale Economies

Multiple Markets and the Rationale for Ramsey Pricing

However, in most cases, including that of railroads, firms serve multiple markets (or sell multiple products). When the firm serves multiple markets with different demands, total social welfare is maximized by choosing prices in each market that minimize social welfare losses in each individual market, while also allowing the firm to earn a rate of return sufficient to allow for reinvestment (this is known as "zero economic profit"). The second best social welfare maximizing rule in such cases is called **Ramsey pricing**. Under Ramsey pricing, the firm prices inversely to the elasticity of demand. This entails charging higher prices in more inelastic markets (customers with fewer alternatives) and lower prices in more elastic markets (customers with more alternatives). A more detailed explanation, along with the intuition of such an approach follows.

In order to understand Ramsey pricing, it is necessary to understand price elasticity of demand. The **price elasticity of demand** is defined as the percentage change in quantity demanded resulting from a one percent change in price, as shown in **Equation 16**:

$$\varepsilon_p = \frac{\%\Delta Q_D}{\%\Delta P} = \frac{\frac{\Delta Q}{Q}}{\frac{\Delta P}{P}} = \frac{\Delta Q}{\Delta P}\frac{P}{Q} \; or \; \varepsilon_p = \frac{\partial Q}{\partial P}\frac{P}{Q} \qquad (16)$$

Since the quantity demanded of any product will always decrease with an increase in price, price elasticity of demand is always negative. Thus, when discussing price elasticity of demand, economists often refer to absolute values. For example, a price elasticity of demand of -2 means that a one-percent increase in price leads to a two-percent decrease in quantity demanded, while an elasticity of

-3 means that a one percent increase in price leads to a three percent decrease in quantity demanded; economists would say the latter is more elastic. Markets with a lower elasticity in absolute value are considered to be more inelastic (demand is less responsive to a change in price), while those with a higher elasticity in absolute value are considered to be more elastic (demand is more responsive to a change in price).

An understanding of elasticity facilitates an understanding of Ramsey pricing and its intuition. Ramsey pricing is a second best solution to maximizing social welfare. It is second best in the sense that it yields the maximum possible social welfare that can be achieved while also allowing the firm to break even; yet, it yields a smaller social welfare than pricing at marginal cost. In order to achieve the maximum social welfare, a price is charged in each market so that it reduces output by the same proportion in each market in comparison to the output that would be produced if price were set equal to marginal cost. This is shown by Sharkey (1982) as in **Equation 17**:[13]

$$\frac{Q_{D1}(p_1)}{Q_{D1}(mc)} = \frac{Q_{D2}(p_2)}{Q_{D2}(mc)} \qquad (17)$$

where $Q_{D1}(p_1)$ represents the quantity demanded in market 1 when charging a price of p_1, $Q_{D1}(mc)$ represents the quantity demanded in market 1 when charging a price equal to marginal cost, and $Q_{D2}(p_2)$ and $Q_{D2}(mc)$ are defined analogously for market 2.

From **Equation 17**, it is obvious that a larger markup above marginal cost should occur in markets that have less elastic demand and a smaller markup should occur in markets that have more elastic demand. This intuition can further be shown through a simple graphic illustration.

Figure 5 provides a simple illustration of why charging a uniform price in markets characterized by different demand conditions does not maximize social welfare. In examining **Figure 5**, assume that a firm that produces its product at a constant marginal cost can sell it in two different markets: an inelastic market shown on the left or an elastic market shown on the right. From the previous analysis, we know that charging a price above marginal cost will result in a deadweight loss in each market. As the figure shows, if the firm charges the same price in each market the deadweight loss is much larger in the elastic market than in the inelastic market. This is the case because any given price increase above marginal cost will lead to a larger percentage decrease in quantity demanded in a market the more elastic the demand curve. Thus, in terms of social welfare, there is a larger underutilization of resources in the elastic market; that is, there are larger quantities valued more by consumers than the cost of producing them that are not being produced in the elastic market.

[13] Sharkey (1982) shows this for multiple products, where the equation would be the same except 1 and 2 would represent different products, and each would have a different marginal cost. However, he also notes that this applies equally if the same product is sold in different markets. For simplicity, the illustration here uses the same product sold in different markets. Another way to state this rule, as shown later, is that the markup above marginal cost should be set inversely to the elasticity of demand. This forces outputs to be reduced by the same proportion for all goods, in comparison to the output that would be sold at marginal cost.

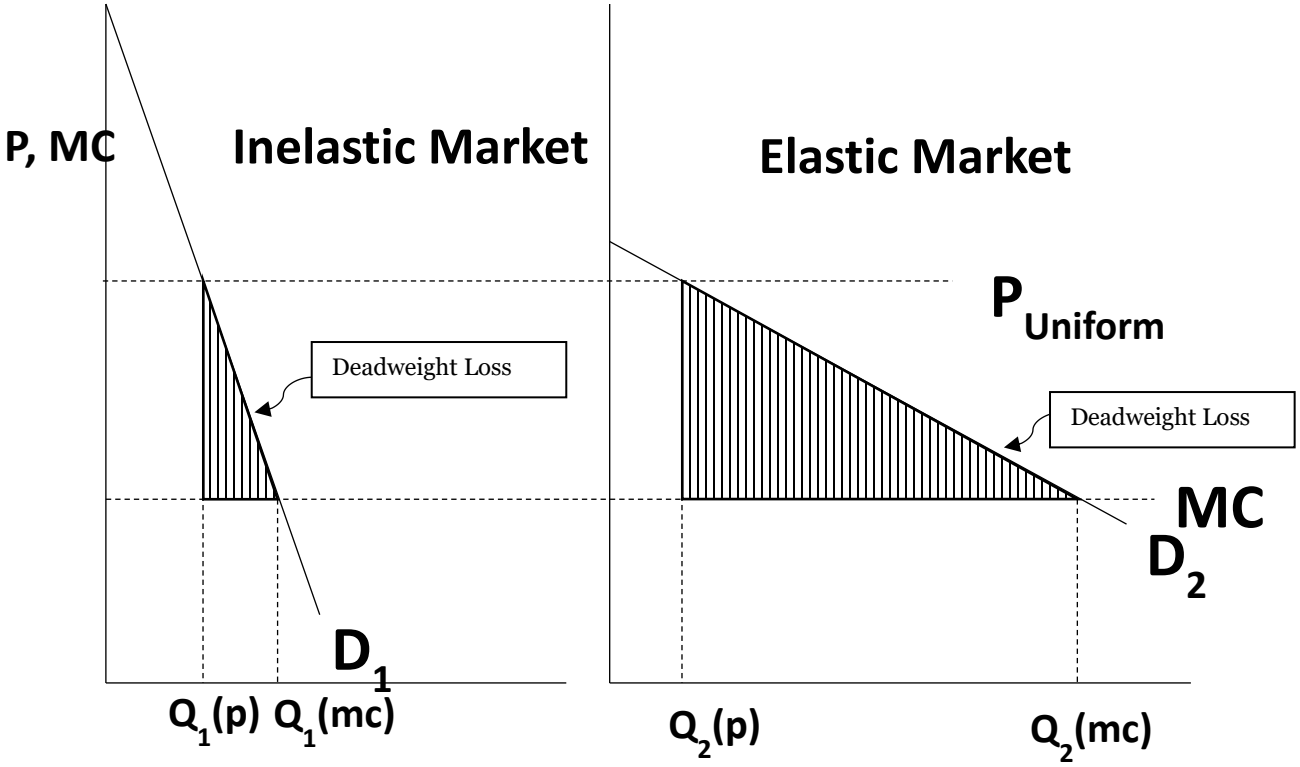

Figure 5: Deadweight Losses Resulting from Uniform Pricing in Different Markets

Although not explicitly shown in **Figure 5**, a higher markup in the inelastic market and a lower markup in the elastic market will result in a smaller social welfare loss. The Ramsey rule of charging a markup that equalizes the proportional reduction in quantity demanded in each market is the one that minimizes the loss in social welfare from charging a price that is different from marginal cost.

Formally, the Ramsey pricing rule is that the markup in each market (or for each good) should be as follows:

$$\frac{P_i - MC_i}{P_i} = \frac{k}{|\varepsilon_{pi}|} \qquad (18)$$

where k is a proportionality constant that is the adjustment of the markup needed in all markets to make the firm break even, and ε_{pi} is the price elasticity of demand in market i. The proportionality constant (k) will be between 0 and 1, depending on the degree of differential pricing needed to allow the firm to break even. As shown by the equation, Ramsey pricing results in pricing inversely to the price elasticity of demand.

Application of Cost Concepts to Railroad Pricing

The previous section provided an introduction to social welfare, the costs of monopoly, and the rationale for Ramsey pricing in markets characterized by increasing returns to scale. This section presents a brief discussion of cost issues specific to the railroad industry, the measurement of railroad costs, and the role that cost has played in evaluating the reasonableness of railroad rates.

In evaluating costs in the railroad industry, an important consideration is that the railroad industry is characterized by large amounts of fixed and common costs. **Fixed costs** are costs that do not vary with output. **Common costs** are those costs that cannot be attributed to particular products or services, but are realized nonetheless.

From the previous discussion, we know that economists define the long run as a period of time where all costs vary with output, and the short run as a period of time where at least one input is fixed. Thus, according to these definitions, there are fixed costs in the short run, but no fixed costs in the long run. The rationale for these definitions is that there is a period of time where some of the firm's inputs cannot be adjusted to meet demand conditions (the short run), but eventually those inputs can be adjusted to meet demand conditions (the long run). For example, a manufacturing firm that experiences a surge in the demand for its product cannot instantaneously increase the size of its plant and equipment, but if the surge in demand is permanent, it can eventually add on to its plant and acquire larger equipment.

In the railroad industry, however, there are indivisibilities in the roadway and structures. For example, in order to accommodate a one-hundred car freight train every day for a year from a particular origin to a particular destination you need to have the roadbed and a rail line in place. If, instead, you want to accommodate a one-hundred car freight train for two days of the year from that same origin to that same destination, you would still need to have the roadbed and rail line in place. In other words, while the quality of the roadbed and rail line can certainly be varied to accommodate different amounts of traffic, there is some minimum amount of roadbed and rail line that is needed regardless of the amount of traffic hauled. In this sense, these costs might be considered fixed, even in the long run.

Closely related to the notion of fixed costs is the concept of common costs. The railroad industry is a multiproduct industry, where the products provided by railroad firms are the transportation of a wide variety of different products from a wide variety of origins to a wide variety of destinations. When railroads produce any product, there are costs that are clearly attributable to that particular product (e.g. the extra fuel cost resulting from handling the additional tons of that product). In addition, however, there are also costs that cannot be attributed to any individual product (e.g. if two different products are being carried from the same origin to the same destination, the line-haul costs of the train crew are the result of handling both products, and cannot be attributed to either product individually). These costs are called common costs. Just as common costs cannot be attributed to individual products in the multiproduct case, fixed costs cannot be attributable to individual amounts of the same product in the single product case.

These characteristics of the railroad industry have made the measurement of movement specific costs very difficult. Economists have successfully estimated aggregate cost functions that have allowed for an assessment of scale economies and the impacts on costs resulting from producing multiple products. However, these aggregate cost studies have not been successful at identifying the costs of specific movements.[14] Similarly, while the Interstate Commerce Commission (ICC) and subsequently the Surface Transportation Board (STB) have made good progress in developing a practical approach to approximating the costs of individual rail movements, the approach has been criticized on multiple grounds. One criticism is based on some arbitrary allocations of common costs. While various approaches to estimating marginal, incremental, variable, and fixed railroad costs are examined in another section, this section explores difficulties in using costs to determine appropriate rail prices (rates).

Using Railroad Costs to Regulate Pricing

Historically, costs have played an important role in railroad pricing. Willig and Baumol (1987) point out that in regulatory history, "fully-allocated" or "fully-distributed" costs served as both an indicator of minimum reasonable rates and as an indicator of maximum reasonable rates.[15] **"Fully-allocated" cost** is an accounting concept meant to approximate average cost.

As mentioned previously, many costs of providing railroad services are fixed and common. Some inputs used in providing railroad services are not completely divisible (e.g. in order to provide one railroad shipment between a particular origin and destination, you need to have a minimum amount of track in place, and this same amount of track could accommodate many more shipments). Moreover, many railroad costs are common among different shipments (e.g. a railroad movement that handles cars with different commodities shares the costs of the locomotive, the signaling, the roadway and structures, the train crew, and the fuel). As a result of both phenomena, many costs cannot be attributed to particular shipments. "Fully-allocated" cost assigns these unattributable costs to individual shipments, based on the total attributable costs of the shipment, the volume or weight of the shipment, or some other measure.[16]

The logic behind using "fully-allocated" or average costs as a measure of maximum reasonable rates is based on a notion of equity, with two components. First, is the idea that no shipper should have to pay more than the cost of providing the service. Second is the idea that no shipper should have to cross-subsidize services to other shippers.

While ensuring that no shipper pays more than the cost of providing the service and that no shipper should have to cross-subsidize services provided to other shippers are reasonable goals for the regulator to pursue, there are two problems with using "fully-allocated" costs to achieve these goals.

---

[14] A subsequent section reviews various econometric cost studies.

[15] Willig and Baumol (Regulation, 1987). In this section, the terms "fully allocated" and "fully distributed" are used interchangeably.

[16] In the railroad industry, the allocation of these unattributable costs by regulators has been based on the total attributable costs of the shipment.

First, the notion that the "fully-allocated" cost of the shipment is the cost of providing the service is fundamentally flawed. By definition, a portion of the costs included in "fully-allocated" costs are unattributable. Thus, any attempt to assign these costs to individual shipments is arbitrary and without economic meaning.[17]

Second, following from the attempt to apply economic meaning to this allocation is an error in defining a cross-subsidy. The economically correct definition of a cross-subsidy notes that as long as the price charged for a service exceeds (or at least equals) the additional cost of providing the service (incremental cost), then the service is not being cross-subsidized by any other service. Faulhaber (1975) illustrates the economic definition of a cross-subsidy by using an example of four neighborhoods that are served by a water company. A brief description of Faulhaber's example will make the appropriate definition of cross-subsidy clear.

An Example: Incremental Cost, Fully-Allocated Cost, and Cross-Subsidy

Faulhaber's example includes two neighborhoods (1 and 2) that are east of a common well and two neighborhoods (3 and 4) that are west of the well. The components of the cost of providing service include: (1) a well and storage tank that cost $160; (2) one trunk line and pumping station dedicated to eastern neighborhoods, and one dedicated to western neighborhoods, each with a cost of $100; and (3) a distribution system for each neighborhood, costing $100.

Using this example, the cost of serving all neighborhoods is $760, calculated as follows:

$$C_{1,2,3,4} = C_{WS1,2,3,4} + C_{TL1,2} + C_{TL3,4} + C_{DS1} + C_{DS2} + C_{DS3} + C_{DS4} \quad (19)$$
$$= 160 + 100 + 100 + 100 + 100 + 100 + 100$$

where C_{WS} is the cost of the well and storage tank, C_{TL} is the cost of the trunk line and pumping station, C_{DS} is the cost of distribution systems, and the subscripts 1, 2, 3, and 4 refer to each of the neighborhoods.

Similarly, the cost of serving two eastern neighborhoods alone and the cost of serving two western neighborhoods alone would each be $460 calculated as (shown for the eastern neighborhoods):

$$C_{1,2} = C_{WS1,2} + C_{TL1,2} + C_{DS1} + C_{DS2} \quad (20)$$
$$= 160 + 100 + 100 + 100$$

The cost of serving one eastern neighborhood and one western neighborhood (any two non-adjacent neighborhoods) would be $560, calculated as:

$$C_{1,3} = C_{WS1,3} + C_{TL1} + C_{TL3} + C_{DS1} + C_{DS3} \quad (21)$$
$$= 160 + 100 + 100 + 100 + 100$$

[17] An excellent and entertaining illustration of the flaws associated with using "fully-allocated" costs for regulatory purposes is included in Baumol, W.J., Koehn, M.F., and R.D. Willig, "How Arbitrary is 'Arbitrary'? – or, Toward the Deserved Demise of Full Cost Allocation," *Public Utilities Fortnightly*, September 3, 1987. In the paper, the authors show widely varying assessments of "costs" based on different methods of allocating unattributable costs. They argue that each method of allocating unattributable costs could be deemed "reasonable" and show that such allocation is void of any economic meaning.

The cost of serving any three neighborhoods would be $660, calculated as:

$$C_{1,2,3} = C_{WS1,2,3} + C_{TL1,2} + C_{TL3} + C_{DS1} + C_{DS2} + C_{DS3} \quad (22)$$
$$= 160 + 100 + 100 + 100 + 100 + 100$$

Finally, the cost of serving any neighborhood alone would be $360, calculated as (shown for neighborhood 1):

$$C_1 = C_{WS1} + C_{TL1} + C_{DS1} \quad (23)$$
$$= 160 + 100 + 100$$

As mentioned above, a cross-subsidy does not exist as long as the revenue charged to any service (and to any set of services) is at least as great as the incremental cost of providing that service (and any set of services). In the context of this example, the incremental cost of providing any service (or any subset of services) is the total cost of providing all services less the cost of providing all other services except that service (or subset of services). Thus, the incremental cost of providing service to any individual neighborhood is $100 ($C_{1,2,3,4} - C_{1,2,3}$), the incremental cost of providing service to any two adjacent neighborhoods is $300 ($C_{1,2,3,4} - C_{1,2}$), the incremental cost of providing service to any two non-adjacent neighborhoods is $200 ($C_{1,2,3,4} - C_{1,3}$), and the incremental cost of providing service to any three neighborhoods is $400 ($C_{1,2,3,4} - C_1$).

This example can be used to show the importance of considering demand in pricing, and the harm that can be placed on consumers by using "fully-allocated" costs to ascertain whether a cross-subsidy is occurring; even when consumers who pay the highest prices may seem to be the ones who would benefit from eliminating such a "cross subsidy." If we use the notion of "fully-allocated" costs and assume that the use of water is the same for each neighborhood, the "fully allocated" cost per neighborhood would be $190, calculated as follows (using neighborhood 1 as an example):

$$FAC_1 = C_{attrib\ DS1} + C_{alloc\ TL1,2} + C_{alloc\ WS1,2,3,4} \quad (24)$$
$$= 100 + 50 + 40$$

where 100 percent of the costs of its distribution system ($100) are allocated to neighborhood 1, 50 percent of the common costs of the trunk line and pumping station for neighborhoods 1 and 2 are allocated to neighborhood 1 ($50), and 25 percent of the common costs of the well and storage tank for all neighborhoods are allocated to neighborhood 1 ($40).

Now, suppose that the water company is a regulated monopoly that is constrained to earn zero profits, and suppose neighborhood 1 is only willing to pay $160 for water services. Further, suppose that each of the remaining neighborhoods is charged a price of $200 each for water services. The other neighborhoods complain of a "cross-subsidy," and say that neighborhood 1 is not paying its fair share of expenses based on the notion of "fully-allocating" common costs.

When the alternative to this pricing scheme of charging all services the same price (based on the notion of equity embedded in the idea of "fully-allocated" costs) is considered, charging the lower price to neighborhood 1 is not a cross-subsidy. Suppose the regulator requires the water company to charge "fully allocated" cost to each neighborhood. In this scenario, neighborhood 1 stops buying water from the water company, and the total cost of providing service to the remaining three neighborhoods is $660 ($C_{WS2,3,4} + C_{TL2} + C_{TL3,4} + C_{DS2} + C_{DS3} + C_{DS4}$). Then, the "fully-allocated" cost

of serving neighborhood 2 becomes $253.33, calculated as (where 1/3 of the costs of the well and storage tank are allocated to each neighborhood = $160/3 = $53.33):

$$FAC_2 = C_{attrib\ DS2} + C_{attrib\ TL2} + C_{alloc\ WS2,3,4} \quad (25)$$
$$= 100 + 100 + 53.33$$

and the "fully-allocated" cost of serving neighborhoods 3 and 4, each become $203.33 (where ½ of the neighborhood trunk line is allocated to each neighborhood = $100/2 = $50):

$$FAC_3 = C_{attrib\ DS3} + C_{alloc\ TL3,4} + C_{alloc\ WS2,3,4} \quad (26)$$
$$= 100 + 50 + 53.33$$

The price charged to the neighborhoods that were deemed to be treated "unfairly" is increased by not having neighborhood 1 being served by the water company, showing that the lower price paid by neighborhood 1 is clearly not a cross-subsidy. The example shows that as long as any customer is paying a price that is above the incremental costs of providing that service, and as long as the seller's profits are constrained, the other customers benefit from the contribution to common costs.

Similar examples can be constructed to show that as long as the price charged to any one neighborhood is above $100 (the incremental cost of serving it), the remaining neighborhoods pay a lower price as a result of having the neighborhood included. In discussing early economic arguments for pricing inversely with elasticity of demand, Baumol and Bradford (1970) state "For (particularly if the firm is subject to a constraint on its overall profit) the opening of a market which makes any net contribution may permit or may even require a reduction in prices elsewhere."

Applications to the Railroad Industry

The example of the water company has many applications to the railroad industry. As stated previously, railroads are multiproduct firms. They transport a wide variety of products from a wide variety of origins to a wide variety of destinations. Consequently, the different "outputs" or services provided by railroad firms share many of the same inputs, including rights of way, rail track, structures, labor, and equipment.

To the extent these inputs are common between individual services, they cannot be allocated in a meaningful way without considering demand. It is well known that railroad shippers (those that hire rail carriers to transport their products) have varying alternatives to shipping their products by rail. In some cases, shipping alternatives and perceived advantages of various modes depend on the commodity shipped. For example, some shippers may have a commodity where improved timeliness and security afforded by truck or air are an advantage, while other shippers may transport a commodity where this is not an advantage. When shippers have a commodity that is fragile and/or high value, the shipper may require a lower rate by rail in order to overcome the perceived disadvantages associated with rail service in comparison to the alternatives. As another example, shippers of some products may be able to sell their product(s) in many markets, while shippers of others have few options in which to sell their product(s). When a shipper has many options in which to sell their product(s), it is also likely that they have more transport options to consider in delivering their product(s).

In other cases, the available alternatives for shipping a firm's products depend on the geographic location of the shipper. Some shippers may have nearby options for delivering their products to markets, increasing the viability of truck as a rail alternative, while in other cases such options are not available. Similarly, some shippers may have nearby options for using alternative modes (or alternative rail carriers) for transporting their product(s), such as nearby access to a water loading facility, while other shippers do not have such options.

Because of varying alternatives for delivering products to final destinations, shippers have varying price elasticities of demand for transporting their products; that is, they have differences in their willingness to pay for rail services. In many cases, those shippers' willingness to pay is less than "fully-allocated" or average costs for shipping by rail. At that level, they would seek a cheaper alternative. To the extent that the railroad is constrained to charging a maximum rate of average costs on any shipment and that some (relatively more elastic) shippers are not willing to pay that amount (and will instead use an alternative), railroad revenues would be inadequate for recovering the full costs of operation. Moreover, as the above example of the water company makes clear, as long as the prices charged to those shippers with more transportation alternatives (those with more elastic demand) are above the incremental costs of providing services to those shippers, "captive shippers" (those with fewer transportation options) are not harmed by the lower prices charged to competitive shippers. To the extent that railroad profits are constrained, such lower rates to competitive shippers may also mean lower rates to captive shippers.

The prevalence of fixed and common costs, and the varying alternatives faced by customers that characterize the railroad industry, make differential pricing a necessity. Multiple outputs using shared inputs lead to a large amount of common costs that cannot be allocated to any individual shipment in an unarbitrary way. Moreover, because of varying transportation alternatives, many shippers would choose not to ship with the railroad if they were charged a price that attempted to "average out" these common costs among shipments. This suggests that any attempt to charge a uniform price to all customers would lead to losses for railroads. As highlighted in a previous section, the most economically efficient solution to pricing in this situation is Ramsey Pricing—pricing inversely to the elasticity of demand.[18]

The previous background on costs, their application in the railroad industry, and the role they play in pricing provides the foundation for understanding the cost analysis that follows. The next section provides a brief history of railroad cost analysis. A more detailed history of railroad cost analysis by economists, as well as a discussion of the Uniform Rail Costing System (URCS), is included in **Appendix B**.

[18] While Ramsey Pricing is desirable, it requires detailed information on price elasticity of demand for every shipment. Thus, the Surface Transportation Board has adopted an alternative to Ramsey Pricing for regulating railroad rates, known as Constrained Market Pricing.

Roots of Railroad Cost Analysis

As highlighted by Waters and Woodland (1984), railroad cost analysis has generally followed along two distinct paths. On the one hand, academic economists have examined relationships between total railroad costs, outputs, input prices, and system characteristics—estimating a **cost function** to examine questions related to the structure of railroad costs and policy issues. On the other hand, railroads and regulatory agencies have examined the relationships between specific railroad cost accounts and activity measures in order to measure the costs of specific rail movements—a process known as **railroad costing**.

Both of these approaches have distinct advantages and disadvantages. The cost function approach used by academic economists is one that has been used successfully to identify and measure a number of cost concepts (e.g. economies of density, economies of size, and economies of scope) that have important policy implications.[19] Moreover, the approach has been refined over time to ensure its consistency with economic theory and to consider advancements in econometric methods and computing technology. However, while the approach has yielded success in identifying broad cost characteristics in a way that is consistent with economic theory, it has not been successful in estimating the costs of specific railroad movements.

Railroad costing has been successful in generating reasonable approximations of movement specific costs in a way that has not been possible with cost functions. However, it has been criticized heavily for its noncompliance with economic and statistical theory.[20] Despite such criticisms, most analysts recognize that it is a practical alternative to cost analysis consistent with economic theory, when movement specific costs are needed.[21]

While the first railroad cost analyses that took place in the late 1800s and early 1900s used much less sophisticated techniques than those used today, they addressed the same issues of how costs vary with output, network size, and the types of outputs produced. Moreover, they addressed the problems of common and unattributable costs, their allocation among shipments, and the need for differential pricing when the industry has economies of scale.

Since then, a number of important innovations to cost analysis have taken place, including utilizing statistical techniques to measure cost/output/input price/system characteristic relationships, matching cost analysis with economic theory, utilizing appropriate data for analyzing railroad costs, and distinguishing between two different types of scale economies in the railroad industry—economies of density and economies of firm size. **Economies of density** refer to reductions in average cost resulting from increased traffic over a network of a given size, while **economies of**

[19] Economies of density are reductions in average cost resulting from increased traffic over a network of a given size. Economies of size are reductions in average cost resulting from increased traffic due to expansions in the size of the network. Economies of scope are cost savings resulting from diversifying the product mix.

[20] For example, Wilson and Wolak (2016) list a number of criticisms of the Uniform Railroad Costing System (URCS) that is used by the Surface Transportation Board to measure movement specific costs for regulatory purposes.

[21] InterVISTAs (2016) points out that the system used to estimate individual movement costs by the U.S. Surface Transportation Board, URCS, can be improved. They also point out, however, that it is a reasonable approach to identifying movement-specific costs, given the difficulty of this task.

size refer to reductions in average cost resulting from increased traffic due to expansions in the size of the network. Other innovations in cost analysis have included utilizing flexible functional forms, distinguishing between way & structures capital and route mileage, considering firm-specific effects, and including measures to account for differences in railroad system and traffic characteristics. In addition to including the innovations introduced in these previous studies, the current study introduces a new innovation of introducing commodity-specific output measures. The approach used is described in the next section.

Data and Methodology for Estimating Railroad Costs

As highlighted previously, economies of scale imply marginal costs are below average costs; therefore, marginal cost pricing will not recover total costs. As a result, the social welfare maximizing solution for multiproduct or multimarket firms that realize economies of scale is Ramsey pricing; that is, pricing inversely to the elasticity of demand for the product or service. Moreover, the extent of differential pricing needed for the firm to recover total costs and earn a return on investment necessary to attract capital is greater the more extensive the scale economies realized.

This section describes the methodology used to estimate scale economies in the railroad industry. Any changes in regulation of maximum rates for the industry (particularly any changes that could lead to a more homogeneous rate structure) should consider the extent of differential pricing that is necessary to ensure the continued investment needed to maintain the future health of the industry.

In examining economies of scale in the railroad industry, it is important to distinguish between two different types of scale economies that may exist. **Economies of density** refer to cost savings resulting from transporting more traffic over a fixed network, while **economies of size** refer to cost savings resulting from transporting more traffic as the network size expands.[22] Because this study focuses on examining the need for differential pricing over current railroad networks, and not on cost implications of network expansion, the relevant measure of economies of scale is economies of density; that is, how railroad costs change as traffic is expanded over current networks. To the extent that average costs decline as traffic is expanded over existing networks, this suggests that marginal costs are below average costs and that marginal cost pricing will not recover the total costs of operation. Thus, to the extent that there are differences in the elasticity of demand for various shipments, second-best pricing requires differential pricing.

In order to examine the extent of economies of density in the U.S. railroad industry, this study uses railroad financial data to estimate a short-run variable cost function. Not only does this study use more recent data than any previous study analyzing costs in the railroad industry, but it also more accurately captures the multiproduct nature of railroads by including ton-miles hauled of the four largest commodities in terms of tonnage over the last 33 years.

[22] This distinction was pointed out by Keeler in 1974.

The estimated short-run variable cost function, includes four factor prices: (1) the price of labor, (2) the price of equipment (rail cars and locomotives), (3) the price of fuel, and (4) the price of materials. (See the sidebar for the general form of the short-run variable cost function using mathematical notation.)

Generalized Short-run Variable Cost Function

The generalized short run variable cost function is defined as follows:

$$C^{SR} = C^{SR}(w, k, y, t) \qquad (27)$$

where C^{SR} represents short-run variable costs, w is a vector of factor prices, k is the fixed factor, y is a vector of outputs, and t is a vector of technological attributes (including a time trend).

Since the cost function estimated is short run, it also includes the amount of capital employed (the fixed input), measured as way and structures capital per route mile. The model also includes technological attributes to account for differences in network structures, service obligations, and the nature of service provided. These attributes include route-miles, average length of haul, and a time trend. Although some studies have erroneously used route miles as a measure of capital, it is important to note that route miles represent the extent of the carrier's network and an opportunity to serve more shippers.[23] Average length of haul is included to account for the increased efficiency of longer hauls due to the spreading of fixed terminal costs over more miles. The time trend is included to account for technological change over time.

The biggest innovation this cost function includes is in its measurement of multiple output variables. Previous studies have represented the output produced by railroads using revenue ton-miles. A revenue ton-mile is measured as one ton of a commodity hauled for one mile.[24] While revenue ton-miles seem to be the best measure of output for railroad services, it is well known that revenue ton-miles are not a homogenous output. Railroads are truly multiproduct firms, carrying a variety of products from a variety of origins to a variety of destinations. Technically, each product carried from each origin to each destination represents a different railroad output.

Given the lack of data on individual railroad shipments of different commodities from different origins to different destinations, and the infeasibility of measuring relationships between each of these outputs and costs, researchers have made some innovations that attempted to capture more of the multiproduct nature of railroads. These innovations have included capturing the percent of ton-miles in various types of service (e.g. way/through trains vs. unit trains), the percent of tons accounted for by different commodities, or the quantities of car-miles accounted for by different car

[23] According to economic theory, in the short run a firm (railroad in this case) chooses variable inputs to minimize costs, knowing that there is some input that is fixed (usually capital). In the long run, the firm adjusts its fixed factor (capital) to the amount that is necessary to minimize costs, given the output it is producing. Suppose that a railroad operates only from Minneapolis to Chicago, and it realizes a big increase in its traffic. If the railroad believes the increase in traffic is long-lived, it is likely to increase its capital stock to accommodate the increase in traffic at the lowest possible cost. The railroad could do so by building additional side-by-side track or by improving the quality of the track, but it would not do so by adding a route from Chicago to Milwaukee. This illustrates why additional route miles enable service to more markets, but they are not a measure of fixed capital that can be adjusted to minimize costs in the long run.

[24] The term revenue ton-mile is used because these are ton-miles in revenue service. They do not include the weight of the equipment or empty mileage.

types.[25] However, no previous studies have explicitly used railroad ton-miles accounted for by various commodities as individual outputs.

This study uses five output variables: revenue ton-miles of coal, revenue ton-miles of chemicals, revenue ton-miles of farm products, revenue ton-miles of non-metallic minerals, and all other revenue ton-miles carried. The four commodity-specific output variables are the four largest carried by Class I railroads in terms of tonnage over the last 33 years, accounting for 64 percent of all tonnage carried. This separation of outputs allows us to capture the varying impacts carrying different commodities has on costs resulting from different shipment sizes, different types of equipment, different shipment lengths, different levels of fragility of products carried, and other factors.

To estimate the generalized short-run variable cost function, this study uses the translog functional form. The translog cost function is a second-degree polynomial in logs of the variables. **Appendix A** includes the detailed cost function specification, and econometric details.

Data

To estimate the short-run variable cost function, data obtained from each Class I railroad's Annual Reports (also known as R1 data) to the Surface Transportation Board (STB) are used from 1984 through 2016. Data from these reports are used to calculate input prices, the amount of way and structures capital, route miles, and average length of haul on a particular railroad's system.

Ton-miles by two-digit Standard Transportation Commodity Code (STCC) are obtained using a combination of data obtained from the Class I Annual Reports and from the Surface Transportation Board's confidential Carload Waybill Sample. Specifically, commodity tonnage data from the Class I Railroad Annual Report is multiplied by commodity-specific average length of haul from the Waybill Sample for each railroad and year. A detailed description of variables and data sources used is included in **Table A1** in **Appendix A**.

Table A2 of **Appendix A** shows the railroads and years included in the dataset. Because of mergers, a series of firm dummies are included. Each railroad has a dummy for the original pre-merged firm. This dummy maintains a value of one for the merged firm, as well. In addition, a new dummy is created for each merged firm that has a value of zero prior to the merger. This approach to including dummies ensures that the unobserved characteristics of the original firms, as well as those unique to the newly merged system, are captured in the analysis.

Table 1 shows descriptive statistics of the variables included in the short-run variable cost function. As the table shows, there are large differences between the largest and smallest railroads in the sample, with the largest railroad in terms of route miles comprising more than 35,000 miles and the smallest 442. Similarly, the largest railroad in terms of coal ton-miles hauls 295 billion coal ton-miles, while the smallest only hauls 5 million coal ton-miles.

[25] For example, Bitzan (2003), Bitzan and Keeler (2003), Bitzan and Wilson (2007), and Wilson (1997) account for differences in types of services; Ivaldi and McCullough (2007) and Bitzan and Keeler (2011 and 2014) account for different types of car miles; and Friedlaender and Spady (1981) account for traffic mix.

Table 1: Descriptive Statistics of Data Used in Short-Run Variable Cost Function*				
	Mean	Standard Dev.	Minimum	Maximum
Variable Cost	$3,707,470,743	$3,606,876,114	$136,277,457	$14,326,081,542
Labor Price	$37.71	$5.43	$16.52	$57.84
Equipment Price	$57,250	$35,378	$4,320	$176,431
Fuel Price	$1.34	$0.69	$0.58	$3.34
Materials Price	$266.86	$43.79	$209.83	$379.51
W&S Capital per Route Mile	$664,797	$359,144	$62,357	$1,737,271
Coal Ton-Miles	48,423,004,735	75,058,658,450	4,645,318	295,195,133,246
Chemicals Ton-Miles	13,900,902,522	15,859,156,501	30,873,169	78,640,012,438
Farm Products Ton-Miles	13,459,116,720	19,213,869,046	16,846,281	104,035,428,747
Nonmetallic Minerals Ton-Miles	3,931,262,959	5,125,323,758	15,086,344	37,706,936,051
Other Ton-Miles	57,028,314,392	64,843,767,708	1,342,778,970	289,900,626,778
Route Miles	12,413	10,511	442	35,208
Average Length of Haul	499.44	234.88	175.11	1139.79
Time	13.44	9.84	1.00	33.00
*All monetary variables are in 2009 prices				

Results: Cost Estimation

Table 2 presents the first-order terms of the estimation results for the short-run variable cost function (full results are shown in **Table A3** of **Appendix A**).[26] As the table shows, all first order terms have their expected signs, and all but one are significant at conventional levels. Because all variables are divided by their mean values, and because they are in natural logarithms, all first-order terms can be interpreted as the elasticity of cost with respect to that variable at the means of all variables.

For example, the elasticity of cost with respect to labor price is .395, meaning that a one percent increase in labor price would lead to a .395 percent increase in costs, if all other variables were at their mean. For factor prices, these elasticities also show the factor's share of variable costs, meaning that labor accounts for about 40 percent of variable costs for the average railroad.

[26] The cost function is increasing in factor prices, continuous in factor prices by assumption, and concave in factor prices at the means of all variables. However, in testing for concavity in factor prices for individual observations, the condition is met for 34.9% of the observations (116 out of 332 observations). As noted by Pels and Rietveld (2008) failure to find global concavity is common in empirical studies. To test for concavity in factor prices, the characteristic root of the Hessian matrix are taken for every observation in the sample. If all characteristic roots are non-positive, this suggests the Hessian matrix is negative semi-definite, and therefore, the cost function is concave in factor prices.

Table 2: Estimation Results – Short-Run Variable Cost Function (First-Order Terms)		
Variable	**Parameter Estimate**	**Standard Error**
Intercept	21.97686*	0.0839
w_L (Labor Price)	0.394859*	0.00390
w_E (Equipment Price)	0.150928*	0.00324
w_F (Fuel Price)	0.134597*	0.00200
k (Way and Structures Capital per Mile)	-0.15814**	0.0657
RTM_{COAL} (Coal Revenue Ton-Miles)	0.086642*	0.0288
RTM_{CHEM} (Chemicals Revenue Ton-Miles)	0.100709***	0.0556
RTM_{FARM} (Farm Products Revenue Ton-Miles)	0.098845*	0.0376
RTM_{NONMET} (Nonmetallic Minerals Revenue Ton-Miles)	0.091412*	0.0326
RTM_{OTH} (Other Revenue Ton-Miles)	0.36618*	0.0574
RM (Route Miles)	0.269533***	0.1391
ALH (Average Length of Haul)	-0.1172	0.1270
T (Time)	-0.2724*	0.0452
All variables are in natural logarithms, except the intercept. # of observations = 332 *significant at the 1% level, **significant at the 5% level, ***significant at the 10% level Adj. R^2 Cost = .9966, Adj. R^2 Labor Share = .6949, Adj. R^2 Equip Share = .2534, Adj R^2 Fuel Share = .8781		

Interestingly, the output variables show elasticities that are fairly similar at the means of all variables; the exception is for other commodities (RTM_{OTH}), which has an elasticity of .37. Although the elasticities are similar at the point of means, they vary more by commodity for the average railroad today (as shown in the next section). Differences in elasticities may reflect differences in costs of carrying various commodities due to differences in shipment characteristics and due to differences in the density of the routes such commodities travel. Commodities traveling on more densely traveled routes will have higher elasticities, as cost savings from additional traffic tend to diminish with more traffic on the route.

The sum of output elasticities at the point of means is 0.7438, suggesting significant short-run economies of density. A one percent increase in all ton-miles (output) leads to a .74 percent increase in variable costs. The amount of way and structures capital per route mile has the expected negative sign, suggesting that capital is productive. This means an increase in capital decreases variable costs. In addition, costs (1) increase with more route miles due to an increased network size, and (2) decrease with longer average lengths of haul due to reductions in terminal costs per mile with increased distance. The time trend also shows that real variable costs have been decreasing over time. The following section explores the results regarding returns to density in more detail, including examining implications for differential pricing.

Results: Returns to Density

As shown in the previous section, the estimated cost function shows strong short-run returns to density. However, in making an assessment of returns to density, it is the long run that is relevant. Friedlaender and Spady (1981) and Caves, Christensen, and Swanson (1981) show that long-run elasticity of costs with respect to output (and therefore, returns to density) can be derived from a short-run variable cost function using the following formula:

$$\varepsilon_C{}^{LR} = \sum_i \frac{\partial ln\, C^{LR}}{\partial ln\, Q_i} = \sum_i \frac{\partial ln\, C^{SR}}{\partial ln\, Q_i}\; x \left[\frac{1}{1-\partial ln\, C^{SR}/\partial ln\, k}\right] \quad (28)$$

Friedlaender and Spady (1981) note that this measure of long-run returns to scale assumes that railroads are operating at a point of long-run equilibrium; that is, at a point where the reduction in variable costs from using another unit of capital are equal to the increase in fixed costs from using another unit. This condition is shown in the following equation:

$$\frac{\partial\, C^{SR}}{\partial\, k} = -w_k \quad (29)$$

Since the percentage change in variable costs from a one percent change in way and structures capital has been estimated in the seemingly unrelated system above, it is possible to obtain the change in variable costs resulting from a one dollar change in the amount of way and structures capital employed using the following equation:

$$\frac{\partial C^{SR}}{\partial k} = \frac{\partial ln\, C^{SR}}{\partial ln\, k}\; x\; \frac{C^{SR}}{k} \quad (30)$$

At the means of all variables, this implies that a one dollar increase in way and structures capital per mile would decrease variable costs by $881.92 (in 2009 prices). This is shown as follows:

$$\frac{\partial C^{SR}}{\partial k} = \frac{\partial ln\, C^{SR}}{\partial ln\, k}\; x\; \frac{C^{SR}}{k} = -.15814\; x\; \frac{3{,}707{,}470{,}743}{664{,}797} = -881.92 \quad (31)$$

With the average railroad comprising 12,413 miles, this suggests that an increase in way and structures investment of $12,413 would yield a variable cost savings of $881.92. This implies that the investment in way and structures capital is optimal if the opportunity cost of capital is 7.1 percent ($881.92/$12,413). This is lower than the Surface Transportation Board's recent estimates of the industry cost of capital (ranging between 8.88 percent and 11.75 percent over the last twelve years), implying that railroads may have some overinvestment in way and structures capital, and not be at a point of long-run equilibrium. However, given the controversy and uncertainty about the true cost of capital in the industry, long-run returns to scale are estimated using the equation given above. The implied long-run elasticity of costs with respect to output is 0.642 at the point of means. As shown below, this value says a 1 percent increase in outputs is associated with a 0.64 percent increase in costs. This is less than one, which points to economies of density.

$$\sum_i \frac{\partial ln C^{LR}}{\partial ln Q_i} = \frac{0.7438}{1.15814} = 0.642 \quad (32)$$

Table 3 shows estimated short-run and long-run output elasticities at the yearly means of all variables. As the table shows, U.S. Class I railroads realize large scale elasticities in all years. This suggests that railroad firms would fall well short of recovering costs if they priced at marginal costs. As shown in **Appendix C**, given the degree of scale economies realized in the railroad industry substantial markups are needed on captive traffic to allow railroads to generate a rate of return adequate to continue to attract investment. Although the extent of scale economies realized in the industry has diminished slightly over time, they are still large. The next section presents a summary and implications.

Table 3: Output Elasticities and Elasticity with Respect to Capital at Yearly Means of All Variables (standard errors in parentheses)

Year	Coal	Chemicals	Farm Products	Nonmetallic Minerals	Other	SR Output Elast	W & S Capital	LR Output Elast
1984	0.0869** (0.0368)	0.1644* (0.0535)	0.0769*** (0.0465)	0.0962*** (0.0501)	0.1827** (0.0830)	0.6071	-0.0659 (0.0928)	0.5696
1985	0.0927** (0.0361)	0.1968* (0.0493)	0.0779** (0.0331)	0.1093* (0.0390)	0.1607* (0.0573)	0.6374	-0.0093 (0.0730)	0.6316
1986	0.0606*** (0.0316)	0.1341* (0.0426)	0.0820* (0.0314)	0.0578 (0.0389)	0.3653* (0.0510)	0.6998	-0.1147*** (0.0619)	0.6278
1987	0.0664** (0.0292)	0.1174* (0.0417)	0.0840** (0.0325)	0.0591 (0.0371)	0.3801* (0.0519)	0.7070	-0.1343** (0.0600)	0.6233
1988	0.0756** (0.0303)	0.1606* (0.0438)	0.0729** (0.0306)	0.0726** (0.0319)	0.2822* (0.0461)	0.6639	-0.0731 (0.0552)	0.6187
1989	0.0798* (0.0290)	0.1358* (0.0423)	0.0834* (0.0297)	0.0806* (0.0305)	0.3120* (0.0449)	0.6916	-0.0902*** (0.0544)	0.6344
1990	0.0809* (0.0290)	0.1331* (0.0442)	0.0838* (0.0299)	0.0812* (0.0303)	0.3202* (0.0444)	0.6991	-0.0957*** (0.0540)	0.6380
1991	0.0802* (0.0293)	0.1338* (0.0430)	0.0756** (0.0300)	0.0861* (0.0294)	0.3090* (0.0589)	0.6846	-0.1025*** (0.0538)	0.6210
1992	0.0848* (0.0297)	0.1179** (0.0459)	0.0901* (0.0326)	0.0942* (0.0311)	0.3519* (0.0501)	0.7391	-0.1343** (0.0604)	0.6515
1993	0.0707** (0.0277)	0.0982** (0.0486)	0.0913* (0.0322)	0.0720** (0.0287)	0.3616* (0.0478)	0.6939	-0.1085*** (0.0567)	0.6259
1994	0.0655** (0.0281)	0.0880*** (0.0530)	0.0978* (0.0350)	0.0686** (0.0301)	0.3737* (0.0511)	0.6937	-0.1190** (0.0597)	0.6200
1995	0.0723* (0.0275)	0.1091** (0.0474)	0.0784** (0.0334)	0.0749* (0.0284)	0.3659* (0.0485)	0.7005	-0.1216** (0.0551)	0.6246
1996	0.0663** (0.0291)	0.0715 (0.0570)	0.1041* (0.0391)	0.0664*** (0.0354)	0.4370* (0.0620)	0.7452	-0.1886* (0.0695)	0.6270
1997	0.0667** (0.0297)	0.1023*** (0.0585)	0.0924** (0.0381)	0.0672** (0.0330)	0.3906* (0.0574)	0.7192	-0.1536** (0.0626)	0.6234
1998	0.0737** (0.0306)	0.1228** (0.0587)	0.0905** (0.0379)	0.0853* (0.0317)	0.3561* (0.0566)	0.7285	-0.1408** (0.0621)	0.6386
1999	0.0771** (0.0323)	0.1080*** (0.0642)	0.1188* (0.0418)	0.0868** (0.0357)	0.3971* (0.0657)	0.7878	-0.1874* (0.0719)	0.6635

Table 3: Output Elasticities and Elasticity with Respect to Capital at Yearly Means of All Variables (standard errors in parentheses)								
Year	Coal	Chemicals	Farm Products	Nonmetallic Minerals	Other	SR Output Elast	W & S Capital	LR Output Elast
2000	0.0813** (0.0317)	0.0940 (0.0654)	0.1235* (0.0427)	0.0865** (0.0360)	0.4093* (0.0670)	0.7946	-0.1939* (0.0730)	0.6656
2001	0.0783** (0.0319)	0.0932 (0.0650)	0.1232* (0.0424)	0.0816** (0.0355)	0.4189* (0.0668)	0.7952	-0.1922* (0.0720)	0.6670
2002	0.0733** (0.0309)	0.1158*** (0.0654)	0.1080* (0.0413)	0.0730** (0.0325)	0.3806* (0.0602)	0.7508	-0.1501** (0.0613)	0.6528
2003	0.0731** (0.0305)	0.0973 (0.0663)	0.1136* (0.0422)	0.0700** (0.0334)	0.4085* (0.0628)	0.7625	-0.1691* (0.0641)	0.6522
2004	0.0699** (0.0303)	0.0974 (0.0684)	0.1033** (0.0431)	0.0624*** (0.0334)	0.3913* (0.0626)	0.7243	-0.1533** (0.0640)	0.6280
2005	0.0733** (0.0301)	0.0845 (0.0689)	0.1081** (0.0441)	0.0652*** (0.0341)	0.4103* (0.0650)	0.7414	-0.1715** (0.0667)	0.6329
2006	0.0786* (0.0301)	0.0773 (0.0690)	0.1075** (0.0450)	0.0712** (0.0345)	0.4142* (0.0672)	0.7488	-0.1806** (0.0702)	0.6342
2007	0.0792* (0.0301)	0.0883 (0.0683)	0.1068** (0.0437)	0.0733** (0.0321)	0.3856* (0.0638)	0.7331	-0.1513** (0.0661)	0.6368
2008	0.0862* (0.0304)	0.1007 (0.0678)	0.1022** (0.0431)	0.0802* (0.0304)	0.3691* (0.0624)	0.7384	-0.1423** (0.0647)	0.6464
2009	0.0649** (0.0322)	0.1499** (0.0678)	0.0769*** (0.0404)	0.0534*** (0.0299)	0.3298* (0.0558)	0.6749	-0.0476 (0.0542)	0.6443
2010	0.0770** (0.0315)	0.1458** (0.0692)	0.0764*** (0.0417)	0.0633** (0.0295)	0.30328 (0.0559)	0.6657	-0.0594 (0.0581)	0.6284
2011	0.0867* (0.0307)	0.1310*** (0.0703)	0.0839*** (0.0431)	0.0739** (0.0297)	0.3123* (0.0583)	0.6878	-0.0832 (0.0610)	0.6350
2012	0.0857* (0.0299)	0.1141 (0.0714)	0.0943** (0.0441)	0.0741** (0.0297)	0.3329* (0.0598)	0.7012	-0.0964 (0.0606)	0.6395
2013	0.0911* (0.0298)	0.1087 (0.0735)	0.0969** (0.0459)	0.0789** (0.0306)	0.3253* (0.0621)	0.7008	-0.0952 (0.0648)	0.6399
2014	0.1037* (0.0299)	0.1036 (0.0743)	0.0996** (0.0481)	0.0960* (0.0335)	0.3203* (0.0680)	0.7232	-0.1273*** (0.0754)	0.6415
2015	0.0980* (0.0304)	0.1379*** (0.0728)	0.0788*** (0.0461)	0.0894* (0.0309)	0.2709* (0.0626)	0.6750	-0.0753 (0.0712)	0.6277
2016	0.0865* (0.0304)	0.1518** (0.0710)	0.0673 (0.0449)	0.0639** (0.0293)	0.2674* (0.0584)	0.6370	-0.0410 (0.0679)	0.6119
*significant at the 1% level, **significant at the 5% level, ***significant at the 10% level								

Summary and Implications

Recent recommendations in the way railroad rates are regulated have been advocated by the National Academy of Sciences (TRB, 2015) and by the STB's Rate Reform Task Force (2019). These proposed changes are the result of recent concerns over increasing rail rates since 2003, an increasingly revenue adequate railroad industry, concerns over poor railroad service, and a concern that the costs of pursuing rate cases are too high (especially for smaller shippers).

In assessing the desirability of various recommended changes, it is important to consider the degree of scale (density) economies realized in the industry, and consequently the need for differential pricing. This study finds that while economies of scale (density) have declined slightly over time, there are still substantial economies of scale (density) in the U.S. railroad industry.

Estimation of "Modified Polar Ramsey Markups" (**Appendix C**), shows that while scale economies have not changed much over time, the average markup that railroads would need to charge relatively captive traffic to continue to attract investment has declined in recent years. Due to an increase in the average markup on traffic with revenue-to-URCS variable cost ratios below 1.8 and an increase in the proportion of traffic moving at revenue-to-variable cost ratios of 1.8 or greater, the average revenue burden needed from relatively captive traffic to guarantee revenue adequacy has declined. This has corresponded with more railroads achieving revenue adequacy in recent years.[27]

Nonetheless, the persistence of scale (density) economies in the industry suggests that extensive differential pricing is still necessary. Moreover, as shown by a variety of other studies (e.g. Gallamore, 1999, Bitzan and Keeler, 2007 and 2011, Morrison and Winston, 1999), the increased pricing flexibility afforded railroads as a result of deregulation has resulted in innovation and cost savings that have benefited shippers.

This suggests that policymakers should be cautious in implementing policies that limit differential pricing in the industry. Certainly, there is room to improve rate relief processes for relatively captive shippers, including making it easier to contest rates that might be unreasonably high. However, policies that attempt to make broad changes in the extent of differential pricing have the potential to limit industry investment, and the cost saving innovations and improved service quality that comes with it. In the context of the significant scale economies that exist in the railroad industry, as long as the prices charged to shippers with many transportation alternatives are above the incremental costs of providing services to those shippers, "captive shippers" benefit through higher quality service and increased capacity enabled through the ability of railroads to attract investment.

[27] See STB, Ex Parte 552, Railroad Revenue Adequacy. Since captivity is relative (and not absolute), this makes intuitive sense. In years with high "Modified Polar Ramsey Markups," the theoretical average revenue burden placed on relatively captive traffic (traffic with R/VC $\geq$1.8) was higher than such shippers were willing to pay. As a result, railroads were unable to achieve revenue adequacy. More recently, the lower theoretical average revenue burden placed on such shippers was likely within the limits that they were willing to pay.

In summary, scale (density) economies in U.S. railroad industry persist, though they have declined slightly over time. The large scale economies that exist suggest that marginal cost pricing would not come close to recovering railroad costs and that differential pricing is needed to ensure continued industry investment and innovation. The recent realization of revenue adequacy by many of the nation's railroads (along with the smaller average markup needed from "relatively captive shippers" to achieve revenue adequacy as demonstrated by the "Modified Polar Ramsey Markups"), is one manifestation of the benefits that differential pricing has had for the industry.

References

Baumol, William J., John C. Panzar, and Robert D. Willig. *Contestable Markets and the Theory of Industry Structure*. Harcourt Brace Jovanovich, New York, 1988.

Baumol, William J. and David F. Bradford. "Optimal Departures from Marginal Cost Pricing," *American Economic Review*, 60(3) (1970), pp. 265-283.

Berndt, E.R., Friedlaender, A.F., Wang Chiang, J.S., and C.A. Vellturo. "Cost Effects of Mergers and Deregulation in the U.S. Rail Industry." *Journal of Productivity Analysis*, 4 (1994), pp.127-44.

Bitzan, John D., "Railroad Costs and Competition: The Implications of Introducing Competition to Railroad Networks, *Journal of Transport Economics and Policy* 37 (May 2003), pp. 201-225.

Bitzan, John D. and Theodore E. Keeler, "Economies of Density and Regulatory Change in the U. S. Railroad Freight Industry," *Journal of Law and Economics*, 50 (February 2007), pp156-179.

Bitzan, John D. and Theodore E. Keeler, "Intermodal Traffic, Regulatory Change and Carbon Energy Conservation in U.S. Freight Transport," *Applied Economics* 43(August 2011), pp. 3945-3963.

Bitzan, John D. and Theodore E. Keeler, "Productivity Growth and Some of its Determinants in the Deregulated U. S. Railroad Industry," *Southern Economic Journal* 70 (October 2003), pp. 232-253.

Bitzan, John D. and Theodore E. Keeler, "The Evolution of U.S. Rail Freight Pricing in the Post-Deregulation Era: Revenues versus Marginal Costs for Five Commodity Types," *Transportation* (2014) pp. 305-324.

Bitzan, John D. and Wesley W. Wilson, "Industry Costs and Consolidation: Efficiency Gains and Mergers in the U.S. Railroad Industry," *Review of Industrial Organization* 30 (2007), pp. 81-105.

Borts, George H. "Increasing Returns in the Railway Industry, *Journal of Political Economy* 62 (1954), pp. 316-333.

Borts, George H. "The Estimation of Rail Cost Functions," *Econometrica* 28(1) (1960, pp. 108-131.

Braeutigam, Ronald R., Daughety, Andrew F., and Mark A. Turnquist. "A Firm Specific Analysis of Economies of Density in the Railroad Industry." *The Journal of Industrial Economics*, 33(1) (1984), pp. 3-20.

Brown, Randal S., Caves, Douglas W., and Laurits R. Chistensen, "Modelling the Structure of Cost and Production for Multiproduct Firms," *Southern Economic Journal* 46(1) (1979), pp. 256-273.

Caves, Douglas W., Christensen, Laurits R., and Joseph A. Swanson, "Productivity Growth, Scale Economies, and Capacity Utilization in U.S. Railroads, 1955-74," *American Economic Review*, 71 (December 1981), pp. 994-1002.

Caves, Douglas W., Christensen, Laurits R., and Michael W. Tretheway," Flexible Cost Functions for Multiproduct Firms," *Review of Economics and Statistics* 62(3) (1980), pp. 477-481.

Caves, Douglas W., Christensen, Laurits R., Tretheway, Michael W., and Robert J. Windle (1985). "Network Effects and the Measurement of Returns to Scale and Density," in A.F. Daughety, ed., *Analytical Studies in Transport Economics.* Cambridge: Cambridge University Press, 1985.

Christensen, Laurits R., Jorgenson, Dale W., and Lawrence J. Lau, "Transcendental Logarithmic Production Frontiers," *Review of Economics and Statistics* 55(1) (1973), pp. 28-45.

Diewert, Walter, "An Application of Shephard Duality Theorem: A Generalized Leontief Production Function," *Journal of Political Economy* 79(3) (1971) pp. 481-507.

Faulhaber, Gerald R. "Cross-Subsidization: Pricing in Public Enterprises," *American Economic Review*, 65(5) (1975), pp. 966-977.

Friedlaender, Ann F. "Coal Rates and Revenue Adequacy in a Quasi-Regulated Rail Industry," *Rand Journal of Economics* 23 (Autumn 1992), pp. 376-394.

Friedlaender, Ann F., Berndt, Ernst R., Chiang, Judy Shaw-Er Wang, Showalter, Mark, and Christopher A. Vellturo, "Rail Costs and Capital Adjustments in a Quasi-Regulated Environment," *Journal of Transport Economics and Policy* 27(2) (1993), pp. 131-152.

Friedlaender, Ann F., Chaing Judy Shaw-Er Wang, and Christopher A. Velturo, "Cost Effects of Mergers and Deregulation in the U.S. Rail Industry," *Journal of Productivity Analysis* 4(1-2) (1993), pp. 137-144.

Friedlaender, A.F. and R. H. Spady. *Freight Transport Regulation: Equity, Efficiency, and Competition in the Rail and Trucking Industries*. Cambridge, MA: The MIT Press, 1981.

Gallamore, Robert E., "Regulation and Innovation: Lessons from the American Railroad Industry," in Jose Gomez-Ibanez, William B. Tye, and Clifford Winston (Eds), *Essays in Transportation Economics and Policy: A Handbook in Honor of John R. Meyer*, Brookings Institution Press, Washington, D.C., 1999.

Grimm, Curtis and Clifford Winston, "Competition in the Deregulated Railroad Industry: Sources, Effects, and Policy Issues," in Peltzman, Sam and Clifford Winston, eds. *Deregulation of Network Industries: What's Next?*. Brookings Institution Press, American Enterprise Institute, 2000.

Harris, Robert .G. "Economies of Density in the Railroad Freight Industry." *Bell Journal of Economics*, 8(2) (1977), pp. 556-64.

Interstate Commerce Commission, "Coal Rate Guidelines – Nationwide," Ex Parte No. 347 (sub-no. 1), 1985.

Ivaldi, Marc and Gerard McCullough, "Density and Integration Effects on Class I U.S. Freight Railroads, " Journal of Regulatory Economics 19(2) (2001), pp. 161-182.

Ivaldi, Marc and Gerard McCullough, "Railroad Pricing and Revenue-to-Cost Margins in the Post-Staggers Era," *Research in Transportation Economics* 20 (2007), pp. 153-178.

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

Keeler, Theodore E. *Railroads, Freight, and Public Policy*. The Brookings Institution, Washington, D.C., 1983.

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

Miller, Edward, "Economies of Scale in Railroading," *Proceedings, Transportation Research Forum* 14 (1973), pp. 683-701.

Morrison, Steven A, and Clifford Winston, "Regulatory Reform of U.S. Intercity Transportation," in Jose Gomez-Ibanez, William B. Tye, and Clifford Winston (Eds), *Essays in Transportation Economics and Policy: A Handbook in Honor of John R. Meyer*, Brookings Institution Press, Washington, D.C., 1999.

National Academy of Sciences, *Modernizing Freight Rail Regulation, Special Report 318*, Transportation Research Board, Washington, D.C., 2015.

Oum, Tae H. and William G. Waters II, "A Survey of Recent Developments in Transportation Cost Function Research," Logistics and Transportation Review 32 (1996), pp. 423-463.

Pels, Eric, Rietveld, Piet, "Cost functions in transport," in: Hensher, D., Button, K. (eds.) *Handbook of transport modelling, 2nd edn*. Bingley, Emerald, 2008.

Rate Reform Task Force, *Report to the Surface Transportation Board*, April 25, 2019.

Rhodes, George F. and M. Daniel Westbrook, "Econometric Analysis of Costing System Components in Rail Rate Regulation," *Journal of Business & Economic Statistics* 4(3) (1986, pp. 289-303.

Schmalensee, Richard L. and Wesley W. Wilson, "Modernizing U.S. Freight Regulation," *Review of Industrial Organization*, 49 (2016), pp. 133-159.

Sharkey, William W. *The Theory of Natural Monopoly*, Cambridge University Press, New York, 1982.

Shephard, Ronald W. *Cost and Production Functions*, Princeton University Press, Princeton, N.J., 1953.

Surface Transportation Board, "Expanding Access to Rate Relief," Ex Parte 665 (sub-no. 2), 2018.

Surface Transportation Board, "Expediting Rate Cases, Ex Parte 733, 2018.

Surface Transportation Board, "Rail Transportation of Grain, Rate Regulation Review," Ex Parte 665 (sub-no. 1), 2016.

Surface Transportation Board, *Surface Transportation Board Report to Congress Regarding the Uniform Rail Costing System*, May27, 2010.

Surface Transportation Board, "Simplified Standards for Rail Rate Cases," Ex Parte No. 646 (sub-no. 1), 2007.

Surface Transportation Board, U.S. Department of Transportation, *Class I Annual Reports (R1)*, various years.

Waters, William G. and A.D. Woodland, *Econometric Analysis and Railway Costing*, North Oxford Academic Publishing, Oxford, England, 1984.

Willig, Robert D. and William J. Baumol, "Railroad Deregulation: Using Competition as a Guide," Regulation 1 (1987), pp. 28-35.

Wilson, Wesley W., "Cost Savings and Productivity in the Railroad Industry," *Journal of Regulatory Economics* 11 (1997), pp. 21-40.

Wilson, Wesley W. and Frank A. Wolak, "Freight Rail Costing and Regulation: The Uniform Rail Costing System," *Review of Industrial Organization* 49 (September 2016), pp. 229-261.

Glossary

accounting costs – explicit financial statement value of expenses incurred in doing business.

average costs – total costs divided by the amount of output produced.

average fixed costs – total fixed costs divided by amount of output produced.

average total costs – total costs divided by the amount of output produced.

average variable costs – total variable costs divided by the amount of output produced.

common costs – costs that cannot be attributed to particular products or services.

consumer's surplus – the value placed on a good or service by all consumers in excess of the price that they have to pay for it.

deadweight loss – reduction in social welfare from an inefficient allocation of resources.

economic cost – a cost concept based on the principle of opportunity costs – that is, the value of a resource in its best alternative use.

economies of density – reductions in average cost resulting from increased traffic over a network of a given size.

economies of scale – reductions in average cost from producing more output.

economies of scope – cost savings from producing more than one type of output.

economies of size – reductions in average cost resulting from increased traffic due to expansions in the size of the network.

first best pricing – pricing that maximizes social welfare.

fixed costs – costs that do not vary with output.

fixed inputs – inputs that cannot be adjusted in the short run.

fully-allocated cost – an accounting concept meant to approximate average cost.

incremental cost – the change in total cost resulting from some larger change in output.

inputs – the factors of production (e.g. labor, fuel, materials) needed to produce a given amount of the firm's product or service (output).

long run – a period of time where all inputs can be adjusted.

long-run cost function – shows the minimum cost of producing any output level, given input prices.

marginal cost – the change in total cost resulting from a one-unit change in output.

marginal revenue – the change in total revenue from a one-unit change in quantity sold.

natural monopoly – a market, where the product or service can be provided at a lower cost by one firm than by more than one firm.

output – the quantity of the firm's product or service that it produces.

price elasticity of demand – the percentage change in quantity demanded resulting from a one percent change in price.

producer's surplus – total revenues received by producers in selling the good or service in excess of the costs of producing the good or service.

production function – shows the maximum amount of output that can be produced with different quantities of inputs.

railroad costing – measuring the relationships between specific railroad cost accounts and activity measures in order to measure the costs of specific rail movements.

Ramsey pricing – a second best solution to maximizing social welfare, where prices are set inversely to elasticity of demand; a price is charged in each market so that it reduces output by the same proportion in each market in comparison to the output that would be produced if price were set equal to marginal cost.

second best – solution that maximizes social welfare, subject to allowing the firm to break even.

short run – a period of time when at least one of the inputs of the firm is fixed.

short-run conditional input demand functions – cost minimizing quantities of inputs for any output, given input prices and the quantity of the fixed factor employed.

short-run cost function – the minimum cost of producing any output level, given input prices and the quantity of the fixed factor employed.

social welfare – the value placed on goods and services by society in excess of the costs of resources used to produce those goods and services.

sunk costs – costs that are incurred and cannot be recovered.

transformation function – shows the maximum possible vector of outputs that can be produced with given quantities of inputs.

variable costs – costs that vary directly with output.

variable inputs – inputs that can be adjusted, even in the short run.

Appendix A – Data Definitions, Firms, and Cost Function Specification

Table A1: Data Definitions and Sources Used to Estimate the Railroad Cost Function*

Variable	Source
Cost Variable and Construction	
Real Variable Cost	(OPERCOST-CAPEXP-ANNDEPRD +ROILCM+ROICRS)/GDPPD
OPERCOST	Railroad Operating Cost (R1, Sched. 410, ln. 620, Col F)
CAPEXP	Capital Expenditures Classified as Operating in R1 (R1, Sched 410, lines 12-30, 101-109, Col F)
ANNDEPRD	Annual Depreciation of Road (R1, Sched 335, line 30, Col C)
COSTKAP	Pre-Tax Cost of Capital (AAR *Railroad Facts*)
ROILCM	Return on Investment in Locomotives [(IBOLOCO+LOCINVL)-(ACDOLOCO+LOCACDL)]*COSTKAP
IBOLOCO	Investment Base in Owned Loc. (R1, Sched 415, line 5, Col. G)
LOCINVL	Investment Base in Leased Loc. (R1, Sched 415, line 5, Col. H)
ACDOLOCO	Accum. Depr. Owned Loc. (R1, Sched 415, line 5, Col. I)
LOCACDL	Accum. Depr. Leased Loc. (R1, Sched 415, line 5, Col. J)
RENTLOCO	Lease/Rental Payments Locomotives (R1, Sched 415, Line 5, Col. F)
ROICRS	Return on Investment in Cars [(IBOCARS+CARINVL)-(ACDOCARS+CARACDL)]*COSTKAP
IBOCARS	Investment Base in Owned Cars (R1, Sched 415, line 24, Col. G)
CARINVL	Investment Base in Leased Cars (R1, Sched 415, line 24, Col. H)
ACDOCARS	Accum. Depr. Owned Cars (R1, Sched 415, line 24, Col. I)
CARACDL	Accum. Depr. Leased Loc. (R1, Sched 415, line 24, Col. J)
RENTCARS	Lease/Rental Payments Freight Cars (R1, Sched 415, Line 24, Col. F)
Output Variables	
Coal Revenue Ton-Miles	(QCS, line corresponding to STCC 11, Col. K) x Coal ALH from Waybill
Chemical Revenue Ton-Miles	(QCS, line corresponding to STCC 28, Col. K) x Chemicals ALH from Waybill
Farm Revenue Ton-Miles	(QCS, line corresponding to STCC 01, Col. K) x Farm ALH from Waybill
Nonmetallic Revenue Ton-Miles	(QCS, line corresponding to STCC 14, Col. K) x Nonmetallic Minerals ALH from Waybill

Table A1: Data Definitions and Sources Used to Estimate the Railroad Cost Function*

Variable	Source
Other Revenue Ton-Miles	RTM – sum of Coal, Chemicals, Farm Products, and Nonmetallic Minerals Ton-Miles
RTM	Revenue Ton-Miles (R1, Sched 755, line 110, Col. B)
Factor Prices (all divided by GDPPD)	
Labor Price	Labor Price per Hour (SWGE+FRINGE-CAPLAB) / LBHRS
SWGE	Total Salary and Wages (R1, Sched 410, line 620, Col B)
FRINGE	Fringe Benefits (R1, Sched 410, lns. 112-114, 205, 224, 309, 414, 430, 505, 512, 522, 611, Col E)
CAPLAB	Labor Portion of Cap. Exp. Class. as Operating in R1 (R1, Sched 410, lines 12-30, 101-109, Col B)
LBHRS	Labor Hours (Wage Form A, Line 700, Col 4+6)
Equipment Price	Weighted Average Equipment Price (ROI, Ann. Depr, and Lease/Rental payments. per Car and Locomotive - weighted by that type of equipment's share in total equipment cost)
Fuel Price	Price per Gallon (R1, Sched 750)
FUEL Expenses	Fuel Expenses (R1, Sched 410, line 409, Col F + R1, Sched 410, line 425, Col F)
FUEL Gallons	Fuel Gallons (R1, Sched 750, line 4, Col B)
Materials and Supply Price	AAR Materials and Supply Index
Technological Conditions	
Route Miles	(R1, Sched 700, line 57, Col. C)
Way & Structures Capital	(ROADINV-ACCDEPR) / Route Miles
ROADINV	Road Investment (R1, Sched 352B, line 31) + CAPEXP from all previous years
ACCDEPR	Accumulated Depreciation in Road (R1, Sched 335, line 30, Col. G)
PCTORG	Percent of Tons Originated ((TONSOT+TONSOD)/TOTTONS)
TONSOT, TONSOD, TOTTONS	Tons Originated and Terminated, Tons Originated and Delivered, Total Tons (QCS, line 900, Cols. C, E, and K)
Average Length of Haul	RTM / REVTONS
REVTONS	Revenue Tons (R1, Sched 755, line 105, Col. B)

Note: * *Italics* indicate that the variable is used directly in the translog estimation

Table A2: Firms in the Data Set, with Merger Definitions

Atchison, Topeka & Santa Fe (ATSF)	1984-1995 - Merged into BN
Boston & Maine (BM)	1984-1986 – lost Class I status after 1988 (missing data in 1987-1988)
Burlington Northern (BN)	1984-2016 - From 1996-2016 includes merged ATSF, BN System
Canadian National (CN)	1999-2016 - Formed with the Merger of ICG, GTW
Chesapeake & Ohio (CO)	1984-1985 – merged with BO, SCL to form CSX
Chicago & Northwestern (CNW)	1984-1994 - merged into UP
Consolidated Rail Corporation (CR)	1984-1997 - Merged into CSX, NS in 1999 (missing input price data in 1992)
CSX Transportation (CSX)	1986-2016 - from 1999 - 2016 includes merged CSX, CR System
Delaware & Hudson (DH)	1984-1987 – lost Class I status
Denver, Rio Grande & Western (DRGW)	1984-1993 - Merged into the SP
Florida East Coast (FEC)	1984-1991 – lost Class I status
Grand Trunk & Western (GTW)	1984-1998 - Merged with ICG into the CN
Illinois Central Gulf (ICG)	1984-1998 - Merged with GTW into the CN
Kansas City Southern (KCS)	1984-2016 (Missing input price data from 1992-1994)
Milwaukee Road (MILW)	1984 – acquired by SOO
Missouri-Kansas-Texas (MKT)	1984-1987 - Merged into UP
Missouri Pacific (MP)	1984-1985 – Merged into UP
Norfolk Southern (NS)	1985-2016 - from 1999-2016 includes the merged NS, CR System
Norfolk & Western (NW)	1984 – merged with SRS to form NS
Seaboard Coast Line (SCL)	1984-1985 – merged with BO, CO to form CSX
SOO Line (SOO)	1984-2016
Southern Pacific (SP)	1984-1996 - From 1990-1993 includes merged SP, SSW - From 1994 - 1996 includes merged SP, SSW, DRGW - Merged into UP (missing data on carloads by commodity in 1991-1992)
Southern Railway System (SRS)	1984 – merged with NW to form NS
Saint Louis, Southwestern (SSW)	1984-1989 - Merged into SP (missing data from 1987-1988)

Table A2: Firms in the Data Set, with Merger Definitions	
Union Pacific (UP)	1984-2016 - From 1988-1994 includes merged UP, MKT system - From 1995-1996 includes merged UP, CNW system - From 1997-2007 includes merged UP, SP system (missing data on carloads by commodity in 1995)
Western Pacific (WP)	1984-1985 – merged into UP
TOTAL OBSERVATIONS = 332	

Cost Function Specification and Approach

The translog cost function specification is written as follows:

$$lnC^{SR} = \alpha_0 + \sum_i a_i \ln\ w_i + \alpha_k \ln k + \sum_m \beta_m \ln y_m + \sum_z \emptyset_z \ln\ t_z$$
$$+ \frac{1}{2}\sum_{ij} \alpha_{ij} \ln w_i \ln w_j + \frac{1}{2}\sum_{mn} \beta_{mn} \ln y_m \ln y_n + \frac{1}{2}\sum_{zx} \emptyset_{zx} \ln t_z \ln t_x$$
$$+ \frac{1}{2}\vartheta_k \ln k \ln k + \sum_i \theta_i \ln w_i \ln k + \sum_{im} \gamma_{im} \ln w_i \ln y_m + \sum_{iz} \varphi_{iz} \ln w_i \ln t_z$$
$$+ \sum_m \sigma_m \ln y_m \ln k + \sum_{mz} \tau_{mz} \ln y_m \ln t_z + \sum_z \omega_z \ln t_z \ln k + \epsilon$$

where ϵ is a disturbance term. All variables are divided by their sample means, which serves as the base point of approximation. With an assumption of cost minimization, Shephard's lemma generates conditional factor demands as follows:

$$x_i = \frac{\partial C^{SR}}{\partial w_i}$$

With variables in natural logarithms, this generates factor share equations. These factor share equations are estimated jointly with the cost function in a seemingly unrelated system, and shown as follows[28]:

$$s_i = \frac{\partial lnC^{SR}}{\partial lnw_i} = \alpha_i + \sum_j \alpha_{ij} \ln w_j + \theta_i ln\, k + \sum_m \gamma_{im} \ln y_m + \sum_{iz} \varphi_{iz} ln\ t_z + \epsilon_i$$

[28] Seemingly unrelated regression (SUR) is an econometric technique introduced by Zellner (1962). The equations are estimated as a system because the errors associated with the estimation of the cost function are likely correlated with those associated with the share equations. This improves the efficiency of the estimates.

Consistent with previous research in this area, we also assume symmetry of relevant cross-parameter terms, and we impose homogeneity of degree one in factors prices. This implies that a one percent increase in factor prices leads to a one percent increase in costs. Homogeneity of degree one in factor prices is imposed by the following parameter constraints:

$$\sum_i \alpha_i = 1, \qquad \sum_i \alpha_{ij} = \sum_i \theta_i = \sum_i \gamma_{im} = \sum_i \varphi_{iz} = 0\,, \qquad \alpha_{ij} = \alpha_{ji},\ \beta_{mn} = \beta_{nm},\ \emptyset_{zx} = \emptyset_{xz}$$

The cost function and factor share equations are estimated jointly in a seemingly unrelated regressions system using Zellner's procedure. In order to avoid perfect collinearity, one of the factor share equations is eliminated. The parameter estimates are asymptotically equivalent to maximum likelihood estimates, and thus are invariant to the equation deleted.
Consistent with most previous cost estimations, the cost function used in this study includes firm dummies to account for fixed effects.[29]

Table A3: Full Estimation Results – Short-Run Variable Cost Function

Variable	Parameter Estimate	Standard Error
Intercept	21.97686*	0.0839
w_L (Labor Price)	0.394859*	0.00390
w_E (Equipment Price)	0.150928*	0.00324
w_F (Fuel Price)	0.134597*	0.00200
k (Way and Structures Capital per Mile)	-0.15814**	0.0657
RTM_{COAL} (Coal Revenue Ton-Miles)	0.086642*	0.0288
RTM_{CHEM} (Chemicals Revenue Ton-Miles)	0.100709***	0.0556
RTM_{FARM} (Farm Products Revenue Ton-Miles)	0.098845*	0.0376
RTM_{NONMET} (Nonmetallic Minerals Revenue Ton-Miles)	0.091412*	0.0326
RTM_{OTH} (Other Revenue Ton-Miles)	0.36618*	0.0574
RM (Route Miles)	0.269533***	0.1391
ALH (Average Length of Haul)	-0.1172	0.1270
T (Time)	-0.2724*	0.0452
½ $(w_L)^2$	0.104707*	0.0135
w_L x w_E	-0.0372*	0.00494

[29] As noted by Oum and Waters (1996), there is some disagreement among authors over whether fixed effects should be included. Some authors argue that collinearity between output or network variables and firm dummies may reduce statistical significance or change the size of output and network parameter estimates. Nonetheless, because unobserved network variables likely influence costs, and they are correlated with included variables, their exclusion can create biased parameter estimates. As a result, we estimate the model with fixed effects.

Table A3: Full Estimation Results – Short-Run Variable Cost Function		
Variable	**Parameter Estimate**	**Standard Error**
w_L x w_F	-0.04366*	0.00479
½ $(w_E)^2$	0.026035*	0.00463
w_E x w_F	-0.01262*	0.00252
½ $(w_F)^2$	0.105142*	0.00328
w_L x RTM_{COAL}	0.004948**	0.00229
w_L x RTM_{CHEM}	-0.00537	0.00504
w_L x RTM_{FARM}	0.01679*	0.00428
w_L x RTM_{NONMET}	-0.00681***	0.00366
w_L x RTM_{OTHER}	0.012611	0.00845
w_E x RTM_{COAL}	-0.00201	0.00191
w_E x RTM_{CHEM}	0.010302**	0.00415
w_E x RTM_{FARM}	0.010797*	0.00358
w_E x RTM_{NONMET}	0.005622***	0.00307
w_E x RTM_{OTHER}	-0.01973*	0.00698
w_F x RTM_{COAL}	0.008019*	0.00118
w_F x RTM_{CHEM}	-0.00882*	0.00256
w_F x RTM_{FARM}	0.008414*	0.00219
w_F x RTM_{NONMET}	-0.00062	0.00188
w_F x RTM_{OTHER}	0.007152***	0.00430
w_L x k	-0.00065	0.00821
w_E x k	0.021291*	0.00686
w_F x k	-0.0005	0.00424
w_L x RM	-0.00607	0.0128
w_L x ALH	-0.06004*	0.0103
w_L x T	-0.03272*	0.00403
w_E x RM	-0.01125	0.0105
w_E x ALH	-0.01664**	0.00845
w_E x T	-0.02105*	0.00332
w_F x RM	-0.03044*	0.00649
w_F x ALH	0.078142*	0.00523
w_F x T	0.013389*	0.00205

Table A3: Full Estimation Results – Short-Run Variable Cost Function		
Variable	**Parameter Estimate**	**Standard Error**
½ $(RTM_{COAL})^2$	0.016815	0.0120
RTM_{COAL} x RTM_{CHEM}	-0.00228	0.0222
RTM_{COAL} x RTM_{FARM}	0.022174	0.0162
RTM_{COAL} x RTM_{NONMET}	0.055847*	0.0130
RTM_{COAL} x RTM_{OTHER}	-0.02385	0.0305
½ $(RTM_{CHEM})^2$	0.03978	0.0650
RTM_{CHEM} x RTM_{FARM}	-0.00691	0.0342
RTM_{CHEM} x RTM_{NONMET}	0.002638	0.0298
RTM_{CHEM} x RTM_{OTHER}	-0.21314*	0.0542
½ $(RTM_{FARM})^2$	-0.01719	0.0446
RTM_{FARM} x RTM_{NONMET}	0.01754	0.0204
RTM_{FARM} x RTM_{OTHER}	0.043469	0.0537
½ $(RTM_{NONMET})^2$	0.077055*	0.0267
RTM_{NONMET} x RTM_{OTHER}	-0.0312	0.0448
½ $(RTM_{OTHER})^2$	0.290657*	0.0842
RTM_{COAL} x k	-0.04343	0.0369
RTM_{COAL} x RM	-0.08163***	0.0490
RTM_{COAL} x ALH	0.005989	0.0583
RTM_{COAL} x T	-0.00338	0.0119
RTM_{CHEM} x k	0.159716*	0.0712
RTM_{CHEM} x RM	0.221975**	0.0889
RTM_{CHEM} x ALH	0.000255	0.1148
RTM_{CHEM} x T	-0.05966**	0.0273
RTM_{FARM} x k	-0.12325**	0.0523
RTM_{FARM} x RM	-0.03857	0.0966
RTM_{FARM} x ALH	0.07669	0.0756
RTM_{FARM} x T	0.020503	0.0257
RTM_{NONMET} x k	-0.08762	0.0564
RTM_{NONMET} x RM	-0.15544**	0.0746
RTM_{NONMET} x ALH	0.00075	0.0674
RTM_{NONMET} x T	-0.00706	0.0199

Table A3: Full Estimation Results – Short-Run Variable Cost Function		
Variable	**Parameter Estimate**	**Standard Error**
RTM_{OTHER} x k	-0.27968*	0.1013
RTM_{OTHER} x RM	0.007225	0.1415
RTM_{OTHER} x ALH	-0.2222***	0.1321
RTM_{OTHER} x T	0.194916*	0.0461
½ $(k)^2$	0.337681**	0.1375
k x RM	0.254654	0.1761
k x ALH	0.364353**	0.1488
k x T	-0.0836***	0.0473
½ $(RM)^2$	0.046279	0.2291
RM x ALH	0.066054	0.2180
RM x T	-0.12138***	0.0647
½ $(ALH)^2$	0.801452**	0.3551
ALH x T	-0.16965**	0.0697
½ $(T)^2$	-0.04988	0.0321

All variables are in natural logarithms, except the intercept.

Firm dummies not shown.

# of observations = 332

*significant at the 1% level, **significant at the 5% level, ***significant at the 10% level

Adj. R^2 Cost = .9966, Adj. R^2 Labor Share = .6949, Adj. R^2 Equip Share = .2534, Adj R^2 Fuel Share = .8781

Appendix B – Review of Literature on Railroad Cost Analysis

The first railroad cost estimations are often attributed to M.O. Lorenz in 1916.[30] Lorenz (1916) plotted railroad costs per unit of output against the amount of output produced, showing that unit costs fell sharply with more output at low levels of output and less sharply at higher levels of output.

Although Lorenz (1916) and his predecessors did not perform statistical estimations of costs, the studies are noteworthy for at least two reasons. First, the focus of these studies on how railroad costs vary with output and how costs change with changes in traffic, the size of the network, and the types of outputs produced are at the roots of both lines of cost analysis used today: (1) aggregate cost analysis that examines the overall structure of railroad costs to examine issues such as economies of scale, economies of scope, and cost subadditivity, and their implications for policy, and (2) individual movement costing used by railroads and regulators. Second, the cost issues being considered and the motivation for examining these cost issues in these studies is very similar to the current study.

Although these studies use different terminology than is used today, it is clear that they are addressing the same problems of common and unattributable costs, their allocation among shipments, and the need for differential pricing when faced with increasing returns to density. For example, Lorenz (1916) states "If it were possible to trace the connection between the particular service and particular portions of the outgo, we could solve the rate question by simply charging each service with the outgo occasioned. But the connecting of service and outgo is not possible for all the items of the outgo." In this statement, Lorenz is referring to costs when he says "outgo," and the problem he identifies is the problem of common or unattributable costs. Moreover, later in the paper, he goes on to state "the greater the economy to be regarded as resulting from a mere increase in traffic, the greater is the proportion of expenses which may be considered as fixed or independent of that traffic and the greater will be the scope to be given to the ability to pay element in rate-making…." This is another way of saying that the greater the extent of economies of density, the greater the need for differential pricing.[31]

This review of literature focuses on aggregate cost analyses, since the focus of this study is on the cost structure in the railroad industry and its implications for pricing. Moreover, given the ad hoc and/or linear specifications of most early cost studies, most of this review focuses on cost studies performed over the last 50 years.

Before going into these studies, however, it is worth noting an important study of railroad costs done in 1954 by George Borts. Borts (1954) made a significant contribution to the modern study of railroad costs by emphasizing the importance of distinguishing the short run from the long run in railroad cost analysis, and by grounding railroad cost analysis in economic theory.

---

[30] While Lorenz (1916) is often credited as being the first to estimate railroad costs, he cites a number of previous studies examining railroad cost issues, including studies by Sax (1879), Rank (1895), Launhardt (1890), Talcott (1904), Shinn (1875), Erickson (1910), Millard (1915), and others.

[31] In reviewing previous analyses of railroad costs, Lorenz (1916) makes it clear that others were examining similar issues as well.

Borts (1954) examined early cost studies that found some conflicting results and used economic theory to reconcile these results. He noted that many of the early cost studies, including those by Wellington (1893), Acworth (1904), Ripley (1927), and Jones (1931) found substantial economies of density with cost elasticities (percentage change in cost resulting from a one percent change in output) below .5 in most cases, meaning a 1 percent increase in output was associated with no more than a .5 percent increase in cost. However, these contrasted with other cross-sectional and time-series estimates showing cost elasticities that were substantially higher.

Borts (1954) went on to examine the studies of Clark (1923) and Daniels (1932) who explained this contradiction by stating that the short run cost elasticity was about .5 or less while the long run elasticity was greater than .5, but less than or equal to one. Daniels (1932) explained that the very low elasticities are reasonable in the short run, but that elasticity was higher in the long run because investment would have to increase with more traffic. Borts (1954) carefully explained with the economic theory of the relationship between short-run and long-run costs that such a finding implied that there was an overinvestment, not an underinvestment in rail facilities as implied by Daniels. Borts noted that railroad firms would expand to increased output by using facilities more intensively, not by investing. Although the study was performed 65 years ago, Borts' emphasis on understanding the implications of short-run returns and long-run returns and their implications is one that has shaped subsequent cost analyses and one that is still important today.

Another important early study by Borts (1960) highlighted a potential problem in estimating cost-output relationships using cross-sectional data. Borts referred to the problem that led to an underestimate of elasticity as regression fallacy. The regression fallacy occurs because the actual output produced by a railroad may be more or less than the output the railroad planned on when budgeting expenses for output. As Borts noted, in some years the railroad's budgeted (accounting) expenses are less than the true expenses realized (e.g. the railroad may incur an inadequate amount of maintenance expenses for the current traffic and have to make up with more future maintenance), due to anticipated output being less than actual output. In other years, the budgeted expenses are greater than those actually realized, due to anticipated output being greater than actual output. As a result, when output is high, measured costs are lower than the real costs, and when output is low, measured costs are higher than the real costs. This results in estimates of cost elasticity that are biased downward. This problem can be addressed by pooling cross sectional and time series data (Keeler, 1974), as was done in this study.

Another landmark study in the estimation of railroad cost functions was by Keeler (1974). In a study aimed at identifying the extent of excess capacity in the U.S. railroad industry, Keeler identified important problems with previous cost estimations. Similar to the findings of Borts, the problems identified by Keeler were identified by examining economic theory. As Keeler pointed out, most previous cost studies used one of two approaches: (1) estimated total costs as a function of output without including a measure of capacity, or (2) estimated total costs as a linear function of output and track mileage. He correctly noted that the first approach would only yield a long-run cost function if railroads had all adjusted to long-run equilibrium. There was strong evidence that this was not the case for railroads under regulation. Keeler argued that the second approach's assumption that factor proportions between track and other inputs were fixed was not appropriate, as increased intensity in use of railroad plant should lead to marginal maintenance and operating costs rising.

Although a couple of previous studies did allow for variable factor proportions and excess capacity, Keeler noted that the studies estimated short-run costs by assuming firms were above the long-run cost curve and estimated long-run costs by assuming firms were on the long-run cost curve—assumptions that contradicted each other. He remedied these problems by formulating a short-run cost function from neoclassical economic theory using a Cobb-Douglas production function, deriving the optimal capital stock for a given traffic level, and substituting the optimal capital stock into the short-run cost function to obtain the long-run cost function. Like the previous contributions of Borts, this approach of grounding cost results in economic theory was an extremely important contribution that has guided more recent cost studies.

A final important contribution by Keeler (1974) was that he distinguished between two different types of scale economies in the railroad industry—economies of density and economies of firm size. **Economies of density** refer to reductions in average cost resulting from increased traffic over a network of a given size, while **economies of size** refer to reductions in average cost resulting from increased traffic due to expansions in the size of the network. Keeler found substantial excess capacity in the rail industry, substantial economies of density with marginal cost pricing recovering less than 70 percent of costs for most firms, and constant returns to firm size. He noted that while a lot of track could be abandoned, thus eliminating the large economies of density, indivisibilities meant many density economies were likely to remain.

Another important study that has had a big influence on cost function estimation was by Harris (1977). In addition to reinforcing the importance of the innovations made by Keeler, Harris introduced other important innovations in cost function estimation that are still used today. Harris identified several problems in previous rail cost estimations. Despite the distinction made between economies of density and economies of size made by Keeler, there was continued confusion between the two concepts. Harris noted that the critical concept for pricing and investment policies was economies of density, not size. Other problems identified by Harris were inappropriate measures of output and capacity, an inadequate division of costs between passenger and freight services, failure to include a return on capital investment in costs, no clear rationale behind regional stratification, and a failure to include average length of haul as an explanatory variable of costs. While issues associated with dividing costs between passengers and freight (or issues related to regional stratification) are no longer issues in estimating railroad costs, all the remaining criticisms are important in guiding current and future cost analyses.

In addition to identifying problems in previous cost analyses, Harris estimated costs under several specifications, finding significant economies of density in the railroad industry and identifying some of their sources. An important finding by Harris (and one that is still misunderstood) was that many of the economies of density realized were not simply the result of spreading fixed way and structures costs among more traffic. Subsequent authors, such as Keeler (1983), have pointed out that a substantial portion of density economies are the result of railroads introducing longer and more frequent trains with increased traffic.

An important development in the estimation of railroad costs was the introduction of flexible functional forms in the 1970s. Flexible functional forms, like the generalized Leontief function that was introduced by Diewert (1971) and the translog function that was introduced by Christensen, Jorgenson, and Lau (1973), do not place the heavy restrictions on production structure that other forms such as the Cobb-Douglas do. Moreover, as is explicitly shown by Diewert in the context of the

generalized Leontief cost function, the flexible cost functions can be used to obtain (through Shephard's 1953 Duality Theorem) the production technology facing the firm. The translog function is particularly appealing, because it does not place any restrictions on the substitutability of inputs; allows returns to scale to vary by firm size, different factor prices, and other technological characteristics; and allows the researcher to capture the multiproduct nature of firms.

Brown, Caves, and Christensen (1979) were the first to use the translog cost function to estimate costs in the railroad industry. In an attempt to empirically evaluate the advantages of the translog over more restrictive forms, the authors estimated a railroad cost function using 1936 data for U.S. railroads with an unrestricted translog functional form and compared the results to more restricted functions. They showed that the translog cost function was a significant generalization of the other functional forms, and that restricted models can result in large errors in identifying scale economies and in measuring marginal costs. While the authors found significant multiproduct economies of scale for 66 out of the 67 railroads in their sample, there was no way to distinguish economies of density from economies of size since the study did not include route miles.

Friedlaender and Spady (1981) made several important innovations in the estimation of railroad costs in a book that was aimed at examining potential impacts of railroad and trucking deregulation. Like the previous authors that made improvements in the estimation of railroad costs, they examined railroad costs in the context of economic theory. One notable innovation made by Friedlaender and Spady based on economic theory was to distinguish between way and structures capital and route mileage. Way and structures capital is a factor of production, while route miles represents the extent of the railroad's network. By including way and structures capital in addition to route miles, Friedlaender and Spady were able to specify a short run cost function consistent with economic theory. Other innovations introduced by Friedlaender and Spady included accounting for more output heterogeneity by including the percentage of ton-miles accounted for by manufactured products and by introducing hedonic output functions, and distinguishing between different types of route miles based on their density levels. Unlike previous studies, Friedlaender and Spady were able to distinguish short-run returns to density (keeping way and structures capital fixed) from long-run returns to density. They found long-run increasing returns to density but decreasing returns to size.

Many of the same innovations present in the cost functions estimated by Friedlaender and Spady (1981) were also used in Berndt, Friedlaender, Chiang, and Velturo (1993) and in Friedlaender, Berndt, Chiang, Showalter, and Velturo (1993). In the context of examining the impacts of deregulation and mergers on railroad costs, and in the context of examining the industry's capital adjustments in response to deregulation, these studies estimated short-run railroad variable cost functions that included the amount of way & structures capital and accounted for differences in the mix of traffic among different railroads. These studies found extensive short-run and long-run returns to density, and mild returns to size. The authors suggested returns to density were not primarily the result of excess capital; rather they were an inherent characteristic of railroad technology.

In the first estimation of railroad costs using the translog functional form, Brown, Caves, and Christensen (1979) noted that a drawback of the functional form was that it was not able to accommodate zero output observations. Caves, Christensen, and Tretheway (1980) proposed a generalized translog multiproduct cost function that allowed the use of zero outputs through a Box-Cox transformation. In comparing this cost function to the translog multiproduct cost function, a

hybrid Diewert multiproduct cost function, and a quadratic multiproduct cost function, the authors showed several advantages over the others. Moreover, when comparing an estimated railroad cost function using the generalized translog multiproduct cost function to the translog multiproduct cost function, the authors find significant differences resulting from using observations with non-zero outputs. The authors found economies of scale in the railroad industry, although they were not able to distinguish economies of density from economies of size due to not including route miles in their estimation.

Caves, Christensen, and Swanson (1981) used the translog cost function to estimate productivity growth in the railroad industry between 1955 and 1974. They showed that their estimates of productivity growth were much different than those generated by previous studies that used restrictive index procedures to measure productivity growth. In particular, they showed important improvements in measuring productivity growth using the translog approach in comparison to previous approaches which implicitly assumed constant returns to scale and static equilibrium. In evaluating economies of scale for the railroad industry, they found strong scale economies when output increases resulted from increases in shipment distances. They were not able to distinguish returns to density from returns to scale, however, as their estimation did not include route miles.

As noted above, although Keeler noted the importance of distinguishing economies of density from economies of size in 1974, several subsequent studies failed to make such a distinction. Braeutigam, Daughety, and Turnquist (1984) pointed out this problem, in addition to a problem with other studies that failed to account for firm differences when using cross-sectional or panel data. To highlight the potential bias resulting from the failure to account for firm differences, the authors estimated a cost function for an individual firm using time series data. In addition, they included two other innovations. First, they included speed of service as a proxy for service quality. Second, they included a measure of "effective track," considering mileage and the amount invested in existing track above that required to offset normal deprecation. They found significant economies of density for the railroad studied, and that these economies of density were understated when not accounting for service quality.

Another study highlighting the importance of considering firm differences when estimating economies of density was Caves, Christensen, Trethway, and Windle (1985). Using a long-run cost function to estimate returns to density for 1951-1974 railroads, the authors found large economies of density and slightly increasing or constant returns to overall scale. In addition to distinguishing route miles from way and structures capital as was done by Friedlaender and Spady, they also highlighted the bias that can occur from estimating returns to density without considering firm effects.

All of the studies mentioned above used data that was prior to the major railroad regulatory reform in the 1970s and in 1980. A number of studies have examined railroad costs using post-deregulation data. For the most part, these studies maintained the innovations made in previous studies, continuing to find strong economies of density in the railroad industry.

In an attempt to measure cost savings and productivity gains from deregulation, Wilson (1997) estimated a short-run variable cost function using 1978-1989 data. In addition to standard variables included in cost function analyses, he added other variables meant to capture differences in railroad services and operations, such as the percent of traffic handled in unit trains, the percent of traffic

interlined, and the average speed rating of railroad track. Wilson found substantial cost reductions and productivity gains from deregulation. Moreover, he found large economies of density and roughly constant returns to size.

Although there were significant advances in cost function estimation since the studies of Borts in the 1950s, one problem that still existed (and still exists today) was that cost functions were unable to capture the multiproduct nature of railroad firms. While railroads handle a variety of products from a variety of origins to a variety of destinations every day, most railroad cost functions assume that railroads have one output (ton-miles), attempting to capture differences in railroad operations by including technological variables such as average length of haul and percent of shipments in unit trains. In a study aimed at examining cost implications associated with an open access framework for railroads, Ivaldi and McCullough (2001) estimated a railroad cost function with three different outputs: bulk traffic car-miles, high-value traffic car-miles, and general traffic car-miles. Using 1978-1997 data, they found significant economies of density and cost complementarities among different types of outputs. The study made a significant contribution by more closely capturing the multiproduct nature of railroads.

Since the study by Ivaldi and McCullough, there have been a number of railroad cost studies involving one of the authors of the current study in collaboration with others. Bitzan (2003), and Bitzan and Keeler (2003) accounted for three different types of outputs—unit train ton-miles, through train ton-miles, and way train ton-miles—and found significant economies of density using 1983-1997 data. Bitzan and Keeler (2007) found that deregulation led to substantial increases in density of U.S. railroads, along with substantial cost savings. In a study quantifying the impacts of railroad mergers in the U.S., Bitzan and Wilson (2007) found significant economies of density, as well as cost savings from mergers. Finally, studies by Bitzan and Keeler (2011 and 2014) show substantial economies of density in models that accounted for car-miles of different commodities.

While most studies of railroad costs support the existence of strong economies of density, most of these studies use data that is more than 20 years old. The study by Bitzan and Keeler (2014) is an exception. However, the study still uses data that only go through 2008. The next section describes the Uniform Railroad Costing System – the approach used by the STB to estimate individual movement costs.

Uniform Railroad Costing System (URCS) – Explanation and Criticisms

The Uniform Rail Costing System (URCS) was adopted by the Interstate Commerce Commission (ICC), the predecessor to the Surface Transportation Board, as its costing methodology in 1989.[32] It plays an important role in the STB's regulation of railroad rates, as it is used in annual revenue adequacy determinations, in determining whether railroads have market dominance for specific shipments, and in determining the reasonableness of rates when using the simplified rate reasonableness procedures.

[32] Surface Transportation Board Report to Congress Regarding the Uniform Rail Costing System (2010).

Variable costs are estimated by URCS in a three-step procedure.[33] The three steps in URCS are referred to as phases. In Phase I, various expenses incurred by railroads are grouped into different categories (e.g. transportation overhead expenses, transportation fuel expenses, yard operations) and those categories are related to output and capacity measures through linear regression. In Phase II, parameter estimates obtained in Phase I are combined with railroad outputs, capacity measures, and costs in each category to obtain variability ratios for each cost category (the proportion of that cost category's expenses that vary with output) and unit costs that consist of the variable cost of each category associated with units of railroad output. In Phase III, the units of various outputs encompassed in individual railroad shipments are combined with railroad unit costs to estimate the variable costs of a railroad shipment.

The STB states that the purpose of URCS is "to estimate that portion of the variable costs of providing rail service that can be attributed to any given rail movement." However, as highlighted by Rhodes and Westbrook (1986) and by Wilson and Wolak (2016), there are several problems with using URCS variable costs as a meaningful measure of the economic costs of providing a particular shipment.

Rhodes and Westbrook (1986) highlight several problems with the way the URCS regressions are performed. One criticism is that the regressions use a linear formulation. This implies that the relationship between costs and output does not vary with output. In contrast, flexible formulations used to estimate costs in academic studies allow these relationships to vary with output, as one would expect. As economic theory and logic would suggest, the relationship between costs and output will change as the railroad handles more and more traffic. A second criticism by Rhodes and Westbrook is that the regressions do not include input prices, and therefore assume that inputs are used in fixed proportions in producing output. To the extent that different railroads experience different input prices and different types of shipments and different system configurations, and to the extent that technology and relative input prices change over time, the fixed proportions assumption may be extremely inaccurate. Third, Rhodes and Westbrook criticize an assumption of constant variability functions across carriers. As they point out, variability ratios may vary among railroads because of different input prices and because of different mixes of inputs among railroads due to operating at different scales and with different networks.

Wilson and Wolak (2016) also criticize URCS on multiple dimensions. They argue that URCS is not really a method for estimating the increase in railroad costs resulting from producing another shipment, but rather a method for "allocating railroad costs to a generic shipment type." In addition to showing that URCS variable costs do not correspond to the relevant economic cost measures for railroad pricing—for example, marginal and incremental costs (highlighted in a previous section)—Wilson and Wolak combine URCS variable costs with revenues from the Carload Waybill Sample to show that the costs are not consistent with rational pricing decisions in many cases and to show that estimated URCS variable costs vary widely for different railroads providing the same shipment.

In addition to the criticisms of URCS by Rhodes and Westbrook (1986) and Wilson and Wolak (2016), a STB (2010) report prepared for Congress highlights other flaws associated with URCS. First, the report highlights the fact that URCS Phase I regressions have not been updated since the

[33] The URCS methodology is highlighted by Rhodes and Westbrook (1986), Wilson and Wolak (2016), and in STB (2010).

original Westbrook regressions done for 14 Class I railroads between 1979 and 1987. As railroad technology changes, output-cost relationships change as well. Another criticism is that the variability of 78 percent of railroad expenses is estimated using regressions, while the remaining 22 percent of railroad expenses are assigned by default variability factors (50 percent fixed and 50 percent variable) according to the judgement of the ICC. The extent to which this assumption is accurate is unknown. Moreover, the report also highlights the fact that the assignment of some railroad activities (e.g. train switching miles) to specific movements is based on engineering relationships determined from studies that date back to the 1930s. It is obvious that railroad operations and technologies have changed a lot since then. Finally, the report suggests that STB should be more careful to select output and capacity variables for Phase I regressions that are consistent with economic theory. STB says that data in Railroad Annual Reports to the Surface Transportation Board (known as R-1 data) show that there is a high correlation between operations and expenses, and any output or capacity variable put into the model might be significant, even when it is not the theoretically correct variable.

Appendix C – Returns to Scale and Differential Pricing

It is well known that under conditions of increasing returns to scale, a "second best" pricing solution is needed. The second best solution that maximizes social welfare subject to a break even constraint is known as Ramsey pricing. Ramsey pricing calls for pricing inversely to the elasticity of demand, and results in reducing outputs by the same proportion in each market in comparison to the output that would be produced if price were set at marginal cost. Formally, the Ramsey pricing rule is that the markup in each market should be as follows:

$$\frac{P_i - MC_i}{P_i} = \frac{\lambda}{|\varepsilon_{pi}|}$$

where λ is a proportionality constant that is the adjustment of the markup needed in all markets to make the firm break even, and ε_{pi} is the price elasticity of demand in market i. The proportionality constant (λ) will be between 0 and 1, depending on the degree of differential pricing needed to allow the firm to break even. As shown by the equation, Ramsey pricing results in pricing inversely to the price elasticity of demand.

An important question to ask is to what extent is the degree of differential pricing needed in the industry influenced by the extent of returns to scale? More specifically, in the railroad industry the important question is how big must the markup charged to relatively "captive" shippers be in order to ensure the railroad is able to earn a rate of return to ensure continued viability and investment?

If the assumption is made that "captive" shippers will bear the burden of any revenue shortfall, there are three things that influence the size of the markup that the railroad needs to charge "captive" shippers in order to ensure continued viability. These are (1) the degree of scale economies, (2) the price elasticity of demand in "competitive markets", and (3) the proportion of traffic that is "captive". Obviously, larger scale economies mean that there is a larger difference between marginal and average costs, and consequently a larger markup above marginal costs to recoup total costs. In addition, given any degree of scale economies and the competitive-captive traffic mix, a higher price elasticity of demand for transportation in competitive markets means higher markups for captive traffic due to the limited ability to mark up prices in competitive markets. Finally, a smaller

proportion of traffic that is captive will also mean a larger required markup to such shippers since the revenue burden is shared among fewer shippers.

Friedlaender (1992) developed a theoretical model to determine the necessary markups in captive sectors to recover full railroad costs, given the proportion of traffic that is captive and the degree of scale economies realized. This model, which is called the "Polar Ramsey" model by Friedlaender, shows the markup that would be required in the "captive" sector if there were only two types of traffic (captive and competitive), if competitive market traffic were charged a price equal to marginal cost, and if the captive sector demand were insensitive to price changes (perfectly inelastic demand).

The Friedlaender (1992) model can be adjusted to give a better approximation of the true average markup needed on captive traffic by utilizing the waybill sample to estimate the percentage of traffic that is captive (using a benchmark of revenue-to-URCS variable costs of 1.8 or higher) and to estimate the average markup for non-captive traffic. This adjustment is discussed after presenting the full model. The Polar Ramsey markups (as adjusted) illustrate the degree of differential pricing necessary to ensure revenue adequacy. Moreover, they can be used to show the importance of scale economies in necessitating differential pricing. The model developed by Friedlaender is shown below.

Multiproduct economies of scale are defined as (Baumol, et. al, 1988) as the inverse of the elasticity of cost with respect to all outputs (with a captive and competitive sector, there are two outputs):

$$S = \frac{C(y_{Comp}, y_{Captive})}{y_{Comp} x \frac{\partial C}{\partial y_{Comp}} + y_{Captive} x \frac{\partial C}{\partial y_{Captive}}}$$

In examining the equation for multiproduct economies of scale above, it can be seen that the denominator is equal to the revenues the firm would generate if it pursued marginal cost pricing of all outputs. Thus, if we define RMC (y_{Comp}, $y_{Captive}$) as the revenues the firm would generate under marginal cost pricing, it is apparent that the extent of multiproduct economies of scale multiplied by RMC is equal to the firm's total costs:

$$S\ x\ RMC(y_{Comp}, y_{Captive}) = C(y_{Comp}, y_{Captive})$$

If, however, the firm charges a price on each output (P_{Comp} , $P_{Captive}$) so that the railroad earns revenues equal to costs:

$$P_{Comp}\ x\ y_{Comp} + P_{Captive}\ x\ y_{Captive} = C(y_{Comp}, y_{Captive})$$

then, we get the following:

$$S\ x\ RMC(y_{Comp}, y_{Captive}) = P_{Comp}\ x\ y_{Comp} + P_{Captive}\ x\ y_{Captive}$$

If we solve this for S, we get:

$$S = \frac{P_{Comp}\ x\ y_{Comp} + P_{Captive}\ x\ y_{Captive}}{y_{Comp} x \frac{\partial C}{\partial y_{Comp}} + y_{Captive} x \frac{\partial C}{\partial y_{Captive}}}$$

This is also equal to:

$$S = \frac{\frac{P_{Comp}}{MC_{Comp}} MC_{Comp}\ x\ y_{Comp}}{y_{Comp} x MC_{Comp} + y_{Captive} x MC_{Captive}} + \frac{\frac{P_{Captive}}{MC_{Captive}} MC_{Captive}\ x\ y_{Captive}}{y_{Comp} x MC_{Comp} + y_{Captive} x MC_{Captive}}$$

The equation can be defined more compactly as:

$$S = \lambda_{Comp}\,\gamma_{Comp} + \lambda_{Captive}\,\gamma_{Captive}$$

where

λ_{Comp} = price/marginal cost ratio in the competitive market allowing the firm to break even
$\lambda_{Captive}$ = price/marginal cost ratio in the captive market allowing the firm to break even
γ_{Comp} = share of marginal cost revenues accounted for by the competitive sector
$\gamma_{Captive}$ = share of marginal cost revenues accounted for by the captive sector

This can be solved for the markup needed in the captive sector to allow the railroad to break even, as follows:

$$\lambda_{Captive} = \frac{S - \lambda_{Comp}\,\gamma_{Comp}}{\gamma_{Captive}}$$

If it is assumed that there is perfect competition in the competitive sector, so that the price elasticity of demand for rail service is equal to negative infinity, then the price/marginal cost ratio in the competitive sector would be equal to one. That is, price would equal marginal cost in the competitive sector, making the relevant equation for determining the markup needed in the captive sector independent of elasticity of demand in either sector, only depending on the degree of scale economies realized, and the proportion of marginal cost revenues accounted for by each sector.

Obviously, there is no clear demarcation between captive and competitive shippers. Because of differences in geography and products shipped, shippers have varying alternatives for delivering or receiving products. In some cases, shippers may be close to terminal markets, have nearby options for using alternative modes (or alternative rail carriers), have the ability to deliver products to or receive products from alternative locations, and/or have the ability to substitute different products for the one received. In such cases, shippers are likely to have relatively elastic demand for a particular railroad's services and be considered more competitive. Alternatively, shippers with few alternatives for delivering or receiving products are likely to have relatively inelastic demand for a particular railroad's services and be considered more captive.

The polar Ramsey markup assumes that captivity is an absolute concept, such that shippers are either captive or competitive. We can make an adjustment to the Polar Ramsey Markup to give a more realistic approximation of the average markup that needs to be charged to all traffic with a revenue-to-URCS variable cost ratio of 1.8 or higher (the STB's initial indicator of market dominance) in order for railroads to break even by utilizing the STB's Carload Waybill Sample (CWS). Specifically, we estimate a Modified Polar Ramsey Markup that uses actual average markups above URCS variable costs in the competitive sector. Thus, in the equation below, we use the following: λComp = revenue-to-URCS variable cost for all traffic with R/VC ratios below 1.8 from the

CWS, γCaptive = proportion of ton-miles from the CWS that have revenue-to-URCS variable cost ratios at 1.8 or above.[34]

$$Modified\ Polar\ Ramsey\ Markup\ (\lambda_{Captive}) = \frac{S - \lambda_{Comp}\,\gamma_{Comp}}{\gamma_{Captive}}$$

where S = degree of economies of scale

λComp = revenue-to-URCS variable cost for all traffic with R/VC ratios below 1.8 from the waybill sample.

γCaptive = proportion of ton-miles from the waybill sample that have revenue-to-URCS variable cost ratios at 1.8 or above.

γComp = (1- γCaptive)

Table C1 shows polar Modified Ramsey Markups using each year's average railroad characteristics, along with each year's proportion of traffic with R/VC ratios at 1.8 or above and each year's average markup for competitive traffic.[35] As the table shows, an increasing percentage of traffic with revenue-to-variable cost ratios at 1.8 or higher and a larger average markup on all other traffic have reduced the average markup that would need to be charged to all "potentially captive" traffic in order for railroads to earn a rate of return adequate to attract continued investment. However, it is estimated that the average revenue-to-variable cost ratio for "potentially captive" traffic would still need to exceed 2.25 in any year.

Table C1: Modified Polar Ramsey Markups

Year	Economies of Scale	Percent of Traffic with Revenue to URCS VC ≥1.8*	Average Revenue to URCS VC for all traffic with R/VC < 1.8**	Average Markup Needed on Traffic with R/VC ≥1.8
1988	1.62	26.01%	1.17	2.89
1989	1.58	22.57%	1.15	3.04
1990	1.57	21.59%	1.14	3.11
1991	1.61	19.71%	1.15	3.50
1992	1.53	19.60%	1.14	3.17
1993	1.60	20.79%	1.14	3.33
1994	1.61	22.79%	1.13	3.26
1995	1.60	22.82%	1.13	3.19
1996	1.60	20.63%	1.14	3.33
1997	1.60	20.18%	1.15	3.41
1998	1.57	19.63%	1.15	3.25

[34]The STB uses a revenue to URCS variable cost ratio of 1.8 as an initial indicator of market dominance. Our revenue-to-URCS variable cost ratio for competitive traffic is the weighted average for all traffic with ratios below 1.8, with ton-miles used as the weighting factor.

[35] Copies of the waybill sample available to the authors do not have URCS variable costs included for the years 1984 through 1987.

Table C1: Modified Polar Ramsey Markups				
Year	Economies of Scale	Percent of Traffic with Revenue to URCS VC ≥1.8*	Average Revenue to URCS VC for all traffic with R/VC < 1.8**	Average Markup Needed on Traffic with R/VC ≥1.8
1999	1.51	21.32%	1.19	2.68
2000	1.50	20.19%	1.16	2.84
2001	1.50	21.73%	1.17	2.69
2002	1.53	21.09%	1.18	2.84
2003	1.53	20.78%	1.22	2.71
2004	1.59	18.18%	1.16	3.53
2005	1.58	17.21%	1.12	3.80
2006	1.58	19.19%	1.14	3.40
2007	1.57	17.38%	1.11	3.78
2008	1.55	16.76%	1.09	3.81
2009	1.55	24.59%	1.17	2.73
2010	1.59	24.81%	1.21	2.76
2011	1.57	22.35%	1.19	2.91
2012	1.56	23.49%	1.23	2.66
2013	1.56	26.37%	1.26	2.41
2014	1.56	29.25%	1.27	2.27
2015	1.59	31.18%	1.26	2.34
2016	1.63	31.36%	1.26	2.46

*percent of traffic with R/VC of 1.8 or higher is based on percent of ton-miles with different R/VC ratios

**average R/VC ratios for traffic with R/VC<1.8 is the weighted average, where ton-miles are used as the weighting factor.

The results of the cost function estimation show that there continue to be strong returns to scale in the railroad industry, suggesting that marginal cost pricing would not yield the revenues necessary to generate continued investments needed in the railroad industry. The illustration of the extent of differential pricing necessary to generate break-even revenues from "Modified Polar Ramsey" markups for railroads suggests that large markups from traffic that is relatively captive is necessary to ensure continued railroad viability.