# Natural Monopoly in Transport

André de Palma, Julien Monardo

# Natural Monopoly in Transport

**André de Palma**, CREST, ENS Paris-Saclay, University Paris-Saclay
**Julien Monardo**, CREST, ENS Paris-Saclay, University Paris-Saclay

**Keywords** Natural Monopoly, Regulation, Subadditivity of Costs, Economies of Scale, Average Cost, Ramsey-Boiteux, Incentive, Multiproduct Firm.

**Abstract** Transportation networks, such as railways, roads and highways provide standard examples of natural monopolies. Since the introduction of the term "natural monopoly" by T. Malthus in 1815, this concept has been defined in different ways by several authors (F. Bastiat, J. S. Mill or L. Walras). The current formal definition is due to Baumol (1977) and is based on the subadditivity of the cost functions. After estimating the cost functions, the researcher can test whether subadditive holds or not. Natural monopolies are associated to market efficiencies, which call for regulation (e.g., price cap regulation and Ramsey-Boiteux regulatory policy). As a key example, the econometric study of the British railways in the 19$^{th}$ century shed light on the difficulty of regulating natural monopolies.

## History

The concept of natural monopoly appeared with Smith (1776) who, without naming it, explicitly provided the main characteristics of what scholars after him refers to as "natural monopoly". Its definition has then evolved through time and has attracted the attention of several famous scholars of the 17th-18th centuries, such as Thomas Malthus, Frédéric Bastiat, John Stuart Mill and Léon Walras.[1]

In the earliest explicit use of the concept, natural monopolies referred to as monopolies derived from natural factors of production, which are supplied in fixed quantity, with the idea that the limited supply of such factors constitutes barriers to entry.[2] The first definition, however, was given by J. S. Mill: natural monopolies were "those which are created by circumstances, and not by law". At that time, natural monopolies were therefore those created by nature, due to the presence of production factors supplied in given, and potentially limited, quantity; *natural monopolies* were thus distinguished from *artificial monopolies* created by law, i.e., by government measures. For J. S. Mill, natural monopolies encompassed many situations, including for instance barriers to entry due to capital requirement. J. S. Mill was also the first to recognize that natural monopolies could arise due to the production process, that is, due to technological reasons.

Afterwards, natural monopolies were meant to arise due to the presence of economies of scale, that is, when the average total cost is decreasing. This happens,

[1] See Mosca (2008) for an excellent history of the concept of natural monopoly.

[2] In 1815, Malthus, in his essay *The Nature of Rent*, made the distinction between « natural » monopoly and « artificial » monopoly. For instance, he mentioned as natural monopoly the case of "certain vineyards in France, which, from the peculiarity of their soil and situation, exclusively yield wine of a certain flavour".

in particular, when there are fixed (potentially sunk) costs and low or zero marginal costs. In this situation, the cost of the incumbent firm is lower than the cost of any other firm that would wish to enter the market, and, in turn, that firm remains alone in the market. Then, price is not equal to the marginal cost, as in the case of perfect competition, since profit maximization requires the monopoly to equalize marginal revenue to marginal cost; and the monopoly produces too little with respect to the social optimum conditions, so that the government may wish to regulate it.

The current formal definition used in the academic literature is due to Baumol (1977) and is closely related to the subadditivity of the cost functions, i.e., natural monopolies arise when the production cost associated to any set of outputs is less than the sum of the costs of producing separately all the different products in this set of outputs (see the formal definition below).

Very soon, academic scholars recognized that monopolies were unavoidable in transport networks, such as railways, roads and highways. For Jules Dupuit, a French engineer, monopolies in transport networks are due to their need to build a large infrastructure before operations could start. This makes the entry of a new firm impossible because only a very limited number of entrepreneurs can have access to a sufficiently huge capital. Moreover, if a new firm entered the market, it would extract profits from the incumbent monopoly, making both of them unprofitable. By contrast, for L. Walras, monopolies arise because only the government can decide the expropriation of the lands required to build the transport networks. Note also that, in transport networks, the presence of several small businesses is inefficient: as highlighted by Walras (1875), “building a second network of roads in a country where there is already one that is enough for all the communications would be an absurd way of chasing economies”.

However, many monopolies we know remain unchallenged given that strong regulations often protect them. Productions of electricity, of nuclear weapons, of military defence involve large fixed cost, and have been (and are still, for since several decades) protected by governments. By contrast, several economists belonging to the Austrian School such as Ludwig von Mises or Friedrich von Hayek, have advocated that natural monopolies do not really exist (Thomas J. DiLorenzo speaks about “myth of natural monopolies”) but are often the outcome of regulation or of some kind of State protection. The libertinism of the Austrian School is here a bit confusing. What is true, for sure, is that governments often play a role in protecting some natural monopolies. However, other monopolies, even “natural”, could be challenged by firms using improved technologies.[3]

## Formal Definition of the Natural Monopoly

[3] Entry in the taxi market, for example, has been historically difficult in France, especially in Paris, Île-de-France; but Uber managed to break (more or less successfully) this market in December 2011 (see https://en.wikipedia.org/wiki/Timeline_of_Uber) and started to capture customers, even when facing low network externalities, because they developed a revolutionary technology and were prepared to face (at least initially) negative profits.

A monopoly is a market structure in which a single firm produces a good or service without any close substitutes. Monopolies may have several sources, such as legal barriers (e.g., patents), capital requirements, economies of scales, etc. One particular form of monopoly is the natural monopoly, which arises when a single firm is able to offer that good or service to an entire market at a lower cost than two or more firms could. This means that a natural monopoly can be the outcome of an unrestricted competition.

The current formal definition of the natural monopoly is due to Baumol (1977) and is closely linked to the strict subadditivity of the cost function. A cost function $C(y)$ is *strictly subadditive* if for any vector, $(y_1, \ldots, y_M)$:

$$C\left(\sum_{m=1}^{M} y_m\right) < \sum_{m=1}^{M} C(y_m),$$

where the quantities $y_m$ are either quantities of different outputs or different quantities of the same output. A necessary and sufficient condition for a natural monopoly to exist is that the cost function is subadditive, which means that a single firm could produce at a cheaper cost compared to several firms.

The definition is also related to the concepts of economies of scale and economies of scope, which are cost efficiencies formed by quantity and by variety, respectively. *Economies of scale* correspond to a decreasing average total cost, while *economies of scope* arise when it is cheaper to produce several products together than to produce them separately.

For single product cost functions, economies of scale and economies of scope are sufficient but not necessary for subadditivity. This means that a natural monopoly arises when there are economies of scale or economies of scope over the relevant range of output (i.e., the range of output between the first unit of output produced and the output which consumers would demand at a zero price). For multiproduct cost functions, however, these conditions are neither necessary nor sufficient.[4]

The subadditivity of the cost function is also related but must be put in perspective with the concept of sustainability of monopoly, which refers to "[a]n industry to which entrants are not "naturally" attracted, and are incapable of survival even in the absence of "predatory" measures by the monopolist" (Baumol, 1977). In particular, Faulhaber (1975) shows that subadditivity of the cost function does not imply sustainability of the monopoly, while Baumol et al. (1977) shows the converse.

To illustrate what happens, consider a monopoly that produces and sells a single good or service at single price (i.e., absent of price discrimination). The monopoly produces a quantity $y$ so as to maximize its profits $\pi(y)$ defined by the difference between its total revenues $R(y)$ and its total costs, $C(y)$:

---

[4] Consider for example the multiproduct cost function, when there are two outputs,1 and 2:

$$C(y_1, y_2) = y_1 + y_2 + (y_1 y_2)^{1/3}.$$

Clearly, this cost function exhibits economies of scale when productions are strictly positive, but is never subadditive.

$$\pi(y) = R(y) - C(y) = p(y)y - C(y),$$

where $p(y)$ is the (decreasing) inverse demand function, which gives the price at which the quantity $y$ can be sold.

Assuming that price and cost are differentiable and well behaved, profits will be maximum when marginal revenues equal marginal costs, i.e., when $R'(y) = C'(y)$.[5] Since total revenues are equal to price multiplied by demand, this first-order condition leads to the monopoly pricing formula, also known as the inverse elasticity rule:

$$\frac{p(y) - C'(y)}{C'(y)} = \frac{1}{\varepsilon},$$

where $\varepsilon$ represents the elasticity of demand (in absolute value). The RHS of the above equation is referred to as the Lerner index and measures the market power of the monopoly.

As illustrated in the top panel of Figure 1, where marginal costs are assumed to be constant, profits are maximum at the point of intersection denoted by $E$, where the monopolist produces a quantity $y_m$ and sells at a price $p_m$.

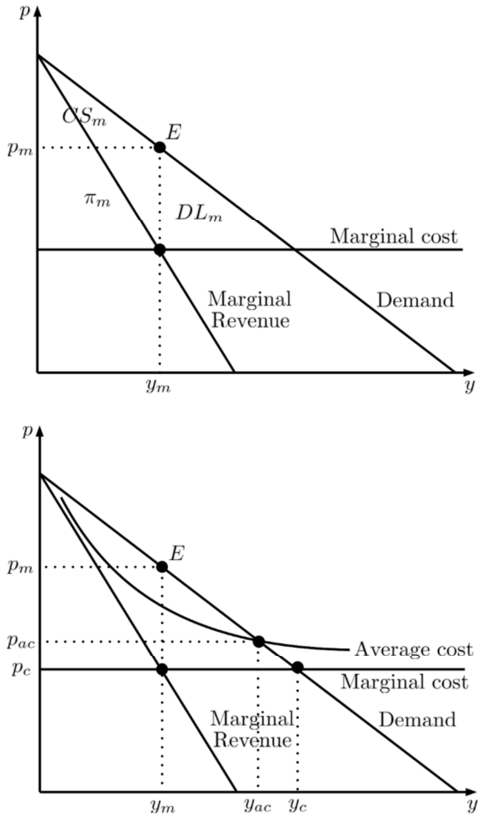


Figure 1: Monopoly and Natural Monopoly

At this optimum, the monopoly obtains profits equal to $\pi_m$ and consumers enjoy a surplus of $CS_m$. The society incurs a deadweight loss of $DL_m$: the social surplus, equal to $\pi_m + CS_m$, is lower than its socially efficient level (obtained under

[5] The maximum is attained provided that the second-order condition of the profit maximization program is satisfied.

perfect competition), since the monopoly sets a price strictly higher than marginal cost.

For the natural monopoly, the situation gets more complicated. This is because the natural monopoly typically exhibits a decreasing total average cost, which implies that its marginal cost is lower than its average total cost. This situation is illustrated in the bottom panel of Figure 1 for a firm with (large) fixed costs and (low or zero) constant marginal costs. In this case, the monopoly serves the entire market at a lower cost than multiple firms could achieve. For the natural monopoly, profits are still maximum at the point of intersection denoted by $E$, which is however the most undesirable situation for the society since it leads to high prices, small outputs and a large welfare loss.

## Regulating a Natural Monopoly

The inefficiency of the (natural) monopoly justifies its regulation, which aims to reduce its price and therefore increase its output. To address the inherent inefficient behaviour of the monopoly, policymakers or governments can resort to regulation or public ownership (i.e., in the limiting case, they can decide to run the monopoly themselves, i.e., opt for nationalization).

The choice of the regulated price is not easy. The government may want to set the price equal to the monopoly's marginal cost (marginal-cost pricing), so that efficiency is restored. However, this regulatory scheme faces two drawbacks.

First, the monopoly facing the marginal-cost pricing policy would incur losses and may, in turn, exit the market, since this policy leads to a price lower than average total cost (marginal cost being lower than average total cost). The government can address this problem, for example, by subsidizing the monopoly. However, in this case, the government incurs the loss, which can be covered by a tax that is associated itself to a deadweight loss. Alternatively, the government can allow a price higher than the marginal cost, for example by choosing an average-cost pricing rule so that the monopoly just makes zero profit, which is associated to a lower deadweight loss.

Second, marginal cost pricing does not provide the monopoly the incentives to reduce its costs. In a competitive market, firms can make higher profits by reducing their costs. By contrast, with the marginal-cost pricing rule, the regulated monopoly will not obtain higher profits by reducing its costs. The government can address this problem by designing a contract to induce the monopoly to reduce its cost as much as possible. Such incentives schemes are not simple to implement since effort of the monopoly is not directly observable.

To be more specific, consider the regulation of a monopoly producing $M$ goods or services, indexed $m = 1, \ldots, M$, when regulated prices are linear.[6] In the Ramsey-Boiteux problem, the social surplus is maximized under the constraint that

[6] Linear prices are unit prices that are constant for each product and that therefore depend neither on the quantity sold (no second degree price discrimination, involved for example in quantity discount), nor on the identity of the customers (no third degree price discrimination, where customers with different characteristics pay different prices for the same good or service).

the firm here the monopoly) breaks even. Let $S(y)$ denote the surplus that consumers derive from purchasing a vector of quantities $y = (y_1, \ldots, y_M)$. The government solves:

$$\max_y \{S(y) - C(y)\}$$
$$s.t.\ R(Y) - C(y) \geq 0,$$

where, as above, $R(y)$ is total revenues and $C(y)$ is total costs. First, consider the simple case in which demands for the products are independent. The first-order conditions lead to the Ramsey-Boiteux pricing:

$$\frac{p_m(y) - C_m(y_m)}{p_m(y)} = \frac{\lambda}{1+\lambda}\frac{1}{\eta_m}, m = 1, \ldots, M,$$

where $\eta_m$ denotes the own-price elasticity of good or service *m*, $p_m$ its price and $C_m$ its marginal cost, and where $\lambda > 0$, the Lagrange multiplier represents the shadow price of the budget constraint (or the shadow cost of public funds with government transfers).

Accordingly, for each good or service, its Lerner index is inversely proportional to its own-price elasticity. However, it should be noted that the Lerner index is smaller than the inverse elasticity of the demand since $\lambda > 0$, whereas, as seen above, in the unregulated monopoly, the Lerner index is just equal to its corresponding inverse own-price elasticity of demand.[7]

In practice, the regulator sets a price cap at the beginning of each period. The regulated price in period 1, $p_1$, is given by:

$$p_1 = p_0(1 + RPI - X),$$

where $p_0$ is the regulated price in period 0, $RPI$ is the inflation rate, and $X$ is the *efficiency factor* (i.e., the expected efficiency improvements).[8] One period is typically between 3 and 5 years. The $RPI$ can be measured by the Consumer Price Index (CPI) – or Retail Prices Index (RPI) as used in the United Kingdom.[9] The evaluation of $X$ is trickier, since it depends on the evolution of the inputs price and on the expected change in productivity. In practice, the regulator resorts to some heuristic rules, rather than to a full econometric analysis, which may be hard to accomplish. Benchmarking is another alternative, although studies may not always be comparable, so that econometric analysis is required.

For example, Gagnepain and Ivaldi (2002) examine the impact of incentives in the case of public transportation (buses) in France. They examine how incentive compatible contracts (à la J.-J. Laffont) may induce the bus companies to lower their costs, and compare two different regulator contracts, the *cost-plus contracts* (based on observed costs and *ex-post* deficits are covered) and *fixed price contracts* (based

[7] This analysis can be extended to the case where products are not independent. If they are substitutes, the Ramsey-Boiteux prices are higher; if they are complements, prices are lower.

[8] For a discussion on how to measure efficiency, see Gagnepain and Ivaldi (2002).

[9] See http://oa.upm.es/43724/1/Mariana_Rodrigues_Brochado.pdf

on expected costs and expected deficits). They empirically show that fixed price contracts are more efficient to reduce costs than cost-plus contracts.

This study shed useful light on the importance of incentives in the design of the contracts. The efficiency of the contracts varies significantly according to the size of the network, the density of the customers and the geographical characteristics. This is a common trait to many studies in transportation areas, such airline, maritime, railroads, rail freights or highways. Much work remains to be done to better understand the best way to regulate monopolies.

We cannot close this section without alluding to the fact that regulation may potentially reduce product innovation and process innovations. Lastly, note, as shown by Deneckere et al. (2019), that risk aversion of the principal (here the government) and the agent (here the monopoly) changes significantly the optimal contracts.

## Econometrics of Natural Monopoly

With data on costs and input at hand, the cost function can be estimated to determine whether it is subadditive or not, i.e., whether the industry under consideration is a natural monopoly or not. However, subadditivity is difficult to verify empirically. Fortunately, for the multiproduct case, a sufficient condition for the cost function to be subadditive is that its second partial derivatives are not positive over the relevant range of output. This condition, called “cost complementarity”, means that an increase in the production $y_i$ of good or service $i$ decreases the incremental cost of producing the quantity $y_j$ of good or service $j$.

Cost complementarity may be hard to test empirically over the relevant range of output, but can easily be tested at the data point. Then, from an econometric point of view, we are interesting in local conditions:

$$\frac{\partial^2 C}{\partial y_i \partial y_j} < 0.$$

The first step consists in assuming a functional form for the cost function that is able to identify whether or not there are cost complementarities. Flexible functional forms are usually used.[10] For example, Foreman-Peck (1987) uses the generalized translog (GTL) multiproduct cost function to estimate the cost function of the British railways.

The GTL multiproduct cost function model is defined as follows:

$$lnC(Y,w) = \alpha_0 + \sum_{i=1}^{N} \alpha_i ln(w_i) + \sum_{i=1}^{M} \beta_i Y_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_{ij}\, ln(w_i) ln(w_j)$$

[10] The flexible functional form were introduced by Diewert (1974). A flexible cost function is able to approximate an arbitrary twice continuously differentiable cost function to the second order at the data point. This is the reason for which only a local measure of cost complementarity may be tested. See Diewert (1974) and the literature that follows for more details and for other examples of flexible functional forms.

$$+\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}\beta_{ij}\,Y_i Y_j+\sum_{i=1}^{N}\sum_{j=1}^{M}\rho_{ij}\,Y_j ln(w_i)+\ \varepsilon,$$

where $w=(w_1,\ldots,w_N)$ denotes the price vector of the $N$ inputs, where $Y=\left(y^{\lambda_k}-1\right)/\lambda_k$ denotes the Box-Cox transformation of the output vector of the $M$ goods or services $y=(y_1,\ldots,y_M)$, where $\varepsilon$ denotes the error term of the model and where the $\alpha$'s, the $\beta$'s and the $\gamma$'s are parameters to be estimated.[11]

Given the large number of parameters to be estimated, the statistical precision of parameter estimates can be improved by assuming that firms minimize their (input) costs to produce the exogenously predetermined levels of output. In turn, Shephard's Lemma can be applied to the cost function $C$ in order to obtain the following cost share equations:

$$S_i=\alpha_i+\sum_{j=1}^{N}\alpha_{ij}\,ln\left(w_j\right)+\sum_{j=1}^{M}\rho_{ij}\,Y_j+\ \varepsilon_i, i=1,\ldots,N,$$

where $\varepsilon_i$ denotes the error term of the model.[12] The cost function can then be jointly estimated with $N-1$ of the $N$ share equations by using the method developed by Zellner (1962) for estimating seemingly unrelated regression models.

For the GTL multiproduct cost function model, the local cost complementarity is given by:

$$\frac{\partial^2 C}{\partial y_i \partial y_j}=\frac{C}{y_i y_j}\left[\frac{\partial \ln C}{\partial \ln y_i}\frac{\partial \ln C}{\partial \ln y_j}+\beta_{ij}y_i^{\lambda_i}y_j^{\lambda_j}\right],$$

which can be computed after estimation.

This methodology was used by Foreman-Peck (1987), who studies the railway industry in the United Kingdom during the 19th Century.

## Natural Monopolies in Transports: Panorama and Case Study

In transport, natural monopolies are important phenomena and arise, amongst other reasons, because the transport sector is capital intensive and needs large infrastructure to start producing. However, once fixed costs have been covered, the marginal cost to provide an extra unit of service is typically low. Since fixed costs are sunk, if an incumbent firm wishes to enter the market, the existing firm can easily cut prices to protect its market. On the other hand, if learning by doing is important, the incumbent firm may benefit from lower costs. Moreover, the incumbent firm can scream the market and serve the most profitable customers. For example, the

[11] The GTL function of cost generalizes the translog function of cost by using the Box-Cox transformation, rather than the logarithm, for the output levels. It therefore allows to include zero outputs. The Box-Cox transformation reduces to the logarithm as $\lambda_k$ approaches zero

[12] Note also that linear homogeneity (in input prices) of the cost function and symmetry of its Hessian matrix can be imposed by using the following linear restrictions on parameters:

$$\sum_{i=1}^{N}\alpha_i=1,\quad \sum_{j=1}^{N}\alpha_{ij}=0,\quad \sum_{j=1}^{M}\rho_{ij}=0,\quad \alpha_{ij}=\alpha_{ji},\quad \beta_{ij}=\beta_{ji}.$$

intercity railways are more profitable that the regional railways, where demand is sparser.

With the opening of the market in the railway market in France (December 2019 for local train and December 2020 for intercity train), it is likely that the non-French competitors will first enter the most profitable niches. However, practice is somewhat different. It should be noted that the First European Railway Directive, which dates back to 1991, allowed open access for passengers and freight trains. In 2019, still not much competition occurs. Breaking State monopoly is in the agenda, but political and institutional barriers still remain very strong. The study of natural monopolies should not ignore their most important facet: the political economy dimension, such as electoral competition, centralized versus decentralized decisions, etc. Deregulation has been so far more successful in the airline industry or in the truck industry, even if several imperfections remain, as widely discussed by Joskow (2007).

The case of the British railways in the 19th century provides an interesting case study (see Foreman-Peck, 1987). Competition has virtue to lower the price, while possibly leading to either duplication or underutilisation of tracks. Moreover, competing firms may deny and make difficult interconnections. The Railway clearing House, created in 1947, encouraged interconnection and fair competition. The estimations of Foreman-Peck (1987) suggest that before regulation, construction costs were 50% higher and national income per capita 0.75% lower than if would have been in a properly regulated market.[13] History teaches us that nationalisation does not solve all problems. In 1911, the British railways were heavily regulated, yet the performance were poor since competition was absent. However, privatisation of British Rail, 20 years ago, was not a full success either, with high fares, low reliability and little customer's support.[14]

The study of natural monopoly is by far not completed and raises questions opened for deep debates. Possibly, the divorce of ownership and control may provide a solution to a problem that seems to never end.[15]

## Bibliography

[1] Baumol, W. (1977). On the Proper Cost Tests for Natural Monopoly in a Multiproduct Industry. *The American Economic Review*, 67(5), 809-822.

[2] Baumol, W., Bailey, E., & Willig, R. (1977). Weak Invisible Hand Theorems on the Sustainability of Multiproduct Natural Monopoly. *The American Economic Review*, 67(3), 350-365.

[3] Cournot, A.A. ([1838] 1960). *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette. Translation: N. T. Bacon, (Trans.) (1960).

[13] Interestingly, he says that in "1856 Belgian third class fares per miles were one quarter lower than the British fare and in 1883 40% less, while freight was similarly cheaper"

[14] The majority of UK wish to revert the privatization of British Rail. According to the Office of Rail and Road, as of 2016 there was 62% support for public ownership of train-operating companies. See https://en.wikipedia.org/wiki/Impact_of_the_privatisation_of_British_Rail

[15] See e.g. http://www.cirje.e.u-tokyo.ac.jp/research/dp/2012/2012cf864.pdf

*Researches into the Mathematical Principles of the Theory of Wealth*. New York: Kelley.
[4] Deneckere, R., A. de Palma and L. Leruth (2019). Risk Sharing in Procurement. *International Journal of Industrial Organization*, forthcoming.
[5] Diewert, W. E. (1974), "Applications of duality theory", in: M. D. Intriligator and D. A. Kendrick, eds., *Frontiers of Quantitative Economics*, Vol. II. Amsterdam: North-Holland Publishing Company.
[6] Faulhaber, G. (1975). Cross-Subsidization: Pricing in Public Enterprises. *The American Economic Review*, 65(5), 966-977.
[7] Foreman-Peck, J. (1987). Natural Monopoly and Railway Policy in the Nineteenth Century. *Oxford Economic Papers*, 39(4), new series, 699-718.
[8] Gagnepain, P., & Ivaldi, M. (2002). Incentive Regulatory Policies: The Case of Public Transit Systems in France. *The RAND Journal of Economics*, 33(4), 605-629.
[9] Joskow, P. L. (2007). Regulation of natural monopoly. *Handbook of law and economics*, *2*, 1227-1348.
[10] Malthus, T.R. ([1815] 1969). *An Inquiry into the Nature and Progress of Rent, and the Principles by which it is regulated*. New York: Greenwood Press.
[11] Mosca, M. (2008). On the origins of the concept of natural monopoly: Economies of scale and competition. *The European Journal of the History of Economic Thought*, *15*(2), 317-353.
[12] Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: W. Strahan and T. Cadell.
[13] Train, K. (1991). *Optimal regulation: the economic theory of natural monopoly*. Cambridge, MA: MIT Press.
[14] Walras, L. ([1875] 1936). L'Etat et le chemin de fer. Reprinted: (1936). *Etudes d'économie politique appliquée*. Paris: R. Pichon et R. Durand-Auzias, pp. 193-236.
[15] Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association*, 57(298), 348-368.