



European Transport \ Trasporti Europei

Editorial Advisory Board

Editor-in-chief/Direttore responsabile:

Giacomo Borruso,
University of Trieste, Italy
President of ISTIEE.

Editorial Advisory board/Comitato Scientifico:

Aurelio Amodeo
University of Trieste, Italy.

Michel Bierlaire
Swiss Institute of Technology, Lausanne,
Switzerland.

Sergio Caracoglia
University of Trieste, Italy.

Roberto Camus
University of Trieste, Italy.

André de Palma
Cergy-Pontoise University / Ecole Nationale des
Ponts et Chaussées. France.

Jadranka Jovi
University of Belgrade, Serbia and Montenegro.

Enrico Musso
University of Genoa, Italy.

Esko Niskanen
STARResearch, Helsinki, Finland.

Stratos Papadimitriou
University of Piraeus, Greece.

Piet Rietveld
Free University, Amsterdam, The Netherlands.

Fabio Santorini
University of Trieste, Italy.

Bernhard Wieland
Technical University, Dresden.

*Managing Editor/Coordinatore Comitato di
Redazione:*

Romeo Danielis
University of Trieste, Italy.

Associate Editors/Comitato di Redazione:

Elena Maggi
Politecnico di Milano.

Edoardo Marcucci
University of Urbino, Italy.

Marco Mazzarino
IUAV, Venice, Italy.

Vittorio A. Torbianelli
University of Trieste, Italy.

European Transport \ Trasporti Europei
Quarterly Journal of Transport Law, Economics and Engineering
Rivista Quadrimestrale di Diritto, Economia e Ingegneria dei Trasporti

ISSN 1825-3997

YEAR X, NUMBER 29, april 2005

Publisher/Direzione, redazione e amministrazione

I.S.T.I.E.E.

Istituto per lo Studio dei Trasporti nell'Integrazione Economica Europea

Via Lazzaretto Vecchio, 13
34123 TRIESTE
Tel. +39.040.311464
Fax. +39.040.311465
Mail: istiee@units.it

www.istiee.org/te

Print/Stampa
ARTIGRAFICHERIVA SRL TRIESTE

Autorizzazione del Tribunale di Trieste, n. 915 del 31 ottobre 1995

Spedizione in A.O. 70% DC TS
Tassa Pagata/Taxe Perçue



Index

YEAR X, NUMBER 29, April 2005

<i>Cantos, Pedro</i> <i>Campos, Javier</i>	Recent changes in the global rail industry: facing the challenge of increased flexibility	1
<i>Cantos, Pedro</i> <i>Campos, Javier</i>	Recent changes in the global rail industry: evaluating the new regulatory instruments	22
<i>Liliopoulou, Anastasia</i> <i>Roe, Michael</i> <i>Pasukeviciute, Irma</i>	Trans Siberian Railway: from inception to transition	46
<i>Kapsa, Elzbieta</i> <i>Roe, Michael</i>	The development of the highway network in Poland and the future development of Polish ferry shipping	57
<i>Gundaliya, P. J.</i> <i>Mathew, T. V.</i> <i>Dhingra, S. L.</i>	Methodology for finding optimum cell size for a grid based cellular automata traffic flow model	71



Recent changes in the global rail industry: facing the challenge of increased flexibility

Pedro Cantos^{1*}, Javier Campos²

¹ *Departamento de Análisis Económico (Universidad de Valencia)*

² *Departamento de Análisis Económico Aplicado (Universidad de Las Palmas de Gran Canaria)*

Abstract

This paper discusses how the current trend towards increased private participation in the rail industry is reshaping the way in which Governments should address the main regulatory challenges arising from the particular economic and technical characteristics of this industry. We review the role of railroads in the last two decades and examine the characteristics of the most relevant processes of private participation around the world. The lessons learned from these changes suggest that many of the traditional regulatory paradigms in this industry are being replaced by more flexible schemes of public intervention. Although this change does not fully preclude direct participation by the Government, it seems that the traditional monopolistic rail company is dead as the dominant model around the world, and new forms, such as franchises or concessions competing on the tracks are progressively gaining relevance.

Keywords: Railways; Restructuring; Regulation; Privatization.

1. What makes rail regulation different?

The rail industry poses a number of specific problems for transport economists and regulators that are only partially shared with other transport modes. These elements are the multi-product nature of the activity, the particular cost structure of railroad companies, the role of infrastructure and networks, the existence of indivisibilities in inputs and outputs, the organization of rail transport as a public service, and the existence of externalities in the transport system as a whole. These characteristics define a descriptive framework for this sector, and jointly determine the main factors that should be considered when studying in detail the appropriate economic regulation for the rail industry.

* Corresponding author: Pedro Cantos. Departamento de Análisis Económico (Universidad de Valencia). Campus dels Tarongers, s/n. 46022 Valencia (Spain). E-mail: Pedro.Cantos@uv.es.

1.1. The multi-product nature of the rail activity

Rail companies are, in most cases, multi-product firms that provide different types of freight and passenger transport services. In the case of freight, along with the usual transport of bulk freight, rail operators also supply complete cargo wagons or trains, parcel and postal services, as well as other services of intermodal transport. In the case of passenger transport, long-distance traffic usually co-exists with local services (suburban and commuter trains), regional services, and in certain cases, even with high-speed trains.

The multi-product nature of railways has different implications. In accounting, for example, it is often difficult to allocate total operating costs among services. Many of the costs of running a long-distance train (including not only infrastructure costs but also variable costs) are shared by different types of traffic and these joint costs co-exist with other costs not affected by changes in output. For instance, the common costs of signal maintenance along a line section usually do not increase if the proportions of traffic of the different services change. Although some cost elements may be attributable to a particular traffic (for example, passengers), most of them (wagons, energy, staff, etc.) are not. Thus, cost interdependence requires simultaneous decisions on prices and services, which, in practice, makes any regulatory task much harder.

At the cost level, another important aspect to consider in the multi-product setup of the rail industry is the sub-additivity of the cost function faced by a railroad.¹ This idea conveys two relevant implications for the rail industry. First, is it more efficient for a single firm, rather than two separate firms, to supply both infrastructure and transport services? Second, if the infrastructure and services are separated, is the supply of such services more efficient within the context of a monopoly, or should two or more firms participate. This analysis, connected to the advantages and disadvantages of the separation of infrastructure from services, will be discussed in depth.

1.2. The pervasive structure of railway costs

Waters (1985) broadly distinguishes four railway cost categories: (i) *train working costs*, including the cost of providing transport services (fuel, crew, maintenance and depreciation of rolling stock); (ii) *track and signalling costs* (including operation, maintenance and depreciation of infrastructures); (iii) *terminal and station costs*; and finally, (iv), *administration costs*.

The first two categories are prevalent in most companies and change according to several factors. Among train working costs, for example, rolling stock costs depend on both their number and the distance they run. Fuel costs depend on car-kilometres run for each type of vehicle, while train crew costs vary according to train-kilometres run. Track and signalling costs usually rely on the length of the route (since they typically request a single, standard-quality track). The amount of track and signalling needed, however, changes with the number of trains requiring paths, although this relationship is not constant. Terminal and station costs depend on traffic volumes, but vary considerably with the type of traffic. For instance, bulk freight handling requires more

¹ According to Baumol (1977), a cost function is sub-additive when the provision of services by a single firm is more efficient (in terms of a lower unit cost) than the same production carried out by two or more companies.

terminal expenses than parcel services. Similarly, long distance passengers require more services (ticketing, reservations, luggage, etc.) than short distance users. Administration costs, finally, fluctuate depending on the overall size of the firm, although the precise nature of this dependence is generally difficult to determine.

Allocating all of these costs to the multiple outputs or inputs it produces is complex. It often involves a degree of arbitrariness that demands, from a regulatory point of view, a clear distinction between avoidable and unavoidable costs. The avoidable costs are uniquely associated with a particular output: were this output not produced, no cost would be incurred. Avoidable costs may therefore be considered as a regulatory price floor (if any), since charging less would be equivalent to operating at an economic loss.

1.3. The economic role of rail infrastructure

Since the birth of the rail industry in the last century, mainstream economists have always considered that the larger the size of a railway company, the greater its efficiency. The existence of substantial fixed costs (particularly, those associated with infrastructure) traditionally led economists to assume the presence of important economies of scale, and thus to regard rail transport service as a textbook example of a natural monopoly.

However, this notion has been heavily challenged in recent decades by the introduction of new ideas into the industry's economic analysis. Particularly, the upheaval of the theory of contestable markets (Baumol, Panzar and Willig, 1982) contributed to clarifying the proper definition of the natural monopoly concept, in terms of the sub-additive cost function (see note 1). This concept implies that duplicating rail infrastructure is generally inefficient (and therefore is subject to natural monopoly conditions), but once the network has been deployed, the cost of operating rail transport services and rolling stock can be efficiently covered by more than one company, either as actual or potential competitors.

Therefore, from the regulatory point of view, the conclusion is that infrastructure and services can be dealt with in different ways: the former, as a natural monopoly,² but also as a potential provider of adequate access to any willing-to-serve operator; the latter, as any other competitive economic activity that could be provided by multiple competing operators or by a single firm under some sort of concession or license arrangement.

1.4. The implications of asset indivisibilities

Even though this potential vertical separation alleviates some of the natural monopoly problems, the rail industry remains very capital-intensive, with several other indivisibilities within its productive process. Specifically, the capital units (rolling stock, tracks and stations) can only be expanded in discrete, indivisible increments (the addition of a train or wagon, for example), while demand fluctuates in much smaller units. Consequently, increases (decreases) in supply can exceed increases (decreases) in demand, resulting in excess capacity. This lumpiness has several important implications for investment and pricing. For example, the transportation costs of an additional unit of

² At least, when the infrastructure has not yet been built, although not necessarily after that moment.

traffic (freight or passengers) may be insignificant when there is idle capacity, but may become substantial when the capital is being used to its fullest.

Firms can also be forced to employ fixed assets with differing economic lives, whose reliability spans over a large time horizon and heterogeneously affects the cost items described above, modifying investment decisions, and requiring a complete accounting and management information system. Therefore, dynamic price and output considerations become crucial in order to recover the real costs associated with each period of activity.

A final implication of the indivisibilities in the rail industry's capital assets is that innovation and infrastructure improvement projects are usually deferred and only carried out in small, discrete amounts. Railway firms seldom change the entire definition of their existing network, which in most countries corresponds to an inherited burden from past decades when the traffic structure was very different than today. Instead, they opt for partial renovations that often introduce technical asymmetries between tracks within a country or region, and accentuate indivisibilities and inflexibilities.

1.5. The role of rail transport as a public service

Although not derived from historical and organizational reasons and not from technical characteristics, the concept of rail transportation as a public or social service, irrespective of profitability, is another defining element that has determined the industry's organization and performance around the world. The low rolling resistance of steel wheels on steel rails made railroad transportation extremely fuel efficient and relatively cheap. This allowed railroads to rapidly grow as the first mass transportation system, particularly for passengers, beginning in the years of the industrial revolution.

For military and industrial reasons, some form of public control was envisaged in most countries, and many imposed their control by legal mandate. Public control over the rail industry occurred both with and without accompanying subsidies, public service obligations to transport providers in the form of compulsory (often unprofitable) routes, organized timetables or particular services for strategic products or areas. The ultimate reason behind this control, which remains the same today, is that this industry is regarded as an integral mechanism to overcome geographical barriers in certain areas, aid in the economic development of undeveloped zones, and even as a guarantee of minimum transport services for a particular segment of the population.

1.6. Externalities and the rail system

The policy goal of public service obligation is often supported by the idea that rail transportation contributes less to negative externalities than other transport modes, especially roads. There is abundant empirical evidence showing that under high demand conditions, the external costs of traffic congestion, accidents and environmental impact (noise, visual impact, pollution, etc.) could be reduced by transferring a substantial part of road traffic to rail.

The current intermodal misallocation (more road users than rail users) arises from the fact that road transport does not fully internalize all of the social costs that it generates.

Economists often recommend the use of congestion and/or pollution rates to account for this. However, when these mechanisms are not feasible or politically viable, it might be preferable to decrease railway fares to improve the overall intermodal balance, which is an additional consideration for rail regulation.

In summary, all of the above-mentioned characteristics suggest that an analysis of the regulation of railway transport should be carried out within a general context, taking into account the industry's technological and organizational features, beginning with a detailed evaluation of recent performance.

2. Recent regulatory trends in the global rail industry

The overall evolution of rail transportation in recent years as compared to other transport modes is summarized in Table 1 for OECD countries. There was a substantial fall of the market share in both freight and passenger markets during the 1970s and 1980s, which stabilized (even with a slight increase) during the last five years. In relative terms the decline is particularly relevant because it was during a period when the total volume in both markets grew about 50%, implying that the railroads were not able to take advantage of growing demand in these years.

Table 1: Market shares of different transport modes (1970-2002).

	<i>Passenger traffic</i>						
	1970	1980	1985	1991	1994	1998	2002
Rail (%)	10.43	8.64	7.33	6.92	6.85	6.83	7.04
Private car (%)	77.30	79.97	83.37	84.37	84.38	84.48	84.64
Bus (%)	12.26	11.38	9.29	8.70	8.75	8.68	8.30
	<i>Freight traffic</i>						
	1970	1980	1985	1991	1994	1998	2002
Rail (%)	31.3	23.2	21.2	17.9	15.5	14.3	15.0
Road (%)	55.2	65.9	69.3	74.0	76.2	78.5	77.6
Waterways (%)	13.5	10.9	9.5	8.1	7.9	7.2	7.4

Source: CEMT. *Evolution des Transports. OECD Countries.*

The substantial reduction in market share is not particular to OECD countries but a common trend around the world. It can be attributed to both exogenous and endogenous causes. The former includes the rapid development of alternative modes of transport, especially road. For passengers, economic growth fostered the development of the automobile market, leading to enormous growth in motorization. In freight transport, the expanding, competitive trucking sector gained a growing percentage of transport in many countries. For example, in 1970 in Europe, there were 150 cars per 1,000 inhabitants, a figure that now is 424. Similarly, the number of heavy vehicles and trucks increased from 7 to 20 million from 1970-2000.

The endogenous causes of the decline can be summarized in the inability of the sector to adapt to the changing conditions of the economic environment. Regulation remained obsolete and the rail industry was slow to react. The policies adopted during the 1980s, as described below, did not halt the steady loss of market share, the growing financial deficits, and in some countries, the impossibility of raising the low productivity indices of the industry. Thus, more radical restructuring processes were put into practice.

2.1. The traditional railroad model

During the past fifty years, the most common market structure in many countries' rail sectors was a single, state-owned firm, entrusted with the unified management of both infrastructure and services. Despite some differences in their degree of commercial autonomy, the traditional methods of regulation and control of this sort of company have been relatively homogeneous. In general, it was assumed that the monopoly power of the national company required price and service regulation to protect the general interest. In addition, there was an obligation on the part of the companies to meet any demand at those prices. The closure of existing lines or the opening of new services required government approval. Thus, competition was rare and often discouraged, and the preservation of the national character of the industry was considered the key factor governing the overall regulatory system.

Under this protective environment, most national rail companies incurred growing trade deficits during the 1970s and 1980s. Furthermore, social obligations to their staff made it nearly impossible to reach any agreement on redundancies or even wage adjustments. In some countries, the companies were forced to finance their deficits by borrowing, so their accounts lost all resemblance to reality. The main problems associated with the traditional policies for railways were: *(i)* increasing losses, which were usually financed by public subsidies; *(ii)* a high degree of managerial inefficiency; and *(iii)* business activities oriented exclusively toward production targets rather than commercial and market targets.³

These distortions did not come from any artificial reduction in the range of services provided, nor from excessively high fares, but more commonly, from an unjustified increase in the supply of services (and hence, of costs). Such behaviour implied larger public subsidies. In many cases, the lack of commercially-oriented tariffs and investment policies explained many of the difficulties faced by the companies. Together with the burden imposed by the technical characteristics of the sector, this placed most railways in a very weak position to compete against alternative transport modes. However, fierce intermodal competition was not able to improve the competitiveness of the railway system by itself. It was necessary to adopt measures affecting the internal behaviour and structure of the sector itself. Therefore, the sector's overall decline sparked a widespread restructuring movement around the world.

³ On this point, Oum and Yu (1994) and Gathon and Pestieau (1995) have empirically shown that the companies that achieve the greatest efficiency were those that had been run with a higher level of autonomy and independence from state intervention.

2.2. Main features of the rail restructuring policies

The worldwide restructuring process of the rail industry began with timid reforms. Many countries began by replacing their national railways with autonomous commercial bodies possessing independent, realistic balance-sheets, in which only public service obligations could be explicitly subsidized by the government. Other countries opted to substitute their old geographically-based management with a multi-divisional structure, defined by the companies' different lines of business or services.

A common feature in most cases is that some countries have carried out a relatively long-term restructuring, whereas others have preferred a quicker implementation. For example, privatization in New Zealand and Japan was phased over several years, while Argentina and the United Kingdom took less than two years. Another common characteristic is that all restructuring processes were undertaken to make the companies attractive to private investors, although full privatization has been less preferred than concessioning.

The changes have involved the revision of laws and other regulations affecting railways, reducing staff, dealing with pension issues, and deciding how much property should be sold and how much should be retained by the state. In addition, several arrangements for paying for unprofitable (but socially needed) train services were put into place, together with a precise definition of the concession contracts and their main terms.

With regard to results, in general, most of the restructuring experiences detailed below seem to have been positive. The objectives of stopping the industry's drain on the state's resources, along with the stabilization of market share for both passengers and freight, were achieved in most countries. Likewise, the companies succeeded in raising their levels of productivity.

3. New organizational models for the rail industry

Despite all these changes, the most salient characteristic of the restructuring process of the rail industry in the last decades has been the consolidation of different and alternative organizational structures for the industry as a whole. These structures differ along three main features to be analyzed in detail: how are access and infrastructure and multimodal competition considered, what is the extent of vertical separation introduced after the change, and what is the amount of private participation allowed in the industry after the reform.

3.1. Access to rail infrastructure and intermodal competition

The management of rail infrastructure not only includes simple pricing principles, but also access rights and long-term development provisions. Each country addresses these differently: most have opted to publicly retain infrastructure, creating state management agencies (Sweden's *Banverket*) to regulate private train operators (as in Argentina); others (France, Germany or Spain) have established nominally independent but state-owned enterprises to manage stations and tracks. Only the United Kingdom privatized

infrastructure and operations in 1996, although part of the changes were later reverted in 2001. The financial collapse of *Railtrack* (the private owner of the infrastructure) and the poor infrastructure maintenance (that provoked serious accidents and significant disruptions to service) were the main reasons to dissolve *Railtrack* and substitute it for a *not for profit* body, *Network Rail*. This suggests that, whether in public or private hands, rail infrastructure regulation must include minimum investment requirements to avoid short-term myopia and to ensure that key investments are prioritized over dividend increases or defence against potential takeover.

On the other hand, the separation of infrastructure from services also implies that the new models should focus on the issue of access, which is particularly relevant in the case of highly integrated trans-national networks (as in Europe) or privately or publicly managed dense networks (as in the United States, Canada and some Asian countries). In the European Union, for example, Directive 91/440 directs each member state to grant international access and transit rights to international groups where stakes are held by railway undertakings in that or other member states. There have been no directives or resolutions related to domestic traffic, although the European Commission advocates the extension of these provisions to all freight and international passenger services. In January 2002 the Commission adopted a new communication: *towards an integrated European railway area* (known as the *second railway package*). Open access to the infrastructure for national services is promoted in order to completely open up the rail freight market. It has been agreed and open access in the domestic freight market will be introduced in 2007. However, open access in the passenger market is a much slower process.

In the privatized structure of the United Kingdom, open access to passenger services has been limited by a number of provisions regarding that moderate competition. Initially designed to protect rail franchisees from new entrants and from each other, these provisions were anticipated to be gradually reduced over time. In other countries (Argentina and Côte d'Ivoire-Burkina Faso), access rights are also clearly specified in the contract. In certain large cities, like Mexico, D.F. or Buenos Aires, operators share a common network under a unique transport authority.

The final aspect regarding access rights to rail infrastructure lies in the removal of existing or potential barriers to entry that might distort competition by favouring some competitors over others. These barriers also include technical requirements (for example, those related to incompatible rolling stock and tracks) and safety standards (in terms of a common minimum level). In summary, the general rule should be to promote open access as widely as possible once the separation between the natural monopoly infrastructure and train operations has been effectively achieved. However, this process must depend upon a detailed analysis of infrastructure costs and the prices charged to cover them.

Barriers to entry are also related with intermodal competition. As already mentioned, modal choices can be heavily distorted due to different cost coverage ratios and the use of different cost input bases. A solution is to follow an integrated, multi-modal approach. Basic principles will have to apply to all transport operators, irrespective of the mode in which they operate. For example, in countries like Argentina and Chile, the extent of road freight transport competition was considered in designing the rail concession contracts. The general rule was that operators undertaking business at their own commercial and financial risk should not be at an undue disadvantage to those who enjoy public aid or indirectly benefit from huge externalities.

In the case of rail infrastructures, the principles envisaged to avoid these distortive effects should be solidified in the coordination of existing networks (particularly in dense rail areas) and the establishment of mechanisms that facilitate inter-operability and international links. However, not even the most advanced infrastructure regulations, such as the Swedish and the British systems, offer much help since they were conceived for a single-country environment. In other countries, such as Argentina before the restructuring process, railways attempted to solve national transport problems (by offering under-priced passenger services or subsidized low-quality freight transport). As a result, their financial performance rapidly deteriorated in an isolated framework. Therefore, the infrastructure pricing strategy in these areas should be compatible with the achievement of both local and international objectives, by establishing, if needed, a system of slot assignments in more congested corridors.

3.2. The degree of vertical separation

One of the most clearly defined patterns emerging from deregulation and restructuring is that they carry out two critical dimensions summarized in Table 2: the degree of *vertical separation* between infrastructure and services, and the involvement of *private management* in the sector. With respect to the first dimension, there are three main options for the vertical organization of the railway industry: (i) *vertical integration*, (ii) *competitive access*, and (iii) *vertical separation*.

The first option corresponds to the traditional, historic model of railway organization described above, where a single (usually public) entity controls all the infrastructure facilities as well as the operating and administrative functions. Less frequent, competitive access is characterized by the existence of an integrated operator, who is required to make rail facilities (tracks, stations, etc.) available to other operators on a fair and equal basis through the trading of, for example, circulation rights. This has the advantages of integration (economies of scope, coordinated planning and reduction of transaction costs), but its overall effectiveness may be jeopardized if the integrated company has incentives to leave out other operators.

Alternatively, in the complete vertical separation scenario, the management (and, possibly, the ownership) of facilities is fully separated from other rail functions. This is very attractive because although infrastructure may remain a natural monopoly, it is separated from rail services, where potential competition among different operators is possible. In general, the main advantage of this vertical unbundling is that rail transport is placed in a similar situation as road transport, especially regarding the tariff system and infrastructure planning. Governments could study investment proposals on the basis of a cost-benefit analysis, while pricing policies could be based on social cost.⁴ In addition, separating infrastructure from services greatly facilitates the entry of more than one operator on a single route. For profitable services this would permit notable

⁴ Note that an important problem here is the difficulty of defining the social cost of use of railway infrastructure. Determining the marginal or incremental costs of the use and wear and tear of one additional train is not, in principle, any more difficult than the equivalent calculation for road transport. The problem, however, is greatly complicated for the railway when this cost is evaluated in a congested environment. In pure economic terms, this cost is the opportunity cost of the stretch of track in question, but in practice, it is difficult to quantify this opportunity cost, especially if there is a mixture of social and commercial services.

improvements in efficiency by allowing direct competition among operators. For non-profitable services, infrastructure separation can be accompanied by tendering, to stimulate increased efficiency through competition for the market, the introduction of innovations, and marketing improvements.

Table 2: Alternative organizational structures for the rail industry.

		<i>VERTICAL UNBUNDLING</i>		
		Total vertical integration	Competitive access	Vertical separation
PRIVATE PARTICIPATION	Government Department	India, China, former socialist countries.		
	Public Enterprise	European railways		
	Reformed Public Enterprise	Many European railways at present		Sweden
	Service Contract with Private Sector		Japan (HSR) US (rolling stock) Pakistan (ticket sales)	UK (rolling stock)
	Management Contract with Private Sector	Nigeria (1980)	US small railways	
	Leasing to Private Sector		Amtrak (USA) (track) VIA (Canada) (track) Japan (track) Cameroon (baggage)	
	Leasing from Private Sector		US and Europe (wagons and cars)	
	Concession (franchising)		Argentina, Brazil, Chile, Cote d'Ivoire	UK (passengers)
	Joint Venture		Canada US (pipe and wire)	UK
	Private Company	New Zealand	Japan, US (Class I), Canada	UK (freight, infrastructure)

Source: Elaborated from Galenson and Thompson (1993).

However, the vertical unbundling of the rail industry also implies several disadvantages. The main problem is the potential loss of economies of scope derived from the joint operation of tracks and services. It is often noted that the relationship between the services supplied and the rolling stock used, as well as the quality, quantity and technical characteristics of the infrastructure, is so close that both aspects need to be planned together. Thus, assigning different services to several operators may decrease the utilization of the sector's staff and physical assets. Another negative factor is the higher risk that the new system becomes less attractive to the user than an integrated system.⁵ It is also mentioned that vertical separation requires such a complex institutional arrangement that the resulting transaction costs will be often prohibitive for many countries. A final disadvantage of vertical separation is the reduction of investment incentives. For example, an infrastructure owner considering an investment on a facility with only one potential buyer will anticipate bargaining away some of the benefit from the new service once it comes on line. This problem becomes less relevant with more competition in the market, since competition weakens the bargaining position of individual operators by reducing the specificity of the assets.

⁵ For example, because of the lack of interchangeable ticketing, an integrated national network, etc.

3.3. *The amount of private participation*

With respect to the dimension of private participation in the industry, Galenson and Thompson (1993) provide a list (ordered in terms of increasing private participation) of the different situations that can be found in the world's rail industry. The first situation is a *government department*, where the railroad is fully controlled and financed by the government and therefore subordinated to its interests.

The second example is a *public enterprise*, where the railway is characterized by a higher managerial autonomy, but is still requires government approval for many decisions. Normally, these railways sign *contracts* (or have sectoral laws) with the government, specifying each party's objectives and attributions and the financing rules. Similarly, the case of a *reformed public enterprise* corresponds to a situation where the railway is incorporated (into a shareholding company), commercialized (financially and managerially autonomous), and made subject to the country's company law. However, the government, as the main owner, determines pricing policies and investment levels, while guaranteeing the supply of non-economical social services with the necessary subsidies.

There are other situations that include mixed forms of cooperation between private and public capital. For example, rail service in some countries is provided through a *service contract* with the private sector, where, maintaining full ownership, governments or public enterprises can contract activities to be performed by private sector entities, including food catering, medical services, ticket sales, maintenance of physical assets, etc. Related to these there are *management contracts* with the private sector, where the contractor assumes responsibility for the operations and maintenance of certain activities. One variation is *leasing* to the private sector, where the contractor pays a fee for the use of the fixed assets. The lease contractor has more autonomy than in management contracts, controlling aspects such as the working capital and staff, but also assumes more risk. The owner maintains responsibility for investment and debt service. In many countries, locomotives and wagons are sold or leased to non-railway entities for transporting very specialized goods.

Concessions are a broader form of lease where the contractor also agrees to make certain fixed investments and maintains the use of the assets for a longer period. This is currently the preferred restructuring method in the rail industry and will be extensively discussed in the rest of this chapter. Finally, *joint ventures* entail the largest degree of private participation. Private partners contribute development capital and planning and management expertise to develop land or other real estate owned by a railway. Also, under full *private ownership*, certain services or whole companies are operated by private firms.

4. **New regulatory scenarios for the rail industry**

The vertical separation/private participation bi-dimensional space discussed above creates a new regulatory framework in the rail sector whose most relevant characteristic is the flexibility. It introduces significant new roles and functions for the regulator and modifies the number of possible regulatory structures and models. In practice, choosing a particular method for railway restructuring depends on a number of particular

objectives or goals that the Government must balance according to the economic environment in which it operates.

One of the first elements to consider is the existence of *financial constraints*. If they are important, the maximization of the proceeds obtained from the restructuring process will be a primary goal. A second element to consider is the pursuit of *internal (or cost) efficiency* in terms of providing services at the lowest possible cost, and therefore generating an efficient use of resources. Similarly, there is the goal of attaining *allocative efficiency* by setting optimal prices equal to the marginal social cost, which from an intermodal viewpoint, facilitates the best distribution of traffic. The objective of *dynamic efficiency* requires the long run minimization of cost through active, technology-improving investment policies. There can also be *equity objectives*, such as facilitating transport for all citizens, independent of income level. Finally, the government can also consider the optimal *allocation of capacity*, which favours management of railway capacity, coordination with other modes of transport, and overall *minimization of risks* in terms of service maintenance over time, risk of default, etc.

Table 3: Different rail regulatory scenarios and their objectives.

Scenario	<i>Objectives</i>						
	Financial constraints	Internal Efficiency	External Efficiency	Dynamic Efficiency	Risk Minimizing	Capacity Allocation	Equity
(1) Vertical Integration and Government Department	-	-	+	+	+	+	+
(2) Vertical Integration and reformed public enterprise	-	-	+	+	+	+	+
(3) Vertical Separation and Reformed public enterprise	-	-	+	+	+	-	+
(4) Competitive Access and Concession regime	+	+	+	unclear	unclear	-	+
(5) Vertical separation and Concession regime	+	+	+			-	+
(6) Vertical integration and Private enterprise	+	+	-		-	+	-
(7) Competitive access and Private enterprise	+	+	-		-	-	-
(8) Vertical Separation and Private company	+	+	-	-	-	-	

Note. A “+” sign means that the objective is easily achievable within the corresponding scenario. The “-” sign implies the opposite.

Table 3 summarizes the combination of these objectives, creating at least eight different possible regulatory scenarios, grouped in decreasing order of private

participation. Some additional scenarios, such as the mixed forms described above, have not been included.⁶

It is important to note that the objectives could be given a different weight. For example, financial and cost efficiency objectives are now valued above all others, which explains the privatization boom, through concessions and direct sales to the private sector. In addition, as the degree of privatization increases, there is a trade-off between social and financial efficiency objectives. The public company scenarios serve social objectives (equity, reduction of risk on the service, intermodal coordination, etc.), but there are inefficient, leading to huge commercial deficits. As we have already indicated, this was the main reason for the restructuring of the sector.

The deregulation measures that define scenarios 4 and 5 (*concessions*) have the advantage of favouring the efficiency and solvency of the companies, as well as reducing the state's financial burden (although these effects are possibly not as great as with direct privatization). In addition, concession contracts allow the cushioning of some of the negative effects that may arise from the actions of the private company. Thus, it is habitual to establish maximum prices and minimum service levels so that impact on equity can be minimized. Likewise, many routes which, though not profitable, are beneficial from a social viewpoint can continue to be served: concessioning them to operators who request lower public subsidies meets both efficiency and equity objectives.

In regard to dynamic efficiency, the first results of the investments implemented by the restructured companies or bodies are ambiguous. In Argentina, the investment levels of some operators have been below those foreseen in their concession contracts, though at the aggregate level, investment levels seem to have improved. Something similar has occurred with some passenger franchises in the United Kingdom. At any rate, the effective investment levels should be compared with those that existed in the regulated context. In this sense, other experiences have indeed led to a substantial recovery in investments in both infrastructure and rolling stock, as well as an improvement in service quality. In other countries, such as Japan, privatization does not seem to have slowed the technological development of the railway industry (Fujimori, 1997).

Apart from other considerations, operational risks are minimized when entrusted to a public enterprise. With a private company, there is obviously a greater risk of closure of certain services, or of larger instability. Again, concession systems allow the risks inherent to the action of private enterprise to be reduced.

Finally, the problem associated with managing capacity is easily eliminated in the case of vertically integrated companies, although this is not so simple for systems of competitive access or separation. In this case, the problem is increased for companies with high traffic densities and conflicting capacity demands. Modern computer technology can reduce the problem through real-time management of electronic systems, but when connecting systems have different informational qualities and dispatching priorities it is very difficult for anyone to plan and manage integrated services across several systems.

⁶ This is because many of these forms of private participation are related to very specific services (e.g., the case of *service* or *management contracts*) and on occasions some of the forms of contracting (e.g., *leasing*) are very similar to those established in a concession or franchising system.

5. The role of concession contracts in the rail industry

In spite of the number of potential regulatory scenarios just described, few railways around the world have been fully privatized. Instead, most countries have opted to concession rail services and even rail infrastructures in some cases, to private firms in exchange for a fixed payment. This has been the favoured form of restructuring because it allows the government to retain ultimate control over the assets while the private sector carries out day-to-day operations according to pre-specified rules devised in a contract that transforms the problems associated with traditional regulation into issues of contract enforcement.⁷

Since there are many variables to consider, rail concession contracts cannot be reduced into a single standard model. However, according to existing experiences, Table 4 proposes four key variables to consider.

Table 4: Some key variables in rail concession contracts design.

Type of contract	<p>Package size depends on economies of scale/scope and existing potential for competition</p> <p>Horizontal concessions (geographic) according to country's characteristics</p> <p>Vertical concessions (functional) according to network's characteristics (including current state of infrastructure and new investment needed)</p> <p>Mixed packages depending on profitability and bidders' financial constraints</p> <p>Freight vs. passenger concessions depending on relative traffic shares</p>
Award and duration	<p>Pre-qualification requirements to reduce risks</p> <p>Type of auction (sealed, one-shot) and explicit rules for auctioning</p> <p>Selection based on government's objectives (fiscal, equity or efficiency)</p> <p>Short periods (favour competition; diminish investment incentives) versus long periods (favour investment; diminish enforceability)</p> <p>Termination: re-auction preferable to automatic renewal</p>
Operating general contents	<p>Concessionaire:</p> <ul style="list-style-type: none"> obligations: services (with adequate performance) and payments rights: exclusivity and compensation for public service obligations <p>Government:</p> <ul style="list-style-type: none"> risk sharing (net cost/gross cost mechanisms) asset ownership rules
Regulation mechanisms	<p>Price control rules (services and infrastructure)</p> <p>Principles regarding price discrimination and cross-subsidization</p> <p>Definition of quality targets and quality control</p> <p>Issues regarding safety and externalities</p>

The first critical aspect of a concession is determining its *type*, both in vertical (functional) size and horizontal (geographical) size. Recent concessions in the rail

⁷ The list of countries with actual or planned rail concessions include, among others, United Kingdom, Argentina, Chile, Brazil, Bolivia, Peru, Colombia, Guatemala, Mexico, Côte d'Ivoire-Burkina Faso, Cameroon, Congo, Malawi, Jordan and Mozambique.

industry have created smaller horizontal packages throughout the country. For example, rail freight systems in Argentina, Brazil, Mexico and Colombia were split into several regional companies, and Chilean railways were broken down into four passenger companies and two freight companies with a separate infrastructure firm. All of these countries also used economic criteria to design the size of the concessioning package, accounting for the profitability of different lines.

In Europe, functional separation between infrastructure and services has been preferred, especially after European Commission Directive 91/440. At its most extreme, this form of concessioning was used in the privatization of British Rail, which also included the private provision and management of rail infrastructures. A less extensive vertical separation has been developed in Sweden and other European countries, where infrastructure has not been auctioned off to private firms (Lundberg, 1996). However the debate about the advantages and disadvantages of the separation of infrastructure and operations is not closed. There is a perception that separation is an essential condition for non discriminatory access, and that it is very difficult to increase competition in a situation in which the major operator controls the infrastructure. But the problems of British process have increased the doubts about its advisability. Nash (2004) points that perhaps the Swedish model, combining a stated-owned infrastructure entity and a charging based on short run marginal cost, is the system that has worked with best results in Europe.

The second key issue in designing rail service and infrastructure concession contracts is defining the *award process* and *duration* of the concession. This includes the auction rules and, particularly, the criteria defining how each concession will be awarded to a private operator. There are a number of possibilities to choose from as the award criteria (for example, maximum payment to government or minimum tariff). There is also a choice between unrestricted bidding and bidding that could involve some pre-selection (see Guislain and Kerf, 1995, and Kerf *et al.*, 1997). In the privatization of the former British Rail, for example, the concession process began with a pre-qualification stage, followed by a formal invitation to tender for a particular package. After indicative bids were received, four bidders were short-listed. One of these was subsequently named the preferred bidder, and was given a fortnight to complete financing and other organizational arrangements before being confirmed as the winner. At that point, the regulator gave public details of the bid, in terms of the required subsidy and promised service improvements.

With respect to bidding mechanisms, there is extensive literature on experiences and results in different auction forms. Single, sealed-envelope bids is the simplest, avoiding collusion and obtaining higher bids. However, more complex approaches, such as real-time auctions, have been used in some transport concessions. Once the rules have been set up and the bids requested, bidders should have a study period to form their own evaluation of the potential gains to be extracted from the concession. Early research by Preston *et al.* (1996) for the United Kingdom indicated that key issues for bidders were the length of franchises, the level of competition they would face from other operators, the separation of infrastructure from services, the costs (including new investments) associated with maintenance and the selection criteria for the bidding process.

Although the guiding principle should be to maximize competition so that the most efficient firm ends up winning the award, it is clear that there is no single method for selecting the winner once bids have been submitted. The final choice depends on the Government's objectives, which should be explicit and built on transparent criteria.

Thus, if the government intends private participation to be a means of reducing the burden on the public sector, it must use fiscal benefits as the main criterion, looking at who requires the lowest subsidy or who offers the highest auction price. In Brazil, for example, the six regional rail concessions were successfully auctioned to the highest bid above the Government's minimum price. Concessionaires were required to make an up-front payment immediately after the auction, followed by a stream of pre-determined payments over the life of the concession. Similarly, in Britain, minimizing subsidy payments appeared to drive the regulator's choice of bidders, especially in the first concessions. Other criteria were the financial position of the tenderer, its managerial competence and its operational proposals.

Alternatively, if tariffs and quality of service are defined in the contract, bids can be evaluated on the basis of the lower cost provider, simultaneously including penalties for not achieving certain performance objectives. Social objectives can be also targeted by focusing on the bids that propose to monopolize the industry for the lowest number of years or to charge the lowest fare to final users. Sometimes, as in the case of rail freight, the traffic mix makes the price structure very complex, so that this mechanism becomes impractical. Moreover, using tariffs as an award criterion for rail concessions limits the later possibility of regulatory intervention in prices and demands an adequate definition of quality standards.

Many concessions in the rail industry have been awarded using formulas with multiple criteria, which can account for a larger number of objectives. For example in Argentina, the bids for the six freight packages that were concessioned were evaluated using the net present value of the canon to be paid to the government during the first fifteen years of the concession, the quality of business and investment plans, staffing levels, the proposed track fee for passenger trains, and the share of Argentine interest in the consortium. The weights of these criteria reflected both the importance attributed to investment in the railways and political compromises on employment. However, for the award of metropolitan commuter railways, the Argentinean authorities kept things simpler to make the bidding process and final selection as transparent as possible. They learned from the freight concession that selecting the winning bid through numerous cumbersome criteria with discretionary weights was more likely to reduce the efficiency of the bidding process than to improve it. Instead, the terms of the concession should be made clear to all potential bidders and bidding should take place on the basis of a single parameter encompassed in the bidders' economic assumptions in terms of the concession.⁸

With regard to the optimal *duration* of the concession contract, the trade-off is evident in terms of efficiency, since the shorter the concession, the more immediate the competitive pressure, but the lower the incentive to invest and develop the business. Longer concessions, in contrast, tend to diminish the regulator's enforcement capacity and soften the incentives to promote efficient outcomes. The general rule is to adapt the concession period to the economic life of the assets and to make this compatible with the government's objectives. This balance often creates conflict: while concessionaires generally argue for long contracts that provide them with incentives to build up the business and purchase or replace long-lived assets, concessioning authorities prefer shorter lengths to favour the achievement of efficiency (by the implicit threat of not

⁸ In the case of the metropolitan railway concession, for instance, each concessionaire calculated her expected revenue from operations, then compared it with the capital investment programs and finally estimated the subsidy amount to be requested (The World Bank, 1996).

renewal) and fiscal goals (since the canon or auction price may be increased after the first few years of the concession). Only if sunk investments are minimal and asset reutilization is possible, are shorter periods advisable for particular rail services (those related to signals, track and station maintenance).

Shaw, Gwilliam and Thompson (1996) point out that the average duration of a rail service concession is five to ten years, increasing up to thirty when network investment and development are included. In Argentina, for example, the six freight packages were concessioned on a thirty year term, with an optional ten year extension, due to the poor state of infrastructure and the huge investment that was required. For similar reasons, the international rail link between Côte d'Ivoire and Burkina Faso was awarded in a fifteen year concession. Conversely, train operating companies in the United Kingdom were granted a concession to run passenger services for a period of only seven to fifteen years.

After the duration period has expired, the contract must also specify several *termination arrangements* to avoid any disruption in services. One possibility is to make automatic renewals in case new candidates for the concession do not exist. The regulator should not compromise on this before the concession ends in order to ensure that the incumbent has the correct incentives. New auctioning seems to be the standard procedure after a concession has ended, but most rail operators will seek a renegotiation of duration terms while the contract is still in force. Examples of this strategy are some United Kingdom rail franchises who argued that they made long-lived investments in high-quality wagons and locomotives when they asked for a license extension.

Since renegotiation costs money, but a lack of renegotiation might cause performance deterioration, concession contracts should specify the circumstances for renegotiation, and which party should initiate the process. If intermediate objectives are achieved, a pre-scheduled revision process might help to reduce both parties' risks. Although the contract will always be incomplete, standard clauses should include behaviour in unforeseen changes in demand conditions, responses to unanticipated rises in energy or labour costs, etc. For example in Argentina, freight concessions could not fulfil their promise to invest \$1.2 billion in the rail network over fifteen years due to unexpected falling traffic levels.

A flexible contract renegotiation mechanism is a good idea in any case since the Government may face the dilemma of enforcing contracts to the detriment of the operating companies and the national rail system or rescheduling investment and making other compromises at the cost of undermining his credibility for enforcing future agreements (Carbajo and Estache, 1996).

This is why one of the most critical issues in designing a rail concession contract is the *attribution of rights and obligations* to the parties. On one side, the private operator pays a regular canon or receives a subsidy and is awarded the right to operate train services and/or manage its infrastructure (including future investments) with (total or partial) exclusivity rights that protect her from other competitors. On the other hand, in exchange for the payment or the compensating subsidy, there is a regulatory activity by means of which the overall performance of the sector is monitored and a stable framework for current and future rail operations is provided.

These operations may include infrastructure provisions if they were auctioned off to private firms. In fact, a large part of railway activities might be concessioned. These include infrastructure: track, signals, stations, yards and shops; operating equipment: locomotives, wagons, carriages; and general service access to track, route and schedule

information and maintenance. The exact form in which this process is developed in practice depends on the parties' risk sharing agreements. According to a service contract, for example, train operators provide rail transport services for passengers or (rarely) freight according to specific routes, levels of quality and technology as established by the regulator. The operators may cover some investment costs and carry some commercial risk, which can be integrated into a net cost contract, where the operator keeps all revenues generated by passenger or freight traffic. This type of contract, where the operator carries revenue as well as cost risk, often generates more traffic and is let to the most attractive bid, but offers a higher incentive to predate. Alternatively, gross cost contracts specify that all revenue accrues to the government and the contracts are let on the basis of the least total cost supplier so operators carry cost but not revenue risk. The experience in the United Kingdom with regard to passenger franchises suggests that gross cost contracts generate more bids per tender (particularly from new entrants), offer greater incentives to public revenue generation, reduce the administrative cost for the regulatory authority, and support any fare scheme with modal integration and quality control.

The regulator may retain control over and responsibility for common functions, and its main roles should be restricted to regulating quality (in terms of service, safety, environmental and technical standards), controlling monopolistic behaviour (in terms of abusive prices or services), and determining the overall characteristics of the function of the sector (in terms of coordination at the national and international level) according to established the competition rules or rights and anti-trust and commercial legislation.

The implementation of rail concession exclusivity rights varies in each country. In Argentina, freight concessionaires have exclusive use of tracks but must grant access to passenger operations in return for a compensatory track fee. In Chile, passenger services and infrastructure initially remained in public hands, while freight concessions were awarded to private competing firms. The fifteen-year concession for the Côte d'Ivoire-Burkina Faso trans-national railway was awarded with a seven year exclusivity period, after which the operator should grant access to third-parties specified by the regulator for an agreed fee. Thus, exclusivity rights should be viewed as another instrument for regulatory control, and not taken for granted by the firms *ex-ante*. Limiting the duration of the monopoly period balances the regulator's desire to reap the benefits of competitive access to the tracks and the private train operators' preference for full control of the market to generate profit and facilitate revenue forecasting. In general, most railways have been concessioned on an exclusive basis in geographical areas, as in Argentina or Brazil, possibly with some access rights for connecting railways to certain central or strategic track segments. This has been due to the geopolitical configuration of the country, the density of the existing network, and the need to promote competition in major markets (as in Mexico) or for non-competing services (such as passenger services on freight tracks in Chile).

With respect to the concessionaires' obligations, the private provision of rail transport services, particularly in less developed areas or zones with a structural lack of network, cannot always be separated from public subsidization or reciprocal compensation for politically motivated public service obligations. Arrangements for these loss-making but socially necessary services must be included in concession contracts, in terms of detailed performance levels to be attained by the firm, possibly even be designed to be awarded to the company willing to provide the specified services for the lowest level of subsidy (negative concessions), as in Argentina.

A final feature of defining the rights and obligations of the concessionaires, the current experience of rail concessions in South America shows that restructuring has often lowered employment levels. This is, in practice, one of the toughest obstacles hindering the private participation process in certain countries and often requires difficult political decisions. In Brazil, for example, large redundancies were inevitable and were dealt with in two phases. Before concessioning, incentive schemes for early retirements were in place; after the concession was awarded, the former national rail operator paid involuntary separation grants to the remaining staff not hired by the concessionaire. After that point, compensation for additional laid off employees is the responsibility of the private operator. Undoubtedly, any such employment constraints will be reflected in the auction price of the concession.

In summary, in their general form, rail concessions are the most advantageous solution to the challenges posed by the current regulatory environment of the rail industry. It usually adopts the form of a long or medium term contract where a vertically or horizontally integrated package of (passenger and/or freight) rail services is auctioned off to private firms, while economic assets remain public property. Three of its key features – type, duration and contents – have been described in this section, but there are other particular aspects of the concession contract design in the rail industry that, based on their importance, deserve a more detailed treatment. These include price regulation, in terms of defining the most important issues in establishing effective and well-oriented price control mechanisms; quality regulation, in both its static dimension (quality of service, safety and environmental issues) and dynamic dimension (rules for infrastructure investment and financing), and coordination between infrastructure and superstructure.

6. Conclusions

The increasing role of private sector in the rail industry is one of the most relevant characteristics of the evolution of this industry in recent years. This change is reshaping the way in which Governments are addressing the main regulatory challenges derived from the economic and technical characteristics of railways.

In this paper we have showed that the industry regulation is moving accordingly towards more flexible schemes of public intervention. Although this does not fully preclude direct participation by the Government, it seems that the monopolistic rail company is progressively disappearing as the dominant model around the world. There is no unique form of rail regulation to address these new challenges, but the general rule is to maintain flexibility and simplicity whenever possible.

Two key issues in the new regulatory environment of the rail industry are that private participation is included in license contracts and the organization of the industry is adapted to each country's needs and characteristics. In turn, the use of these mechanisms also changes the role of the rail regulator, whose actions should now be governed by principles that foster competition and market mechanisms and simultaneously provide a stable legal and institutional framework for economic activity. The regulator should refrain from intervention unless the ultimate goal of achieving economic efficiency subject to the socially demanded level of equity is in jeopardy.

Nevertheless, two important caveats for future regulation must be taken into account. First, the process of privatization chosen in each country depends on the basic objectives sought: to maintain an industry with one operator or a small number of operators, or to facilitate a process of competition on the track. Second, legacies from the traditional mechanisms of regulation should be avoided. In particular, high debt levels and overstaffing are two common problems that must be dealt with before starting any privatization policy.

In any case, future researches will be necessary to evaluate the advantages and difficulties of the current rail restructuring processes in the world. Some of these costs and benefits have been described in the paper. The new regulatory schemes will be essential in order to preserve the advantages of these new rail systems and to reduce its potential problems and costs.

References

- Baumol, W. J. (1977). "On the Proper Cost Tests for Natural Monopoly in a Multiproduct Industry". *American Economic Review*, 65, 5, pp.810-822.
- Baumol, W. J., Panzar, J. C. and Willig R. D. (1982) *Contestable Markets and the Theory of Industry Structure*, Harcourt Brace: Jovanovich, New York.
- Carbajo, J. C. and Estache A. (1996) "Railway Concessions – Heading down the Right Track in Argentina", *Viewpoint* 88. Finance and Private Sector Development Department. The World Bank: Washington, D.C.
- Fujimori, Y. (1997) "Ten Years of JNR's Privatization (Technological Development after JNR's Reformation)", *Japanese Railway Engineering*, 138, pp.3-5.
- Galenson, A. and Thompson, L.S. (1993) "Forms of Private Sector Participation in Railways", *TWU Papers* No. 9. The World Bank: Washington, D.C.
- Gathon, H. J. and Pestieau, P. (1995) "Decomposing Efficiency into its Managerial and its Regulatory Components: The Case of European Railways", *European Journal of Operational Research*, 12, pp.500-507.
- Guislain, P. and Kerf, M. (1995) "Concessions – The Way to Privatize Infrastructure Sector Monopolies", *Viewpoint* 59. Finance and Private Sector Development Department. The World Bank: Washington, D.C.
- Kopicki, R. and Thompson, L. S. (1995) *Best Methods of Railway Restructuring and Privatization*, CFS Discussion Paper Series No. 111. The World Bank: Washington, D.C.
- Lundberg, A. (1996) "Restructuring of the Swedish State Railways", *Japan Railways and Transport Review*, 218. pp.22-26.
- Nash, C. A. and Rivera-Trujillo, C. (2004) "Rail regulatory reform in Europe- principles and practice", STELLA Focus Group 5 Synthesis Meeting, Athens.
- Oum, T. H. and Yu, C. (1994) "A Comparative Study of OECD Countries' Railways", *Journal of Transport Economics and Policy*, 38, pp.121-138.
- Preston, J., Whelan, G., Nash, C. A. and Wardman, M. (1996) *The Franchising of Passenger Rail Services in Britain*, Institute of Transport Studies: University of Leeds.
- Shaw N. L., Gwilliam, K. M. and Thompson, L. S. (1996) "Concessions in Transport", *TWU Papers*, No. 27. The World Bank: Washington, D.C.
- Thompson, L. S. and Budin, K. J. (1997) "Global Trend to Railway Concessions Delivering Positive Results", *Public Policy for the Private Sector*, No. 134. The World Bank; Washington, D.C.
- Waters II, W. G. (1985) "Rail Cost Analysis". In Button, K. J. and Pittfield, D. E. (eds.). *International Railway Economics*. Gower: Aldershot.

Acknowledgements

Pedro Cantos thanks financial support from Spanish Ministry of Science and Technology under project SEJ 2004-00110. Javier Campos gratefully acknowledges financial support from the Spanish Ministry of Science and Technology and from FEDER through grant BEC2002-02527. Both authors are grateful to The World Bank for collaborating in previous versions of this paper.



Recent changes in the global rail industry: evaluating the new regulatory instruments

Pedro Cantos^{1*}, Javier Campos²

¹ *Departamento de Análisis Económico (Universidad de Valencia)*

² *Departamento de Análisis Económico Aplicado (Universidad de Las Palmas de Gran Canaria)*

Abstract

The changes faced by the global rail industry in recent years have brought a redefinition of some of the traditional regulatory instruments available in this sector. This paper, focusing on price and quality regulation, discusses how these instruments have been applied in several countries where private sector participation in railways has been introduced mainly through concession contracts, and where some form of vertical and/or horizontal unbundling has been implemented.

Keywords: Railways; Price regulation; Quality regulation.

1. Introduction

After enjoying an unchallenged position for more than 100 years as the dominant means of transport, the rail industry has globally faced a dramatic change both in terms of economic relevance and organizational structure during the last decades. The decline of the railways has been partially explained by the government involvement in its management and the pervasive effects of an obsolete regulatory framework, which impeded, or at least slowed, the necessary adaptation to a changing environment dominated by more flexible transport alternatives.

Narrowly classified as natural monopolies since the XIX century, railways' management around the world widely relied on an undisputed model based on a vertically integrated firm, heavily protected from competition which acted as a national provider of a public service and received generous support from the Government. With very few exceptions, this was the paradigm until the 1980s, when a series of reforms, in the UK, Chile, New Zealand or Japan proved that competition could be introduced in this model through horizontal and/or vertical unbundling, and the subsequent increase in

* Corresponding author: Pedro Cantos. Departamento de Análisis Económico (Universidad de Valencia). Campus dels Tarongers, s/n. 46022 Valencia (Spain). E-mail: Pedro.Cantos@uv.es

private participation in the provision of services and, although less successfully, in the management of infrastructure.

Within the traditional railway model, pricing and quality decisions were heavily regulated and political interference in managerial decisions often affected these aspects of the railway companies. In fact, pricing rules were relatively simple: in most cases the overall scheme was characterized by maximum prices with little connection with costs, combined with cross subsidization, through which some profitable services pay above their avoidable costs maintaining unprofitable services paying below their avoidable costs. Subsidies, not necessarily associated to public service obligations completed the picture. With respect to quality, few commercial provisions were in practice, since the Government-owned nature of most companies prevented them from making a real effort on improving this issue.

For these reasons, the main aim of this paper is to discuss some of the new regulatory instruments on price and quality regulation that have recently become of common use in the countries which have opted for a change in their railway organizational paradigm. In section 2, we will first review the principles behind the price regulation mechanisms governing the provision of (mostly, passenger) services in a context of a possibly unbundled rail industry company enjoying a significant degree of private participation (usually, through a concession contract). In section 3, we specifically study in detail two of the major problems arising in the regulation of rail infrastructure, provision and access. Finally, since tariff controls can easily be cheated on quality grounds, quality requirements become essential for monitoring overall performance of rail concessionaires. We thus address the issue of quality regulation, including safety concerns in section 4; this includes not only the adequate definition of quality targets, but also a review of the most relevant mechanisms for quality control currently used in the rail industry. The final section describes some performance indicators that could be applied to monitor the behaviour of the regulated rail companies, thus providing a useful device aimed at moving from the definition of the regulation theoretical principles to the problem of how to implement them.

2. Price regulation of rail services: principles and mechanisms

According to standard economic principles, prices for rail transport services should match the opportunity cost of providing it in order to make the most efficient use of the economy's resources. This is the economic efficiency or *first best* criterion which has defined the traditional regulation of the rail industry during the last fifty years. The main focus of government regulation was controlling market power by setting prices that limited the monopolistic abuse of any particular railroad. The exact form of tariff control (official approval of rates with little or no degree of financial autonomy) in each case depended on the nature of the industry, the ownership of the assets, the complexity of the regulated service, and the social and political pressures to maintain financial equilibrium in the medium and long run.

In practice, however, opportunity cost pricing implies some measurement difficulties and often conveys economic losses, especially in industries with large economies of scale (Armstrong, Cowan and Vickers, 1994). Therefore, this form of regulation was complemented by a number of standard price mechanisms that economic theorists

devised to substitute the ideal efficiency criterion of pricing each unit of service at the exact cost of its provision.

Price discrimination policies, either by type (student and senior prices, frequent traveller and commuter passes), number of consumers (group discounts), type or volume of freight (cargo rebates for some goods) or time of day or season (peak-load prices), have always been common in transport. The use of two-part tariffs, with fixed and variable components, is also a common tariff policy in which each unit of consumption (for example, a single trip) is priced differently. These mechanisms allow greater flexibility for railways and increase revenues without a great effect on costs. However, their social acceptability and information requirements can limit the extent of their application.

In the new regulatory environment defined by the changes experienced in the rail industry since the 1980s, where separation of the infrastructure from services has been widely implemented in diverse forms, and a notable degree of private participation in rail management exists through, for example, concession contracts, pricing principles must be put into practice by means of concrete rules within the contract itself.

In general, as private operators, rail concessionaires are allowed to set prices relatively freely, price regulation has a different nature: instead of price-setting, it becomes more price-supervision. To carry out this task, most concession contracts awarded in the rail industry (for example, in Argentina or Brazil) routinely include a specific procedure to control and evaluate the prices set by operators. These price control mechanisms are generally set according to three key factors: *(i)* the degree of monopoly power effectively conferred to the operator; *(ii)* the extent of government non-commercial objectives in the concession award procedure; and *(iii)* the possible existence of other limiting factors, such as intermodal competition. This latter element is relevant in rail freight operations (intermodal competition from trucking), but in the case of passenger traffic (especially commuter and regional), social pressure for low fares usually dominates many price interventions. In practice, the most common alternatives (*second best criteria*) for price control in rail concessions adopt the form of a rate of return regulation or a price cap mechanism.

2.1. Rate of return mechanisms

Rate of return regulation is used in railroads in Canada, Japan and the United States. The principle behind this type of regulation is to constrain prices so that the regulated rail transport operator earns only a fair rate of return on its capital investment. The regulator typically determines a revenue requirement based on a firm's total costs during a test year, according to the variable costs and an estimate of the cost of capital to the firm, given by a "reasonable" rate level multiplied by a base rate (Liston, 1997).

Thus, rate of return regulation has three components: the base rate, the allowed rate level, and the rate structure. The base rate refers to the investments that are allowed to earn a rate of return, the rate level refers to the relation of overall revenues to costs, and the rate structure determines how individual prices are set for different services or customers. Determining the first of these three components is often the most important regulatory task under this form of regulation, since inadequate calculations of the base rate may either jeopardize the survival of the firm or allow it to earn excessive profits.

In practice, the base rate usually includes most fixed costs less depreciation and working capital.

Three characteristics should govern the definition of the asset base rate. First, with respect to the treatment of past investments carried out by the railroad before the regulatory period,¹ it should be consistent and transparent in order to ensure that assets are not expropriated *ex post* by opportunistic regulatory behaviour, which would increase the cost of capital required by investors. Second, with future investments and expected operating expenditures and costs should be considered in the asset base definition inasmuch as they do not imply “excessive” investment and only when they are fully incorporated into the firm. Finally, with respect to current investments, a problem lies in determining the value of the firm’s capital. If the existing assets were transferable to other activities without cost, then the conceptual problem of determining their value would be simple: their replacement cost or resale value. At the other extreme, and more frequent in the rail industry, is that existing assets are sunk, so the opportunity cost of using them in their present activity is zero. If the regulator seeks maximum efficiency, it should ensure that the rate of return structure (and, indirectly, the prices) are set to cover future avoidable costs.

Since most of the assets currently used by railways are financed before the concessioning process, both of these solutions are troublesome. Market values are much lower than replacement costs so this valuation would yield large price increases and windfall gains for private shareholders at the expense of consumers. On the other hand, in attributing a zero value to the existing assets, windfall gains would go in the opposite direction and the proprietors would be reluctant to finance future investments with such a lower real return. A possible way to address this problem is to use some average procedure that considers either a financial projection of what will happen with the future base rate or calculates indicative values by estimating the cash flows that the firm would have earned had the regulatory regime remain unchanged.

Despite its advantages within the traditional price regulation mechanisms (mainly its simplicity), three additional problems are associated with this sort of regulation. First, there is little incentive for productive efficiency, since firms can pass production costs on to final users in the form of higher prices; second, it leads to excessive investment and capital use because the firm is guaranteed a return on investment;² and, finally, the high degree of discretion enjoyed by the regulator in determining the base rate and the rate of return reduces the incentive for rent-seeking behaviour on the part of the regulated firm.

2.2. Price cap regulation mechanisms

The most common alternative to the standard rate of return regulation is the use of cost-plus incentives that, in practice, take the form of a menu of cost reimbursement rules that firms themselves select according to their preferences for sharing operating costs with the regulator. The basic aim of these mechanisms is the achievement of dynamic efficiency (in the sense of the regulated firm achieving the lowest unit cost in

¹ This is often the case in many restructuring processes when a former state-owned railway transfers its assets to private concessionaires.

² This is the so-called Averch-Johnson or capital-bias effect, which is not particularly adverse in less developed economies whose capital needs are seldom fulfilled.

the long run) by sharing some of the efficiency improvement rents between the firm and the regulator.³

Alternatively, price cap regulation is another incentive used in both railways and other privatized utilities. In its most standard form, it consists of setting traditional maximum price schemes based on long-run marginal costs in order to offer a firm an incentive to achieve the goal of dynamic efficiency while maintaining all or part of the gains associated with the firm's future increases in efficiency. This mechanism came as a consequence of the criticism directed at the lack of cost minimization embedded in rate of return regulation and other traditional price regulation mechanisms. However, its efficiency gains have to be balanced with the higher information rents that it implies.

There are a number of minor variations of the price cap system. In the rail industry, one of the most developed is the RPI-X formula. In this setup, the price for a basket of the firm's prices can increase in any one year by no more than the increase in the retail price index (RPI) for that year, minus some fixed-cost (efficiency related) parameter X. In the case of multi-product activities, such as railways output, this expression can be easily adapted by requiring that a certain weighted average of percentage price increases not exceed the rate of growth of the RPI less X percent. The weight for each price can be defined according to the share in total revenue of each product or, alternatively, it can be imposed that the average revenue (calculated with accounting figures) can grow at most by RPI-X. Thus, the regulator can control the prices of multi-product firms by focusing on their revenues and correcting them according to adequate weights. It starts with a reference price, often calculated with rate of return criteria, and set the price for a fixed number of years.

In the United Kingdom, for example, the price cap mechanism, in its RPI-X formula, has been applied to passenger traffic franchises. Commuter fares are regulated with respect to a basket containing all relevant fares, weighed broadly by the income that the operator derives from each. For three years from January 1996, increases in the capped fares are not permitted to be more than the retail price index increase from the 1995 base price; after January 1999, the price cap was set at RPI-1%.

The goal of this method is to increase the efficiency of the regulated rail operator, allowing the firm to earn substantial profits by improving efficiency while simultaneously financing current and future operations. This implies that, in practice, when setting the level of a price cap, the rail regulator must consider several factors: the cost of capital, the value of the existing assets, future investment programs, expected changes in productivity, estimates of demand growth, and, perhaps, the effect of X on actual and potential competitors. Some of these are common to other price regulation mechanisms and, in particular, they are needed when using rate of return regulation, as described above.

There are different procedures and rules to deal with each mechanism. The cost of capital and the value of existing assets are calculated using standard financial techniques. The future investment program and its implications depends on both expected changes in productivity and estimated demand that can be obtained from econometric techniques or simpler projection and analysis of historical data. Finally, the

³ There are several ways to accomplish this goal and implement its results. For example, the sliding scale plans used in the United Kingdom's Railtrack regulation consist of a price adjustment mechanism through which the actual rate of return earned by the firm is adapted to changes in productivity according to a variable parameter.

effect of the price cap on the future shape of the market is conjectured from past experiences or yardstick comparisons.

One of the most critical issues is the setting and resetting of the productivity X-factor. A possible method consists of using indexes or indicators (as described below) to measure the difference between aggregate rates of growth of outputs and inputs and therefore calculate productivity from the residual. Econometrics also provides alternatives for estimating cost functions and their corresponding productivity parameters. Once the X-factor is determined, the initial price ceiling that is imposed on the firm after a switch of regime is critical. If the caps are too high, then too little surplus is transferred to consumers and deadweight losses are huge. If they are set too low, the firm may not be able to break even and may then have difficulty attracting capital, leading to a deterioration of quality of service.

Another important element of RPI-X regulation is the existence of cost pass-through provisions, through which the firm can transfer to customers unexpected increases in certain factors outside of its control. Although these clauses are standard in the regulation of other utilities, they are not in the rail industry. The most plausible case could be given by energy costs, for which a certain percentage (100% or less) of the cost pass-through onto customers could be established in the concession contract.

3. Regulation and rail infrastructure

After reviewing the principles and mechanisms of price regulation for rail services, this section addresses the two most relevant problems of infrastructure regulation nowadays in a context of vertical unbundling and private participation. We first focus on the recovery cost problem and then study the issue of access pricing.

3.1. How to recover infrastructure costs?

Rail infrastructure provision and management are characterized by a high ratio of fixed to marginal costs, the existence of avoidable costs and unavoidable or common costs. Avoidable costs are uniquely associated with a particular output: if this output is not produced, no cost is incurred. This guiding principle relates to the idea of cost recovery for particular outputs. Avoidable costs may thus be considered as a floor to regulated prices (if any), since charging less than the avoidable cost is equivalent to operating at an economic loss. This makes standard pricing rules inoperable in this sector, since first best or efficient principles of marginal cost pricing may result in large deficits that jeopardize the long run survival of the firm. Three particular problems then arise with respect to the allocation of the rail infrastructure costs: cross-subsidization issues, cost-recovery problems, and the possibility of setting inefficient prices (Talley, 1988).

The existence of cross-subsidization problems in pricing rail services or infrastructure produced in the presence of common costs can be illustrated with the case of a profit-regulated railroad connecting two large cities and also providing rail service to a smaller town along the route between the two cities. The fares charged for passage from the small town generate revenues exceeding the additional cost of serving it, such as

ticketing and station costs, but not sufficient to cover an equal or proportionate (however defined) share of the common costs, such as trackage, signalling, and trainyard costs. The issue is how to allocate common costs among customers and services. In many cases, cost sub-additivity and efficiency require joint production and allocation of fixed costs among all services, without cross-subsidization (accounting for externalities whenever present).

Cross-subsidization is not only an equity problem for rail services, as in this example, but also a relevant issue for efficient pricing of infrastructure like railbeds, signals or stations. The standard procedure is the so-called fully distributed costs method, under which common costs are allocated on the basis of some common measure of utilization, such as gross tons/km, or other measure of relative output or gross revenue. Alternatively, common costs can also be allocated in proportion to costs that can be directly assigned to the various services (Braeutigam, 1989). The arbitrary nature of fully distributed cost methods and its lack of a conceptual foundation have been criticized, but they remain a useful measure for recovering common costs.

However, the treatment of the cross-subsidization problem should not be based on excessively rigid criteria, particularly for developing countries with few alternative finance mechanisms. The analysis should be made on a case-by-case basis, since, for example, stand-alone cost tests do not apply if railroads are not allowed to abandon unremunerative facilities or services (Kessides and Willig, 1995). If that freedom is denied, a railroad cannot earn adequate revenues if its rates on potentially remunerative activities are constrained by stand-alone cost ceilings.

The cost recovery principle should be a central issue in the design of any rail infrastructure pricing procedure. The theoretical and political debate focuses on two options. Many public firms still advocate the use of the efficient price mechanisms and propose marginal cost rules with the simultaneous use of public subsidies to cover fixed costs. Alternatively, a growing literature patronizes the use of full cost recovery prices, including price discrimination, multiple part tariffs or cross-subsidization schemes, if needed. Although it is thought that it might yield inefficient outcomes for the theoretical efficiency principles, it constitutes the second best available alternative in most cases.

Similarly, with respect to access pricing of a rail network, it is clear that it should be based on marginal cost pricing rules in a first best world. In practice, however, the achievement of this objective is difficult due to at least three reasons: the above described cost structure of the rail network, which cannot always be recovered with simple price rules; the asymmetric information problem faced by the regulator with respect to these costs; and the subsidy level that can be sustained in the long-run.

Several econometric studies have shown that in the case of the rail industry, the marginal cost of those railways that are still vertically integrated lies in the range of 60%-70% of average cost; where rail services are separated from infrastructure, the marginal social cost of rail infrastructure alone often is well below the 60%-70% range (see, for example, Friedlander *et al.*, 1993). Price discrimination, if feasible and politically acceptable, may help to raise cost recovery to around 60% of total cost without driving demand off the market. Thus, full cost recovery would require a further price mark-up of more than 60% above the efficient price. Alternative proposals, in terms of the so-called Ramsey pricing principle, have been defended for infrastructures

with high fixed costs and low marginal costs.⁴ However, they rarely work in practice, since they arouse consumers' suspicions of unfair treatment and undue discrimination. Moreover, under Ramsey pricing rules all unattributable fixed and common costs are apportioned on the basis of the services' demand characteristics.

In the current debate, a reasonable conclusion is to advocate a balance between the cost recovery issue and the efficient pricing rules, giving preferential treatment to one or the other according to the case. However, the issue remains unsolved and depends on how different countries have faced their access pricing problem. Whether a country's government is willing to assume these differences or not is, in most cases, a political question. In many cases, the ultimate challenge is how to price access to rail infrastructure in a transparent, efficient and non-discriminatory way. In Europe, for example, Directive 95/19 requires infrastructure managers to balance revenues with expenditures. In countries where revenues from operations and compensation from government for public service obligations are insufficient to provide a surplus for depreciation and investment, railways will be dependent on the state to fund or guarantee repayment of investment loans. This continues to be the case in many of the countries of Central and Eastern Europe.

3.2. The access pricing question

The development of tariffs for accessing rail infrastructure varies greatly among different countries according to the stage of their railway restructuring process. Some countries have already identified procedures for setting fees, and a number of them have laid down precise rules for the structure and level of fees. In others, business unit or infrastructure companies (either in public or private hands) are responsible for setting charges. In fact, access charges are mostly relevant in countries where traditional railroads have been vertically unbundled by the separation of the potentially competitive area of service operations from the naturally monopolistic area of infrastructure management.

Apart from the already discussed problem of cost recovery, access pricing may create a market structure problem regarding its effects on competition and barriers to entry. This problem arises in network industries where a single, vertically integrated dominant firm (either private or public) controls the supply of a key input (in this case, railway tracks) to its competitors. It is obvious that in these cases, there are incentives for the firm to set prices high to raise rivals' costs, but it could also be the case in which the regulator sets access prices too low in order to favour the entrants.

Depending on the discretion allowed to the integrated firm, potentially distortive effects on access prices can be determined in several ways. First, when infrastructure is still publicly owned or managed, the regulator can determine the price as an integral part of the access terms defined in a contract with one of several private train operators. Secondly, the regulator may allow the firm to choose from a menu of alternative regulatory schemes, usually rooted in incentive-based price regulation mechanisms (to favour the firm that achieves higher levels of efficiency). Thirdly, the firm may have discretion over aspects of access pricing subject to some overall regulatory constraint,

⁴ Ramsey pricing refers to charging higher prices above unit costs to more inelastic market segments. When infrastructure and services are separated, their use becomes more complicated and still is not clearly solved, since different demands for services – as well as for tracks – must be estimated.

and finally, the firm may have full discretion over the price and is only restricted by the country's anti-trust law.

In all of these cases, there are two main approaches to setting access prices when the principles of cost recovery plus the normal rate of return are required. First, some countries use the current dominant paradigm for setting access charges: *cost-related charges*, which are based on the optimal first-best principle of pricing according to marginal cost (considered the forward-looking long-run incremental cost). The higher the proportion of common costs, the more complex the principle. It is based on the so-called *efficient-component rule*, which determines that optimal access charge is equal to the direct cost plus the opportunity cost of providing access (given by the reduction in the dominant firm's profit). To compute these costs, the regulator has to consider economic depreciation (physical depreciation plus technological progress) and forecast future usage.

The first problem to be solved is that of the actual value of capital assets: nominal value versus potential to generate cash. While the latter is clearly a function of the privatization and regulation methods and the extent of competition envisaged in bidding for the right to operate concessioned infrastructure services, the former is more likely to reflect a past situation that domestic reforms are trying to overcome.

The second method of setting access prices consists of developing *usage-related charges*. Once-avoidable costs are covered by increasing prices that are inversely related to demand elasticity. Another option (less controversial) is the use of a two-part tariff to avoid service cuts by train operators to save charges even when the network has no cost saving. The British infrastructure provider until 2001, Railtrack, is a well-studied example of how access prices functioned in practice. In a industry context where operating companies were franchised, Railtrack managed the infrastructure (track, signalling systems, electric power supply and stations) and was responsible for its maintenance, new investments and train operations (timetables, coordination, etc.). It also sold access to infrastructure to passenger and freight operators.

Railtrack owned the rail network and set track charges that had to be agreed upon with the rail regulator under the criteria openly published in a number of regulatory policy statements. The price control system operated through a simple RPI-X formula that was revised every five years, remaining fixed between revisions. The structure of Railtrack's access charges for passenger services was based on the usage-related charges and was made up of multiple-part tariffs with at least four elements. First, track usage charges, which tend to reflect short run effects on maintenance and the renewal costs of running trains of different types for different distances. Second, traction current charges, to recover the costs of electric current, varying geographically and temporally and reflecting distance covered and type of vehicle. Third, the long run incremental cost, which indicated the long run costs imposed on Railtrack in delivering the total access rights of a train operator. Finally, common costs, as the remainder of the fixed charge, designed to recover the rest of Railtrack's costs at the sub-zonal, zonal or national level. This was apportioned among train operators on the basis of budgeted passenger vehicle miles for sub-zonal costs and budgeted passenger revenue for zonal and national costs. The first two elements amounted on average to only about 9% of total track access charges, and given the current structure of charges, these were the only elements that vary directly. The remaining 91% of the aggregate charge was in the form of a fixed charge, which did not vary with the number or type of trains run or with passenger revenue.

In the case of freight services, access prices were more flexible. The rail regulator had simply established several principles to be considered by Railtrack in its relationship with private operators. First, prices must cover the avoidable costs incurred by Railtrack as a direct result of carrying that particular freight flow; second, prices must be lower than the stand-alone cost that would be incurred by a national efficient competitor; third, no undue discriminatory charges are possible; and finally, charge structure should reflect the value to users of access to the rail network and enable Railtrack recover its total cost

As opposed to the British case, the setting of access charges in other European countries is still underdeveloped. In 1995, the European Union passed two directives concerning the application of Directive 91/440 on the separation of infrastructure management and transport operations. Directive 95/18 regulated the licensing of railway undertakings, and Directive 95/19 established several general principles on the allocation of railway infrastructure capacity and infrastructure fee charges. These principles were designed to ensure an optimum, non-discriminatory use of infrastructure and guarantee an access charging policy according to EC rules, but they were received by member states with various degrees of enthusiasm. The objective of most governments that have set rules for infrastructure fees is to cover costs and differentiate fees to reflect different cost factors. In 2001 a new rail package clarified the principles on which rail infrastructure management should be based on and, very recently, through EC Directive 2004/51, the deadlines for implementing 'third party access' have been shortened (to January 2007). However, many of these changes have been slowly implemented in most countries.

In France, for example, several principles were introduced to give access to railway infrastructure to licensed international groupings of transport services and operators of combined transport, but present arrangements seem more inclined to promote conventional international rail groupings rather than new entrants into the rail market. With centrally planned timetables, only the domestic operator pays a fixed amount to the (also public) infrastructure manager. User fees are fixed, accounting for a wide set of criteria including: infrastructure costs, the transport market situation, supply and demand characteristics, imperatives based on optimized use, and standard conditions for intermodal competition. In 2004 the access fee system was changed in several important ways. First, the access fees per unit of traffic were set two years in advance instead of essentially being negotiated after the fact. Second, the structure of the fees was changed to sharply increase fees for local passenger trains, freight, and ancillary services (such as stops in stations or the use of marshalling yards). Third, the projected total volume of fees was forecast to increase more gradually at a rate of about 300 million euros per year.

Similarly, in Germany, the federal government owns the track infrastructure and is responsible for its preservation and for securing a certain level of public transport service by means of the *Deutsche Bahn* (DB), an independent joint-stock holding whose sole shareholder is the state. The infrastructure division of DB bears operating and maintenance costs and is in charge of stations, ticket sales, passenger attention, etc. It is also responsible for setting charges for track usage, which are supposed to cover all infrastructure costs, including investment. These charges are based on prices per train/km on the different line sectors, resulting in a number of different fee combinations (Häfner, 1996).

In Spain, the 2003 Railroad Law introduced charges for the usage of rail infrastructure, stations and other track elements that conform to EC Directive 2001/14. These charges intend to recover infrastructure's full costs, and include four components: access, capacity reservation, circulation and traffic. The access charge is a general payment to be made by all licensed operators for the right to use the infrastructure. The capacity reservation and circulation fees depend on the kilometres of track used and vary with the type of service or train, the hour of the day and the characteristics of the track. Finally, the traffic charge is levied on the operators depending on the economic value of their service as measured by the number of seat-kilometres or ton-kilometres operated.

4. Quality and safety regulation in the rail industry

Quality performance is not neutral for the economic contribution of the rail transport sector to the social welfare. The particular level of quality achieved by train operators and particular features in regard to three main dimensions that broadly define quality in the rail industry (service, externalities and investment) critically determine the value added by this transport mode. The first question that naturally arises is why quality regulation is needed at all in this industry, and to what extent this regulation relates to the standard price regulation mechanisms described in the previous section. Economic theory provides a well-known argument to answer these questions: real world transport activities are characterized by market failures due to information problems.

In an ideal world with a large number of competitive rail transport service providers and well-informed consumers of passenger and freight services, quality regulation would not be required since market forces would adjust consumer demand (in terms of prices, levels of output and of quality of service) to firm supply. If no price correction took place, less reliable rail companies would be driven out the market and only those whose price-quality ratios were in accordance with demand would remain. However, when full information doesn't exist, markets cannot exert this disciplinary role on firms and purely competitive solutions do not always positively affect quality, prices, or output. Pure competition may result in unsafe, unreliable or unpleasant services since limited availability of resources and lack of adequate control mechanisms make it impossible to adjust consumer and producer interests.

In the traditional organization of the rail industry some years ago, a monopolistic structure with a single firm providing services at the national or local level, price-quality adjustment problems may have increased since the monopoly's privately optimal level of quality may not have coincided with social standards. Simple price regulation is seldom a solution. Any regulated, multi-product monopolist in an environment of asymmetric information tends to degrade quality in order to achieve higher profits once it enters the market. Railway firms are not immune to this temptation, for example, in terms of punctuality and cancellation standards. The quality outcome of any monopolist, not just in the rail sector, heavily depends on the specific regulation adopted. For example, with rate of return regulation, over-investing in non-required technological quality may accentuate the Averch-Johnson effect. Alternatively, with price cap regulation, a subtle cut in quality can be a very tempting way to cut costs (Carbajo, Estache and Kennedy, 1997).

Therefore, the price regulation mechanisms analyzed above are considered incomplete if they do not include quality provisions. This is not always easy, since adjusting price mechanisms by quality may render them inoperative or excessively difficult for the firm to manage or the regulator to monitor. Therefore, most regulators set quality standards or targets for train operators instead of correcting price control mechanisms.

4.1. Definition of quality targets

In setting up those quality standards incorporated in concession contract designs, the regulator often uses the principles of yardstick competition.⁵ These quality standards may be constructed at the national or regional level with inter-industry comparisons (as in Brazil and Chile for many of their public utilities) or by establishing international benchmarks or best practices (as in Australia for transport services and infrastructures).

Three elements are considered in detail when designing this process. First, as in other transport modes, quality is mainly measured in service levels or specified service standards. However, this measurement is suited more for factors such as train punctuality, the reliability of services and the waiting time at stations or platforms, than it is for other factors.⁶ Simultaneously, the services provided before the transport itself, such as ticketing, reservations, and luggage or cargo handling are often ignored as part of the rail industry's value chain, although they may constitute relevant aspects of both intramodal and intermodal competition. For these reasons, the first element to consider in designing a quality control in the rail industry is an integrated vision of transport service that includes not only the ride itself, but all aspects related to infrastructure (track and stations), stations and pre- and post-transport services provided to clients.

A second aspect of quality regulation that is particularly relevant to railways is the flexibility with which scheduled services can be changed and new services introduced in response to changes in demand. Here, the rail industry has always been at a disadvantage to road transport because of the need to coordinate working timetables and operations with certain technical requirements due to the lack of alternative routes between points.

Hence, it is not usually easy (with a few increasing exceptions in many countries) for rail transport to offer on-demand services to passengers (for example, as done by charter airlines) or freight customers (door-to-door services). Thus, coordination is relevant for quality of service regulation within the rail firms, and must also be considered in the design of the industry structure. For example, one potential disadvantage of the split between infrastructure and operations is that coordination might be even more difficult when changes have to be negotiated between different organizations, especially where timetable approvals also need to be secured from other train operators using conflicting train paths.

Intermodal coordination with other industries is also necessary, since social quality performance is always evaluated in relation to feasible alternatives. Saturated corridors

⁵ This is done to avoid the problem of regulator's capture and the discretionary nature of the regulatory action. However, there is a risk of making undue comparisons between different rail systems.

⁶ For example, railway tracks can deteriorate with respect to the smoothness of the ride or the noise or vibration generated to passengers and third parties (buildings close to tracks) even though punctuality and/or safety are not jeopardized, so there may be an incentive to reduce maintenance standards in this respect.

(where investment in roads, railways and airports clearly overcomes demand) are a waste of resources that few economies can assume. This almost general equilibrium approach constitutes the third element of the quality regulation process, although it is not particular to this industry. The socio-political implications of quality regulation (in terms of equity or public service obligations and the social acceptance of quality standards) determine the overall targets to be established in each industry.

Table 1: Quality dimensions of the rail industry.

<i>Dimension</i>		<i>Definition</i>	<i>Measurement Variables</i>
Quality of Service	Vehicle	Onboard quality (wagons, locomotives)	<ul style="list-style-type: none"> - Age of vehicle/number of years in service - Vehicle size and load factor - Availability of seats - Accessibility - Travel comfort <ul style="list-style-type: none"> - noise - vibration - temperature - tidiness
	Route	Route quality (travel of passengers and cargo)	<ul style="list-style-type: none"> - Distribution and number of stations - Timetable: <ul style="list-style-type: none"> - peak trains - first-last train - weekend-commuter services - Frequency (number of trains per hour) - Punctuality/reliability (waiting at stations) - Cargo services (reliability)
	Service	Pre-transport and post-transport service quality (added value to service)	<ul style="list-style-type: none"> - Ticket sales/reservations - Handling - Staff adequacy and competence - Inquiries and general information - Response to complaints
External Quality		Externalities (safety and environment)	<ul style="list-style-type: none"> - Public service obligations - Safety procedures - Liability regimes - Environment protection (noise, pollution) - Congestion

Taking into account these three characteristics, Table 1 summarizes the five most important quality dimensions for the railway industry (vehicle, route, service, social and dynamic quality) along with a number of standard performance measurement instruments for them. The first three (vehicle, route and service) are related to what is usually named (internal) *quality of service*, whereas the last one refers to externalities.

4.1.1. *Quality of service*

Regulation of the quality of rail transport services in regarding vehicle quality, the transport service itself (aboard trains) and the pre- and post-transport services has been dealt with in different depths in different countries although there is a positive correlation between the extent of the restructuring activity in the rail industry (in terms of private participation and/or separation of infrastructure from services) and the quality regulation requirement imposed on the industry post-restructuring.

In general, countries where the sector is still heavily dependent on government or public agencies (such as in Eastern Europe and Asia) have done less to establish separate quality control frameworks than in those where private participation has been significant (for example, the United Kingdom) and detailed quality control systems have been set up. In all cases, the basic principle governing the design of quality mechanisms is that customer service should be paramount if railways are to maximize profitability and compete with alternative modes of transport. The economic relationship between separate units in a railway enterprise should be structured to ensure the preservation of incentives to maximize customer service (see Swift, 1997a, 1997b).

This is particularly relevant to the separation of infrastructure and operations. Vertical unbundling in railways distances infrastructure management from the end-user customer and could yield undesirable side effects or contradictions. For example, the density of traffic (trains per day) that maximizes returns on infrastructure investment is likely to be greater than the optimal level from the operators' point of view. This is because at high densities, passenger service is likely to suffer due to congestion. Therefore, no matter whether the separation is institutional or only financial, mechanisms to compensate infrastructure units that run below optimal capacity must be incorporated into contracts in order to maximize end-user customer performance as a whole. Since the particular characteristics of the rail industry in each country require fine tuning of any regulatory or contract enforcement mechanism, Table 2 proposes a simple scheme that identifies and separates the roles to be assigned to the regulator and the operator (either franchisees or public or private monopolies) with regard to quality of service regulation.

Table 2: Role assignment in railways quality of service regulation.

<i>Role</i>	<i>Regulator</i>	<i>Operator</i>	<i>Both</i>
Design of adequate quality of service standards	✓	×	×
Level of application of these standards	✓	×	×
Punishments, fines, sanctions, etc.	✓	×	×
Information to passengers about quality standards	✓	✓	✓
Variables to be controlled	✓	×	×
Inspection and reporting procedures	✓	✓	✓
Responsibility for achieving quality standards	×	✓	×
Risk sharing of service quality fluctuations	×	✓	✓
Technical quality	✓	×	✓

After its reform and the full privatization of services and track provision, the United Kingdom's rail system constitutes one of the most practical examples of a detailed quality of service regulatory framework (see Table 2). For example, in the case of passenger transport, the regulatory agency (Office for Passenger Rail Franchising, OPRAF) defined what level of service is tendered for particular routes and corridors and sets the minimum level of service for every route in the country (not only timetable specifications, but also journey time, first and last departure times, etc.) If franchises operated a poorer service than specified then the OPRAF had the right to withhold the license.

Operators awarded with licenses, Train Operating Companies (TOCs), are obliged to include in their timetable certain passenger service requirements set out in the franchise

agreement. These are the minimum standards of quality that operators need to achieve to ensure the basic provision of services. However, in order to avoid excessively limiting the freedom of the operators, these requirements do not specify detailed timetables for each route, but instead set parameters within which each company must design its own timetable. Passenger service requirements are set out by route and are largely based on the former British Rail timetable, specifying frequency of trains, stations to be served, maximum journey times, first and last trains, weekend services, through services, and load factors/peak train capacity (for commuter services). Passenger service requirements also include limits on the number of train cancellations and, where applicable, the level of capacity that needs to be provided. These limits apply in any 28-day reporting period, with three levels determined: (i) a call-in level, where OPRAF reviews the performance of the operator; (ii) a second level, where the operator is in breach of the franchise agreement, and (iii) a third level, which can trigger default of the agreement.

For example, load factor requirement compliance is measured by the ratio of passengers exceeding capacity to the total number of passengers (PIXC). The maximum acceptable PIXC level is 3% for morning and evening peak together, or 4.5% for either peak considered alone. If extra capacity is needed to meet load factor specifications, the cost is shared by the operator and OPRAF according to the following criteria: (i) up to a certain capacity limit, the franchise payment does not change; (ii) between the initial limit and a second limit, OPRAF bears a share of costs, and (iii) above the second limit, all costs are paid by OPRAF.

In practice, not all of the quality dimensions defined in Table 1 can be incorporated in the same proportion to any service quality mechanism. The British system mainly focuses on the route dimension and is based on their extensive experience with deregulation. When the role assignment proposed in Table 2 is not considered, or its components cannot be easily separated, several quality regulation failures may arise. The most important is the failure to define adequate independent quality measures. This is the case of several rail concessionaires in Argentina, where the level of vertical integration between the train service providers and the maintenance firms (in the form of subsidiaries or units integrating a larger industrial group) has distorted the incentive to provide the optimal price-quality ratio in favour of more frequent repairs and technical updates.

4.1.2. Safety and externalities

Regulation of the quality of service is only one of the two static aspects of quality regulation to be considered in designing a global framework for quality regulation in the rail industry. The *social or external dimension* of quality regulation, including all issues related to safety and externalities (pollution, congestion, etc.) must also be considered, and it specifically differs from level of service quality regulation in at least four aspects.

The first element is the scope of regulation. Since non-compliance with social quality standards may affect users and non-users of transport services, these standards should always be exogenously set, by national or supranational legislation with intermodal implications, in the case of the rail industry. This is not always the case for timetables, load factors or vehicle size, variables that usually have simple intra-firm consequences. In the European railway industry, for example, three levels of quality regulation can be

found. Directive 91/440 determined the overall principles, and the obligation to comply was envisaged in mode-specific regulation (e.g. Railways Act in the United Kingdom) or in legislation that applies to all sectors of the economy (e.g. Health and Safety Act).

The second factor that makes service quality regulation different from social quality regulation in the rail industry is that a regulatory approach must be used in the latter. Since the risks associated with accidents or potential environmental damages not only directly affect the private benefit, but also the social benefit of this transport mode, there is a need for an external regulator or agency to coordinate safety and reliability. This coordination is particularly important when firms move from a public to a deregulated system. Furthermore, in the rail industry, separation of infrastructure from services and the introduction of open access have made it necessary for a rail track controller to ensure safe coordination between different operators who are using the same tracks or stations.

Again using the British railway system as an example, their safety regulator is the Health Safety Executive (HSE), which informs and advises the Office of the Rail Regulator (ORR). Operators of railway services, stations and networks must have an accepted *safety case* before the ORR approves their license. A safety case is a complete resource, control and management plan for delivering safety and defining safety procedures, organizations and systems. The private infrastructure provider, Railtrack, is required to have its own safety case, a fundamental component of which is Railtrack's Safety Management System, which is a system of operational and technical standards to ensure safety and safe interworking in Railtrack's infrastructure.

The third aspect of particular interest to safety regulation in the railway industry is the assessment and assignment of risk. Given the inherent difficulties associated with strict monitoring, incentives exist for quality-regulated private providers of rail transport services to place compliance with safety requirements below the attainment of financial objectives.

In fact, despite recent tragedies, railways traditionally have a good reputation for safety, a perception that converges with statistical proof in most countries. Therefore, one could conclude that safety levels and management are quite sufficient and no particular safety precautions or measures should be taken. However, public outcry, negative social effects and adverse public opinion from a single catastrophe, together with the persistence of regular fatalities (staff accidents, passengers joining and alighting trains, etc.) make it impossible for the regulator to avoid designing measures and policies to diminish individual and social risk.

One of these policies relates to the compulsory insurance against third-party liability, since it may correct the operators' incentives to take excessive risk. In Europe, for example, Directive 95/18 required that operators of train services must obtain, together with the operating license and path allocations, a safety certificate and insurance. The insurance arrangements in the privatized British railway industry provide another example of scope of liability cover: the basis and conditions for self-insurance. In this case, licenses for the private operators of railway assets (passenger trains, freight trains, stations, and maintenance depots) contain a condition requiring the operator to maintain insurance against third-party liability for licensed activities. The type, cover, level and identity of the insurer need the approval of the regulator, who sets guidelines on minimum insurance requirements that operators must meet. The operation of licensed activities without insurance approved by the regulator is considered a breach of the license.

Finally, the fourth element where service quality regulation differs from social quality regulation is externality issues and, in particular, those connected with the environment (engine pollution, noise, transport of hazardous goods, etc.) Again, in this case, social quality regulation should be concerned with rail operators' internal and external factors, and should have several differences and similarities to other transport modes.

For example, air pollution is one of the most regulated areas in the road and air transport modes, but is not a critical issue in the rail industry though, there are some notable exceptions in certain countries and routes. Noise pollution in suburban neighbourhoods, areas close to stations and depots and delicate countryside ecosystems has attracted more attention from both the public and regulators. Most countries, therefore, incorporate into their regulation the design and specification of measures to reduce noise produced by rolling stock and stationary sources (fans, compressors, and generators) and shunting noise.

The final issues related to environmental regulation are measuring, analyzing and predicting the emissions of chemical substances (heavy metals, lubricants, dust, etc.) where railway lines are present and assessing the risk to the safety of local residents as a result of rail-related activities (transport of dangerous goods, explosions, etc.) In these cases, most countries subordinate their social quality standards and the role of their regulators to the overall technical principles emanating from supranational organisms or professional associations. Private and public rail transport operators are obliged to comply with national and supranational environmental standards. In Europe, for example, there are EC Directives on air pollution from vehicles that specify environmental standards for vehicle engines and fuel qualities which apply to both vehicles (wagons, locomotives) and transport operations.

4.2. Instruments for quality control

Once the objectives for service and social quality are well established, the next step in devising a quality regulation system for railways is designing control instruments. In principle, there are three alternative mechanisms for regulating quality in the rail industry.

First, the firm can simply be required to publish and report measures of quality every pre-defined period. This information can also be made public to inform consumers and/or actual or potential rivals about the operator's current performance. As in any other type of regulatory process, access to public information is a very delicate issue since it can serve as a disciplinary device for the rail provider and as a strategic instrument to undermine or strengthen the ability of the firm to survive in the market.

A second quality control mechanism is including a direct, explicit measure of quality in the price control mechanism. For example, when subject to rate of return regulation, a rail service provider may be obliged to calculate its asset base according to certain average values and/or obtain authorization to carry out certain technological improvements in order to avoid overinvestment and make use of the Averch-Johnson effect. Similarly, under price cap restrictions, the basket of products whose average price increase is controlled by the regulator can be defined to avoid changes in quality (and consequently, cost reductions) that could be used by the regulated firm to increase profit, even if the same price caps are maintained.

The third mechanism that can be used to control quality is a customer compensation scheme, where grants or payments are awarded to people affected by non-compliance with quality standards. In practice, these mechanisms only work if quality failures can be easily verified. This requires a detailed regulation not only of quality standards, but also of monitoring rules and guarantees for both the regulator and the regulated that the inspection process will be transparent and objective. Moreover, if the compensation is distributed to consumers, either directly by the firm or through an intermediary body, sharing rules must be also defined. The practical difficulties associated with this quality control mechanism have made it common in many countries to instead specify minimum quality standards for certain parameters of the rail industry, backed by explicit legal sanctions that may include fines or the revocation or withdrawal of the operating license.

Table 3: Instruments for quality control in the privatized rail industry.

<i>Regulation stage</i>	<i>Instrument</i>	<i>Additional characteristics</i>
Stage I: Before entering The market	- Pre-tender qualification Requirements	- Experience - Financial strength - Technical ability
	- Specification of service characteristics in licenses	- Routes and frequencies - Timetables - Vehicle capacities and load factor - Punctuality and reliability
	- Specification of financing rules and investment plans	- Investment plans - Fleet and track renewal rates
Stage II: During market Operation	- Quality of price-control Mechanisms	- Rate of return regulation vs. Price cap Regulation
	- Information revelation obligation	- Control of access to critical information
	- Audit processes	- Internal and/or external
	- Company reporting	- Frequency - Format
	- Regulator's direct monitoring	- Setup of monitoring mechanisms and rules
	- Technological control	- Tacograph readings, electronic controls.
Stage III: After market Operation	- Incentive payments	- Customer compensation schemes
	- Penalties	- Fines for underperformance
	- Enforcement and binding rules	- Contract withdrawal as a last resource

Finding the adequate mix of these control mechanisms is often the most difficult task in the design of the quality regulation process. The approach followed by most countries is outlined in Table 3, with a summary of the most important instruments. Thus, the quality regulation process consists of three stages. First, before entering the market (Stage I), the goal is to anticipate and minimize future conflicts between the regulator and the concessionaire.⁷ Licenses must specify the expected characteristics of the service in terms of, for example, routes and frequencies of trains or timetables. For passenger services, particularly in the case of urban and suburban trains, vehicle

⁷ To achieve this, pre-tender qualification requirements can be used in order to ensure a minimum level of technical and practical expertise and financial solvency, as described in the previous section.

capacities and punctuality can also be set. Finally, in order to not forget the dynamic dimension of quality described above, Stage I must also specify investment plans and financing rules. Afterwards, during market operation (Stage II), instruments for quality control in the rail industry should mostly be related to the direct monitoring of the firm's performance. Thus, this is the time to introduce quality incentives in price-mechanisms, to establish the firm's obligation to reveal information and the auditing (external or internal) processes to be carried out. In most cases, the use of technical control instruments (such as tacographs or track electronic controls) complements the standard instruments. Finally, after the transport activity has already occurred (Stage III), compensations or punishments can be implemented according to any of the schemes described above. Both penalties and incentives must be graded according to the expected future evolution of the relationship, since severe fines or large subsidies may alter the behaviour of the operator in the market.

5. Performance indicators

Performance indicators are used in the rail industry to monitor the behaviour of one or more regulated firms in order to evaluate the effectiveness of the regulatory measures to which they are subjected.⁸ The main advantage of these indicators or indices is that they provide a periodical assessment and control of the firm's activity and continuously update information, simply, quickly, and at a relatively low administrative cost for the regulator.

The most important disadvantage of performance indicators is that their use is only valid when comparisons (whether between different firms or the same firm over time) are constructed on a similar basis. For inter-firm comparisons, the companies must belong to countries with similar characteristics (e.g. the participation of transport in the economy as a whole, the degree of economic development, or the regulatory framework, etc.). For intra-firm comparisons, indicators must account for external and internal changes produced during each period (e.g. new management, changes in demand, etc.)

Comparisons across companies usually provide interesting, persuasive results that can help the regulator set objectives and design future license contracts. However, extreme care should be used in drawing normative conclusions from these results. What constitutes a benchmark of desirable practice for some objectives may differ among companies. For example, countries with very liberalized frameworks in their rail industry (the United States, for example) could set desirable productivity indicator levels (or quality of service) that clearly differ from the levels in other more regulated frameworks (such as in Europe).

Similarly, simple indicators should be carefully interpreted over time to avoid contradictions and inappropriate measurements. For example, when assessing railway output, the number of trains/km may be relatively high, while passengers/km or tons/km may be relatively low (if the firm specializes in one type of traffic). Given this conflict, overall performance can be ambiguous. The most practical solution is to jointly interpret

⁸ For example, quality indicators can be established in a contract and reviewed regularly to confirm that the terms of the license are being fulfilled.

the indicators and the objectives that they serve. For example, a service quality objective, such as the number of trains per hour, may conflict with both financial objectives, reflected in a high cost recovery rate, and objectives based on the maintenance of low prices.

Thompson and Fraser (1996) point out that monetary and productivity variables should be carefully defined for inter-firm comparisons. Fares, wages, outputs and inputs vary widely among countries for a large number of reasons that are not necessarily related to the firm's operations, but to measurement or statistical errors. For example, average passenger fares are based on the overall mix of passenger classes (each with a different price). Tariffs are often higher per passenger/km for short trips than for long ones, and they must also depend on the existence of government subsidies or artificial compensations. Similarly, common freight tariff mistakes include not accounting for the different mix of commodities, size of shipment or length of haul. The latter also affects passenger traffic and is particularly relevant since some costs (ticketing, billing and station maintenance, for example) are fixed with respect to the length of the trip but vary with size or distance.

These difficulties are increased when measuring productivity, since a simple comparison among partial measurements of output cannot capture the complexity of relationships or the variety of productive structures that take place within a rail operator. For example, a commonly used productivity indicator, the number of passengers/km or tons/km per employee,⁹ depends on such diverse factors (e.g., regulatory environment, structure of the labour market, availability and quality of infrastructure, alternative transport modes, etc.), that it could be seriously misleading if interpreted without care.

To elude these sorts of problems, the construction of performance indicators should avoid excessively simple data management, and use statistical techniques that account for the different relative environments of each company. Oum and Yu (1994), for example, estimated different efficiency levels for a sample of OECD railway companies by introducing internal factors (such as the characteristics of outputs) and external factors (difference in the legal and regulatory framework between companies).

Despite these difficulties, a large number of indicators are commonly used to monitor the performance of firms within the rail industry around the world. The definition of each particular indicator depends on its objectives and its informative value.

Several external factors that vary widely from country to country and firm to firm substantially influence comparisons. Therefore *contextual indicators* assist in comparative analysis and define desirable performance levels. They include social and economic characteristics of the railways as well as other elements associated with the economy as a whole. Directed mainly at the regulator, they control for the exogenous factors in inter-firm and intra-firm comparisons. Table 4 presents several examples from international statistical sources.¹⁰ Simultaneously, there are many indicators (particularly those for prices and quality of service) that are informative to transport users and provide input for the regulator's control tasks. Jointly with the contextual indicators, these *management indicators* provide the necessary instruments to judge the management and behaviour of the company, and can be grouped at three different levels, summarized in Table 5.

⁹ The term *employee* can also refer to terminal staff, administrative staff, train crew or maintenance staff. Similarly, capital can be disaggregated into trains, wagons, terminals, platforms, routes, etc.

¹⁰ In particular, the International Union of Railways (UIC) publishes a yearly summary of the main statistics of its affiliated railways, although not all of them are always available for all railroads.

Table 4: Contextual indicators in the rail industry.

<i>Type</i>	<i>Examples</i>
Overall economic activity	GDP GDP per capita Urbanization degree Industry structure Energy costs Private cost of capital
Transportation sector importance	Participation of transport in GDP Intermodal market share (passengers and freight)
Overall rail sector indicators	<ul style="list-style-type: none"> • <u>Output</u> Passenger trains/km Freight trains/km Passengers/km Tons/km • <u>Revenues</u> Passenger revenue Freight revenue • <u>Network indicators</u> Length of line Length of track Electrified track (%) Route/km/km² • <u>Density and service</u> Train routes/km per capita Trains/km per routes/km Average size of shipment Average length of haul • <u>Organization of the industry</u> Regulatory agencies (number) Separation of infrastructure and services (type) Access and entry system (type)
Regulatory and institutional system	State involvement in economy (in % of GDP) Tax and Judiciary system (corruption index)

Some final practical rules that could be helpful in this process are as follows: *(i)* each indicator should have at least a function or objective, *(ii)* the relationship between each indicator and its objective must be clear and direct, although *(iii)* multiple objectives can be addressed by multiple indicators (jointly interpreted); and finally, in order to assure the utility of the indicators, *(iv)* appropriate data must be provided and *(v)* the management of the indicators' information should be part of the regulatory process.

For the regulator, price indicators can be a control mechanism over the activities of the operators, despite the difficulties mentioned. This control may be established not only in terms of the comparison between companies with similar characteristics, but through monitoring over a period of time. In any event, the regulator must ensure that any variation in price corresponds to a proportionate variation in costs or level of efficiency. The operational and efficiency indices therefore are instruments that help the regulator. Improvements in company productivity and efficiency levels combined with increases in price levels are clear signs of abuse of market power on the part of railway operators.

Table 5: Management indicators in the rail industry.

<i>Type</i>	<i>Examples</i>
Commercial	<ul style="list-style-type: none"> • <u>Prices</u> Average passenger fare (revenues per passenger/km) Average freight price (revenues per ton/km) • <u>Quality of service</u> Average train-speed (in passengers and freight) Delayed arrivals or departures (as % of scheduled) % of lost or damaged freight Average passenger load factor Traffic density (trains per hour) • <u>Pollution and safety</u> Rate of fuel usage (per train/km) Level of noise Level of emission of pollutants Number of accidents or incidents
Operational	<ul style="list-style-type: none"> • <u>Labour productivity</u> Passengers/km per employee Tons/km per employee Passenger trains/km per employee Freight trains/km per employee Total trains/km per employee • <u>Capital productivity</u> Number and kms. travelled by locomotives Locomotive availability (in %) Tons/km per wagon/km Wagons/km per wagon Tons/km per wagon
Financial	<ul style="list-style-type: none"> • <u>Efficiency</u> Costs per employee. Costs per unit of capital Unit cost (per passenger/km, ton/km, train/km) • <u>Profits</u> Revenues/costs Subsidies

Indicators of service quality that were earlier presented should serve the same way as price indices to establish evaluations of different companies, as well as dynamic or time evaluations. These measurements should be analyzed together with price indices because of the possibility of finding different feasible combinations of price and service quality. For example, a high number of trains per hour, i.e. a high traffic density, could only be financed by means of high prices.

The simultaneous implementation of control systems for prices and service quality may limit the firm management and reduce operability. Placing an emphasis on price control or service quality depends on whether it prefers to offer services at the lowest possible price, or offer services with certain standards of quality. All of these indicators allow the regulator to monitor the operators' activities as defined in Phase II of Table 3. Unjustified or systematic breaches of quality standards (insufficient number of trains per hour, lack of punctuality, unreliability, very high indices of load factor, etc) should be accompanied by an appropriate system of penalties, as described above.

6. Conclusions

This paper has reviewed the theoretical foundations to regulate prices and quality service levels in the rail sector. Also the paper describes how the current changes in this industry have provoked the necessity to modify the old mechanisms to control prices and quality decisions. The institutional separation between infrastructure and operations, the horizontal unbundling process and the increasing contribution of private participation have promoted the introduction of novel and new regulatory systems. But the definition and the application of these systems present problems and difficulties which must be appropriately evaluated. The elaboration of a suitable list of measures or indexes which allow to monitor the performance of the industry is crucial in order to reduce these problems.

In conclusion, there is no unique form of rail regulation to address these new challenges, but the general rule is to maintain flexibility and simplicity whenever possible. Two key issues in the new regulatory environment of the rail industry are that private participation is included in license contracts and the organization of the industry is adapted to each country's needs and characteristics. In turn, using these mechanisms also changes the role of the rail regulator, whose actions should now be governed by principles that foster competition and market mechanisms and simultaneously provide a stable legal and institutional framework for economic activity. The regulator should refrain from intervention unless the ultimate goal of achieving economic efficiency subject to the socially demanded level of equity is in jeopardy.

References

- Armstrong, M., Cowan, S. and Vickers, J. (1994) *Regulatory Reform – Economic Analysis of British Experience*, MIT Press: Cambridge, Mass.
- Braeutigam, R. (1989) "Optimal Policies for Natural Monopolies", *Handbook of Industrial Organisation*, Vol. II.: North Holland.
- Carbajo, J. C., Estache, A. and Kennedy, D. (1997) "Regulating the Quality of Privatized Transport Services", Mimeo, The World Bank: Washington, D.C.
- Friedlander, A., Berndt, E. R., Chiang, J.S., Showalter, M. and Vellturro, C. A. (1993) "Rail Costs and Capital Adjustments in a Quasi-Regulated Environment". *Journal of Transport Economics and Policy*, 32, pp. 131-152.
- Häfner, P. (1996) "The Effects of Railroad Reform in Germany", *Japan Railways and Transport Review*, 15, pp. 27-30.
- Kessides, I. N. and Willig, R. D. (1995) *Restructuring Regulation of the Rail Industry for the Public Interest*, Policy Research Working Paper 1506. The World Bank: Washington, D.C.
- Liston, C. (1997) "Price-Cap versus Rate-of-Return Regulation", *Journal of Regulatory Economics*, 5, pp. 25-48.
- Oum, T. H. and Yu, C. (1994) "A Comparative Study of OECD Countries' Railways", *Journal of Transport Economics and Policy*, 38, pp. 121-138.
- Swift, J. (1997a) *Regulating the Railway in the Public Interest*. Office of the Rail Regulator. Railway Study Association, Meeting at the London School of Economics, 11 June.
- Swift, J. (1997b) *Regulation of the Railways in Great Britain: 1993-1997 to 2000+*, Office of the Rail Regulator. Address to the European Union Delegation of the French National Assembly and the Senate, 13 March: Paris.
- Talley, W. K. (1988) *Transport Carrier Costing*, Gordon and Breach Science Publishers.
- Thompson, L. S. and Fraser, J. M. (1996) "Financial success hinges on productivity", *Database, Rail Business Report 1995*, The World Bank: Washington, D.C.

Acknowledgements

Pedro Cantos thanks financial support from Spanish Ministry of Science and Technology under project SEJ 2004-00110. Javier Campos gratefully acknowledges financial support from the Spanish Ministry of Science and Education and from FEDER through grant SEJ2004-00143/ECON. Both authors are grateful to The World Bank for collaborating in previous versions of this paper.



Trans Siberian Railway: from inception to transition

Anastasia Liliopoulou¹, Michael Roe², Irma Pasukeviciute^{2*}

¹ *Export manager, Combi Trans Hellas Ltd*

² *Centre for International Shipping and Logistics, University of Plymouth Business School*

Abstract

This article presents a detailed historical overview of the existence of the largest railway in the world, which runs for 5,867 miles and connects Far East with Western Europe. Over the years it gained many names, such as Trans Siberian Land Bridge, Trans Siberian Route, Trans Siberian Line and Trans Siberian Railway but each one of the names stand for the longest rail route across the continent of Eurasia. This article provides an opportunity to look at early stages of railway's construction, its uniqueness, interesting path of development, survival of two World Wars and finally its establishment as a vital part of continent's logistics chain. Throughout the years, the Trans Siberian Railway (TSR) has been proven to have the longest history of commercial freight operation between Europe and the Far East.

Keywords: Trans Siberian Railway; Logistics.

Construction of the Trans Siberian Railway

The initial idea of a railway construction, which would open up Siberian region for development, was set out by general governor of Eastern Siberia N.N Muravyov-Amursky in 1857. In the following years, this idea inspired military engineer D. Romanof to create a project, which involved building a railway line that would ultimately connect Russia to Siberia. The idea was highly thought of however the cost of such a construction ensured no support from the Russian government. This was mainly due to the lack of funds and the insufficient number of railways which connected Russia to its mining interests (Soviet Geography, 1990). Only in 1873 when the Ural Railway Company was established to link iron and coal rich Ural mines with central Russia, the Russian government started working in earnest for the Trans-Siberian Railway. There were many suggestions from foreign entrepreneurs to fund the construction, but nonetheless the Russian government decided to use its own funds, because capitalists could have strengthened foreign influence on Siberia and the Far East of Russia whilst building the railway there and it was unacceptable at the time (Slepven, 1996).

* Corresponding author: Irma Pasukeviciute (ipasukeviciute@plymouth.ac.uk).

The first real impulse to start construction works on the new Trans Siberian Railway line was given by Tsar Alexander III in 1886 but in reality it did not come into effect until 1891 when the construction actually began from both ends, Vladivostok (East Siberia) and Chelyabinsk (West Siberia), and worked towards the centre (see figure 1).

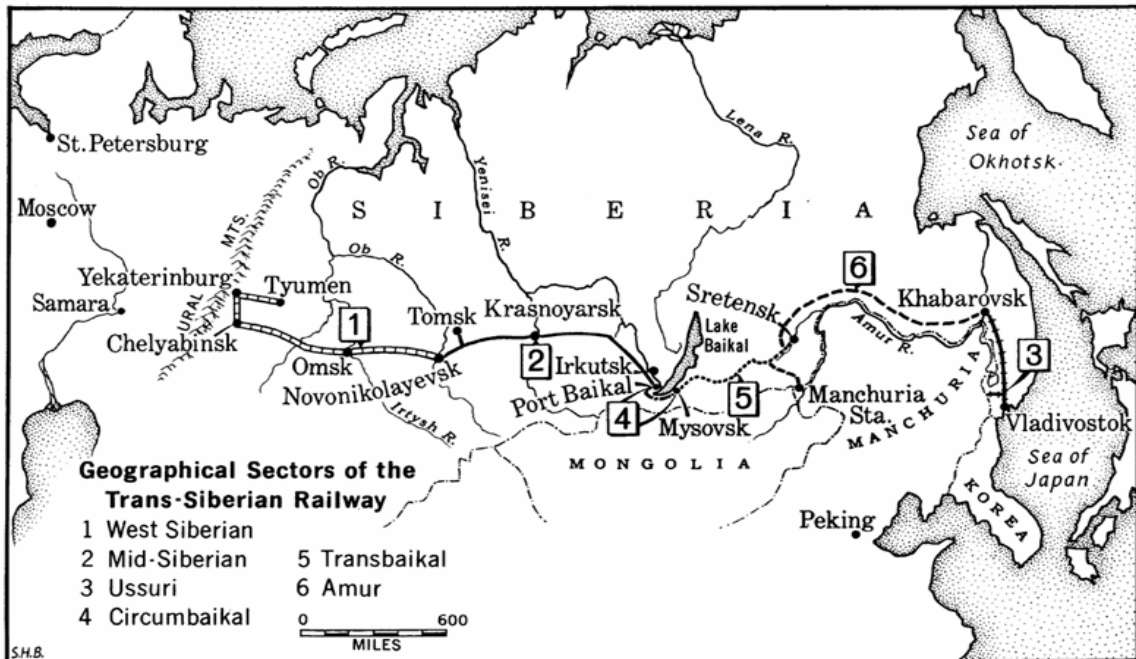


Figure 1: historical map of the Trans Siberian Railway.
Source: <http://www.transsib.ru/Map/transsib-building.gif>.

The project was built in several sections. From 1891 until 1897 a railway line connecting Vladivostok and Khabarovsk was completed because by 1880 Vladivostok had grown into a major port city, and the lack of adequate transportation links between European Russia and its Far Eastern provinces became an obvious problem.

A treaty with China in 1896 enabled the Russians to construct an 800 mile railway line through Manchuria, this way shortening the distance to Vladivostok. Therefore between 1897 and 1903 Russian Government constructed the Chinese Eastern Railway, across Manchuria, in Northern China (Soviet Geography, 1990), which connected Vladivostok with sections of the TSR in Western and Central Siberia (see figure 1). By 1904 the TSR stretched from Vladivostok across China and Siberia to the Ural Mountains (Moore, 1980). Construction continued, despite the fact that building the railroad was a hugely challenging task for the Russian Government due to the difficult terrain and extremes of temperature in Siberia. One of the main obstacles to the completion of the Trans Siberian line was the Baikal Lake, but a way around the lakeshore was completed in 1905. By 1916 the Amur River line, situated north of the Chinese border was finished together with the continuous railway line within Russian territory from Moscow across Siberia.

Tsar Alexander III started the construction of the Trans Siberian Railway with the vision of providing a reliable communication and transportation system, linking the Russian empire with Siberia and making long scale-immigration possible (Jorre, 1961). He knew that Siberia was at the mercy of strong Asiatic powers if there wasn't a reliable

communication and transportation system linking the empire with Siberia. This leads to a conclusion that the TSR was initially built to protect the Siberian borders and hence “the replacement of the normal 4ft gauge by the unusual 5ft gauge was adopted, whose adoption aimed at isolating conservative Russia from progressive Europe as well as at hindering possible invasions” (Jorre, 1961, p189).

America was one of the first countries to express their enthusiasm for the longest railway line in the world and its advantages to be gained from Russia’s opening up of Siberia. The Americans believed in no doubt that “the Siberian Railway would invigorate the extensive and richly endowed territories, and create favourable conditions for American exports” (Slepnev, p37).

In the years to follow, the contracts for the equipment of the Trans Siberian Railway were given to American firms including the supply of rails, locomotives, freight-car bodies, air brakes and engines. The Siberian line not only opened a dependable route to the Pacific Ocean but was also the key to the miraculous natural wealth (oil and coal) of Siberia and the Far East (Karbonski, 1992).

TSR during the world wars

From the earliest days of their introduction railways have been regarded as offering the most efficient means for meeting the special needs of military transport in time of war and it is a fact, that Russia was one of the countries with a long history of wars. Therefore, originally, the Trans Siberian Railway was used during the First World War by the Allied powers to transport troops and supplies across the vast territory. In 1914 the war with Germany worn out the Russian Empire and the Trans Siberian Railway had almost ceased to function, in a critical period where German submarines had effectively precluded shipment of arms to Russia through the Baltic. As a result military supplies were purchased from the US and shipped to Vladivostok awaiting movement via TSR. The reorganization of the Trans Siberian Railway meant “access to massive stockpiles of munitions, food fuel, coal and other war supplies that the Allied shipping had stockpiled in the ports of both Archangel and Vladivostok” (Giffin, 1998) for the Allied Russian armies in Europe.

After the government of Tsar Nicolas fell in 1917, the new pro-western Provisional Government needed a substantial amount of money to purchase more supplies (Gaddis, 1990). Later that year they turned to the American Government for help in order to maintain functionality of the Trans Siberian Railway (Jacqueline, 1969). Three hundred men from American railway companies were selected to form the Russian Railway Service Corps (RRSC) (Johnson, 1923). The American team inspected the railway and reached the conclusion that “Trans Siberian Railway was the only usable railway into Russia from the outside world” (Culloton, 2002). The RRSC established 14 station units distributed along the Chinese Eastern Trans Siberian Railroad to Omsk (Graves, 1931). Their duty was to inspect the Trans Siberian line and advise the Russian government on how to improve the railway and increase its carrying capacity (Johnson, 1923).

Siberia became a very important strategic area during the First World War and thus in May 1918 United States intervened into the Russian war, with the aim to remove the Czech Legion from Siberia, salvage a front against the Germans, prevent Germans from seizing Allied supplies and keep a watch upon Japanese who were also intervening in Siberia (Unterberger-Miller, 1989). The TSR suffered severe damages as a result of the

war – over one hundred bridges, numerous depots, water towers, and other railway facilities were in need of considerable repairs or even rebuilding (Gaddis, 1990).

The poor condition of the railway, tension caused by the Allied Intervention, the civil war and the Czech control over a large portion of the railway, led to the Inter-Allied Railway Agreement (Kennan, 1967). The US Government developed a plan for the creation of a commission that would operate the railway until the Russians were able to resume control. Therefore, in 1919 the Inter-Allied Railway Committee (IARC) was established. It included representatives from governments of the following countries: Russia, United States, Japan, China, Great Britain, France, Italy and Czechoslovakia. Its first task was to divide the railway into sections to be guarded by American, Chinese and Japanese powers.

IARC was the policy-making committee, and therefore a number of specialized agencies were established in order to utilize its decisions. The most important agency was the Technical Board, which was responsible for the technical and economic management of the Trans Siberian line.

Between 1919 and 1922 the Technical Board contributed substantially to the improvement in both railway's physical condition and its efficiency. Under the Board's direction over one hundred bridges were repaired or rebuilt, entrances to major tunnels blocked by explosives were cleared, and depots that had been destroyed were replaced (Giffin, 1998). Furthermore, locomotives and cars were repaired quickly and the freight tonnage was increased through the use of daily reports on train movements and the heavier loading of freight cars.

In 1922 the Allied troops evacuated Siberia and the Inter-Allied Railway Technical Board was dissolved. During its existence the Board achieved a number of goals; they managed to re-organize, revitalize, and co-ordinate operations of the Trans Siberian Railway (Giffin, 1998).

In the years to follow Soviet Government realized that Trans Siberian Railway provided a major logistics and communication line and therefore, since 1936 TSR was used for the movement of freight with the ultimate aim of earning hard currency (Helmer, 1999). Transit via the Trans Siberian Railway was favoured by customers, if compared to the deep sea route, due to the fact that goods of origin in transit through the USSR were not charged export and import duties. In addition to that, TSR route was also about 4,375 miles shorter than the route by sea via the Suez Canal from Far East to Europe (Soviet Shipping Journal, 1982).

During the Second World War TSR was mainly used for the movement of military cargoes delivered from the United States. The end of the war brought two major changes in terms of the historical development of the TSR. Firstly, the Trans-Manchurian line, connecting Vladivostok and Siberia, came under the Chinese control and was renamed as Chand-Chu'nn Railway (Mellor, 1975) and secondly, Eastern and Central European countries were taken under Soviet control, as agreed by the West at the Yalta Conference in 1945 (Karbonski, 1992).

TSR under the Soviet Union rule

This section provides short overview of trade principles in the Soviet Union (SU), its economic development and the role of the Trans Siberian Railway in that context.

Soviet foreign trade was dominated by three principles: (1) the trading partners had to be neighbouring communist countries or the developing third world countries to which the Soviet Union has given aid involving trade credits, (2) sale of diamonds, gold, furs etc., was carried out to acquire foreign exchange and (3) the acquisition of required raw materials and essential consumer imports had to be conducted on the most favourable terms (Jain, 1993).

At that time, the emphasis towards industrialisation, favoured the development of rail networks at the expense of other modes (Cullinane and Toy, 1998). Rail transport flourished in the middle of 1960's mainly due to the growth in the agricultural and industrial sectors and more particularly to its geographical position, serving even the vast areas of Siberia (Mathieson, 1975).

Despite the fact that it contradicted the main trade principles of the Soviet Union, in the years to follow trade with West Europe flourished. The introduction of containerization in the world market benefited both the shipping and railway sector in the Soviet bloc due to the fact that it led to the co-operation and co-ordination of these two sectors and provided through combined transport, the fastest and cheapest route from Far East to Western Europe and vice versa (Slepven, 1996). More specifically, in 1967 the progress of containerization and intermodal transport helped the Trans Siberian Railway to gain value as the shortest and cheapest alternative route compared to the deep-sea route from Europe to the Far East (Soviet Shipping Journal, 1982).

In the late 1960's experimental shipments between Japan and Europe using the Trans Siberian Railway were conducted (Zheleznodorozhnyy Transport, 1975). In 1971 a formal agreement was signed between the Soviet agency, Soyuzvneshtans, and the Japanese and European freight forwarding agents for the operation of the Trans Siberian container route and the creation of a modern system of intermodal container shipping known as the Trans Siberian Container Service (TSCS) (Miller, 1978). The first recipients of licenses to operate on the TSR were Y.S Van Gend and Loos, and C.T.I (JEURO)/ M.A.T. Transport.

According to this agreement the Soviet agency was responsible for most of the facilities of the intermodal service. Furthermore, the agreement provided that the containerized cargo heading to Western Europe would be moved using Japanese and Soviet vessels from Japan to the Russian port of Nakhoda (U.S News and World Report, 1975). Once the containers reached Nakhoda they would be loaded on flatcars and move across TSR to Moscow. From there the containers would be routed to the container terminals on the Baltic Sea coast in Tallinn (Estonia) and Riga (Latvia). There the containers would be transferred to ships heading to the country of final destination.

The co-operation and co-ordination among the countries of the self-sufficient communistic Soviet Union was vital for the TSR's success to follow. The block trains which were used to carry the containers were built at the ports of Leningrad, Tallinn and Vostochny and at the rail station at Brest (Belarus). Moreover, the USSR focused on transforming its Baltic ports from naval bases to water gateways for containers in transit. The port of Leningrad was the first of the Soviet ports to develop as a container port providing all kinds of container facilities including repairs to damaged containers and equipment (Queiroz, 2001).

In addition, Soviet ship owners established four companies to serve the Western European market and maintain a network of TSCS liner services. These were: the Baltic Shipping Company, the Estonian Shipping Company, the Latvian Shipping Company and the Azov Shipping Company.

The Baltic Shipping Company was based in Leningrad operating a fleet of container vessels and Ro-Ro ships. The Company was providing weekly services to/from Tilbury, Hull in the United Kingdom, Antwerp in Belgium, Rotterdam in The Netherlands, Hamburg and Bremen in Western Germany (Lukov, 2000).

Tallinn was the most important transshipment port for the TSCS due to its geographical location, close to the Scandinavian countries. The Estonian Shipping Company was operating frequent services to ports in Stockholm in Sweden, Oslo and Drammen in Norway, and Aarhus and Copenhagen in Denmark (Zurek, 2001).

The Latvia Shipping Company was based in Riga. The Company was providing regular container services to Dublin in Ireland, Ellesmere, Rostok and Le Havre in France.

The Azov Shipping Company was situated in the port of Zhdanov. This Company was covering all the Mediterranean countries. The services were operated from/to Valencia, Barcelona, Ravenna, Savona, Venice, Rijeka, Piraeus, Istanbul, and Alexandria. All four Companies are closely co-operating with the TSCS by providing an extensive sea network for containers in transit from Far East to Western Europe and vice versa.

In the early 1970's, the joint efforts of the Soviet and foreign parties, involved in the improvement of the Trans Siberian Container Service, helped the Trans Siberian Railway to win popularity and confidence with Japanese and Western European transport communities (Helmer, 1999). By 1979 the Trans Siberian Railway had won more than 20% of Japan's westbound export cargo (Lloyd's Shipping Economist, 1979). Moreover, the improvement in political relations between the USA and Soviet Union in 1980 lead shippers to view the Trans Siberian Railway more positively (Lloyd's Shipping Economist, 1980).

In July 1971 the Council for Mutual Economic Assistance (CMEA), which was a multilateral economic alliance responsible for promoting the economic development of the countries that were under the Soviet rule, adopted a plan for complex Socialist economic integration, which led to more joint projects and information exchange between members (Jorre, 1961). The plan included the exchange of goods among communist countries frequently by bilateral negotiations. However, as mentioned before, trade drifted outside the bloc countries, mainly to the third world countries for imports of raw materials against capital goods or towards a varied trade with Western European countries (Mellor, 1975).

The lack of convertible currency in the USSR lead to further expansion of energy trade with Western Europe in order to earn hard currency (Jain, 1993). This was easily achieved because the price system in the Soviet bloc did not depend on the principles of supply and demand, but rather upon a series of state-controlled prices for all commodities (Roe, 2001). The transport sector was continually subsidized and that meant that certain sub sectors, such as shipping and railways could develop as major hard currency earners (Lavigne, 1999).

Even though this opening to the West, against the policy of self-sufficiency, was opposed in the Soviet bloc, it provided the opportunity to develop true commercial skills and experience of the free market (Soviet Shipping, 1989). The new order occurring in the Soviet bloc required new changes in order to be able to trade and benefit from the West. Therefore, the most vital change involved the establishment of a convertible currency (Bernard, 1966). That was mainly due to the fact that Eastern European countries were constrained by the lack of foreign currency, and that meant that these countries had to export first in order to be in a position to import later.

The problem of non-convertible currency was partially resolved in 1976 by introduction of a commercial rate, which was calculated as the average amount of domestic currency needed to earn a unit of foreign currency.

Further more in 1976 the port of Vostochny was developed into the terminal gate of the Trans Siberian Railway with the ultimate aim to accommodate TSCS liner services in the Far East (Bergstrant and Doganis, 1987). The Far Eastern Shipping Company allocated in the port maintained all the TSCS container services in the Far East, including Japan, Hong Kong, Philippines, and Thailand.

In the late 1970's the FIATA Congress observed that *'the organization of shipments by the Trans Siberian Containers Service was a major achievement on the part of European and Japanese forwarding agents'* (Soviet Shipping Journal, 1982). In December 1979 the V/O Soyouztransit association was formed which was the Soviet Foreign Trade Self-Supporting Corporation, the sole forwarder of transit shipments via the USSR territory (Bergstrant and Doganis, 1987). V/O Soyouztransit offered three transit routes by TSCS: TRANSRAIL, TRANSEA and TRANSCONS.

The TRANSRAIL route was providing transit for cargo moved from the Soviet border stations, Luzhaika, Brest, Chop, Ungeny, Djulfa and Kushka to ports in Japan and other countries in South – East Asia, and vice versa. V/O Soyouztransit was responsible for the receipt of containers from the European rail at the Soviet border stations and their movement by rail to destination via the ports of Vostochny and Nakhodka (Soviet Shipping Journal, 1982). The transit time of containers was approximately 25-30 days.

The TRANSEA route was providing arrangements for the transportation of containers from European ports to Soviet ports in the Baltic Sea and Black Sea, shipment by rail to the ports of Vostochny and Nakhodka and further transshipment to a vessel for delivery in its destination. The transit time on this route was between 35 and 40 days.

Finally, the TRANSCONS route was responsible for the carriage of containers by road from the Europe to Vysoko-Litovsk near Brest. Once the containers reached Brest, they were transhipped onto rail heading to the ports of Vostochny and Nakhodka. From there the containers were delivered in their destination by sea. The transit time was approximately 40-45 days.

In order to provide an efficient service on these routes, V/O Soyouztransit was closely co-operating with many Forwarding Agents, shipping carriers in Europe and Japan, the Soviet railway, and trucking association Sovtransavto (Bergstrand and Doganis, 1987).

In the years to follow, the joint efforts of the Soviet and foreign parties involved in the improvement of the Trans Siberian Container Service, Trans Siberian Railway had won popularity and confidence with Japanese and Western European transport communities (Helmer, 1999). Figure 2 indicates the movement of containers via the Trans Siberian Railway from 1980 to 1989.

At this point it should be emphasized that the actual figures for the amount of containers carried by the Trans-Siberian Container Line (TSCL) were difficult to compile accurately due to a number of reasons such as different sources, for their own reasons, providing different statistics of the amount of containers that the Trans Siberian Container Line handled since its establishment (Lloyd's Shipping Economist, 1980). In addition, the figures were, in some instances, confused by the number of different countries that transport their containers via the TSR route, making the comparison of origin/destination areas a difficult task (Lloyd's Maritime Asia, 1990). The main reasons behind these problems were the lack of co-operation between the operators involved due to competition, the lack of co-ordination among the parties involved in this

trade, and lack of organization of the TSCL. Also another reason behind the publication of different figures at the time was the strict regulations included in the old Soviet economic system. More specifically, the Soviet economic system was based on targets. If a company succeeded fulfilling the targets imposed then the company could earn more money and more security. However, if a company failed to meet the targets imposed by the Soviet Rule, then the consequences were very unpleasant. Hence, all companies during Soviet times were publishing results that showed they had met targets, even if such results did not reflect the reality.

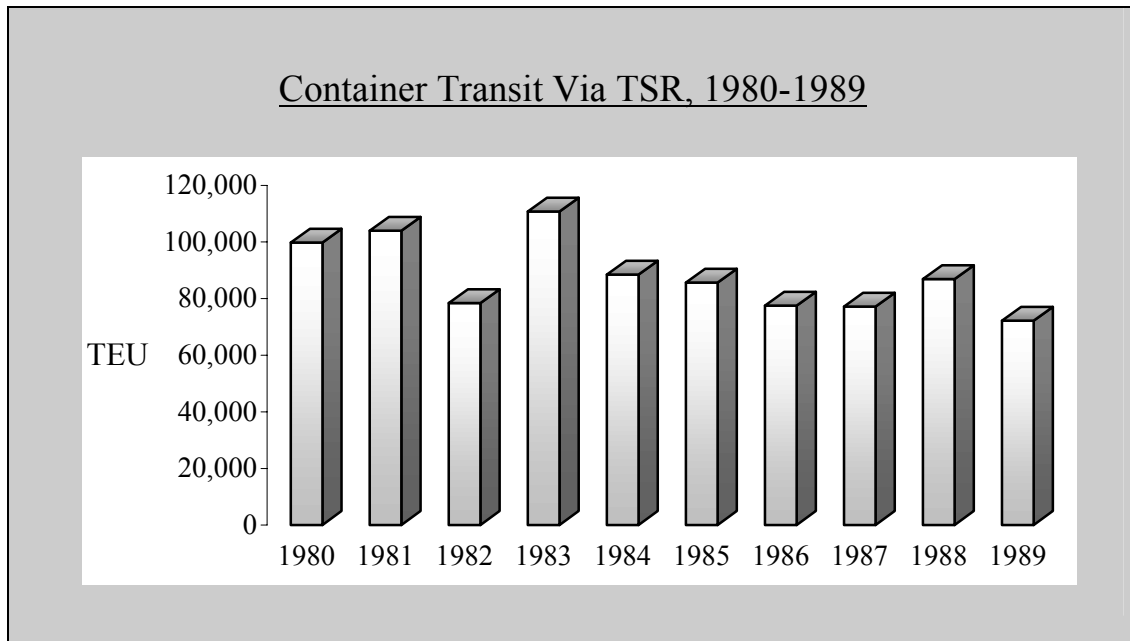


Figure 2: container transit via TSR, 1980-1989.

Source: <http://www.erina.or.jp/Forum/Forum2000/eSession1/eNagasawa.htm> and Nikolai Lukov, Secretariat of the CCTST

However, experts of the region have managed to compile a set of figures that were reliable enough to provide a general picture of the TSR route (presented in Figure 2). This was achieved partly through collating a wide range of sources and partly by comparing the trade between Europe and Japan carried by TSCL and that carried by the Far East Freight Conference (Lloyd's Shipping Economist, 1980). In terms of figures or more specifically in terms of TEU capacity, the numbers of containers transported by the TSCL between Europe and the Far East increased from 55,000 TEU in 1978 to over 100,000 TEU in 1979. However in 1980 the movement of containers from Europe to Japan and Korea via the Trans Siberian Container Line suffered a considerable decline of 10,000 TEU.

Nevertheless, as it is indicated in Figure 2, the total movement of containers via the Trans Siberian Railway had increased in the 1980's, reaching 110,683 TEU in 1983. The main reason for this sharp increase between 1980 and 1983 is mainly due to the war between Iran and Iraq that started in 1980 (Lloyd's Shipping Economist, 1981a). More specifically, the number of containers destined for Iran in 1980, increased by 400% when compared with 1979, reaching 24,000 TEU (Lloyd's Shipping Economist, 1981b). This was mainly due to both the effects of the war with Iraq, which resulted in the

closure of the Iranian ports, and the shortage of shipping space in most trades heading to Europe (Lloyd's Shipping Economist, 1981b).

From 1982 onwards the competition between sea and rail operators on the Europe/Far East route became increasingly intensive due to the announcement of the Trans Siberian Container Line's intention to reduce tariffs, which was expected to increase the container traffic going via railways rather than sea (Lloyd's Shipping Economist, 1982). The main reason behind this policy adopted by the TSCL for the reduction of tariffs had two aims: one was to increase volumes and the second one to maximize the hard currency earnings in the Soviet Union.

During the same period the competition between shipping lines on the Far East to Europe route was increasing, which resulted in the TSCL causing a major concern among shipping lines. Most of them were strongly opposing to the Soviet Union's regulation of competition between the Trans Siberian Railways and maritime routes, as well as to the subsidies provided to the TSR (Lloyd's Shipping Economist, 1981b).

In the first half of 1983 the trade to Iran increased, and so did the volumes of containers from the Far East to Iran via the Trans Siberian Railway. This resulted in shipping lines, which operated in the Arabian Gulf, having to introduce better quality services to Iran (Lloyd's Shipping Economist, 1983a). At the same time, Soyuztransit (SOTRA), the authority which operates the transit system in the TSR, was facing problems due to congestion at the Far East ports of Vostochny and Nakhodka and the inland station of Djulfa on the Iranian border. The congestion led to delays, which consequently raised a number of questions about the future capacity of the land bridge. By the end of 1983, the total volume on the Iran-bound TSR cargo was approximately 44,600 TEU (Lloyd's Shipping Economist, 1983b).

In order to keep an advantage over sea transport, Trans Siberian Railway introduced regular express block train services at the beginning of 1985 (Lloyd's Maritime Asia, 1990). Each block train had 52-55 wagons, carrying up to 110 TEU and these were dispatched to five Soviet border destinations: Leningrad (for UK/Baltic traffic), Chop (Czechoslovakia/Hungary), Brest (Poland/Germany), Djulfa (Iran) and Lujaika (Finland) within 20-21 days from Japan. This newly established service was co-ordinated by an expert non-vessel operating common carrier, Jeuro Container Transport Inc.

"In its capacity as general agent for v/o Soyuztransit (SOTRA)-the Soviet body responsible for operating the Trans-Siberian Container Service-Jeuro arranges all block trains bookings through its Yokohama office placed by fellow members of the Trans Siberian Intermodal Operators Association of Japan (TSIOAJ)" (Lloyd's Maritime Asia, 1990).

Moreover, along with the new block train system came further improvements in the TSR's operational system, including the introduction of a computer tracking system to monitor the movement of containers along the railway.

The overall movement of containers via the Trans Siberian Railway remained stable between 1980 and 1989 however after 1989 the major political differences between the Soviet Union and its East European satellites, led to a great uncertainty over its future.

The collapse of the Soviet Union in 1989 negatively influenced the development of the Trans Siberian Railway and its co-operation with foreign partners, to a large extent because all of the Baltic ports which were purposely developed in Soviet times in order to facilitate Soviet trade were located in countries, which chose to leave the Soviet Union and become independent (Cargo Systems, 2002).

Conclusion

Over the years the TSR served a number of purposes and played a significant role in the Russian as well as Soviet economy. A lot of the issues regarding the development of this railway line have been discussed within this article, however plenty more remain yet to be investigated. Since the collapse of the Soviet Union the dramatic political and economic developments in the former communist countries were closely followed by the rest of the world (Estrin, 1994). In particular the transport sector attracted a lot of interest. Western governments, companies and international organizations were eager to be kept informed, to understand, to advise, to trade, to invest and to be involved in one way or another in this region.

This renewed interest has focussed perhaps more than anything else on the potential that the TSR possesses for transporting containers to Europe from the Far East and to provide an alternative service to the ocean route and a source of income for Russia.

This historical discussion creates the base for a future article in which issues relating to the operational future of the TSR, following the collapse of the USSR will be analyzed.

References

- Bergstrant, S. and Doganis, R. (1987) *The impact of Soviet Shipping*, Allen & Unwin Publishing, London
- Bernard, P.J. (1966) *Planning in the Soviet Union*, Pergamon Press, London.
- Cargo Systems (2002) "Baltic Ports" March, p.13.
- Cullinane, K. and Toy, N. (1998) "Planned Road Network Developments in the Baltic Sea Region", *Transport Reviews*, Vol. 18, 1, pp.37-49.
- Culloton, J. (2002) "American Troops in Northern Russia and Siberia, World War I 1918-1920", <http://www.militaria.com/8th/WW1/siberia.html>.
- Estrin, S. (1994) *Privatisation in Central & Eastern Europe*, Longman Publications, London.
- Gaddis, J. L. (1990) *Russia, the Soviet Union, and the United States: An Interpretive History*, Columbia University Press, New York.
- Giffin, F. C. (1998) "Trans-Siberian Railway in the world history", June, <http://www.icc.ru/fed/transsib.html>.
- Graves, W. S. (1931) *America's Siberian Adventure, 1918-1920*, Longmans, New York.
- Helmer, J. (1999) "Moller weights future of Russian intermodal route", *Journal of Commerce*, 2 August.
- Jacqueline, D. St. J. (1969) "John E Stevens: American Assistance to Russian and Siberian Railroads" 1917-1922, *PhD thesis*, University of Oklahoma.
- Jain, R. (1993) *Germany, The Soviet Union and Eastern Europe, 1949-1991*, Sangam Books Ltd., London.
- Johnson, B. O. (1923) "American Railway Engineers in Siberia", *The Military Engineer*, May-June, 15, 81, p.191.
- Johnson, B. O. (1923) "The Trans Siberian Railway", *The Journal of the Worcester Polytechnic Institute*, July, 180-182.
- Jorre, G. (1961) *The Soviet Union*, Longman Publications, London.
- Karbonski, A. (1992) *The Columbia History of Eastern Europe in the Twentieth Century*, Columbia University Press, New York.
- Kennan, G. F. (1967) *Soviet-American Relations, 1917-1920: The decision to Intervene*, New York.
- Lavigne, M. (1999) *The Economics of Transition*, Mac Millan Press, London.
- Lloyd's Maritime Asia (1990) "Trans Siberian Railway: Contest on the Orient Express", December.
- Lloyd's Shipping Economist (1979) "Threat from Rail Link", February, p. 13.
- Lloyd's Shipping Economist (1980) "Russians Unjustly Persecuted?", January, p. 46.
- Lloyd's Shipping Economist (1980) "Soviet Rail Link Delay", October, p. 17.
- Lloyd's Shipping Economist (1981a) "Trans Siberian Container Line", July, p. 43.
- Lloyd's Shipping Economist (1981b) "Trans Siberian Railway", April, p. 79.

- Lloyd's Shipping Economist (1982) "Struggling into Container Age", July, p. 38.
- Lloyd's Shipping Economist (1983a) "Iran Trade Starts to Recover", August, p. 31.
- Lloyd's Shipping Economist (1983b) "South Korea and North Europe Liner Trade", November, p. 64.
- Lukov, B. E. (2000) *Interview*, General Secretary of the Coordinating Council on Trans Siberian Transportation, June.
- Mathieson, R. S. (1975) *The Soviet Union*, Heinemann Educational Books, London.
- Mellor, R. (1975) *Eastern Europe: A Geography of the Comecon Countries*, Macmillan Press Ltd., London.
- Miller, E. (1978) "The Trans-Siberian Land Bridge, A new Trade Route between Japan & Europe", *Soviet Geography*, Vol. XIX, 4, pp.34-38.
- Moore, K. A. (1980) *Development of the USSR*, Greenwich Forum VI, p.138.
- Queiroz, C. (2001) "Major Trends in the Transport Sector and Impact on the Baltic States", *Trans Baltica 2001 International Conference*, Riga, Latvia.
- Roe, M. S. (2001) *Polish Shipping Under Communism*, Ashgate Ltd, London.
- Slepven, I. (1996) "The Trans Siberian Railway", *History Today*, 46, pp.134-145.
- Soviet Geography (1990) "Analysis of a Railway's Past, Present, and Future", May, Vol. XXI, No. 5.
- Soviet Shipping Journal (1982) "Trans Siberian Container Service", February, 25-27.
- Soviet Shipping (1989) "CMEA Jubilee and Prospects", January, p12.
- U.S News and World Report (1975) "A Land bridge Across Russia-How It's Working", December 15.
- Unterberger-Miller, B. (1989) *The United States, Revolutionary Russia, and the Rise of Czechoslovakia*, Chapel Hill Publishing, London.
- Zheleznodorozhnyy Transport (1975) "Automated Control of Container Transport", July, 36-37.
- Zurek, J. (2001) *The Role of Seaports in Region Development*, University of Gdansk Press, Gdansk.



The development of the highway network in Poland and the future development of polish ferry shipping

Elzbieta Kapsa¹, Michael Roe^{2*}

¹ *Technical University of Gdansk, Gdansk, Poland*

² *Centre for International Shipping and Logistics, University of Plymouth, United Kingdom*

Abstract

The aim of this paper is to emphasize the role of the developing network of roads and motorways in the Polish economy and its impact upon expanding competitive solutions in alliance with the ferry shipping industry. The paper begins with an assessment of Poland within its geographical location in Europe and of its economic connections with other European nations. The first part of the paper presents the current situation of road infrastructure in Poland. The second part presents the proposed transport corridors which cross a series of European countries and the plans for development of the sections of these routes located in Poland. The next part assesses the advantages for Poland and for ferry shipping in particular.

Keywords: Highways; Ferry shipping; Poland.

Introduction

One of the Baltic Sea countries, Poland is located on four European transit routes North–South and East–West. Its strategic transit location provides an important opportunity for development for Poland as a whole. The ports of the Scandinavian countries and Denmark in particular assume the foreground for Polish ports and these countries are characterised by a high level of economic development and living standard for their inhabitants. Meanwhile the countries lying along the North-South transport corridor such as: Poland, the Czech Republic, Slovakia, Hungary and Austria (the Central European states), and Romania, Bulgaria, Greece, Turkey, the countries of the former Yugoslavia and North Italy (the South–East European states) are the hinterland for Polish ports. This basic hinterland should also include Lithuania, Latvia, Estonia, Russia (Kaliningrad), Belarus and West Ukraine. The hinterland for the port of Szczecin-Swinoujscie in north-west Poland is also Germany and includes about 40 million citizens. The countries belonging to the near hinterland for the Polish based ferry shipping industry occupies an area almost 50% bigger than the area of the foreground (over 1780 km²) and more than 143 million people live there.

* Corresponding author: Michael Roe (mroe@plymouth.ac.uk)

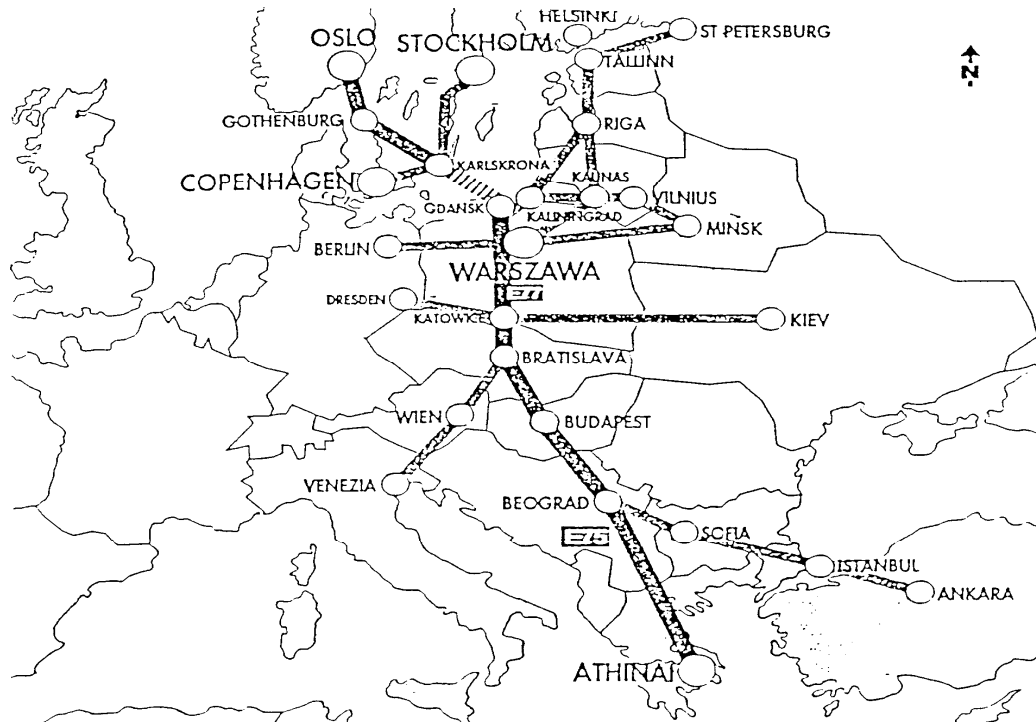


Figure 1: location of Gdansk and Gdynia ports in the North–South transport corridor.

Source: Szwanowski, S. (1996) *Adaptation of Gdansk agglomeration ports to the service of fast ferry connections*, in Roe, M.S. *International studies in shipping policy and management*. University of Plymouth, Plymouth.

From 1st of May 2004, Poland has been a member of the European Union. Thanks to the continuous integration of Poland with the other European countries it can now play a part in the whole EU area characterised without internal borders where free movement of people, goods, services, capital and ideas is encouraged and sustained.

The current situation of road infrastructure in Poland

The present condition of road infrastructure in Poland is one of the greatest barriers to the growth of the Polish economy and more specifically it has serious impacts upon specific industrial and commercial activities. One such activity vital to the improvement of Polish overseas trade is Polish ferry shipping centred on the three main Polish Baltic Sea ports of Gdansk, Gdynia and Szczecin-Swinoujscie. The existing highway network does little to ensure the provision of a suitable quality of service for both passenger and freight carriers. Poland still has less than 300 km of motorways whilst Germany, a neighbour and of comparable physical size, has over 11,000 km. In addition, the physical and technical condition of most of the other existing roads in Poland is very poor and it has been estimated that substantial repairs are needed to 63% of the length of all Polish highways to make them reach general European standards (Figure 1).

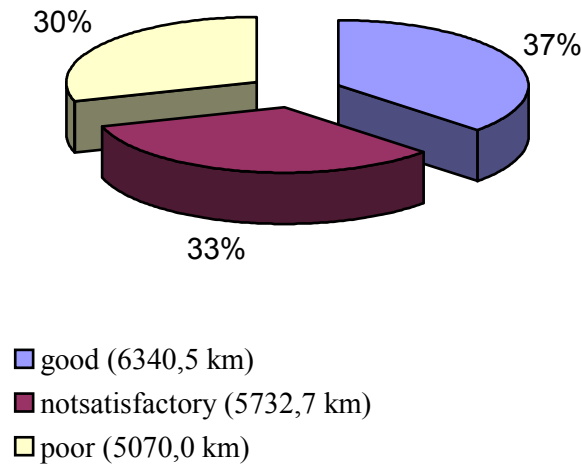


Figure 2: Estimation of the technical situation of highway pavements in Poland (2005).
 Source: Polish Ministry of Infrastructure.

The improvements needed to the non-motorway highway network includes a range of different repair works: reinforcements, smoothness, improvements to non-skid properties and ensuring that pavements are watertight.

In addition to this undesirable situation, both the quantitative and qualitative situations that exist in the Polish economy inhibits growth in both passenger travel and freight transport, something that is becoming a severe problem following Poland's entry into the European Union in 2004. Poland actually possesses a far too small highway capacity with at present only 3% of Polish roads with a breadth of more than eight meters. Such parameters are far worse than general European standards or the situation found in most EU countries. In addition there are almost no domestic nor international roads in Poland that provide standards in operating conditions that meet the demands of TIR movements. Within the European Union, the standard axle pressure that roads should provide for heavy goods vehicles is 115 kN, whilst Polish roads are constructed in the main for axle pressures ranging from 60 to 80 kN. Only a very small proportion of Polish roads can accommodate truck axle pressures of 100 kN.

The demands in terms of the range of road movements continues to grow. The estimates for 2005 suggest that about 9000 vehicles drive a day on what Polish international motorways there are, about 6500 on other international roads and 3000 on domestic roads.¹ During the last 10 years these movements have more than doubled. This is a result amongst others of the fact that road vehicle transport is now responsible for over 80% of freight carried. In addition (and perhaps more significantly) car ownership has also increased by about 70% in the same time period so that now every third Polish citizen is an owner of a car.

As a consequence of this situation, Poland is not attractive for freight transit movements. The shortage of motorways, the substandard and deteriorating condition of

¹ Klimek, H. (1999) *Motorways in Poland – opportunities or threats for success for maritime ports*, (in) Competition of maritime transport. Economy of maritime transport. Scientific exercises of University of Gdansk, Gdansk.

road surfaces, the shortage of grade separated road junctions, the failure to design roads for heavy truck movements, the absence of by-passes around city centres and the small number of bridges and viaducts causes foreign exporters to avoid Poland where they can and to direct their goods (particularly in Eastern Europe for the Swedish and Norwegian markets) to the much better developed network of German motorways with clear ramifications for Baltic Sea ferry operators operating out of Poland. This in turn leads to a reduction in competitiveness for Polish Baltic ports and also a reduction in opportunities for the development of the Polish economy resulting in deficiencies in international trade, worker mobility and earning foreign capital.

Transport corridors

Poland, thanks to its geographical location and the potential capacity of its domestic market, has enormous potential for economic development serving markets to both the east and west and also as a transit route for north-south movements between South-East Europe (The Balkans, Greece, Turkey), and the Middle East and Scandinavia.

Taking advantage of the potential that Poland displays depends however, on taking the decision to develop the existing system of roads and motorways. This in turn will have undoubted benefits for the ferry industry located in Poland in the three major ports of Gdansk, Gdynia and Szczecin-Swinoujscie. The key improvements needed for highways, particularly for the future development of Polish maritime transport, rests with the A1 motorway, representing the Polish stretch of the International North-South (TEM) motorway and also the A3 motorway, which will connect the ports of Swinoujscie and Szczecin with the southern border of Poland. The A1 motorway is one section of the third North-South Transport Corridor passing through the Polish Baltic ports of Gdansk and Gdynia, and eventually reaching the countries of the Near East, the basin of the Black Sea and the Mediterranean Sea.

The idea of pan-European transport corridors came into being at the 1st EU National Transport Conference in Prague in 1991. During that conference, the initial plan for European transport networks was agreed. As well as new proposals (such as those which incorporated Polish motorways), the plan also included existing agreements concerning the rebuilding and modernization of many major roads connecting a substantial number of European Union countries and beyond.²

The concept of building a North-South Trans-European Motorway had already been developed many years earlier in 1972 under the pre-transition regime. The motorway was to have its origin in Gdansk and would pass in the direction of the Mediterranean Sea and farther on via Turkey to the Ports of the Persian Gulf.³

² Andruszkiewicz, W. (1997) *Port in Gdansk and the other Polish maritime ports in multimodal land and maritime transport*, (in) Gdansk on transport map of Europe. The Scientific Session, 50th Anniversary. Polish Economic Company, Gdansk.

³ Andruszkiweicz, W. (1997) Why building of Transeuropean Motorway North-South (TEM/A-1) has its beginning in Rusocino located on the South form Pruszcz Gdanski instead in Port of Gdansk? *Spedycja i Transport*, Nr 3/97, pp. 31 - 34.

The European Economic Commission and the United Nations agreed plans for the building of three main transport corridors which would run North–South across European space. The first of them has its beginnings in the United Kingdom near to London, crosses the English Channel, France, Spain, the Mediterranean Sea in the region of the Gibraltar Strait and then through North Africa and ultimately in the direction of West Africa. The second corridor begins in Denmark, crosses Germany, Italy, Sicily and the Mediterranean Sea to North Africa. The third corridor, the most important for Poland, begins in Finland, crosses via the Baltic Sea to Gdansk and Gdynia and then passes through Poland to the Czech Republic, Slovakia, Hungary, Austria, the Balkans, Greece, Turkey and on eventually to the Persian Gulf and Middle East. This proposed motorway route is one of the longest in the world stretching some 10,000 km including a number of formal branches. By 2005, some 4,000 km of the motorway has been constructed and the next 3,000 km are in the building stage. Another 3,000 km are currently being designed. The beginning of the motorway in Poland is located in the ports of Gdansk and Gdynia and the Polish stretch of this TEM is numbered motorway A1. The construction of this motorway would create the shortest traffic artery connecting Southern Europe with Sweden and the remainder of Scandinavia. In 1997 at the 3rd EU Transport Conference in Helsinki, the concept of transport corridors was completed with the introduction of Motorways of the Sea representing the sections of proposed trans European motorways that crossed seas including the Poland-Scandinavia section of the A1. The total number of corridors was increased to 10.

Unfortunately, the Polish part of the project has not been finished some 30 years after the inception of the idea of building motorway A1. One of the main reasons for this has been the shortage of financial resources. Since 1993, from the moment when it was realised that Poland could not afford free motorways, it was decided, like many other West European countries including France and Italy, to construct tolled roads. The new programme was agreed in 1994 taking into consideration building of 2,600 km of tolled roads during the next 20 years – an average of 160 km of motorways a year. According to this new plan there was to be built the four following motorways:

- A1 (597 km) from Gdansk via Lodz, Katowice to Gorycze near to Rybnik;
- A2 (626 km) Swiecko – Poznan – Warsaw – Terespol;
- A3 (440 km) Szczecin – Gorzow – Zielona Gora – Legnica – Lubawka;
- A4 (738 km) Zgorzelec – Wroclaw – Katowice – Krakow – Medyka.

However it soon became clear that this ambitious plan was not realizable and as a result the Polish government had to reduce the length of planned roads by 600 km. However, this reduction in highway length could not help sufficiently. The severity of the problem became clear when it was realized that in order to adapt Polish roads (not including motorways) to European standards there was needed up to 2015 about 90 milliards zlotys, that is to say more than a half of the entire budget of the country. Motorway construction would be in addition to this.

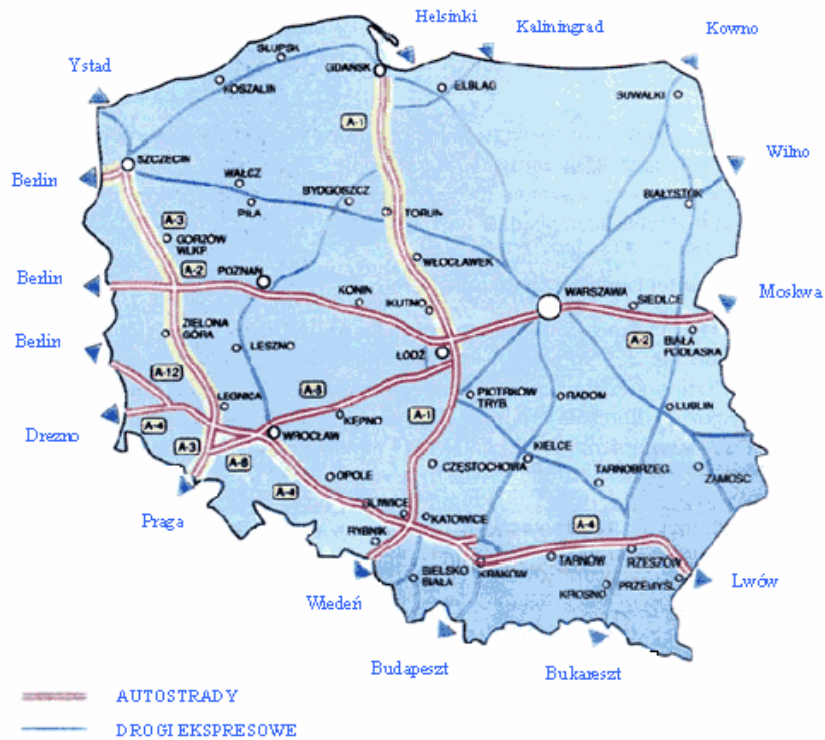


Figure 3: planned motorways and express roads in Poland.
 Source: Ministry of Infrastructure.

The latest plan for the roads and motorways network of Poland

In 1998 the Polish Ministry of Transport and Maritime Economics concluded a project entitled “Transport Policy of the Country for 2000–2015; Years for Eco-development”. The project examined the ways of creating the necessary conditions for the integration of the Polish highway transport network with the European network and suggested possibilities for improving new technologies for multi-modal transport and the construction of European transport chains. It was concluded that 7% of Polish gross domestic product a year should be spent upon the development of transport in order to achieve these aims within the next 10 years. Contained within the conclusions of the project were aims at integrating the Polish economy including building the following planned transport networks:⁴

- the modernization and construction of road and railway networks forming part of the Trans-European transport corridors (TINA) including activating a programme of raising standards and consolidating road surfaces and bridges;
- the improvement of the management system for roads and traffic movements and improved control of compliance with rules concerning safety of highways.

⁴ Polish Ministry of Infrastructure

The programme of constructing road infrastructure in Poland consists of three main parts - motorways, express roads and accelerated roads. The aim of the programme is the creation of an efficient and effective transport network, including contributions to four priority European corridors:

- I Helsinki – Tallinn – Riga – Warsaw.
- II Berlin – Warsaw – Minsk – Moscow – Niznyj Nowograd.
- III Berlin – Drezno – Wroclaw – Krakow – Przemysl – Lwow – Kiev.
- IV Gdansk – Warsaw – Katowice – Zilina.

As part of the Polish contribution to these networks it was decided to build the three following motorways by the year 2015:

- A1 – Gdansk – Torun – Czestochowa.
- A2 – Swiecko – Poznan – Warszawa – Terespol (on the border with Belarus).
- A4 – Zgorzelec – Wroclaw – Gliwice – Katowice – Krakow.

However, the plan to build motorway A3 in the near future, the most important investment within the western maritime region, was almost immediately interrupted. The motorway was to have its beginning in Szczecin and pass by Gorzow, Zielona Gora, Legnica and Lubawka to reach the border with the Czech Republic. The motorway was subsequently limited to the stretch between Szczecin and the beginning of motorway A2 (see Figure 2) and farther on was to be replaced by Express Road S3 from Zielona Gora. The building of the S3 road should be finished earlier than the motorway would have been and the cost of its construction is likely to be lower by about 30%. Besides this plan, the building of Express Roads connecting Wroclaw (via Poznan) with Bydgoszcz, Bydgoszcz (via Warsaw) with Lublin and the border of the country, and the border of Poland with Slovakia (via Krakow, Kielce, Warsaw, Bialystok) with the border of Lithuania, is planned according to the transport network programme. This key network is to be completed by a number of domestic roads amongst others including: Szczecin–Bydgoszcz, Katowice–Kolobrzeg, Gdansk–Warsaw, and Rzeszow–Bialystok. It is planned to build 1572 km such roads by the end of 2015. Earlier it had been the plan to build about 23,000km of such roads but because of the shortage of financial resources, this aim was clearly unrealistic.

During the activities of the project - in the period between 2004 and 2006, and with the long-term perspective up to 2008 – there is planned to be realized the following tasks:⁵

- connecting Warsaw with Swiecko and the western border of Poland by Motorway A1;
- connecting Tarnow, Krakow, Katowice and Wroclaw with Germany by Motorway A4 and A18 (In Germany Motorways nr 4 and 15);

⁵ Ibidem.

- connecting Gdansk, Grudziadz, Motorway A2, Lodz, Czestochowa with the Czech Republic (at Gorzyce) by Motorway A1;
- modernisation of Motorway A6 from Szczecin to the Polish border at Kolbaskowo (with connection with Germany and Motorway nr 11) [plans for 2004 and 2005];
- building Express Road nr S8 from Warsaw to Wyszkow;
- building Express Road nr S22 connecting Gdansk with the Polish border at Grzechotki (and then onwards to Russia).

In addition it is intended to reconstruct a part of a number of roads in order to adapt them to a standard of pressure of 10 tons per axle between 2004 and 2006. The roads to be modernized are as follows:⁶

- nr 1 Torun – Lodz (144 km).
- nr 2 Warsaw - Terespol (132 km).
- nr 4 Krakow - Tarnow (56 km).
- nr 4 Rzeszow - Radzynie (70 km).
- nr 50 Sochaczew – Mszczonow – Grojec – Minsk Mazowiecki (140 km).

Further developments in the Polish plans include related improved facilities for the domestic roads network, which are proposed in the Treaty between Poland and the European Union. These include rebuilding roads nr 5, 7, 12 and 17 between 2007 and 2013. The roads will be adapted to accommodate pressures of 11.5 tons per axle.⁷

In order to realize all these aims in the re-building of the highway infrastructure in Poland, there is needed a degree of substantially increased financial resources. The following sources of finance have been suggested and will be followed up: public sources (taxation), planned fuel payments, EU grants and loans; and credits from the other international financial institutions.

The cost of building these 1572 km of roads is about 10bn euro including an average price of building 1 km of motorway in Poland of 4.1m to 5m euro. In Silesia, because of the damages caused by the mining industry over many decades, the price of 1 km of motorway is twice as high. Constructing tolled motorways can help recover these costs and also create profits for the operators. According to the USA consulting company Wilbur Smith, between 75,000 and 94,000 vehicles a day will use the motorway facilities partly dependant upon the price chosen. According to more optimistic forecasts it could even be between 15,000 and 22,000 vehicles a day.⁸

Only decisive actions at a political level, particularly in terms of creating appropriate financial conditions, will create the basis for realizing these plans for the highway network in Poland. An appropriately designed and powerful organizational structure will also be needed to activate and direct motorway construction or else little will occur. Finally, a sufficiently streamlined and well designed regulatory regime will also be important.

⁶ Ibidem.

⁷ Ibidem.

⁸ Ibidem.

The advantages for Poland

The creation of a well developed and appropriate highway network for Poland should be a priority for the Polish government and is now essential following membership of the European Union. It is motivated by many different and undeniable advantages. In particular ensuring the good condition of Polish roads and increasing of number of cross-roads free from the possibility of collisions will have a major influence for improvement of safety and the reduction in the number of accidents which currently occurs.

Economic advantages also clearly play a very important role. Included among them is the activation of international trade, an improvement in the accessibility of particular cities and regions, the possibilities for developing new areas for the employment market and a general reduction in the time for travel. Development of the road network encourages integration between different centres of economic and cultural development and increases access to high level services and different branches of industry.

However, apart from these advantages it is perhaps most necessary to acknowledge the transit advantages, both in the range of passenger transport and goods carriers and it is here that the links with current and potential ferry operations from Polish ports is most apparent. An increase in transit traffic has a clear influence upon the activation of development of Polish maritime ports in addition to its impact upon ancillary services including amongst others: hotels, gastronomy, fuel suppliers and tourism. It also has a significant and beneficial influence upon the development of other branches of industry (for example food, clothes) which may be given advantages by additional transit facilities encouraged by highway improvements.

The Polish motorway network and the ferry shipping sector

In 2005, in Poland two Polish (Polferries and Unity Line) and one Swedish ferry company (Stena Line) are operating across the Baltic Sea. Polferries' services operate from Gdansk to Nynashamn (Sweden) and also compete with Unity Line on the line from Swinoujscie (Poland) to Ystad in Sweden. The Swedish operator Stena Line serves the connection between Gdynia and Karlskrona (Sweden), where two passenger/car units and one passenger/car/trailer ferry are in operation.

These ferry carriers from Poland are mainly attractive for movements in two transport corridors which run North–South:⁹

- Western – connecting Scandinavia and Denmark with Germany and Poland, where about 9 million passengers, 2 million cars, more than 1 million trailers and more than 130,000 rail wagons are carried each year;

⁹ *ShipPax Statistics 01. The Yearbook for Passenger Shipping Traffic Figures*. Halmstad, Sweden 2001, s. 83 – 126.

- Central - connecting Sweden and Finland with the agglomerations of Gdansk, Riga (Latvia) and Klaipeda (Lithuania), where more than 400,000 passengers, 60,000 cars and 21,000 trailers are carried each year.

The share of Polish ferry operators in the German market is currently very small with nearly all carriage on ferries serving Polish ports peripheral in character with limited freight transport and tourist traffic. The most important route in this western sector for Poland is the connection from Gdynia to Karlskrona in Sweden, on which the movement of passengers has increased by about 40% over the last 20 years.

The biggest role played by ferry shipping in Poland is that connected with marine tourism. The most important customers are citizens from Sweden, Finland, Norway and Denmark. In last few years the number of passengers arriving from those countries increased substantially and with the expansion of the EU, these numbers are bound to increase. The number of travellers from Scandinavia to the countries of South Europe (for example to Italy, Greece and Spain) has also increased dramatically. In 2004, about 20% of all inhabitants of the Scandinavian countries crossing Poland traveled to South Europe. In all in 2004, there were around 450,000 foreign passengers (mainly from Sweden amounting to around 50 % of the total).¹⁰

In recent years the Scandinavian countries have become increasingly more popular with travelers from the rest of the European continent as well attracted by the environment offered by Norway and Sweden. Every year over 4 million passengers travel on ferry lines connecting German and Polish ports with Scandinavia although the large majority of these use German ports. Poland is transited only by about 5% of passengers from the south of Europe. However, thanks to the integration of Poland with the EU, the number of passenger movements via Polish ports is likely to increase because of its strategic location lying along the shortest communication routes in a North-South direction.

Within the Polish ferry market two basic groups of passengers can be identified. The first of these are residents from Poland who make up about 40% of all ferry passenger movements passing through Polish ports.¹¹ Amongst this group it is worth noting:¹²

- tourists going on vacation to Scandinavian countries and Denmark (mainly in the summer season);
- participants of marine cruises;
- Polish drivers of trucks and other passengers going to Scandinavia and Denmark for business.

The second group consists of foreign passengers making up around 50% of those carried. Amongst this group should be noted:¹³

¹⁰ Institute of Tourism.

¹¹ Unity Line.

¹² Urbanyi - Popiołek, I. (1998) *Market of ferry shipping in North Europe*. The University of Gdansk, Sopot.

¹³ Ibidem.

- Scandinavian and Danish citizens traveling to Poland for vacation (commonly with Polish connections with their family living abroad) mainly in the summer season;
- inhabitants of Scandinavian countries, Denmark and Central and South Eastern Europe transiting through Poland;
- foreign drivers of trucks;
- foreign participants of marine cruises.

The other identifiable group of customers is made up of institutions. Amongst them we should mention first of all companies organising conferences and courses for their workers on board ferries and also private schools organising different meetings for pupils.

In the case of cargo traffic, Scandinavian countries and Denmark play the key role for Polish based ferry shipping. However these countries have only a small share of their foreign trade with Poland totaling in terms of exports and imports between 8 and 8.5%. Comparing the trade of Poland with these countries, the biggest share – also in terms of total exports and imports is with Denmark which has 30% of Polish imports and 32 % of exports in 2004. Sweden co-incidentally has the same pattern of trade. Finland is less significant with about 29% of Polish imports and 8% exports. The smallest share is with Norway with about 9% of imports and 8% exports. Electrical machinery, food, chemicals and light industrial goods play the most important role in the structure of cargo movements.

The share of goods trade from the Eastern Baltic countries of Lithuania, Latvia and Estonia remains of little importance in Polish foreign trade with about 2.3% of Polish exports and 0.4% of imports¹⁴.

A very important market for Polish based ferry shipping is the transit of general cargo. In Polish ports this category represents about 80% of total cargo each year.

The main countries for Poland with respect to international transit trade are the Czech and Slovak Republics. Other countries, which commonly use Poland for transit are: Belarus, Ukraine, Russia, Hungary, Romania and Austria (Table 1). Taking into consideration the export and import trade of Poland, trade with the Baltic States and Baltic Sea transit crossings, the share of Polish based ferry shipping of the general cargo trade is about 12%.

According to various forecasts, the demand for ferry operations from Poland should increase as a result of the intensification of international trade between Poland and Scandinavia, Denmark and other European Union countries located along the North-South transport corridor. Passenger movements between these countries will increase also. According to the Polish Organization of Tourism, the number of foreign citizens arriving in Poland will increase by about 2.5% a year for the next 10 years. This increasing trend is also forecast for Polish citizens traveling abroad.¹⁵ The growth up to 2010 is expected to be about 1.9 million passengers, 450,000 cars, 200,000 trailers, 75,000 rail wagons and from 3.5 to 4.0 million tons of general cargo.¹⁶

¹⁴ *Transport – results in 2000 years*. GUS, Warsaw 2001.

¹⁵ *Tourism in 1998*. GUS, Warsaw 1999.

¹⁶ Tubielewicz, A. (ed.) (1994) *Forecast of development of container in Port of Gdynia*. Technical University of Gdansk, Gdansk.

Table 1: general cargo transit reloading in Polish ports 1997 [000tons].

<i>Transit Countries</i>	<i>Gdynia</i>			<i>Szczecin-Swinoujscie</i>					
	<i>Gdansk</i>								
	Import	Export	Total	Import	Export	Total	Import	Export	Total
Slovakia	3.2	34.3	37.5	9.6	0.7	10.3	7	441	448
Czech Rep.	2.1	27.7	29.8	8.1	2.0	10.1	8	305	313
Russia	1.5	0.1	1.6	2.5	1.0	3.5	1	1	2
Belarus	0	0.9	0.9	-	0.1	0.1	-	-	-
Hungary	0.8	-	0.8	0.05	0.04	0.09	-	-	-
Lithuania	0.2	-	0.2	-	-	-	-	-	-
Latvia	-	-	-	-	-	-	-	-	-
Ukraine	-	-	-	0.09	0.02	0.11	-	-	-
Austria	-	-	-	0.01	-	0.01	-	-	-
Estonia	-	-	-	-	-	-	2	-	2
Luxembourg	-	-	-	-	-	-	0	1	1
Other Countries	-	-	-	0.05	0.3	0.35	-	-	-
Total	7.9	63.0	70.8	20.4	4.16	24.56	18.0	748.0	766

Source: Anonymous (1999) *Bearings for Sea and Trade*, Nr 18/1999, p. 14.

Table 2: forecast of passenger and cargo turnover in Polish ports in 2010.

<i>Specification</i>	<i>Swinoujscie</i>	<i>Gdansk/Gdynia</i>
Passengers [thousands]	1000	800-900
Cars [thousands]	250	180-200
Trailers [thousands]	180	70
Rail wagons [thousands]	45	30
General cargo [millions of tons]	2.8-3.0	0.7-1.2

Source: Tubielewicz, A. (ed.) (1994) *Forecast of development of container in Port of Gdynia*, Technical University of Gdansk, Gdansk.

In the near future the potential exists for new inland motorway based routes to be developed through Poland which will attract traffic away from the Polish ferry sector. This is especially the case on the East-West route running from Russia, through the Baltic States and Belarus and Poland and on to Germany and Western Europe. Currently some of this traffic uses Polish based ferry operators for some of their route (for example trucks make great use of the Tallinn-Helsinki ferry operations rather than Poland-Finland links because of the problems of transiting Poland). Meanwhile, the development of North-South motorway links targeting Polish ports would have a dramatic effect upon Polish based ferry operations as it would provide a viable road based alternative to the current trend of avoiding Polish ports and using those in Germany, Lithuania and beyond which have better road links. Polish ferry carriers will continue to have to function under pressure of a developing transport system for some years: in other words freight taking land and marine (ferry and ro-ro) services from

Moscow, St. Petersburg, Riga and Tallinn traveling via the ports of Finland to West Europe avoiding Polish ports. The development of new motorways would make routes via Poland more attractive.

Consequently, in order to develop Polish ferry shipping operations beyond their currently limited activities there is a substantial need to build a network of motorways and in particular to complete the North-South proposals which remain largely unfinished despite a planning period now exceeding 30 years. Sources of finance remain the main problem and need to be acquired from other than the Polish state which in the foreseeable future will not have the resources to devote to this issue. Grants and loans from the EU, income from fuel taxation, and proposals for road pricing and private sector investments remain alternatives much talked about but with little progress. Without this investment, Polish ferry shipping activities will remain insignificant despite Poland's strategic location and port facilities.

If we accept the prognosis of the Polish Ministry of Infrastructure, the forecast increase of cars and truck/trailer movements in Poland in the next 10 years confirms the necessity to build the A1, A2 and A4 motorways in particular. If this does not happen we can expect that much of the traffic between the Balkans, South-East Europe, Slovakia, the Czech Republic, the Middle East and Scandinavia/Finland will continue to use ferry services across the Baltic which avoid Polish ports and the services that they offer.

Conclusions

Without a modern and fully developed motorway system which mirrors that found in the established countries of the EU, Poland will have few opportunities for the effective functioning and revitalization of its ferry shipping industry which has the potential to create wealth and employment for the country. Currently financial problems make the construction of these roads and associated ferry facility developments unlikely leading to the continued diversion of North-South movements of cargo and passengers to the Baltic States and Germany highly probable.

References

- Andruszkiewicz, W. (1997) Why building of Transeuropean Motorway North-South (TEM/A-1) has its beginning in Rusocino located on the South from Pruszcz Gdanski instead in Port of Gdansk? *Spedycja i Transport*, Nr 3/97.
- Andruszkiewicz, W. (1997) *Port in Gdansk and the other Polish marine ports in multimodal land and marine transport*, In: Gdansk on transport map of Europe. Scientific Session; 50th Anniversary. Polish Economic Company, Gdansk.
- Anonymous (1999) *Bearings for Sea and Trade*, Nr 18/1999, p.14.
- Klimek, H. (1999) *Motorways in Poland – opportunities for success or threats for marine ports*, (in) Competition of marine transport. Economy of marine transport. Scientific Exercises of University of Gdansk, Sopot.
- ShipPax Statistics 01. The Yearbook for Passenger Shipping Traffic Figures*. Halmstad, Sweden 2001, s. 83-126.

- Szwankowski, S. (1996) *Adaptation of Gdansk agglomeration ports to the service of fast ferry connections. International studies in shipping policy and management.* University of Plymouth, Plymouth.
- GUS (1999) *Tourism in 1998*, Warsaw.
- GUS (2001) *Transport – results in 2000 years*, Warsaw.
- Tubielewicz, A. (ed.) (1994) *Forecast of development of container in Port of Gdynia.* Technical University of Gdansk, Gdansk.
- Urbanyi-Popiołek, I. (1998) *Market of ferry shipping in North Europe.* The University of Gdansk, Sopot.



Methodology for finding optimum cell size for a grid based cellular automata traffic flow model

P. J. Gundaliya¹, Tom V. Mathew^{2*} and S. L. Dhingra³

¹Civil Engineering Department, L. D. College of Engineering, Ahmedabad, INDIA

²Transportation Systems Engineering, Civil Engg. Dept., I.I.T Bombay, INDIA

³Transportation Systems Engineering, Civil Engineering Department, I.I.T Bombay, INDIA

Abstract

A methodology for determining optimum cell size for a grid based traffic flow model for heterogeneous traffic is proposed in this paper. The cell size is an important factor to determine as it affects the computational efficiency and model accuracy. The objective function minimizes three aspects namely the difference of distance headway in case of cellular automata and grid based traffic flow model, the total number of cells to represent different types of vehicles and multiple of cell width that gives closer representation of the different road widths. The presented method is found better than the previous attempt which tries to find the cell size by trial and error.

Keywords: Heterogeneous traffic flow; Simulation; Grid based approach; Cellular Automata.

Introduction

Deciding a cell size is crucial for the cell-based simulation approach. In the present study, a systematic approach is given to decide the cell size in cellular automata based modelling for heterogeneous traffic flow. In this model, size of the cell is carefully decided according to the types of vehicle. It is decided in such a way that it represents the actual size of vehicles and the total width of the road as close as possible. The physical representation of the vehicle should be kept slightly more than the actual size of vehicle to provide some clearance. The cell length also depends upon the dynamic characteristics of the vehicular movement, as in the cellular automata, distance-headway and speed is considered in terms of number of cells. The cellular automata (CA) traffic flow model developed by Nagel and Schreckenberg (1992) is used for comparison of

* Assistant Professor of Transportation Systems Engineering, Dept. of Civil Engg., IIT Bombay, Mumbai-76, India. Ph: 91-22-25767349, Fax: 91-22-25767302 E-mail: vmtom@iitb.ac.in.

distance headway at different speed. If cell size is taken small, it can represent the physical features of vehicles more accurately but at the same time it will reserve the huge memory for computation. However, in CA based model, the length of the cell affects the dynamic characteristics of vehicle like speed, acceleration, and deceleration. This restricts cell size selection unlike the other heterogeneous models developed where there is no bar except the computational limitations. Considering all these aspects, the maximum possible cell size is decided which gives minimum and maximum desired clearance and the dynamic characteristics as close as to NaSch model.

The optimal cell size selection results in better model performance and satisfies the minimal modelling concept. It also gives better physical representation of the vehicle.

Background

Several simulation studies for heterogeneous traffic have been conducted in the past to capture the mixed nature of traffic flow prevailing in developing countries. These studies are different from those models of homogeneous traffic dominated by cars and heavy vehicles. These models have to capture the manoeuvres and interactions among the various types of vehicle, with varied in dimensions, speed and acceleration characteristics. There is a need to further understand these complex interactions. It is difficult to get the analytical solution in case of high heterogeneity condition; many researchers have used simulation based modelling approach for the heterogeneous traffic flow.

Considerable researches have been conducted for heterogeneous traffic flow in developing countries. Many simulation models were developed based on the grid based approach. Pillai (1975), Marwah and Ramaseshan (1978), Marwah and Bandyopadhyay (1983), Palaniswamy (1983), Isaac (1995), Chalapati (1987), Kumar and Rao (1996), and Ramamayya (1988) have developed simulation model for the heterogeneous traffic flow. Various cell sizes are taken by different researchers depending on the vehicles used for simulation run. However, none of them have given specific procedure for deciding cell size. They have used trial and error methods for deciding a cell size. Sing (1999) has taken a cell size as 1m x 1m to represent the road stretch. Roy (2000) developed a simulation model for the heterogeneous traffic flow. He considered the cell size as 0.28 m x 0.28 m considering update periods of 0.1 seconds. Korlapati (2003) has taken a similar concept for representing vehicles on the road grid. Gundaliya et al. (2004) has taken cell size as 0.9 x 1.9 meters in their simulation models. They have used cellular automata for the first time in grid-based approach for various types of vehicles including motorised and non-motorised vehicles. Lan and Chang (2004) have developed similar model for two wheelers and cars using 1.25 x 1.25 meters cell size.

The cell size decided in the above studies is based on the vehicle types. They have taken the cell size based on the vehicle types and the accuracy of the model performance. No specific guideline is given for determining the cell size in most of the cases. If the cell size is taken small, the model accuracy for the physical representation is high. However, the time taken for the simulation is increased. Singh (1999) has suggested that the cell size should be decided based on the computational criteria as well as lateral and longitudinal clearance of the vehicle. Hence, in the present study, a systematic approach is developed for determining optimal cell size for the heterogeneous traffic flow model using genetic algorithm.

Methodology

In the grid-based traffic simulation model road is divided into a number of uniform cells. The vehicles are then physically represented on the grid as per their sizes. The physical representation of the vehicles in the single lane is shown in Fig. 2. The vehicles then move according to the model criteria taken. In the present study the model developed by Gundaliya et al. (2004) is taken to formulate the objective function for optimum size of cell of the grid. The model developed by Gundaliya et al. (2004) and Lan and Chang (2004) have used CA concept to update the vehicle position. However, the cell size was decided based on type of vehicles. In these models speed is termed as number of cells per time-step. Hence, 1 cell/time-step means that vehicle advances 1 cell length in one time step. Therefore, the cell size also plays a key role for vehicle dynamics and model accuracy in these types of models.

A minimization problem is formulated considering headway distance, total number of cells and road width. The first term in the objective function (Eq. 1) represents the headway difference such that the model reacts similar to the NaSch basic model at all the cell speeds. Hence, as per NaSch model, the speed is taken from 1 to 5 cells / time-step. According to CA, the minimum headway required is the number of cells ahead of the vehicles as per the vehicle speed. This headway is fixed in the case of NaSch model for each speed as shown in column 6 of Table 2. The difference between these headways will be compared with that of grid-based model at the same speed for the given cell. Minimum difference indicates that the model closely represents the dynamic characteristics of NaSch model. The second term in objective function represents the total number of cells of different types of vehicles. The third term represents the difference between the actual width of the road and the width obtained by multiplying number of cells with width of each cell. This objective function is subjected to following constraints. The constraints (Eq. 2 and Eq. 3) ensure that all the vehicle types are having clearance within the limits of desired minimum and maximum clearance. Constraints (Eq. 4 and Eq. 5) take care of cell length and width such that it is always positive. C_1 , C_2 , and C_3 are the appropriate constant weights for all three objectives in Eq. 1.

$$\begin{aligned} \text{Min } f(L_c, W_c) = & C_1 \times \sum_{i=1}^5 (d_i^N \times 7.5 - d_i^G \times L_c)^2 + C_2 \times \sum_{k=1}^K d_k^W \times d_k^L \\ & + C_3 \times (d_i^{L_n} \times W_c - W_j)^2 \end{aligned} \quad (1)$$

Subjected to

$$C_{\min}^l < C_k^l < C_{\max}^l \quad \forall K \quad (2)$$

$$C_{\min}^w < C_k^w < C_{\max}^w \quad \forall K \quad (3)$$

$$L_c > 0 \quad (4)$$

$$W_c > 0 \quad (5)$$

Where

$$d_i^G = (d_i^N \times \frac{L_c^N}{L_c}) \quad (6)$$

$$d_i^{L_n} = \frac{W_j}{W_c} \quad (7)$$

$$C_k^l = (d_k^W \times L_c - L_k) \quad (8)$$

$$C_k^l = (d_k^L \times W_c - W_k) \quad (9)$$

$d_i^G, d_i^N, d_k^W, d_k^L, d_k^{L_n}$ are integers indicating number of cells.

d_i^N is the distance headway at speed i considering NaSch model, d_i^G is the distance headway at speed i considering cell size of Grid based model, d_k^W is the width of vehicle type k , d_k^L is the length of vehicle type k , $d_k^{L_n}$ is the width of lane (j) considering L_c , where L_c is optimum cell length in meters, L_n is total number of lanes of width W_j in meter of lane (j), L_k is the length of vehicle-type k in meters, W_k is the width of vehicle-type k in meters, L_c^N is the cell length in meters taken as 7.5 meter (NaSch model), W_c is the optimum cell width in meters, C_k^l is the clearance of vehicle type k in meters (lengthwise), C_k^w is the clearance of vehicle type k in meters (width wise), C_{\min}^l is the minimum lengthwise clearance in meters, C_{\min}^w is the minimum widthwise clearance in meters, C_{\max}^l is maximum allowed lengthwise clearance in meters, C_{\max}^w is maximum allowed widthwise clearance in meters, and K is total number of type of vehicle.

The above function is having number of variables and constraints, which are of conflicting nature. Hence, a Genetic Algorithm is used to solve the problem. The objective is to minimize the total square meter for the best representation of cells in grid based modelling approach.

Genetic algorithm

Genetic algorithms (GAs) are a family of computational models inspired by evolution (Goldberg, 1989). These algorithms encode a potential solution to a specific problem on a simple chromosome-like data structure and apply recombination of parameters

(reproduction, crossover, and mutation) to these structures for storing critical information. GAs are often viewed as function optimizers, although the ranges of problems to which GAs have been applied are quite broad. An implementation of GAs begins with a population of (typically random) chromosomes, then evaluates these structures and allocates reproductive opportunities in such a way that those chromosomes that represents a better solution to the target problem are given more chances to 'reproduce' than those chromosomes, which are poorer solutions. The 'goodness' of a solution is typically defined with respect to the current population. The working principle of a GAs is illustrated in Fig. 1. The major steps involved are the generation of a population of solutions, finding the objective function and fitness function and the application of genetic operators. In the present study a cell length and cell width are taken as function variables and the optimum function value is obtained. This methodology is explained in the subsequent section.

```

/* GENETIC ALGORITHM */
formulate initial population
randomly initialize population
repeat
evaluate objective function with the constraint penalties
find fitness function
apply genetic operators
reproduction
crossover
mutation
until stopping criteria

```

Figure 1: working principle of genetic algorithm.

The GA parameters are tuned to get the near optimum solution to determine the cell size which satisfies the constraints. The total chromosome length is taken as 30 units for representing the cell size variable, length and width. Poll size is taken as 40 units and crossover rate and mutation rate is found to be 0.9 and 0.1 respectively. The libGA software is used for getting near optimum solution for determining cell size.

Case study

In this simulation model, seven types of vehicles differing in size have been considered as shown in Table 1. These seven types of vehicles are again classified as per their dynamic characteristics like maximum speed, acceleration etc.. Therefore, non-motorised vehicles like bicycle, bullock-cart, and pedal rickshaws can be considered by selecting appropriate size and dynamic characteristics. This problem has been solved using Genetic Algorithm. To restrict search area for the width and length of cell the search window is given for width as 0.9 to 1.0 meter and length as 1.0 to 2.2 meter. The distance headway for the NaSch model is taken as shown in column 7 of Table 2 for

each speed option. The minimum clearances assumed for this problem are: $C_{\min}^l = 0.1$ meters, $C_{\min}^w = 0.1$ meters, $C_{\max}^l = 1.2$ meters and $C_{\max}^w = 1.0$ meters.

Table 1: type of vehicle and dimension details taken for the study.

<i>Sr.No</i>	<i>Vehicle Type</i>	<i>Width (meter)</i>	<i>Length (meter)</i>
1	2W	0.6	1.8
2	3W	1.4	2.6
3	Car	1.7	4.7
4	LCV1	1.9	5
5	LCV2	1.9	6.8
6	HCV1	2.5	8.5
7	HCV2	2.5	10.3

Table 2: speed and distance headway for different discrete models.

<i>Speed in</i> c/ts	<i>GBTFM</i> (0.9 x 1.9 meters)		<i>NaSch(CA-7.5)</i>	
	Speed (kmph)	Headway (meters)	Speed (kmph)	Headway (meter)
(1)	(2)	(3)	(6)	(7)
1	6.84	1.9	27	7.5
2	13.68	3.8	54	15
3	20.52	5.7	81	22.5
4	27.36	7.6	108	30
5	34.2	9.5	135	37.5
6	41.04	11.4	-	-
7	47.88	13.3	-	-
8	54.72	15.2	-	-
9	61.56	17.1	-	-
10	68.4	19	-	-
11	75.24	20.9	-	-
12	82.08	22.8	-	-
13	88.92	24.7	-	-
14	95.76	26.6	-	-
15	102.6	28.5	-	-
16	109.44	30.4	-	-
17	116.28	32.3	-	-
18	123.12	34.2	-	-
19	129.96	36.1	-	-
20	136.8	38	-	-

Where 2W stands for two wheelers, 3W for three wheelers, HCV1 for heavy commercial vehicle type 1, HCV2 for heavy commercial vehicle type 2, LCV1 for light commercial vehicle type 1, and LCV2 for light commercial vehicle type 2.

Total type of lanes considered as L_n is a two-lane road of 7.0 meters width. The single lane length is taken as $W_1 = 3.6$ meters and two lane width is taken as $W_2 = 7.0$ meters. In the present study, the weightings of all three constants are taken as equal and hence the value for C_1 , C_2 , and C_3 are the same. The values of C_1 , C_2 , and C_3 can be taken as different giving appropriate weightings to the terms in the objective functions. The GA converged after 130 iterations and cell size found as 0.9 meters as width and 1.9 meters as length. The closeness of distance headway of grid based model and NaSch model is given in Table 2. However, this cell size will change as the new vehicles with different sizes are added as well as the minimum and maximum clearance is changed. In the present study the cell size is taken as 0.9 meters and 1.9 meters for the vehicle under consideration.

The dynamic characteristics of the vehicle speed and the minimum distance headway required for the one-second time-step are shown in column 2 and 3 of Table 2 for cell length of 1.9 meters. The cell size is decided based on the different vehicle types and other considerations discussed earlier.

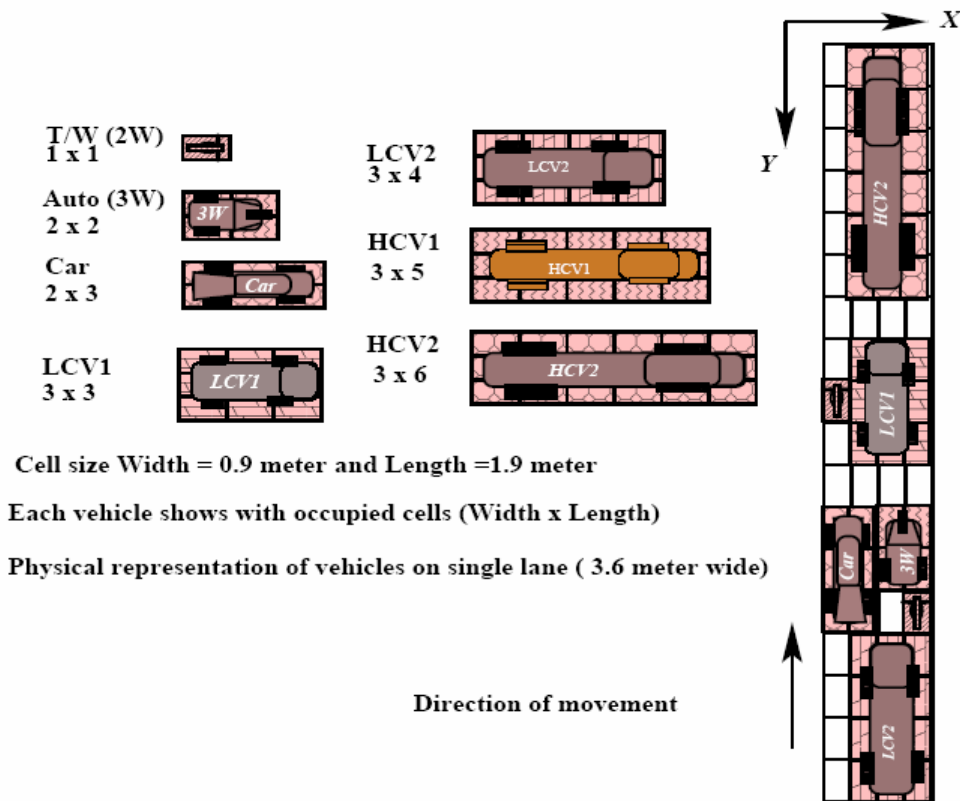


Figure 2: physical representation of vehicles on single lane road.

Results and discussion

The above-formulated problem has been used to find out the optimum cell size for the vehicle types given in Table 1 for the grid based traffic flow modelling. After deciding the cell size, the length and width of the vehicle in terms of number of cells can be obtained by adding clearance to the vehicle's actual length and width. After deciding the size of the vehicle in terms of number of cells in lateral and longitudinal, vehicle can be

physically represented on occupied number of cells. This physical representation of the vehicle in the single lane is shown in Fig 2. The left most corner of the each vehicle represents the position of the vehicle in each time-step. Column 2 of Table 3 shows vehicle type, column 3 and 4 show vehicle actual dimensions of width and length taken in model in meters respectively, column 5 and 6 show the dimensions of vehicles in width and length of vehicles in cells, column 7 and 8 are width and length of vehicle representation taken in present study in meters and column 9 and 10 show the minimum clearance on width and length in meters.

Conclusions

The methodology developed is useful to define the optimum cell size, which represents the vehicles as the nearest as the actual size in terms of number of cells. Moreover it also represents the number of lanes as a multiple of cell width as close as possible. The objective function also takes care of dynamic characteristics of the vehicles where the vehicles gaps are represented in terms of number of cells. However, it needs to decide the weightage of all three objectives numerically, which can be universally applied for any type of function. The methodology applied here is for deciding cell size for the seven different categories of the vehicles. This method is further useful to define the cell size in case of heterogeneous traffic flow particularly where vehicles move forward according to the principle of CA based simulation.

Table 3: Vehicle dimensions details with cell size (0.9 x 1.9 meters).

S.No	Vehicle type	Actual (meters)		Taken in model (cells)		Taken in model (meters)		Clearance (meters)	
		Width	Length	Width	Length	Width	Length	Width	Length
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	2W	0.6	1.8	1	1	0.9	1.9	0.3	0.1
2	3W	1.4	2.6	2	2	1.8	3.8	0.4	1.2
3	Car	1.7	4.7	2	3	1.8	5.7	0.1	1
4	LCV1	1.9	5	3	3	2.7	5.7	0.8	0.7
5	LCV2	1.9	6.8	3	4	2.7	7.6	0.8	0.8
6	HCV1	2.5	8.5	3	5	2.7	9.5	0.2	1
7	HCV2	2.5	10.3	3	6	2.7	11.4	0.2	1.1

References

- Chalapati, M. V. (1987) Simulation of multi-lane unidirectional traffic, *Ph. D. Thesis*, Indian Institute of Technology, Kanpur.
- Goldberg, D. E. (1989) *Genetic algorithm in search optimization and machine learning*, Addison Wesley Publishing Company, Mass London.
- Gundaliya, P. J., Tom, V. Mathew, and Dhingra, S. L. (2004) "Heterogeneous traffic flow modelling using Cellular Automata", *World Transport Research –Proceedings from the 10th World conference on Transportation Research*, Istanbul, Turkey.

- Issac, K. P. (1995) Study of mixed traffic flow characteristic under varying composition, *Ph. D. Thesis*, Bangalore University, Bangalore.
- Khan, S. I. and Maini, P. (2000) "Modelling Heterogeneous Traffic Flow", *Transportation Research Record* 1678, pp.234-241.
- Korlapati, D. R. (2003) Evaluation of diversion strategies for an urban traffic corridor with heterogeneous traffic, *Ph. D. Thesis*, Indian Institute of Technology, Madras.
- Kumar, V. M. and Rao, S. K. (1996) "Simulation modelling of traffic operations on two-lane highways", *Indian Highways, Indian Roads Congress, New Delhi* 54, pp.211-237.
- Lawrence, W. Lan and Chang Chiung-Wen (2004) "In homogeneous Particles hopping models for mixed traffic with motorcycles and car", International conference on application of Information and communication technology in Transportation systems in developing countries, University of Maratuway, Srilanka.
- Marwah, B. R. and Bandyopadhyay, S. A. (1983) "Development of a traffic simulation model for an Indian city street", *Journal of Indian Roads Congress* 46(3), pp.655-695.
- Marwah, B. R. and Ramaseshan, S. (1978) "Development of a traffic simulation model for an Indian city street", *Journal of Indian Roads Congress, Highway research Bulletin* 8, pp.1-15.
- Nagel, K. and Schreckenberg. M. (1992) "Cellular automaton model for freeway traffic." *Journal de physique* 2(20), pp.2212-2229.
- Pillai, K. S. (1974) "Simulation techniques to study road traffic behaviour", *Indian Highways*, 29(2), pp.6-17.
- Ramanayya, T. V. (1988) "Highway capacity under mixed traffic conditions", *Traffic Engineering and Control* 29(5), pp.284-287.
- Roy, S. (2000) Effectiveness of auto rickshaw as a fixed route public transport mode – case study Calcutta metropolitan area, *Ph. D. Thesis*, Indian Institute of Technology, Khargapur.



Instructions to Authors

Papers should be written in English and submitted electronically to trasportieuropei@istiee.org and, by regular mail, in printed form in duplicate to:

prof. Romeo Danielis
Dipartimento di Scienze Economiche e Statistiche, Facoltà di Economia
Università degli Studi di Trieste. P.le Europa, 1, 34100 Trieste, Italy
Phone: +39-0405587076 - Fax: +39-040567543

Submission of a manuscript is considered to be a representation that it has been neither copyrighted (or if copyrighted is clearly marked so that the appropriate permission can be obtained) nor published, that it is not being submitted for publication elsewhere, and that, if the work results from a military contract, it has been released for open publication. As a condition of final acceptance of a paper for publication in European Transport (Trasporti Europei), the author(s) must indicate if their paper is posted on a working paper website, other than their own. They are responsible for assuring that, if any part of the paper has been copyrighted for prepublication as a working paper, the copyright can and will be transferred to ISTIEE when the paper has been accepted. This includes both print and electronic forms of the paper. On acceptance, the text, or any link to full text, must be removed from the working paper websites, other than the author's own website. Other material such as book reviews and announcements should also be sent to the Editor.

Manuscripts should contain only endnotes. Figures are required in a form suitable for photographic reproduction. Any one of a number of forms will be acceptable, e.g., laser printer drawing, original black ink drawings, or high-quality glossy prints. Lettering should be uniform in size and style and sufficiently large to be legible after reduction. Figures should be designated by arabic numbers and referred to in the text by number. Figure legends should be collectively provided on a separate sheet rather than placed on the figures themselves. Tables may be typed on sheets separated from the text. Each table should have a caption that makes the table entries clearly independent of the text; complicated column headings should be avoided. All tables should be numbered and referred to in the text by number.

In mathematical expressions, authors are requested in general to minimize unusual or expensive typographical requirements; for example: authors are requested to use the solidus wherever possible in preference to built-up fractions, to write complicated exponentials in the form $\exp()$ and to avoid subscripts and superscripts on subscripts or superscripts. Subscripts and superscripts should be shown large and clear, Greek letters and unusual symbols should be labeled on first occurrence, as should subscript "zero", to distinguish it from the letter "oh". Whether each letter is capital or lower case should be unambiguous. Equation numbers must be at the right.

Each paper must be accompanied by an abstract of about 100 words. The abstract should be adequate as an index and should summarize the principal results and conclusions. The first section of the article should not be numbered. References to related previous work should be reasonably complete, and grouped at the end of the paper. References in the text should be cited by the author's surname and the year of publication, e.g.: (Jansson, 1980), (Marguier and Ceder, 1984). The following format should be used for references:

Article in a journal:

David, P. and Bunn, J. (1988) "The Economics of Gateway Technologies and Network Evolution: Lessons from Electricity Supply History", *Information Economics and Policy* 3: 165-202.

Chapter in a book:

Regan, A. and Garrido, R. (2001) "Modelling Freight Demand and Shipper Behaviour: State of the Art, Future Directions", In: Hensher, D. (eds) *Travel Behaviour Research: The Leading Edge*, Pergamon, Amsterdam.

Working paper:

Gavish, B. and Graves, S.C. (1981) "Scheduling and Routing in Transportation and Distribution Systems: Formulations and New Relaxations", *Working Paper 8202*, Graduate School of Management, University of Rochester, Rochester, NY.

Book:

Urban, D. (1993) *Logit - Analyse. Statistische Verfahren zur Analyse von Modellen mit qualitativen Response-Variablen*, Gustav Fischer, Stuttgart.

Dissertation:

Jaillet, P. (1985) Probabilistic Traveling Salesman Problems, *Ph.D. thesis*, Massachusetts Institute of Technology, Cambridge, MA

Presentation at a conference:

Maggi, R. and Bolis, S. (1999) "Adaptive Stated Preference Analysis of Shippers' Transport and Logistics Choice", *World Transport Research - Proceedings from the 8th World Conference on Transport Research*, (H. Meersman, E. Van de Voorde, W. Winkelmanns eds.), Pergamon, Amsterdam.

Authors are responsible for revising their proofs, and should limit alterations to the strict minimum. The editorial management of ISTIEE reserves the right to accept only those changes that affect the accuracy of the text.