1 Transportation and geography

Movements of people, goods and information have always been fundamental components of human societies. Contemporary economic processes have been accompanied by a significant increase in mobility and higher levels of accessibility. Although this trend can be traced back to the industrial revolution, it significantly accelerated in the second half of the twentieth century as trade was liberalized, economic blocs emerged and the comparative advantages of global labor and resources were used more efficiently. However, these conditions are interdependent with the capacity to manage, support and expand movements of passengers and freight as well as their underlying information flows. Societies have become increasingly dependent on their transport systems to support a wide variety of activities ranging, among others, from commuting, supplying energy needs, to distributing parts between factories. Developing transport systems has been a continuous challenge to satisfy mobility needs, to support economic development and to participate in the global economy. The goal of this introductory chapter is to provide a definition of the nature, role and function of transport geography and where the discipline stands in regard to other disciplines. It also underlines the importance of specific dimensions such as nodes, locations, networks and interactions. A historical perspective on the evolution of transport systems underlines the consequences of technical innovations and how improvements in transportation were interdependent with contemporary economic and social changes.

Concept 1 – What is transport geography?

The purpose of transportation

> The ideal transport mode would be instantaneous, free, have an unlimited capacity and always be available. It would render space obsolete. This is obviously not the case. Space is a constraint for the construction of transport networks. Transportation appears to be an economic activity different from the others. It trades space with time and thus money.
>
> (translated from Merlin, 1992)

As the above quotation underlines, the purpose of transportation is to overcome space, which is shaped by a variety of human and physical constraints such as distance, time, administrative divisions and topography. Jointly, they confer a friction to any movement, commonly known as the friction of distance. However, these constraints and the friction they create can only be partially circumscribed. The extent to which this is done has a cost that varies greatly according to factors such as the distance involved and the nature of what is being transported. There would be no transportation without geography and there would be no geography without transportation. The goal of transportation is thus to transform the geographical attributes of freight, people or information, from an origin to a destination, conferring them an added value in the process. The convenience at which this can be done varies considerably.

Transportability. Refers to the ease of movement of passengers, freight or information. It is related to transport costs as well as to the attributes of what is being transported (fragility, perishability, price). Political factors can also influence transportability such as laws, regulations, borders and tariffs. When transportability is high, activities are less constrained by distance.

The specific purpose of transportation is to fulfill a demand for mobility, since transportation can only exist if it moves people, freight and information around. Otherwise it has no purpose. This is because transportation is the outcome of a derived demand (Figure 1.1).

What takes place in one sector has impacts on another; demand for a good or service in one sector is derived from another. For instance, a consumer buying a good in a store will likely trigger the replacement of this product, which will generate demands for activities such as manufacturing, resource extraction and, of course, transport. What is different about transport is that it cannot exist alone and a movement cannot be stored. An unsold product can remain on the shelf of a store until a customer buys it (often with discount incentives), but an unsold seat on a flight or unused cargo capacity in the same flight remains unsold and cannot be brought back as additional capacity later. In this case an opportunity has been missed since transport supply is higher than transport demand. The derived demand of transportation is often very difficult to reconcile with an equivalent supply. There are two major types of derived transport demand:

Direct derived demand. Refers to movements that are directly the outcome of economic activities, without which they would not take place. For instance, work-related activities commonly involve commuting between the place of residence and the workplace. There is a supply of work in one location (residence) and a demand of labor in another (workplace). For freight transportation, all the components of a supply chain require movements of raw materials, parts and finished products on modes such as trucks, rail or containerships.

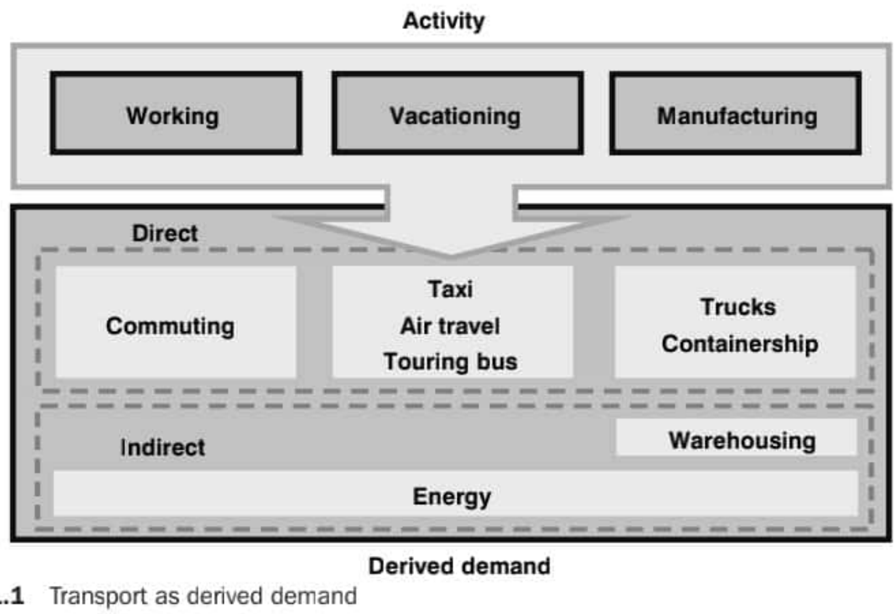


Figure 1.1 Transport as derived demand

Indirect derived demand. Considers movements created by the requirements of other movements. The most obvious example is energy where fuel consumption from transportation activities must be supplied by an energy production system requiring movements from zones of extraction to refineries and storage facilities and, finally, to places of consumption. Warehousing can also be labeled as an indirect derived demand since it is a "non-movement" of a freight element. Warehousing exists because it is virtually impossible to move commodities instantly from where they are produced to where they are consumed.

Consequently, the fundamental purpose of transport is geographic in nature, because it facilitates movements between different locations. Transport thus plays a role in the structure and organization of space and territories, which may vary according to the level of development. In the nineteenth century, the purpose of the emerging modern forms of transportation, mainly railways and maritime shipping, was to expand coverage, and create and consolidate national markets. In the twentieth century, the objective shifted to selecting itineraries, prioritizing transport modes, increasing the capacity of existing networks and responding to the mobility needs and this at a scale which was increasingly global. In the twenty-first century, transportation must cope with a globally oriented economic system in a timely and cost-effective way, but also with several local problems such as congestion.

The importance of transportation

Transport represents one of the most important human activities worldwide. It is an indispensable component of the economy and plays a major role in spatial relations between locations. Transport creates valuable links between regions and economic activities, between people and the rest of the world. Transport is a multidimensional activity whose importance is:

- **Historical**. Transport modes have played several different historical roles in the rise of civilizations (Egypt, Rome and China), in the development of societies (creation of social structures) and also in national defense (Roman Empire, American road network).
- **Social**. Transport modes facilitate access to healthcare, welfare, and cultural or artistic events, thus performing a social service. They shape social interactions by favoring or inhibiting the mobility of people. Transportation thus supports and may even shape social structures.
- **Political**. Governments play a critical role in transport as sources of investment and as regulators. The political role of transportation is undeniable as governments often subsidize the mobility of their populations (highways, public transit, etc.). While most transport demand relates to economic imperatives, many communication corridors have been constructed for political reasons such as national accessibility or job creation. Transport thus has an impact on nation building and national unity, but it is also a political tool.
- **Environmental**. Despite the manifest advantages of transport, its environmental consequences are also significant. They include air and water quality, noise level and public health. All decisions relating to transport need to be evaluated taking into account the corresponding environmental costs. Transport is a dominant factor in contemporary environmental issues.

- **Economic**. The evolution of transport has always been linked to economic development. The construction of transport infrastructures also permitted the development of a corresponding transport industry (car manufacturing, air transport companies, etc.). The transport sector is also an economic factor in the production of goods and services. It contributes to the value-added of economic activities, facilitates economies of scale, influences land (real estate) value and the geographic specialization of regions. Transport is a factor shaping economic activities, but is also shaped by them.

Substantial empirical evidence indicates that the importance of transportation is growing. The following contemporary trends can be identified regarding this issue:

- **Growth of the demand**. The twentieth century, more than any other, has seen a considerable growth of the transport demand related to individual (passengers) as well as freight mobility. This growth is jointly the result of larger quantities of passengers and freight being moved, but also the longer distances over which they are carried. Recent trends underline an ongoing process of mobility growth, which has led to the multiplication of the number of journeys involving a wide variety of modes that service transport demands.
- **Reduction of costs**. Even if several transportation modes are very expensive to own and operate (ships and planes for instance), costs per unit transported have dropped significantly over recent decades. This has made it possible to overcome larger distances and further exploit the comparative advantages of space. As a result, despite the lower costs, the share of transport activities in the economy has remained relatively constant in time.
- **Expansion of infrastructures**. The above two trends have obviously extended the requirements for transport infrastructures both quantitatively and qualitatively. Roads, harbors, airports, telecommunication facilities and pipelines have expanded considerably to service new areas and add capacity to existing networks. Transportation infrastructures are thus a major component of land use, notably in developed countries.

Facing these contemporary trends, an important part of the spatial differentiation of the economy is related to where resources (raw materials, capital, people, information, etc.) are located and how well they can be distributed. Transport routes are established to distribute resources between places where they are abundant and places where they are scarce, but only if the costs are lower than the benefits.

Consequently, transportation has an important role to play in the conditions that affect global, national and regional economic entities. It is a strategic infrastructure that is so embedded in the socio-economic life of individuals, institutions and corporations that it is often invisible to the consumer, but always part of all economic and social functions. This is paradoxical, since the perceived invisibility of transportation is derived from its efficiency. If transport is disrupted or ceases to operate, the consequences can be dramatic. The paradox gives rise to several fallacies about transportation; two major ones shown on Figure 1.2 are:

- **Access is not accessibility**. Many transport systems have universal access; no specific user can have a competitive advantage over others since access is the same for anyone. For instance, a public highway system can in theory be accessed by anyone, for example by a major trucking company having a large fleet, its competitors, or

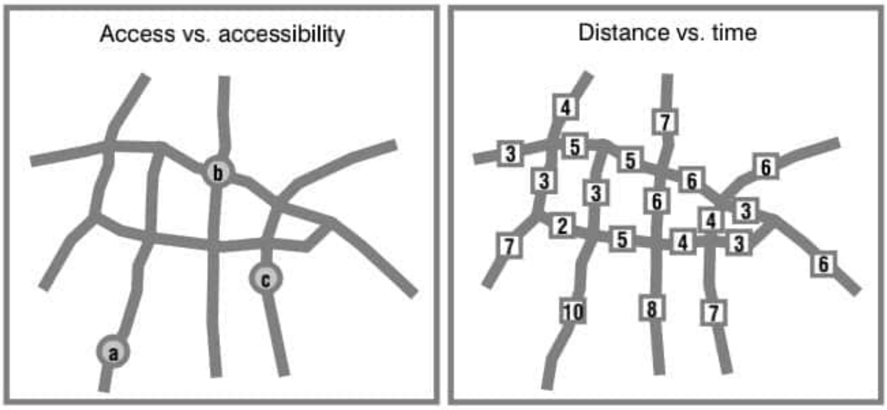


Figure 1.2 Two common fallacies in transport geography

by an individual driving an automobile. Thus, access is uniform wherever one is located in regard to the transport system as long as there is a possibility to enter or to exit. On the other hand, accessibility varies according to one's location within the transport system. Access is thus uniform while accessibility is not; the latter is a relative concept. On the transport network shown in Figure 1.2, locations a, b and c all have access to the system. However, location b appears to be more accessible than the other two due to its central location in relation to the network.

- **Distance is not time**. Distance often tends to be interchanged with time when measuring the performance of transport systems, which is a conceptual error. While distance remains constant, time can vary due to improvements in transport technology or because of congestion. Driving one kilometer through Manhattan is not the same as driving one kilometer through an Interstate in Iowa even if in both cases the same unit of distance has been traveled. Distance is thus a uniform attribute of the geography, while time is relative. On the above transport network shown in Figure 1.2, while distance is a uniform attribute, each segment has a travel time, which due to congestion, varies differently from distance.

Concepts and dimensions of transport geography

Transportation interests geographers for two main reasons. First transport infrastructures, terminals, equipment and networks occupy an important place in space and constitute the basis of a complex spatial system. Second, since geography seeks to explain spatial relationships, networks are of specific interest because they are the main support of these interactions.

Transport geography is a sub-discipline of geography concerned about movements of freight, people and information. It seeks to link spatial constraints and attributes with the origin, the destination, the extent, the nature and the purpose of movements.

Transport geography, as a discipline, emerged from economic geography in the second half of the twentieth century. Traditionally, transportation has been an important factor over the economic representations of geographic space, namely in terms of the location of economic activities and the monetary costs of distance. The growing mobility of passengers and freight justified the emergence of transport geography as a specialized field of investigation. In the 1960s, transport costs were recognized as key factors in location theories. However, from the 1970s globalization challenged the centrality of transportation in many geographical and regional development investigations. As a result, transportation became under-represented in economic geography in the 1970s and 1980s, even if mobility of people and freight and low transport costs were considered as important factors behind the globalization of trade and production.

There are twelve key concepts related to transport geography among which transportation networks, transportation nodes and transportation demand are at its core (Figure 1.3). They are closely linked to economic, political, regional, historical and population geography, among others. Several other concepts, such as regional planning, information systems, operations research and location theory are commonly used in transport geography, notably as tools and methods for the spatial analysis of transportation. At a wider level, links exist with several major fields of science including natural sciences, mathematics and economics. Indeed, like geography, transport geography is at the intersection of several concepts and methods initially developed outside the discipline that have been adapted to its particular interests and concerns.

Since the 1990s, transport geography has received renewed attention, especially because the issues of mobility, production and distribution are interrelated in a complex geographical setting. It is now recognized that transportation is a system that considers the complex relationships between its core elements: networks, nodes and demand (Figure 1.4). Demand for the movement of people, freight and information is a derived function of a variety of socio-economic activities. Nodes are the locations where movements are originating, ending and being transferred. The concept of nodes varies according to the geographical scale being considered, ranging from local to global (poles of the global economy). Networks are composed of a set of linkages derived from transport infrastructures. The three core relationships and the impedance (friction) they are subject to are:

- **Locations**. The level of spatial accumulation of socio-economic activities jointly defines demand and where this demand is taking place. Impedance is mostly a function of the accessibility of nodes to the demand they service.
- **Flows**. The amount of traffic over the network, which is jointly a function of the demand and the capacity of the linkages to support them. Flows are mainly subject to the friction of space with distance being the most significant impedance factor.

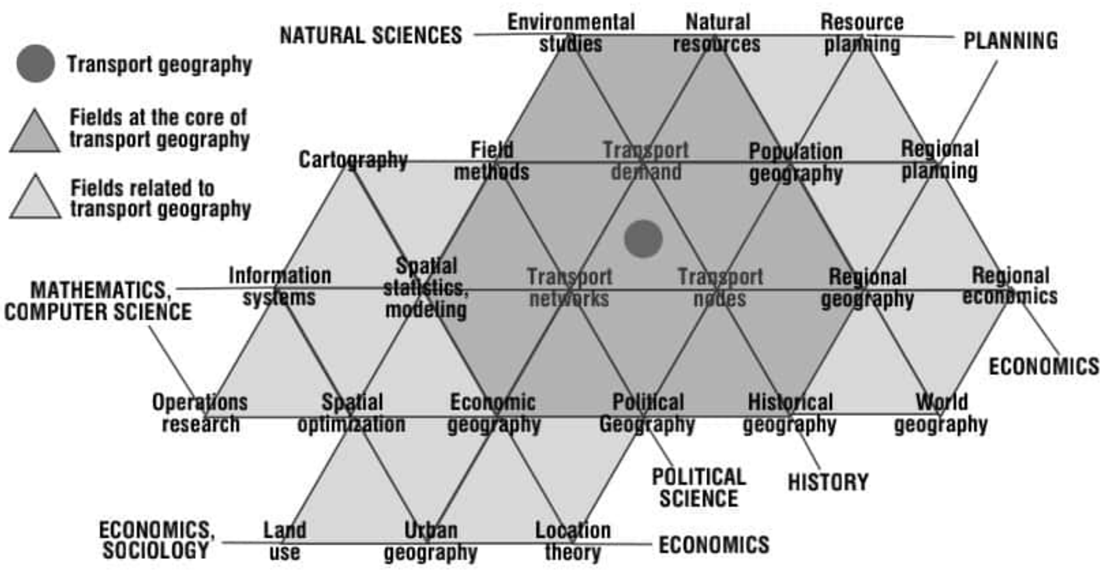


Figure 1.3 Fields of transport geography (Source: Haggett 2001)

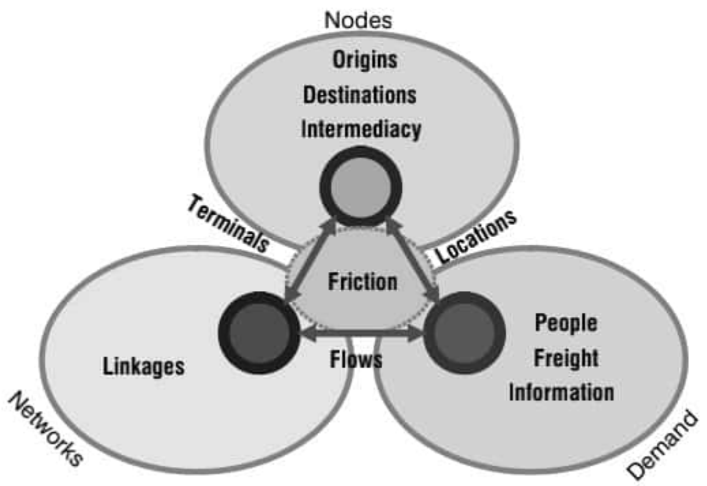


Figure 1.4 The transport system

- **Terminals**. The facilities enabling access to the network as terminals are jointly characterized by their nodality and the linkages that are radiated from them. The capacity of transport terminals to handle flows is the main impedance factor.

The analysis of these concepts relies on methodologies often developed by other disciplines such as economics, mathematics, planning and demography. For instance, the spatial structure of transportation networks can be analyzed with graph theory, which was initially developed for mathematics. Further, many models developed for the analysis of movements, such as the gravity model, were borrowed from physical sciences. Multidisciplinarity is consequently an important attribute of transport geography, as in geography in general.

The role of transport geography is to understand the spatial relations that are produced by transport systems. A better understanding of spatial relations is essential to assist private and public actors involved in transportation mitigate transport problems, such as capacity, transfer, reliability and integration of transport systems. There are three basic geographical considerations relevant to transport geography:

- **Location**. As all activities are located somewhere, each location has its own characteristics conferring a potential supply and/or a demand for resources, products, services or labor. A location will determine the nature, the origin, the destination, the distance and even the possibility of a movement to be realized. For instance, a city provides employment in various sectors of activity in addition to consuming resources.
- **Complementarity**. Locations must require exchanging goods, people or information. This implies that some locations have a surplus while others have a deficit. The only way an equilibrium can be reached is by movements between locations having surpluses and locations having demands. For instance, a complementarity is created between a store (surplus of goods) and its customers (demand of goods).
- **Scale**. Movements generated by complementarity are occurring at different scales, pending the nature of the activity. Scale illustrates how transportation systems are established over local, regional and global geographies. For instance, home-to-work journeys generally have a local or regional scale, while the distribution network of a multinational corporation is most likely to cover several regions of the world.

Consequently, transport systems, by their nature, consume land and support the relationships between locations.

Concept 2 – Transportation and space

Physical constraints

Transport geography is concerned with movements that take place over space. The physical features of this space impose major constraints on transportation systems, in terms of what mode can be used, the extent of the service, its costs and capacity. Three basic spatial constraints of the terrestrial space can be identified:

- **Topography**. Features such as mountains and valleys have strongly influenced the structure of networks, the cost and feasibility of transportation projects. The main land transport infrastructures are built usually where there are the least physical impediments, such as on plains, along valleys, or through mountain passes. Water transport is influenced by water depths and the location of obstacles such as reefs. Coastlines exert an influence on the location of port infrastructure. Aircraft require airfields of considerable size for takeoff and landing. Topography can impose a natural convergence of routes that will create a certain degree of centrality and may assist a location in becoming a trade center as a collector and distributor of goods. Topography can complicate, postpone or prevent the activities of the transport industry. Land transportation networks are notably influenced by the topography, as highways and railways tend to be impeded by grades higher than 3 percent and 1 percent respectively. Under such circumstances, land transportation tends to be of higher density in areas of limited topography.
- **Hydrography**. The properties, distribution and circulation of water play an important role in the transport industry. Maritime transport is influenced greatly by the availability of navigable channels through rivers, lakes and shallow seas. Several rivers such as the Mississippi, the St. Lawrence, the Rhine, the Mekong or the Yangtze are important navigable routeways into the heart of continents and historically have been the focus of human activities that have taken advantage of the transport opportunities. Port sites are also highly influenced by the physical attributes of the site where natural features (bays, sand dunes, and fjords) protect port installations. Since it is at these installations that traffic is transshipped, the location of ports is a dominant element in the structure of maritime networks. Where barriers exist, such as narrows, rapids, or land breaks, water transport can only overcome these obstacles with heavy investments in canals or dredging. Conversely, waterways serve as barriers to land transportation necessitating the construction of bridges, tunnels and detours, etc.
- **Climate**. Its major components include temperature, wind and precipitation. Their impacts on transportation modes and infrastructure range from negligible to severe. Freight and passenger movement can be seriously curtailed by hazardous conditions such as snow, heavy rainfall, ice or fog. Jet streams are also a major physical component that international air carriers must take into consideration. For an aircraft, the speed of wind can affect travel costs. When the wind is pushing the airplane towards its destination, it can reduce flight time by up to several hours for intercontinental flights. Climate also affects transportation networks by influencing construction and maintenance costs.

Physical constraints fundamentally act as absolute and relative barriers to movements (Figure 1.5):

- **Absolute barriers** are geographical features that entirely prevent a movement. They must either be bypassed or be overcome by specific infrastructures. For instance, a river is considered as an absolute barrier for land transportation and can only be overcome if a tunnel or a bridge is constructed. A body of water forms a similar absolute barrier and could be overcome if ports are built and a maritime service (ferry, cargo ships, etc.) is established. Conversely, land acts as an absolute barrier for maritime transportation, with discontinuities (barriers) that can be overcome with costly infrastructures such as navigation channels and canals.
- **Relative barriers** are geographical features that force a degree of friction on a movement. In turn, this friction is likely to influence the path (route) selected to link two locations (A and B on Figure 1.5). Topography is a classic example of a relative barrier that influences land transportation routes along paths having the least possible friction (e.g. plains and valleys). For maritime transportation, relative barriers, such as straits, channels or ice, generally slow down circulation.

From a geometrical standpoint, the sphericity of the Earth determines the great circle distance. This feature explains the paths followed by major intercontinental maritime and air routes (Figure 1.6). Since the Earth is a sphere, the shortest path between two points is calculated by the great circle distance, which corresponds to an arc linking two points on a sphere. The circumference inferred out of these two points divides the Earth in two equal parts, thus the great circle. The great circle distance is useful to establish the shortest path to use when traveling at the intercontinental air and maritime level. The great circle route follows the sphericity of the globe; any shortest route is the one following the curve of the planet, along the parallels.

Because of the distortions caused by projections of the globe on a flat sheet of paper, a straight line on a map is not necessarily the shortest distance. Ships and aircraft usually follow the great circle geometry to minimize distance and save time and money to customers. For instance, Figure 1.6 shows the shortest path between New York and Moscow (about 7,540 km). This path corresponds to an air transportation corridor. Air travel over the North Atlantic between North America and Europe follows a similar

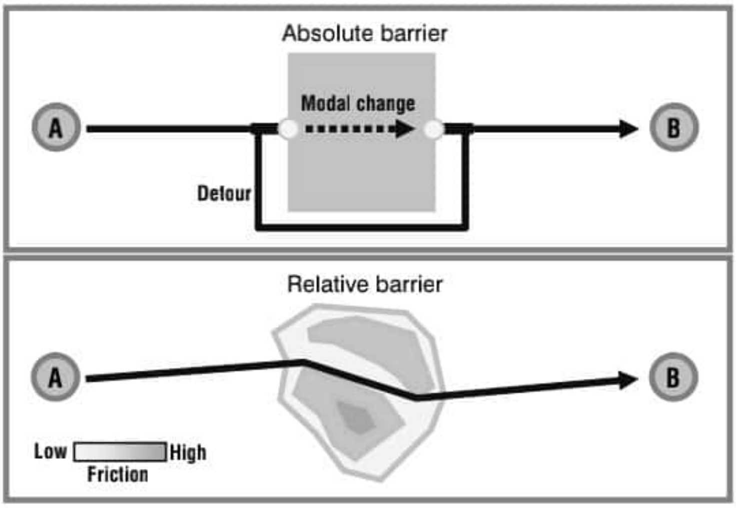


Figure 1.5 Absolute and relative barriers

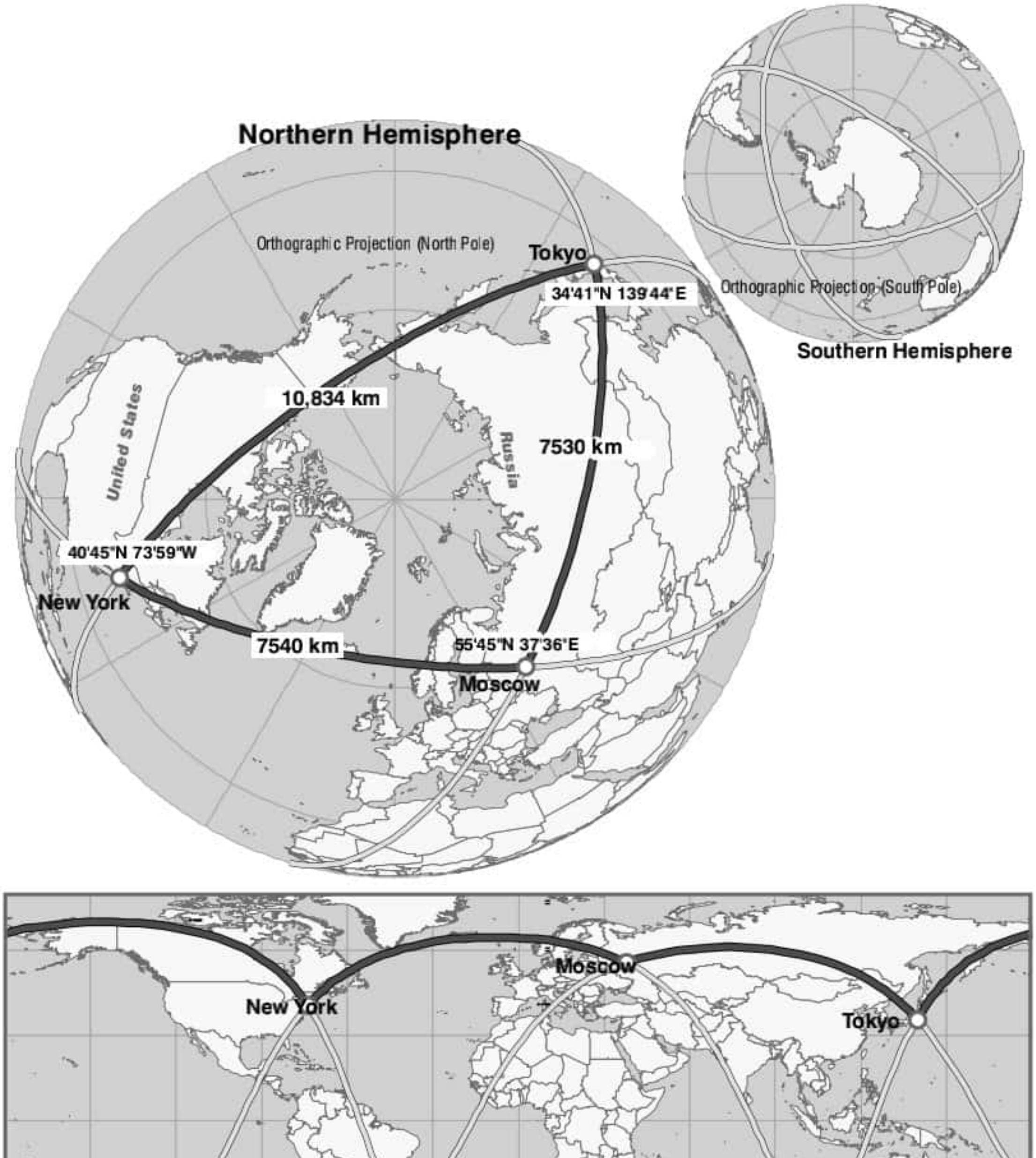


Figure 1.6 The great circle distance

path. To calculate the great circle distance (D) between two coordinates the following formula is used: $\cos(D) = (\sin a \sin b) + (\cos a \cos b \cos |c|)$, where a and b are the latitudes (in degrees) of the respective coordinates and $|c|$ is the absolute value of the difference of longitude between the respective coordinates. The results of this equation are in degrees. Each degree on the Earth's surface equals about 111.32 km, so the result must be multiplied by this number.

Transportation and the spatial structure

All locations are relative to one another. However, locations are not constant as transportation developments change levels of accessibility, and thus the relations between locations. The development of a location reflects the cumulative relationships between transport infrastructure, economic activities and the built-environment. The following factors are particularly important in shaping the spatial structure:

- **Costs**. The spatial distribution of activities is related to factors of distance, namely its friction. Locational decisions are taken in an attempt to minimize costs, often related to transportation.
- **Accessibility**. All locations have a level of accessibility, but some are more accessible than others. Thus, because of transportation, some locations are perceived as more valuable than others.
- **Agglomeration**. There is a tendency for activities to agglomerate to take advantage of the value of specific locations. The more valuable a location, the more likely agglomeration will take place. The organization of activities is essentially hierarchical, resulting from the relationships between agglomeration and accessibility at the local, regional and global levels.

Many contemporary transportation networks are inherited from the past, notably transport infrastructures. Even if over the last 200 years new technologies have revolutionized transportation in terms of speed, capacity and efficiency, the spatial structure of many networks has not much changed. This inertia in the spatial structure of some transportation networks can be explained by two major factors:

- **Physical attributes**. Natural conditions can be modified and adapted to suit human uses, but they are a very difficult constraint to escape, notably for land transportation. It is thus not surprising to find that most networks follow the easiest (least cost) paths, which generally follow valleys and plains. Considerations that affected road construction a few hundred years ago are still in force today, although they are sometimes easier to circumscribe.
- **Historical considerations**. New infrastructures generally reinforce historical patterns of exchange, notably at the regional level. For instance, the current highway network of France has mainly followed the patterns set by the national roads network built early in the twentieth century. This network was established over the Royal roads network, itself mainly following roads built by the Romans. At the urban level, the pattern of streets is often inherited from an older pattern, which itself may have been influenced by the pre-existing rural structure (lot pattern and rural roads).

While inertia is important in transport networks, the introduction of new transport technologies or the addition of new transport infrastructures are leading to a transformation of existing networks. Recent developments in transport systems such as container shipping, jumbo aircraft and the extensive application of information technology to transport management are creating a new transport environment and a new spatial structure. These transport infrastructures have intensified global interactions and modified the relative location of places. In this highly dynamic context, two processes are taking place at the same time:

- **Specialization**. Linked geographical entities are able to specialize in the production of commodities for which they have an advantage, and trading for what they do not produce. As a result, efficient transportation systems are generally linked with higher levels of regional specialization. The globalization of production clearly underlines this process as specialization occurs as long as the incurred savings in production costs are higher than the incurred additional transport costs.
- **Segregation**. Linked geographical entities may see the reinforcement of one at the expense of others, notably through economies of scale. This outcome often contradicts regional development policies aiming at providing uniform accessibility levels within a region.

The continuous evolution of transportation technology may not necessarily have expected effects on the spatial structure, as two forces are at play: **concentration** and **dispersion**. A common myth tends to relate transportation solely as a force of dispersion, favoring the spread of activities in space. This is not always the case. In numerous instances, transportation is a force of concentration, notably for business activities. Since transport infrastructures are generally expensive to build, they are established first to service the most important locations. Even if it was a strong factor of dispersion, the automobile has also favored the concentration of several activities at specific places and in large volumes. Shopping centers are a relevant example of this process where central locations emerge in a dispersed setting.

Space/time relationships

One of the most basic relationships of transportation involves how much space can be overcome within a given amount of time. The faster the mode, the larger the distance that can be overcome within the same amount of time. Transportation, notably improvements in transport systems, changes the relationship between time and space. When this relationship involves easier, faster and cheaper access between places, this result is defined as a space/time convergence because the amount of space that can be overcome for a similar amount of time increases significantly. Significant regional and continental gains were achieved during the eighteenth and nineteenth centuries with the establishment of national and continental railway systems as well as with the growth of maritime shipping, a process which continued into the twentieth century with air and road transport systems. The outcome has been significant differences in space/time relationships, mainly between developed and developing countries, reflecting differences in the efficiency of transport systems.

At the international level, globalization processes have been supported by improvements in transport technology. The result of more than 200 years of technological improvements has been a space/time collapse of global proportions in addition to the regional and continental processes previously mentioned. This enabled the extended exploitation of the advantages of the global market, notably in terms of resources and labor. Significant reductions in transport and communication costs occurred concomitantly. There is thus a relationship between the rate of a space/time collapse on the integration of a region in global trade. Four major factors are of particular relevance in this process:

- **Speed**. The most straightforward factor relates to the increasing speed of many transport modes, a condition that particularly prevailed in the first half of the twentieth century. More recently, speed has played a less significant role as many modes are not going much faster. For instance, an automobile has a similar operating speed today than it had 60 years ago and a commercial jet plane operates at a similar speed than one 30 years ago.
- **Economies of scale**. Being able to transport larger amounts of freight and passengers at lower costs has improved considerably the capacity and efficiency of transport systems.
- **Expansion of transport infrastructures**. Transport infrastructures have expanded considerably to service areas that were not previously serviced or were insufficiently serviced. A paradox of this feature is that although the expansion of transport infrastructures may have enabled distribution systems to expand, it has also increased the average distance over which passengers and freight are being carried.
- **Efficiency of transport terminals**. Terminals, such as ports and airports, have shown a growing capacity to handle large quantities of traffic over a short time period in a timely manner. Thus, even if the speed of many transport modes has not increased, more efficient transport terminals may have helped reduce transport time.

The space/time convergence process investigates the changing relationship between space and time, and notably the impacts of transportation improvements on such a relationship. It is closely related to the concept of speed, which indicates how much space can be traveled over a specific amount of time (Figure 1.7). To measure space/time convergence (STC), travel time information is required for at least two locations and two time periods. Variation in travel time (ΔTT) is simply divided by the time period (ΔT) over which the process took place. Figure 1.7 provides an example of space/time convergence between two locations, A and B. In 1950, it took 6.2 hours to travel between A and B. By 2000, this travel time was reduced to 2.6 hours. Consequently, STC is –0.072 hours per year, or –4.32 minutes per year. The value is negative because the time value is being reduced; if the value was positive, a space/time divergence would be observed.

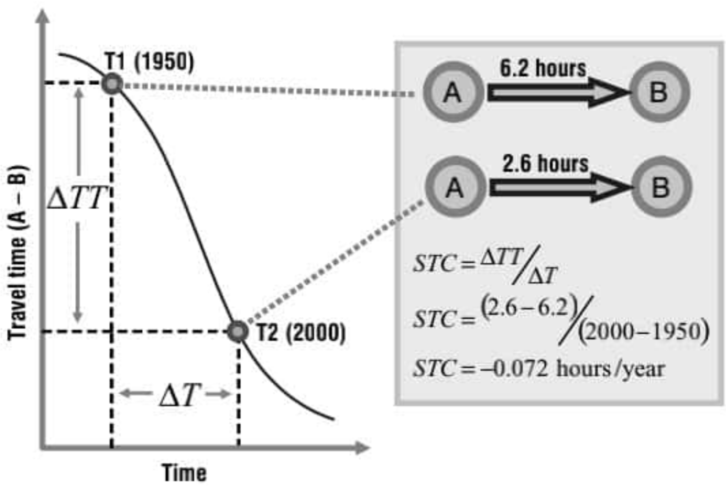


Figure 1.7 Space/time convergence

However, space/time convergence can also be inverted under specific circumstances. For instance, congestion is increasing in many metropolitan areas, implying additional delays for activities such as commuting. Traffic in congested urban areas is moving at the same speed that it did 100 years ago on horse carriages. Air transportation, despite having dramatically contributed to the space/time collapse convergence is also experiencing growing delays. Flight times are getting longer between many destinations, mainly because of takeoff, landing and gate access delays. Airlines are simply posting longer flight times to factor in congestion. An express mail package flown from Washington to Boston in about an hour (excluding delays at takeoff and landing due to airport congestion) can have an extra one hour delay as it is carried from Logan Airport to downtown Boston, a distance of only two miles. The "last mile" can be the longest in many transport segments.

Concept 3 – Historical evolution of transportation

Transportation in the pre-industrial era (pre-1800s)

Efficiently distributing freight and moving people has always been an important factor for maintaining the cohesion of economic systems from empires to modern nation states. With technological and economic developments, the means to achieve such a goal have evolved considerably. The historical evolution of transportation is very complex and is related to the spatial evolution of economic systems. It is possible to summarize this evolution, from the pre-industrial era to transportation in the early twenty-first century, in five major stages, each linked with specific technological innovations in the transport sector.

Before the major technical transformations brought forward by the industrial revolution at the end of the eighteenth century, no forms of motorized transportation existed. Transport technology was mainly limited to harnessing animal labor for land transport and to wind for maritime transport. The transported quantities were very limited and so was the speed at which people and freight were moving. The average overland speed by horse was between 8 and 15 kilometers per hour and maritime speeds were barely above these figures. Waterways were the most efficient transport systems available and cities next to rivers were able to trade over longer distances and maintain political, economic and cultural cohesion over a larger territory. It is not surprising to find that the first civilizations emerged along river systems for agricultural but also for trading purposes (Tigris–Euphrates, Nile, Indus, Ganges, Huang He).

Because the efficiency of the land transport system of this era was poor, the overwhelming majority of trade was local in scope. From the perspective of regional economic organization, the provision of cities in perishable agricultural commodities was limited to a radius of about 50 kilometers, at most. The size of cities also remained constant in time. Since people can walk about 5 km per hour and they are not willing to spend more than one hour per day walking, the daily space of interaction would be constrained by a 2.5 km radius, or about 20 square kilometers. Thus, most rural areas centered around a village and cities rarely exceeded a 5 km diameter. The largest cities prior to the industrial revolution, such as Rome, Beijing, Constantinople, or Venice never surpassed an area of 20 square kilometers. International trade did exist, but traded commodities were high-value (luxury) goods such as spices, silk, wine and perfume, notably along the Silk Road (see Figure 1.8).

The Silk Road was the most enduring trade route of human history, being used for about 1,500 years. Its name is taken from the prized Chinese textile that flowed

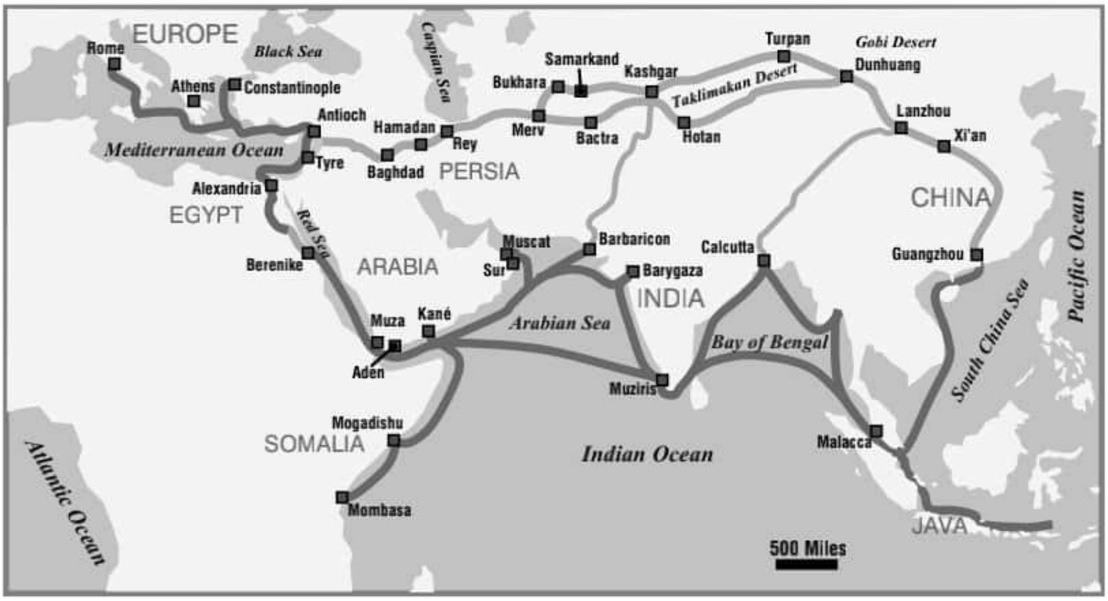


Figure 1.8 The Silk Road and the Arab sea routes

from Asia to the Middle East and Europe. The Silk Road consisted of a succession of trails followed by caravans through Central Asia, about 6,400 km in length. Travel was favored by the presence of steppes, although several arid zones had to be bypassed such as the Gobi and Takla Makan deserts. Economies of scale, harsh conditions and security considerations required the organization of trade into caravans slowly trekking from one stage (town and/or oasis) to the other.

Although it is suspected that significant trade occurred for about 1,000 years beforehand, the Silk Road opened around 139 BC once China was unified under the Han dynasty. It started at Changan (Xian) and ended at Antioch or Constantinople (Istanbul), passing by commercial cities such as Samarkand and Kashgar. It was very rare that caravans traveled for the whole distance since the trade system functioned as a chain. Merchants with their caravans were shipping goods back and forth from one trade center to the other.

The initial use of the sea route linking the Mediterranean basin and India took place during the Roman Era. Between the first and sixth centuries, ships were sailing between the Red Sea and India, aided by summer monsoon winds. Goods were transshipped at the town of Berenike along the Red Sea and moved by camels inland to the Nile. From that point, river boats moved the goods to Alexandria, from which trade could be undertaken with the Roman Empire. From the ninth century, maritime routes controlled by the Arab traders emerged and gradually undermined the importance of the Silk Road. Since ships were much less constraining than caravans in terms of capacity, larger quantities of goods could be traded. The main maritime route started at Canton (Guangzhou), passed through Southeast Asia, the Indian Ocean, the Red Sea and then reached Alexandria. A significant feeder went to the Spice Islands (Mollusks) in today's Indonesia. The diffusion of Islam was also favored through trade as many rules of ethics and commerce are embedded in the religion.

During the Middle Ages, the Venetians controlled the bulk of the Mediterranean trade which connected to the major trading centers of Constantinople, Antioch and Alexandria. As European powers developed their maritime technologies from the fifteenth century, they successfully overthrew the Arab control of this lucrative trade route to replace it by their own. Ships being able to transport commodities faster and cheaper marked the downfall of the Silk Road by the sixteenth century.

The transport system of the Roman Empire was a reflection of the geographical characteristics and constraints of the Mediterranean basin (Figure 1.9). The Mediterranean Ocean provided a central role to support trade between a network of coastal cities, the most important of the Empire (Rome, Constantinople, Alexandria, Carthage, etc.). These cities were serviced by a road network permitting trade within their respective hinterlands. Little fluvial transportation took place since the major pan-European rivers, the Rhine and the Danube, were military frontiers, not the core, of the Empire. The roads served numerous functions, such as military movements, political control, cultural and economic (trade).

Under such conditions, it was difficult to speak of an urban system, but rather of a set of relatively self-sufficient economic systems with very limited trade. The preponderance of city-states during this period can a priori be explained by transportation, in particular the difficulties of shipping goods (therefore to trade) from one place to another. Among the most notable exceptions to this were the Roman and Chinese empires, which committed extraordinary efforts to building transportation networks and consequently maintained control over an extensive territory for a long time period.

The economic importance and the geopolitics of transportation were recognized very early, notably for maritime transportation, since before the industrial revolution it was the most convenient way to move freight and passengers around. Great commercial empires were established with maritime transportation. Initially, ships were propelled by rowers and sails were added around 2500 BC as a complementary form of propulsion. By Medieval times, an extensive maritime trade network, the highways of the time,

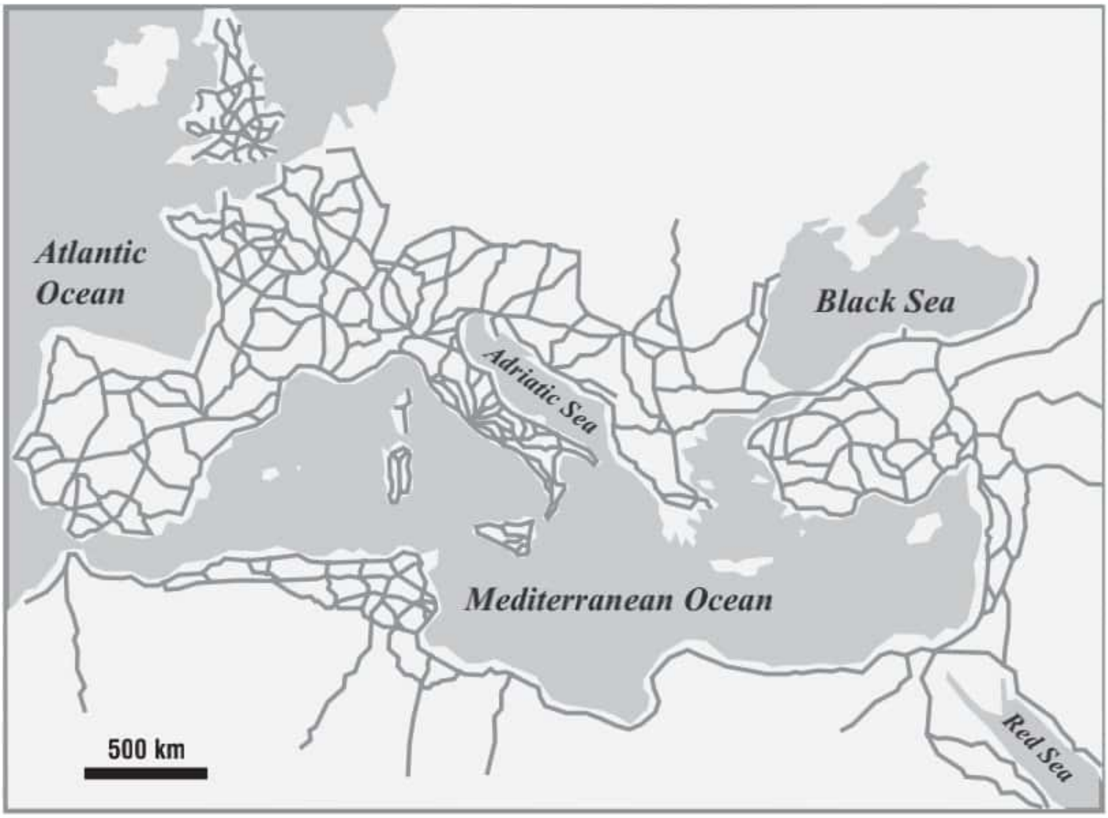


Figure 1.9 Roman road network, 200 AD

centered along the navigable rivers, canals, and coastal waters of Europe (and also China) was established. Shipping was extensive and sophisticated using the English Channel, the North Sea, the Baltic and the Mediterranean where the most important cities were coastal or inland ports (London, Norwich, Königsberg, Hamburg, Bruges, Bordeaux, Lyon, Lisbon, Barcelona, and Venice). Trade of bulk goods, such as grain, salt, wine, wool, timber, and stone was taking place. By the fourteenth century galleys were finally replaced by fully fledged sailships (the caravel and then the galleon) that were faster and required smaller crews. The year 1431 marked the beginning of European expansion with the discovery by the Portuguese of the North Atlantic circular wind pattern, better known as the trade winds. A similar pattern was also found on the Indian and Pacific oceans with the monsoon winds.

The fall of Constantinople, the capital of the Byzantium Empire (Eastern Roman Empire), to the Turks in 1453 disrupted the traditional land trade route from Europe to Asia. Europe was forced to find alternate maritime routes. One alternative, followed by Columbus in 1492, was to sail to the west and the other alternative, followed by Vasco daGama in 1497, was to sail to the east. Columbus stumbled upon the American continent, while Gama found a maritime route to India using the Cape of Good Hope. These events were quickly followed by a wave of European exploration and colonization, initially by Spain and Portugal, the early maritime powers, then by Britain, France and the Netherlands. The traditional trade route to Asia no longer involved Italy (Venice) and Arabia, but involved direct maritime connections from ports such as Lisbon. European powers were able to master the seas with larger, better armed and more efficient sailing ships and thus were able to control international trade and colonization. By the early eighteenth century, most of the world's territories were controlled by Europe, providing wealth and markets to their thriving metropolises through a system of colonial trade (Figure 1.10).

By the early eighteenth century, a complex network of colonial trade was established over the North Atlantic Ocean. This network was partially the result of local conditions

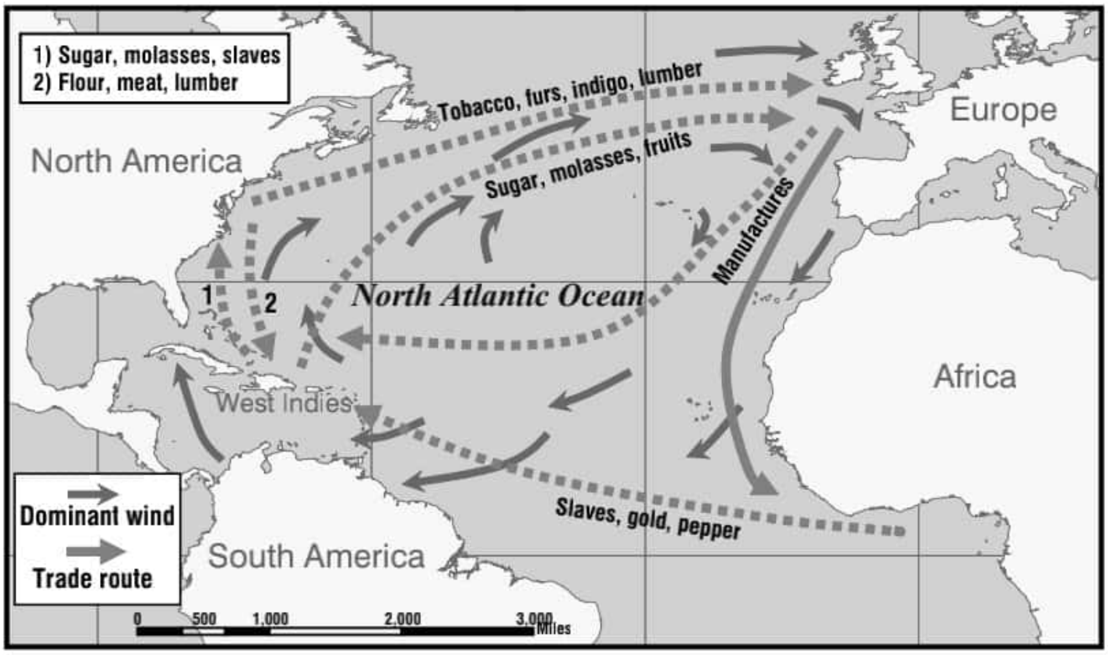


Figure 1.10 Colonial trade pattern, North Atlantic, eighteenth century

and of dominant wind patterns. It was discovered in the fifteenth century, notably after the voyages of Columbus, that there is a circular wind pattern over the North Atlantic. The eastward wind pattern, which blows on the southern part, came to be known as the "trade winds" since they enabled ships to cross the Atlantic. The westward wind pattern, blowing on the northern part, came to be known as the "westerlies".

Since sailing ships were highly constrained by dominant wind patterns, a trade system followed this pattern. Manufactured commodities were exported from Europe, some towards the African colonial centers, some towards the American colonies. This system also included the slave trade, mainly to Central and South American colonies (Brazil, West Indies). Tropical commodities (sugar, molasses) flowed to the American colonies and to Europe. North America also exported tobacco, furs, indigo (a dye) and lumber (for shipbuilding) to Europe. This system of trade collapsed in the nineteenth century with the introduction of steamships, the end of slavery and the independence of many of the colonies of the Americas.

Prior to the industrial revolution, the quantity of freight transported between nations was negligible by contemporary standards. For instance, during the Middle Ages, French imports via the Saint-Gothard Passage (between Italy and Switzerland) would not fill a freight train. The total amount of freight transported by the Venetian fleet, which dominated Mediterranean trade for centuries, would not fill a modern cargo ship. The volume, but not the speed, of trade improved under mercantilism (fifteenth to eighteenth centuries), notably for maritime transportation. In spite of all, distribution capacities were very limited and speeds slow. For example, a stagecoach going through the English countryside in the sixteenth century had an average speed of two miles per hour; moving one ton of cargo 30 miles (50 km) inland in the United States by the late eighteenth century was as costly as moving it across the Atlantic. The inland transportation system was thus very limited, both for passengers and freight. By the late eighteenth century, canal systems started to emerge in Europe, initially in the Netherlands and England. They permitted the beginning of large movements of bulk freight inland and expanded regional trade. Maritime and fluvial transportation were consequently the dominant modes of the pre-industrial era.

The industrial revolution and transportation (1800–70)

It was during the industrial revolution that massive modifications of transport systems occurred in two major phases, the first centered along the development of canal systems and the second centered along railways. This period marked the development of the steam engine that converted thermal energy into mechanical energy, providing an important territorial expansion for maritime and railway transport systems. Much of the credit of developing the first efficient steam engine in 1765 is attributed to the British engineer Watt, although the first steam engines were used to pump water out of mines. It was then only a matter of time before the adaptation of the steam engine to locomotion. In 1769, the French engineer Cugnot built the first self-propelled steam vehicle, along with being responsible for the first automobile accident ever recorded. The first mechanically propelled maritime vehicle was tested in 1790 by the American inventor Fitch as a mode of fluvial transportation on the Delaware River. This marked a new era in the mechanization of land and maritime transport systems alike.

From the perspective of land transportation, the early industrial revolution faced problems over bottlenecks, as inland distribution was unable to carry the growing quantities of raw materials and finished goods. Roads were commonly unpaved and could not be used to effectively carry heavy loads. The first Turnpike Trust was established in 1706. Each Trust was responsible to construct and maintain a specific road segment, which required capital. Capital was publicly raised and revenues were generated by charging tolls on users. This came as a somewhat unwelcome change as road users were used to freely make use of any public roads. Some would even jump over toll gates to avoid paying the fare. Spikes (or pikes) were installed on top of toll gates to prevent this, thus the name turnpike. The most potentially profitable roads became Trusts, which at their peak never accounted for more than 20 percent of Britain's road network. Turnpike Trusts were a success and improved land circulation substantially. Figure 1.11 depicts this evolution in rather typical phases of introduction, fast growth, maturity and then obsolescence. Between 1750 and 1800, the average time for a journey from London to Edinburgh was reduced from 12 to 4 days. Also, the time of a journey from Manchester to London fell from 3 days in 1760 to 28 hours in 1788. Road freight transportation also improved due to the introduction in the 1760s of "flywagons": a system of freight distribution involving changing horses and crews at specific stages and thus permitting day-long movements. By 1770, there were 25,000 km of turnpike roads in England and most of the country was within 12.5 miles of one. The turnpike system reached a peak of 32,500 km by 1836, but by then rail transportation started to emerge, which marked the downfall of turnpikes.

Although improvements were made to road transport systems in the early seventeenth century this was not sufficient to accommodate the growing demands on freight transportation. From the 1760s a set of freight shipping canals were slowly built in emerging industrial cores such as England (e.g. Bridgewater Canal, 1761) and the United States (e.g. Erie Canal, 1825). These projects relied on a system of locks to overcome changes in elevation, thus linking different segments of fluvial systems into a comprehensive waterway system. Barges became increasingly used to move goods at a scale and a cost that were not previously possible. Economies of scale and specialization, the foundation of modern industrial production systems, became increasingly applicable through fluvial canals. Physical obstacles made canal construction expensive, however, and the network was constrained. In 1830 there were about 2,000 miles of canals in Britain and by the end of the canal era in 1850, there were 4,250 miles of navigable waterways. The canal era was however short-lived as a new mode which would revolutionize and transform inland transportation appeared in the second half of the nineteenth century.

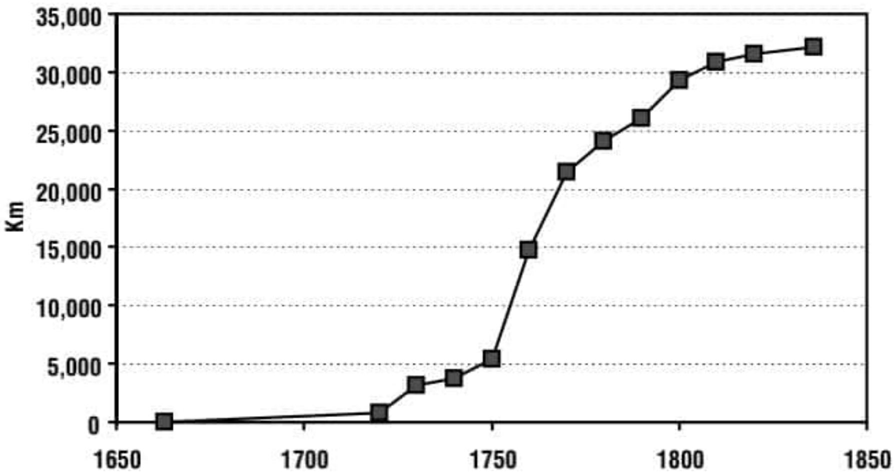


Figure 1.11 Turnpikes in Great Britain, late 18th and early 19th century (Source: adapted from D. Bogart 2004)

Steam railway technology initially appeared in 1814 to haul coal. It was found that using a steam engine on smooth rails required less power and could handle heavier loads. The first commercial rail line linked Manchester to Liverpool in 1830 (a distance of 40 miles) and shortly after rail lines began to be laid throughout developed countries. By the 1850s, railroad towns were being established and the railways were giving access to resources and markets of vast territories. Six thousand miles of railways were then operating in England and railways were quickly being constructed in Western Europe and North America. Railroads represented an inland transport system that was flexible in its spatial coverage and could carry heavy loads. As a result many canals fell into disrepair and were closed as they were no longer able to compete with rail services. In their initial phase of development, railways were a point-to-point process where major cities were linked one at a time by independent companies. Thus, the first railroad companies bore the name of the city pairs or the region they were servicing (e.g. the Camden and Amboy Railroad Company, chartered in 1830). From the 1860s, integrated railway systems started to cohesively service whole nations with standard gauges and passenger and freight services. The journey between New York and Chicago was reduced from three weeks by stagecoach to 72 hours by train. Many cities thus became closely interconnected. The transcontinental line between New York and San Francisco, completed in 1869, represented a remarkable achievement in territorial integration made only possible by rail. It reduced the journey across the continent (New York to San Francisco) from six months to one week, thus opening for the Eastern part of the United States a vast pool of resources and new agricultural regions. This was followed by Canada in 1886 (trans-Canada railway) and Russia in 1904 (trans-Siberian railway).

In terms of international transportation, the beginning of the nineteenth century saw the establishment of the first regular maritime routes linking harbors worldwide, especially over the North Atlantic between Europe and North America. These routes were navigated by fast clipper ships, which dominated ocean trade until the late 1850s. Another significant improvement resided in the elaboration of accurate navigation charts where prevailing winds and sea current could be used to the advantage of navigation. Composite ships (a mixture of wood and iron armature) then took over a large portion of the trade until about 1900, but they could not compete with steamships which had been continually improved since they were first introduced 100 years before. Regarding steamship technology, 1807 marks the first successful use of a steamship, Fulton's North River / Clermont, on the Hudson servicing New York and Albany. The gradual improvement of steam engine technology slowly but surely permitted longer and safer voyages. In 1820, the Savannah was the first steamship (used as auxiliary power) to cross the Atlantic, taking 29 days to link Liverpool to New York. The first regular services for transatlantic passenger transport by steamships was inaugurated in 1838, followed closely by the usage of the helix, instead of the paddle wheel, as a more efficient propeller (1840). Shipbuilding was also revolutionized by the usage of steel armatures (1860), enabling to escape the structural constraints of wood and iron armatures in terms of ship size. Steel armature ships were 30 to 40 percent lighter and had 15 percent more cargo capacity.

The main consequence of the industrial revolution was a specialization of transportation services and the establishment of large distribution networks of raw materials and energy.

Emergence of modern transportation systems (1870–1920)

By the end of the nineteenth century, international transportation undertook a new growth phase, especially with improvements in engine propulsion technology and a gradual shift from coal to oil in the 1870s. Although oil has been known for centuries for its combustion properties, its commercial use was only applied in the early nineteenth century. Inventors started experimenting with engines that could use the cheap new fuel. Oil increased the speed and the capacity of maritime transport. It also permitted to reduce the energy consumption of ships by a factor of 90 percent relative to coal, the main source of energy for steam engines prior to this innovation. An equal size oil-powered ship could transport more freight than a coal-powered ship, reducing operation costs considerably and extending range. Also, coal refueling stages along trade routes could be bypassed. Global maritime circulation was also dramatically improved when infrastructures to reduce intercontinental distances, such as the Suez (1869) and the Panama (1914) canals, were constructed. With the Suez Canal, the far reaches of Asia and Australia became more accessible (Figure 1.12).

The Panama Canal, completed in 1914, considerably shortens the maritime distances between the American East and West coasts by a factor of 13,000 km. Planned by the French but constructed by the British, the Suez Canal opened in 1869. It represents, along with the Panama Canal, one of the most significant maritime "shortcuts" ever built. It brought a new era of European influence to Pacific Asia by reducing the journey from Asia to Europe by about 6,000 km. The region became commercially accessible and colonial trade expanded as a result of increased interactions because of a reduced friction of distance. Great Britain, the maritime power of the time, benefited substantially from this improved access. For instance, the Suez Canal shortened the distance of a maritime journey from London to Bombay by 41 percent and shortened the distance of the journey from London to Shanghai by 32 percent.

The increasing size of ships, the outcome of advances in shipbuilding, imposed massive investments in port infrastructures such as piers and docks to accommodate them. Ship size grew dramatically, from the largest tonnage of 3,800 gross registered tons (revenue making cargo space) in 1871 to 47,000 tons in 1914. The harbor, while integrating production and transshipping activities, became an industrial complex around which agglomerated activities using ponderous raw materials. From the 1880s, liner services linked major ports of the world, supporting the first regular international passenger transport services, until the 1950s when air transportation became the dominant mode. This period also marked the golden era of the development of the

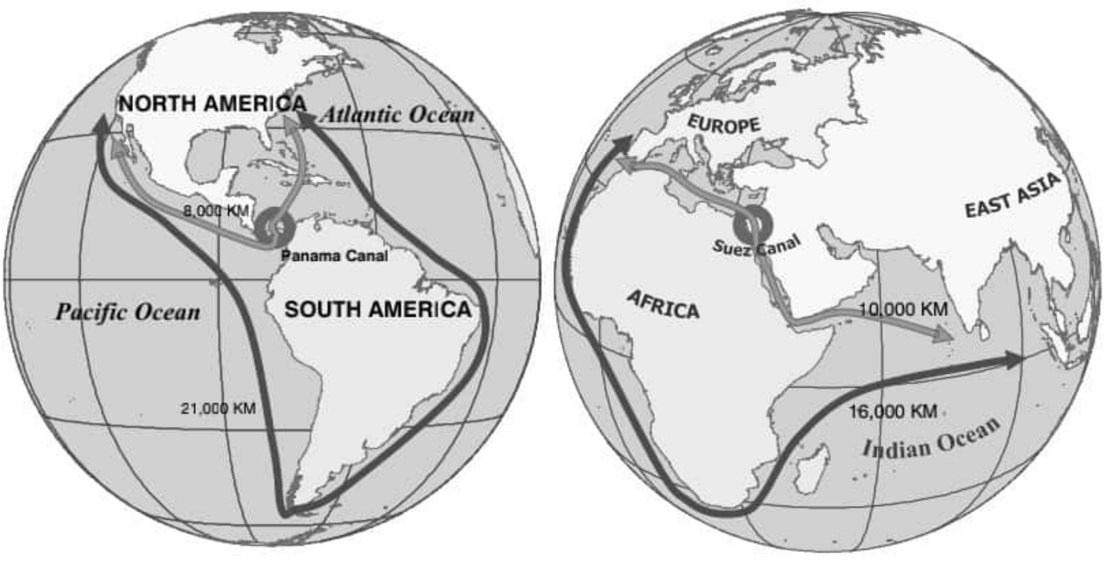


Figure 1.12 Geographical impacts of the Panama and Suez canals

railway transport system as railway networks expanded tremendously and became the dominant land transport mode for both passengers and freight. As the speed and power of locomotives improved and as the market expanded, rail services became increasingly specialized, with trains entirely devoted to passengers or freight. Rail systems reached a phase of maturity.

Another significant technological change of this era involved urban transportation, which until then solely relied on walking and different types of carriages (mainly horse drawn). The significant growth of the urban population favored the construction of the first public urban transport systems. Electric energy became widely used in the 1880s and considerably changed urban transport systems with the introduction of tramways (streetcars), notably in Western Europe and in the United States. They enabled the first forms of urban sprawl and the specialization of economic functions, notably by a wider separation between the places of work and residence. In large agglomerations, underground metro systems began to be constructed, London being the first in 1863. The bicycle, first shown at the Paris Exhibition of 1867, was also an important innovation which changed commuting in the late nineteenth century. Initially, the rich used it as a form of leisure, but it was rapidly adopted by the labor class as a mode of transportation to the workplace. Today, the bicycle is much less used in developed countries (outside of recreational purposes), but it is still a major mode of transportation in developing countries, especially China.

This era also marked the first significant developments in telecommunications. In 1844, Samuel Morse built the first experimental telegraph line in the United States between Washington and Baltimore, opening a new era in the transmission of information. By 1852, more than 40,000 km of telegraph lines were in service in the United States. In 1866, the first successful transatlantic telegraph line marked the inauguration of an intercontinental telegraphic network. The growth of telecommunications is thus closely associated with the growth of railways and international shipping. Managing a rail transport system, especially at the continental level, became more efficient with telegraphic communication. In fact, continental rail and telegraphic networks were often laid concomitantly. Telecommunications were also a dominant factor behind the creation of standard times zones in 1884. From a multiplicity of local times, zones of constant time with Greenwich (England) as the reference were laid. This improved the scheduling of passenger and freight transportation at national levels. By 1895, every continent was linked by telegraphic lines, a precursor of the global information network that would emerge in the late twentieth century. Business transactions became more efficient as production, management and consumption centers could interact with delays that were in hours instead of weeks and even months.

Transportation in the Fordist era (1920–70)

The Fordist era was epitomized by the adoption of the assembly line as the dominant form of industrial production, an innovation that benefited transportation substantially. The internal combustion engine, or four-stroke engine by Daimler (1889), which was a modified version of the Diesel engine (1885), and the pneumatic tire (1885) by Dunlop made road vehicle operations faster and more comfortable. Compared with steam engines, internal combustion engines have a much higher efficiency and use a lighter fuel: petrol. Petrol, previously perceived as an unwanted by-product of the oil refining process, which was seeking kerosene for illumination, became a convenient fuel. Initially, diesel engines were bulky, limiting their use to industrial and maritime propulsion, a purpose which they still fulfill today. The internal combustion engine permitted an extended flexibility of movements with fast, inexpensive and ubiquitous (door to door) transport modes such as automobiles, buses and trucks. Mass producing these vehicles changed considerably the industrial production system, notably by 1913 when Ford began the production of the Model T car using an assembly line. From 1913 to 1927, about 14 million Ford Model T cars were built, making it the second most important production car, behind the Volkswagen Beetle. The rapid diffusion of the automobile marked an increased demand for oil products and other raw materials such as steel and rubber.

Economies of scale also improved transportation in terms of capacity, which enabled to move low-cost bulk commodities such as minerals and grain over long distances. Oil tankers are a good example of the application of this principle to transport larger quantities of oil at a lower cost, especially after World War II when global demand surged. Maritime routes were thus expanded to include tanker routes, notably from the Middle East, the dominant global producer of oil. The very long distances concerned in the oil trade favored the construction of larger tankers. In the 1960s, tanker ships of 100,000 tons became available, to be supplanted by VLCCs (Very Large Crude Carriers) of 250,000 tons in the 1970s and by ULCCs (Ultra Large Crude Carriers) of 550,000 tons at the end of the 1970s. A ship of 550,000 tons is able to transport 3.5 million tons of oil annually between the Persian Gulf and Western Europe.

Although the first balloon flight took place in 1783, due to the lack of propulsion no practical applications for air travel were realized until the twentieth century. The first propelled flight was made in 1903 by the Wright brothers and inaugurated the era of air transportation. The initial air transport services were targeted at mail since it was a type of freight that could be easily transported and initially proved to be more profitable than transporting passengers. The year 1919 marked the first commercial air transport service between England and France, but air transport suffered from limitations in terms of capacity and range. Several attempts were made at developing dirigible services, as the Atlantic was crossed by a Zeppelin dirigible in 1924. However, such technology was abandoned in 1937 after the Hindenburg accident, in which the hydrogen filled reservoirs burned. The 1920s and 1930s saw the expansion of regional and national air transport services in Europe and the United States with successful propeller aircrafts such as the Douglas DC-3. The post-World War II period was however the turning point for air transportation as the range, capacity and speed of aircraft increased as well as the average income of the passengers. A growing number of people were thus able to afford the speed and convenience of air transportation. In 1958, the first commercial jet plane, the Boeing 707, entered service and revolutionized international movements of passengers, marking the end of passenger transoceanic ships.

Basic telecommunication infrastructures, such as the telephone and the radio, were mass marketed during the Fordist era. However, the major change was the large diffusion of the automobile, especially from the 1950s as it became a truly mass consumption product. No other mode of transportation has so drastically changed lifestyles and the structure of cities, notably for developed countries. It created suburbanization and expanded cities to areas larger than 100 km in diameter in some instances. In dense and productive regions, such as the Northeast of the United States, the urban system became structured and interconnected by transport networks to the point that it could be considered as one vast urban region: the Megalopolis.

A new context for transportation: the post-Fordist era (1970–)

Among the major changes in international transportation since the 1970s are the massive development of telecommunications, the globalization of trade, more efficient distribution systems, and the considerable development of air transportation.

Telecommunications enabled growing information exchanges, especially for the financial and service sectors. After 1970, telecommunications successfully merged with information technologies. As such, telecommunication also became a medium of doing business in its own right, in addition to supporting and enhancing other transportation modes. The information highway became a reality as fiber optic cables gradually replaced copper wires, multiplying the capacity to transmit information between computers. This growth was however dwarfed by the tremendous growth in processing power of computers, which are now fundamental components of economic and social activities in developed countries. A network of satellite communication was also created to support the growing exchanges of information, especially for television images. Out of this wireless technology emerged local cellular networks which expanded and merged to cover whole cities, countries, regions and then continents. Telecommunications have reached the era of individual access, portability and global coverage.

In a post-Fordist system, the fragmentation of production, organizing an international division of work, as well as the principle of "just-in-time" increased the quantity of freight moving at the local, regional and international levels. This in turn required increasing efforts to manage freight and reinforced the development of logistics, the science of physical distribution systems. Containers, the main agents of the modern international transport system, enabled an increased flexibility of freight transport, mainly by reducing transshipment costs and delays. Handling a container requires 25 times less labor than its equivalent in bulk freight. They were introduced by the American entrepreneur, Malcolm McLean who initially applied containerization to land transport. However, the true potential of containerization became clear when interfacing with other modes became possible, mainly between maritime, rail and road transportation.

The first containership (the Ideal-X, a converted T2 oil tanker) set sail in 1956 from New York to Houston and marked the beginning of the era of containerization. The Sea-Land Company established the first regular maritime container line in 1965 over the Atlantic between North America and Western Europe. In 1960, the Port Authority of New York/New Jersey, foreseeing the potential in container trade, constructed the first specialized container terminal next to Port Newark: the Port Elizabeth Marine Terminal. The first international container shipping services began in 1966 between the East Coast of the United States and Western Europe. By the early 1980s, container services with specialized ships (cellular containerships, first introduced in 1967) became a dominant aspect of international and regional transport systems. However, the size of those ships remained for 20 years constrained by the size of the Panama Canal, which de facto became the panamax standard. In 1988, the first post-panamax containership was introduced, an indication of the will to further expand economies of scale in maritime container shipping.

Air and rail transportation experienced remarkable improvements in the late 1960s and early 1970s. The first commercial flight of a Boeing 747 between New York and London in 1969 marked an important landmark for international transportation (mainly for passengers, but freight became a significant function in the 1980s). This giant plane can transport around 400 passengers, depending on the configuration. It permitted a considerable reduction of air fares through economies of scale and opened intercontinental air transportation to the mass market. Attempts were also undertaken to establish faster-than-sound commercial services with the Concorde (1976; flying at 2,200 km/hr). However, such services proved to be financially unsound and no new supersonic commercial planes have been built since the 1970s. The Concorde was finally retired in 2003. At the regional level, the emergence of high-speed train networks provided fast and efficient inter-urban services, notably in France (1981; TGV; speeds up to 300 km/hr) and in Japan (1964; Shinkansen; speeds up to 275 km/hr).

Major industrial corporations making transportation equipment, such as car manufacturers, have become dominant players in the global economy. Even if the car is not an international transport mode, its diffusion has expanded global trade of vehicles, parts, raw materials and fuel (mainly oil). Car production, which used to be mainly concentrated in the United States, Japan and Germany, has become a global industry with a few key players part of well integrated groups such as Ford, General Motors, Daimler Chrysler, Toyota and Mitsubishi. Along with oil conglomerates, they have pursued strategies aimed at the diffusion of the automobile as the main mode of individual transportation. This has led to growing mobility but also to congestion and waste of energy. As of the twenty-first century begins, the automobile accounts for about 80 percent of the total oil consumption in developed countries.

The second half of the twentieth century has seen a major shift in car production (Figure 1.13). In 1950, the United States accounted for more than 80 percent of global car production. However, this share declined to about 9.6 percent in 2004, reflecting the loss of competitiveness of the American car manufacturing system. The United States, even if it represents the largest car market in the world, has been thoroughly motorized, which means that its market is mainly one of replacement with acute competition between manufacturers for market share. Roughly the same number of cars was produced in the United States during the 1990s than during the 1950s. In the 1960s, two major players in the car industry emerged, Japan and Germany. They respectively accounted for 19.7 percent and 11.7 percent of global car production in 2004. A growing proportion of cars are being manufactured in newly industrialized countries, but the main consumption market still remains the developed world, under the control of American, Japanese and German car manufacturers.

The current period is also one of transport crises, mainly because of a dual dependency. First, transportation modes have a heavy dependence on fossil fuels and second, road transportation has assumed dominance. The oil crisis of the early 1970s, which saw a significant increase in fuel prices, induced innovations in transport modes, the reduction of energy consumption and the search for alternative sources of energy (electric car, adding ethanol to gasoline and fuel cells). However, from the mid-1980s to the end of the 1990s, oil prices declined and attenuated the importance of these initiatives.

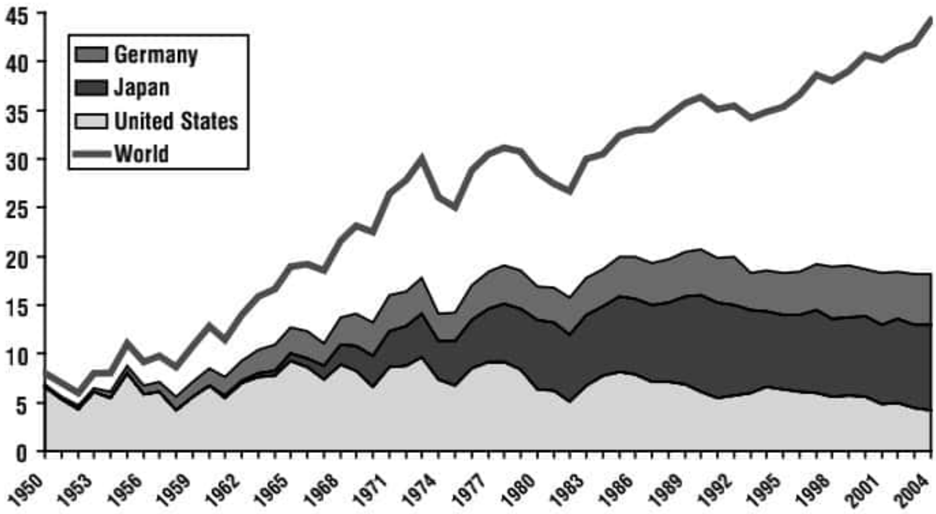


Figure 1.13 Automobile production, United States, Japan and Germany, 1950–2004 (in millions)
Source: Worldwatch Institute; International Organization of Motor Vehicle Manufacturers, http://www.oica.net

The reliance on fossil fuels continued unabated with a particularly strong growth of motorization in developing countries.

Future transportation

In the 200 years since the beginning of mechanized transportation, the capacity, speed, efficiency and geographical coverage of transport systems has improved dramatically. These processes can be summarized as follows:

- Each mode, due to its geographical and technical specificities, was characterized by different technologies and different rates of innovation and diffusion. A **transport innovation** can thus be an additive/competitive force where a new technology expands or makes an existing mode more efficient and competitive. It can also be a destructive force when a new technology marks the obsolescence and the demise of an existing mode, often through a paradigm shift.
- **Technological innovation** is linked with faster and more efficient transport systems. This process implies a space–time convergence where a greater amount of space can be exchanged with lesser amount of time. The comparative advantages of space can thus be more efficiently used.
- **Technological evolution** in the transport sector is linked with the phases of economic development of the world economy. Transportation and economic development are consequently interlinked as one cannot occur without the other.

Technological developments have two significant consequences over transportation modes. The first involves the emergence of new modes and the second concerns an improvement of their operational speeds (Figure 1.14). Many modes follow a similar pattern where a significant growth of their operational speed takes place in their introduction phase. Once technical constraints are solved and modal networks expand, the operational speeds reach a threshold which remains until the mode becomes obsolete and is abandoned (stagecoaches, clipper ships and liners) or a new technology is introduced and a new wave of technical improvements occurs (jet planes, high-speed trains).

Since the introduction of commercial jet planes, high-speed train networks and containers in the late 1960s, no significant technological changes have impacted on passenger and freight transport systems. The early twenty-first century is an era of car and truck dependency, which tends to constrain the development of alternative modes of transportation, as most of the technical improvements aim at insuring the dominance of oil as a source of energy. However, with dwindling oil reserves, the end of the dominance of the internal combustion engine is approaching. As oil production is expected to peak by 2008–10 and then decline, energy prices are expected to soar, triggering the most important technological transition in transportation since the automobile. Among the most promising technologies are:

- **Maglev**. Short for magnetic levitation, a maglev system has the advantage of having no friction with its support and no moving parts, enabling to reach operational speeds of 500–600 km per hour (higher speeds are possible if the train circulates in a low pressure tube). This represents an alternative for passenger and freight land movements in the range of 75 to 1,000 km. Maglev improves from the existing technology of high-speed train networks which are limited to speeds of 300 km per hour. In fact, maglev is the first fundamental innovation in railway transportation since the industrial revolution. The first commercial maglev system opened in Shanghai in 2003 and has an operational speed of about 440 km per hour.

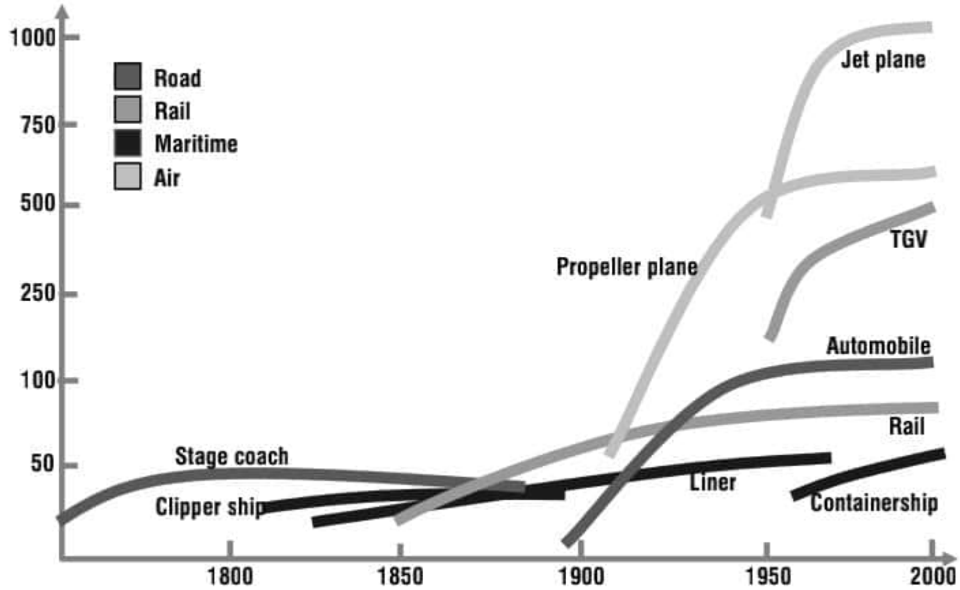


Figure 1.14 Development of operational speed for major transport modes, 1750–2000 (km per hour)

- **Automated transport systems**. Refers to a set of alternatives to improve the speed, efficiency, safety and reliability of movements, by relying upon complete or partial automation of the vehicle, transshipment and control. These systems could involve the improvement of existing modes such as automated highway systems, of the creation of new modes and new transshipment systems such as for public transit and freight transportation. The goal of this initiative is mainly to use existing infrastructures efficiently.
- **Fuel cells**. An electric generator using the catalytic conversion of hydrogen and oxygen. The electricity generated can be used for many purposes, such as supplying an electric motor. Current technological prospects do not foresee high output fuel cells, indicating they are applicable only to light vehicles, notably cars, or to small power systems. Nevertheless, fuel cells represent a low environmental impact alternative to generate energy and fuel cell cars are expected to reach mass production by 2010. Additional challenges in the use of fuel cells involve hydrogen storage (especially in a vehicle), as well as establishing a distribution system to supply the consumers.

A fundamental component of future transport systems, freight and passengers alike, is that they must provide increased flexibility and adaptability.

Method 1 – The notion of accessibility

Definition

Accessibility is a key element to transport geography, and to geography in general, since it is a direct expression of mobility either in terms of people, freight or information. Well-developed and efficient transportation systems offer high levels of accessibility (if the impacts of congestion are excluded), while less-developed ones have lower levels of accessibility. Thus accessibility is linked with an array of opportunities, economic and social.

Accessibility is defined as the measure of the capacity of a location to be reached by, or to reach different locations. Therefore, the capacity and the structure of transport infrastructure are key elements in the determination of accessibility.

All places are not equal because some are more accessible than others, which implies inequalities. The notion of accessibility consequently relies on two core concepts:

- The first is **location** where the relativity of places is estimated in relation to transport infrastructures, since they offer the mean to support movements.
- The second is **distance**, which is derived from the connectivity between locations. Connectivity can only exist when there is a possibility to link two locations through transportation. It expresses the friction of space (or deterrence) and the location which has the least friction relative to others is likely to be the most accessible. Commonly, distance is expressed in units such as in kilometers or in time, but variables such as cost or energy spent can also be used.

There are two spatial categories applicable to accessibility problems, which are interdependent:

- The first type is known as **topological accessibility** and is related to measuring accessibility in a system of nodes and paths (a transportation network). It is assumed that accessibility is a measurable attribute significant only to specific elements of a transportation system, such as terminals (airports, ports or subway stations).
- The second type is know as **contiguous accessibility** and involves measuring accessibility over a surface. Under such conditions, accessibility is a measurable attribute of every location, as space is considered in a contiguous manner.

Last, accessibility is a good indicator of the underlying spatial structure since it takes into consideration location as well as the inequality conferred by distance. Due to different spatial structures, two different locations of the same importance will have different accessibilities. On example A of Figure 1.15, representing a spatial structure where locations are uniformly distributed, locations 1 and 2 have different accessibilities, with location 1 being the most accessible. As distance (Euclidean) increases, location 1 has access to a larger number of locations than location 2. To access all locations, location 2 would require a longer traveled distance (roughly twice) than location 1. This is particularly the case when the spatial structure is one concentrated around location 1 (Example B). In this case, the number of locations that can be reached by location 1 climbs rapidly and then eventually peaks. The third example (C) has a spatial structure with roughly two foci. Although the number of locations that can be reached from location 2 initially climbs faster than for location 1, location 1 catches up and is actually the most accessible, but by a smaller margin.

Connectivity

The most basic measure of accessibility involves network connectivity where a network is represented as a connectivity matrix ($C1$), which expresses the connectivity of each node with its adjacent nodes. The number of columns and rows in this matrix is equal to the number of nodes in the network and a value of 1 is given for each cell where this is a connected pair and a value of 0 for each cell where there is an unconnected pair. The

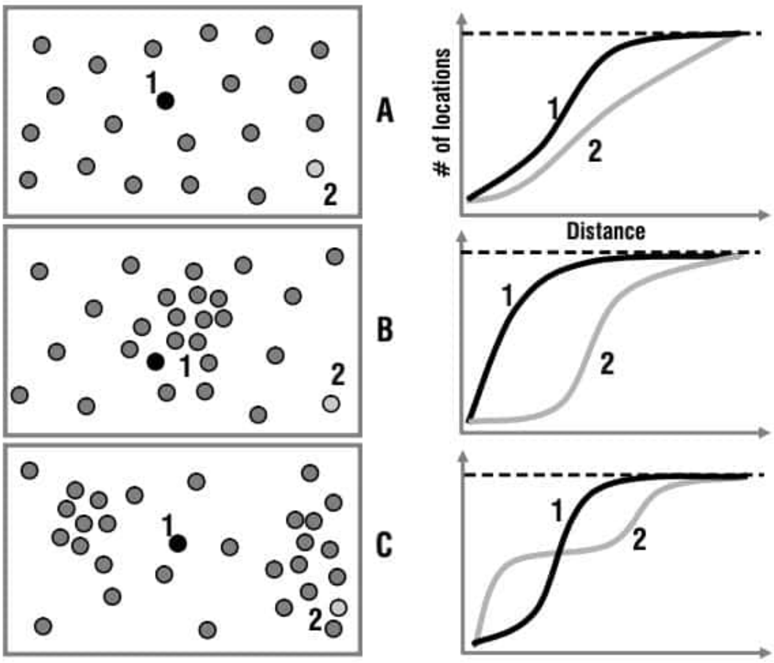


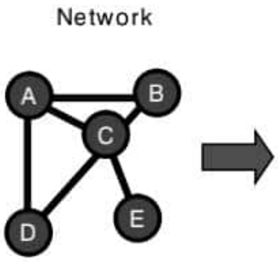
Figure 1.15 Accessibility and spatial structure

summation of this matrix provides a very basic measure of accessibility, also known as the degree of a node:

$$C1 = \sum_{j}^{n} c_{ij}$$

- $C1$ = degree of a node.
- c_{ij} = connectivity between node i and node j (either 1 or 0).
- n = number of nodes.

The network on Figure 1.16 can be represented as a connectivity matrix, which is rather simple to construct. The size of the connectivity matrix involves a number of rows and cells equivalent to the number of nodes in the network. Since the network on Figure 1.16 has five nodes, its connectivity matrix is a 5 × 5 grid. Each cell representing a connection between two nodes receives a value of 1 (e.g. cell B–A). Each cell that does not represent a connection gets a value of 0 (e.g. cell D–E). If all connections in the network are bi-directiona, the connectivity matrix is transposable. Adding up a row or a column gives the degree of a node. Node C is obviously the most connected since it has the highest summation of connectivity compared to all other nodes. However, this assumption may not hold true on a more complex network because of a larger number of indirect paths which are not considered in the connectivity matrix. The connectivity matrix does not take into account all the possible indirect paths between nodes. Under such circumstances, two nodes could have the same degree, but may have different accessibilities.



Connectivity matrix

	A	B	C	D	E
A	0	1	1	1	0
B	1	0	1	0	0
C	1	1	0	1	1
D	1	0	1	0	0
E	0	0	1	0	0

Figure 1.16 Connectivity matrix

Geographic and potential accessibility

From the accessibility measure developed so far, it is possible to derive two simple and highly practical measures, defined as geographic and potential accessibility. Geographic accessibility considers that the accessibility of a location is the summation of all distances between other locations divided by the number of locations.

$$A(G) = \sum_{i}^{n}\left(\sum_{j}^{n} d_{ij}\right) / n$$

- $A(G)$ = geographical accessibility matrix.
- d_{ij} = shortest path distance between location i and j.
- n = number of locations.

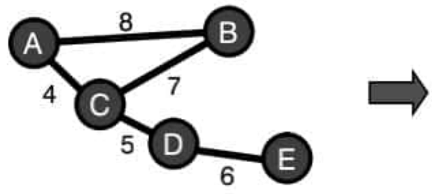
In this measure of accessibility, the most accessible place has the lowest summation of distances. As shown on Figure 1.17, the construction of a geographic accessibility matrix, $A(G)$, is a rather simple undertaking. First, build a matrix containing the shortest distance between the nodes (node A to node E), here labeled as the L matrix. Second, build the geographic accessibility matrix $A(G)$ with the summation of rows and columns divided by the number of locations in the network. The summation values are the same for columns and rows since this is a transposable matrix. The most accessible place is node C, since it has the lowest summation of distances.

Although geographic accessibility can be solved using a spreadsheet (or manually for simpler problems), Geographic Information Systems have proven to be a very useful and flexible tool to measure accessibility, notably over a surface simplified as a matrix (raster representation). This can be done by generating a distance grid for each place and then summing all the grids to form the total summation of distances (Shimbel) grid. The cell having the lowest value is thus the most accessible place.

Potential accessibility is a more complex measure than geographic accessibility, since it includes the concept of distance weighted by the attributes of a location. All locations are not equal and thus some are more important than others. Potential accessibility can be measured as follows:

$$A(P) = \sum_{i}^{n} P_i + \sum_{j}^{n} P_j / d_{ij}$$

- $A(P)$ = potential accessibility matrix.
- d_{ij} = distance between place i and j (derived from valued graph matrix).
- P_j = attributes of place j, such as its population, retailing surface, parking space, etc.
- n = number of locations.



L

	A	B	C	D	E
A	0	8	4	9	15
B	8	0	7	12	18
C	4	7	0	5	11
D	9	12	5	0	6
E	15	18	11	6	0

A(G)

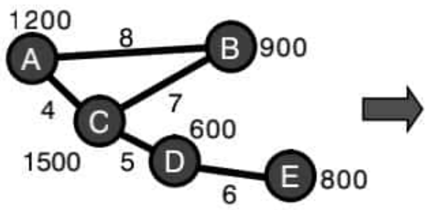
	A	B	C	D	E	Σ/n
A	0	8	4	9	15	7.2
B	8	0	7	12	18	9.0
C	4	7	0	5	11	5.4
D	9	12	5	0	6	6.4
E	15	18	11	6	0	10.0
Σ/n	7.2	9.0	5.4	6.4	10.0	38.0

Figure 1.17 Geographic accessibility

The potential accessibility matrix is not transposable since locations do not have the same attributes, which brings the underlying notions of emissiveness and attractiveness:

- **Emissiveness** is the capacity to leave a location, the sum of the values of a row in the *A*(*P*) matrix.
- **Attractiveness** is the capacity to reach a location, the sum of the values of a column in the *A*(*P*) matrix.

By considering the same shortest distance matrix (L) as on Figure 1.17 and the population matrix P, the potential accessibility matrix, *P*(*G*), can be calculated (Figure 1.18). The value of all corresponding cells (A–A, B–B, etc.) equals the value of their respective attributes (*P*). The value of all non-corresponding cells equals their attribute divided by the corresponding cell in the L matrix. The higher the value, the more a location is accessible, node C being the most accessible. The matrix being non-transposable, the summation of rows is different from the summation of columns, bringing forward the issue of implying different levels of attractiveness and emissiveness. Node C has more attractiveness than emissiveness (2525.7 versus 2121.3), while Node B has more emissiveness than attractiveness (1358.7 versus 1266.1). Likewise, a Geographic Information System can be used to measure potential accessibility, notably over a surface.



L

	A	B	C	D	E
A	0	8	4	9	15
B	8	0	7	12	18
C	4	7	0	5	11
D	9	12	5	0	6
E	15	18	11	6	0

P

A	1200
B	900
C	1500
D	600
E	800

P (G)

i\j	A	B	C	D	E	Σi
A	1200.0	150.0	300.0	133.3	80.0	1863.3
B	112.5	900.0	128.6	75.0	50.0	1266.1
C	375.0	214.3	1500.0	300.0	136.4	2525.7
D	66.6	50.0	120.0	600.0	100.0	936.6
E	53.3	44.4	72.7	133.3	800.0	1103.7
Σj	1807.4	1358.7	2121.3	1241.6	1166.4	7695.4

Figure 1.18 Potential accessibility

Method 2 – Geographic information systems for transportation (GIS-T)[1]

Introduction

In a broad sense a geographic information system (GIS) is an information system specializing in the input, storage, manipulation, analysis and reporting of geographical (spatially related) information. Among the wide range of potential applications GIS can be used for, transportation issues have received a lot of attention. A specific branch of GIS applied to transportation issues, commonly labeled as GIS-T, has emerged.

> Geographic information systems for transportation (**GIS-T**) refers to the principles and applications of applying geographic information technologies to transportation problems.

The four major components of a GIS, encoding, management, analysis and reporting, have specific considerations for transportation (Figure 1.19):

- **Encoding**. Deals with issues concerning the representation of a transport system and its spatial components. To be of use in a GIS, a transport network must be correctly encoded, implying a functional topology composed of nodes and links. Other elements relevant to transportation, namely qualitative and quantitative data, must also be encoded and associated with their respective spatial elements. For instance, an encoded road segment can have data related to its width, number of lanes, direction, peak hour traffic, etc.
- **Management**. The encoded information is stored in a file structure and can be organized along spatial (by region, country, census bloc, etc.), thematic (for road,

1 Dr Shih-Lung Shaw (University of Tennessee) is the major contributor of this section.

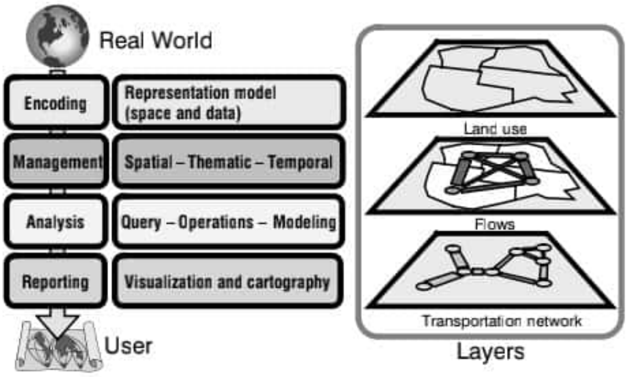


Figure 1.19 Geographic information systems and transportation

transit, terminals, etc.) or temporal (by year, month, week, etc.) considerations. This process is commonly automatic and has established conventions. For instance, many government agencies use specific file formats and deploy their information along predetermined spatial (their jurisdiction), thematic (their field of interest) and temporal (their frequency of data collection) considerations.

- **Analysis**. Considers the wide array of tools and methodologies available for transport issues. They can range from a simple query over an element of a transport system (what is the peak hour traffic of a road segment?) to a complex model investigating the relationships between its elements (if a new road segment was added, what would be the impacts on traffic and future land use developments?).
- **Reporting**. A GIS would not be complete without all its visualization and cartographic capabilities. This component is particularly important as it gives the possibility to convey complex information in a symbolic format. A GIS-T thus becomes a tool to inform and convince actors who otherwise may not have the time or the capability for non-symbolic data interpretation.

Information in a GIS is often stored and represented as layers, which are a set of geographical features linked with their attributes. On Figure 1.19 a transport system is represented as three layers related to land use, flows (spatial interactions) and the network. Each has its own features and related data.

GIS-T research can be approached from two different, but complementary, directions. While some GIS-T research focuses on issues of how GIS can be further developed and enhanced in order to meet the needs of transportation applications, other GIS-T research investigates the questions of how GIS can be used to facilitate and improve transportation studies. In general, topics related to GIS-T studies can be grouped into three categories:

- **Data representations**. How can various components of transport systems be represented in a GIS-T?
- **Analysis and modeling**. How can transport methodologies be used in a GIS-T?
- **Applications**. What types of applications are particularly suitable for GIS-T?

GIS-T data representations

Data representation is a core research topic of GIS. Before a GIS can be used to tackle real world problems, data must be properly represented in a digital computing environment.

One unique characteristic of GIS is the capability of integrating spatial and nonspatial data in order to support both display and analysis needs. There have been various data models developed for GIS. The two basic approaches are **object-based data models** and **field-based data models**.

- An object-based data model treats geographic space as populated by **discrete and identifiable objects**. Features are often represented as points, lines, and/or polygons.
- On the other hand, a field-based data model treats geographic space as populated by **real-world features** that vary continuously over space. Features can be represented as regular tessellations (e.g. a raster grid) or irregular tessellations (e.g. a triangulated irregular network – TIN).

Representing the "real world" in a data model has been a challenge for GIS since their inception in the 1960s. A GIS data model enables a computer to represent real geographical elements as graphical elements. As shown on Figure 1.20, two representational models are possible: raster (grid-based) and vector (line-based):

- **Raster**. Based on a cellular organization that divides space into a series of units. Each unit is generally similar in size to another. Grid cells are the most common raster representation. Features are divided into cellular arrays and a coordinate (X,Y) is assigned to each cell, as well as a value. This allows for registration with a geographic reference system. A raster representation also relies on **tessellation**: geometric shapes that can completely cover an area. Although many shapes are possible (triangles and hexagons), the square is the most commonly used. The problem of resolution is common to raster representations. For a small grid, the resolution is coarse but the required storage space is limited. For a large grid the resolution is fine, but at the expense of a much larger storage space. On Figure 1.20, the real world (shown as an aerial photograph) is simplified as a grid where each cell color relates to an entity such as road, highway and river.
- **Vector**. The concept assumes that space is continuous, rather than discrete, which gives an infinite (in theory) set of coordinates. A vector representation is composed of three main elements: points, lines and polygons. **Points** are spatial objects with no area but can have attached attributes since they are a single set of coordinates (X and Y) in a coordinate space. **Lines** are spatial objects made up of connected points (nodes) that have no width. **Polygons** are closed areas that can be made up of a circuit of line segments. On Figure 1.20, the real world is represented by a series of lines (roads and highway) and one polygon (the river).

GIS-T studies have employed both object-based and field-based data models to represent the relevant geographic data. Some transportation problems tend to fit better with one type of GIS data model than the other. For example, **network analysis** based on the graph theory typically represents a network as a set of nodes interconnected with a set of links. The object-based GIS data model therefore is a better candidate for such transportation network representations. Other types of transportation data exist which require extensions to the general GIS data models. One well-known example is linear referencing data (e.g. highway mileposts). Transportation agencies often measure locations of features or events along transportation network links (e.g. a traffic accident occurred at the 52.3 milepost on a specific highway). Such a one-dimensional linear referencing system (i.e. linear measurements along a highway segment with respect to a pre-specified starting point of the highway segment) cannot be properly handled

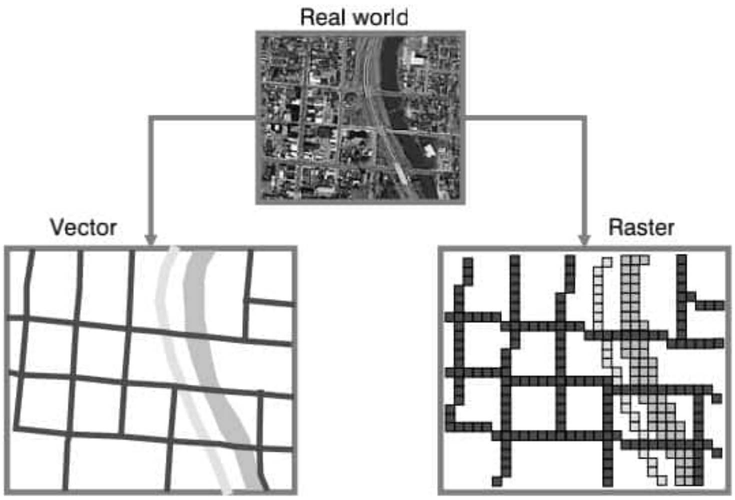


Figure 1.20 GIS data models

by the two-dimensional Cartesian coordinate system used in most GIS data models. Consequently, the dynamic segmentation data model was developed to address this specific need of the GIS-T community. Origin-destination (O-D) flow data are another type of data that are frequently used in transportation studies. Such data have been traditionally represented in matrix forms (i.e. as a two-dimensional array in a digital computer) for analysis. Unfortunately, the relational data model widely adopted in most commercial GIS software does not provide adequate support for handling matrix data. Some GIS-T software vendors therefore have developed additional functions for users to work with matrix data within an integrated GIS environment. The above examples illustrate how the conventional GIS approaches can be further extended and enhanced to meet the needs of transportation applications.

In recent years, the development of enterprise and multidimensional GIS-T data models has occurred. Successful GIS deployments at the enterprise level (e.g. within a state department of transportation) demand additional considerations to embrace the diversity of application and data requirements. An enterprise GIS-T data model is designed to allow "each application group to meet the established needs while enabling the enterprise to integrate and share data". The needs of integrating 1-D, 2-D, 3-D, and time for various transportation applications also have called for the implementation of multidimensional transportation location referencing systems.

In short, one critical component of GIS-T is how transportation-related data in a GIS environment can be best represented in order to facilitate and integrate the needs of various transportation applications. Existing GIS data models provide a good foundation of supporting many GIS-T applications. However, due to some unique characteristics of transportation data, many challenges still exist of developing better GIS data models that will improve rather than limit what we can do with different types of transportation studies.

GIS-T analysis and modeling

GIS-T applications have benefited from many of the standard GIS functions (query, geocoding, buffer, overlay, etc.) to support data management, analysis, and visualization needs. Like many other fields, transportation has developed its own unique analysis methods and models. Examples include shortest path and routing algorithms (e.g. traveling salesman problem, vehicle routing problem), spatial interaction models (e.g.

gravity model), network flow problems (e.g. user optimal equilibrium, system optimal equilibrium, dynamic equilibrium), facility location problems (e.g. p-median problem, set covering problem, maximal covering problem, p-centers problem), travel demand models (e.g. the four-step trip generation, trip distribution, modal split, and traffic assignment models), and land use–transportation interaction models.

While the basic transportation analysis procedures (e.g. shortest path finding) can be found in most commercial GIS software, other transportation analysis procedures and models (e.g. facility location problems) are available only selectively in some commercial software packages. Fortunately, the recent trend of moving towards component GIS design in the software industry provides a better environment for experienced GIS-T users to develop their own custom analysis procedures and models.

It is essential for both GIS-T practitioners and researchers to have a thorough understanding of transportation analysis methods and models. For GIS-T practitioners, such knowledge can help them evaluate different GIS software products and choose the one that best meets their needs. It also can help them select appropriate analysis functions available in a GIS package and properly interpret the analysis results. GIS-T researchers, on the other hand, can apply their knowledge to help improve the design and analysis capabilities of GIS-T.

GIS-T applications

GIS-T is one of the leading GIS application fields. Many GIS-T applications have been implemented at various transportation agencies over the last two decades. They cover much of the broad scope of transportation, such as infrastructure planning, design and management, transportation safety analysis, travel demand analysis, traffic monitoring and control, public transit planning and operations, environmental impacts assessment, hazards mitigation, and intelligent transportation systems (ITS). Each of these applications tends to have its specific data and analysis requirements. For example, representing a street network as centerlines and major intersections may be sufficient for a transportation planning application. A traffic engineering application, however, may require a detailed representation of individual traffic lanes. Turn movements at intersections also could be critical to a traffic engineering study, but not to a region-wide travel demand study. These different application needs are directly relevant to the GIS-T data representation and the GIS-T analysis and modeling issues discussed above. When a need arises to represent transportation networks of a study area at different scales, what would be an appropriate GIS-T design that could support the analysis and modeling needs of various applications? In this case, it may be preferable to have a GIS-T data model that allows multiple geometric representations of the same transportation network. Research on enterprise GIS-T data model and multidimensional, multimodal GIS-T data model discussed above aims at addressing these important issues of better integrating various GIS-T applications.

With the rapid growth of the Internet and wireless communications in recent years, a growing number of Internet-based and wireless GIS-T applications can be found. Such applications are especially common for ITS and for location-based services (LBS). Another trend observed in recent years is the growing number of GIS-T applications in the private sector, particularly for logistics applications. Since many businesses involve operations at geographically dispersed locations (e.g. supplier sites, distribution centers/ warehouses, retail stores, and customer sites), GIS-T can be useful tools for a variety of logistics applications. Again, many of these logistics applications are based on GIS-T analysis and modeling procedures such as the routing and the facility location problems.

GIS-T is interdisciplinary in nature and has many possible applications. Transportation geographers, who have appropriate backgrounds in both geography and transportation, are well positioned to pursue GIS-T studies.

Bibliography

Adams, T.M., N.A. Koncz and A.P. Vonderohe (2001) *Guidelines for the Implementation of Multimodal Transportation Location Referencing Systems*, NCHRP Report 460, Transportation Research Board, National Research Council. Washington, DC: National Academy Press.

Ausubel, J.H. and C. Marchetti (2001) "The Evolution of Transportation", *The Industrial Physicist*, April/May, pp. 20–24. http://www.aip.org/tip/INPHFA/vol–7/iss–2/p20.pdf.

Banister, D. (2002) *Transport Planning*, 2nd edn, London: Spon.

Black, W. (2003) *Transportation: A Geographical Analysis*, New York: Guilford.

Bogart, D. (2004) "Turnpike Trusts and the Transportation Revolution in Eighteenth Century England", http://orion.oac.uci.edu/~dbogart/transportrev_oct13.pdf.

Butler, J.A. and K.J. Dueker (2001) "Implementing the Enterprise GIS in Transportation Database Design", *Urban and Regional Information Systems Association* (*URISA) Journal*, 13(1), 17–28.

Haggett, P. (2001) *Geography: A Modern Synthesis*, 4th edn, New York: Prentice Hall.

Harrington, R. (1999) "Transport: Then, Now, and Tomorrow", *Royal Society of Arts Journal*, vol. CXLVI, no. 5488. http://www.york.ac.uk/inst/irs/irshome/papers/carmen.htm.

Hoover, E.M. (1948) *The Location of Economic Activity*, New York: McGraw-Hill.

Hoyle, B. and R. Knowles (1998) "Transport Geography: An Introduction", in B. Hoyle and R. Knowles (eds) *Modern Transport Geography*, 2nd edn, London: Wiley, pp.1–12.

Hoyle, B. and J. Smith (1998) "Transport and Development: Conceptual Frameworks", in B. Hoyle and R. Knowles (eds) *Modern Transport Geography*, 2nd edn, London: Wiley, pp. 13–40.

Hugill, P.J. (1992) *World Trade since 1431*, Baltimore, MD: Johns Hopkins University Press.

Lo, C.P. and A.K.W. Yeung (2002) *Concepts and Techniques of Geographic Information Systems*, Upper Saddle River, NJ: Prentice Hall.

Merlin, P. (1992) *Géographie des Transports, Que sais-je?*, Paris: Presses Universitaires de France.

Miller, H.J. and S.L. Shaw (2001) *Geographic Information Systems for Transportation: Principles and Applications*, New York: Oxford University Press.

Rimmer, P. (1985) "Transport Geography", *Progress in Human Geography*, 10, 271–7.

Rioux, J.-P. (1989) *La révolution industrielle, 1780–1880*, Paris: Éditions du Seuil.

Shaw, S.L. (2002) Book Review: Geographic Information Systems in Transportation Research, *Journal of Regional Science*, 42(2), 418–21.

Taaffe, E.J., H.L. Gauthier and M.E. O'Kelly (1996) *Geography of Transportation*, 2nd edn, Upper Saddle River, NJ: Prentice Hall.

Thill, J.C. (ed.) (2000) *Geographic Information Systems in Transportation Research*, Oxford: Elsevier Science.

Tolley, R. and B. Turton (1995) *Transport Systems, Policy and Planning: A Geographical Approach*, Harlow: Longman.

Victoria Transport Policy Institute (2005) "Defining, Evaluating and Improving Accessibility", *Transport Demand Management Encyclopedia*, http://www.vtpi.org/tdm/tdm84.htm.

Williams, A. (1992) "Transport and the Future", in B.S. Hoyle and R.D. Knowles (eds) *Modern Transport Geography*, London: Belhaven Press, pp. 257–70.

2 Transportation systems and networks

Transportation systems are composed of a complex set of relationships between the demand, the locations they service and the networks that support movements. They are mainly dependent on the commercial environment from which are derived operational attributes such as transportation costs, capacity, efficiency, reliability and speed. Such conditions are closely related to the development of transportation networks, both in capacity and in spatial extent. Transportation systems are also evolving within a complex set of relationships between transport supply, mainly the operational capacity of the network, and transport demand, the mobility requirements of a territory. This chapter consequently investigates the relationships between transportation networks and their spatial structure.

Concept 1 – Transportation and commercial geography

Trade and commercial geography

Economic systems are based on trade and transactions since specialization and efficiency require interdependency. People trade their labor for a wage while corporations trade their output for capital. Trade is the transmission of a possession in return for a counterpart, generally money. The exchange involves a **transaction** and its associated **flows** of capital, information, commodities, parts, or finished products. All this necessitates the understanding of commercial geography.

> **Commercial geography** investigates the spatial characteristics of trade and transactions in terms of their cause, nature, origin and destination. It leans on the analysis of contracts and transactions. From a simple commercial transaction involving an individual purchasing a product at a store, to the complex network of transactions maintained between a multinational corporation and its suppliers, the scale and scope of commercial geography varies significantly.

Commercial geography is concerned with transactions (Figure 2.1). As each transaction involves movements of people, freight and information, there is a close relationship between the sphere of transactions (the geographical setting of transactions) and the sphere of circulation (the geographical setting of movements). This implies transaction costs and transportation costs. The main transaction costs are: 1) Search and information costs: costs related to finding the appropriate goods on the market, who has them and at what price. 2) Negotiation costs: costs involved in reaching an agreement with the other party to the transaction, the contract being the outcome. 3) Policing and enforcement costs: costs related to ensuring that both parties respect the terms of the contract and, if this is not the case, taking legal actions to correct the situation.

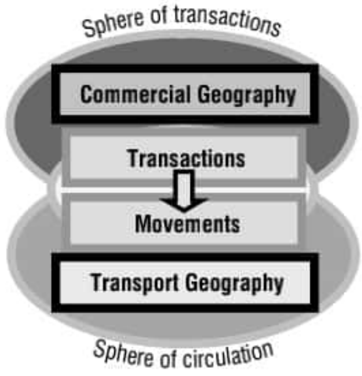


Figure 2.1 Commercial and transport geography

Trade, in terms of its origins and destinations, has a spatial logic. It reflects the economic, social and industrial structure of the concerned markets, but also implies other factors such as transport costs, distance, political ties, exchange rates and the reciprocal economic advantages proponents get from trade. For trade to occur several conditions must be met:

- **Availability**. Commodities, from coal to computer chips, must be available for trade and there must be a demand for these commodities. In other terms, a surplus must exist at one location and a demand in another. A surplus can often be a simple matter of investment in production capabilities, such as building an assembly plant, or can be constrained by complex environmental factors like the availability of resources such as fossil fuels, minerals and agricultural products.
- **Transferability**. There are three major impediments to transferability, namely policy barriers (tariffs, custom inspections, quotas), geographical barriers (time, distance) and transportation barriers (the simple capacity to move the outcome of a transaction). Transport infrastructures, in allowing commodities to be moved from their origins to their destinations, favor the transferability of goods. Distance often plays an important role in trade, as does the capacity of infrastructures to route and to transship goods.
- **Transactional capacity**. It must be legally possible to make a transaction. This implies the recognition of a currency for trading and legislations that define the environment in which transactions are taking place, such as taxation. In the context of a global economy, the transactional environment is very complex but is important in facilitating trade at the regional, national and international levels.

Once these conditions are met, trade is possible and the outcome of a transaction results in a flow. Three particular issues relate to the concept of flow:

- **Value**. Flows have a negotiated value and are settled in a common currency. The American dollar, which has become the major global currency, is used to settle and/or measure many international transactions. Further, nations must maintain reserves of foreign currencies to settle their transactions and the relationship between the inbound and outbound flows of capital is known as the balance of payments. Although nations try to maintain a stable balance of payments, this is rarely the case.
- **Volume**. Flows have a physical characteristic, mainly involving a mass. The weight of flows is a significant variable when trade involves raw materials such as petroleum

or minerals. However, in the case of consumption goods, weight has little significance relative to the value of the commodities being traded. With containerization, a new unit of volume has been introduced: the TEU (Twenty-foot Equivalent Unit), which can be used to assess trade flows.

- **Scale**. Flows have a range which varies significantly based on the nature of a transaction. While retailing transactions tend to occur at a local scale, transactions related to the operations of a multinational corporation are global in scale.

Tendencies in commercial geography

The contemporary commercial setting is marked by increasing free trade and profound technological, industrial and geopolitical changes. The liberalization of trade, as confirmed by the implementation of the World Trade Organization, has given a strong impetus and a positive trend in the growth rate of world trade and industrial production. However, in a true free trade environment, regulatory agencies would not be required. In spite of attempts at deregulation, transactions and trade are prone to disputes, litigations and perceived imbalances concerning who benefits the most. Although these issues mainly apply to international trade, there are also situations where trade is constrained between the provinces/states of a nation.

In spite of globalization, much trade is still dominantly regional. An overview of world trade flows indicates that trade within regions is more significant than trade between regions, but long distance trade is steadily growing. Figures indicate the increasing share of East Asia, especially China, in world trade, in terms of both exports and imports. Flows of merchandise have also been accompanied by a substantial growth in foreign direct investments. There is thus a remarkable reallocation of production capacities following changes in comparative advantages around the world. This trend goes in tandem with mergers and acquisitions of enterprises that are increasingly global in scope. The analysis of international trade thus reveals the need to adopt different strategies to adapt to this new trading environment. As production is being relocated, there is a continuous shift in emphasis in the structure of export and import of world economies.

Recent decades have seen important modifications in international trading flows (Figure 2.2). The bulk of international trade occurs within economic blocs, especially the European Union and NAFTA. Other significant flows are between Asia/Pacific and

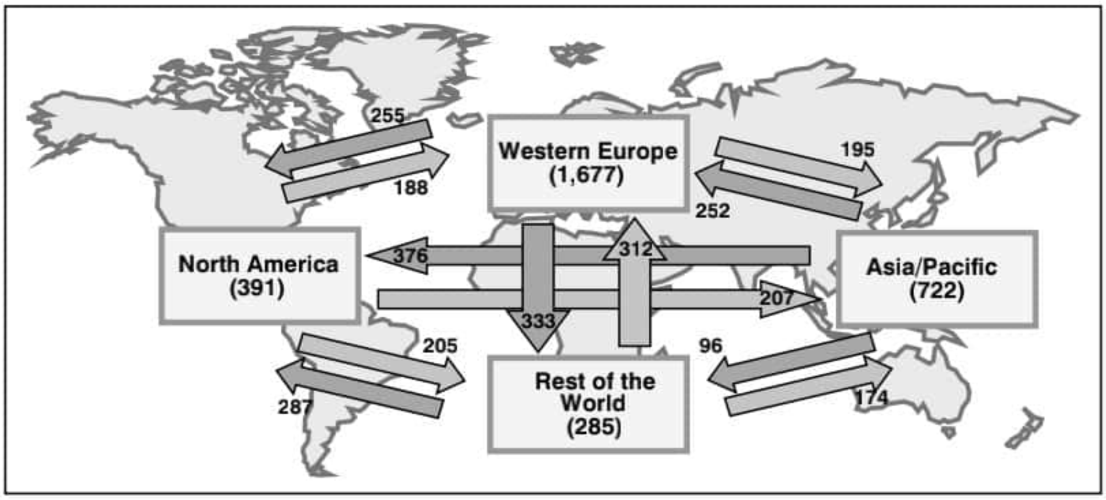


Figure 2.2 World trade flows, 2001 (billion $US) (Source: WTO)

North America (especially the United States), between Europe and North America and between Europe and Asia/Pacific. For several reasons, such as geographical proximity (Eastern Europe), energy (Middle East) and colonial legacy (Africa), the European Union has significant trading linkages with the rest of the world. North America also maintains important trade linkages with Latin America. Another important characteristic of the contemporary commercial setting concerns imbalances in trade flows. For instance, it is clear on Figure 2.2 that the Asia/Pacific region exports more than it imports and that North America imports more than it exports.

Major changes have occurred in the organization of production. There is a noticeable increase in the division of labor concerning the design, planning and assembly in the manufacturing process of the global economy. Interlocking partnerships in the structure of manufacturing have increased the trade of parts and the supply of production equipment around the world. One-third of all trade takes place among parent companies and their foreign affiliates. Part of this dynamism resides in the adoption of standards, a process which began in the late nineteenth century to promote mass production. It permitted the rapid development of many sectors of activity, including railways, electricity, the automobile and the telecommunication industry more recently (Internet, Electronic Data Interchange). In the realm of globalization of economic activities, the International Standards Organization developed the ISO norms that serve as comparison between various enterprises around the world. These norms are applicable to the manufacturing and services industries and are a necessary tool for growth.

Another significant force of change in commercial geography implies the growth of personal consumption, although this is not taking place uniformly. The bulk of consumption remains concentrated in a limited number of countries, with the G7 countries alone accounting for two-thirds of the global Gross Domestic Product. As a result, the commercial geography is influenced by the market size, the consumption level of an economy (often measured in GDP per capita), but also by the growth potential of different regions of the world. Economic growth taking place in East and Southeast Asia has been one of the most significant forces shaping changes in the contemporary commercial environment. The commodification of the economy has led to significant growth in retail and wholesale and the associated movements of freight.

Commercialization of the transport industry

The liberalization of trade was accompanied by a growth of transportation since transactions involve movements of freight, capital, people and information. Developments in the transport sector are matched by global and regional interdependence and competition. Transportation, like commodities, products and services, is traded, sometimes openly and subject to full market forces, but more often subject to a form of public control or ownership. The core component of a transport-related transaction involves costs that either have to be negotiated between the provider of the service and the user or are subject to some arbitrary decree (price fixing such as public transit). Since transportation can be perceived as a service to people, freight or information, its commercialization, how it is brought to the market, is an important dimension of its dynamics (Figure 2.3).

The extension of the operational scale of freight distribution insures that a production system reaches its optimal market potential, namely by a combination of strategies related to the exploitation of comparative advantages and a wider market base. Although an optimal market size can never be attained due to regulations preventing monopolies and differences in consumer preferences, the trend to insure maximal market exposure is unmistakable. The emergence of global brands and global production networks clearly

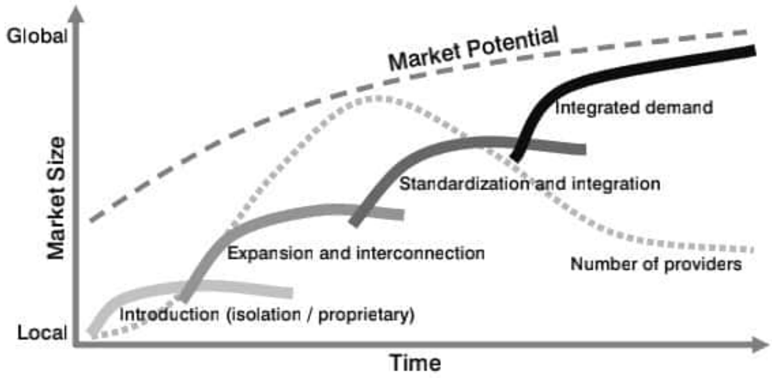


Figure 2.3 Commercialization of transportation

underlines this. Within freight distribution, four distinct cyclic phases of extension and functional integration can be identified:

- **Introduction**. Initially, a transport system is introduced to service a specific opportunity in an isolated context. The technology is often "proprietary" and incompatible with other transport systems. Since they are not interconnected, this does not represent much of an issue.
- **Expansion and interconnection**. As the marketability and the development potential of a transport system becomes apparent, a phase of expansion and interconnection occurs. The size of the market serviced by these transport systems consequently increases as they become adopted in new locations. At some point, independently developed transport systems connect. This connection is however often subject to a function of transshipment between two incompatible transport systems.
- **Standardization and integration**. This phase often involves the emergence of a fully developed distribution system servicing vast national markets. The major challenge to be addressed involves a standardization of modes and processes, further expanding the commercial potential of the concerned supply chains. Modal flows are moving more efficiently over the entire network and are able to move from one mode to another through intermodal integration. A process of mergers and acquisitions often accompanies this phase for the purpose of rationalization and market expansion.
- **Integrated demand**. The most advanced stage of extension of a distribution system involves a system that is fully able to answer freight mobility needs under a variety of circumstances, either predicted or unpredicted demand. As this system tends to be global, it commonly operates close to market potential. In such a setting, a distribution system expresses an integrated demand where the distribution capabilities are tuned to the demand in an interdependent system.

Each of these phases tends to be sequential and related to a historical process of transport development. For instance, up to the mid-nineteenth century, most distribution systems were isolated and developed independently from one another. Even global maritime transport was fragmented by national flags and trading systems. As regional transport systems grew in the second half of the nineteenth century, they gradually interconnected, but moving from one system to another required a form of transshipment. By the early twentieth century, most national transport systems were integrated, but interconnection between modes was difficult. The next challenge resided in the development of intermodal transportation, accelerated by containerization and information technologies.

One important component of the commercialization of transportation concerns **investments** in infrastructure, modes and terminals, as well as marketing. This task is performed either to expand the geographical extent and/or the capacity of a transport system or to maintain its operating conditions. The public and private sectors have contributed to the funding of transport investments depending on economic, social and strategic interests. For obvious reasons, the private sector seeks transport investments that promise economic returns while the public sector often invests for social and strategic reasons. In many cases private transport providers have difficulty in acting independently to formulate and implement their transport investments. Various levels of government are often lobbied by transport firms for financial and/or regulatory assistance in projects that are presented as of public interest and benefit. The consolidation of regional markets and the resulting increase in transborder traffic has led transport firms to seek global alliances and greater market liberalization in the transport and communication sector as a means to attract investments and to improve their productivity.

Deregulation and **divestiture** policy in the transport industry has led governments to withdraw from the management, operation and ownership of national carriers, ports and airports. This has given rise to a major reorganization of the international and national transport sectors with the emergence of transnational transport corporations that govern the global flow of air, maritime and land trade and the management of airports, ports and railyards (see Chapter 9).

Concept 2 – Transport costs

Transport costs and rates

Transport systems face requirements to increase their capacity and to reduce the costs of movements. All users (e.g. individuals, enterprises, institutions, governments, etc.) have to **negotiate** or **bid** for the transfer of goods, people, information and capital because supplies, distribution systems, tariffs, salaries, locations, marketing techniques as well as fuel costs are changing constantly. There are also costs involved in gathering information, negotiating, and enforcing contracts and transactions, which are often referred as the cost of doing business. Trade involves transaction costs that all agents attempt to reduce since transaction costs account for a growing share of the resources consumed by the economy.

Frequently, enterprises and individuals must take decisions about how to route passengers or freight through the transport system. This choice has been considerably expanded in the context of the production of lighter and high value consumer goods, such as electronics, and less bulky production techniques. It is not uncommon for transport costs to account for 20 percent of the total cost of a product. Thus, the choice of a transportation mode to route people and freight within origins and destinations becomes important and depends on a number of factors such as the nature of the goods, the available infrastructures, origins and destinations, technology, and particularly their respective distances. Jointly, they define transportation costs.

Transport costs are a monetary measure of what the transport provider must pay to produce transportation services. They come as fixed (infrastructure) and variable (operating) costs, depending on a variety of conditions related to geography, infrastructure, administrative barriers, energy, and on how passengers and freight are carried. Three major components, related to transactions, shipments and the friction of distance, impact on transport costs.

As shown on Figure 2.4, a movement between locations A and B involves three cost components in the assessment of its transport cost. The friction of distance represents how many units of space can be traded per unit of cost. Distance is a common attribute used to measure it. Shipment implies the mode used, the frequency as well as economies of scale.

Transport costs have significant impacts on the structure of economic activities as well as on international trade. Empirical evidence underlines that raising transport costs by 10 percent reduces trade volumes by more than 20 percent. In a competitive environment where transportation is a service that can be bidded on, transport costs are influenced by the respective rates of transport companies, the portion of the transport costs charged to users.

> **Rates** are the price of transportation services paid by their users. They are the negotiated monetary cost of moving a passenger or a unit of freight between a specific origin and destination. Rates are often visible to the consumers since transport providers must provide this information to secure transactions. They may not necessarily express the real transport costs.

The difference between costs and rates results in either a loss or a deficit from the service provider. Considering the components of transport costs previously discussed, rate setting is a complex undertaking subject to constant change. For public transit, rates are often fixed and the result of a political decision where a share of the total costs is subsidized by the society. The goal is to provide an affordable mobility to the largest possible segment of the population even if this implies a recurring deficit (public transit systems rarely make any profit). For freight transportation and many forms of passenger transportation (e.g. air transportation) rates are subject to a competitive pressure. This means that the rate will be adjusted according to the demand and the supply. They either reflect costs directly involved with shipping (cost-of-service) or are determined by the value of the commodity (value-of-service).

Among the most significant conditions affecting transport costs and thus transport rates are:

- **Geography**. Its impacts mainly involve distance and accessibility. Distance is commonly the most basic condition affecting transport costs. The more difficult it is to trade space for a cost, the more important is the friction of distance. The friction of distance can be expressed in terms of length, time, economic costs or the amount of energy used. It varies greatly according to the type of transportation mode involved and the efficiency of specific transport routes. Landlocked countries tend to have higher transport costs, often twice as much, as they do not have direct access to maritime transportation.
- **Type of product**. Many products require packaging, special handling, are bulky or perishable. Coal is obviously a commodity that is easier to transport than fresh

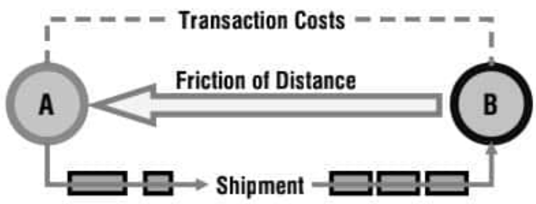


Figure 2.4 Components of transport costs

flowers as it requires rudimentary storage facilities and can be transshipped using rudimentary equipment. Insurance costs are also to be considered and are commonly a function of the value to weight ratio and the risk associated with the movement. As such, different economic sectors incur different transport costs as they each have their own transport intensity. For passengers, comfort and amenities must be provided, especially if long distance travel is involved.

- **Economies of scale**. Another condition affecting transport costs is related to economies of scale or the possibilities to apply them as the larger the quantities transported, the lower the unit cost. Bulk commodities such as energy (coal, oil), minerals and grains are highly suitable to obtain lower unit transport costs if they are transported in large quantities. A similar trend also applies to container shipping with larger containerships involving lower unit costs.
- **Energy**. Transport activities are large consumers of energy, especially oil. About 60 percent of all the global oil consumption is attributed to transport activities. Transport typically accounts for about 25 percent of all the energy consumption of an economy. The costs of several energy intensive transport modes, such as air transport, are particularly susceptible to fluctuations in energy prices.
- **Trade imbalances**. Imbalances between imports and exports have impacts on transport costs. This is especially the case for container transportation since trade imbalances imply the repositioning of empty containers that have to be taken into account in the total transport costs. Consequently, if a trade balance is strongly negative (more imports than exports), transport costs for imports tend to be higher than for exports. The same condition applies at the national and local levels where freight flows are often unidirectional, implying empty movements.
- **Infrastructures**. The efficiency and capacity of transport modes and terminals has a direct impact on transport costs. Poor infrastructures imply higher transport costs, delays and negative economic consequences. More developed transport systems tend to have lower transport costs since they are more reliable and can handle more movements.
- **Mode**. Different modes are characterized by different transport costs, since each has its own capacity limitations and operational conditions. When two or more modes are directly competing for the same market, the outcome often results in lower transport costs.
- **Competition and regulation**. This concerns the complex competitive and regulatory environment in which transportation takes place. Transport services taking place over highly competitive segments tend to be of lower cost than on segments with limited competition (oligopoly or monopoly). International competition has favored concentration in many segments of the transport industry, namely maritime and air modes. Regulations, such as tariffs, cabotage laws, labor and safety impose additional transport costs.

Types of transport costs

Mobility tends to be influenced by transport costs. Empirical evidence for passenger vehicle use underlines the relationship between annual vehicle mileage and fuel costs, implying the higher fuel costs are, the lower the mileage. At the international level, doubling of transport costs can reduce trade flows by more than 80 percent. The more affordable mobility is, the more frequent the movements and the more likely they will take place over longer distances. A wide variety of transport costs can be considered.

- **Freight on board (FOB)**. The price of a good is the combination of the factory costs and the shipping costs from the factory to the consumer. In the case of FOB, the consumer pays for the freight transport costs. Consequently, the price of a commodity will vary according to transportation costs and distance.
- **Costs–Insurance–Freight (CIF)**. Considers the price of the good, insurance costs and transport costs. It implies a uniform delivered price for all customers everywhere, with no spatially variable shipping price. The average shipping price is built into the price of a good. The CIF cost structure can be expanded to include several rate zones, such as one for local, another for the nation and another for exports.
- **Terminal costs**. Costs that are related to loading, transshipment and unloading. Two major terminal costs can be considered: loading and unloading at the origin and destination, which are unavoidable, and intermediate (transshipment) costs that can be avoided.
- **Linehaul costs**. Costs that are a function of the distance over which a unit of freight or passenger is carried. Weight is also a cost function when freight is involved. They commonly exclude transshipment costs.
- **Capital costs**. Costs applying to the physical assets of transportation, mainly infrastructures, terminals and vehicles. They include the purchase or major enhancement of fixed assets, which can often be a one-time event. Since physical assets tend to depreciate over time, capital investments are required on a regular basis for maintenance.

With an FOB cost structure, customers located nearby will have a lower overall cost than customers that are further away (Figure 2.5). Under the CIF cost structure, every consumer is charged the same price, which commonly reflects the average transport cost. Customers located close to production are "subsidizing" the costs paid by customers located further away. This price structure is common for consumer goods.

Real freight rates can be complicated to calculate for a transport company, especially when there are numerous customers. A common answer to this problem is to establish a set of geographic zones where freight rates are equal (Figure 2.6). The rate is commonly set through the CIF principle where the closest customers in a zone are partially subsidizing the furthest customers. For instance, under a zonal rate system a customer located at D1 pays the same rate as a customer located at D2. Under a distance-based system, the customer at D1 would have paid a lower rate than a customer located at D2. Many transit systems also use a zonal rate structure.

Transport providers make a variety of decisions based on their cost structure, a function of all the above types of transport costs. The role of transport companies has

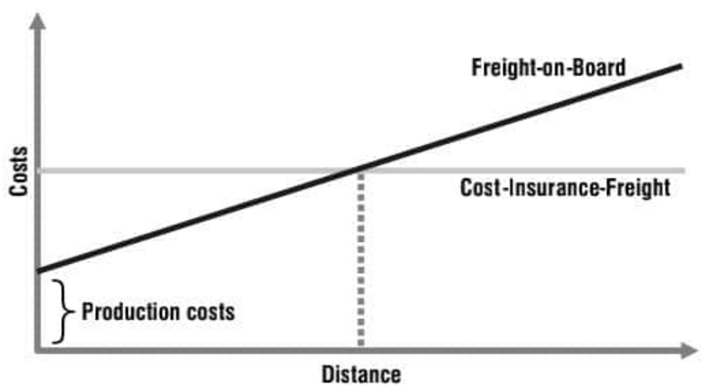


Figure 2.5 FOB and CIF transport costs

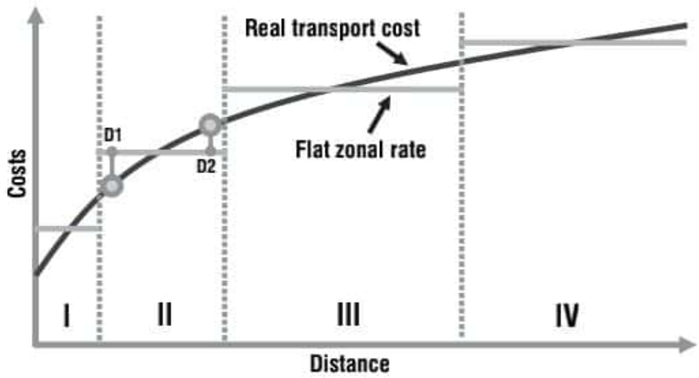


Figure 2.6 Zonal freight rates

sensibly increased in the general context of the global commercial geography. However, the nature of this role is changing as a result of reduction of transport costs but growing infrastructure costs, mainly due to greater flows and competition for land. Each transport sector must consider variations in the importance of different transport costs. While operating costs are high for air transport, terminal costs are significant for maritime transport.

Technological changes and their associated decline in transport costs have weakened the links between transport modes and their terminals. There is less emphasis on heavy industries and more importance given to manufacturing and transport services (e.g. warehousing and distribution). Indeed, new functions are being grafted on to transport activities that are henceforward facilitating logistics and manufacturing processes. Relations between terminal operators and carriers have thus become crucial, notably in containerized traffic. They are needed to overcome the physical and time constraints of transshipment, notably at ports.

The requirements of international trade gave rise to the development of **specialized and intermediary firms** providing transport services. These are firms that do not physically transport the goods, but are required to facilitate the grouping, storage and handling of freight as well as the complex paperwork and financial and legal transactions involved in international trade. Examples included freight forwarders, customs brokers, warehousing, insurance agents and banking, etc. Recently, there has been a trend to **consolidate** these different intermediate functions, and a growing proportion of global trade is now being organized by multi-national corporations that are offering door-to-door logistics services.

Concept 3 – The geography of transportation networks

Transport networks

Transportation systems are commonly represented using networks as an analogy for their structure and flows.

> The term **network** refers to the framework of routes within a system of locations, identified as nodes. A route is a single link between two nodes that are part of a larger network that can refer to tangible routes such as roads and rails, or less tangible routes such as air and sea corridors.

The territorial structure of any region corresponds to a network of all its economic interactions. The implementation of networks, however, is rarely premeditated but the consequence of continuous improvements as opportunities arise and as conditions change. They result from the influence of various strategies, such as providing access and mobility to a region, and technological developments. A transport network denotes either a permanent track (e.g. roads, rails and canals) or a scheduled service (e.g. airline, transit, train). It can be extended to cover various types of links between points along which movements can take place.

In transport geography, it is common to identify several types of transport structures that are linked with transportation networks. Network structure ranges from centripetal to centrifugal in terms of the accessibility they provide to locations. A centripetal network favors a limited number of locations while a centrifugal network does not convey any specific locational advantages. Recent decades have seen the emergence of transport hubs, a strongly centripetal form, as a privileged network structure for many types of transport services, notably for air transportation. Although hub-and-spoke networks often result in improved network efficiency, they have drawbacks linked with their vulnerability to disruptions and delays at hubs, an outcome of the lack of direct connections.

Hubs, as a network structure, allow a greater flexibility within the transport system, through a concentration of flows. For instance, on Figure 2.7, a point-to-point network involves 16 independent connections, each to be serviced by vehicles and infrastructures. By using a hub-and-spoke structure, only 8 connections are required. The main advantages of hubs are:

- Economies of scale on **connections** by offering a high frequency of services. For instance, instead of one service per day between any two pairs in a point-to-point network, four services per day could be possible.
- Economies of scale at the **hubs**, enabling the potential development of an efficient distribution system since the hubs handle larger quantities of traffic.
- Economies of scope in the use of **shared transshipment facilities**. This can take several dimensions such as lower costs for the users as well as higher quality infrastructures.

Many transportation services have adapted to include a hub-and-spoke structure. The most common examples involve air passenger and freight services which have developed such a structure at the global, national and regional levels, such as those used by UPS, FedEx and DHL. However, potential disadvantages may also occur such

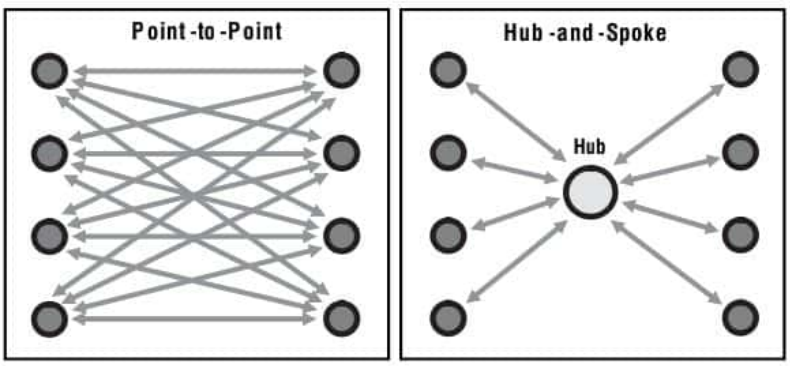


Figure 2.7 Transport networks

as additional transshipment as less point-to-point services are offered, which for some connections may involve delays and potential congestion as the hub becomes the major point of transshipment.

The efficiency of a network can be measured through graph theory and network analysis. These methods rest on the principle that the efficiency of a network depends partially on the lay-out of points and links. Obviously some network structures have a higher degree of accessibility than others, but careful consideration must be given to the basic relationship between the revenue and costs of specific transport networks. Rates thus tend to be influenced by the structure of transportation networks. Inequalities between locations can often be measured by the quantity of links between points and the related revenues generated by traffic flows. Many locations within a network have better accessibility and higher opportunities. However, economic integration processes tend to change inequalities between regions. This in turn has impacted on the structure and flows of transportation networks at the transnational level (Figure 2.8).

Prior to economic integration processes (such as a free trade agreement) networks tended to service their respective national economies with flows representing this structure. With economic integration, the structure of transportation networks is modified with new transnational linkages. Flows are also modified. In some cases, there could be a relative decline of national flows and a comparative growth of transnational flows.

The typology and topology of networks

Transportation networks, like many networks, are generally embodied as a set of locations and a set of links representing connections between those locations. The arrangement and connectivity of a network is known as its topology. Each transport network has consequently a specific topology indicating its structure. The most fundamental elements of such a structure are the network geometry and the level of connectivity. Transport networks can be classified in specific categories depending on a set of topological attributes that describe them. It is thus possible to establish a basic typology of a transport network that relates to its geographical setting, and its modal and structural characteristics.

There are many criteria that can be used to classify transportation networks (Figure 2.9). The level of abstraction can be considered with concrete network representations

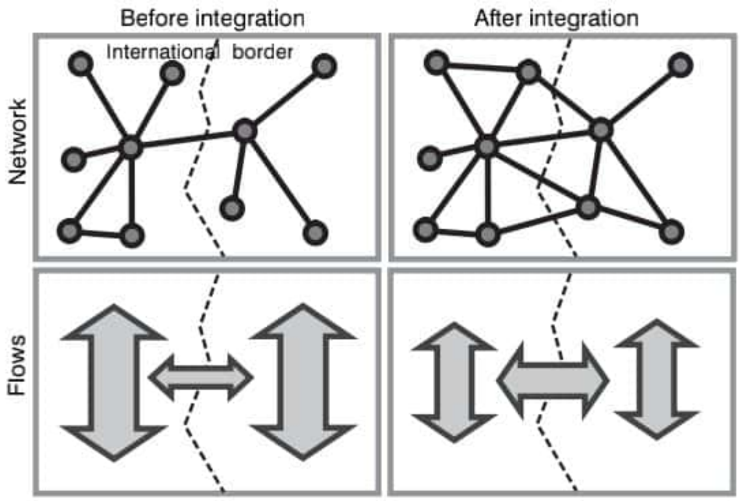


Figure 2.8 Impacts of integration processes on networks and flows

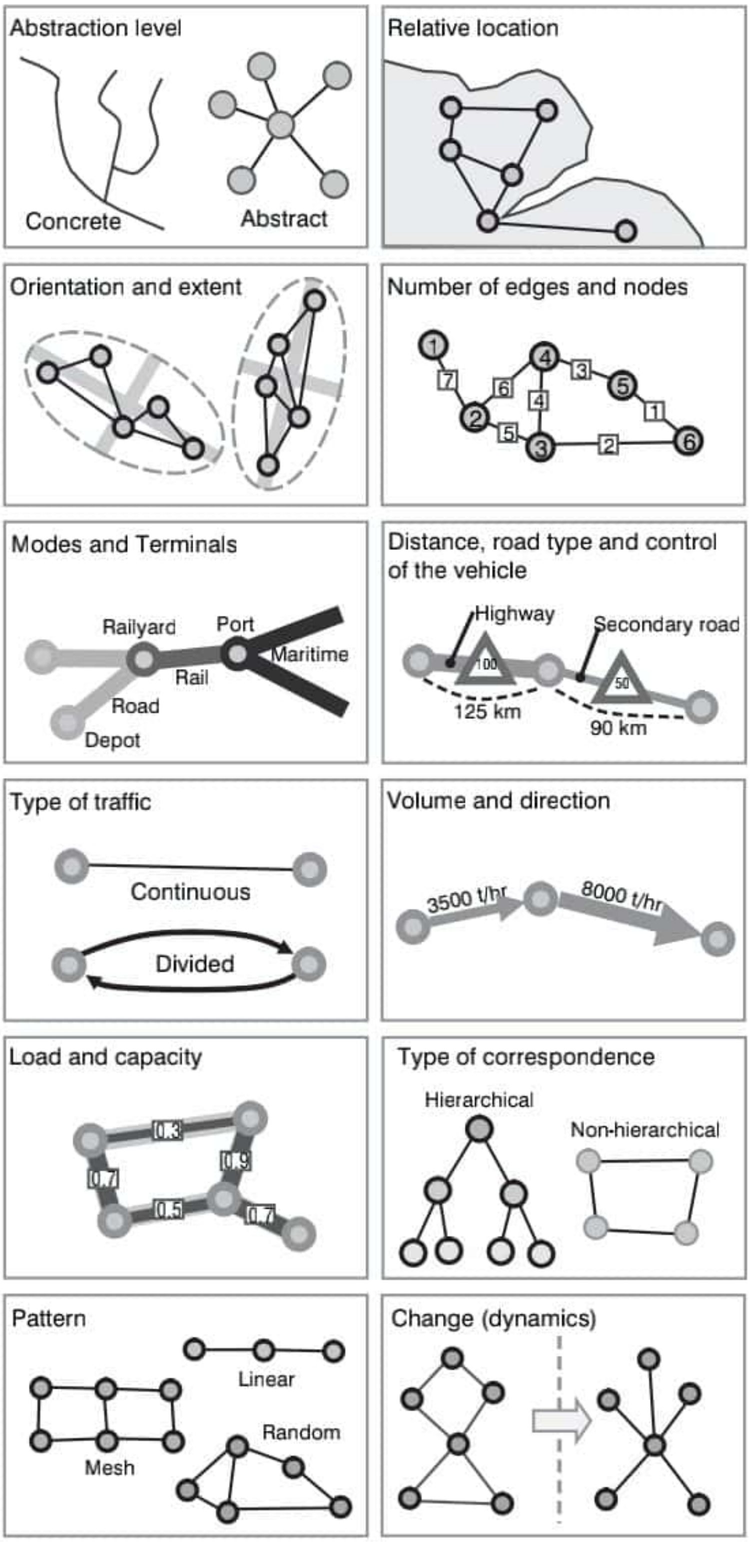


Figure 2.9 Typology of transportation networks

closely matching the reality (such as a road map) while conversely an abstract network would only be a symbolization of the nodes and flows (such as the network of an airline). Since transportation networks have a geographical setting, they can be defined according to their location relative to the main elements of a territory (such as the Rhine delta). Networks also have an orientation and an extent that approximates their geographical coverage or their market area. The numbers of nodes and edges are relevant to express the complexity and structure of transportation networks with a branch of mathematics, graph theory, developed to infer structural properties from these numbers. Since networks are the support of movements they can be considered from a modal perspective, their edges being an abstraction of routes (roads, rail links, maritime routes) and their nodes an abstraction of terminals (ports, railyards). Specific modes can further be classified in terms of types of road (highway, road, street, etc.) and level of control (speed limits, vehicle restrictions, etc.). Flows on a network have a volume and a direction, enabling to rank links by their importance and evaluate the general direction of flows (e.g. centripetal or centrifugal). Each segment and network has a physical capacity related to the volume it can support under normal conditions. The load (or volume to capacity) is the relation between the existing volume and the capacity. The closer it is to a full load (a ratio of 1), the more congested it is. The structure of some networks imposes a hierarchy reflecting the importance of each of its nodes and a pattern reflecting their spatial arrangement. Finally, networks have a dynamic where both their nodes and links can change due to new circumstances.

Further, three types of spaces on which transport networks are evolving are found. Each of these spaces represents a specific mode of territorial occupation:

- **Clearly defined and delimited**. In this case the space occupied by the transport network is strictly reserved for its usage and can be identified on a map. Ownership can also be clearly established. Major examples include road, canal and railway networks.
- **Vaguely defined and delimited**. The space of these networks may be shared with other modes and it is not the object of any particular ownership, only rights of way. Examples include air and maritime transportation networks.
- **Without definition**. The space has no tangible meaning, except for the distance it imposes. Little control and ownership are possible, but agreements must be reached for common usage. Examples are radio, television and cellular networks, which rely on specific frequencies granted by governing agencies.

Networks provide a level of transport service which is related to its costs. An optimal network would be a network servicing all possible locations but would have high capital and operational costs. Transport infrastructures are established over discontinuous networks. Therefore, operational networks are not servicing every part of the territory directly. Some compromise must often be found among a set of alternatives, considering a variety of route combinations and level of service.

Networks and space

Transportation networks illustrate the territorial organization of economic activities and the efforts incurred to overcome distance. These efforts can be measured in absolute (distance) or relative (time) terms and are proportional to the efficiency and the structure of the networks they represent. The relationships that transportation networks establish with space are related to their continuity, their topographic space and the spatial control they establish. The territory is a topological space with two or three dimensions, depending on the transport mode considered (road moves roughly over a two-dimensional space while air transport evolves over a three-dimensional space).

Figure 2.10 represents the same network topology but with different distance units of measurement between nodes. In an absolute context, distance in a network is a fixed attribute that does not change. For instance, the straight distance between New York and Boston is about 310 km, which has not changed in time and will not change. The location of the elements of such a network is also absolute and fixed. In a relative context, distance is a variable attribute that depends on numerous factors, such as technology, the mode being used and its efficiency. Under such circumstances, some nodes of the network are "closer" than others. So, while it took about 44 hours to travel between New York and Boston in around 1800, this figure is just above an hour today using air travel.

However, flows and infrastructures are linear; they have one dimension since they conceptually link two points. The establishment of a network is thus a logical outcome for a one-dimensional feature to service a territory by forming a lattice of nodes and links. In order to have such a spatial continuity in a transport network, three conditions are necessary:

- **Ubiquity**. The possibility to reach any location from any other location on the network, thus providing a general access. Access can be a simple matter of vehicle ownership or bidding on the market to purchase a thoroughfare from one location to another.
- **Fractionalization**. The possibility for a traveler or a unit of freight to be transported without depending on a group. It becomes a balance between the price advantages of economies of scale and the convenience of a dedicated service.
- **Instantaneity**. The possibility to undertake transportation at the desired or most convenient moment. There is a direct relationship between fractionalization and instantaneity since the more fractionalized a transport system is, the more likely time convenience can be accommodated.

These three conditions are never perfectly met as some transport modes fulfill them better than others. For instance, the automobile is the most flexible and ubiquitous mode for passenger transportation, but has important constraints such as low capacity and high levels of space and energy consumption. In comparison, public transit is more limited in the spatial coverage of its service, implies batch movements (bus loads, train

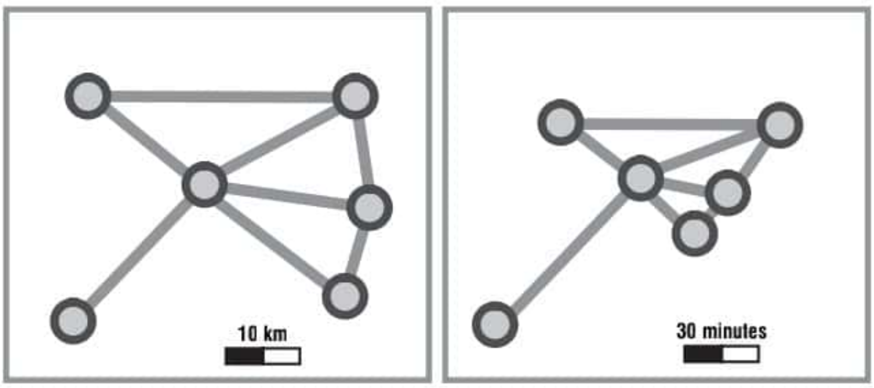


Figure 2.10 Absolute and relative distance in a network

loads, etc.) and follows specific schedules (limited instantaneity), but is more cost and energy efficient. Freight transportation also varies in its spatial continuity, ranging from massive loads of raw materials (oil and ores) that can be handled only in a limited number of ports to highly flexible parcel movements. Containerization has been a remarkable attempt to address the issue of ubiquity (the system permits intermodal movements), fractionalization (each container is a load unit) and instantaneity (units can be loaded by trucks at any time of the day and containerships make frequent port calls).

An important cause of discontinuity is linked to the spatial distribution of economic activities, notably industrial and urban, which tend to agglomerate. Congestion may also alter those conditions. Road congestion in a metropolitan area may impair ubiquity as some locations may be very difficult to reach since their accessibility is reduced. Fractionalization may also be reduced under such circumstances as people would consider public transit and carpooling and would thus move as batches. Further, as commuters cope with increasing congestion, several trips may be delayed or cancelled altogether, reducing instantaneity.

Transportation networks have always been a tool for spatial control and occupation. The Roman and Chinese empires relied on transportation networks to control their respective territories, mainly to collect taxes and move commodities and military forces. During the colonial era, maritime networks became a significant tool of trade, exploitation and political control, which was later expanded by the development of modern transportation networks within colonies. In the nineteenth century, transportation networks also became a tool of nation building and political control. For instance, the extension of railways in the American hinterland had the purpose to organize the territory, extend settlements and distribute resources to new markets. In the twentieth century, road and highways systems (such as the Interstate system in the United States and the autobahn in Germany) were built to reinforce this purpose. For the early twenty-first century, telecommunication networks have become means of spatial cohesion and interactions.

Network expansion

As transport networks expand, existing transport infrastructures are being upgraded to cope with spatial changes. Airports and ports are being transformed, expanded or relocated. In the air transport sector, emphasis is being given to integrate airports within fully-fledged multimodal transport systems, networking air with rail and road transport. In maritime transport, networks are also being modified with increasing attention being paid to:

- Exploiting sea leg routes across the Arctic Ocean
- Expanding the Panama and Suez canals
- Increasing traffic on inland maritime waterways
- Creating new inland passages between semi-enclosed or enclosed seas.

The growing competition between the sea and land corridors is not only reducing tariffs and encouraging international trade but also prompting many governments to reassess their land-based connections and seek shorter transit routes.

Existing land routes are also being extended. Passages through extremely rigorous terrain are being investigated with a view to creating fully-fledged land-based continental connections, notably through railways. These land network expansions are driven by economic globalization and inter-regional cooperation and eventually become multimodal transcontinental corridors for rail, road, pipelines and trunk telecommunications routes. But the impact of increasing world trade on land network expansion, notably the railway transportation network, is scale specific. The expansion of railways has permitted inter- and intra-continental connections in the form of:

- **Landbridges**, a land movement across a continent linking origin and destination overseas
- **Minibridges** to cover movement linking two ends of a continent
- **Microbridges** covering traffic from/to a port to an inland destination or origin.

Over the last twenty years, new rail routes in North America, Eurasia, Latin America and Africa trade routes have been developed or are being considered. There is scope for shippers to increase their trade through these new routes, particularly if rising insurance premiums, charter rates and shipping risks prompt them to opt for a land route instead of the sea route through the Suez or Panama canal. These developments linked to the integration of regional economies to the world market are part of a rationalization and specialization process of rail traffic presently occurring around the world. But the success of these rail network expansions depends on the speed of movement and the unitization of general cargo by containerization. Railways servicing ports tend more and more to concentrate on the movement of container traffic. This strategy followed by some rail transport authorities allows on the one hand an increase in the delivery of goods, and on the other hand the establishment of door-to-door services through a better distribution of goods among different transport modes.

New arterial links are constructing and reshaping new trade channels, underpinning outward cargo movements and the distribution of goods. As some coastal gateways are now emerging as critical logistics service centers that rationalize distribution systems to fit new trading patterns, the land network development and cross-border crossings throughout the world have far-reaching geopolitical implications.

Concept 4 – Transport supply and demand

Context

What are the differences between a Boeing 747, an oil tanker, a car and a bicycle? Many indeed, but they each share the common goal of fulfilling a **derived transport demand**, and they thus all fill the purpose of **supporting mobility**. Transportation is a service that must be utilized immediately and thus cannot be stored. Mobility must occur over transport infrastructures, providing a transport supply. In several instances, transport demand is answered in the simplest means possible, notably by walking. However, in some cases elaborate and expensive infrastructures and modes are required to provide mobility, such as for international air transportation.

An economic system including numerous activities located in different areas generates movements that must be supported by the transport system. Without movements infrastructures would be useless and without infrastructures movements could not occur, or would not occur in a cost-efficient manner. This interdependency can be considered according to two concepts:

- **Transport supply**. This is the expression of the capacity of transportation infrastructures and modes, generally over a geographically defined transport system and for a specific period of time. Therefore, supply is expressed in terms of infrastructures (capacity), services (frequency) and networks. The number of passengers, volume (for liquids or containerized traffic), or mass (for freight) that can be transported per unit of time and space is commonly used to quantify transport supply.

- **Transport demand**. This is the expression of the transport needs, even if those needs are satisfied fully, partially or not at all. Similar to transport supply, it is expressed in terms of number of people, volume, or tons per unit of time and space.

Transport supply is generally expressed by A_{ij}; the transport supply between locations *i* and *j* (Figure 2.11). Indirectly it combines modal supply, the capacity of a mode to support traffic, and intermodal supply, the capacity to transship traffic from one mode to the other. Transport demand is represented by T_{ij}, which expresses the transport demand between locations *i* and *j*. The potential transport demand would be the amount of traffic if transport costs were negligible. The realized transport demand, a subset of the potential transport demand, is the traffic that actually takes place, namely in view of costs between the origins and the destinations.

There is a simple statistical way to measure transport supply and demand for passengers or freight:

> The **passenger-km** (or passenger-mile) is a common measure expressing the realized passenger transport demand as it compares a transported quantity of passengers with a distance over which it gets carried. The ton-km (or ton-mile) is a common measure expressing the realized freight transport demand. Although both the passenger-km and the ton-km are most commonly used to measure realized demand, the measure can equally apply for transport supply.

For instance, the transport supply of a Boeing 747-400 flight between New York and London would be 426 passengers over 5,500 kilometers. This implies a transport supply of 2,343,000 passenger-km. In reality, there could be a demand of 450 passengers for that flight, or of 2,465,000 passenger-km, even if the actual capacity would be of only 426 passengers (if a Boeing 747-400 is used). In this case the realized demand would be 426 passengers over 5,500 kilometers out of a potential demand of 450 passengers, implying a system where demand is at 105 percent of capacity.

Transport demand is generated by the economy, which is composed of persons, institutions and industries and which generates movements of people and freight. When these movements are expressed in space they create a pattern which reflects mobility and accessibility. The location of resources, factories, distribution centers and markets is obviously related to freight movements. Transport demand can vary under

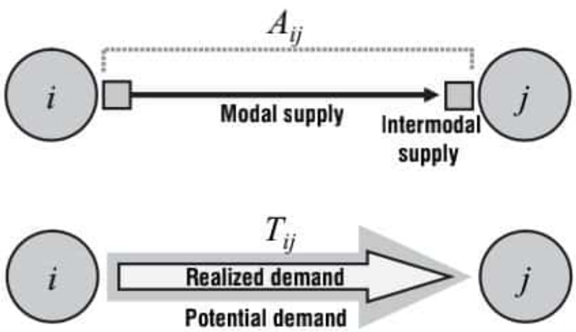


Figure 2.11 Transport supply and demand

two circumstances that are often concomitant: the quantity of passengers or freight increases or the distance over which these passengers or freight are carried increases. Geographical considerations and transport costs account for significant variations in the composition of freight transport demand between countries. For the movements of passengers, the location of residential, commercial and industrial areas tells a lot about the generation and attraction of movements.

The realized transport demand, expressed in passenger-km or ton-km, can increase for two reasons (Figure 2.12). The first is obviously that more passengers or freight are being carried. This is an outcome of growth in population, production, consumption and income. The second is a growth in the average distance over which passengers or freight are being carried. Industrial relocation, economic specialization (factors linked with globalization) and suburbanization are relevant factors behind this trend. These two factors often occur concomitantly: more passengers and freight being carried over longer distances.

Supply and demand functions

Transport supply can be simplified by a set of functions representing the main variables influencing the capacity of transport systems. These variables are different for each mode. For road, rail and telecommunications, transport supply is often dependent on the capacity of the routes and vehicles (modal supply), while for air and maritime transportation transport supply is strongly influenced by the capacity of the terminals (intermodal supply).

- **Modal supply**. The supply of one mode influences the supply of others, such as for roads where different modes compete for the same infrastructure, especially in congested areas. For instance, transport supply for cars and trucks is inversely proportional since they share the same road infrastructure.
- **Intermodal supply**. Transport supply is also dependent on the transshipment capacity of intermodal infrastructures. For instance, the maximum number of flights per day between Montreal and Toronto cannot be superior to the daily capacity of the airports of Montreal and Toronto, even though the Montreal–Toronto air corridor has potentially a very high capacity.

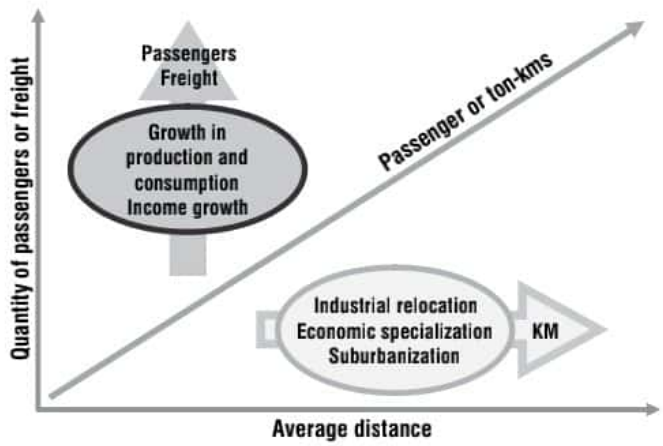


Figure 2.12 Growth factors in transport demand

Transport demand tends to be expressed at specific times that are related to economic and social activity patterns. In many cases, transport demand is stable and recurrent, which allows a good approximation in planning services. In other cases, transport demand is unstable and uncertain, which makes it difficult to offer an adequate level of service. For instance, commuting is a recurring and predictable pattern of movements, while emergency response vehicles such as ambulances are dealing with an unpredictable demand. Transport demand functions vary according to the nature of what is to be transported:

- **Passengers**. For the road and air transport of passengers, demand is a function of demographic attributes of the population such as income, age, standard of living, race and sex, as well as modal preferences.
- **Freight**. For freight transportation, the demand is a function of the nature and the importance of economic activities (GDP, commercial surface, number of tons of ore extracted, etc.) and of modal preferences. Freight transportation demand is more complex to evaluate than passengers.
- **Information**. For telecommunications, the demand can be a function of several criteria including the population (telephone calls) and the volume of financial activities (stock exchange). The standard of living and education levels are also factors to be considered.

Supply/demand relationships

Relationships between transport supply and demand continually change, but they are mutually interrelated. From a conventional economic perspective, transport supply and demand interact until an equilibrium is reached between the quantity of transportation the market is willing to use at a given price and the quantity being supplied for that price level.

Many transport systems behave in accordance with supply and demand, which are influenced by cost variations. In Figure 2.13 the demand curve assumes that if transport costs are high, demand is low as the consumers of a transport service (either freight or passengers) are less likely to use it. If transport costs are low, the demand would be high as users would get more services for the same cost. The supply curve behaves inversely. If costs are high, transport providers would be willing to supply high quantities of services since high profits are likely to arise under such circumstances. If costs are

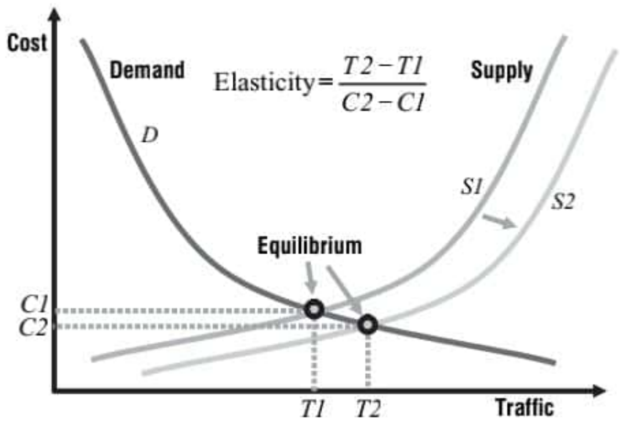


Figure 2.13 Classic transport demand/supply function

low, the quantity of transport services would be low as many providers would see little benefit of operating at a loss.

The equilibrium point represents a compromise between what users are willing to pay and what providers are willing to offer. Under such circumstances, an amount of traffic *T1* would flow at an operating cost *C1*. If because of an improvement a larger amount of service is possible for the same cost (the supply curve moves from *S1* to *S2*), a new equilibrium will be reached with a quantity of traffic *T2* at a price *C2*. Elasticity refers to the variation of the demand in accordance with the variation of the price. The higher it is, the more the traffic in a transport system is influenced by costs variations.

However, several considerations are specific to the transport sector which complexify supply/demand relationships:

- **Entry costs**. These are the costs incurred to operate at least one vehicle in a transport system. In some sectors, notably maritime, rail and air transportation, entry costs are very high, while in others such as trucking, they are very low. High entry costs imply that transport companies will consider seriously the additional demand before adding new capacity or new infrastructures (or venturing in a new service). In a situation of low entry costs, the market sees companies coming in or dropping out, fluctuating with the demand. Consequently, transport activities with high entry costs tend to be oligopolistic while transport activities with low entry costs tend to have many competitors.
- **Public sector**. Few other sectors of the economy have seen such a high level of public involvement than transportation, which creates many disruptions in conventional price mechanisms. The provision of transport infrastructures, especially roads, was massively funded by governments, namely for the sake of national accessibility and regional equity. Transit systems are also heavily subsidized, namely to provide accessibility to urban populations and more specifically to the poorest segment judged to be deprived in mobility. As a consequence, transport costs are often considered as partially subsidized. Government control (and direct ownership) was also significant for several modes, such as rail and air transportation in a number of countries. Recent years have however been characterized by less governmental involvement and deregulation.
- **Elasticity**. Refers to the variation of demand in response to a variation of cost. For example, an elasticity of –0.5 for vehicle use with respect to vehicle operating costs means that an increase of 1 percent in operating costs would imply a 0.5 percent reduction in vehicle mileage or trips. Variations in transport costs have different consequences for different modes, but transport demand has a tendency to be inelastic. While commuting tends to be inelastic in terms of costs, it is elastic in terms of time. For economic sectors where freight costs are a small component of the total production costs, variations in transport costs have limited consequences on the demand. For air transportation, especially the tourism sector, price variations have significant impacts on the demand.

The concept of elasticity is very useful to understand the economic behavior of transport supply and demand (Figure 2.14). Depending on the transport activity, a movement is linked with different elasticities. Emergencies tend to have low, if any, elasticity. Commuting also has a very low elasticity as this category of movements is related to a fundamental economic activity that provides income. This fact is underlined by empirical evidence which shows that drivers are marginally influenced by variations in the price of fuel in their commuting behavior, especially in highly motorized societies. Since work is a major, if not the only, source of income, commuting can simply not

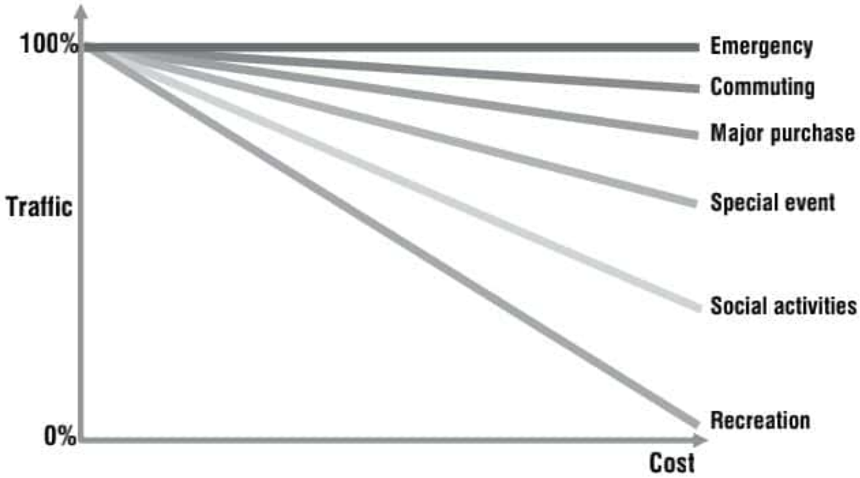


Figure 2.14 Transport elasticity by activity (Source: adapted from Victoria Transport Policy Institute 2002)

be forfeited under any circumstances short of being cost prohibitive. Activities that confer limited economic benefits tend to have high elasticities. Social and recreation-oriented movements are commonly those whose users have the least cost tolerance. Consequently, as transport costs increase, recreational movements are those which experience the fastest decline.

Generally, transport demand is variable in time and space whereas transport supply is fixed. When demand is lower than supply, transit times are stable and predictable, since the infrastructures are able to support the demand. When transport demand exceeds supply for a period in time, there is congestion with significant increases in transit times and higher levels of unpredictability. A growth of the transport demand increases the load factor of a transport network until transport supply is reached. Speed and transit times drop afterwards. The same journey can thus have different durations according to the time of the day.

Method 1 – Definition and properties of graph theory

Basic graph definition

A graph is a symbolic representation of a network and of its connectivity. It implies an abstraction of the reality so it can be simplified as a set of linked nodes.

> **Graph theory** is a branch of mathematics concerned with how networks can be encoded and their properties measured.

The goal of a graph is to represent the structure, not the appearance of a network. The conversion of a real network into a planar graph is a straightforward process which follows some basic rules: 1) The most important rule is that every terminal and intersection point becomes a node. 2) Each connected node is then linked by a straight segment.

The outcome of this abstraction, as portrayed on Figure 2.15, is the actual structure of the network. The real network, depending on its complexity, may be confusing in terms of revealing its connectivity (what is linked with what). A graph representation

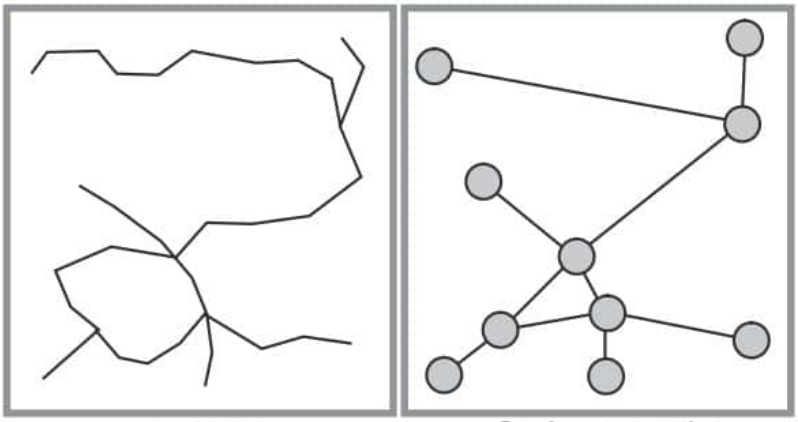


Figure 2.15 Graph representation of a real network

reveals the connectivity of a network in the best possible way. Other rules can also be applied, depending on the circumstances: 3) A node that is not a terminal or an intersection point can be added to the graph if along that segment an attribute changes. For instance, it would be recommended to represent as a node the shift from two lanes to four lanes along a continuous road segment, even if that shift does not occur at an intersection or terminal point. 4) A "dummy node" can be added for esthetical purposes, especially when it is required that the graph representation remains comparable to the real network. 5) Although the relative location of each node can remain similar to its real world counterpart (as in Figure 2.15), this is not required.

In transport geography most networks have an obvious spatial foundation, namely road, transit and rail networks, which tend to be defined more by their links than by their nodes. This is not necessarily the case for all transportation networks. For instance, maritime and air networks tend to be defined more by their nodes than by their links since the links are often not clearly defined. A telecommunication system can also be represented as a network, while its spatial expression can have limited importance and would actually be difficult to represent. Mobile telephone networks or the Internet, possibly the most complex graphs to be considered, are relevant cases of networks having a structure that can be difficult to symbolize. However, cellular phones and antennas can be represented as nodes while the links could be individual phone calls. Servers, the core of the Internet, can also be represented as nodes within a graph while the physical infrastructure between them, namely fiber optic cables, can act as links. Consequently, all transport networks can be represented by graph theory in one way or another.

The following elements are fundamental in understanding graph theory:

- **Graph**. A graph G is a set of vertexes (nodes) v connected by edges (links) e. Thus $G = (v, e)$.
- **Vertex (Node)**. A node v is a terminal point or an intersection point of a graph. It is the abstraction of a location such as a city, an administrative division, a road intersection or a transport terminal (stations, terminuses, harbors and airports).
- **Edge (Link)**. An edge e is a link between two nodes. The link (i, j) is between initial extremity i and terminal extremity j. A link is the abstraction of a transport infrastructure supporting movements between nodes. It has a direction that is commonly represented as an arrow. When an arrow is not used, it is assumed the link is bi-directional.

The graph on Figure 2.16 has the following definition: $G = (v, e)$; $v = (1, 2, 3, 4, 5)$; $e = (1, 2), (1, 3), (2, 2), (2, 5), (4, 2), (4, 3), (4, 5)$.

Sub-graph. A subset of a graph *G* where *p* is the number of sub-graphs. For instance $G' = (v', e')$ can be a distinct sub-graph of *G*. Unless the global transport system is considered as a whole, every transport network is in theory a sub-graph of another. For instance, the road transportation network of a city is a sub-graph of a regional transportation network, which is itself a sub-graph of a national transportation network.

Buckle. A link that makes a node correspond to itself.

Planar graph. A graph where every intersection of two edges is a vertex. Since this graph is located within a plane, its topology is two-dimensional.

Non-planar graph. A graph where there are no vertexes at the intersection of at least two edges. This implies a third dimension in the topology of the graph since there is the possibility of having a movement "passing over" another movement, such as for air transport. A non-planar graph has potentially many more links than a planar graph.

Links and their structures

A transportation network enables flows of people, freight or information, which occur along links. Graph theory must thus offer the possibility of representing movements as linkages, which can be considered over several aspects:

- **Connection**. A set of two nodes. Considers if a movement between two nodes is possible, whatever its direction. Knowledge of the connections within a graph means it is possible to find whether a node can be reached from another node.
- **Path**. A sequence of links that are traveled in the same direction. For a path to exist between two nodes, it must be possible to travel along an uninterrupted sequence of links. Finding all the possible paths in a graph is a fundamental attribute in measuring accessibility and traffic flows.

On graph A of Figure 2.17 there are five links [(1, 2), (2, 1), (2, 3), (4, 3), (4, 4)] and three connections [(1–2), (2–3), (3–4)]. On graph B, there is a path between 1 and 3, but on graph C there is no path between 1 and 3.

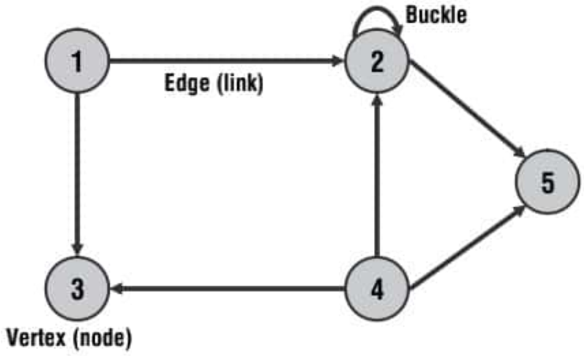


Figure 2.16 Basic graph representation of a transport network

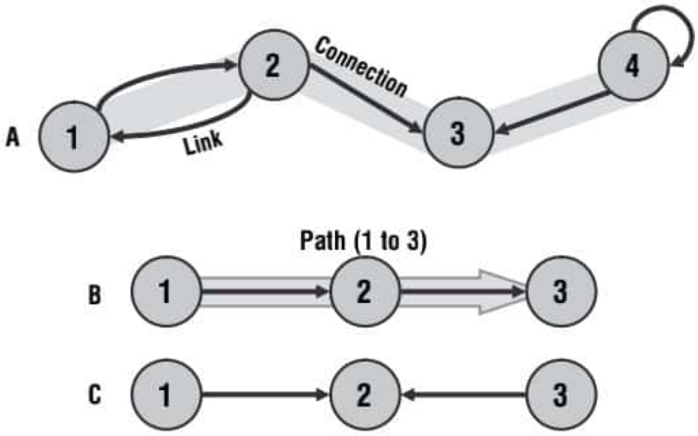


Figure 2.17 Connections and paths

Chain. A sequence of links having a connection in common with each other. Direction does not matter.

Length of a link, connection or path. Refers to the label associated with a link, a connection or a path. This label can be distance, the amount of traffic, the capacity or any attribute of that link. The length of a path is the number of links (or connections) in this path.

Cycle. A chain where the initial and terminal node is the same and which does not use the same link more than once.

Circuit. A path where the initial and terminal node corresponds. It is a cycle where all the links are traveled in the same direction. Circuits are very important in transportation because several distribution systems use circuits to cover as much territory as possible in one direction (delivery route).

On the graph of Figure 2.18, 2–3–6–5–2 is a cycle but not a circuit. 1–2–4–1 is a cycle and a circuit.

Basic structural properties

The organization of nodes and links in a graph convey a structure that can be labeled. The basic structural properties of a graph are:

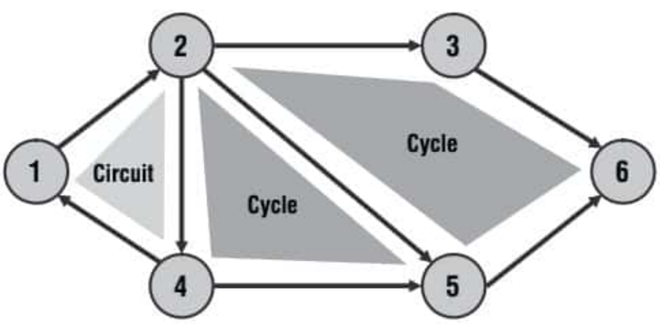


Figure 2.18 Cycles and circuits

- **Symmetry and asymmetry**. A graph is symmetrical if each pair of nodes linked in one direction is also linked in the other. By convention, a line without an arrow represents a link where it is possible to move in both directions. However, both directions have to be defined in the graph. Most transport systems are symmetrical but asymmetry can often occur as is the case for maritime (pendulum) and air services. Asymmetry is rare on road transportation networks, unless one-way streets are considered.
- **Completeness**. A graph is complete if two nodes are linked in at least one direction. A complete graph has no sub-graph.
- **Connectivity**. A complete graph is described as connected if for all its distinct pairs of nodes there is a linking chain. Direction is not important for a graph to be connected, but may be a factor for the level of connectivity. If $p > 1$ the graph is not connected because it has more than one sub-graph. There are various levels of connectivity, depending on the degree to which each pair of nodes is connected.

Two sub-graphs are **complementary** if their union results in a complete graph. Multimodal transportation networks are complementary as each sub-graph benefits from the connectivity of other sub-graphs.

- **Root**. A node r where every other node is the extremity of a path coming from r is a root. Direction is important. A root is generally the starting point of a distribution system, such as a factory or a warehouse.
- **Trees**. A connected graph without a cycle is a tree. A tree has the same number of links than nodes plus one ($e = v - 1$). If a link is removed, the graph ceases to be connected. If a new link between two nodes is provided, a cycle is created. A branch of root r is a tree where no links connect any node more than once.
- **Articulation node**. In a connected graph, a node is an articulation node if the sub-graph obtained by removing this node is no longer connected. It therefore contains more than one sub-graph ($p > 1$). An articulation node is generally a port or an airport, or an important hub of a transportation network, which serves as a bottleneck.
- **Isthmus**. In a connected graph, an isthmus is a link that, when removed, creates two sub-graphs with at least one connection.

Method 2 – Measures and indices of graph theory

Measures

Several measures and indices can be used to analyze network efficiency. Many of them were initially developed by Kansky (1963) and can be used for:

- Expressing the relationship between values and the network structures they represent
- Comparing different transportation networks at a specific point in time
- Comparing the evolution of a transport network at different points in time.

As well as the numbers of nodes and edges, three basic measures are used to define the structural attributes of a graph: the diameter, the number of cycles and the order of a node.

Diameter (*d*). The length of the shortest path between the most distanced nodes of a graph is the diameter. *d* measures the extent of a graph and the topological length between two nodes.

The diameter enables to measure the development of a network in time. The greater the diameter, the less linked a network tends to be. In the case of a complex graph, the diameter can be found with a topological distance matrix (Shimbel distance), which computes the minimal topological distance for each node pair. Graphs in which the extent remains constant, but with a higher connectivity, have lower diameter values.

Number of cycles (*u*). The maximum number of independent cycles in a graph. This number (u) is estimated through the number of nodes (v), links (e) and sub-graphs (p): $u = e - v + p$.

For trees and simple networks $u = 0$ since they have no cycles. The more complex a network is, the higher the value of u, so it can be used as an indicator of the level of development and complexity of a transport system.

Order (degree) of a node (*o*). The number of attached links in a graph. This is a simple but effective measure of nodal importance. The higher its value, the more a node is important in a graph as many links converge to it. Hub nodes have a high order, while terminal points have an order that can be as low as 1. A perfect hub would have its order equal to the summation of all the orders of the other nodes in the graph and a perfect spoke would have an order of 1.

Indexes

Indexes are more complex methods to represent the structural properties of a graph since they involve the comparison of one measure over another.

Detour index. A measure of the efficiency of a transport network in terms of how well it overcomes distance or the friction of space. The closer the detour index gets to 1, the more the network is spatially efficient. Networks with a detour index of 1 are rarely, if ever, seen and most networks would fit on an asymptotic curve getting close to 1, but never reaching it.

$$DI = {}^{DT}\!/_{DD}$$

For instance, the straight distance (DD) between two nodes may be 40 km but the transport distance (DT; real distance) is 50 km. The detour index is thus 0.8 (40/50). The complexity of the topography is often a good indicator of the level of detour.

Network density. Measures the territorial handhold of a transport network in terms of km of links (L) per square kilometer of surface (S). The higher it is, the more a network is developed.

Pi index. The relationship between the total length of the graph $L(G)$ and the distance along its diameter $D(d)$. It is called the pi index because of its similarity with the constant pi (3.14), which expresses the ratio between the circumference and the

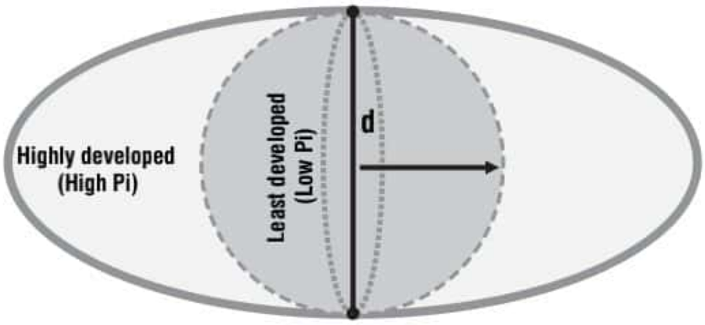


Figure 2.19 Pi index and the shape of transportation networks

diameter of a circle. A high index shows a developed network. It is a measure of distance per units of diameter and an indicator of the shape of a network.

Figure 2.19 provides an abstraction between the diameter (d; vertical axis) and length of the network (horizontal axis). A low pi index is linked with a low level of network development and a high pi index is linked with a more extensively developed network.

Eta index. Average length per link. Adding new nodes will cause the eta index to decrease as the average length per link declines.

$$\eta = \frac{L(G)}{e}$$

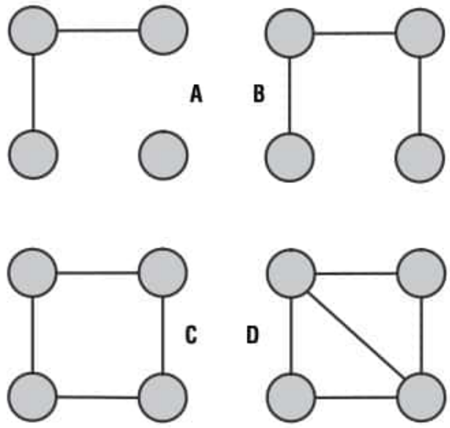
Theta index. Measures the function of a node, that is the average amount of traffic per intersection. The higher theta is, the greater the load of the network.

$$\theta = \frac{Q(G)}{v}$$

Beta index. Measures the level of connectivity in a graph and is expressed by the relationship between the number of links (e) over the number of nodes (v). Trees and simple networks have beta index values of less than 1. A connected network with one cycle has a value of 1. More complex networks have a value greater than 1. In a network with a fixed number of nodes, the higher the number of links, the higher the number of paths possible in the network. Complex networks have a high beta index.

The four graphs of Figure 2.20 are of growing connectivity. Graphs A and B are not fully connected and their beta value is less than 1. Graph C is connected and has a beta value of 1. Graph D is even more connected with a beta value of 1.25.

Alpha index. A measure of connectivity which evaluates the number of cycles in a graph in comparison with the maximum number of cycles. The higher the alpha index, the more a network is connected. Trees and simple networks will have a value of 0. A value of 1 indicates a completely connected network. The alpha index measures the level of connectivity independently of the number of nodes. It is very rare for a network to have an alpha value of 1, because this would imply very serious redundancies.



$$\beta = \frac{e}{v}$$

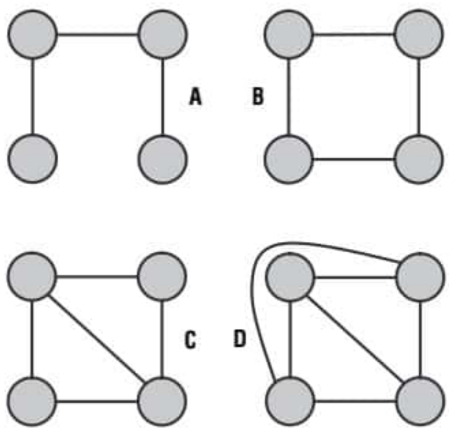
	e	v	Beta
A	2	4	0.5
B	3	4	0.75
C	4	4	1.0
D	5	4	1.25

Figure 2.20 Beta index

The graphs of Figure 2.21 have a growing level of connectivity. While graph A has no cycles, graph D has the maximum possible number of cycles for a planar graph.

Gamma index (*g*). A measure of connectivity that considers the relationship between the number of observed links and the number of possible links. The value of gamma is between 0 and 1, where a value of 1 indicates a completely connected network and is extremely unlikely in reality. The gamma index is an efficient way to measure the progression of a network in time.

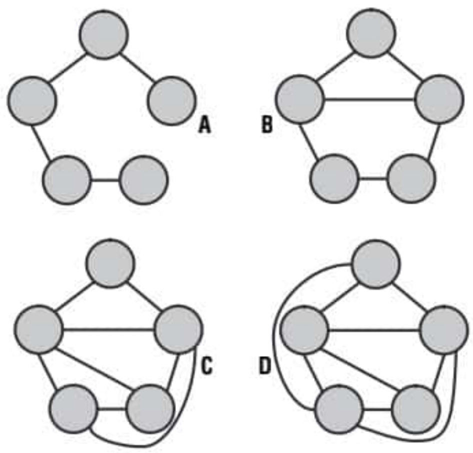
The graphs of Figure 2.22 have a growing level of connectivity with graph D having the maximum number of links (10) and a gamma index of 1.0.



$$\alpha = \frac{u}{2v-5}$$

	$u(e-v+p)$	$2v$-5	Alpha
A	0	3	0.0
B	1	3	0.33
C	2	3	0.66
D	3	3	1.0

Figure 2.21 Alpha index



$$\gamma = \frac{e}{3(v-2)}$$

	e	$3(v-2)$	Gamma
A	4	9	0.44
B	6	9	0.66
C	8	9	0.88
D	9	9	1.0

Figure 2.22 Gamma index

Method 3 – Network data models

Nature and utility

Graph theory gives a topological and mathematical representation of the nature and structure of transportation networks. However, graph theory can be expanded for the analysis of real world transport networks by encoding them in an information system. In the process, a digital representation of the network is created. This digital representation is highly complex, since transportation data is often multimodal, can span several local, national and international jurisdictions and has different logical views depending on the particular user.

It is thus becoming increasingly relevant to use a data model where a transportation network can be encoded, stored, retrieved, modified, analyzed and displayed. Obviously, Geographic Information Systems have received a lot of attention over this issue since they are among the best tools to store and use network data models. Network data models are an implicit part of many GIS. There are four basic application areas of network data models:

- **Topology**. The core purpose of a network data model is to provide an accurate representation of a network as a set of links and nodes. Topology is the arrangement of nodes and links in a network. Of particular relevance are the representations of location, direction and connectivity. Even if graph theory aims at the abstraction of transportation networks, the topology of a network data model should be as close as possible to the real world structure it represents. This is especially true for the usage of network data models in a GIS.

Figure 2.23 represents the basic topology of an urban transport network composed of linked nodes. It has been encoded into a network data model to represent the reality as closely as possible, both topologically and geographically. Topologically, each node has been encoded with the connectivity it permits, such as whether a left turn is possible or not (although this attribute is not displayed here). Further, a direction has been encoded in each link (directional or bi-directional) to represent one-ways. Geographically, each node is located at a coordinate which matches, within a tolerated accuracy, the actual intersection it represents. In addition, the links between each node have been decomposed into several segments (not implicitly shown) to respect the positional accuracy of the road they represent.

- **Cartography**. Allows the visualization of a transport network for the purpose of reckoning and simple navigation and serves to indicate the existence of a network. Different elements of the network can have a symbolism defined by some of their attributes. For instance, a highway link may be symbolized as a thick line with a label such as its number, while a street may be symbolized as an unlabeled simple line. The symbolized network can also be combined with other features such as landmarks to provide a better level of orientation to the user. This is commonly the case for road maps used by the general public.

By using attributes encoded in the network data model, such as road type, each segment can be displayed to reflect its importance. For instance, the cartographic representation of a network data model on Figure 2.24 displays three road classes (highway, main street and street) differently. Descriptive labels for the most important elements and directional signs for one-ways have also been added. To enrich the cartographic message, additional layers of information have been added, namely landmarks (City Hall, Central Park and a college campus). Nodal attributes can also have a cartographic utility, such as displaying whether an intersection has traffic lights.

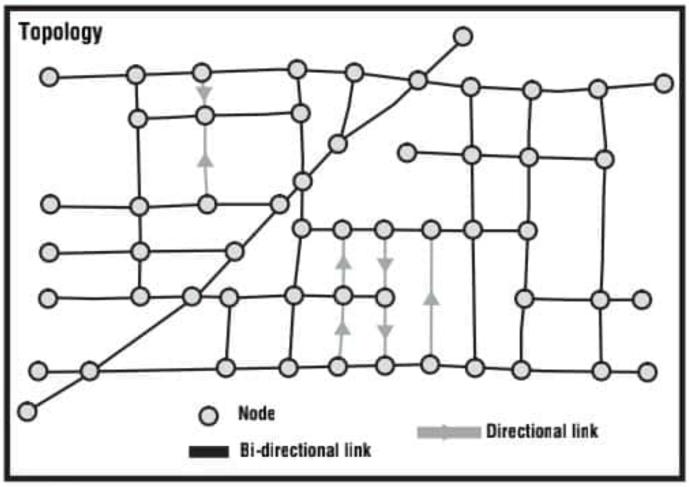


Figure 2.23 Topology of a network data model

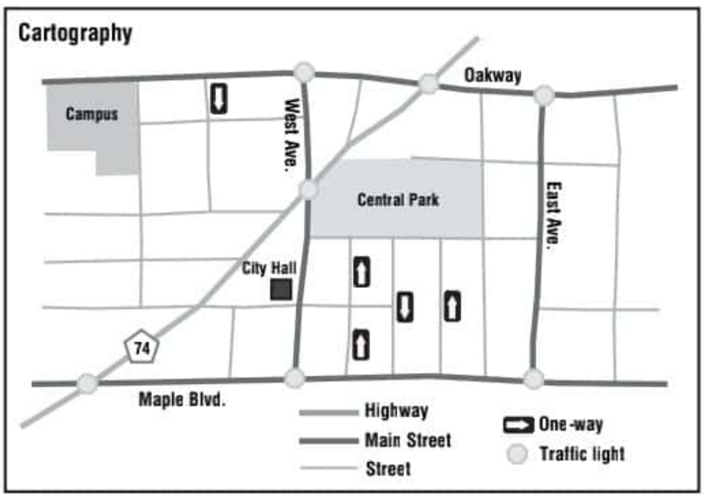


Figure 2.24 Cartography of a network data model

- **Geocoding**. Transportation network models can be used to derive a precise location, notably through a linear referencing system. For instance, the great majority of addresses are defined according to a number and a street. If address information is embedded in the attributes of a network data model, it becomes possible to use this network for geocoding and to pinpoint the location of an address, or any location along the network, with reasonable accuracy.

Geocoding is possible if a linear referencing system is embedded in a network data model. One of the most common linear referencing systems is the address system, where each link has a corresponding street name and address range. The address range of Figure 2.25 illustrates even (right side) and odd (left side) addresses, very common attributes in most network data models such as TIGER (developed by the US Census Bureau). For instance, finding the approximate location of the address "197 East Ave." would first imply querying the network data model to find all the links that have "East Ave." as a name attribute. Then, the appropriate address range is found and the location interpolated. "197" corresponds to the 191–209 address range, located on the left side of East Ave. Its approximate location would be at 1/3 [1 – (209 – 197)/(209 – 191)] of the length of the link that has the 191–209 address range. The same procedure can be applied to the address "188 East Ave.", which in this case would be located at 1/4 of the length of the link that has the 172–210 address range.

- **Routing and assignment**. Network data models may be used to find optimal paths and assign flows with capacity constraints in a network. While routing is concerned with the specific behavior of a limited number of vehicles, traffic assignment is mainly concerned with the system-wide behavior of traffic in a transport network. This requires a topology in which the relationship of each link with other intersecting segments is explicitly specified. Impedance measures (e.g. distance) are also attributed to each link and will have an impact on the chosen path or on how flows are assigned in the network. Routing and traffic assignment at the continental level is generally simple since small variations in impedance are of limited consequences. Routing and traffic assignment in an urban area is much more complex as in determining the impedance of a route it must consider stop signs, traffic lights and congestion.

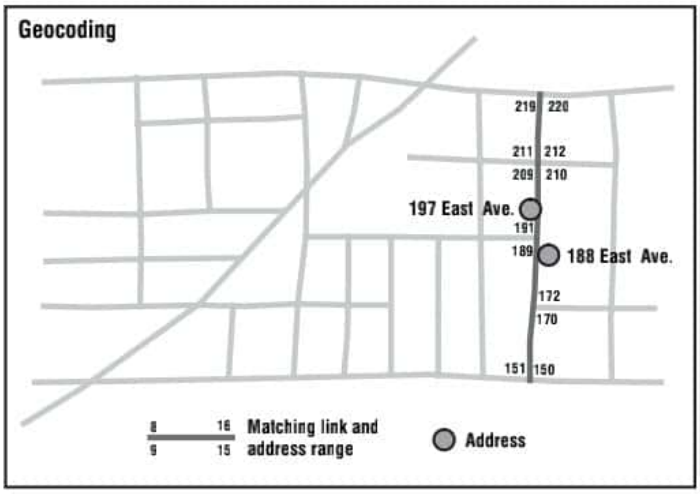


Figure 2.25 Geocoding in a network data model

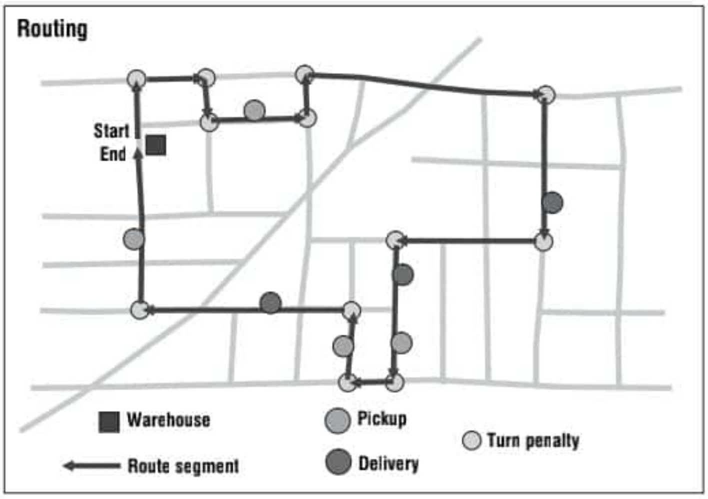


Figure 2.26 Routing in a network data model

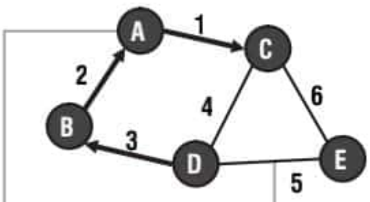
Routing in a network data model can be simulated if impedance is attributed to links and nodes. For links, impedance is often characterized by travel time, while turn penalties are often used to characterize impedance at nodes, that is how difficult (if possible) it is to turn in one direction, as opposed to another. The network in Figure 2.26 represents a typical routing "traveling salesperson" type of problem. Starting and ending at a warehouse, a delivery truck has a set of deliveries and pickups to perform. The locations of those pickup and delivery points could have been derived from address matching (geocoding). Considering link and node (turn penalties) impedance attributes that are encoded in the network data model, it is possible to plot an optimal route minimizing travel time that would satisfy basic constraints related to the start and end points, pickup and delivery points, as well as link and turn penalty impedances.

Basic representation

Constructing the geometry of a network depends on the mode and the scale being investigated. For urban road networks, information can be extracted from aerial photographs or topographic maps. Air transport networks are derived from airport locations (nodes) and scheduled flights between them (links). Two fundamental tables are required in the basic representation of a network data model that can be stored in a database:

- **Node table**. This table contains at least three fields: one to store a unique identifier and the others to store the node's *x* and *y* coordinates. Although these coordinates can be defined by any Cartesian reference system, longitudes and latitudes would insure an easy portability to a GIS.
- **Link table**. This table also contains at least three fields: one to store a unique identifier, one to store the node of origin and one to store the node of destination. A fourth field can be used to state whether or not the link is unidirectional.

Once those two tables are relationally linked, a basic network topology can be constructed and all the indexes and measures of graph theory can be calculated. Attributes such as the connectivity and the Shimbel matrix can also easily be derived from the link table. This basic representation enables to define the topology of networks as structured by graph theory.



Nodes		
ID	Lat	Long
A	40.42345	-75.1245
B	40.31021	-75.2510
C	40.41882	-74.9124
D	40.25908	-75.0031
E	40.28990	-74.7893

Links			
ID	From	To	One-Way
1	A	C	Yes
2	B	A	Yes
3	D	B	Yes
4	D	C	No
5	D	E	No
6	C	E	No

Figure 2.27 Relational database representation of a simple network

A network can be represented by using two tables, one defining nodes and the other defining links (Figure 2.27). The three core elements (fields) of a nodes table are a unique identifier and locational attributes in a coordinate system, such as latitude and longitude values. On Figure 2.27, coordinates are in decimal degrees, meaning that the location of these nodes can be directly imported into a GIS. Additional attributes can also be included in this table.

The links table has four core elements (fields). The first is a unique identifier for each link, the next two are the nodes of origin and destination of the link and the fourth is a directional tag indicating whether or not the link is unidirectional. An alternative would be to assume that all links are unidirectional and define each of them implicitly. This would require the addition of three new records if the directional tag field is not used (C–D, E–D and E–C). However, this would involve serious redundancies on a complex network. As for the nodes table, additional attributes can be included, such as name, number of lanes, maximum speed, etc.

Both the nodes and links tables have little value if they are considered individually, as a network is the combination of the information contained on both tables. A way to combine these tables is by building a relational join between them. In the above example, a relational join can be established between the [From] and [To] fields of the links table with the [ID] field of the nodes table. The resulting relational database contains the basic topological elements of the network.

Many efforts have been made to create comprehensive transportation network databases to address a wide variety of transportation problems ranging from public transit to package distribution. Initially, these efforts were undertaken within transportation network optimization packages (e.g. EMME/2, TransCAD) which created topologically sound representations. Many of these representations were however geographically inaccurate and had limited visual and geocoding capabilities. Using a network data model for the purposes of cartography, geocoding and routing requires further developments.

Layer-based approach

Most conventional GIS data models separate information in layers, each representing a different class of geographical elements symbolized as points, lines and polygons. As such, a network data model must be constructed with the limitation of having points and lines in two separate layers; thus the layer-based approach. Further, an important

4 Transportation modes

Transportation modes are an essential component of transport systems since they are the means by which mobility is supported. Geographers consider a wide range of modes that may be grouped into three broad categories based on the medium they exploit: land, water and air. Each mode has its own requirements and features, and is adapted to serve the specific demands of freight and passenger traffic. This gives rise to marked differences in the ways the modes are deployed and utilized in different parts of the world. Recently, there is a trend towards integrating the modes through intermodality and linking the modes ever more closely into production and distribution activities. At the same time, however, passenger and freight activity is becoming increasingly separated across most modes.

Concept 1 – A diversity of modes

Transport modes are the means by which people and freight achieve mobility. They fall into one of three basic types, depending on what surface they travel over: land (road, rail and pipelines), water (shipping), and air. Each mode is characterized by a set of technical, operational and commercial characteristics (see Figure 4.1).

Road transportation

This has become the dominant land transport system today. Automobiles, buses and trucks require a road bed. Such infrastructures are moderately expensive to provide, but there is

Vehicle	Capacity	1 Barge Equivalency
Barge	1500 Tons 52,500 Bushels 453,600 Gallons	1
15 barges on tow	22,500 Tons 787,500 Bushels 6,804,000 Gallons	0.06
Hopper car	100 Tons 3,500 Bushels 30,240 Gallons	15
100 car train unit	10,000 Tons 350,000 Bushels 3,024,000 Gallons	0.15
Semi-trailer truck	26 Tons 910 Bushels 7,865 Gallons	57.7

Figure 4.1 Performance comparison for selected freight modes

a wide divergence of costs, from a gravel road to a multi-lane urban expressway. Because vehicles have the means to climb moderate slopes, physical obstacles are less important than for some other land modes. Most roads are provided as a public good by governments, while the vast majority of vehicles are owned privately. The capital costs, therefore, are shared, and do not fall as heavily on one source as is the case for other modes.

All road transport modes have **limited abilities to achieve scale economies**. This is due to the size constraints imposed by governments and also by the technical and economic limits of the power sources. In most jurisdictions, trucks and buses have specific weight and length restrictions which are imposed for safety reasons. In addition, there are serious limits on the traction capacities of cars, buses and trucks because of the considerable increases in energy consumption that accompany increases in the weight of the unit. For these reasons the carrying capacities of individual road vehicles are limited.

Road transport, however, possesses significant advantages over other modes. The **capital cost** of vehicles is relatively small. This produces several key characteristics of road transport. Low vehicle costs make it comparatively easy for new users to gain entry, which helps ensure that the trucking industry, for example, is highly competitive. Low capital costs also ensure that innovations and new technologies can diffuse quickly through the industry. Another advantage of road transport is the **high relative speed of vehicles**, the major constraint being government-imposed speed limits. One of its most important attributes is the **flexibility of route choice**, once a network of roads is provided. Road transport has the unique opportunity of providing door-to-door service for both passengers and freight. These multiple advantages have made cars and trucks the modes of choice for a great number of trip purposes, and have led to the market dominance of cars and trucks for short-distance trips.

The success of cars and trucks has given rise to a number of serious problems. Road congestion has become a feature of most urban areas around the world (see Chapters 7 and 10). In addition, the mode is behind many of the major environmental externalities linked to transportation (see Chapter 8). Addressing these issues is becoming an important policy challenge at all levels of jurisdiction, from the local to the global (see Chapter 9).

Rail transportation

Railways require tracks along which the locomotives and rail cars move. The initial **capital costs** are high because the construction of rail tracks and the provision of rolling stock are expensive. Historically, the investments have been made by the same source (either governments or the private sector). These expenditures have to be made before any revenues are realized and thus represent important **entry barriers** that tend to limit the number of operators. It also serves to delay innovation, compared with road transport, since rail rolling stock has a service life of at least twenty years.

Railway routing is affected by topography because locomotives have limited capacities to mount gradients. As a result, railways either avoid important natural barriers or overcome them by expensive engineering solutions. An important feature of rail systems is the width of the rails. The standard **gauge** of 1.4351 meters has been adopted in many parts of the world, across North America and most of Western Europe for example. But other gauges have been adopted in other areas. This makes integration of rail services very difficult, since both freight and passengers are required to change from one railway system to the other. As attempts are being made to extend rail services across continents and regions, this is an important obstacle, as for example between France and Spain, Eastern and Western Europe, and between Russia and China. The potential of the Eurasian land bridge is limited in part by these gauge differences. Other factors that inhibit the movement of trains between different countries include signaling and electrification standards. These are particular problems for the European Union where the lack of "interoperability" of the rail systems between the member states is a factor limiting the wider use of the rail mode.

The ability of trains to haul large quantities of goods and significant numbers of people over long distances is the mode's primary asset. Once the cars have been assembled or the passengers have boarded, trains can offer a **high speed – high capacity service**. It was this feature that led to the train's pre-eminence in opening the interior of the continents in the nineteenth century, and is still its major asset. Passenger service is effective where population densities are high. Freight traffic is dominated by bulk cargo shipments, agricultural and industrial raw materials in particular. Rail transport is a "green" system, in that its consumption of energy per unit load per km is lower than road modes.

Although sometimes identified as a mode that enjoyed its heyday during the nineteenth century, rail transport is enjoying a resurgence because of technological advances in the latter part of the twentieth century. In passenger transport this has come about through significant breakthroughs in speed. For instance, in Europe and Japan high-speed rail systems reach speeds up to 515 km/hr. This gives rail a competitive advantage over road transport and even with air transport over short and medium distances (see Figure 4.2). Japan saw the first comprehensive development of a high-speed train system, notably used along the Tokyo–Osaka corridor in 1964. By the 1990s, the usage of the system had peaked, in part because of competition from air transport. Europe has been the region where the adoption of the high-speed train has been the most significant since the 1990s. Close to a half of all the world's high-speed passengers-km are now occurring in Europe. South Korea is the latest country to build a high-speed rail system along the Seoul–Pusan corridor, which was inaugurated in 2004.

Unit trains, where trains are made up of wagons carrying one commodity-type only, allow scale economies and efficiencies in bulk shipments, and double stacking has greatly promoted the advantages of rail for container shipments. Rail transport is also enjoying a resurgence as a mode for commuters in many large cities.

Pipelines

Pipelines are an extremely important and extensive mode of land transport, although very rarely appreciated or recognized by the general public, mainly because they are

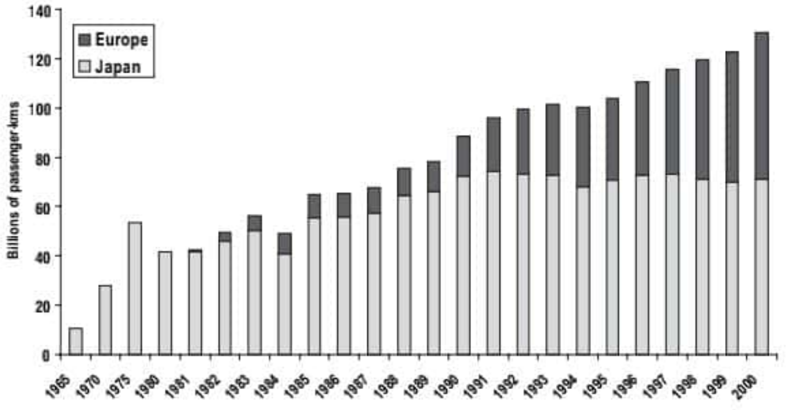


Figure 4.2 Development of high-speed train traffic, Europe and Japan, 1965–2000

buried underground (or under the sea as in the case of gas pipelines from North Africa to Europe). In the USA, for example, there are 409,000 miles of pipelines that carry 17 percent of all ton/miles of freight. The longest oil pipeline is the TransSiberian, extending over 9,344 km to Western Europe from the Russian arctic oilfields in eastern Siberia. Two main products dominate pipeline traffic: oil and gas, although locally pipelines are significant for the transport of water, and in some rare cases for the shipment of dry bulk commodities, such as coal in the form of slurry.

Pipelines are almost everywhere designed for a specific purpose only, to carry one commodity from one location to another. They are built largely with **private capital** and because the system has to be in place before any revenues are generated, represent a significant capital commitment. They are effective in transporting large quantities of products where no other feasible means of transport (usually water) is available. Pipeline routes tend to link isolated areas of production with major centers of refining and manufacture in the case of oil, or major populated areas, as in the case of natural gas.

The **routing** of pipelines is largely indifferent to terrain, although environmental concerns frequently delay approval for construction. In sensitive areas, particularly in arctic/sub-arctic areas where the pipes cannot be buried because of permafrost, the impacts on migratory wildlife may be severe, and be sufficient to deny approval, as was the case of the proposed McKenzie Valley pipeline in Canada in the 1970s. The 1,300 km long Trans Alaskan pipeline was built under difficult conditions and is above the ground for most of its path. Geo-political factors play a very important role in the routing of pipelines that cross international boundaries. Pipelines from the Middle East to the Mediterranean have been routed to avoid Israel, and new pipelines linking Central Asia with the Mediterranean are being routed in response to the ethnic and religious mosaic of the republics in the Caucasus.

Pipeline construction costs vary according to the diameter of the pipe and increase proportionally with the distance and with the viscosity of the fluid (need for pumping stations). **Operating costs** are very low, however, and as mentioned above, pipelines represent a very important mode for the transport of liquid and gaseous products. One major disadvantage of pipelines is the inherent inflexibility of the mode. Once built (usually at great expense), expansion of demand is not easily adjusted to. There exist specific limits to the carrying capacity. Conversely, a lessening of supply or demand will produce a lowering of revenues that may affect the viability of the system. A further limit arises out of geographical shifts in production or consumption, in which a pipeline having been built from one location to another may not be able to easily adjust to changes. For example, the refineries in Montreal, Canada, were served by a pipeline from Portland, Maine in order to receive shipments year-round because of ice on the St. Lawrence River. In the 1980s a pipeline from western Canada was built to provide domestic crude oil at a time when the price of the international supply was escalating. Since then the Portland pipeline has been lying idle.

Water transportation

Shipping exploits the water routes that cross oceans as well as rivers and lakes. Many of the oceanic routes are in international waters and are provided at no cost to the users. In many coastal and inland waters too shipping lanes are "free", although national regulations may exclude foreign vessels from **cabotage** trade. Physical barriers represent a particular problem for shipping in two areas. First are the sections of inland waterways where water depths and/or rapids preclude navigation. The second is where land barriers separate seas. In both cases canals can provide access for shipping, but they may be tolled. An example of the first type is the St. Lawrence Seaway, while the Suez and Panama canals are examples of the latter. Thus, except for canals, shipping enjoys rights of way that are at no cost to the users. Complementing this advantage are the relatively **low operating costs** of ships. Ships have the ability to carry large volumes with small energy consumption and limited manpower requirements. Shipping, therefore, is a mode that can offer very low rates compared with other modes.

Even if maritime transportation has experienced remarkable improvements in safety and reliability, maritime routes are still hindered by dominant winds, currents and general weather patterns. The North Atlantic and the North Pacific (50 to 60 degrees north) are subject to heavy wave activity during the winter that sometimes impairs navigation, and may cause ships to follow routes at lower latitudes, thereby increasing the route lengths (see Figure 4.3). During the summer monsoon season (April to October), navigation may become more hazardous on the Indian Ocean and the South China Sea.

Rivers may not be useful for commercial navigation if their orientations do not correspond to the directions of transport demand. Thus, many of the major rivers of Russia flow north–south, while the main trade and passenger flows are east–west. Shallow draught and extensive obstacles, such as rapids, may also limit navigation. However, many rivers, such as the Rhine or the Chang Jiang, are significant arteries for water transport because they provide access from the oceans to inland markets (see Figure 4.3).

Shipping has traditionally faced two **drawbacks**. It is slow, with speeds at sea averaging 15 knots (26 km/h). Secondly, delays are encountered in ports where loading and unloading takes place. The latter may involve several days of handling. These drawbacks are particularly constraining where goods have to be moved over short distances or where shippers require rapid service deliveries. There are four broad types of ships employed around the world.

- **Passenger vessels** can be further divided into two categories: passenger ferries, where people are carried across relatively short bodies of water in a shuttle-type service, and cruise ships, where passengers are taken on vacation trips of various durations, usually over several days. The former tend to be smaller and faster vessels, the latter are usually very large capacity ships.
- **Bulk carriers** are ships designed to carry specific commodities, and are differentiated into liquid bulk and dry bulk vessels. They include the largest vessels afloat. The largest tankers, the Ultra Large Crude Carriers (ULCC) are up to 500,000 deadweight

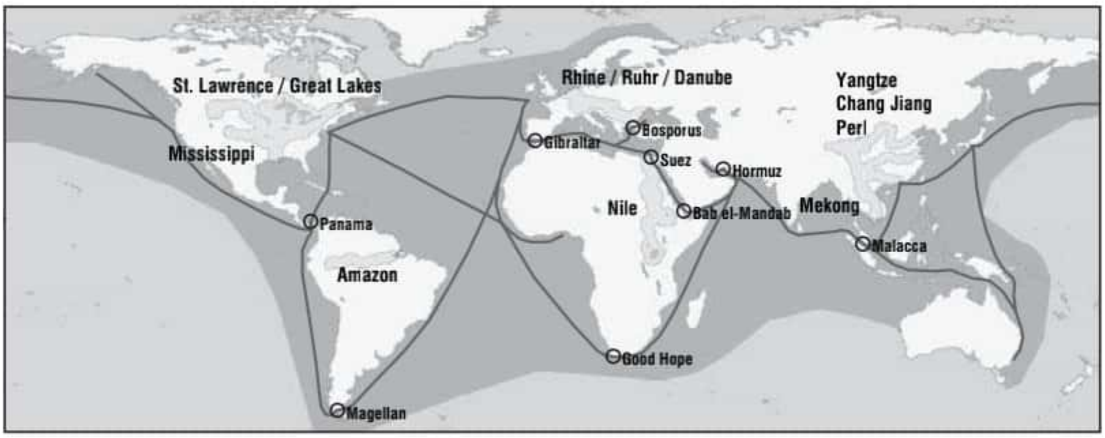


Figure 4.3 Domains of maritime transport

tons (dwt), with the more typical size being between 250,000 and 350,000 dwt; the largest dry bulk carriers are around 350,000 dwt, while the more typical size is between 100,000 and 150,000 dwt.

- **General cargo ships** are vessels designed to carry non-bulk cargoes. The traditional ships were less than 10,000 dwt, because of extremely slow loading and off-loading. More recently these vessels have been replaced by container ships that because they can be loaded more efficiently are becoming much larger, with 80,000 dwt being the largest today.
- **Roll on – roll off (RORO) vessels**, which are designed to allow cars, trucks and trains to be loaded directly on board. Originally appearing as ferries, these vessels are used on deep-sea trades and are much larger than the typical ferry. The largest are the car carriers that transport vehicles from assembly plants to the main markets.

The distinctions in vessel types are further differentiated by the **kinds of services** on which they are deployed. Bulk ships tend to operate either on a regular schedule between two ports or on voyage basis. In the latter case the ship may haul cargoes between different ports based on demand. General cargo vessels operate on liner services, in which the vessels are employed on a regular scheduled service between fixed ports of call, or as tramp ships, where the vessels have no schedule and move between ports based on cargo availability.

An important feature of the economics of shipping is the **capital costs**. Because of their size, ships represent a significant capital outlay. Cruise ships represent the most expensive class of vessels, with the Queen Mary 2 costing $800 million, but even container ships represent initial capital outlays of $75 million. The annual cost of servicing the purchase of these vessels represents the largest single item of operating expenditures, typically accounting for over half of the annual operating costs. Container shipping requires the deployment of many vessels to maintain a regular service (14 ships in the case of a typical Far East – Europe service), which is a severe constraint on the entry of new players. On the other hand, older second-hand vessels may be purchased for much smaller amounts, and sometimes the purchase price can be easily covered by a few successful voyages. In some regards, therefore, the **shipping industry is quite open** and historically has provided opportunities for entrepreneurs to accumulate large fortunes. Many of the largest fleets are in private hands, owned by individuals or by family groups.

The shipping industry has a very international character. This is reflected particularly in terms of ownership and flagging. The ownership of ships is very broad. While a ship may be owned by a Greek family or a US corporation, it may be flagged under another nationality. **Flags of convenience** are means by which ship owners can obtain lower registration fees, lower operating costs and fewer restrictions.

The share of open registry ships operated under a flag of convenience grew substantially after World War II. They accounted for 5 percent of world shipping tonnage in 1950, 25 percent in 1980, and 45 percent in 1995. The usage of a flag of convenience refers to a national owner choosing to register one or more vessels in another nation in order to avoid higher regulatory and manning costs. This enables three types of advantages for the ship owners:

- **Regulation**. Under maritime law, the owner is bound to the rules and regulations of the country of registration, which also involves requisitions in situation of emergency (war, humanitarian crisis, etc.). Being subject to less stringent regulations commonly confers considerable savings in operating costs.

- **Registry costs**. The state offering a flag of convenience is compensated according to the ship's tonnage. Registry costs are on average between 30 and 50 percent lower than those of North America and Western Europe.
- **Operating costs**. Operating costs for open registry ships are from 12 to 27 percent lower than for traditional registry fleets. Most of the savings come from lower manning expenses. Flags of convenience have much lower standards in terms of salary and benefits.

The countries with the largest registered fleets offer flags of convenience (Panama, Liberia, Greece, Malta, Cyprus and the Bahamas) and have very lax regulations (see Figure 4.4). Ship registry is a source of additional income for these governments. Even the landlocked country of Mongolia offers ship registry services.

An important historic feature of oceanic liner transport is the operation of conferences. These are formal agreements between companies engaged on particular trading routes. They fix the rates charged by the individual lines, operating for example between Northern Europe and the East Coast of North America, or eastbound between Northern Asia and the West Coast of North America. Over the years in excess of 100 such conference arrangements have been established. While they may be seen as anti-competitive, the conference system has always escaped prosecution from national anti-trust agencies. This is because they are seen as a mechanism to stabilize rates in an industry that is inherently unstable, with significant variations in supply of ship capacity and market demand. By fixing rates, exporters are given protection from swings in prices, and are guaranteed a regular level of service provision (Brooks, 2000). Firms compete on the basis of service provision rather than price. A new form of inter-firm organization has emerged in the container shipping industry since the mid-1990s. Because the costs of providing ship capacity to more and more markets are escalating beyond the means of many carriers, many of the largest shipping lines have come together by forming strategic alliances with erstwhile competitors. They offer joint services by pooling vessels on the main commercial routes. In this way they are each able to commit fewer ships to a particular service route, and deploy the extra ships on other routes that are maintained outside the alliance. The alliance services are marketed separately, but operationally involve close cooperation in selecting ports of call and in establishing schedules. The alliance structure has led to significant developments in route alignments and economies of scale of container shipping (Slack, 2004).

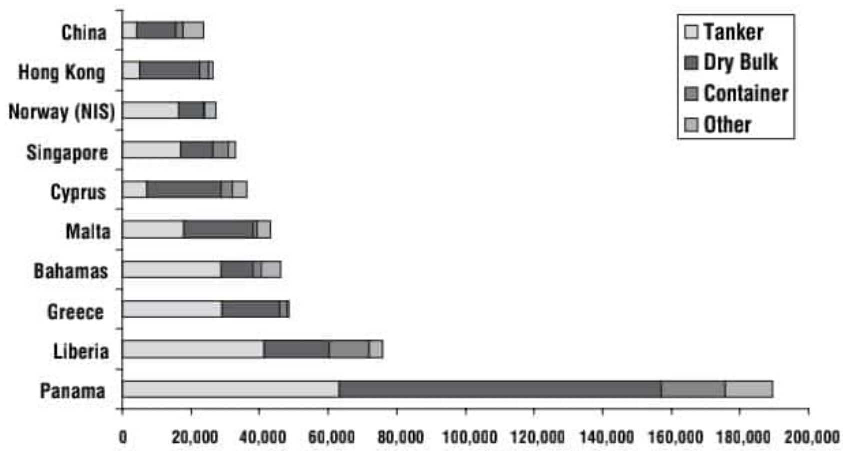


Figure 4.4 Tonnage by country of registry, 2003

Air transportation

Air transport, compared with other modes, has the obvious **advantage of speed**. This feature has served to offset many of its limitations, among which operating costs, fuel consumption and limited carrying capacities are the most significant. Technology has worked to overcome some of the constraints, most notably the growth of capacity, in which aircraft will soon be capable of transporting 500 passengers or 100 tons of freight. Technology has also significantly extended the range of aircraft, so that while 40 years ago aircraft were just beginning to be capable of crossing the Atlantic without stopping at intermediate places such as Newfoundland, they are now capable of making trips of up to 18 hours duration. Surprisingly, the speed of commercial aircraft has not progressed since the 1960s, when the prospect of supersonic speed was being anticipated with the development of the Anglo-French Concorde, which was removed from service in 2003. Figure 4.5 shows the ranges of three major categories of jet planes:

- **Regional**. The airbus A320, with a range of 3,700 km, was designed to service destinations within a continent. From New York, most of North America can be reached. This range can be applied to the European continent, South America, East Asia and Africa. This type of aircraft is also used for high demand regional services needing several flights a day, enabling to improve the quality of service.
- **International**. The Boeing 777-100, with a range of 7,400 km, can link one continent to another. From New York, it is possible to reach Western Europe and most of South America.
- **Intercontinental**. The Boeing 747-400, with a range of 11,400 km, can reach from New York any destination around the world except Australia, South and Southeast Asia. Japan is within range.

Air transport makes use of air space that theoretically gives it **great freedom of route choice**. While the mode is less restricted than land transport to specific rights of way, it is nevertheless much more constrained than might be supposed. In part this is due to physical conditions, in which aircraft seek to exploit (or avoid) upper atmospheric

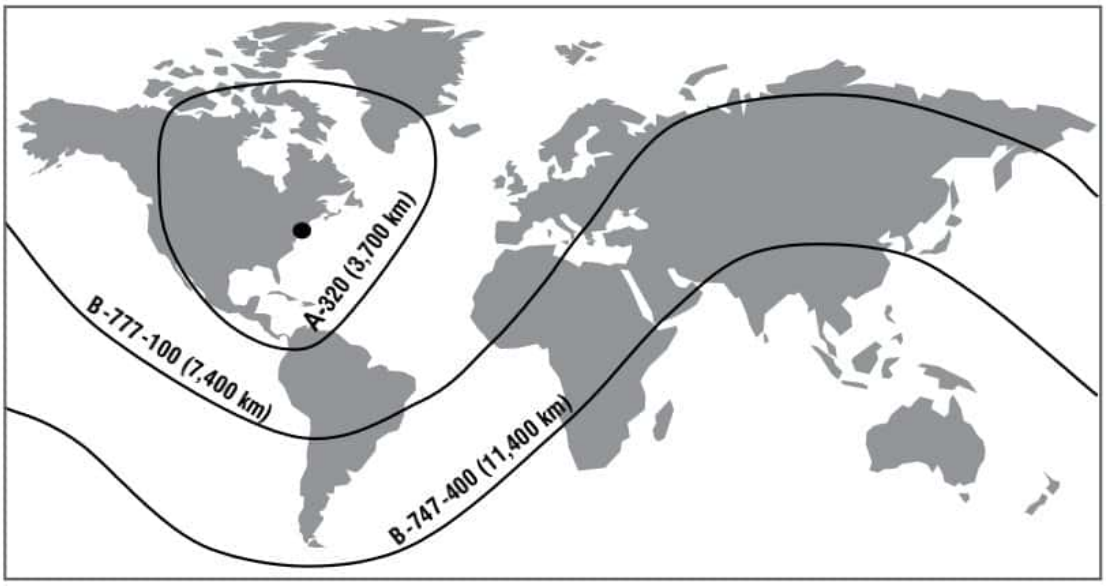


Figure 4.5 Range from New York of different modern commercial jet planes

winds, in particular the jet stream, to enhance speed and reduce fuel consumption. In addition, specific corridors have been established in order to facilitate navigation and safety. Strategic and political factors also influence route choice. For example, the flights of South African Airways were not allowed to over-fly many African nations during the apartheid period, and Cubana Airlines has been routinely prohibited from over-flying the USA.

Like maritime transport, the airline industry is highly **capital intensive**. For instance, a new Boeing 747-400, used for high-volume and long-distance travel, costs approximately $200 million, depending on the configuration, and a new Boeing 737-800, used for regional flights, costs about $60 million. However, unlike the maritime sector, air transportation is labor intensive, with limited room to lower labor requirements, although many airlines are now trying to reduce labor costs by cutting salaries and benefits. The industry has become a powerful factor of development, generating globally more than $700 billion in added value and creating more than 21 million jobs.

The initial development of air transportation took place in the 1920s and 1930s, not always for commercial reasons (Graham, 1995). It was seen as a means of providing a national air mail service (US) and of establishing long-haul air services to colonies and dependencies (UK and France). Airline companies were set up to provide these national goals, a trend that continued in the post-colonial period of the 1950s to the 1970s, as many African, Asian and Caribbean nations created their own airline companies while reserving them for specific markets and for specific routes. By convention, an air space exclusively belongs to the country under it, and this has led to significant government control over the industry.

Traditionally, an airline needs the approval of the governments of the various countries involved before it can fly in or out of a country, or even across another country without landing. Prior to World War II, this did not present too many difficulties since the range of commercial planes was limited and air transport networks were in their infancy and nationally oriented. In 1944, an International Convention was held in Chicago to establish the framework for all future bilateral and multilateral agreements for the use of international air spaces. Five **freedom rights** were designed, but a multilateral agreement went only as far as the first two freedoms (right to over-fly and right to make a technical stop).

Freedoms are not automatically granted to an airline as a right, they are privileges that have to be negotiated. All other freedoms have to be negotiated by **bilateral agreements**, such as the 1946 agreement between the United States and the UK, which permitted limited "fifth freedom" rights. The 1944 Convention has been extended since then, and as shown in Figure 4.6 there are currently nine different freedoms:

- **First Freedom**. The right to fly from a home country over another country (A) en-route to another (B) without landing. Also called the transit freedom.
- **Second Freedom**. The right for a flight from a home country to land in another country (A) for purposes other than carrying passengers, such as refueling, maintenance or emergencies. The final destination is country B.
- **Third Freedom**. The right to carry passengers from a home country to another country (A) for purpose of commercial services.
- **Fourth Freedom**. The right to fly from another country (A) to a home country for purpose of commercial services.

The Third and Fourth Freedoms are the basis for direct commercial services, providing the rights to load and unload passengers, mail and freight in another country.

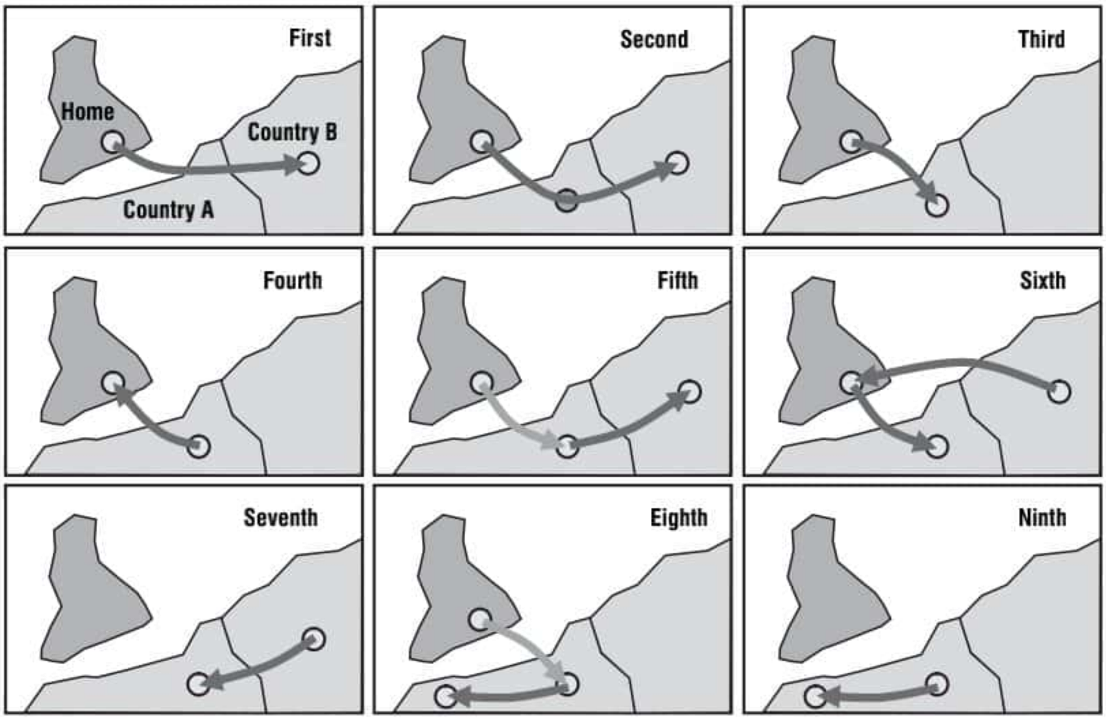


Figure 4.6 Air freedom rights

- **Fifth Freedom**. This freedom enables airlines to carry passengers from a home country to another intermediate country (A), and then fly on to a third country (B) with the right to pick up passengers in the intermediate country. Also referred to as "beyond right". This freedom is divided into two categories: Intermediate Fifth Freedom Type is the right to carry from the third country to the second country. Beyond Fifth Freedom Type is the right to carry from the second country to the third country.
- **Sixth Freedom**. Not formally part of the original 1944 convention, it refers to the right to carry passengers between two countries (A and B) through an airport in the home country. With the hubbing function of most air transport networks, this freedom has become more common, notably in Europe (London, Amsterdam).
- **Seventh Freedom**. Covers the right to operate a passenger service between two countries (A and B) outside the home country.
- **Eighth Freedom**. Also referred to as "cabotage" privileges. It involves the right to move passengers on a route from a home country to a destination country (A) that uses more than one stop along which passengers may be loaded and unloaded.
- **Ninth Freedom**. Also referred to as "full cabotage" or "open-skies" privileges. It involves the right of a home country to move passengers within another country (A).

In the 1970s, the perspective changed and air transport was increasingly seen as just another transport service. Market forces were considered to be the mechanism for fixing prices and it became widely accepted that airline companies should be given freedom within national markets to decide the nature and extent of their services, while the role of governments should be limited to operational and safety regulations. In the United States, the Air Deregulation Act of 1978 put an end to fixed markets and opened the industry to competition. This liberalization process has spread to many other countries, although with important local distinctions. Many of the former private firms in the USA and many former state-owned airlines elsewhere that were heavily protected and subsidized, went bankrupt or have been absorbed by larger ones. Many new carriers have emerged, with several **low-cost carriers** such as Ryan Air and South-West Air, having achieved industry leadership. Internationally, air transport is still dominated by bi-lateral agreements between nations (Graham, 1995).

As in the case of ocean shipping, there has been a significant development of **alliances** in the international airline industry. The alliances are voluntary agreements to enhance the competitive positions of the partners. Members benefit from greater scale economies, a lowering of transaction costs and a sharing of risks, while remaining commercially independent. The first major alliance was established in 1989 between KLM and North West Airlines. The "Star" alliance was initiated in 1993 between Lufthansa and United Airlines. In 1996, British Airlines and American Airlines formed the "One World" alliance. Other national carriers have joined different alliance groupings. They cooperate on scheduling, code sharing, equipment maintenance and schedule integration. It permits airlines that may be constrained by bi-lateral regulations to offer a global coverage (Agusdinata and de Klein, 2002).

Prior to deregulation movements (end of 1970s–early 1980s), many airline services were taking place on a point-to-point basis. Figure 4.7 shows two airline companies servicing a network of major cities. A fair amount of direct connections exists, but mainly at the expense of the frequency of services and high costs (if not subsidized). Also, many cities are serviced, although differently, by the two airlines and connections are likely to be inconvenient. With deregulation, a system of hub-and-spoke networks emerges as airlines rationalize the efficiency of their services. A common consequence is that each airline assumes dominance over a **hub** and services are modified so the two hubs are connected to several spokes. Both airlines tend to compete for flights between their hubs and may do so for specific spokes, if demand warrants it. However, as this network matures, it becomes increasingly difficult to compete at hubs as well as at spokes, mainly because of economies of agglomeration. As an airline assumes

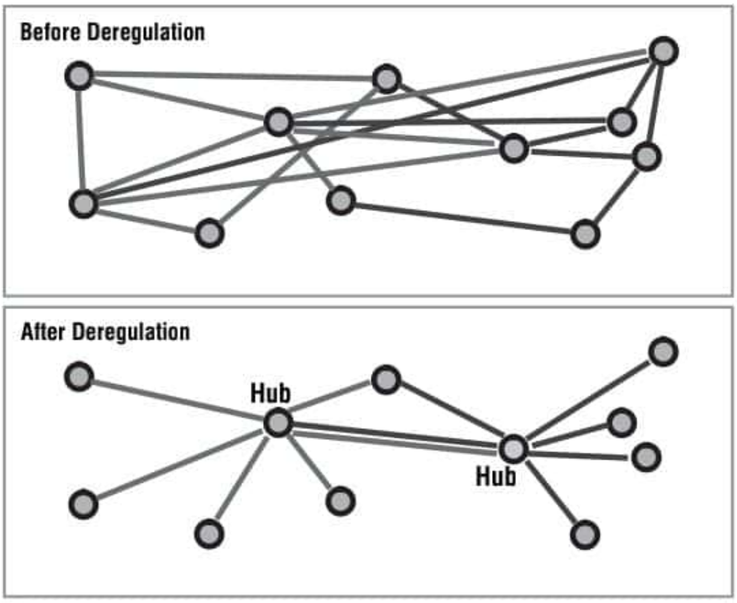


Figure 4.7 Airline deregulation and hub-and-spoke networks

dominance of a hub, it reaches oligopolistic (if not monopolistic) control and may increase airfares for specific segments. The advantage of such a system for airlines is the achievement of a regional market dominance and higher plane loads, while passengers benefit from better connectivity (although delays for connections and changing planes are more frequent) and lower costs.

Air transport is extremely important for both passenger and freight traffic. In 2000, 1.4 billion passengers traveled by air transport, representing the equivalent of 23 percent of the global population. Passenger traffic is made up of business travelers and the general public, many of whom are holiday-makers. Air transport is a very significant factor in the growth of international tourism. Figure 4.8 indicates the continued domination of US carriers in passenger transport.

In 2000, 30 million tons of freight was transported, a figure that represents one third of the value of all international trade. This freight traffic is made up of electronics, parcels and parts with a high value-to-weight ratio that are at the heart of contemporary just-in-time and of flexible production systems. Freight is carried in the belly-hold of passenger airplanes, and provides supplementary income for airline companies. However, with the growth of the freight traffic an increasing share is being accounted for by all-cargo planes and specialized air freight carriers, either as independent companies or as separate ventures by conventional passenger carriers (see Figure 4.9).

Modal competition

A general analysis of transport modes reveals that they each possess key operational and commercial advantages and properties. Modes can **compete or complement** each other in terms of cost, speed, reliability, frequency, safety, comfort, etc. Cost is one of the most important considerations in the choice of mode. Because each mode has its own price/performance profile, the actual competition between the modes depends primarily upon the distance traveled, the quantities that have to be shipped and the value of the goods. Thus, while maritime transport might offer the lowest variable costs, over short distances and for small bundles of goods, road transport tends to be most competitive. A critical factor is the **terminal cost structure** for each mode, where the costs (and delays) of loading and unloading the unit impose fixed costs that are incurred independent of

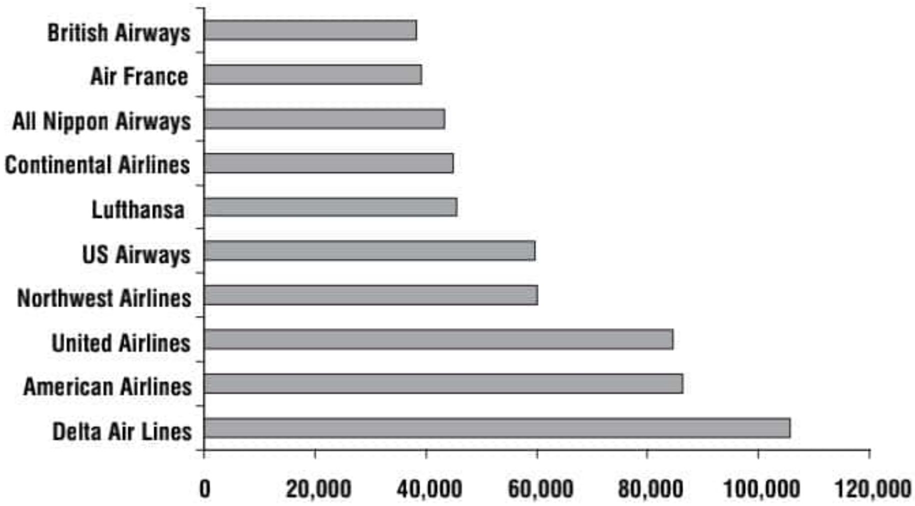


Figure 4.8 World's 10 largest passenger airlines, 2000 (in 1,000 passengers) (Source: IATA, World Air Transport Statistics)

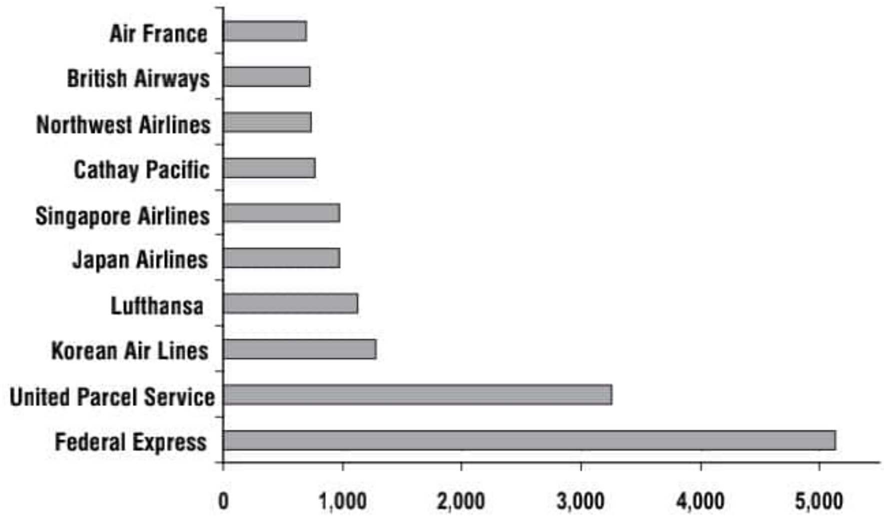


Figure 4.9 World's 10 largest freight airlines, 2000 (in 1,000 tonnes)

the distance traveled (see Chapter 5). As shown in Figure 4.10, different transportation modes have different cost functions. Road, rail and maritime transport have respectively *C1*, *C2*, and *C3* cost functions. While road has a lower cost function for short distances, its cost function climbs faster than rail and maritime cost functions. At a distance *D1*, it becomes more profitable to use railway transport than road transport while from a distance *D2*, maritime transport becomes more advantageous. Point *D1* is generally located between 500 and 750 km of the point of departure while *D2* is near 1,500 km.

With increasing levels of income the propensity for people to travel rises. At the same time, international trade in manufactured goods and parts has increased. These trends in travel demand act differentially upon the modes. The modes that offer faster and more reliable services gain over modes that offer a lower cost, but slower, alternative. For passenger services, rail has difficulty in meeting the competition of road transport over short distances and aircraft for longer trips. For freight, rail and shipping have suffered from competition from road and air modes for high value shipments. While shipping, pipelines and rail still perform well for bulkier shipments, intense competition over the last thirty years has seen road and air modes capture an important market share of the high revenue-generating goods. Figure 4.11 shows the modal split in one major market region, where trucks dominate, particularly in terms of value of shipments.

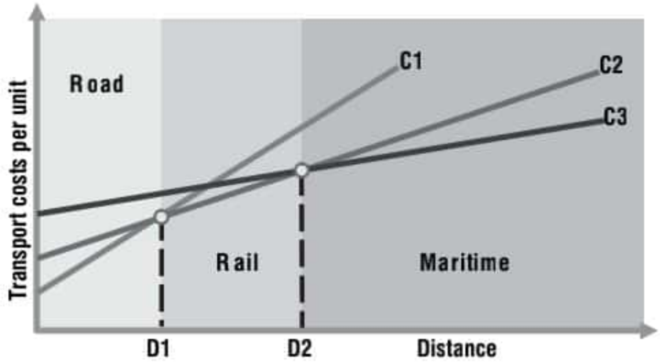


Figure 4.10 Distance, modal choice and transport cost

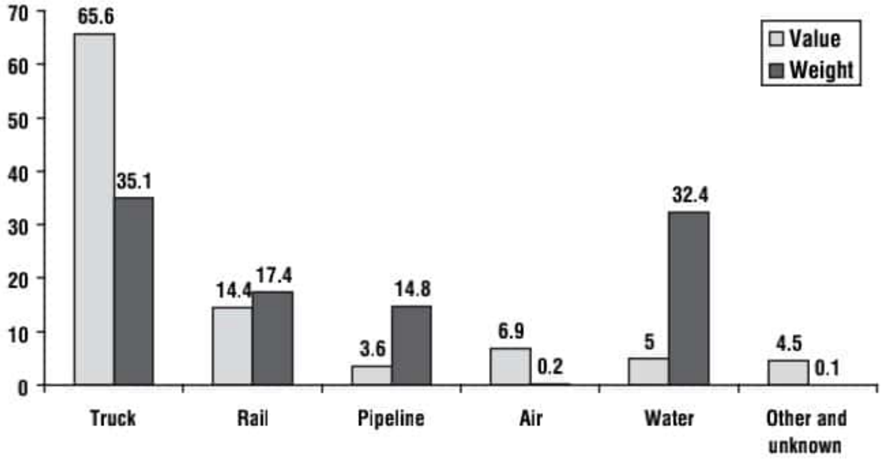


Figure 4.11 Modal shares of US–NAFTA-partner merchandise trade, 2000

There are important geographical variations in modal competition. The availability of transport infrastructures and networks varies enormously. Some regions possess many different modes that in combination provide a range of transport services that ensure an efficient commercial environment. In many parts of the world, however, there are only limited services, and some important modes may be absent altogether. This limits the choices for people and shippers, and acts to limit accessibility. People and freight are forced to use the only available modes that may not be the most economic for the nature of the demand. Goods may not be able to find a market, and people's mobility may be impaired.

For these reasons, transport provision is seen as a major factor in **economic development** (see Chapter 3). Areas with limited modal choices tend to be among the least developed. The developed world, on the other hand, possesses a wide range of modes that can provide services to meet the needs of society and the economy.

Concept 2 – Intermodal transportation

The nature of intermodalism

Competition between the modes has tended to produce a transport system that is segmented and un-integrated. Each mode has sought to exploit its own advantages in terms of cost, service, reliability and safety. Carriers try to retain business by **maximizing the line-haul** under their control. All the modes saw the other modes as competitors, and were viewed with suspicion and mistrust. The lack of integration between the modes was also accentuated by public policy that has frequently barred companies from owning firms in other modes (as in the United States before deregulation), or has placed a mode under direct state monopoly control (as in Europe). **Modalism** was also favored because of the difficulties of transferring goods from one mode to another, thereby incurring additional terminal costs and delays.

The use of several modes of transport has frequently occurred as goods are shipped from the producer to the consumer. When several modes are used this is referred to as **multimodal transport**. Within the last forty years efforts have been made to integrate separate transport systems through **intermodalism**. What distinguishes intermodal from multimodal transport is that the former involves the use of at least two different modes in a trip from origin to destination under a single transport rate. Intermodality enhances the economic performance of a transport chain by using the modes in the most productive manner. Thus, the line-haul economies of rail may be exploited for long distances, with the efficiencies of trucks providing local pick up and delivery. The key is that the entire trip is seen as a whole, rather than as a series of legs, each marked by an individual operation with separate sets of documentation and rates.

Figure 4.12 illustrates two alternatives to freight distribution. The first is a conventional point-to-point multimodal network where origins (A, B and C) are independently linked to destinations (D, E and F). In this case, two modes (road and rail) are used. The second alternative involves the development of an integrated intermodal transport network. Traffic converges at two transshipment points, rail terminals, where loads are consolidated. This can result in higher load factors and/or higher transport frequency, especially between terminals. Under such circumstances, the efficiency of such a network mainly resides in the transshipment capabilities of transport terminals.

The emergence of intermodalism has been brought about in part by technology (Muller, 1995). Techniques for transferring freight from one mode to another have facilitated intermodal transfers. Early examples include piggyback (TOFC: trailers on flat cars), where truck trailers are placed on rail cars, and LASH (lighter aboard ship), where river barges are placed directly on board sea-going ships. The major development undoubtedly has been the container, which permits easy handling between modal systems. Containers have become the most important component for rail and maritime intermodal transportation.

While handling technology has influenced the development of intermodalism, another important factor has been the changes in public policy. Deregulation in the United States in the early 1980s liberated firms from government control. Companies were no longer prohibited from owning across modal types, and there developed a strong impetus towards intermodal cooperation. Shipping lines, in particular, began to offer integrated rail and road service to customers. The advantages of each mode could be exploited in a seamless system. Customers could purchase the service to ship their products from door to door, without having to concern themselves about modal barriers. With one bill of lading clients can obtain one through rate, despite the transfer of goods from one mode to another (Hayuth, 1987).

The provision of through **bills of lading** in turn necessitated a revolution in organization and information control. At the heart of modern intermodalism are data handling, processing and distribution systems that are essential to ensure the safe, reliable and cost-effective control of freight movements across several modes. **Electronic Data**

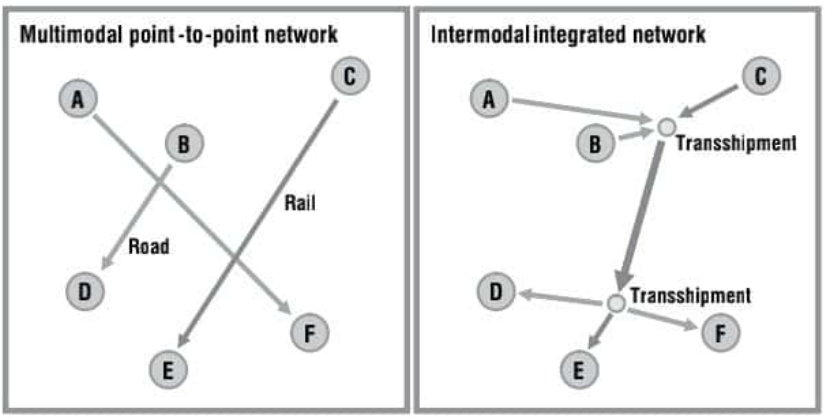


Figure 4.12 Multimodal and intermodal transportation

Interchange (EDI) is an evolving technology that is helping companies and government agencies (customs documentation) to cope with an increasingly complex global transport system.

Intermodalism, the container and maritime transport

Intermodalism originated in maritime space, with the development of the container in the late 1960s and has since spread to integrate other modes. It is not surprising that the maritime sector should have been the first mode to pursue containerization. It was the mode most constrained by the time taken to load and unload the vessels. Containerization permits the mechanized handling of cargoes of diverse types and dimensions that are placed into boxes of standard dimensions. In this way, goods that might have taken days to be loaded or unloaded from a ship can now be handled in a matter of minutes (Slack, 1998).

One of the keys to the success of the container is that the International Standards Organization (ISO) very early on established base dimensions. The reference size is the 20-foot box, 20 feet long, 8 feet high and 8 feet wide, or 1 Twenty-foot Equivalent Unit (TEU). The other major size is the 40-foot box, which has the capacity to carry 4,400 VCRs or 267,000 video games or 10,000 pairs of shoes. Containers are either made of steel or aluminum and their structure confers flexibility and hardiness. Each year, about 1.5 million TEU worth of containers are manufactured. The global inventory of containers was estimated to be around 15.9 million TEU by 2002. The standard 20-foot container costs about $2,000 and a 40-footer about $4,000.

Among the numerous advantages related to the success of containers in international transport, it is possible to note several elements:

- **Standard transport product**. A container can be manipulated anywhere in the world as its dimensions are an ISO standard. Indeed, transfer infrastructures allow all elements (vehicles) of a transport chain to handle it with relative ease. The rapid diffusion of containerization was facilitated by the fact that its initiator, Malcolm McLean, purposely did not patent his invention. Consequently all segments of the industry, competitors alike, had access to the standard. It necessitated the construction of specialized ships and of lifting equipment.
- **Flexibility of usage**. A container can transport a wide variety of goods, ranging from raw materials (coal, wheat), manufactured goods, and cars to frozen products. There are specialized containers for transporting liquids (oil and chemical products) and perishable food items in refrigerated containers or reefers. About 1 million TEUs of reefers were being used by 2002.
- **Management**. The container, as an indivisible unit, carries a unique identification number and a size type code, enabling transport management not only in terms of loads, but in terms of unit. Computerized management reduces waiting times considerably and allows the position of containers to be traced at any time. It enables containers to be assigned according to the priority, destination and available transport capacities.
- **Costs**. Containerization of shipping has reduced costs significantly. Before containerization, maritime transport costs could account for between 5 and 10 percent of the retail price of manufactured products; this share has been reduced to 1.5 percent. The main factors behind costs reductions reside in the speed and flexibility incurred by containerization. It has permitted shipping to achieve ever greater economies of scale through the introduction of larger ships. A 5,000 TEU containership has operating costs per container that are 50 percent lower than a 2,500 TEU vessel.

- **Speed**. Transshipment operations are minimal and rapid. A modern containership has a monthly capacity of three to six times more than a conventional cargo ship. This is notably attributable to gains in transshipment time as a crane can handle roughly 30 movements (loading or unloading) per hour. Port turnaround times have thus been reduced from 3 weeks to about 24 hours. It takes on average between 10 and 20 hours to unload 1,000 TEUs compared with between 70 and 100 hours for a similar quantity of general cargo. A regular freighter can spend between half and two-thirds of its useful life in port. With less time in port, containerships can spend more time at sea, and thus be more profitable to operators. Further, containerships are on average 35 percent (19 knots versus 14 knots) faster than regular freighter ships. System-wide, the outcome has been a reduction of costs by about 30 percent because of containerization.
- **Warehousing**. The container limits the risks for goods it transports because it is resistant to shocks and weather conditions. The packaging of goods it contains is therefore simpler and less expensive. Containers fit together, permitting stacking on ships and on the ground. The container is consequently its own warehouse.
- **Security**. The contents of the container are anonymous to outsiders as it can only be opened at the origin, at customs and at the destination. Thefts, especially those of valued commodities, are therefore considerably reduced.

In spite of numerous advantages in the usage of containers, some drawbacks are evident:

- **Consumption of space**. A containership of 25,000 tons requires a minimum of 12 hectares of unloading space. Conventional port areas are not adequate for container handling. Consequently, containers have modified the local geography of ports (see Chapter 5).
- **Infrastructure costs**. Container handling infrastructures, such as gantry cranes, yard equipment, road and rail access, represent important investments for port authorities and load centers. Several developing countries cannot afford these infrastructures and so cannot participate in international trade.
- **Management logistics**. The management logistics of containers is very complex. This requires high levels of information technology for the recording, positioning and ordering of containers handled.
- **Empty travel**. At the global scale, it is rare for the origins and destinations of containers to be in equilibrium. Most container trade is imbalanced, and thus containers "accumulate" in some places and must be shipped back to locations where there are deficits. Many containers are moved empty. Either full or empty, a container takes the same amount of space on the ship or in a storage yard and takes the same amount of time to be transshipped. As a result, shipping lines waste substantial amounts of time and money in repositioning empty containers.
- **Illicit trade**. By its confidential character, the container is a common instrument used in the illicit trade of drug and weapons, as well as for illegal immigrants. Concerns have also been raised about containers being used for terrorism. Electronic scanning systems are being implemented to remotely inspect the contents of containers at major gateways.

Intermodalism and other modes

With the deregulation and privatization trends begun in the 1980s, containerization, which was already well established in the maritime sector, could spread inland. The shipping lines were among the first to exploit the intermodal opportunities that US deregulation permitted. They could offer door-to-door rates to customers by integrating rail services and local truck pick up and delivery in a seamless network. To achieve this they leased trains, managed rail terminals, and in some cases purchased trucking firms. In this way, they could serve customers across the country by offering door-to-door service from suppliers located around the world. The move inland also led to some significant developments, most notably the **double-stacking** of containers on rail cars. This produced important competitive advantages for intermodal rail transport (Muller, 1995).

Other parts of the world have not developed the same degree of synergies between rail and shipping as is found in North America. However, there appears to be a trend towards closer integration in many regions. In Europe, rail intermodal services are becoming well established between the major ports, such as Rotterdam, and southern Germany, and between Hamburg and Eastern Europe (van Klink and van den Berg, 1998). Rail shuttles are also making their appearance in China.

While rail intermodal transport has been relatively slow to develop in Europe, there are extensive interconnections between **barge services** and ocean shipping, particularly on the Rhine (Notteboom and Konings, 2004). Barge shipping offers a low-cost solution to inland distribution where navigable waterways penetrate to interior markets. This solution is being tested in North America, where the Port Authority of New York and New Jersey is sponsoring barge services to Albany and several other destinations.

While it is true that the maritime container has become the work horse of international trade, other types of containers are found in certain modes, most notably in the airline industry. High labor costs and the slowness of loading planes, which require a very rapid turnaround, made the industry very receptive to the concept of a loading unit of standard dimensions. The maritime container was too heavy and did not fit the rounded configuration of a plane's fuselage, and thus a box specific to the needs of the airlines was required. The major breakthrough came with the introduction of wide-bodied aircraft in the late 1970s. Lightweight aluminum boxes could be filled with passengers' baggage or parcels and freight, and loaded into the holds of the planes using tracking that requires little human assistance.

A unique form of intermodal unit has been developed in the rail industry, particularly in the USA. **Roadrailer** is essentially a road trailer that can also roll on rail tracks. It is unlike the TOFC (piggyback) system that requires the trailer be lifted onto a rail flat car. Here the rail bogies may be part of the trailer unit, or be attached in the railway yard. The road unit becomes a rail car, and vice versa. It is used extensively by a major US rail company, Norfolk Southern, whose "Triple Crown" service provides just-in-time deliveries between the automobile parts manufacturers located in Michigan, and the assembly plants located in Georgia, Texas and Mexico and Canada.

Intermodalism and production systems

NS's Triple Crown Service is but one example of how transport chains are being integrated into production systems. As manufacturers spread their production facilities and assembly plants around the globe to take advantage of local factors of production, transportation becomes an ever more important issue. The **integrated transport chain** is itself being integrated into the production and distribution processes. Transport can no longer be considered as a separate service that is required only as a response to supply and demand conditions. It has to be built into the entire supply chain system, from multi-source procurement, to processing, assembly and final distribution (Robinson, 2002).

While many manufacturing corporations may have in-house transportation departments, increasingly the complex needs of the supply chain are being contracted out to third parties. **Third party logistics providers** (3PL) have emerged from traditional intermediaries such as forwarders, or from transport providers such as FEDEX or Maersk-SeaLand. Because the latter are transporters themselves, they are referred to as fourth party logistics providers (4PL). Both groups have been at the forefront of the intermodal revolution that is now assuming more complex organizational forms and importance. In offering door-to-door services, the customer is no longer aware or necessarily concerned with how the shipment gets to its destination. The modes used and the routing selected are no longer of immediate concern. The preoccupation is with cost and level of service. This produces a paradox, that for the customer of intermodal services geographic space becomes meaningless; but for the intermodal providers routing and modal choice assume an ever greater importance.

Concept 3 – Passengers or freight?

Advantages and disadvantages

With some exceptions, such as buses and pipelines, most transport modes have developed to handle both freight and passenger traffic. In some cases both are carried in the same vehicle, as for example in the airlines where freight is transported in the cargo holds of passenger aircraft. In others, different types of vehicle have been developed for freight and passenger traffic, but they both share the same road bed, as for example in rail and road traffic. In shipping, passengers and freight used to share the same vessel, but since the 1950s specialization has occurred, and the two are now quite distinct, except for ferries and some RORO services.

The sharing by freight and passengers of a mode is not without difficulties, and indeed some of the **major problems** confronting transportation occur where the two seek to co-inhabit. For example, trucks in urban areas are seen as a nuisance and a cause of congestion by passenger transport users. The poor performance of some modes, such as rail, is seen as the outcome of freight and passengers having to **share routes**. This raises the question as to whether freight and passengers are compatible. The main advantages of joint operations are:

- High capital costs can be justified more easily with a diverse revenue stream (rail, airlines, ferries).
- Maintenance costs can be spread over a wider base (rail, airlines).
- The same traction sources can be used for both freight and passengers, particularly for rail.

The main disadvantages of joint operations are:

- Locations of demand rarely match – origin/destination of freight is usually quite distinct spatially from passenger traffic.
- Frequency of demand is different – for passengers the need is for high frequency service, for freight it tends to be somewhat less critical.
- Timing of service – demand for passenger services has specific peaks during the day, for freight it tends to be more evenly spread throughout the day.
- Traffic balance – on a daily basis passenger flows tend to be in equilibrium, for freight, market imbalances produce empty flows.

- Reliability – although freight traffic increasingly demands quality service, for passengers delays are unacceptable.
- Sharing routes favors passenger traffic – passenger trains are given priority; trucks may be excluded from areas at certain times of the day.
- Different operational speeds – passengers demand faster service.
- Security screening measures for passengers and freight require totally different procedures.

A growing divergence

In several modes and across many regions passenger and freight transport is being unbundled.

- **Shipping**. It has already been mentioned that in the maritime sector passenger services have become divorced from freight operations, the exception being some ferry services where the use of RORO ships on high frequency services adapt to the needs of both market segments. Deep sea passenger travel is now dominated by cruise shipping which has no freight-handling capabilities, and bulk and general cargo ships rarely have an interest or the ability to transport passengers.
- **Rail**. Most rail systems still operate passenger and freight business. Where both segments are maintained, the railways give priority to passengers, since rail persists as the dominant mode for inter-city transport in India, China and much of the developing world. In Europe, the national rail systems and various levels of government have prioritized passenger service as a means of checking the growth of the automobile, with its resultant problems of congestion and environmental degradation (see Chapter 8). Significant investments have occurred in improving the comfort of trains and in passenger rail stations, but most notable have been the upgrading of track and equipment in order to achieve higher operational speeds. Freight transport has tended to lose out because of the emphasis on passengers. Because of their lower operational speeds, freight trains are frequently excluded from daytime slots, when passenger trains are most in demand. Overnight journeys may not meet the needs of freight customers. This incompatibility is a factor in the loss of freight business by most rail systems still trying to operate both freight and passenger operations. In Europe, there are signs that the two markets are being separated. First, it is occurring at the management level. The liberalization of the railway system that is being forced by the European Commission is resulting in the separation of passenger and freight operations. This had already taken place in the UK when British Rail was privatized. Second, the move towards high-speed passenger rail service necessitated the construction of separate rights of way for the TGV trains. This has tended to move passenger train services from the existing tracks, thereby opening up more daytime slots for freight trains. Third, the Dutch are building a freight only track, the Betuwe Line, from the port of Rotterdam to the German border, having already sold the freight business of the Netherlands railway (NS) to DB (Deutsche Bahn), and having opened up the freight business to other firms. In North America, the divorce between freight and passenger rail business is most complete. The private railway companies could not compete against the automobile and airline industry for passenger traffic, and consequently withdrew from the passenger business in the 1970s. They were left to operate a freight only system, which has generally been successful, especially with the introduction of intermodality. The passenger business has been taken over by public agencies, AMTRAK in the USA, and VIA Rail in Canada. Both are struggling

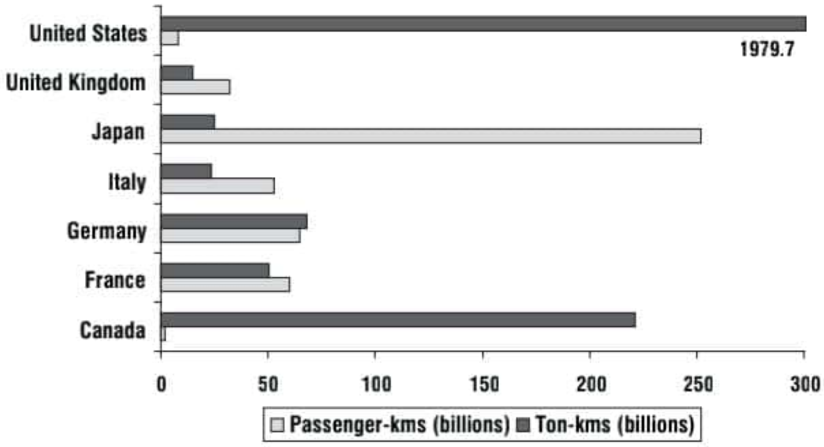


Figure 4.13 Domestic rail passenger travel and freight activity, G7 Countries, 1996 (Source: US Department of Transportation, BTS, G–7 Countries: Transportation Highlights)

to survive. A major problem is that they have to lease trackage from the freight railways, and thus slower freight trains have priority (Figure 4.13).

- **Roads**. Freight and passenger vehicles still share the roads. The growth of freight traffic is helping increase road congestion and in many cities concerns are being raised about the presence of trucks (see Chapters 7 and 9). Already, restrictions are in place on truck dimensions and weights in certain parts of cities, and there are growing pressures to limiting truck access to non-daylight hours. Certain highways exclude truck traffic – the parkways in the USA for example. These are examples of what is likely to become a growing trend – the need to separate truck from passenger vehicle traffic. Facing chronic congestion around the access points to the port of Rotterdam and at the freight terminals at Schiphol airport, Dutch engineers have worked on feasibility studies of developing separate underground road networks for freight vehicles.
- **Air transport**. Air transport is the mode where freight and passengers are most integrated. Yet even here a divergence is being noted. The growth of all-freight airlines and the freight-only planes operated by some of the major carriers, such as Singapore Airlines, are heralding a trend. The interests of the shippers, including the timing of the shipments and the destinations, are sometimes better served than in passenger aircraft. The divergence between passengers and freight is also being accentuated by the growing importance of charter and "no frills" carriers. Their interest in freight is very limited, especially when their business is oriented towards tourism, since tourist destinations tend to be lean freight generating locations.

Method 1 – Technical performance indicators

Indicators

Multimodal transportation networks rest upon the **combinatory costs and performance of transport modes**, or what is referred to as economies of scope. For instance, a single container shipped overseas at the lowest cost from its origin can go from road, to seaway, to railway and to road again before reaching its destination. Freight shippers and carriers therefore require quantitative tools for decision-making in order to compare performances of various transport modes and transport networks. Time-efficiency becomes a set imperative for both freight and passenger transit in private as well as in public sector activities.

Performance indicators are widely used by geographers and economists to empirically assess the **technical performance** (not to be confused with economic performance, for there can exist a lag between the two) of differing transport modes, in other words their capacity to move goods or passengers around. Hence, basic technical performance calculations can be particularly useful for networks' global performance analysis as well as for modal comparison, analysis, and evaluation by bridging both physical attributes (length, distance, configuration, etc.) and time-based attributes (punctuality, regularity, reliance, etc.) of networks. Some indicators are currently used to measure freight and passenger transport. Table 4.1 gives a few of the most common ones.

Passenger-km or ton-km are standard units for measuring travel that consider the number of people traveling or ton output and distance traveled. For example, 120 passenger-km represents 10 passengers traveling 12 kilometers or 2 passengers traveling 60 kilometers, and so on. More specifically, such indicators are of great utility by allowing cross-temporal analysis of a transport nexus or given transport modes.

Economic impact indicators

Undoubtedly, transportation plays a considerable role in the economy with its omnipresence throughout the production chain, at all geographic scales. It is an **integral constituent of the production–consumption cycle**. Economic impact indicators help to appreciate the relationship between transport systems and the economy as well as to inform on the economic weight of this type of activity. Geographers should be familiar with basic econometric impact indexes (see Table 4.2).

Efficiency is usually defined as the ratio of input to output, or the output per each unit of input. Modal variations in efficiency will depend heavily on what is to be carried, the distance traveled, the degree and complexity of logistics required as well as economies of scale. Freight transport chains rest upon the complementarity of cost-efficient and

Table 4.1 Commonly used performance indicators

Indicator	*Passenger*	*Freight*	*Description*
Passenger/freight density	passenger-km/km	ton-km/km	A standard measure of transport efficiency.
Mean distance traveled	passenger-km/passenger	ton-km/ton	A measure of the ground covering capacity of networks and different transport modes.
Mean per capita ton output (freight) Mean number of trips per capita (passenger)	passengers/population	tons/population	Used to measure the relative performance of transport modes.
Mean occupation coefficient	number of passengers aboard/total carrying capacity (%)	actual load (ton)/ overall load capacity (ton) (%)	Especially useful with increasing complexity of logistics associated with containerization of freight (i.e. the problem of empty returns). Can also be used to measure transit ridership.

Table 4.2 Measures of efficiency

Efficiency indicators	*Scale-specific indicators*	
(Factors of production)	*Micro*	*Meso-macro*
output/capital	transport sector income/ local income	output/GDP
output/labor	output/local income	

time-efficient modes, seeking most of the time a balanced compromise rather than an ideal or perfect equilibrium.

Maritime transport is still the most cost-efficient way to transport bulk merchandise over long distances. On the other hand, while air transport is recognized for its unsurpassed time-efficiency versus other modes over long distances, it remains an expensive option. Thus, **vertical integration**, or the absorption of transportation activities by producers, illustrates the search for these two efficiency attributes by gaining direct control over inputs.

Transportation and economic impacts

The relationship between transport systems and their larger economic frame becomes clear when looking at restructuring patterns which carriers and firms are currently undergoing. Structural mutations, best illustrated by the popularity of just-in-time practices, are fuelled by two opposing yet effective forces: transporters seek to achieve economies of scale while having to conform to an increasingly "customized" demand.

Factor substitution is a commonly adopted path in order to reduce costs of production and attain greater efficiency. Containerization of freight by substituting labor for capital and technology is a good illustration of the phenomenon. Measures of capital productivity for such capital-intensive transport means are of central importance; an output/capital ratio is then commonly used. While the output/labor ratio performs the same productivity measurement but for the labor input (this form of indicator can be used for each factor of production in the system), a capital/labor ratio aims at measuring which factor predominates within the relationship between capital and labor productivity. The above set of indicators therefore provides insights on the relative weight of factors within the production process.

More scale-specific indicators can also be used to appreciate the role of transport within the economy. Knowing freight transport both contributes to and is fuelled by a larger economic context, freight output can be confronted against macro-economic indicators: an output/GDP ratio measures the relationship between economic activity and traffic freight, in other words the traffic intensity. At the local level, the status of the transport industry within the local economy is given by a transport sector income / local income ratio. Still at a micro-scale, finally, a measure of the relative production value of freight output is provided by an output/local income ratio.

Underlying objectives of application of such indicators are as varied as they are numerous. Efficiency indicators constitute valuable tools to tackle project viability questions as well as to measure investment returns and cost/subsidy recovery of transport systems. Input–output analyses making use of some of the above indicators are also instrumental to the development of global economic impact indexes and productivity assessment concepts such as the Total Factor Productivity (TFP) and to identify sources of productivity gains.

Specialization index

In transport, to find out if a terminal is specialized in the transshipment and/or handling of a particular kind of merchandise or if, inversely, it transfers a wide variety of merchandise, we can calculate a specialization index. For example, the index can be used to know if a port is specialized in the handling of a certain type of product (e.g. containers) or if it handles a wide range of merchandise. As a consequence, such an index is quite versatile and has a variety of applications; it informs geographers on the activities of any type of terminal (port, train and airport). In the case of an airport terminal, one could ask if a given airport deals with only a single type of flights/passengers (local, national, international, etc.) or if it welcomes several. The specialization index (SI) is calculated using the following formula:

$$SI = \frac{\sum_i t_i^2}{\left(\sum_i t_i\right)^2}$$

which is the total of squares of tonnage (or monetary value) of each type of merchandise i (t_i) handled at a terminal over the square of the total volume tonnage (or monetary value) of merchandise handled at the terminal.

So, if the specialization index tends toward 1, such a result indicates that the terminal is highly specialized. If, inversely, the index tends toward 0, it means that the terminal's activity is diversified. Thus, the specialization index is called upon to appreciate the degree of specialization/diversification of a port, an airport, a train station or any type of terminal.

Location coefficient

Certain kinds of merchandise are often transshipped at particular terminals rather than at others. Thus, the degree of concentration of a certain type of traffic in a terminal (port, airport, train station) compared with the average for all the terminals, can be measured by using the location coefficient.

> The **location coefficient** is the share of traffic occupied by a type of merchandise at a terminal over the share of traffic of the same type of merchandise among the total traffic of all terminals of the same type.

In the field of transportation, the location coefficient (*LC*) is calculated by using the following formula:

$$LC = \frac{\left(\dfrac{M_{ti}}{\sum_t M_{ti}}\right)}{\left(\dfrac{\sum_t M_t}{\sum M}\right)}$$

requirement is that the geometry of the network matches the reality as closely as possible since these networks are often part of a geographic information system where accurate location and visualization is a requisite. This has commonly resulted in the fragmentation of each logical link into a multitude of segments, with most of the nodes of these segments mere intermediate cosmetic elements. The topology of such network data models is not well defined, and has to be inferred. However, these network data models benefit from the attribute linking capabilities of the spatial database models they are derived from. Among the most significant attributes that can be attached to network layers are:

- **Classification and labeling**. Each segment can be classified into categories such as its function (street, highway, railway, etc.), importance (number of lanes) and type (paved, non-paved). Also, a complex labeling structure can be established with prefixes, proper names and suffixes.
- **Linear referencing system**. Several systems to locate elements along a segment have been established. One of the most common is the address system where each segment is provided with an address range. Through linear interpolation, a specific location can be derived (geocoding).
- **Segment travel costs**. Can consider a vast array of impedance measures. Among the most common is the length of the segment, a typical travel time or a speed limit. Congestion can also be assessed, either as a specific value of impedance or as a mathematical function.
- **Direction**. To avoid unnecessary, and often unrealistic duplication of links, especially at the street level, a directional attribute can be included in the attribute table.
- **Overcrossing and undercrossing**. Since the great majority of layer-based network models are planar, they are ill designed to deal with non-planar representations. A provision must be made in the attribute table to identify segments that are overcrossing or undercrossing a segment they intersect with.
- **Turn penalties**. An important attribute to insure accurate routing within a network. Each intersection has different turn constraints and possibilities. Conventionally in road transportation, a right turn is assumed to have a smaller penalty than a left turn.

The TIGER (Topologically Integrated Geographic Encoding and Referencing) model is a notable example of a layer-based structure which has been widely accepted. TIGER was developed by the US Census Bureau to store street information constructed for the 1990 census. It contains complete geographic coordinates in a line-based structure. The most important attributes include street name and address information, offering an efficient linear referencing system for geocoding. The layer-based approach is consequently good to solve the cartography and geocoding issues. However, it is ill suited to comprehensively address routing and assignment transport problems.

The **object-oriented approach** represents the latest development in spatial data models. It assumes that each geographical feature is an object with a set of properties and a set of relationships with other objects, namely membership and inheritance. As such, a transportation network is an object composed of other objects, namely nodes and links. Since topology is one of the core concepts defining transportation networks, relationships expressing it are imbedded in object-oriented representations.