

Appeared in: Karl O. Ott, Ed., Annual Papers of the Institute for Applied Systems Analysis (ASA) 1981-82, Köln: AGF, 1982, pp. 253-262.

DETERMINING THE CAPACITY OF A TRANSPORTATION SYSTEM

Edward K. Morlok*

INTRODUCTION

Questions related to the capacity of transportation systems are increasing in importance, and therefore it is desirable to have a means of measuring the capacity of such systems. At the present time there exists no definition of the capacity of a transportation system and hence no operational means of measuring this. It is the purpose of this paper to summarize briefly research directed toward developing such a definition and measurement procedures.

Although capacity of transportation systems has always been of some interest, its importance has been increasing in recent years, primarily for two reasons. One is that in many situations existing transportation systems are expected to continue to experience increasing traffic that ultimately will exceed that which can be accommodated unless investments are made to expand capacity. For example, in the United States, railroad freight traffic has increased rather steadily for the past two decades, from about 915 billion ton-km (572 billion ton-miles) in 1960 to 1470 billion ton-km (919 billion ton-miles) in 1980 (Assoc. of Amer. Railroads data), an increase of 61 percent. Rail traffic is expected to grow even more rapidly in the future, primarily as a result of deregulation of the railroads and a natural increase in certain rail-oriented traffic such as coal. As a result

* The author is the UPS Foundation Professor of Transportation at the University of Pennsylvania. This study was undertaken while the author was at Angewandte Systemanalyse (ASA), on sabbatical leave from the University of Pennsylvania, as a result of having received the U.S. Senior Scientist Award from the Alexander von Humboldt Foundation at Bonn, West Germany. Appreciation is expressed to ASA, the von Humboldt Foundation, the University of Pennsylvania, and the UPS Foundation for their support, but no endorsement of the findings is implied.

many main lines are experiencing phenomena such as congestion and (probably) the diseconomies of scale often related to systems reaching saturation in terms of capacity. In this sort of situation it is desirable to be able to measure capacity and use that measure to forecast when capacity will be reached (or alternatively when diseconomies will set in) so the proper steps can be taken to alleviate problems that otherwise would occur.

A second reason for the importance of capacity is the desire to understand the cost properties of transportation systems. In an era of freedom with respect to pricing and more generally product characteristics (e.g., travel time, reliability), it is incumbent upon transportation firms to understand their costs well. Economic theory suggests that a productive process such as transportation should experience a "U-shaped" cost curve in the short run, with the location of the minimum point and the slopes of the curve on either side of that point dependent upon the capacity of the system. Thus, in order to develop a reasonably correct and precise characterization of system costs, it presumably will be necessary to be able to include one or more appropriate measures of system capacity. Absent any definition of capacity, this is impossible, and surrogates for capacity must be used.

All this is not to say that the notion of capacity has not been used in connection with transportation systems. The difficulty is simply that the concepts of capacity used to date have been related to 'individual components of the system only and hence say virtually nothing about the capacity of the system as a whole. Examples of such component capacities include the capacity of a link or a terminal facility to accommodate traffic, usually measured in units of vehicles per unit time, the carrying capacity of vehicles as measured by mass or volume, and the capacity of motive power fleets (e.g., a locomotive fleet) measured in total horsepower or total tractive effort (pulling power). Clearly such measures in no way capture the capacity of the system as a whole.

Furthermore, there is no apparent way of combining these measures in some manner so as to yield total system capacity.

In the next section we shall explore various definitions of capacity and select one for use with a transportation system. The definition then makes possible the construction of a model estimating system capacity, this model being sketched in the following section.

DEFINING CAPACITY AND OUTPUT

There are basically two definitions of the capacity of a physical system which produces a product of goods or services. One is that the capacity is equal to the maximum quantity of output which the system can produce, considering only physical limitations on production. For example, one might be interested in the capacity of a system of power plants to produce electric power. This output level is of course limited by both the number and size of generators and the availability of fuel. This definition of capacity focuses solely on maximizing the output, and ignores other factors which may make achievement of such an output unlikely. For example, in some systems marginal costs may become very large when output approaches this capacity. This concept of capacity is often termed ultimate capacity.

The other basic definition of capacity recognizes that the cost may be far too large at the ultimate capacity for such a level of output to be practically or economically attainable. This suggests the other basic definition of capacity: the maximum output at which cost does not exceed a maximum acceptable value. This concept is termed economic capacity in economics literature and practical capacity in the engineering literature. The term cost is deliberately left vague, for the specific measure used varies with the situation, in some cases being average cost, in others marginal cost, etc.

These definitions all refer to the quantity of output of the system of interest. This naturally raises the question of how the output of an entire transportation system is to be defined. In most cases where capacity has been estimated, there is usually defined -- or assumed -- a single, homogeneous output of the system being considered. For example, in the case of a manufactured product, the quantity of output would be simply a count of that particular product produced. If there is only one product, or if the variations in the product are rather minor, such that one measurement can be applied to all of the different products, then there is no difficulty in defining a single measure of output. This is the case for the individual components of transportation systems for which capacity is often estimated, as described previously. But in the case of an entire transportation system, the output is very heterogeneous, encompassing many links as well as other elements of the system. This necessitates a discussion of transportation system output in detail.

If we consider a transportation system from the perspective of a physical system, the product of that system probably is most appropriately considered as the movement of things -- person or other objects -- from one location to another. Thus, at the microscopic level of a trip by an individual person or of a single shipment, the product of a transportation system would be a change in location. Along with this change in location come a number of concomitant changes. One is a change in time from the moment when it departed to that when it arrived at its destination. Usually, associated with movement will be some other changes, some of which may reflect the reason for which the movement is effected. For example, in the case of goods, they are usually moved only when the (monetary) value of the goods is increased sufficiently by virtue of the transport or change in location to more than offset the cost to society of effecting that transport. In addition, other features may change which make the commodity more or less desirable, such as possible deterioration or ripening of foodstuffs as a result of the passage of time during their movement. Thus there may occur changes in the state or condition of the goods between the origin and the destination.

At this microscopic level, we could then consider the output of the transportation system as being expressed by a vector describing the object transported and the changes in attributes associated with that object. For brevity, we shall term the transportation of an object a shipment. The vector describing a single shipment, termed the shipment vector, would then consist of elements of the following types: ones consisting of single values for those attributes that do not change as a result of movement, and ones consisting of pairs of values, the first for the origin and the second for the destination, in those cases where the attribute changes as a result of movement. In particular, shipment vector s_i , describing shipment i , would specify

- Commodity: c_i
- Size of shipment {e.g., weight or volume-assumed here invariant): w_i
- Location of origin and destination: l_{1i}, l_{2i}
- Times of departure at origin and. of arrival at destination: t_{1i}, t_{2i}

Each of these elements could itself consist of a vector if necessary to contain the information desired; for example, location might be given by latitude and longitude.

The total output of a transportation system would then be expressed by the collection of all shipment vectors for that system. This would provide a complete description, but it is obviously too unwieldy for analysis purposes. A more aggregate expression of output is necessary.

From the viewpoint of the transport system as a whole, the problem is that the output is heterogeneous, that is, there are many products. Each shipment is truly unique in some aspects although many of these may be unimportant in terms of overall measures of output or capacity. If categories of attributes can be specified such that every shipment can be classified into one category, then the output could be described by a vector giving the quantity of shipments in each category. This is commonly done in transport, with the output or traffic being described as the total quantity of shipments of a particular type or category originating (or terminating) in a given time period. In the typical network data

bases, for example, each category usually consists of a collection of similar commodities {e.g., all grades of coal, or all types of meat), travelling between the same origin towns or other geographic units. It may include reference to the size of shipment, including only shipments within a certain range of size. The latter is very important in connection with the transport process and costs, a single drum of oil product requiring much more handling per ton than a full unit train load, for example. The precise specification of these categories and their number, will of course depend upon the particular application. Although terminology is by no means identical in the various sectors of transportation, it has become common in some sectors to refer to each of these categories as a traffic lane. We shall use this term here.

Let us assume that there have been identified N traffic lanes on a system. The output of that system could then be described by the quantity of traffic in each lane. A typical measure for freight is weight or mass, tons or kg, and for passengers, persons or possibly tons at a standard conversion factor. For brevity, we shall refer only to output measured by weight in each category or traffic lane, Q_i .

A useful related description of the output is the traffic pattern, which is defined as the fraction of total traffic in each lane. If Q is the total tonnage carried,

$$Q = \sum_{i=1}^N Q_i \quad (3)$$

The traffic pattern, λ , is a vector, each element of which refers to one traffic lane:

$$\lambda = (\lambda_1, \lambda_2, \dots, \lambda_i, \dots, \lambda_N) \quad (4)$$

$$\lambda_i = \frac{Q_i}{Q} \quad (5)$$

The combination of traffic pattern and total output gives us a convenient means of making operational the concept of maximizing output, essential for a measure of capacity, as will be described in the next section.

THE MODEL

The most natural way in which to make an operational procedure for estimating capacity as defined above is to characterize the problem as a mathematical program, in which a scalar quantity is maximized subject to a number of constraints. In this case, the value of Q is maximized, subject to the constraint on traffic pattern as in equation (5) above. In addition other constraints describe the relevant physical or other characteristics of the system. Considering the summary nature of the report of which this paper is a part, it seems most appropriate to describe the general features of this model in prose rather than to specify in abstract mathematical terms each and every equation or inequality of the model. This general description is presented in terms of the basic types of variables and equations.

There are basically four types of relationships involved in the model in addition to the function to be maximized, which is simply the total traffic to be accommodated as measured in either tons originated, ton-miles, or some other suitable measure as discussed in detail above, and the traffic pattern constraint. The four relationships basically deal with the following aspects of the problem. One type of relationships treats the capacity of facilities to accommodate traffic flowing through them. The second type of relationship deals with the necessity of having vehicles available to accommodate all of the traffic which is to be moved. The third relationship deals with the adequacy of various types of essentially continuously consumed resources in order for the system to function, resources such as fuel (or electric power) and labor to operate the system, for example. The fourth relates to costs as a basis for limiting traffic. Each of these types of relationships will be discussed.

The facility capacity relationships are based on the relationship between the time required for a vehicle (or other traffic unit) to move through or over a facility and the flow or volume of vehicles per unit time entering that facility. Such a relationship is often termed a congestion function, for on virtually all facilities it has been observed that as the quantity of vehicles entering that facility increases per unit time, the time required for the vehicles to traverse the facility increases, at an increasing rate. There is also presumably some maximum volume of vehicles which can be accepted. Certain aspects of this sort of relationship will vary with the particular type of facility being represented and with the overall modeling approach. For example, depending upon whether or not it is possible for vehicles to queue up on a facility, this curve may be backward bending, as it is in the highway case.

The second type of relationships deals with fleet size. It requires that sufficient vehicles be available to accommodate the traffic. In the simplest case in which each vehicle is operated independently (as in the case of the motor truck), the basic relationship is one in which the total time vehicles are available in the period, measured by vehicle-hours and equal to the sum of the hours available of all vehicles in the fleet, must be sufficient to cover the total vehicle-hours in which the vehicles are moving loaded plus the total vehicle-hours of required empty movements of these vehicles (in order to position them for accommodating loads). Thus a necessary part of this relationship is one that estimates total vehicle-hours consumed in moving loads from any given origin to any given destination, this being the product of total tons originating at that origin for that destination and the total time required for such movement, the latter being estimated considering the facility capacity relationships. Similar relationships exist for the empty movement. In addition, it is required that there be conservation of vehicle flow at each point at which traffic enters or leaves the system, this in effect stating that the total number of vehicles of a given type entering such a node on the system (loaded and empty), must equal the total number of vehicles leaving (loaded and empty). Such conservation of flow relationships must be written at each node at which traffic may be loaded or unloaded from vehicles.

In addition, in some transportation systems such as railways and towboat barge systems, the load carrying vehicles (railroad cars or barges) can only move if there are adequate propulsion units (locomotives or tow-boats). In such cases additional relationships exist to ensure that the total vehicle-hours of those types of vehicles is sufficient to accommodate the load-carrying vehicle movements (empty and loaded). The ratio of propulsion vehicles to load-carrying vehicles is set on each link of the system for each vehicle type to reflect their particular characteristics.

The third basic type of relationship deals with consumables of which there must be sufficient in order for movement to occur. One example would be fuel. The fuel available in any given time period must not be exceeded by that consumed by the vehicles in moving during that time period. This is accomplished through the use of a fuel consumption estimating relationship for each link of the system and each vehicle type, and using these to estimate total fuel consumption as a function of total vehicle flow. Similar relationships can be developed for electric consumption, for labor of various types, as well as other consumables.

The final type of relationship deals with practical or economic capacity limits. As such, it is optional, and would be employed only when an economic rather than ultimate capacity is desired. Basically these relationships estimate one or more measures of cost as a function of flows (of traffic, vehicles, etc.) and require that costs not exceed upper bounds.

Two versions of the model have been developed. One is as a linear program which can accommodate nonlinear relationships (such as those for congestion) as piecewise linear approximations. This version of the model has the advantage of the computational power of linear programming and the ability to use readily available linear programming solution codes, but has the obvious disadvantage of requiring approximation of nonlinear relationships. The other version is as a general nonlinear program which nevertheless is convex for typical forms of relationships. Given the most

likely forms of the particular types of nonlinearities involved, it should be solvable by standard nonlinear programming methods.

CONCLUSIONS

This model has been developed in such a manner that existing relationships describing such things as congestion functions and vehicle availability can be utilized. Thus the next step would be to apply it to an actual situation. This would serve to test the model formulation for applicability and compatibility with existing relationships, and would upon successful completion also demonstrate the model's usefulness for estimating capacity. Means for carrying out such applications are being pursued.