UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Transportation Operations and Traffic

(Course notes)

Francesc Soriguera Martí

(2022)

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

PREFACE

Contents

1. Aims and scope of the book 2
2. Suggested course schedule 2
3. Acknowledgments and credits 4

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

1. Aims and scope of the book

This book is written as a text book for students taking a first course in transportation operations and traffic, which should be a fundamental block in any transportation engineering degree. The course could also be part of more multidisciplinary degrees (e.g. Civil Engineering) to cover the transportation engineering competences.

The book will present the concepts of transport planning and operations that should be understood by every student of transportation engineering or planning, regardless of his or her background or specific professional interests, and will prepare the student for further study in this field.

The several chapters of the book introduce tools for the analysis and evaluation of transportation systems, including operations research, traffic flow theory, demand modeling and forecasting, network control and flow assignment. The course emphasizes the knowledge of causal and quantitative performance of transport systems as well as the stakeholders' behavior (users, transport agencies and society). Guidelines for an optimal design, performance and management of terminals and transport infrastructures are also provided. Technical and technological aspects are deemphasized in the book, although innovative applications and their relationship with Information and Communication Technologies are outlined when relevant. The concepts presented in the course could be applied in modal interchange terminals for passengers in public transport systems, airport terminals (land- side /air-side management, baggage management system), port terminals (operating container terminals, liquid / solid bulk, ro-ro, etc.), railway terminals, in-land ports, road terminals, freight villages, logistics centers and hubs, amongst others.

2. Suggested course schedule

The contents of this book could be covered in a course with 4 hours of lectures per week in a semester of 15 weeks. This could include some discussion sessions, devoted to reinforce the concepts presented in the lectures with examples and practical application in problems, exams and a course mini-project. The course mini-project (in groups) typically involves field work of the students, in order to develop their observation and measurement capabilities and apply the analysis tools presented in the course to real measured data.

Table 1 summarizes a tentative course schedule.

Table 1. Tentative "Operations in transportation and traffic" course schedule

Week	Session	TOPIC	Chapter	Pages
1	1	Course overview. **Introduction to Transportation Operations.** Components of the transport system. Propose a simple problem involving trajectories. Visualize that simple problems can be challenging without the right tools (i.e. x-t diagram)	1, 2	xx
	2	**Time-space diagram.** Trajectories. Examples. Scheduling problems. Constructing trajectories. Traffic stream properties. Time vs space averages. Fundamental equation of traffic.	2	

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

2	3	**Discussion 1.** Trajectories, time-space diagram and traffic variables.	2	
	4	**Queuing processes I.** Introduction to the analysis of queuing processes. Components of a queuing system. Cumulative plots N-t.	3	
3	5	**Queuing processes II.** Input-Output diagrams. Time and accumulation in the system. Little's formula. Constructing queuing diagrams with incomplete information. Design of queuing systems.	3	
	6	**Queuing processes III.** Stochastic effects. Centralization. Optimization. Psychology of waiting lines.	3	
4	7	**Discussion 2.** Queuing processes.	3	
	8	**Flow conservation.** Introduction to traffic flow theory. Flow conservation equation. Velocity of an interphase. Relative flow seen by a moving observer.	4	
5	9	**Diagrams & Manuals.** Definition of a traffic diagram. Main parameters of the diagrams. Speed-density models: Greenshields, Greenberg, Underwood, Edie. The fundamental diagram. Newell's triangular simplification. Warning about manuals.	4	
	10	**Continuum Theories.** Free & forced regimes. Traffic dynamics. Continuum theory of traffic (LWR). Triangular FD and instantaneous speed change simplification Example of application: Incident on a freeway.	4	
6	11	**Discussion 3.** Continuum theories.	4	
	12	**Microscopic Modelling**. Limitations of continuum theories. Introduction to microscopic traffic analysis. Spacing & reaction time. Car following laws: Pipes, Forbes and General Motors. Traffic stability.	5	
7	13	**Discussion 4.** Microscopic modelling	5	
	14	**Midterm Exam.**	-	
8	15	Sample solution of Midterm Exam **Presentation of the course Mini Project**	-	
	16	**Scheduled Transportation: user costs.** Small vs large headways. Why stay on schedule? Expected wait time. Transfers design.	6	
9	17	**Scheduled Transportation: agency costs.** Model for the vehicle trip time. Required vehicle fleet to serve a route. Headway optimization. Bus capacity assessment. Stochastic effects. Schedule control.	6	
	18	**Discussion 5.** Scheduled transportation.	6	

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

10	19	**Mini-Project proposal submission.** **Useful concepts for the Mini-Project.** Estimation. Stochastic processes. Simulation. Statistical estimation. Sample size issues. Estimating bottleneck capacity.	-	
	20	**Introduction to demand modelling.** 6 steps of the planning process. Demand models. 3 steps of demand modelling. Endogeneity problem & equilibrium solution.	7	
11	21	**Utility theory.** Demographics & aggregation. Multinomial choice. Elasticity. Summary, problems & solutions.	7	
	22	**Random utility models**. Binary logit model. Example of application. Maximum likelihood estimation. Properties of logit models: Elasticity; IIA.	7	
12	23	**Discussion 6.** Demand modelling.	7	
	24	**Traffic control.** The isolated traffic signal. Lost times. Stochastic fluctuations. Webster equation. Coordination.	8	
13	25	**Networks.** Wardrop's principles. Equilibrium. Traffic assignment.	8	
	26	**Network control.** Paradoxes: Braess' paradox. Smith's paradox. Discussion of our network analysis.	8	
14	27	**Discussion 7.** Network control	8	
	28	**Mini-Project presentations**	-	
30	29	**Review for the final exam**	-	
	30	**Final Exam**	-	

3. Acknowledgments and credits

This book is the result of the students' persistence in asking for some course notes for the Transportation Operations and Traffic course I teach in several degrees at the *Universitat Politècnica de Catalunya* (UPC-BarcelonaTech). Every year, students ask for background materials to reinforce my lectures, which are typically held on the blackboard. The books and notes I usually suggest to students either cover the topics in much more depth, so that students feel lost and overwhelmed, or lack the required detail to fully understand and apply the fundamentals. This made me understand that the compilation of my lectures into a textbook would be helpful for students, to have some background material to better follow and understand the course.

Although this book was amongst my objectives for several years, the truth is that it was difficult to find the time and to devote the dedication it requires. The first step was the digitalization of all the figures in the course, which were handwritten in my notes. To this end, students, while in their teaching assistance scholarships, helped me for years. I gratefully acknowledge their help during those years. Then, I used the skeleton provided by the

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

compilation of these figures to support my lectures, without finding the time to dress it up with the text and explanations. Everything changed during the COVID-19 pandemic lockdown, when teaching shifted to an online version. Then, I found the time and motivation to advance in the development of this textbook, which has finally reached its completion, at least in its first version.

It must be understood that nothing in this course is "mine". The concepts covered are well known fundamentals, which were developed many decades ago. Most of them are basics from physics, while those more specific to transportation and traffic theory date back to 1930's -1950's when the science behind transportation operations surged together with the growing traffic flows. Neither the structure, contents of the book, nor the way in which concepts are explained are mine. I learnt and copied everything from my professors, which I was extremely lucky to find along the way. My first exposure to these topics was while taking the transportation courses in the civil engineering degree at the UPC. By then, prof. Francesc Robusté, a young professor who was just landed from UC Berkeley, bombarded us with these transportation operations and traffic concepts. Clearly, the program of the degree at that time was not prepared to absorb the new discipline that prof. Robusté was trying to implement, and the number of concepts per unit time in his lessons overwhelmed me. It took me years to digest everything. Fortunately, today, transportation science is a more relevant part in the civil engineering program at the UPC, and these topics can be covered at a slower pace. Later on, I became aware that prof. Robusté was implementing the knowledge he gained during his MSc and PhD at UC Berkeley, in particular the teachings from his advisor, prof. Carlos F. Daganzo, to whom he always showed great admiration. I read the book entitled "Fundamentals of transportation and traffic operations" (Daganzo, C.F. (1997). Elsevier, New York.) and all the somehow fuzzy knowledge I had acquired on transportation and traffic theory became clear. Later on, during my PhD, directed by prof. Robusté, I had the opportunity to visit UC Berkeley, hosted precisely by prof. Daganzo together with prof. Mike Cassidy, and audit the CE155 and CE251 courses at the Civil Engineering Department, which were based on Daganzo's "Fundamentals" book. That year these courses were taught brilliantly by prof. Cassidy. This experience forged my teaching of the introductory courses to transportation science I teach today at the UPC. Prof. Daganzo says in the preface of the "Fundamentals" book that only topics with "*a solid grounding in physical reality have been included because those are the ones that have the best chance of standing the test of time*." This was written 25 years ago, and I can assure he was right. The topics covered are still fundamental and basic for any student of transportation engineering, they are contemporary to most of the transportation problems we face today, and they are unknown by most engineering students, even at the masters' level.

In conclusion, the present textbook is simply an abridged version of the "Fundamentals of transportation and traffic operations" (Daganzo, C.F. (1997). Elsevier, New York.) book. The interested reader is referred to it to obtain a deeper understanding of many of the topics covered and to check for the relevant citations, which are omitted here. All the credit of the contents and structure of the present textbook must be given to prof. Carlos F. Daganzo, and also to profs. Robusté and Cassidy for disseminating and expanding this ideas, and to whom all I am extremely grateful. Only the errors are mine.

Francesc Soriguera
Barcelona
March 2022

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

INTRODUCTION AND OVERVIEW

Contents

1. Components of the transport system 2
2. Outline of the present book 2

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

1. Components of the transport system

What we define as the "transportation system" has multiple components and dimensions, and each one represents a discipline with its own field of knowledge. Accepting some degree of simplification, we could divide the transportation system in three components: 1) the moving parts; 2) the fixed parts; and 3) the intangible parts.

The moving parts of the transportations system are those which actually move, together with the person or good to be transported. They provide locomotion, mobility, and protection. All these qualities can be grouped in the vehicle (e.g. a car, a bus, etc.) or might be provided by different elements (e.g. a container, on a trailer, towed by a tractor unit or train).

The fixed parts are composed of the paths and guideways used by the moving parts (i.e. the infrastructures). They conform the different transportation networks, with their links and nodes, using the terminology of graph theory. Links are necessary to overcome the distance. Nodes allow connecting links, accessing the network, containers changing vehicles, vehicles changing trains, etc. These nodes are transportation terminals, from a simple bus stop, to a large airport hub.

The level of accessibility of a transportation network is defined by its number of nodes with respect to the total length of the network. The more nodes, the easier the access of a person or good to the network. However, the number of nodes with respect to the total length is inversely proportional to the efficiency of the network (i.e. the ease and average speed at which the users can move in the network). In general, it is easy to travel through links, while difficulties and delays appear when crossing nodes. Bottlenecks appear generally at nodes, creating congestions that spread along the links. This means that there is a trade-off between the accessibility and the efficiency of transportation networks. High access networks are slow, while fast networks need to have a limited accessibility. As an example of this trade-off, just think of the freeway network (fast with limited access) in relation to the network of city streets (slow but with unlimited access). Or similarly, the high-speed rail network with respect to the subway network. In this context, networks are configured with a hierarchical structure in order to make the transportation system efficient and accessible at the same time. The access to the system is accomplished at a lower-level network with high accessibility and then we use terminals to transfer to an upper-level network which provides an adequate efficiency for the characteristics of the trip.

Finally, the last component of the transportation system are the intangible parts. These represent how transportation systems are operated and traffic streams controlled while travelling through transportation networks. The present book mainly addresses this component of the transportation system, as described in the next section.

2. Outline of the present book

In order to operate and control transportation systems ensuring accessibility and efficiency, we must deal with the transportation "supply" and "demand" sides. Transportation supply represents what we make available to people (or goods) for travelling. This includes the infrastructures (i.e. the transportation networks, including all their components) and the transportation services (i.e. public transportation routes, schedules, and services; logistic services, etc.). In turn, transportation demand represents all the people and goods requiring to travel from some origin to some destination at a given time.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Supply side transportation modelling and analysis deals with how the performance of the supply side depends on the transportation demand. In this case, demand is treated as an input, and the objective is to define how to operate the transportation system so that this demand can be served efficiently. However, the actual demand depends on the performance of the transportation system. If the performance improves, it will attract more demand, the opposite being also true. This is precisely what is analyzed in the transportation demand modelling: how the demand for a system is affected by its performance.

In conclusion, transportation supply and demand represent and endogenous problem with bidirectional dependencies, so that in the operation of any transportation system, supply and demand are in equilibrium. The present book mainly addresses the supply side (i.e. Chapters 2 to 6), while an overview of demand modelling is presented in the last chapters (i.e. Chapters 7 and 8).

In particular, Chapters 2 and 3 are devoted to present two of the most fundamental and basic tools of the transportation operations field, namely the trajectories diagrams and the queuing diagrams. Chapter 2 presents the concept of a trajectory and how it can be represented in the time – space coordinate axis (i.e. resulting in the so called trajectories diagram). These diagrams are very helpful in analyzing problems related with the movement of vehicles (e.g. scheduling problems, coordination, etc.). Chapter 3 introduces queuing diagrams (i.e. input – output diagrams constructed from the evolution of the cumulative number of customers to cross one point with respect to time). These diagrams allow determining the excess accumulation and delays suffered by customers when crossing any restriction or service.

Then, Chapters 4 and 5 deal with traffic flow modeling. This is how to predict the evolution of the movement of a large number of granular elements (i.e. any type of vehicles, pedestrians) in a one dimensional guideway (e.g. a street, freeway, pedestrian sidewalk or corridor). Chapter 4 analyzes the macroscopic perspective of traffic flow modelling, analyzing the evolution of flows, densities and average speeds. Macroscopic modelling takes most of the emphasis in this book. This is followed by Chapter 5, introducing the microscopic perspective of traffic analysis which deals with the modeling of the actual trajectories of individual vehicles. Only the most basic car-following models are presented, omitting other components required to simulate traffic at the microscopic scale (e.g. lane-changing models).

Finally, Chapter 6 presents the basics behind the operation of scheduled public transportation services, including the effective vehicle dispatching, headway control strategies and the coordination of transfers. This chapter concludes the analysis of the supply side of the transportation system.

Then, Chapter 7 introduces behavioral modeling, which represents the foundations of transportation demand modelling. Emphasis is given to utility theory and how it can be applied in the context of the four-step Urban Transportation Planning (UTP) procedure. The Logit model for discrete choice analysis is analyzed with some more detail.

Demand modelling and the book itself is concluded in Chapter 8 devoted to the traffic assignment problem, which consists in assigning Origin / Destination demands to a particular transportation network, obtaining link flows. The Wardrop's principles of equilibrium are discussed, and some well-known traffic assignment paradoxes are presented (i.e. the Braess' paradox and the Smith's paradox).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

2 – TRAJECTORIES AND THE TIME-SPACE DIAGRAM

Contents

1. Problems are easier if one uses the right tools 2
2. Trajectories of a traffic stream on a time-space diagram 2
3. Using time-space diagrams in scheduling problems 5
4. Using time-space diagrams in traffic signal coordination 8
5. Other examples of trajectories 9
6. Constructing trajectories 11

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

1. Problems are easier if one uses the right tools

Imagine three friends who want to take a long trip using a tandem bicycle for two persons. Because the bike riders travel at 20 km/h, independent of the number of riders, and all three persons walk at 4 km/h, they proceed as follows:

- To start the journey, friends "A" and "B" ride the bicycle and friend "C" walks.
- After a while, friend "A" drops off friend "B" who starts walking and "A" rides the bicycle alone in reverse direction.
- When "A" and "C" meet, they turn the bicycle and ride forward until they catch up with "B".

At that moment, the 3 friends have completed a basic cycle of their strategy, which they then repeat a number of times until they reach their destination. Could you determine their average travel speed?

In my[1] lectures about trajectories and the time – space diagram, I usually start by posing the previous problem. Despite being easy to understand and only involving movements at constant speed, 3rd year bachelor of engineering students usually[2] cannot solve the problem in a 10-15 min period. By looking at the students' draft paper while trying to solve the puzzle, I observe that many students try to use some graphical constructions in order to formulate the equations which lead to the solution. Most of these graphics show different one-dimensional plots, plotting either the distance or time traveled by each one of the friends. While the engineering instinct of plotting the problems to help in visualizing the solution is well developed among students, they do not realize that the problem has two clear dimensions (i.e. time and space), and that these could be plotted together in two coordinate axis (i.e. a time – space diagram). Using the time – space diagram, the problem could be solved by all students in the given time.

This example illustrates how graphical tools, and in particular the time – space diagram, are helpful in visualizing the solutions of transportation operations problems. Even in the simplest problems, plotting a relevant figure may allow visualizing unnoticed errors or unveiling better solutions.

2. Trajectories of a traffic stream on a time-space diagram

A trajectory is the representation of the movement of one vehicle in space and time (i.e. x, t). Trajectories can be represented in a time – space diagram, namely the plane defined by two coordinate axis, where typically time, t, is represented in the horizontal axis, and space, x, in the vertical axis. This representation is convenient as all the details of the movement (i.e. position, speed, acceleration) can be intuitively visualized (See Figure 1). In particular, the slopes of the trajectories in a time – space diagram represent the instantaneous speed of the

[1] Nothing in this book is mine. The three friends and one tandem puzzle (as almost everything in this book) is taken from prof. Carlos F. Daganzo's book "*Fundamentals of transportation and traffic operations*" (1997). Elsevier, New York. Please refer to the Preface for the credits.

[2] Prof. Daganzo states in his book that only around 5% of the students achieve a satisfactory solution. My experience leads to a similar percentage.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

vehicle. Note from Figure 1 that not all the representations on the time – space diagram are physically feasible trajectories.

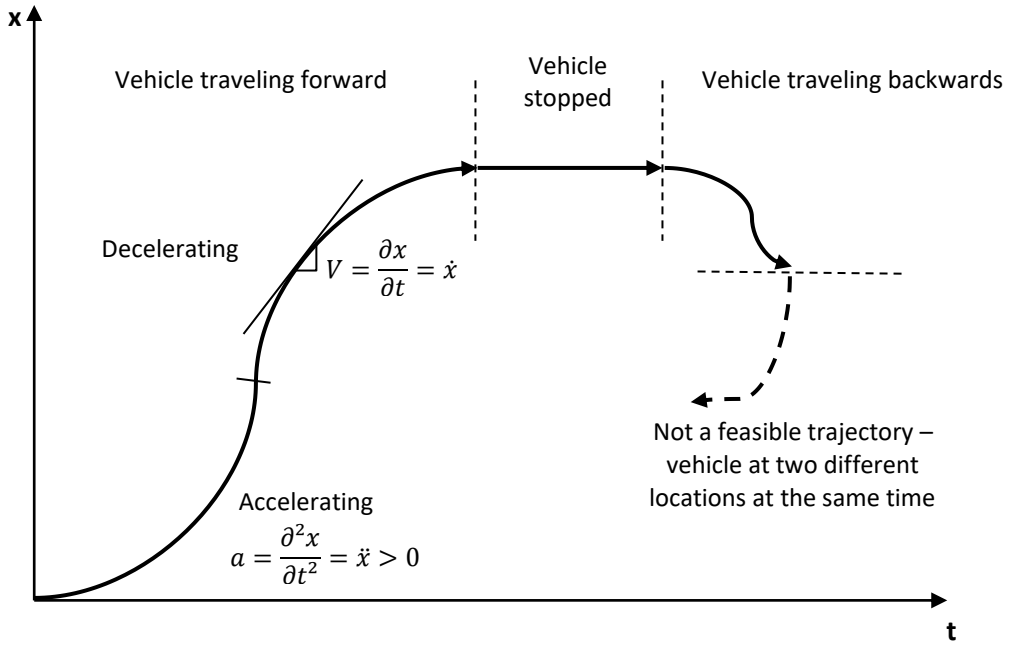


Figure 1. Trajectory definition on a time – space diagram.

In a single time – space diagram, many trajectories could be plotted, representing the movement of a traffic stream. Crossings of trajectories are possible, as far as the considered infrastructure allows different vehicles to be at the same longitudinal position at the same time. Just think of vehicles overtaking each other on the different lanes of a freeway, or the crossing of two vehicles in opposite directions in a conventional two-lane road. In spite of this, there might be situation where the crossing of the trajectories of two vehicles would imply a crash. Just imagine two trains travelling on a single railway track.

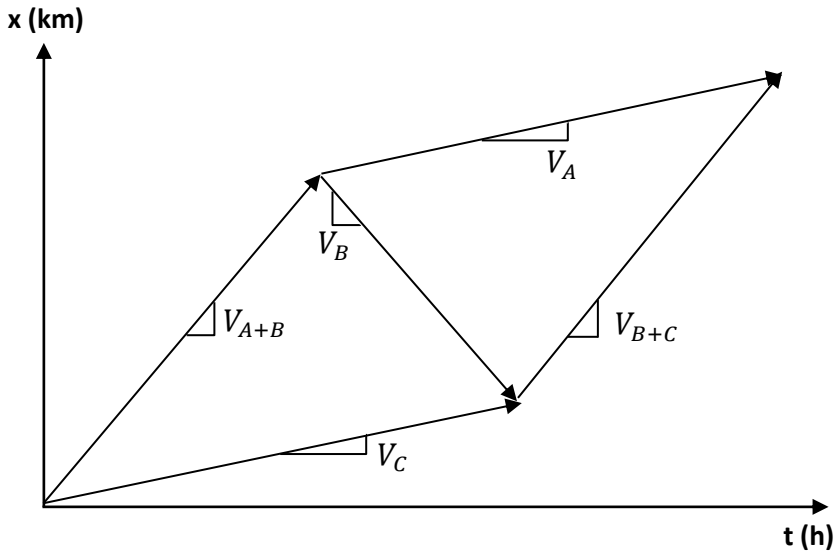
Note that in the time – space diagram, x represents one-dimensional space, typically the dimension corresponding to the longitudinal component of the movement. There might exist transportation systems where the longitudinal dimension of the trip is the vertical one (e.g. elevators in a building).

Using the time – space diagram, the trajectories of the three friends travelling with the tandem bicycle would look like the plot in Figure 2, and the average speed of the group of friends is easily obtained as shown in Figure 3.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

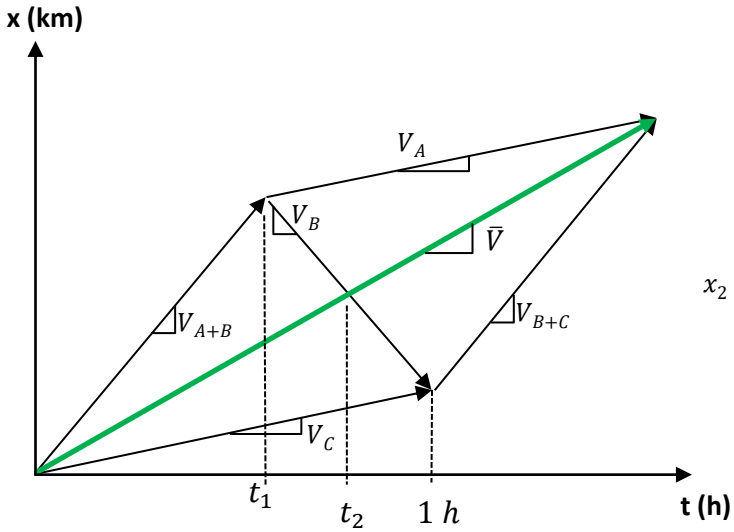
Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >



Trajectory	Speed [km/h]
$V_{A+B} = V_{B+C}$ (bike)	20
V_B (bike)	-20
$V_C = V_A$ (walk)	4

Figure 2. Time - Space representation of the problem of 3 friends and one tandem bike.



$$20t_1 - 20(1h - t_1) = 4 \rightarrow t_1 = 0.6\ h$$

$$t_2 = 0.6 + \frac{0.4}{2} = 0.8\ h$$

$$x_2 = 20 \cdot 0.6 - 20 \cdot 0.2 = 12 - 4 = 8\ km$$

$$\bar{V} = \frac{8\ km}{0.8\ h} = 10\ km/h$$

Figure 3. Solution of the problem of 3 friends and one tandem bike.

The use of the time – space diagram helps in visualizing the problem and determining its solution. This allows and efficient formulation of the equations to solve the problem and minimizes the chances of errors.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

3. Using time-space diagrams in scheduling problems

Plotting trajectories in the time – space diagram is especially helpful in scheduling problems. As an example, imagine that the problem to solve is the dispatching of train expeditions between two cities linked by a single railway track. Note that in this case, trajectories cannot cross. The solution would look like the plot in Figure 4. Note that this solution is simplistic, as acceleration / deceleration of the train is neglected, assuming that the time lost is small compared to the total trip duration. Then, the train travels at its maximum technological speed, V_{tech} (considering the infrastructure and vehicle technological capabilities) during the whole trip. In contrast, the dwell time at both terminals, necessary for the passengers to alight and board, and to clean and prepare the cars of the train for the next trip is included.

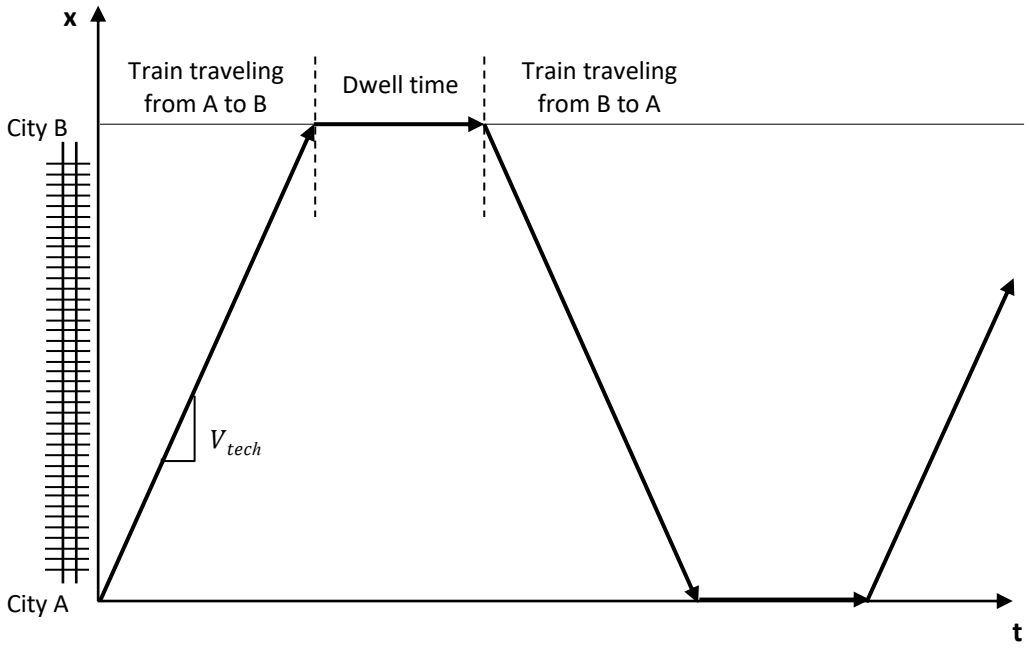


Figure 4. Train trajectories in a single railway track.

A subsequent question that can be addressed from the previous solution is to determine the capacity of the railway line. The capacity of transportation systems is defined as the maximum flow of customers (i.e. [customers/time]) or vehicles (i.e. [vehicles/time]) that can be served. This is:

$$Capacity = \max flow \left[\frac{pax}{h}\right]; \left[\frac{veh}{h}\right]$$

For our single railway track example, we have:

$$Capacity = \max\ vehicle\ flow \left[\frac{expeditions}{h}\right] \cdot train's\ passenger\ capacity \left[\frac{pax}{expedition}\right] = \left[\frac{pax}{h}\right]$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

The maximum number of train expeditions per hour can be obtained from the time – space diagram, by trying to fit as many trajectories as possible while fulfilling all the technological, safety and any other contour conditions. Equally, one could determine the minimum headway, h_{min} (i.e. the minimum time interval between consecutive expeditions) that can be maintained. By definition, the maximum trains' flow is equal to the inverse of h_{min}. This is:

$$\max\ vehicle\ flow\ \left[\frac{expeditions}{h}\right] = \frac{1}{h_{min}}$$

With the help of the time-space diagram it is also easier to visualize different ways in which the capacity of the single railway track could be increased. Possibly the most evident alternative is to build a sidetrack (e.g. in the midpoint between the two cities) so that trains can cross at this point, as illustrated in Figure 5. With the construction of such a sidetrack, the capacity of the line would double. Note that this context would be equivalent to the scheduling of ships to cross a channel with a single widening in the middle (see Figure 6).

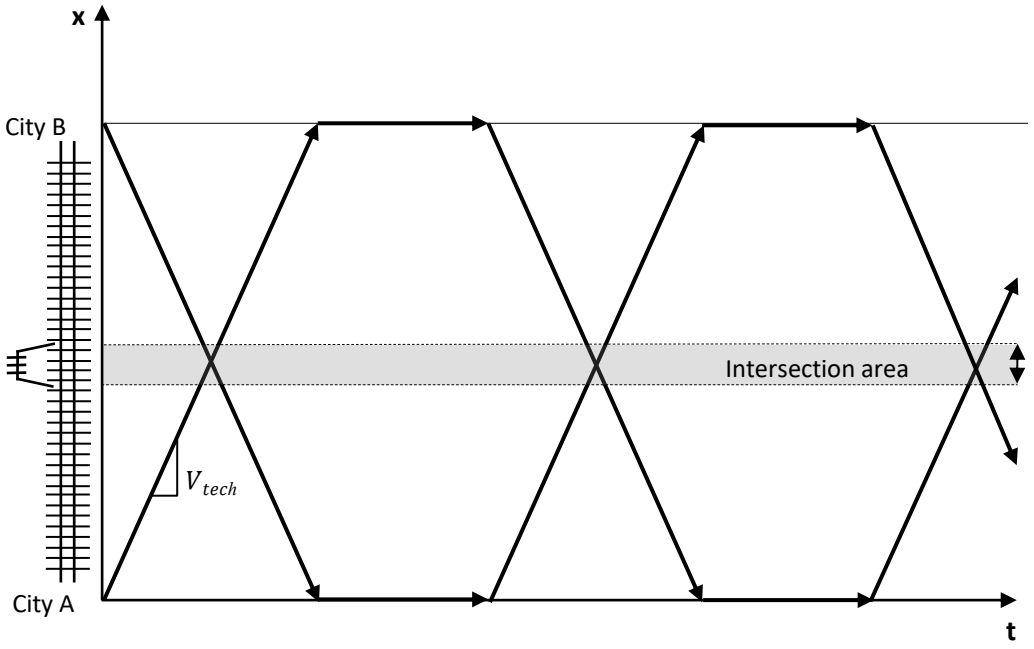


Figure 5. Increasing capacity of a single railway track.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

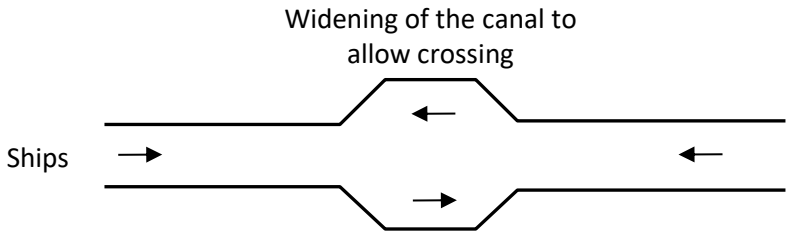


Figure 6. Other scheduling examples: Ships on a one way canal.

Returning to our railway example, more sidetracks could be built along the line, allowing more crossing points and fitting more expeditions per unit time, and therefore increasing the capacity of the line. Time – space diagrams would also help in determining how many sideways, where they should be located, and when to schedule trains. However, the scheduling of the additional expeditions would become increasingly challenging, while the marginal gains (i.e. the gain of each additional sidetrack) would decrease. At some point it would pay-off to construct a double railway track to further increase the capacity, so that each direction of travel is independent of the other. In this scenario, the capacity of the line in each direction would only depend on, V_{tech} (i.e. higher travelling speed implies higher capacity) and on the safety distance necessary between expeditions. Still, sidetracks would be necessary. Just imagine that the scheduling problem includes fast passenger trains traveling together with slower freight trains using the same line. Freight trains may need to divert to the sidetrack to allow being overtaken by the passenger trains, as shown in Figure 7.

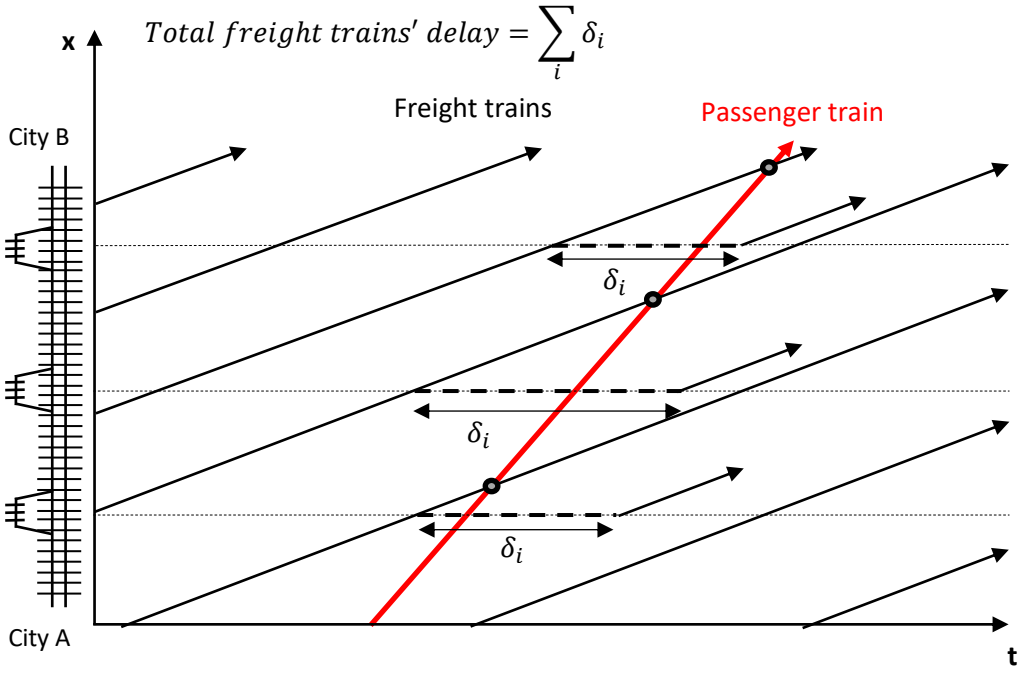


Figure 7. Scheduling trains on a one-way single track with sideways.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Also, the reduction of dwell times would allow an increase in capacity. Note that, in real practice, dwell times do not happen while blocking the main transportation railway link. There might be several platforms in the terminal where vehicles stop without blocking, so that if there are more vehicles available, they could depart once the main line is free.

4. Using time-space diagrams in traffic signal coordination

Traffic signal coordination is another typical scheduling problem where trajectories on the time – space diagram help in visualizing the solution. For example, Figure 8, shows a green wave for cars travelling on a signalized arterial with multiple intersections. The green wave is constructed so that the start of the green at each consecutive intersection is offset a time equal to the free-flow car travel time, from the start of the green at a reference signal (e.g. at Intersection 1).

Note that in case that some bus lines travel also on the arterial, and let's say that there could be one bus stop in the middle of each block between intersections, the signals would not exhibit coordination for the buses, and possibly they would need to make multiple stops. This is because buses have a very different travel time on each leg of the arterial, caused by the time dwelling at stops. Figure 9 shows how the green wave could be adapted to give priority to public transportation, instead of cars, by adapting the green wave to the commercial speed of buses (i.e. their speed including the dwelling time at stops). These pre-timed green waves for public transportation would be, however, unreliable, because of the random nature and high variance of the time buses dwell at stops. This means that the pre-timed offset could be sometimes too short, and sometimes too long to achieve signal coordination for buses. This is the main reason why pre-timed bus signal coordination is not used in practice, and instead active systems are proposed. These are based on the communication between the bus and the signal, which provides green extensions if a bus is approaching and the signal is about to turn red.

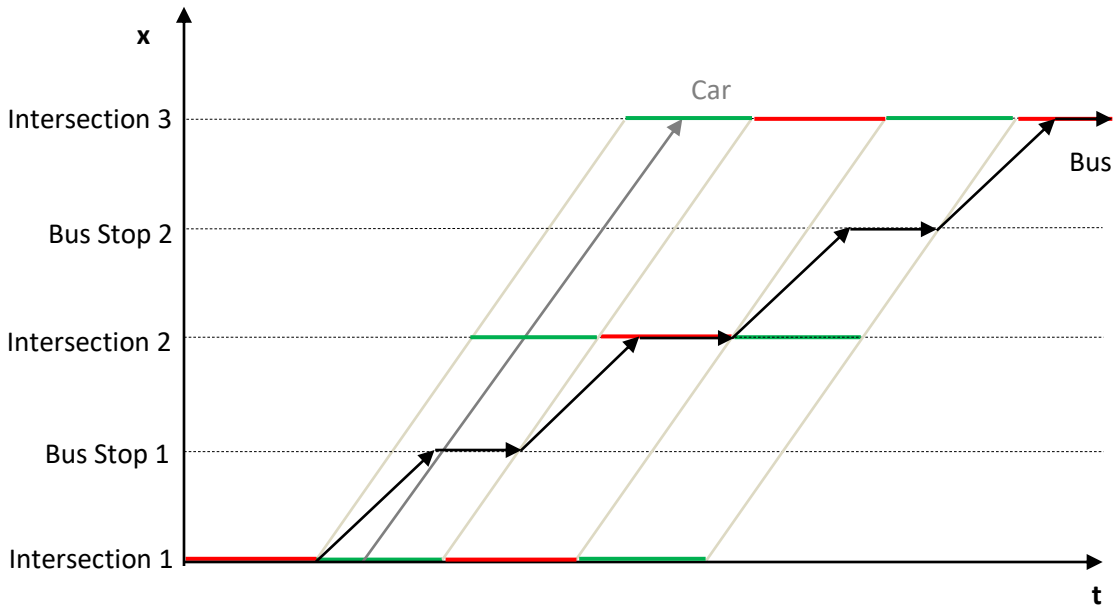


Figure 8. Signal coordination: Green wave for cars.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

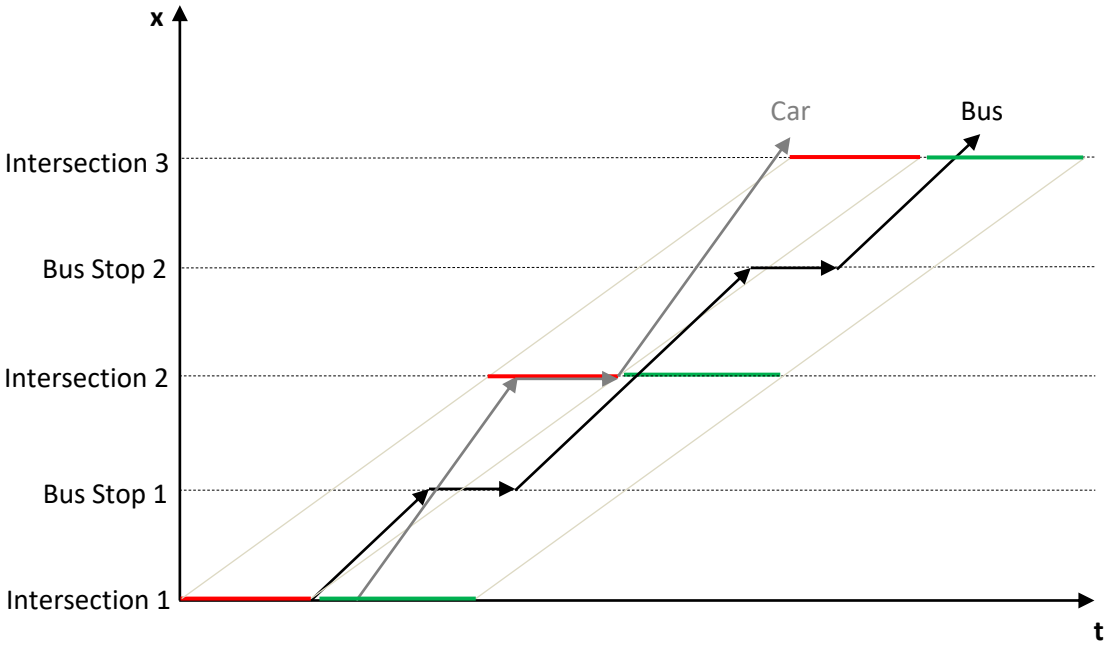


Figure 9. Signal coordination: Green wave for public transportation.

5. Other examples of trajectories

The examples and types of problems that can be visualized by plotting trajectories on a time – space diagram is as large as the number of problems involving longitudinal movement of vehicles in any infrastructure. Only two more examples are provided here.

Figure 10 shows the trajectory of a single bus on two round trips on his route. Note that when the longitudinal movement is circular, the start and the end location of the vehicle are the same. So, in Figure 10 position $x = 0 = L$, and every cycle is visualized as a "jump" in the time space diagram. Figure 10 illustrates also the concept of the commercial speed of a public transportation system. By definition, the commercial speed is the average speed of a public transportation system including the time needed at stops. It is the average speed experienced by the traveler. In contrast, the cruising speed it is the average speed at which the vehicle moves in the infrastructure, without considering the stops.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

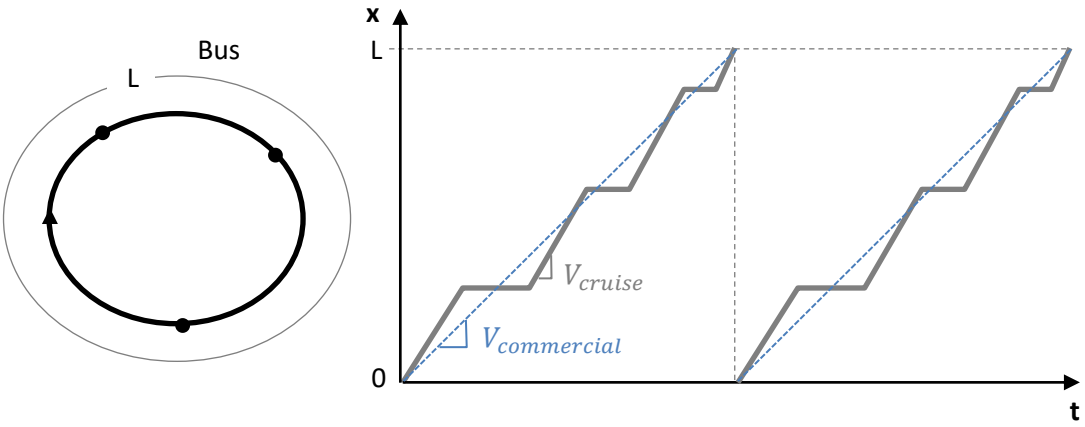


Figure 10. Bus trajectory. Two cycles on a circular route.

Figure 11 shows the trajectories of one landing and one taking-off plane operations in an airport runaway. The trajectories help in visualizing the minimum safety constraints, either in time or in space, that these operations must fulfill. These constrains are the ones steering the capacity of the runaway. Note that the trajectories of the plane in the runaway may leave-off/touch-down at some point, but still the whole trajectory of the longitudinal movement of the plane is represented while it is above the runaway.

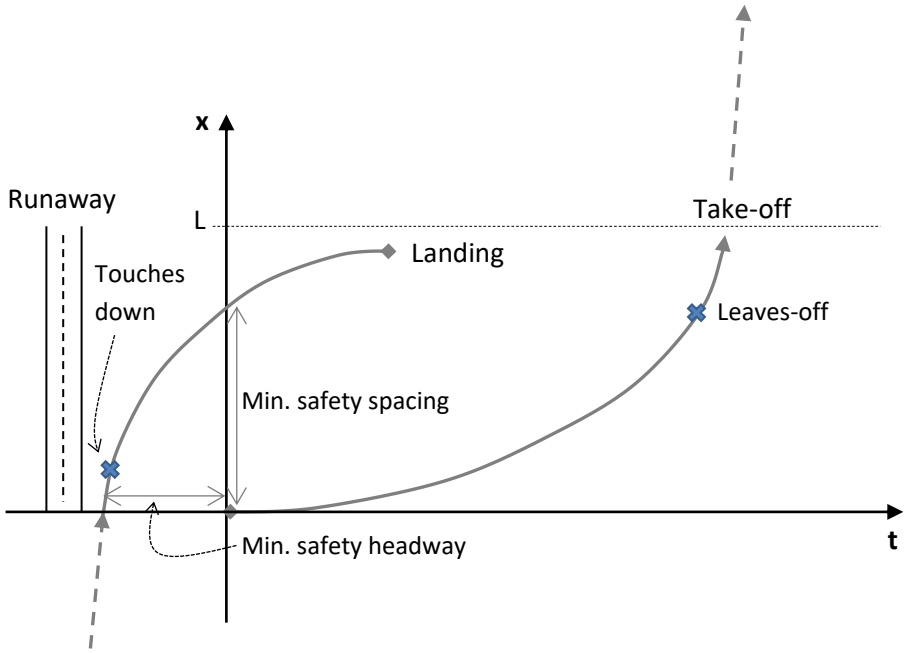


Figure 11. Landing and take-off operations on a time – space diagram.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

6. Constructing trajectories

One trajectory is the most complete representation of the movement of one vehicle. Measuring a real trajectory implies tracking the vehicle to determine its location as a function of time (i.e. $x(t)$).

Considering the traditional infrastructure based vehicle surveillance (e.g. traffic detectors located at some particular spots on the infrastructure), trajectories could be constructed by measuring $t(x_i)$ of individual vehicles at the discrete x_i corresponding to the detectors' locations. Vehicles' should be reidentified at consecutive detectors (or assume First-in / First-out) in order to construct the trajectory by interpolation (e.g. linear interpolation). Clearly, the accuracy of the obtained trajectory would grow with the density of detectors. Figure 12 illustrates this trajectory measurement process for three different vehicles.

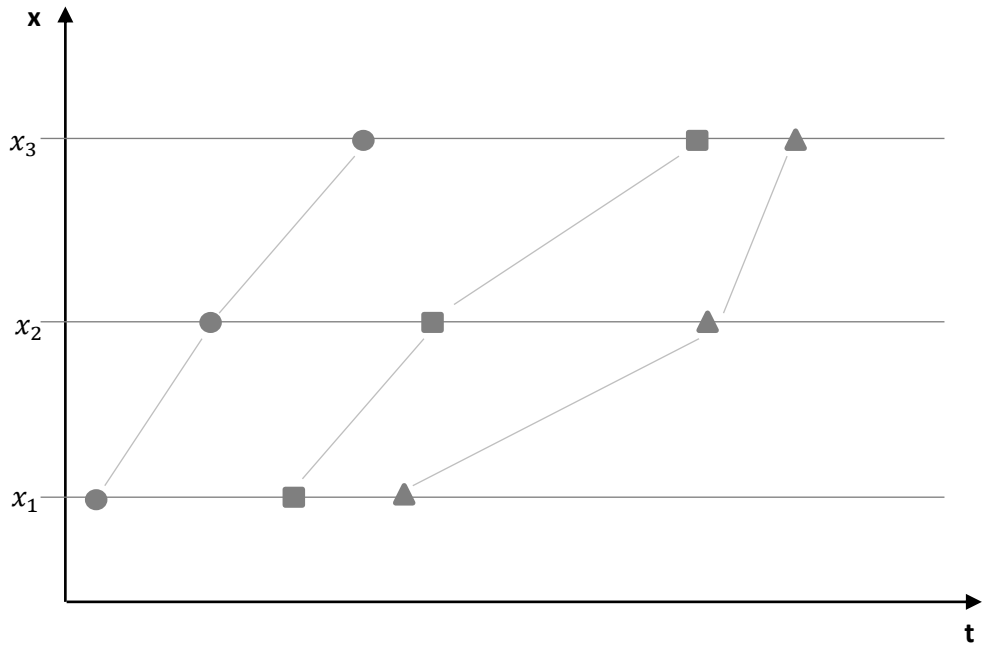


Figure 12. Constructing trajectories from roadside observers / detectors.

Analogously, trajectories could also be constructed by measuring $x(t_i)$ of individual vehicles at discrete t_i corresponding to different "instantaneous pictures" of the whole length of the infrastructure under analysis. In some contexts, these measurements are easily available. Imagine an elevators' control center of a tall building where the control panel shows the positions of all the elevators every t_i. As before, trajectories are constructed by interpolation by reidentifying vehicles (or elevators) in consecutive pictures, as in Figure 13. The accuracy in this case grows with the updating frequency of the "pictures".

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

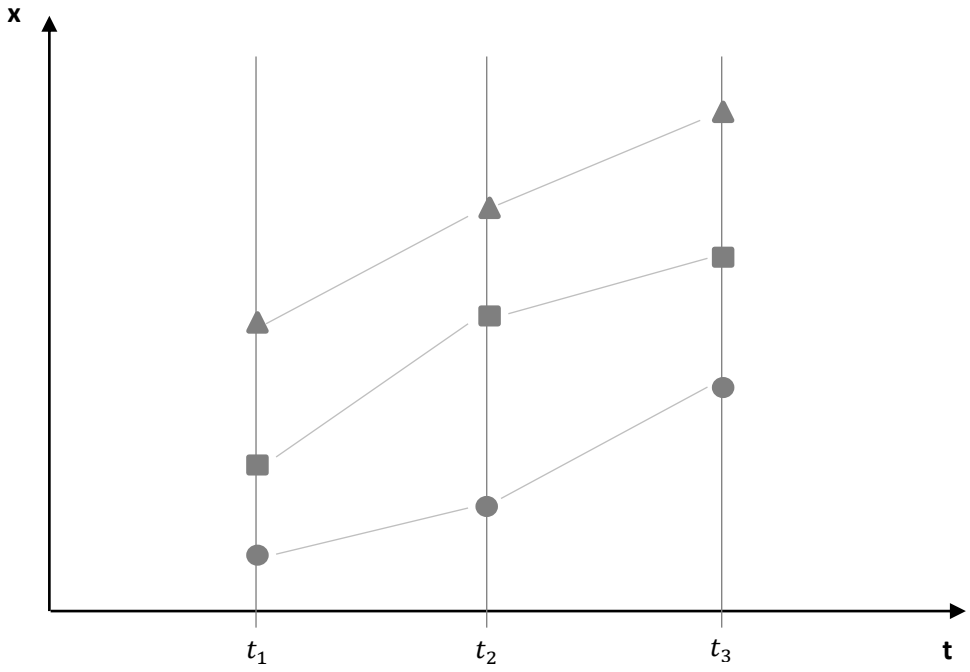


Figure 13. Constructing trajectories from successive instantaneous aerial pictures.

So far we have seen that trajectories can be constructed by measuring continuously in time but discrete in space (i.e. horizontal slices of the time-space diagram), or discrete in time but continuously in space (i.e. vertical slices of the time-space diagram). Alternatively, we could also think in constructing trajectories by measuring from discrete dispatching of moving observers (i.e. which are not static, neither in space nor in time, as in the previous examples). Moving observers can keep track of the crossings with other vehicles while traveling, and trajectories could be constructed by reidentifying these vehicles from the measurements of different observers, as shown in Figure 14. In this case, more data points would be obtained (and thus the accuracy of the trajectory estimation would grow) with more observers and when the speed of the moving observers, V_0, was significantly different than the average speed of traffic.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

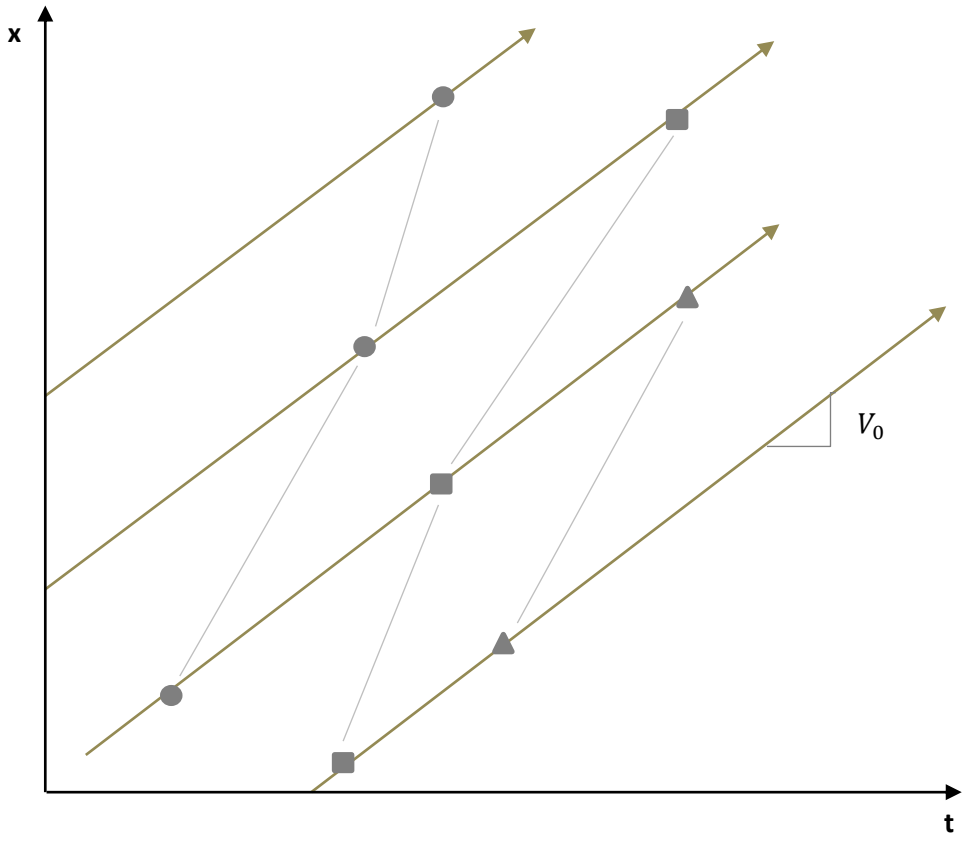


Figure 14. Constructing trajectories from moving observers.

In spite of the previous conceptual approaches for measuring trajectories in traffic streams, today's existing technology allows almost a continuous tracking of vehicles. Trajectories can be measured with much accuracy, while the temporal or spatial "slicing" of the time – space plane becomes imperceptible. Take as an example the video imaging techniques, which allow extracting the vehicles' trajectories from video recordings of a stretch of infrastructure if taken from an elevated platform[3]. Also, the surge of in-vehicle surveillance devices (e.g. navigation apps within the vehicle or in the smartphone) allows the tracking of the vehicles with an extremely short updating frequency. In conclusion, constructing trajectories from traffic streams is easier today than it has never been before, and this will have an impact in the validation and development of our knowledge of traffic flows.

[3] Take as a particular example the Next Generation Simulation (NGSIM) project promoted by the USA Federal Highway Administration (https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

3 – FUNDAMENTALS OF QUEUING THEORY

Contents

1. Introduction to queuing processes 2
2. Components of a queuing system 3
3. Queuing disciplines 6
4. (N, t) Diagrams 7
5. Input-Output diagrams 8
6. 3D representation (x,t,N) 10
7. The virtual arrivals curve, $V(t)$ 12
8. The Little's formula 16
9. Constructing input-output queuing diagrams from incomplete information 17
10. On-off queuing systems 20
11. Serial or tandem queuing systems 22
12. Diverging queuing systems 24
13. A simple strategy to reduce delay 26
14. Stochastic effects in queuing systems 27
15. Centralization effects in queuing systems 31
16. Optimization in queuing systems 32
17. The psychology of waiting lines 35
18. Summary of strategies to improve queuing systems 38

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

1. Introduction to queuing processes

In our daily routines, we face multiple activities that involve queuing. Commuting to university or work, accessing the subway, buying a cup of coffee, taking the elevator… are only some examples of how familiar we are with queues. Queues appear when a flow of vehicles, persons or objects wants to receive a service that implies a restriction. Typically, services have a limited capacity or are offered only with some frequency, which may lead to queues, delays and waits.

The analysis of queuing processes hold a long tradition of research, and define a broad discipline in itself, generally referred as "Queuing Theory". If one dips into queuing theory without caution, she might rapidly become overwhelmed by complexity and mathematics. However, the truth is that the basic laws that steer queuing processes are simple. Only the techniques used to analyze them with detail and little simplification make the topic to appear as complex. This complexity of the detailed analysis is what drives many engineering applications, in practice, to end up using manuals and "recipe" like solutions (e.g. for this type of process, take this solution). This is a dangerous practice, as using "blindly" solutions not fully understood may lead to counterproductive results. In fact, some concepts in queuing theory might seem counterintuitive. For instance, doubling the service rate does not reduce the queue to half, or queues that might exist even if the average demand rate is lower than the average service rate (i.e. due to stochastic queuing).

This chapter presents the fundamentals of queuing theory. The objective is that the reader fully understands simple models and knows how to use adequate tools for obtaining full knowledge of what is going on. This knowledge will unveil the adequate strategies to implement.

Most of the complexities in queuing theory come from the stochastic[1] nature of the processes involved (e.g. demand and service rates). However, most of the phenomena can be explained to a significant degree of accuracy by simplifying the reality, and assuming deterministic[2] processes. In this chapter, the focus is going to be on deterministic queuing, knowing beforehand that the reality is stochastic and that our approach will not be able to pursue the numerical accuracy of the solution. However, the order of magnitude and the conceptual behavior of the solutions will be correct, and these might be more than enough in many applications. In addition, the chapter highlights, when necessary, the effects of stochastic queuing, and when this needs to be taken into account into the solutions.

It is necessary to encourage further the practical use of the queuing theory concepts presented in this chapter, as they could lead to huge savings and avoid operational disasters. Remember that in transportation operations there is always one customer being served. Those who think, "well, I am causing delays, but I do pay nothing, so I do not care", are completely mistaken. Whenever as a manager you face queuing situations, recall a former FedEx advertising, saying "waiting is frustrating, demoralizing, agonizing, aggravating, annoying, time consuming, and incredibly expensive".

As an end note to this introductory section, I want to highlight that in case of services to people, it is important not only to focus on the objective measurable variables (e.g. waits and queue lengths), but also on how these queues and waits are perceived by customers. Human psychology plays a role, and a 2 minute wait can feel like

[1] One variable or process is said to be stochastic when it follows a random probability distribution or pattern that may be analyzed statistically, but may not be predicted precisely with one value.
[2] One variable or process is said to be deterministic when it may be described precisely with a single value.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

nothing, or it may seem to last forever. The last section of this chapter will be devoted to the psychology of waiting.

2. Components of a queuing system

The typical and most elementary queuing system is composed of a "server" which represents the restriction and a "storage" area where customers wait in case of queues. The queuing system is fed by an arrivals' process and yields a departures' process after service (see Figure 1). In transportation systems, the customer arrivals are random. Many servers exhibit random service rates too (e.g. elevator arrival at a particular floor), although, in general, the variance of the service process is smaller than that of the arrivals' process. Even some services may approach a deterministic service rate (e.g. access turnstiles to a metro station). However, and as justified in the previous introduction, we will treat arrivals' and departures' processes as deterministic, acknowledging that we are neglecting the stochastic effects.

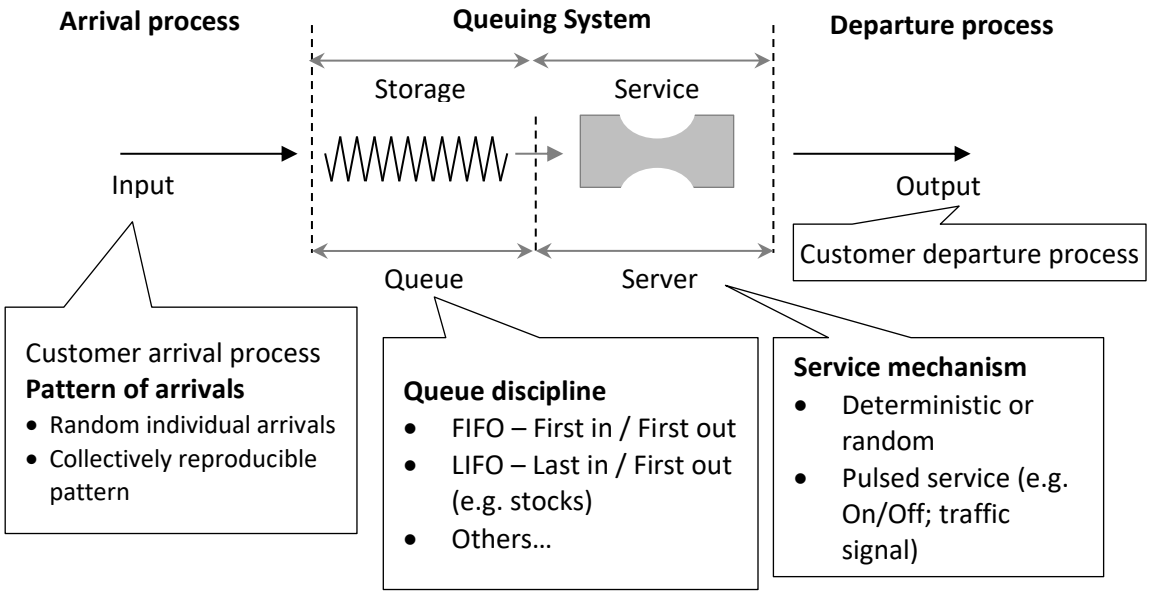


Figure 1. Elements of a queuing system.

A particular case of the previous elementary system appears when the customer departure process is what feeds again the arrivals process, defining a cycle. In this case, the departure rate is equal to the arrival rate. Cycles, which may appear often in industrial processes, are stranger in transportation operations. The closest examples I can think of would be skiers using ski lifts (i.e. once served, they ski down the slopes to queue again for the lift) or customers in amusement park rides.

In addition, queuing systems can be composed of more than one server. The "n" servers in the system could be in parallel (with centralized or decentralized queues) or in series (i.e. "tandem" queues), as shown in Figure 2.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

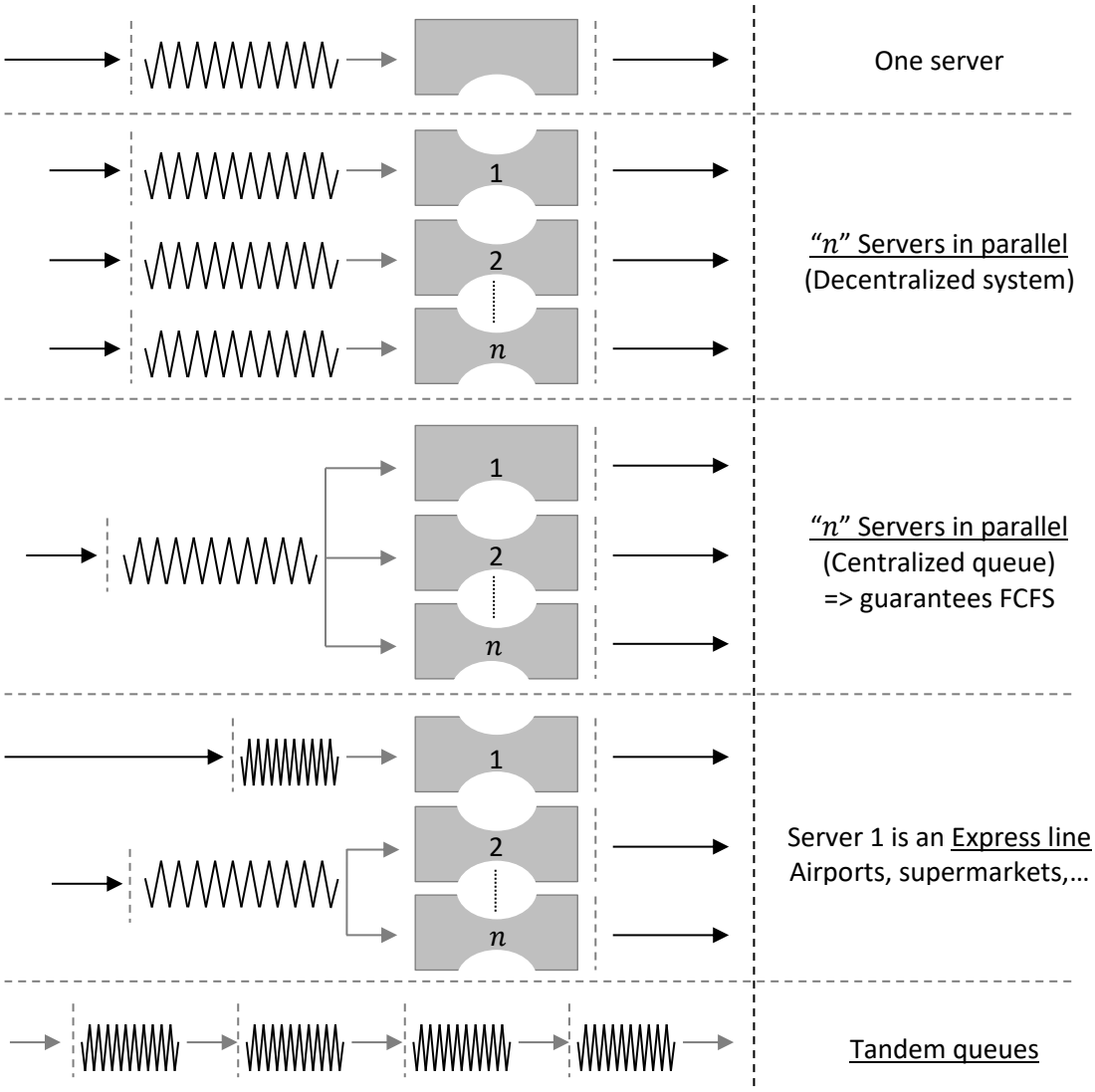


Figure 2. Different configurations of queuing systems.

Centralized queues in multiple server systems are gaining popularity nowadays (e.g. in supermarkets, bank offices, ticket counters …). This is because they exhibit some benefits over decentralized (i.e. independent) queues for each server. Note that in a centralized queue, servers are used more efficiently, as it is not possible that one server being idle while there is queue in the others[3]. We will show later in the chapter that the variance

[3] One might think that this is not a big issue because, even in a decentralized system, if this situation arises the customers could just change queue. Note, however, that this reasoning assumes that customers have information of the status of the other servers and that they can change servers easily. Both assumptions may not be fulfilled. Think for instance of some queuing system where queues are physically located out of sight, or where customers could not move easily, like big containerships waiting for a berth.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

of the arrival' process is reduced if the system is operated in a centralized manner, which allows serving customers more efficiently with less resources. In addition, centralized queues ensure first come / first served (i.e. FCFS) queuing discipline[4], eliminating the stress of customers in the selection of the "best" queue.

In centralized queuing systems, it is critical to implement an adequate information system that assigns customers to servers at due time. This assignment must ensure that there is no idle server while there is queue, and that the customers' wait at the servers is minimal. Note that this might be challenging, especially if it takes long time for the customer to move from the queue to the server (e.g. think of big containerships queuing outside of a port for a berth). In such situation, the end of the service of the previous customer needs to be anticipated and the customer assignment done with uncertain information. This might create a mixed system with a big centralized queue, and small decentralized queues at each server. In conclusion, the benefits of centralization would largely depend on the efficiency of this assignment algorithm.

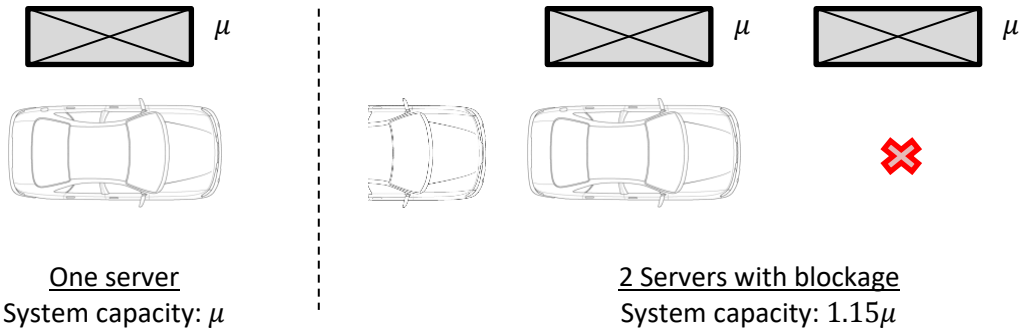


Figure 3. Petrol station blockage problem.

Another typical problem of queuing systems with multiple servers in parallel is that the servers do not fully work in parallel because of a bad design of the queuing storage area. It may happen that the queue blocks some servers, eliminating the full potential of additional servers. Think of a petrol station in a location with limited space. Imagine that there are two petrol pumps, one after the other in the same "lane" without the possibility of vehicles overtaking each other, like illustrated in Figure 3. Note that the capacity of the petrol station will be lower than that of two individual pumps, because one of the servers might be blocked most of the time. Actually, the detailed analysis of such scenario allowed concluding that the capacity of the two servers would be, in this context, only 1.15 times the capacity of a single server.

Serial queuing systems, often called "tandem" queues, are another type of systems with multiple servers. In these systems, customers need to go through several steps to receive complete service. Therefore, they need to go sequentially through several servers until they reach the departure of the system. Think of a customer reaching an airport to take a plane. First, she may need to collect the boarding pass; after, she might clear security, finally she reaches the boarding gate service. The particularity of the analysis of such systems is that the

[4] Note that the first come / first served discipline (i.e. FCFS) is slightly different from the first in / first out (i.e. FIFO). If service rates amongst different servers are different or random, FIFO would not be ensured even in a centralized system.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

departure process of one server constitutes the arrival process of the next. These systems will be analyzed in more detail in this chapter.

3. Queuing disciplines

Queuing discipline is a non-physical component of a queuing system. Queuing discipline refers to the rules which determine in which order are served the customers that are waiting. There most common queuing disciplines are described next.

- *First Come / First Served (FCFS)*. This is the most typical queuing discipline when dealing with human customers. Service is provided in the same order of arrivals. This order of priority is easily understood by the customers, so that it is recommended unless there is a very clear reason to not do so. Sometimes, First In / First Out (FIFO) terminology is used as a synonymous of FCFC. While this is accurate in case of a single server queuing system, in case of several servers in parallel, the first customer who starts services does not necessarily need to be the same as the first customer who goes out of the system. Just keep in mind that different servers may exhibit different service rates.
- *Last Come / First Served (LCFS).* This queuing discipline serves customers in the reverse order of arrivals. This queuing discipline arises whenever there is a "storage" of customers in a closed place, and the exit door happens to be the same as the entrance door. Think of a shuttle bus from the airport terminal to the plane, a direct elevator between two floors, or the products on the shelves of supermarkets[5].
- *Service In Random Order (SIRO)*. In this queuing discipline there is no preset order, and the next customer to be served is selected randomly from the queue. While this would be at least strange if customers waiting are humans, this is the most typical queuing discipline in industrial processes when the customers are different items in a production line, for instance.
- *Round Robin (RR).* This queuing discipline assigns to each customer in the queue an equal amount of service time. If this is not enough to complete the service, the customer returns to the end of the queue. So, the customer takes an equal share of the service capacity in turn. Communication networks and particularly the internet works on the basis of Round-Robin queuing discipline.

The previous queuing disciplines may be modified in particular cases, because of some special customers having priority amongst others. These priorities, which allow some customers to break the general queuing discipline, must be justified, and customers should understand that they are fair. If the priority is adequate from the service provider point of view, but might not be understood by every customer (e.g. paying more for the service and skip the wait), mixing both types of customers should be avoided. Some of the typical accepted priorities include:

- *Early Due Date First (EDDF):* This priority refers to an emergency situation. Customers with an "early due date" may be allowed to skip the general queuing discipline, because if they are not served in short time, they perish. This might apply to the priorities you apply to the products in your fridge, or to patients after the screening process when arriving at the hospital.

[5] When repositioning products, staff is asked to put the new products on the shelves behind the old ones, to avoid the LCFS queuing discipline. Knowing that though, some customers may take the products from behind, especially if they are perishable.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

- *Shortest Service Time First (SSTF):* This priority may be applied when there is a subset of customers with significantly lower service time than average. It can make sense to serve them first, because this would imply short additional time to the rest of customers and could avoid significant wait to them. Think of supermarket lines for customers with less than a number of items, or customers at a doctor office only waiting for a regular medicine prescription.

4. (N, t) Diagrams

(N, t) diagrams, also referred as cumulative plots, are the basic graphical tool to analyze queuing systems. The (N, t) diagram is constructed by plotting the cumulative number of customers, N, to cross a particular location as a function of time, t, as in Figure 4. Note that in order to construct the diagram it is needed to record the passage time of each customer (or vehicle, pedestrian, etc.) over the measurement location. For example, in Figure 4, customer i is observed at time t_i. By definition, the (N, t) curve is discrete, as N grows by one unit at the passage of every customer. Also, the (N, t) curve is an increasing function, as N is a cumulative measurement which cannot decrease. If nobody else goes through the measurement location from a given time onwards, N remains constant. Finally, recall that the average vehicular flow, q, measured at a particular location, is obtained as the total vehicular count over the duration of the observation period. This can be obtained from the (N, t) diagram as:

$$q_{(T)} = \frac{N(t_j) - N(t_i)}{t_j - t_i}$$

Where the observation period T is obtained as: $T = t_j - t_i$.

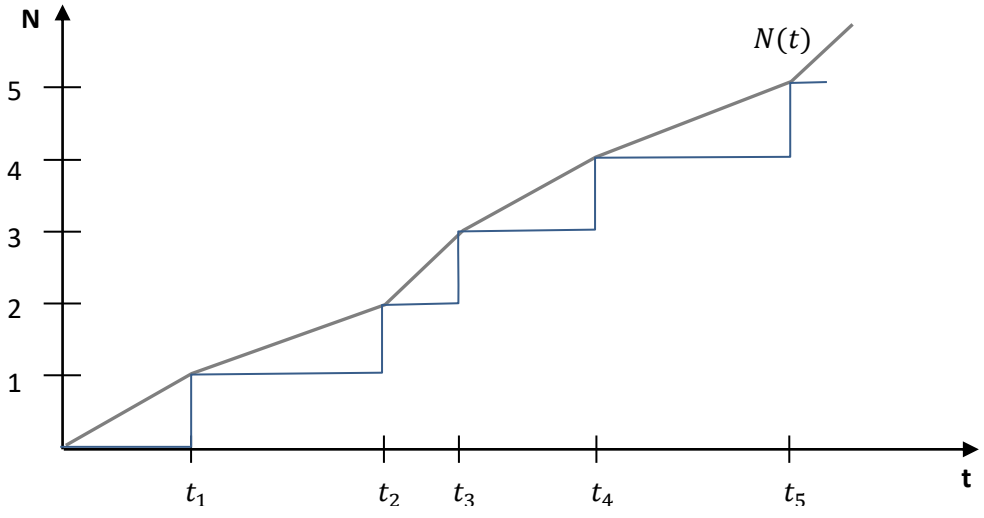


Figure 4. Constructing a (N, t) diagram from individual arrival times.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

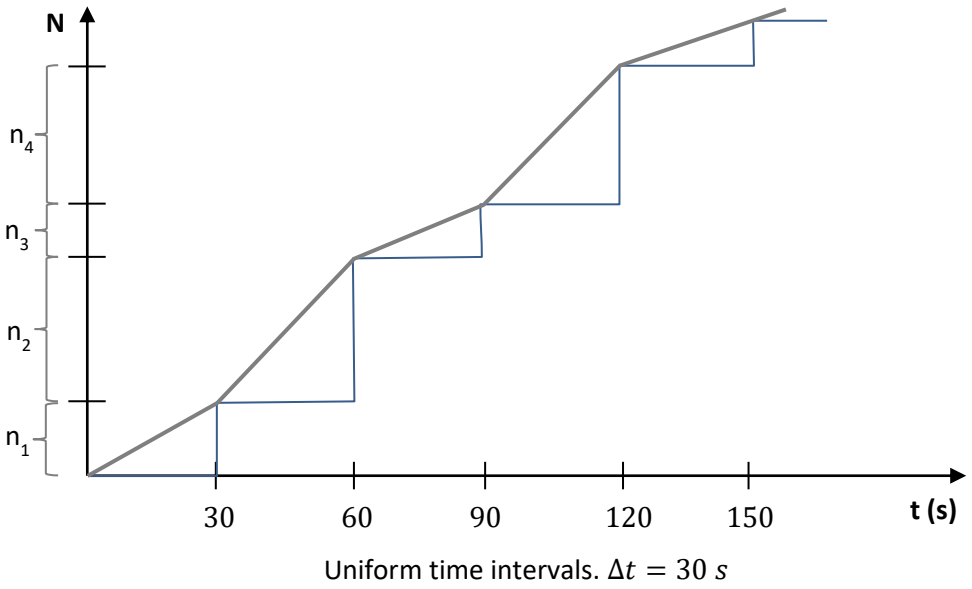


Figure 5. Constructing a (N, t) diagram from vehicle counts at uniform time intervals.

Sometimes, the used measuring devices do not measure the passage time of every individual vehicle, but it aggregates the number of observations every some period of time, Δt. In such cases, the (N, t) diagram can still be plotted, with a reduced level of precision, as in Figure 5. Note that in this case, the discrete jumps of N, are not of one unit, but of the number of observations in the time period i, n_i.

When the number of customers is large, like usually happens when analyzing transportation systems, the discrete jumps of the $N(t)$ curve at particular instants of time are not meaningful, and the stepwise function can be replaced by a continuous interpolation. The continuous approximation to the $N(t)$ curve allows defining the instantaneous flow, $q_{(t)}$ as the time derivative of $N(t)$, or equivalently, the slope of $N(t)$ at a particular time:

$$q_{(t)} = \frac{\partial N}{dt}$$

5. Input-Output diagrams

The usefulness of (N, t) diagrams in the analysis of queuing systems is more evident when working with input – output diagrams. An input – output diagram consists in the plot of two $N(t)$ curves in the same figure. The first of the curves, $N(t, x_0)$ is measured at location x_0, before vehicles join the queue (if any). In turn, the second curve, $N(t, x_1)$ is measured after the service (or bottleneck). This means that $N(t, x_0)$, also referred as $A(t)$, measures the arrivals to a potential queuing system, while $N(t, x_1)$, also referred as $D(t)$, measures the departures after service. In between, we have the potential delay and the service time.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

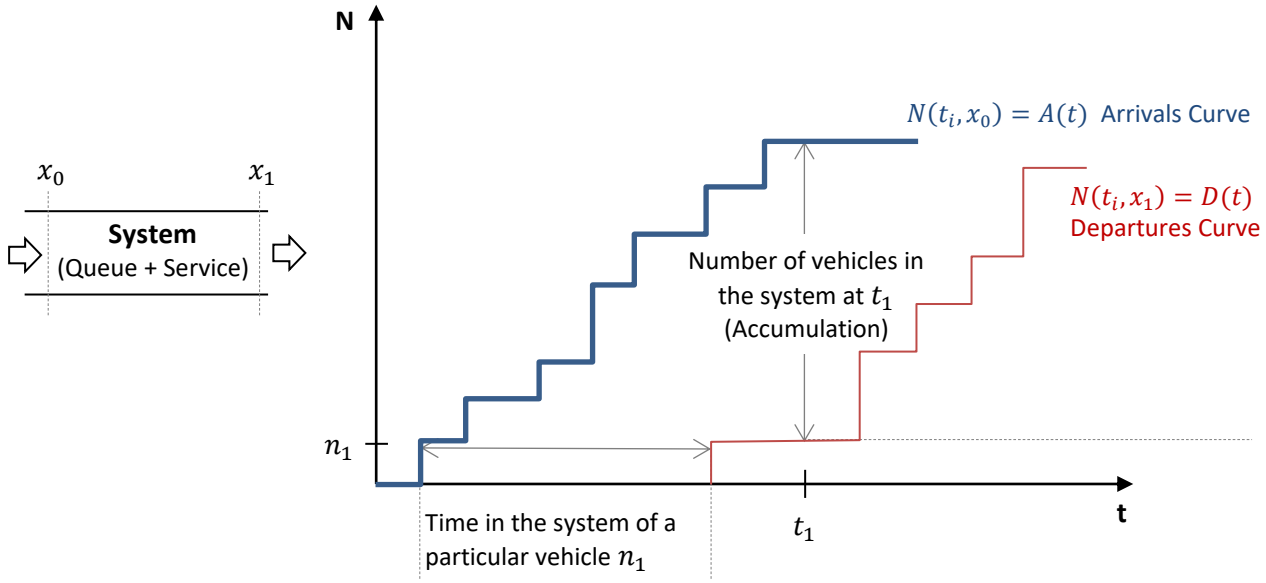


Figure 6. Accumulation and time in the system in a (N,t) diagram.

Figure 6 shows an input – output diagram. Note in this diagram that the time between $A(t)$ and $D(t)$ for a particular customer, n_1, represents the time in the system of this customer. The time in the system includes any potential delay[6] plus the free-flow service time. Also, the difference in the cumulative number of customers between $A(t)$ and $D(t)$ at a particular time, t_1, represents the customers' accumulation in the system. The customers' accumulation includes the customers which are waiting to be served, and those which are being served (i.e. in service). Note that $D(t)$ can never go above $A(t)$, because in order to exit, one customer needs to have entered before. Also, if both curves match, it means that the system is empty of customers. These definitions are clarified in Figure 7 where the input – output diagram is plotted together with the evolution of the number of customers in service and in queue. Note that Figure 7 refers to a queuing system with a single server which can serve one customer at a time.

[6] The customers' delay is defined as the difference between the actual time the customer spends between x_0 and x_1 (i.e. the actual time in the system) and the free-flow service time, defined as the service time any customer experiences in the absence of delay.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

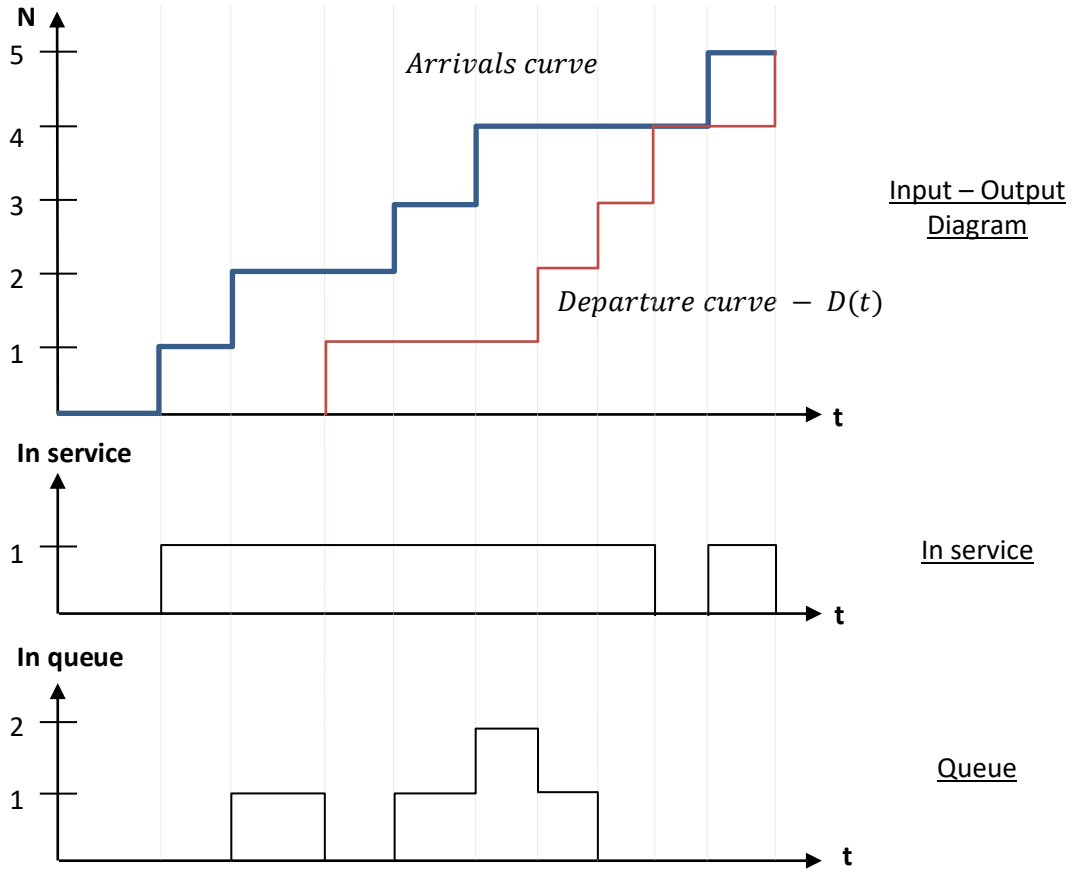


Figure 7. Input – Output diagram for a single server queuing system.

6. 3D representation (x,t,N)

The space-time (x, t) diagram presented in the previous chapter, and the (N, t) diagram analyzed here, conform the same reality and can be plotted together in a 3D coordinate axis (x,t,N) (e.g. Figure 8) and if plotted in 2D, they can be related as in Figure 9.

Note that Figure 8 is a representation of the trajectories of 5 vehicles, which could be visualized as a kind of staircase emerging from the paper. The projection of these trajectories on the (x,t) plane (i.e. in green in Figure 8) results in our typical (x,t) diagram. In turn, the projection of these trajectories on the (N,t) plane (i.e. orthogonal to the paper, in red in Figure 8) results in the stepwise (N,t) diagram. Note that Figure 8 shows the (N,t) projection at $x = 0$, but the projection at any other location could also be obtained, resulting in a different (N,t) curve. This is seen in Figure 9, with two (N,t) projections of the (x,t,N) plot at x_0 and x_1, resulting in an input-output diagram. Realize the correspondence between the times of any vehicle trajectory crossing locations x_0 and x_1, and the increase in one unit of the respective (N,t) curves.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

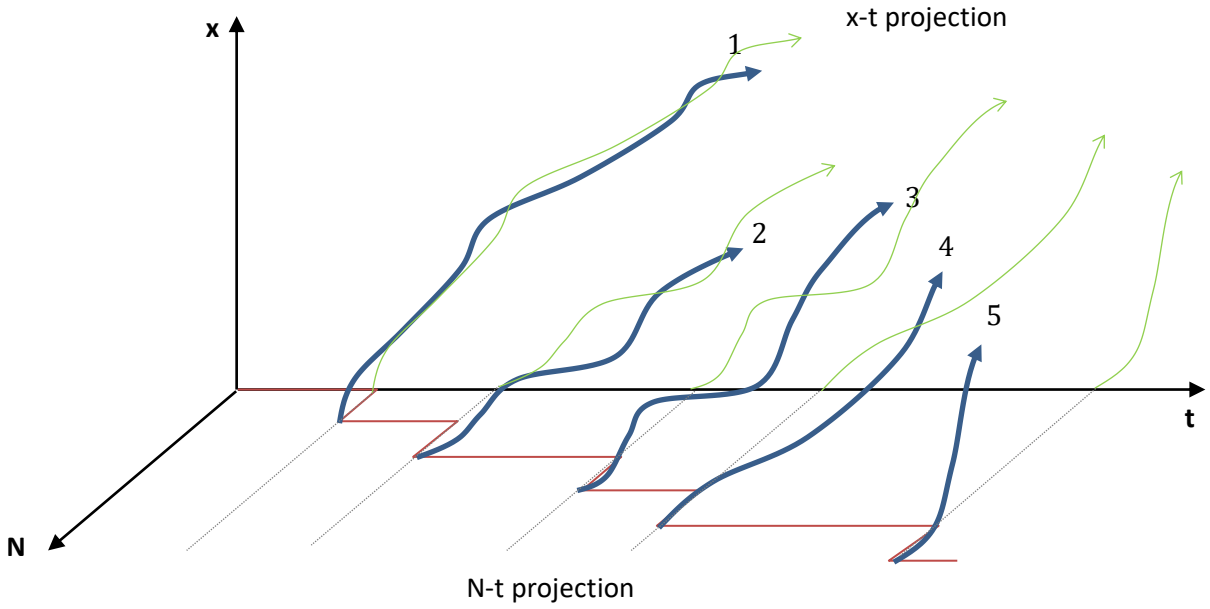


Figure 8. (x, t, N) 3D trajectories representation.

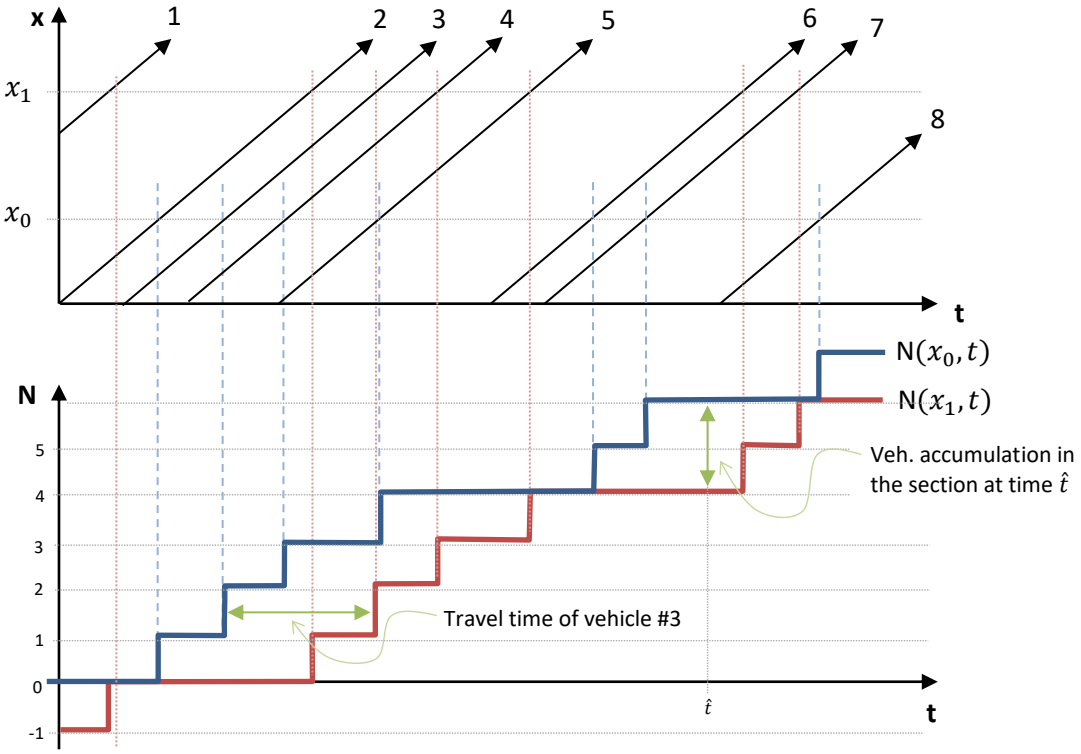


Figure 9. (N, t) from a (x, t) diagram.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

7. The virtual arrivals curve, $V(t)$

Figure 10 shows a typical input – output diagram of a queuing system which deals with a service to human customers. Note the S-shape of the inputs curve, resembling a logistic curve. This is the typical from of the arrivals curve resulting from human behavior when requesting any service. There is a peak period (in the middle) where demand is maximum, while in the extremes (i.e. before and after) the demand rate declines. As an example, think of the demand for lunch at the campus' cafeteria. There are some people who want to have lunch very early (e.g. At 12:30 in Barcelona), but they are few. As time advances, more and more people arrive to have lunch, and the demand rate (i.e. the arriving flow; the slope of the $A(t)$ curve) peaks (e.g. around 14:00 in Barcelona). From then onwards, demand starts declining, and very few people will have lunch after 15:30 (again, for the Barcelona example). This behavior, with clear peak hours yielding the S-shape arrivals curve, results from the vast majority of us wanting the same things at the same times. The departures curve follows the behavior of the arrivals curve, as customers who exit are those who have previously entered. However, there is a clear difference between $A(t)$ and $D(t)$. Note that the maximum slope of $D(t)$ is bounded by the capacity of the system, μ [customers/h] (i.e. the maximum number of customers that can be served by unit time), while, in general, the maximum slope of the arrivals curve is unbounded.

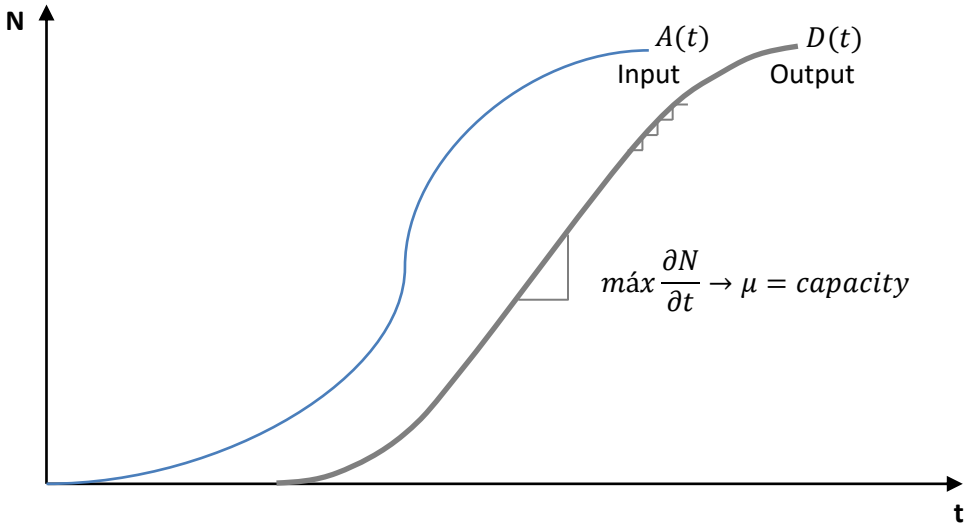


Figure 10. Logistic curves in an input – output diagram.

As described previously, from an input-output diagram (like the one in Figure 10), we can obtain the time and accumulation in the system. These include the service times and the delays, if any. Generally, queuing analysis is more focused on analyzing delays and excess accumulation (i.e. excluding the free-flow service time and the number of customers in-service). Therefore, it would be useful to modify the input-output diagram so that delays, w, and excess accumulation, Q, could be directly obtained. To that end, the virtual arrival curve, $V(t)$, is defined as the departures curve that would be measured in the absence of delay. Note that $V(t)$ is a hypothetical curve which cannot be measured (i.e. virtual) and that can be obtained by translating the arrivals curve, $A(t)$, forward in time a magnitude equal to the free-flow service time (see Figure 11). Then, by definition, the time

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

between $V(t)$ and $D(t)$ for any particular customer, j, represents the delay of customer j, w_j. Also, the difference in the cumulative number of customers between $V(t)$ and $D(t)$ at a particular time t, represents the excess accumulation[7] of customers in the system.

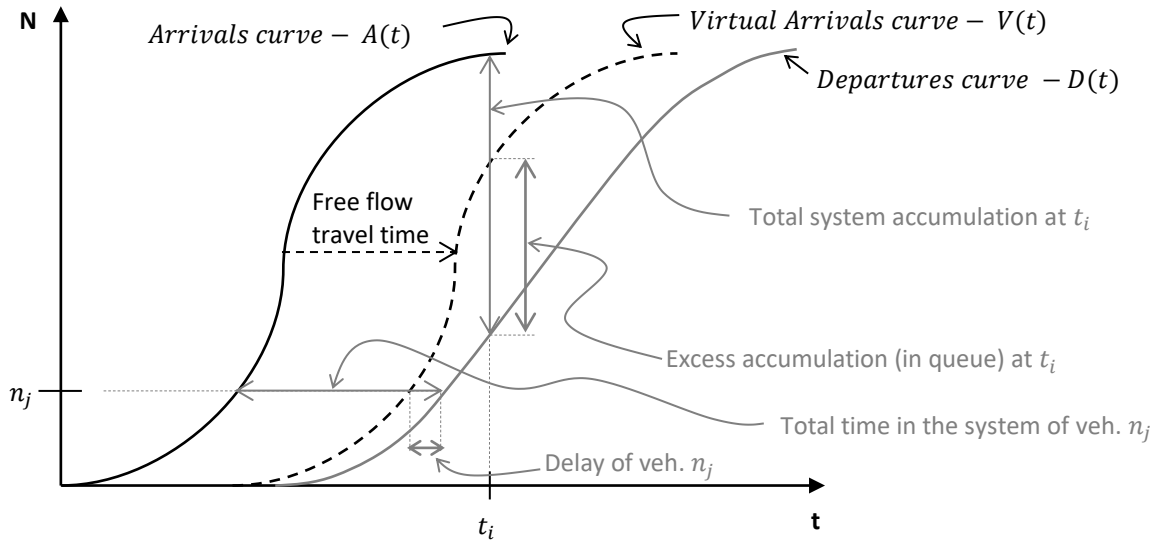


Figure 11. Interpretation of the Virtual Arrivals curve.

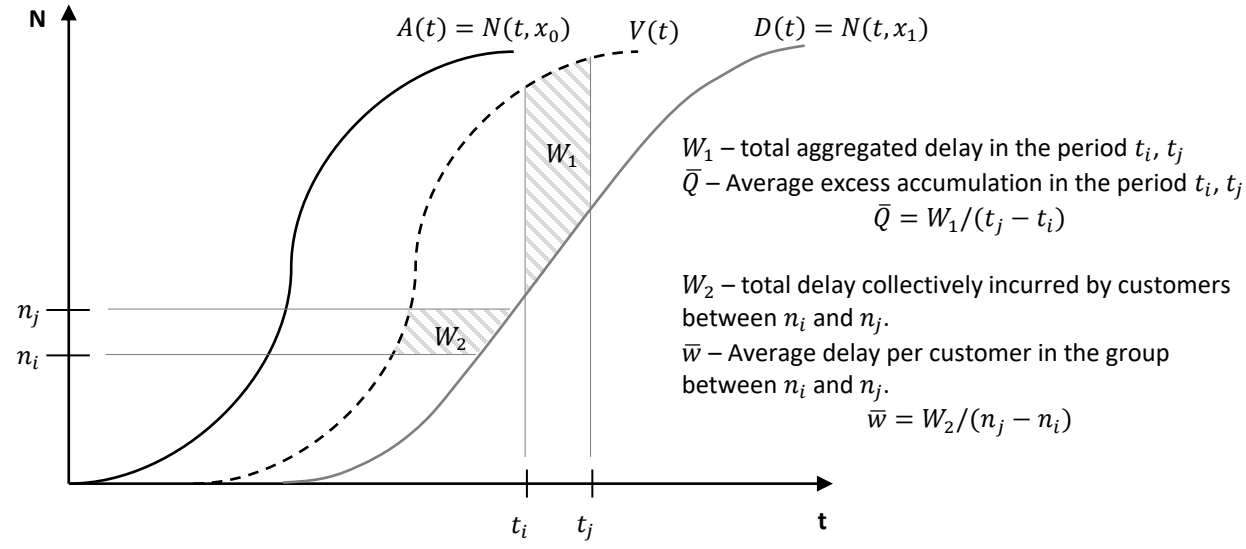


Figure 12. Average delay and average excess accumulation in an input-output diagram.

[7] The excess accumulation is defined as those customers that are in the system, but, in the absence of delay, would already have been served.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Generally, we are more interested in obtaining aggregated or average delays over an extended period of time or group of customers, than individual delays. To that end, we can compute the aggregated[8] delay, W, [customers·time] as the area enclosed between $V(t)$ and $D(t)$, for a given period of time or group of customers (see Figure 12). Then, the average delay, $\overline{w}$, is obtained as the total aggregated delay divided by the total number of customers affected. Similarly, the average excess accumulation during a period of time is obtained as the total aggregated delay divided by the duration of the period (see Figure 12).

The previous construction of the virtual arrivals curve in the input-output diagram holds for any queuing system. In general, the free-flow service time, either it is irrelevant for the queuing analysis or it is much smaller that the delay, so that the arrivals curve, $A(t)$, and the virtual arrivals curve, $V(t)$, are used indistinctively. In spite of this, if for some reason there is an interest in determining when the actual service begins and ends, it might be useful to modify the construction of the input-output diagram according to whether the service happens before (or simultaneously) with the queuing (e.g. freeway traffic when approaching a bottleneck), or after (e.g. queuing for a service, like checking out in a supermarket). Figure 13 and Figure 14 illustrate both constructions.

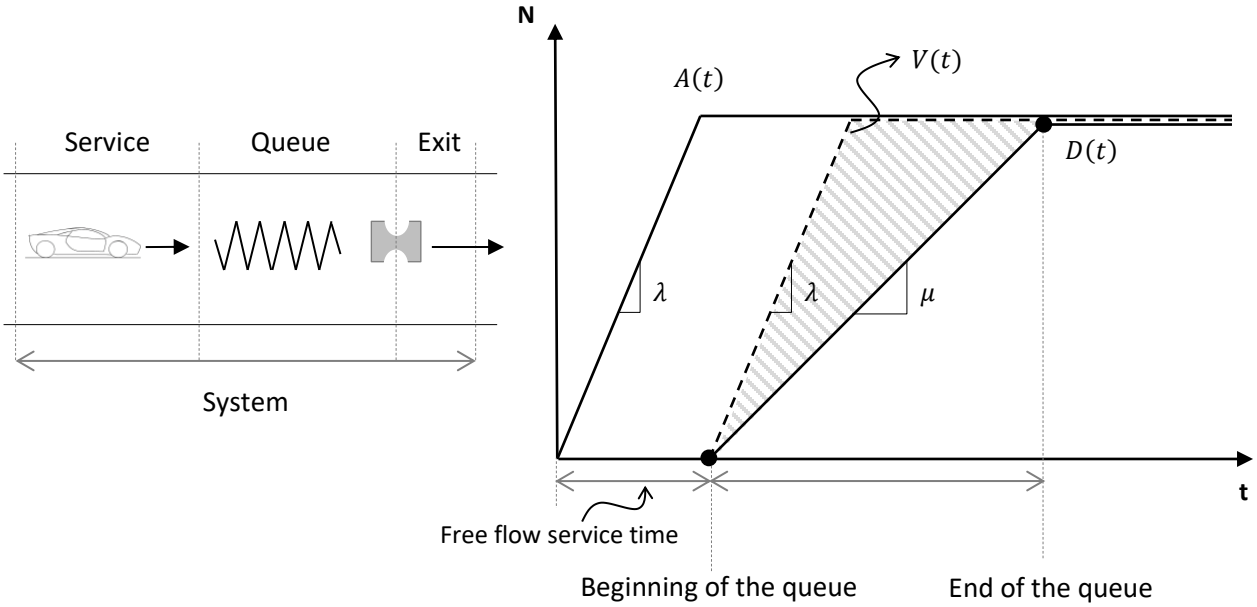


Figure 13. Virtual Arrivals curve. Queue after or during service (i.e. traffic queues).

[8] We use the capital W notation for the aggregated delay, and the lower case w notation for the individual delays.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

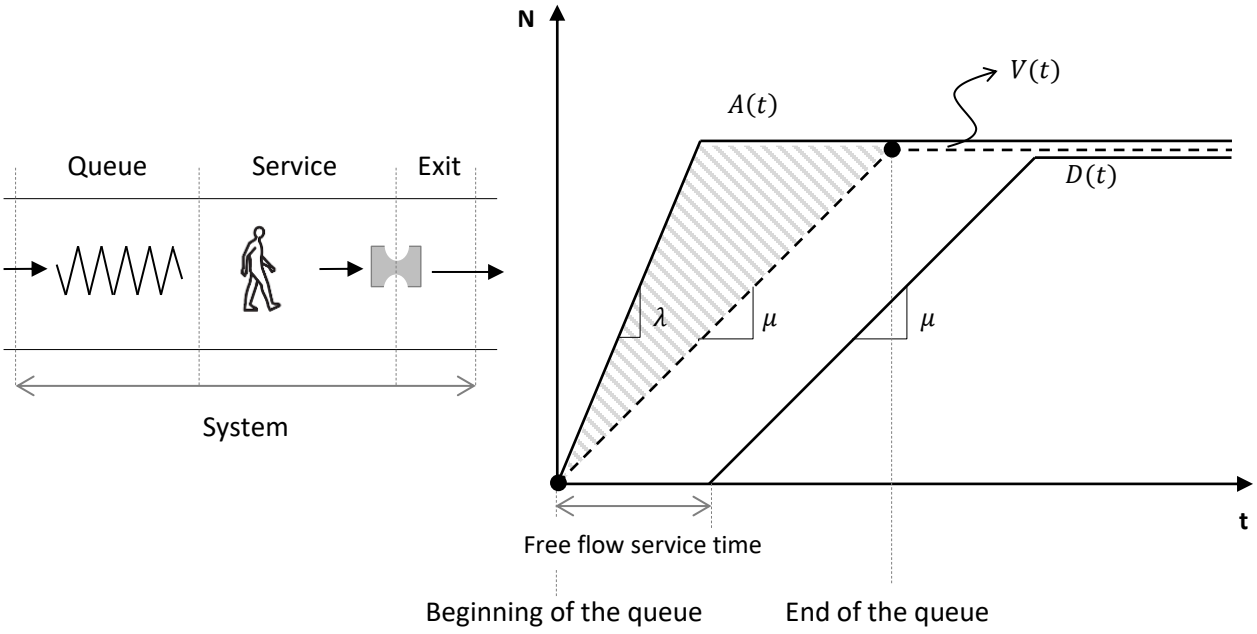


Figure 14. Virtual Arrivals curve. Queue before service.

Usually, the excess accumulation is mistaken for the number of customers in the physical queue. Note that this is not exactly correct, as the physical queue includes those customers who should have already been served in the absence of a queue (i.e. the excess accumulation) and also those customers who, although not being part of the excess accumulation (i.e. even in the absence of delay they would still be in the system) are in the queue because the queue takes physical space and blocks the location they should be in the absence of delay. This means that the number of customers in the physical queue is always larger or equal than the excess accumulation. Figure 15 helps in visualizing the differences in this common confusion. Similarly, the time in the physical queue is always larger than the delay, as the time in the queue includes the time needed to cover the queue length at the free-flow speed (see Figure 16).

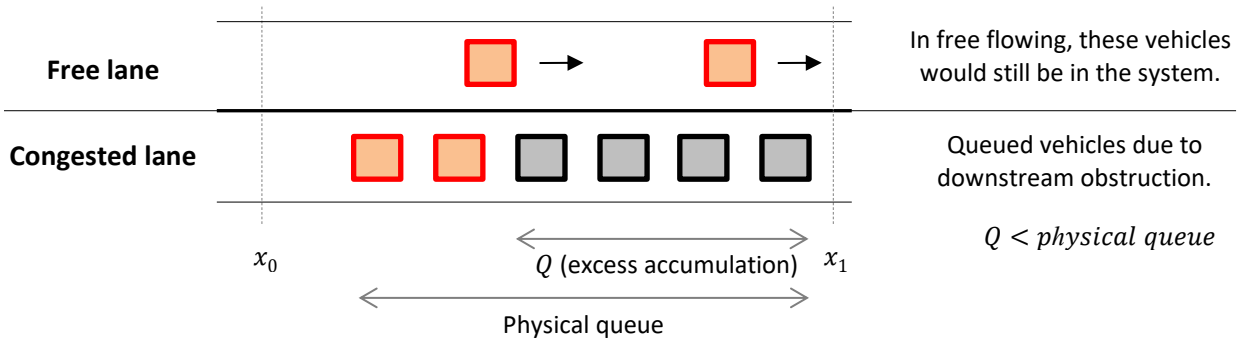


Figure 15. Difference between the excess accumulation and the physical queue.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

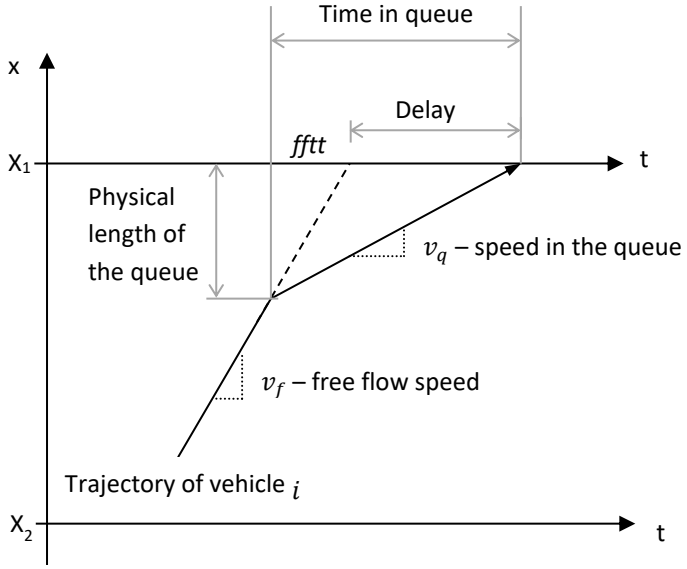


Figure 16. Difference between delay and time in the physical queue.

8. The Little's formula

The Little's formula (or law) is a theorem by John Little (emeritus professor at the Massachusetts Institute of Technology) which states that the long-term average excess accumulation in a stationary system, $\bar{Q}$, is equal to the long-term average delay, $\bar{w}$, multiplied by the long-term average effective arrival rate, $\bar{\lambda}$. This is equivalent to:

$$\bar{\lambda} = \frac{\bar{Q}}{\bar{w}}$$

Figure 17 proves Little's formula for a queuing system in a period, (t_0, t_1), which starts and ends without delays (i.e. $V(t)$ and $D(t)$ define a closed area). In this period, the area enclosed between $V(t)$ and $D(t)$ can be computed equivalently in two ways: *i)* as the average delay, $\bar{w}$, times the total number of affected customers $(n_1 - n_0)$; or *ii)* as the average excess accumulation, $\bar{Q}$, times the duration of the period, $(t_1 - t_0)$. Little formula is directly obtained by recognizing that, in the defined context, the total number of affected customers divided by the duration of the queuing period is equal to the average arrival rate $\bar{\lambda}$. Despite this simple prove assumes that there is no queue at the beginning and end of the period of analysis, Little's formula is equally valid for any system in stationary conditions (i.e. the system is stable and non-preemptive; this rules out transition states such as initial startup or shutdown). Also, Little's equation is equally valid when considering the total time in the sysem and the total accumulation (i.e. including the service time and the customers in service). Although it looks intuitively easy, it is quite a remarkable result, as the relationship is not influenced by the arrival process distribution, the service distribution, the service order, or practically anything else.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

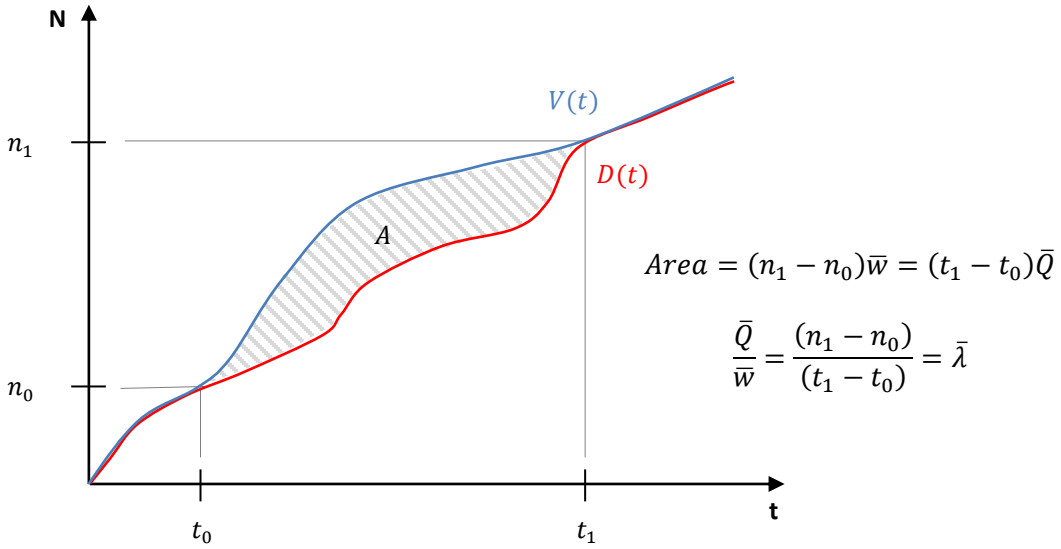


Figure 17. Little's Equation.

The Little formula is quite useful in situations when one of its terms is difficult to measure, while others are available. Imagine an application where there is no easy way to measure the time in the system of some customers. However, the mean number of customers in the system and the throughput are known. Then, according to the Little's formula, the average time in the system can be found as the mean number in system divided by the mean throughput. Or alternatively, imagine a small store with a single counter and an area for browsing, where no one leaves without buying something. Little's formula tells us that the average number of customers in the store is the average arrival rate $\bar{\lambda}$ times the average time that a customer spends in the store.

9. Constructing input-output queuing diagrams from incomplete information

In the typical queuing analysis for the design or improvement of a particular service, the demand is an input. Either the system is in operation and the demand can be measured (e.g. during the rush period), or the demand is predicted by a previous demand analysis for the system. Also, the capacity of the service (i.e. maximum service rate) is known, as the unitary capacity of each server is a fundamental performance parameter which describes the server. Then, the typical question in the queuing analysis is to find the average (or total) delays and excess accumulation. Note that as $V(t)$ and μ are inputs, the only unknown to obtain the desired results is to determine the $D(t)$ curve.

Given $V(t)$ and μ as inputs, $D(t)$ can be obtained following two rules:

- *Rule 1*: Whenever there is queue present, customers pass through at capacity (i.e. $\dot{D}(t) = \mu$)[9]
- *Rule 2*: If queue is not present, customers pass through undisturbed (i.e. $\dot{D}(t) = \min(\dot{V}(t), \mu)$)

[9] The notation $\dot{D}(t)$ refers to the time derivative of $D(t)$, which is graphically represented by the slope of $D(t)$ in the (N, t) diagram.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

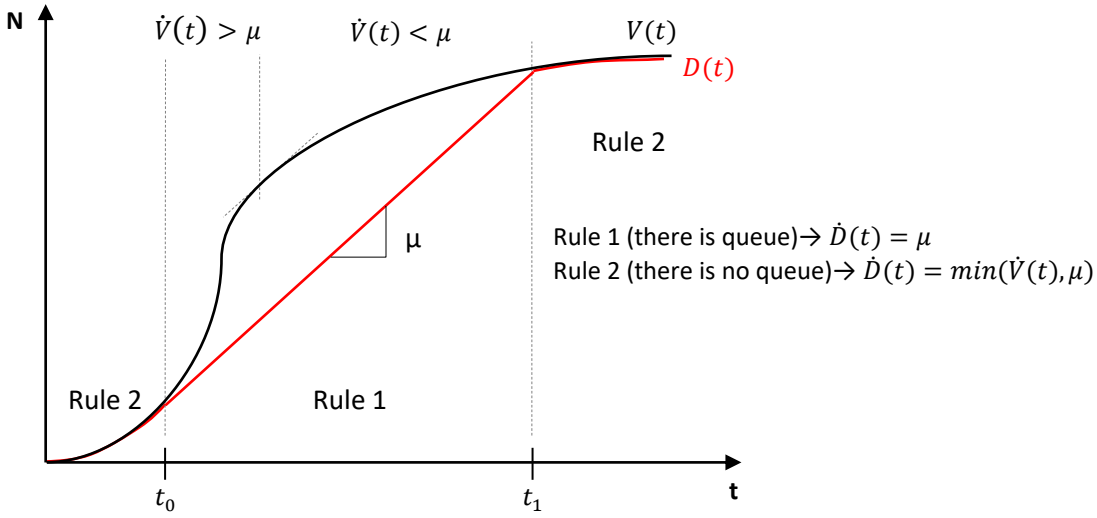


Figure 18. Design of a queuing system.

This means that if initially there is queue in the system, then $\dot{D}(t) = \mu$, as long as the queue prevails (i.e. $V(t) > D(t)$). Otherwise, $V(t) = D(t)$. Graphically in the (N, t) diagram, this process for obtaining $D(t)$ is translated into drawing the "highest" curve possible without going above $V(t)$, and with $\dot{D}(t) \leq \mu$ (see Figure 18).

The application of the previous method for the analysis of a queuing system, must pay attention to ensure that $V(t)$ is representative of the demand which materializes at a particular day. An example will help in clarifying what is meant. Imagine that the objective is to analyze the delays at the security checkpoint when entering a work center. All n workers arrive at the work center sharp at 7:00 am (even days of the month) and at 7:30 am (odd days of the month) (see Figure 19). One could think that the average arrivals curve $V(t)$ to use in the queuing analysis should be the average between that of even and odd days. Note that by constructing this average $V(t)$, what the analyst is doing is to smooth out the peak arrivals (i.e. for each of the day types the peak arrivals were n workers at a particular instant, while in the average $V(t)$ this is $n/2$). Then, by applying the method described to construct $D(t)$, the resulting delays would be largely underestimated with respect to applying the method to each one of the types of day independently. This is just a warning in the averaging of different arrival curves. If the proposed method to construct $D(t)$ is to be applied, averaging of $V(t)$ can only be done for very similar curves, just to increase their significance (as in Figure 20), but not across curves describing different behaviors. In this last case, the actual average $D(t)$ should also be obtained, and therefore the proposed method is not valid.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

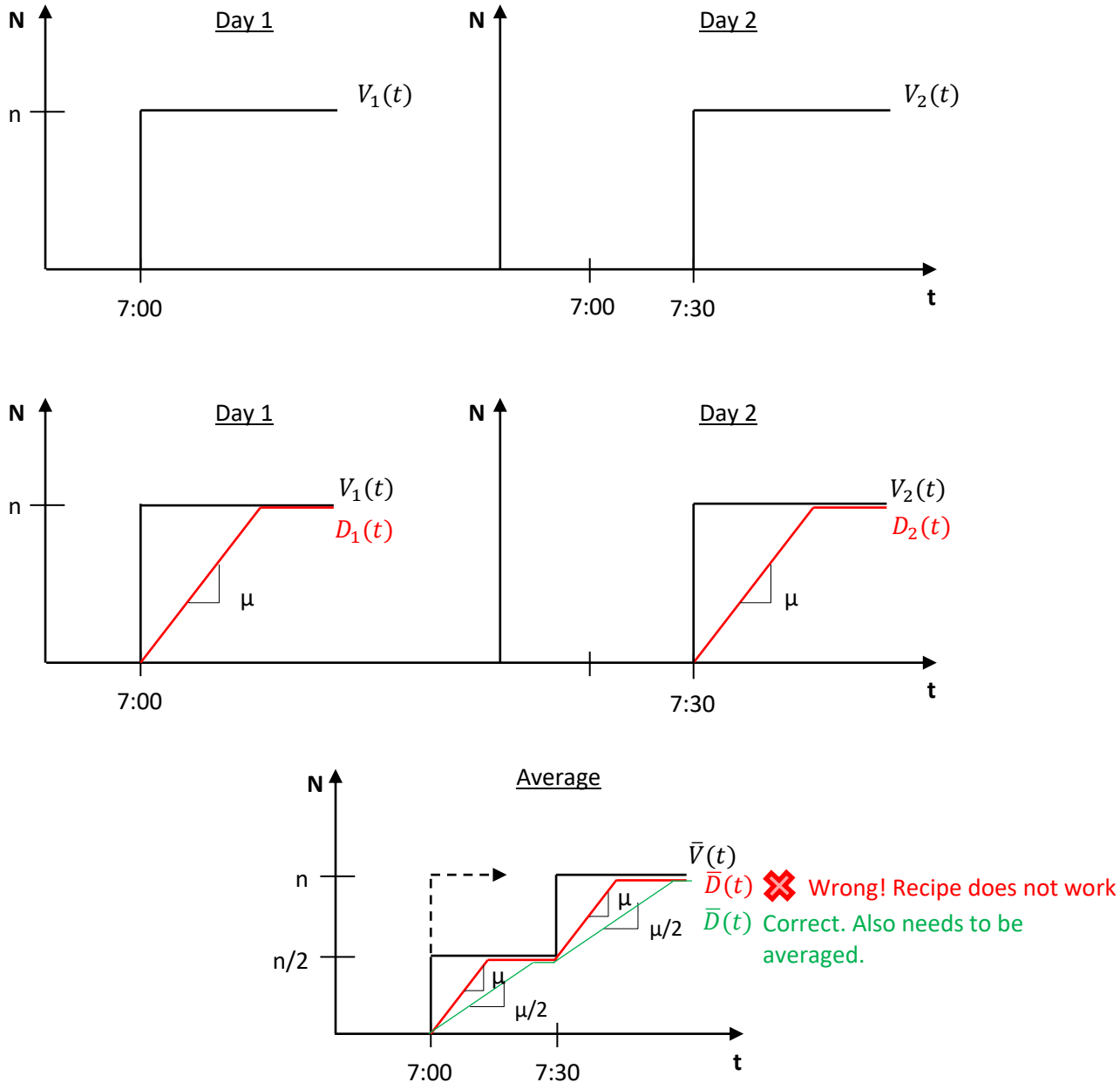
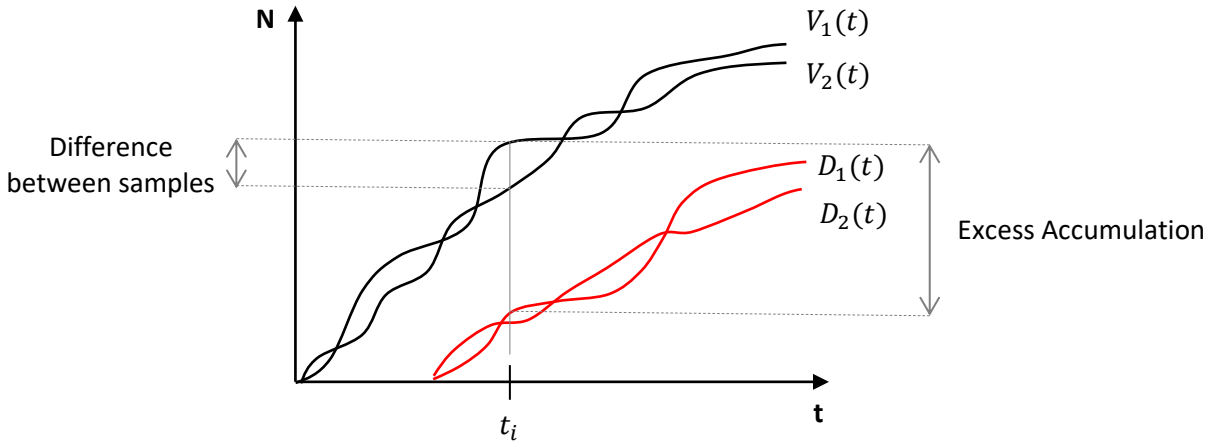


Figure 19. Warning to the proposed method for obtaining $D(t)$.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >



Differences between samples are small in relation to the system excess accumulation

Figure 20. Exception to the warning.

10. On-off queuing systems

On-off queuing systems are a special type of systems where the capacity of the server alternates between its maximum capacity, μ (i.e. the system is "on") and zero (i.e. the system is "off"). A typical example in transportation is the traffic light.

Consider a simplistic traffic light with only two phases: red, with a duration R, when the server is off, and green, with a duration G, when the system is on. The R plus the G durations define the cycle, C. Consider a steady demand rate, λ, and a maximum capacity of the street where the signal is located of μ. Then, the analysis of the queuing system defined by the traffic light is illustrated in Figure 21.

By computing the area enclosed between $V(t)$ and $D(t)$ we can obtain the total aggregated delay, W, as:

$$W = \frac{\frac{1}{2}\lambda\mu R^2}{\mu - \lambda}$$

And the average delay, $\overline{w}$, as:

$$\overline{w} = \frac{W}{n'} = \frac{1}{2}\frac{\mu R^2}{(\mu - \lambda)C}$$

Note the usefulness of this kind of analysis, as the structure of the previous equation directly yields ways of reducing $\overline{w}$ (e.g. reducing R, increasing G, increasing μ, or reducing λ. Also, note that in order to obtain $\overline{w}$, we divide W by the total number of customers served in one cycle, n' (and not n). If W is divided by n, we obtain the average delay of those who stop at the signal, without considering that there are vehicles which do not stop as they find the signal green when they arrive. Obviously, the average delay of those delayed is $R/2$ (see Figure 22).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

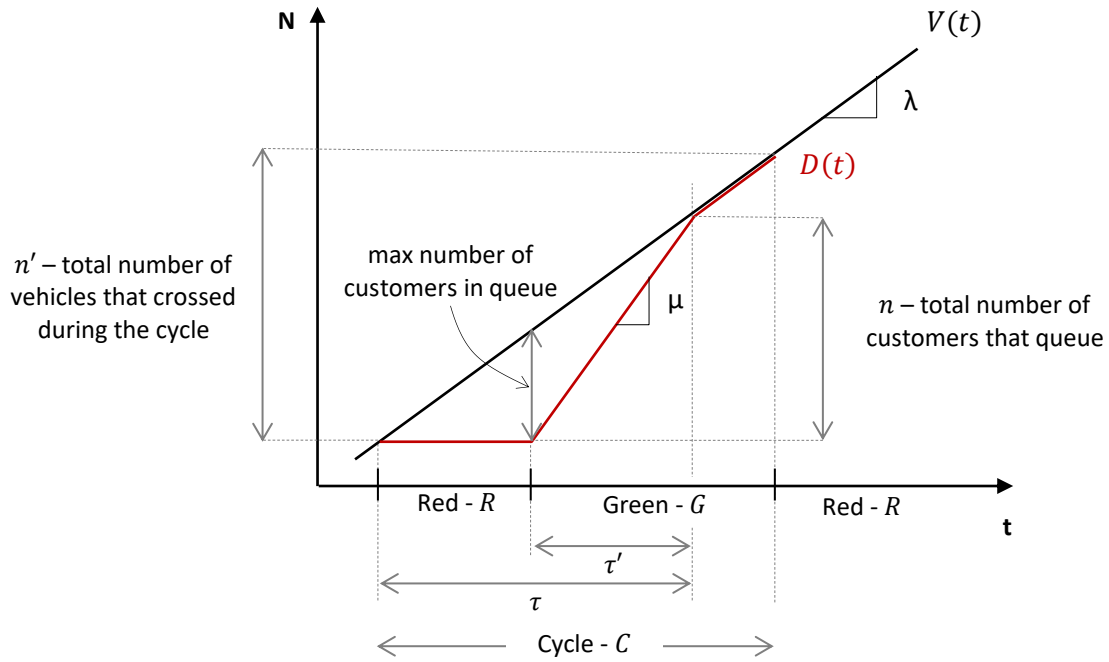


Figure 21. Under-saturated traffic signal.

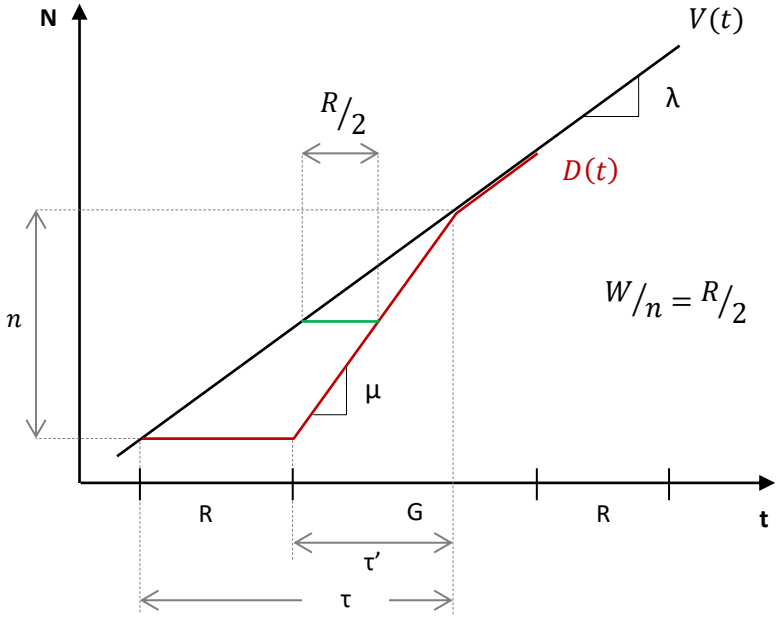


Figure 22. Under-saturated traffic signal: Average delay of those that queue.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

The existence of vehicles which do not stop at the signal (as they find it green), it means that there is some slack in the green time. All the customers arriving in one cycle are served in the same cycle, and still, it remains some green slack time to serve more customers. This is the definition of an undersaturated traffic signal. The limit, (i.e. the exactly saturated signal) happens when the green time is exactly the duration needed to serve the number of vehicles arriving in one cycle. Considering that the number of vehicles arriving in one cycle are λC, and the maximum number of vehicles served during the green are μG, the condition for a traffic signal being undersaturated is:

$$\lambda C \leq \mu G$$

Or equivalently:

$$\frac{\lambda}{\mu} \leq \frac{G}{C}$$

Where the equality holds for the exactly saturated signal.

If $\lambda C > \mu G$, the customers arriving cannot be served in the cycle and some vehicles may need to wait more than one cycle for being served. This means that the queues and delays will grow cycle after cycle, and if the signal configuration does not change, they will not reduce until the demand declines. This signal is said to be oversaturated.

11. Serial or tandem queuing systems

Another type of special queuing systems is that of systems composed of n servers placed in series. This means that the customer needs to go, sequentially, across all servers before completing the service in the system. The particularity of serial (or tandem) queuing systems is that the arrival process to one server is defined by the departure process of the previous server (i.e $D_{i-1}(t) \equiv A_i(t)$) (see Figure 23).

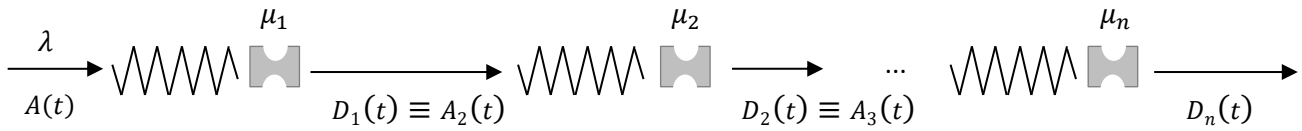


Figure 23. Tandem queues.

Then, one of the questions to answer when designing serial queuing systems is how to place the relative capacities, μ_i, of the different servers in the sequence of the system in order to reduce the total delay (see Figure 24).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

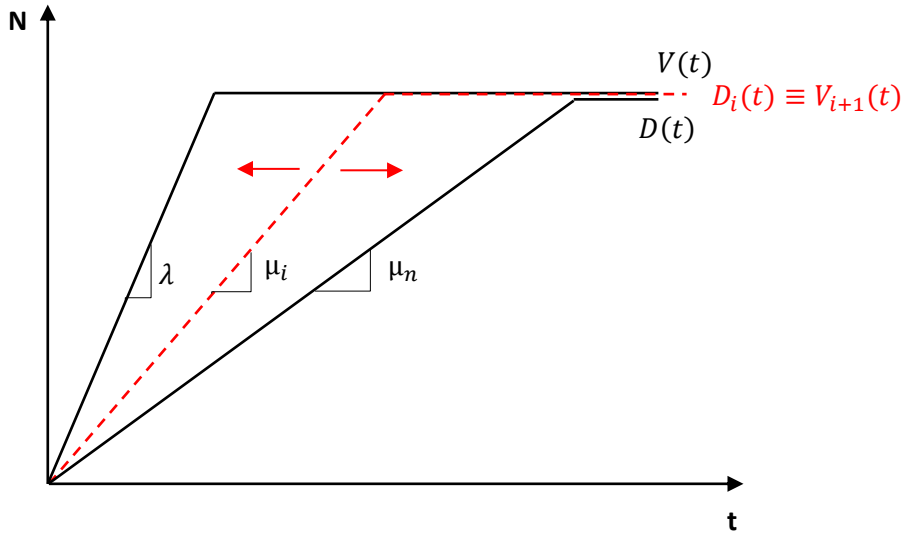


Figure 24. Selecting servers' capacity to reduce total wait.

Imagine the sequence of operations a truck must face in order to discharge a container in the container terminal of some port. Upon arrival, the truck must go through the front gates and clear some paperwork (i.e. Server 1). Then the truck proceeds to clear customs (Server 2). Finally, the truck reaches the landside container yard, a crane discharges its container (Server 3), and the truck is free to go. This sequence of operations clearly defines a tandem queuing system (see Figure 25).

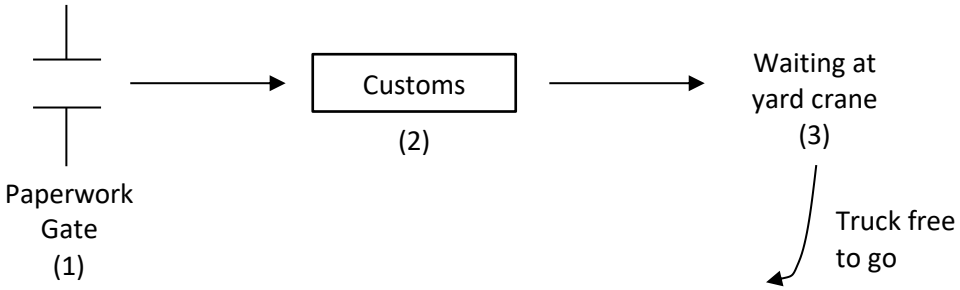


Figure 25. Selecting servers' capacity to reduce total wait.

In this example, the design of the tandem system consists in defining the capacity of the front gate, with respect to that of customs, and with respect to that of the yard cranes. Imagine, that the selected capacities are decreasing as the truck advances in the system (i.e. $\mu_1 > \mu_2 > \mu_3$), as shown in Figure 26.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

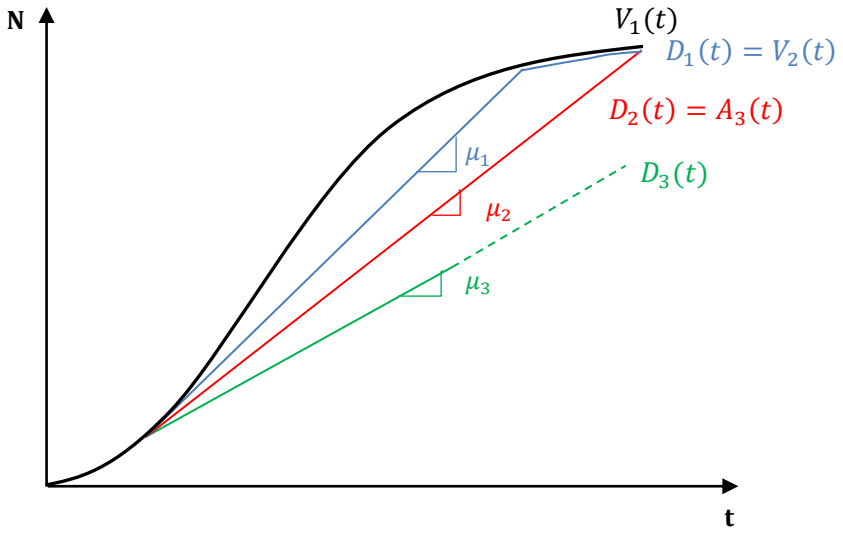


Figure 26. Example of tandem queues – truck delay due to landside operations in a seaport.

Given the previous configuration of decreasing capacities, the truck will wait at the front gates (i.e. time between $V_1(t)$ and $D_1(t)$), at customs (i.e. time between $V_2(t)$ and $D_2(t)$), and at the yard crane (i.e. time between $V_3(t)$ and $D_3(t)$). Then, if an investment in capacity could be made in the system in order to reduce total delay, note that this must be made at the yard cranes (i.e. increase μ_3) so that the slope of $D_3(t)$ grows, reducing the delay at the yard cranes and directly the delay for the whole system. Any other investment strategy (e.g. at the front gates, or at customs) would only yield a delay transfer between the servers, but no benefit for the whole system. Also note that when $\mu_3 = \mu_2$, further investments in μ_3 would yield no additional benefit, as it would be μ_2 which determines the departure curves of the tandem system. From then onwards, it would be necessary to invest equivalently in μ_2 and μ_3 in order to reduce the total delay in the system. This holds until $\mu_3 = \mu_2 = \mu_1$, when investments should be done in all servers. The conclusion for this example is that the optimal design strategy in tandem queuing systems in order to minimize total delay is that all the servers have the same capacity. In practice, however, random fluctuations may affect μ_i. This means that if the design considers the same average $\bar{\mu}_i$ for all servers, at some time could happen that $\mu_i > \mu_{i+1}$ and some stochastic[10] queues would appear inside the tandem system. To avoid large stochastic queues, in practice it is recommended that the average capacities slightly grow through the serial system (i.e. $\bar{\mu}_i < \bar{\mu}_{i+1}$), so that the largest capacity is located at the last server, $\bar{\mu}_n$).

12. Diverging queuing systems

Diverging systems are a variant of tandem systems, where at an intermediate point in the sequence of operations, a fraction of the customers is allowed to leave the system (see Figure 27). Returning to our previous example of the container terminal, imagine that only a fraction p of the trucks needs to go through customs,

[10] The concept of stochastic queues will be further analyzed in the next sections.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

because the remaining $1-p$ fraction corresponds to national freight which follow another sequence of operations and thus diverge after the front gate.

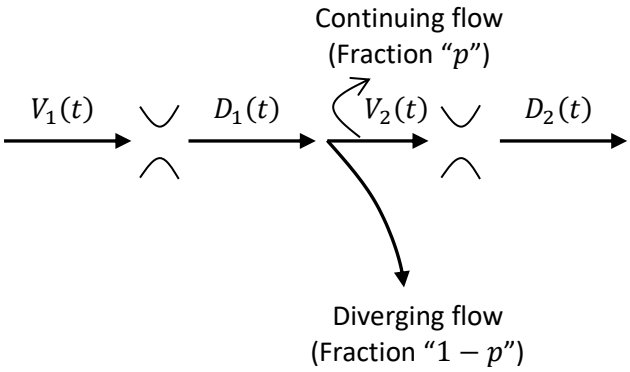


Figure 27. Diverging system.

In this new diverging context, the departures curve from the front gates (i.e. $D_1(t)$) does not coincide with the arrivals curve at customs (i.e. $V_2(t)$). This is because a fraction $1-p$ of the demand diverges. This means that $V_2(t) = pD_1(t)$. This new scenario can be represented in the input – output queuing diagram of the tandem system, as shown in Figure 28. Note that service at the front gates is represented by curves $V_1(t)$ and $D_1(t)$, while service at customs by curves $V_2(t)$ and $D_2(t)$. This means that there is a region on the diagram not related to any service, between curves $D_1(t)$ and $V_2(t)$. Still, the difference in customers accumulation between $D_1(t)$ and $V_2(t)$ represents the cumulative number of customers who have left the system.

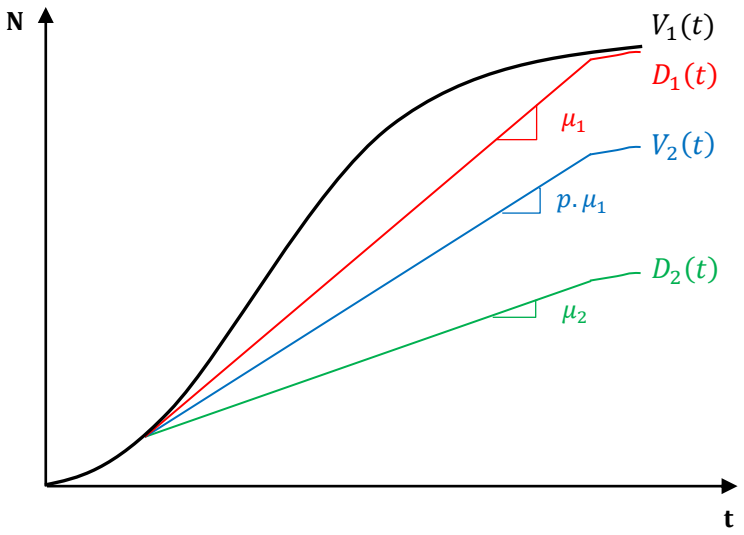


Figure 28. Input – Output diagram of a diverging system.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

The previous construction in Figure 28 can be modified in order to avoid the "empty" region between $D_1(t)$ and $V_2(t)$. This can be achieved by using a secondary vertical axis, N', where the cumulative number of customers is affected by fraction p. Then, by measuring N-curves after the diverging in the N' axis, we have that $D_1(t) = V_2'(t)$. In this construction, however, we must pay attention that all capacities of servers after the diverging (i.e. whose departure curves will be measured in N') need to be divided by p. As an example, the capacity of Server 2, as measured in N' would be μ_2/p, as $N = pN'$.

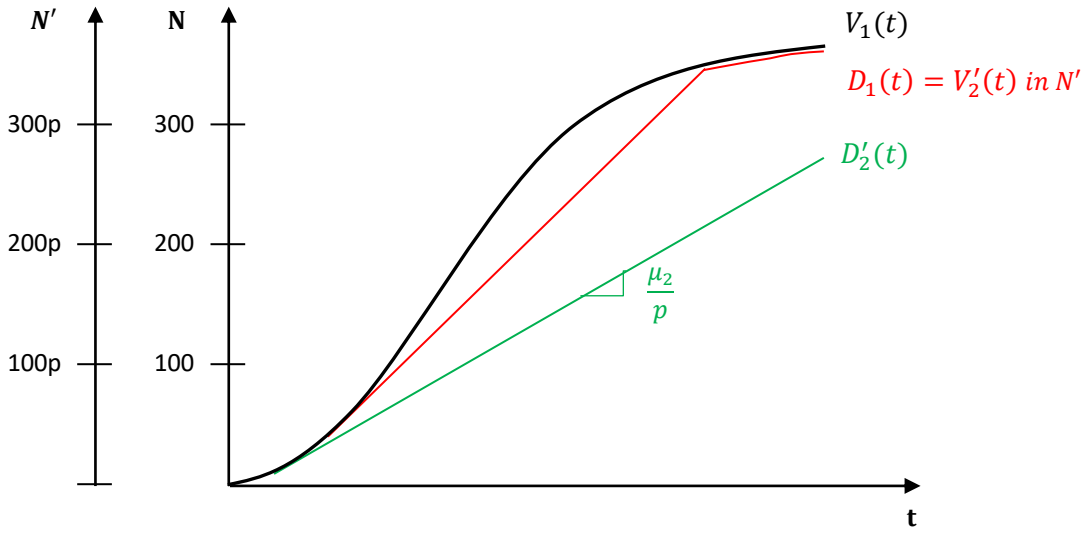


Figure 29. Simplified input – output diagram for diverging systems.

13. A simple strategy to reduce delay

A common strategy to deal with queuing systems is to increase capacity when queue reaches a predefined length. It is typical that managers decide to open an additional server, only when the queue at the working servers is already significant. This strategy is very inefficient, as when the queues are created, they inevitably result in high delay costs.

Imagine a service where a single server is operating with capacity μ. A second identical server is available, but generally it is closed, as the regular demand does not require such capacity level. However, at some instant, demand rate peaks to 2μ for a duration T. Consider that the management strategy in the system is to open the second server only when the queue at the first one reaches a length of μT. This means that in the previous peak demand context, the second server would open at time T, and will take an additional duration $T/2$ to clear the existing queue (see the left hand side of Figure 30). Note the total delay suffered by customers, described by the area between $V(t)$ and $D(t)$. Now imagine a different management strategy. Assume that the second server opens whenever the arrivals rate to the system exceeds μ (i.e. the capacity of one server). Note, that if this early capacity increase strategy is applied, queue will not appear at the start of the peak period. Even if the second server is only open for a duration $T/2$ (in order to compare the results with the previous strategy using the same

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

amount of resources), the total delay generated at the end of the peak period would be much smaller than in the previous (see the right hand side of Figure 30, where the area between $V(t)$ and $D(t)$ is much smaller).

So, the conclusion is that efficient strategies of queuing management imply to act at the early stages of the queue formation, because when the queue has developed it is already too late to significantly reduce the costs in terms of customers' delay.

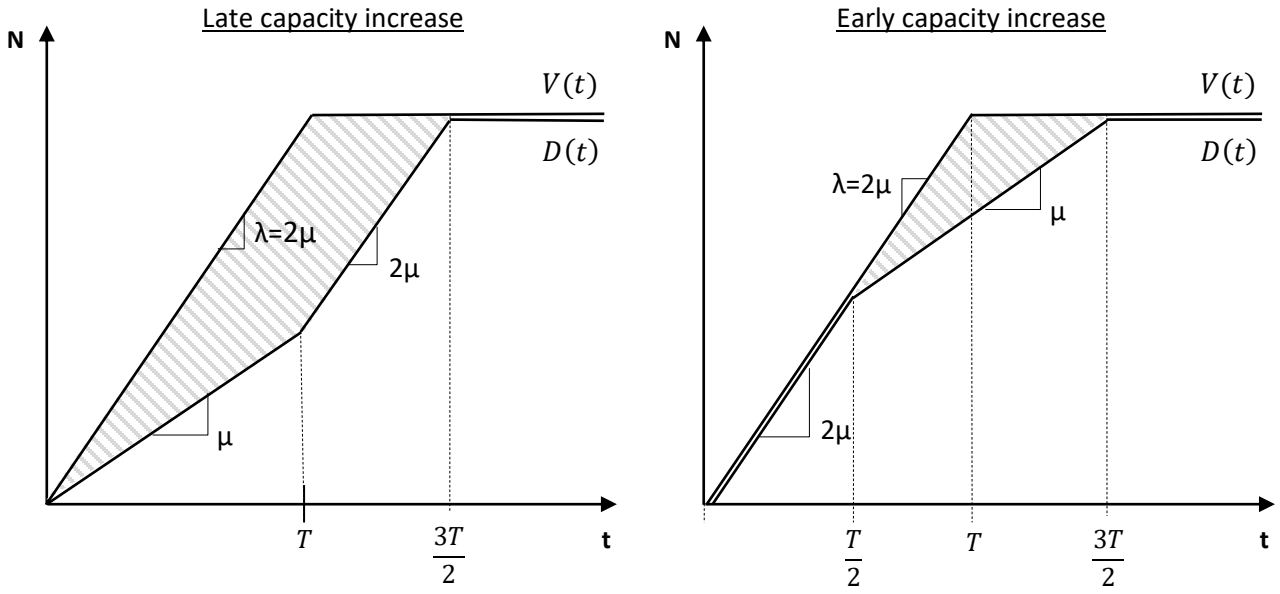


Figure 30. Benefits of acting early in a queuing context.

14. Stochastic effects in queuing systems

So far in our queuing analysis we have treated all the variables as deterministic. In particular, we have considered deterministic values for the arrivals rate, λ, and for the capacity of the service, μ. In this deterministic context, queues only appear when $\lambda > \mu$. However, in reality both the arrival and the service rates are random variables, subject to fluctuations. Let's consider $\bar{\lambda}$ and $\bar{\mu}$ as the average time independent arrivals and service rates over a long observation period T. In such context, if $\bar{\lambda} < \bar{\mu}$, the arrivals, $V(t)$, and the departures, $D(t)$, curves would roughly be two superimposed straight lines (see Figure 31). However, short-lived queues may arise due to fluctuations in the arrivals and service processes, and these would generate (stochastic) delays (see Figure 32).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

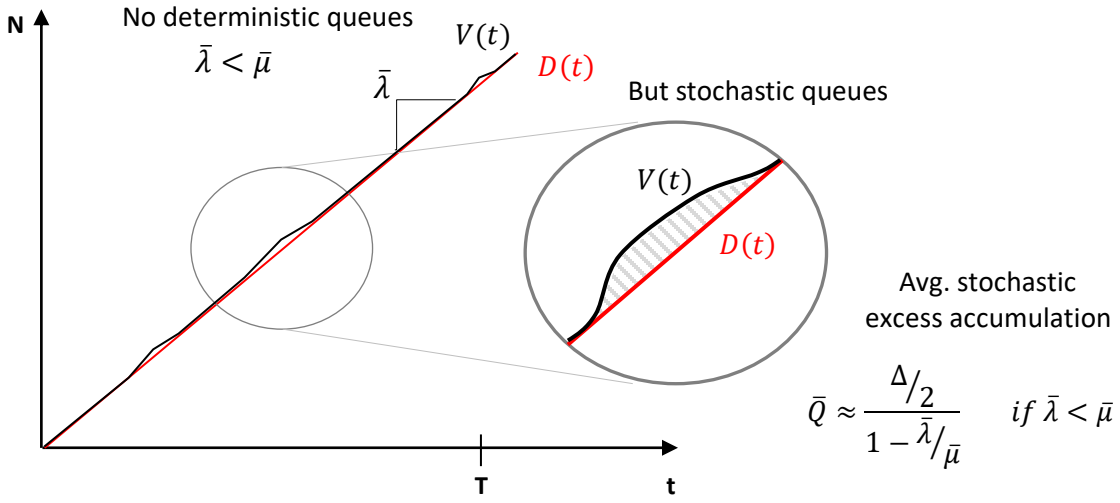


Figure 31. Stochastic effects on queuing.

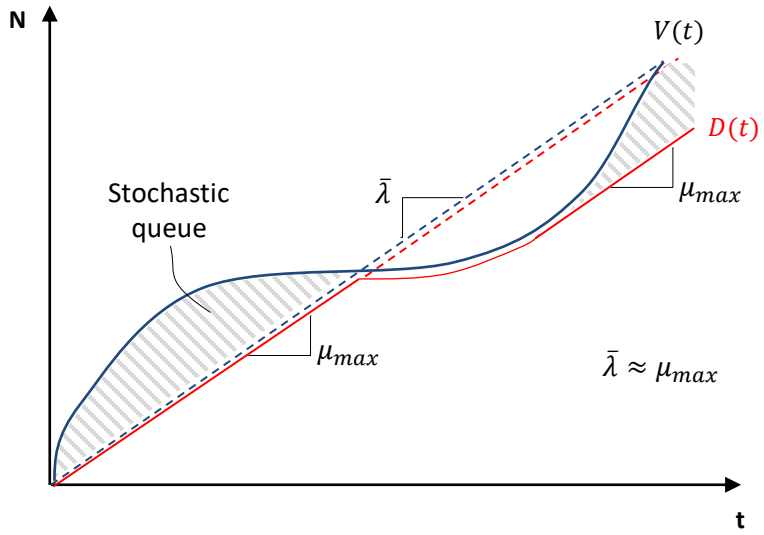


Figure 32. Stochastic queues.

These queues are small during T if compared with the total number of customers that could have been served if the server had been busy all the time. Also, the timing and precise magnitude of these queuing episodes varies from day to day because of their random nature. In spite of this, an approximate expression for the average excess accumulation, $\bar{Q}$, over a period of long duration is given by:

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$\bar{Q} \approx \frac{\Delta/2}{1 - \bar{\lambda}/\bar{\mu}} \quad if\ \bar{\lambda} < \bar{\mu}$$

Where Δ is a constant parameter capturing the variability in the arrival / service processes. $\Delta \approx 1$ whenever customers do not arrive in batches and are not served in groups. Note that stochastic queues will grow as $\bar{\lambda}$ approaches $\bar{\mu}$ (i.e. $\bar{\lambda}/\bar{\mu} \to 1$; $\bar{\lambda}/\bar{\mu}$ is referred as the traffic intensity, or the saturation index). This makes sense as when as $\bar{\lambda}$ approaches $\bar{\mu}$ even a small fluctuation may create queues. In contrast, if $\bar{\lambda} \ll \bar{\mu}$, it would be needed a very large (i.e. and rare) fluctuation to create a queue.

The corresponding average stochastic delay can be obtained from the previous expression by applying Little's formula, yielding:

$$\bar{w} \approx \frac{1/2\ \Delta/\bar{\lambda}}{1 - \bar{\lambda}/\bar{\mu}} \quad if\ \bar{\lambda} < \bar{\mu}$$

The previous expressions will give an approximate estimation of the average magnitude of stochastic excess accumulation and delay assuming that there is no initial queue (or the order of $\bar{Q}$, but not larger), and that the observation period is long enough. The duration of observation T needs to be larger than the relaxation time of the system, T^*. This is:

$$T \gg T^* = \frac{\Delta\mu}{\left(\bar{\mu} - \bar{\lambda}\right)^2} \quad if\ \bar{\lambda} < \bar{\mu}$$

The larger the ratio T/T^* the lesser the variation in $\bar{Q}$ and the more precise the previous expressions are. Note that when the system approaches saturation, the required T increases rapidly. Then, if $T \approx T^*$ or smaller, a more precise description of the random processes involved and stochastic queueing analysis would be required.

Although not being the object of this chapter to enter into the complex stochastic queuing analysis, it is worth presenting the notation typically used in the description and classification of stochastic queuing systems. This is known as the Kendall notation, after the English mathematician D.G. Kendall, and it consists on a series of factors (i.e. $A/S/c/k/N/D$) describing the different properties of the system. Table 1 describes the main components of Kendall's notation. Note that the notation may be simplified to $A/S/c$, where k and N are assumed to be unlimited and the queuing discipline, D, to be First Come / First Served (FCFS). The typical stochastic queuing systems for which exact explicit equations can be derived are $M/M/1$ and $M/D/1$ systems.

As commented, and made evident in the Kendall notation, both the arrivals and the service processes can be random. In real queuing systems, typically the randomness and variance associated with the arrivals process is much larger than that of the service process, which might be often considered as deterministic, or stochastic with a much smaller variance. This implies that most of the stochastic queuing can be associated to the randomness of the arrivals process, as illustrated in Figure 33. Note that for the totally deterministic system (i.e. D/D), there is no stochastic delays, and deterministic delays appear only for $\rho = \bar{\lambda}/\bar{\mu} > 1$.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Table 1. Kendall notation for stochastic queuing systems

Factor	Property	Description
A	Arrival process	M – Markovian or memoryless process. Poisson arrivals (i.e. exponential inter-arrival times) or exponential service times. D – Deterministic. Fixed interarrival times or fixed service times.
S	Service process	E_k - Erlang distribution with k as the shape parameter. G – General distribution, usually referring to uniform independently distributed arrival or service times.
c	Number of servers	The number of service channels.
K	Number of slots in the system	The maximum number of customers allowed in the system. When the number is at this maximum, further arrivals are turned away. If this number is omitted, the capacity is assumed to be unlimited, or infinite.
N	Size of the population	The size of the population from which the customers come. A small population will significantly affect the effective arrival rate, because, as more customers are in system, there are fewer free customers available to arrive. If this number is omitted, the population is assumed to be unlimited, or infinite.
D	Queuing discipline	FCFS/FIFO, LCFS/LIFO, SIRO, Priorities, …

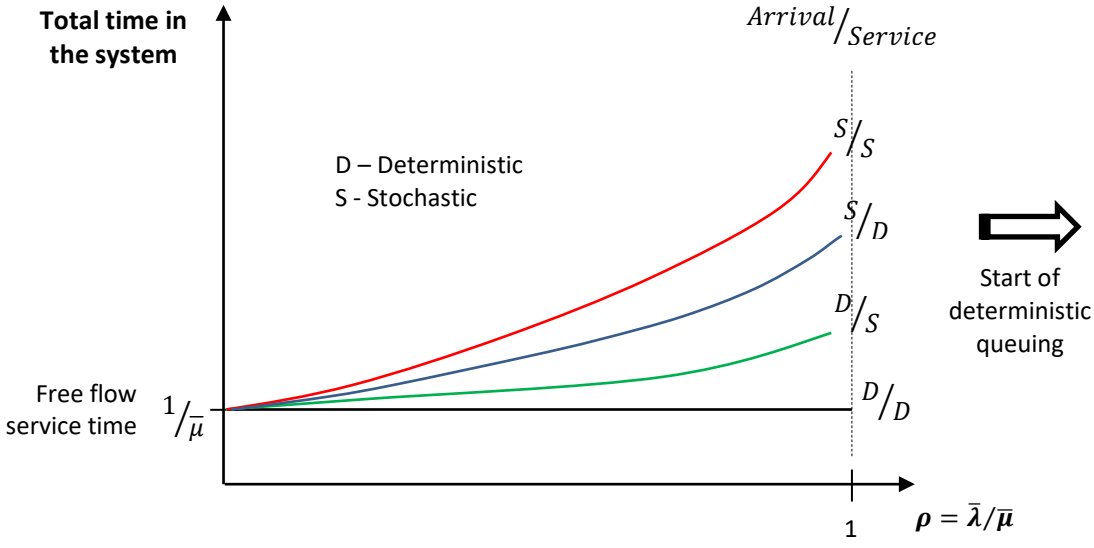


Figure 33. Stochastic arrivals versus stochastic service in queuing systems.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

15. Centralization effects in queuing systems

In the introduction of this chapter, when describing queuing systems with n servers in parallel we already discussed qualitatively the potential benefits of centralizing the queues of the different servers. In this section we are going to prove analytically that centralized systems require less resources to offer the same level of service (or equivalently, with the same resources centralized systems provide better level of service). This higher efficiency of centralized systems resides in the fact that aggregating the demand actually reduces the relative magnitude of the random fluctuations, so that the system is less prone to stochastic queues. The higher the number of servers managed in a centralized fashion, the larger the reduction of stochastic queues (see Figure 34).

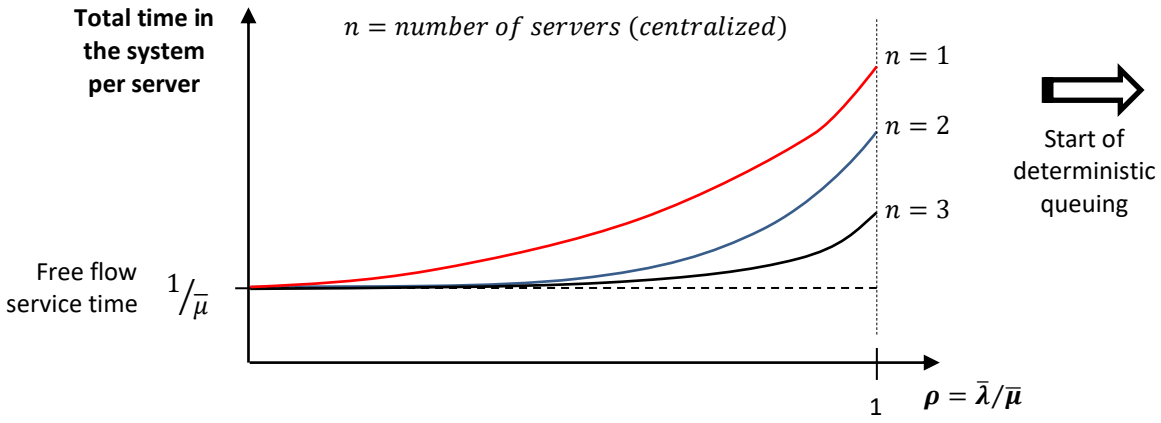


Figure 34. Centralization effects on stochastic queuing.

Imagine a queuing system with n servers in parallel (see Figure 35). The system could be operated with n independent queues (i.e. decentralized), so that each server would receive a demand a. Alternatively, the system could be operated in a centralized fashion, with a single centralized queue whose demand would be A, composed of the sum of all the n demands a. Assume that the system has enough capacity to serve the demand without significant delays, so that only stochastic delays might appear from time to time. Then, the resources needed to serve the system with a given level of service will be proportional to the fluctuations experienced by the demand. We can characterize these fluctuations by the standard deviation of the arrival rates. Then, we have:

$$R_C \propto \sigma_A;\ \ R_D \propto \sum_{i=1}^{n} \sigma_a = n\sigma_a$$

Where σ_A and σ_a are the standard deviations of the aggregated and disaggregated demand rates, respectively. With some mathematical manipulations, and assuming the independence of the n servers in the decentralized case, we can relate σ_A and σ_a as:

$$Var(A) = Var\left(\sum_{i=1}^{n} a\right) = nVar(a)$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Where $Var(\cdot)$ represents the variance, defined as the square of the standard deviation. So, applying the square root at both sides of the equality in the previous equation we obtain:

$$\sigma_A = \sqrt{n}\sigma_a$$

And considering the resources required by the different systems, we have:

$$R_D \propto n\sigma_a = n\frac{\sigma_A}{\sqrt{n}} = \sqrt{n}\sigma_A \propto \sqrt{n}R_C$$

So, the conclusion is that:

$$R_C = \frac{R_D}{\sqrt{n}}$$

Which proves that the resources required to serve the same demand with the same level of service are smaller if the system is operated in a centralized way as $\sqrt{n}$ is always larger than 1 if $n > 1$. Note also that the benefits of centralization grow with the number of servers, n.

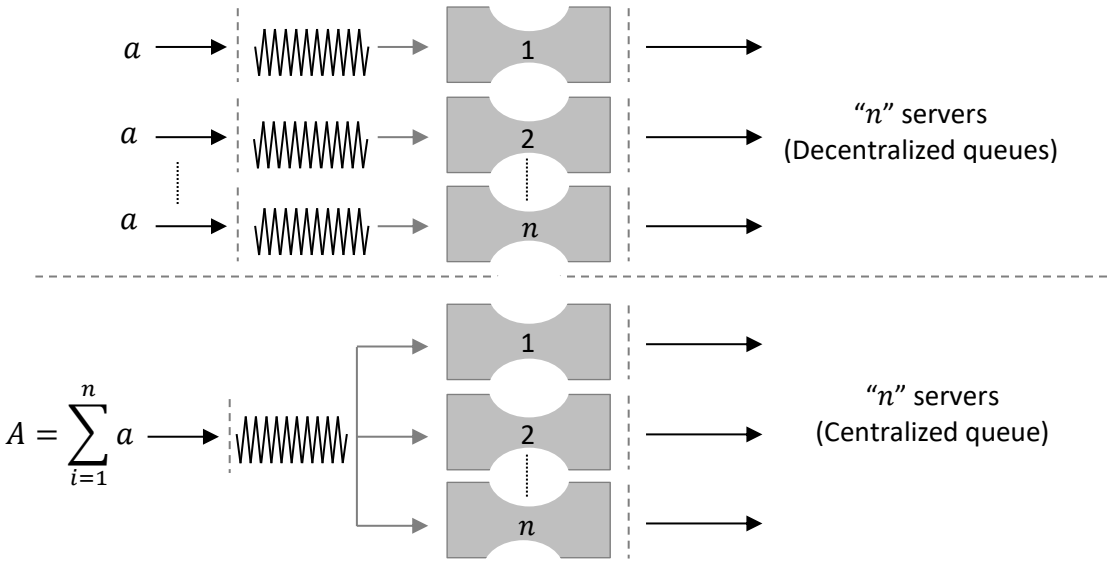


Figure 35. Centralized vs decentralized systems.

16. Optimization in queuing systems

In the planning phase of any service, one of the main decisions is to determine its capacity in order to provide an adequate level of service to users. Let's assume that the demand is known from a previous demand study. Providing higher capacity improves the level of service to users, with reduced delays. However, providing such level of service implies additional costs to the operating agency. There is a clear trade-off between user costs

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

(i.e. delays) and agency costs. Then, the optimal level of service from a social perspective would be the one which minimizes the total costs, composed of both, the user and the agency costs (see Figure 36)

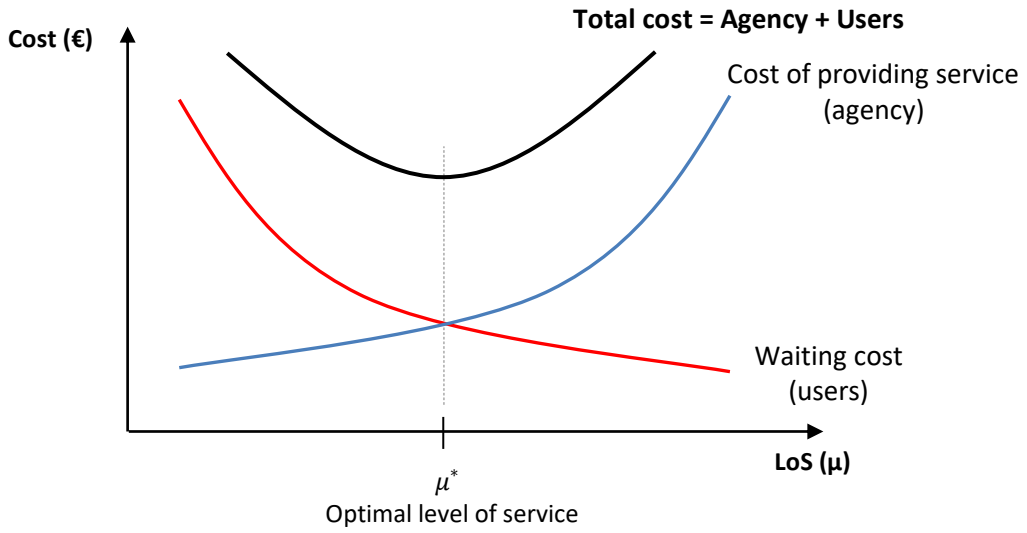


Figure 36. Optimization of the level of service (LoS).

This defines an optimization problem, where the decision variable is μ (i.e. the capacity of the service, as a proxy for the provided level of service) and where the objective function is the total cost. In order to solve the problem, it is needed to formulate the agency and user costs in terms of the decision variable, and apply the first-order condition for optimization (i.e. the first derivative of the total costs with respect to μ is zero). It is not needed to check the second derivative, as total costs are defined by a convex function (i.e. the shape of a smiling face), because user costs decrease with μ, and agency costs behave the opposite.

To clarify this optimization process let's analyze the optimization of the bus headway, h (i.e. the time between consecutive bus expeditions) on a transit line. In this context, h, acts as the decision variable, and it is needed to formulate: *i)* the operating costs of the agency; and *ii)* the user costs resulting from the waits for the bus, both as a function of h.

- *Agency operating cost,* Z_A: Consider β as the cost of operating one bus during one day [€/day]. This cost includes everything (i.e. the prorated cost of the bus acquisition, the fuel, the wage of the driver, administrative costs, etc.). Then, the daily agency cost would be obtained as β times the number of required buses to operate the line. Considering that the bus requires a time C to cover the whole route (i.e. C is the cycle time), and that there is one bus dispatch every h, then the required number of buses is C/h. Note that this is an application of the Little's equation. Then the daily agency operating cost is formulated as:

$$Z_A = \beta \frac{C}{h} \quad [€/day]$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

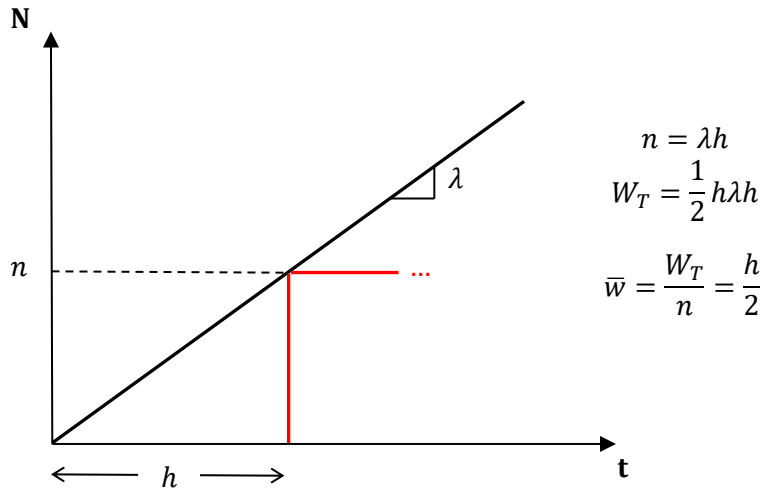
Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

- *User costs,* Z_U: This is the cost of the passengers wait time. Considering the daily demand for the bus service of λ [pax/day], an average value of users' wait time of α [€/pax·day], and the average waiting time of $\overline{w}$, then:

$$Z_U = \lambda \alpha \overline{w} \quad [€/day]$$

Assuming short headways, people arriving uniformly during all day without reference to the schedule, and that buses are perfectly on time, the average waiting time for each customer is $\overline{w} = h/2$ (see Figure 37). Substituting this into the previous expression, we obtain:

$$Z_U = \frac{\lambda \alpha h}{2} \quad [€/day]$$



The expected wait time of those passengers arriving during h is $\frac{h}{2}$

Figure 37. Expected wait time for a given bus headway.

The total cost function is obtained as the sum of the agency and users cost, and is formulated in terms of the decision variable h, as:

$$Z = Z_A + Z_U = \beta \frac{C}{h} + \frac{\lambda \alpha h}{2}$$

Note the convexity of Z, as the first term is proportional to h^{-1} (which is convex) and the second term is linear with h (also convex; limit of convexity). The sum of two convex functions yields another convex function.

Applying the first-order condition of optimization involves:

$$\frac{\partial Z}{dh} = 0$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$\frac{1}{2}\lambda\alpha - \beta C h^{-2} = 0$$

And solving for h, we obtain the optimal h^*:

$$h^* = \sqrt{\frac{2\beta C}{\lambda\alpha}}$$

Again, even if it could be difficult to obtain accurate values for the different parameters involved, the functional structure of the solution allows evaluating decisions. For example, if the fuel price increases so that β grows by a factor of 2, we know that h^* needs to grow by $\sqrt{2}$. The same effect on h^* would have the implementation of a dedicated bus lane which allows reducing the cycle time, C, by a factor of 2.

17. The psychology of waiting lines[11]

In queuing systems where customers are humans, the psychology plays an important role in the customers' satisfaction with the service. This means that all service managers must pay attention to three things: *i)* what was actually done for the customer to improve the service (all the focus in the chapter so far has been to this point); *ii)* what is perceived by the customer; and *iii)* what the customer expected.

Recall the first law in psychology stating that:

$$Satisfaction = Perception - expectation$$

This means that a better perception of the service or lower expectations, both contribute to the satisfaction of the customer. Perception and expectation will have some connection to reality, but they are not the reality and there exist some strategies for managing better these feelings. These are going to be summarized next.

- *Wait after the service feels longer than wait before the service which feels longer than waits in the service*.

  Customers want to get started. The early stages in the service are the most important, as it is very difficult to change a bad initial perception. This happens because the anxiety level is much higher while waiting to be served than it is while being served. There is a fear of being forgotten. How many times you have gone back to a maître in a restaurant to check that your name is still on the waiting list to be seated? So, whenever possible invest most of the attention to the early stages, and convey the sense that the service has started: we know that you are here. Examples of good practice which make people feel reduced waiting times are the restaurants handing out menus, providing drinks in the bar, or an early pass by the table, or the medical triage systems before seeing the doctor.

  But if there is a worst wait, this is the one after the service. After service, customers expect no wait. Also, there is not more added value to be received. This means that after service waits are intolerable. Examples of this extreme long perception of waits after service are check-out at hotels, paying the bill at a restaurant, or waiting for disembark an airplane when it reaches the gate. Note the different perception of wait times in the airplane traveler. The same passenger who sat patiently for hours during the flight, suddenly exhibits an

[11] The contents of this section are obtained from the paper by David H. Maister (2005), *The Psychology of Waiting Lines*, www.davidmaister.com.

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

intolerance for an extra minute or two for disembark, and a fury at an extra few minutes for collecting the baggage.

- *Occupied time feels shorter than unoccupied time.*

  Boredom results from being attentive to the passage of time itself. Note that a watched pot never boils.

  The Barcelona Terminal 2 airport (i.e. the former Barcelona airport before the inauguration of the new Terminal 1) has 3 buildings (i.e. A, B, and C). Arriving passengers could arrive to any of the terminals, depending on the assignment of the airline used. However, all of them need to walk to the basement of Building B to collect the baggage, as the carrousels were located only there. Customer satisfaction surveys at the former airport shown a lack of satisfaction with the "extensive" delays at the baggage pick-up service of those customers arriving at Building B. This claims where not observed for those customers arriving at Building A, not for those arriving at Building C, but only for those arriving at Building B. You may thing that workers moving the baggage from planes arriving at B to the carrousel were less efficient than their colleagues at A or C. Actually, this was empirically measured and was discarded as a possible reason. So, what was happening? The answer was quite simple. Customers arriving at A and C had a ten-minute walk to reach the carrousels, so that in many occasions when they arrive, the baggage was already there, ready for being picked up. In contrast, customers arriving at B, spent these ten minutes just waiting in front of the carrousel and complaining about the service. The solution was imaginative. Managers created a walking detour to the carrousels for customers arriving at B, so that they could also enjoy this ten-minute walk (i.e. for baggage pick-up, follow the yellow line)[12].

  The conclusion from the Barcelona story is that waiting time should be occupied. Handing out menus when waiting for a table in a restaurant, filling up questionnaires and forms at airports / borders, proposing distraction games, explaining some stories in telephone waits, or even the typical reading material (e.g. magazine, newspapers…) at doctors' offices are some examples on how to fill the time. It spite of this, managers should be imaginative on how to occupy waiting times, with activities which provide a benefit in itself, are related to the service the customer is waiting for, and reduce the service time if possible.

- *Anxiety makes waits to seem longer. Unexplained waits feel longer.*

  Waits in anxiety feel longer. When losing a plane, even the most insignificant delay feels like ages. In the advent of low cost airlines, they promoted open-seating flights, and the result was that there appeared agitated queues for boarding. Or even when choosing the (always wrong) queue in the supermarket, and you feel very long waits while the other line moves faster. All of these waits feel longer because waiting in stressed conditions.

  In addition, uncertain waits feel longer that known, finite waits. Waits in ignorance creates a feeling of powerlessness, irritation and rudeness.

  The solution to the previous effects is to provide information to the customer. Information calms (e.g. do not worry, connecting flights are being held; you are indeed in the correct line, and you have sufficient time). If one understands the causes of the delay, the wait is more patient. Examples include informing of bad weather at airports or emergencies at hospitals. The explanation given may or may not exculpate the service

[12] I do not remember who explained me this story, or where I read it. I cannot even assure that this actually happened, however, *se non è vero, è ben trovato*.

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

provider, but it is better than no explanation at all. The exception is when the cause of the wait is unjustifiable and not acceptable from the customer point of view.

The manager must be careful when providing information, as this also affects the customer expectations. When an announcement of 30 minutes delay is made, customer reacts with initial annoyance, which turns into acceptance of inevitable. It is much worse to announce a "short delay" several times, which at the end turns to be 30 minutes. This creates a whole wait in state of nervous, and the customer feels that she is not dealt with honesty. The policy should be to inform of a time slightly in excess, so that the bad expectation will turn into satisfaction when the delay is shorter than expected. Examples of application of this policy include the information of expected wait times in attractions at amusement parks, or the expected flight time of commercial flights.

- *Appointment systems are a troublesome queue management tool.*

  Appointment systems are used to laminate the demand and avoid extensive waits at peak periods. One of the problems of the tool is that having an appointment creates an expectation of no wait, and this is a very promising expectation. A failure to deliver on this premise can happen easily, and then, even the most insignificant delay results in unsatisfaction. This effect is known as the appointment syndrome, where early arrivers sit calmly until the scheduled time, even for long waits, but once the appointment time is passed, even a short wait is incredibly annoying.

  Appointment systems also include other managing difficulties, like the customers' no-show, which is addressed by overbooking, yielding a probability of no-service to some customers and huge unsatisfaction for those affected. Also, it is not easy to schedule customers when their service time is highly random. Scheduling too far apart implies idle server times, and loosing capacity. Scheduling too close together might imply running behind of schedule with cumulative effects (i.e. further and further behind).

  The conclusion is that appointment systems, in some services may have more problems than benefits.

- *Unfair waits feel longer than equitable waits.*

  If customers can "cut in front", violating the queuing discipline, customers become furious. This also affects the typical situation when the customer is waiting in front of a receptionist (for example) answering a telephone call. How it can be that the telephone call can cut in front? Do distant customers have priority with respect to others who made the effort of coming? Also, if the queue does not have any visible order, this makes customers wait in anxiety in order to preserve the priority. All these situations make waits to feel longer.

  So, whatever rule applies, ensure that this matches with the customers' sense of equity. The recommendation is to only break the FCFS discipline when other priorities are clearly accepted by customers (e.g. emergencies, shorter service time). Systems of taking a number, and showing the number currently being served, are helpful in ensuring fair waits, and also provide information about the expected wait time.

- *Customers expect more wait for more valuable services.*

  The more valuable the service, the longer the customer will wait. Customers expect no wait for simple transactions, where even a small wait is intolerable. In contrast, complex activities accept more wait without complain. Think of checking out a supermarket with a full cart versus with a single product. Or obtain cash in

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

a bank versus asking for a mortgage, or checking-in for a flight versus changing the itinerary, buying fast-food versus having an haute-cuisine meal, or waiting in class for an assistant professor versus a full professor.

- *Solo waits feel longer than group waits.*

  There is some comfort in group waiting. Customers wonder collectively what is happening, comment on the expectation, and console each other. Waiting is part of the fun, and in some cases part of the service (e.g. waiting at amusement parks, buying concert tickets, …). Note that before the delay happened, probably they do not talk each other.

  Then, managers should promote group waits, as this increases the tolerance for the waiting time. Solo isolated waits should be avoided. One service where solo waits are typical is at veterinary offices. Because of the possible incompatibility of different pets, customers wait alone (actually, with their pet) in a room waiting for the vet to come in. Is the vet who moves among the different rooms to visit the animals. So, in a vet clinic, there are several rooms closed with one person waiting with his pet inside. After a few minutes of waiting alone, it is typical that the customer feels that he could be there forever without nobody caring. The customer moves to the door to look what is going out outside, just to realize that all the other doors are opening exactly with the same objective.

- *Others…*

  In this section we have tried to provide a summary of the most typical factors affecting the phycology of waiting. This does not exclude that other may exist. Also, cultural and class differences affect the tolerance for waiting. Recall that it is said of the English that if they see a line, they will join.

18. Summary of strategies to improve queuing systems

In this last section we provide a summary of what has been seen in this chapter, particularly summarizing the strategies that can be applied to improve service, adapt arrivals or improve the physiology and psychology of people waiting in queues. The summary is provided in Table 2. The main conclusion from this chapter could be that reductions in delays do not necessarily imply large investments and they do imply huge benefits for customers. It is the adequate analysis which ensures the efficiency and functionality of services and processes.

Table 2. Summary of strategies to improve queuing systems

Strategies to improve the service process	
Centralization	• Centralized queues. • Servers work dependently as a team. • Flexible assignment of custoAmers to servers.
Automation	• Improve maximum service rate, μ.
Avoid physical blockages	• In systems with servers in parallel which actually work in series. • In tandem queues where downstream queues block upstream servers.

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Reduce customers' participation while in service	• Reduce average customer service time. • Transfer customer participation to the queue, to occupy waiting time.
Serve customers in groups	• While one customer participates or decides, serve simultaneously another customer. • Mix FCFS queuing strategy with SSTF. • Serve in groups similar customers.
Increase the number of servers	• Increase the number of servers once arrivals exceed capacity. The earlier the better. • Additional servers should be flexible: part-time servers, reversible servers.
Strategies to improve the arrivals process	
Appointments	• Laminates peak hour demands, but brings additional problems. • Creates no wait expectation. • Difficult to schedule appointments if service time is highly random. • Customer no-show and overbooking.
Pricing	• Implement pricing strategies. Higher prices at peak hours, and lower prices at off-peak periods.
Favor the initial abandonment	• Provide dynamic information of wait times. • Provide wait times slightly in excess, to favor initial abandonment, and create worse expectations than real, which will turn into satisfaction.
Improve physiology during the wait	
Queues must be comfortable	• Drinking water, illumination, adequate temperature, enough space, etc.
Improve psychology during the wait	
Start the service early	• Transmit the feeling to the customer that service has already started. • Initial phases of the service are the most important.
Occupy the waiting time	• Avoid customers being just waiting.
Provide information	• Information calms. • Be careful. Information provides expectations. Avoid creating expectations that will not be fulfilled.
Justify the wait	• Provide explanations about the causes of the delays.
Pay attention to equity	• Waits must be fair to all customers. • Ensure that the queuing discipline is fulfilled.
Favor waits in group	• Waiting is part of the fun. • Avoid solo waits.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Avoid waits after service	• When the service is over, the expectation is of zero wait. • Customers' do not receive any additional value after service.
Valuable services admit more wait.	• Customers expect zero wait for very simple processes. • Only if the service provided is complex, they admit more wait.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

4-FUNDAMENTALS OF TRAFFIC FLOW MODELING

Contents

1. Introduction to traffic flow modeling 2
2. Variables defining a traffic flow: Macroscopic & Microscopic perspective 3
3. The fundamental equation of traffic ($q = kv$) 9
4. The vehicles' conservation equation 11
   5.1. Relative flow seen by a moving observer 13
5. Traffic diagrams 15
   6.1. Greenshields, Greenberg and Edie's $k - v$ models 15
   6.2. Traffic engineering manuals: a warning 21
6. Macroscopic modeling of traffic flow: LWR - Kinematic Wave Theory 22
   7.1. Speed of a traffic wave 25
   7.2. The fundamental diagram of traffic 27
   7.3. Shocks and waves 30
   7.4. Simplifications 34
7. Example of application of the LWR theory: Incident on a freeway 37
8. Limitations of the LWR macroscopic traffic flow theory 42

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

1. Introduction to traffic flow modeling

Traffic flow consists in the movement of a large number of vehicles along a particular infrastructure. Although the individual movement of these vehicles may seem random, the overall behavior of the traffic flow is highly predictable. Take as an example the morning traffic rush accessing big cities, like Barcelona. In normal conditions (i.e. similar demand, no incidents) congestion appears always at same locations, at similar times and evolves in a similar fashion and with similar queue lengths. In this context, it seems reasonable to think that there must exist some physical behavioral laws that steer traffic evolution. The knowledge of these laws would allow to develop traffic flow models, able to predict traffic evolution, congestion and queue lengths. Such models would be very useful in order to assess planning, management and control of road traffic facilities.

This idea of analyzing traffic flow as a science and in quantitative terms appeared in the 1930's with the pioneer works of the first traffic scientist ever, Mr. Bruce Greenshields (https://www.ite.org/about-ite/history/honorary-members/bruce-d-greenshields/). In spite of the early works of Greenshields, it was not until the 1950's that traffic science was popularized, and that the first dynamic models of traffic flow appeared. From the very beginning, two different approaches to flow modelling could be differentiated.

On the one hand, traffic could be seen as a flow, neglecting the fact that it is composed of individual "particles" (i.e. the vehicles). This fluid like approach took researchers to apply hydrodynamic models to traffic, setting the foundations of macroscopic traffic flow modeling. Two physicists in the UK, M.J. Lighthill and G.B. Whitham developed in 1955 the first macroscopic traffic flow model, by comparing "traffic flow on long crowded roads" with "flood movements in long rivers". A year later in the USA, P.I. Richards (1956), physicist and applied mathematician, developed independently the same model, introducing the concept of "shock-waves on the highway" completing the so-called LWR model (i.e. Lighthill-Whitham-Richards model). Since then, this model has received multiple names, like the Continuous Traffic Flow Model, Shock-Wave Theory or Kinematic-Wave Theory. They all refer to the same original and fundamental macroscopic model.

On the other hand, traffic flow could be seen as the aggregated movement of multiple vehicles, so that if individual trajectories are predicted, the overall traffic behavior can be obtained by integration. This trajectory-based approach constitutes the microscopic traffic flow modelling approach. In its most simple form, characterized by the unidimensional movement of vehicles in a single lane, models are referred to as car-following models. The concept is that in dense traffic conditions each vehicle follows the trajectory of the vehicle in front, and reacts to its speed changes. By modeling this car-following behavior of vehicles, for all the vehicles in a traffic stream, we could obtain the overall traffic evolution. Seminal car-following models were developed by L. A. Pipes (1953) based on the California driving code and T.W. Forbes (1958) introducing the concept of the "reaction time", both proposing a linear relationship between the distance between two consecutive vehicles (i.e. the vehicular spacing) with respect to the traveling speed. Building on these initial works, and others developed over the years, a series of models developed by R.E. Chandler, D.C. Gazis, R. Herman, E.W. Montroll, R.B. Potts and R.W. Rothery at the General Motors (GM) Research Laboratories (Detroit, USA) in the period 1958-61 have received maximum attention over the years. These car-following models were the first to incorporate the stimulus-response structure, where the stimulus is a differential speed between the leader and the follower vehicles and the response of the follower is an acceleration or braking maneuver. It has taken more than half a century to evolve from these first analytical car-following models at the GM research labs to the traffic microsimulators we have today (like this one: https://www.youtube.com/watch?v=k_KjM3l295M or this other: https://www.youtube.com/watch?v=OtYby7QnyAE). Obviously, in addition to a car-following model, we need many other components (e.g. lane-changing model, controllers behavior, O/D matrixes, ...) plus an advanced

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

digital graphics animation. In spite of being very fancy, calibrating these microsimulators requires many parameters (e.g. of the order of 30 maybe), some of them without any physical interpretation. This is 10 times more than the 3 parameters required to calibrate the fundamental diagram in the LWR theory (as you will see next in this lecture). This means that microscopic models are much less robust with respect to LWR, and that all results should be validated with the later.

After this brief introduction to traffic flow modeling and its history, this lecture aims to just present the most fundamental concepts of traffic flow modeling. This includes the definition of traffic variables and the fundamental equation, and the postulates of LWR macroscopic model (i.e. the conservation equation and traffic diagrams). We will end the lecture with an example application of this theory.

2. Variables defining a traffic flow: Macroscopic & Microscopic perspective

Recall that a trajectory is the representation of the movement of one vehicle in space and time (i.e. x, t). However, when dealing with traffic flow modeling, we are more interested in the variables defining the movement of large amounts of vehicles. Figure 1 and Figure 2 show the trajectories of several consecutive vehicles in a single lane, from where we can derive the fundamental variables of traffic flow.

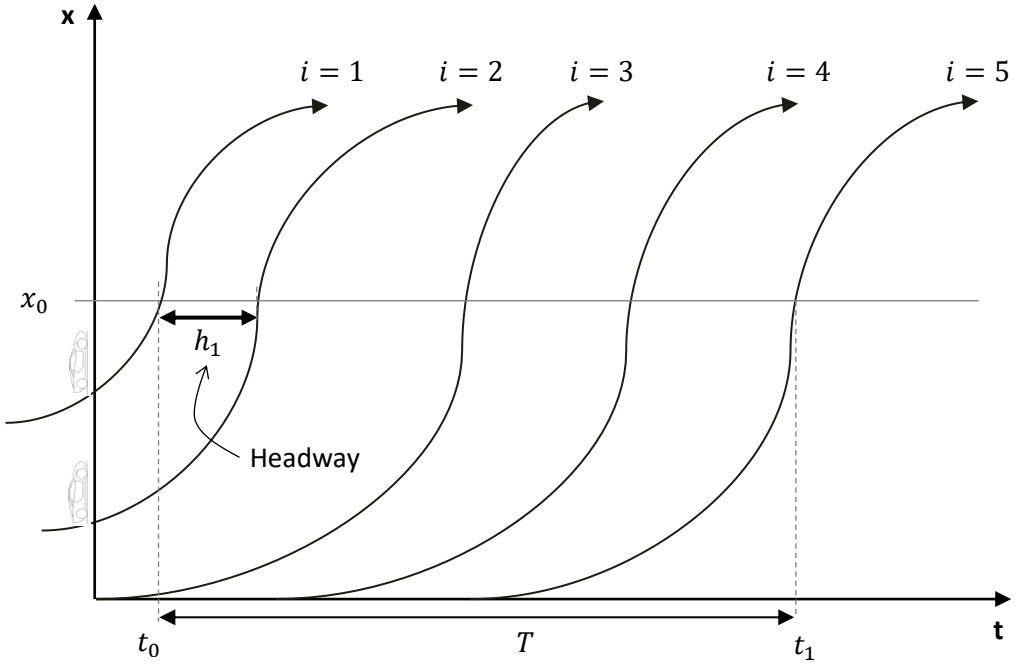


Figure 1. Headway definition on a time – space diagram.

The time interval between the passage of two consecutive vehicles at a given location x_0 is defined as the headway, h, in units of time (e.g. [s]). In Spanish the headway is named "*intérvalo*". Note that the headway is

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

measured taking the same reference point of consecutive vehicles (e.g. the rear bumper in Figure 1), so that, in addition to the empty time gap between vehicles, it includes the time necessary for the passage of the length of the vehicle. Also note that in order to measure the headway, we need to measure at a fixed location continuously in time. This kind of measurement is called a temporal measurement. Consider that we take this temporal measurement during a long period of time, T, and that we measure the passage of m vehicles (e.g. in Figure 1, $m = 5$). Then we can define the traffic flow, q, as:

$$q = \frac{m}{T} \quad [veh/time]$$

Note that the headway is a traffic variable affecting individual vehicles (i.e. a microscopic traffic variable) while the flow is an aggregated or average variable (i.e. a macroscopic traffic variable). Both variables (i.e. h and q) represent different perspectives (i.e. micro or macro) of the same concept, and they can be related as:

$$q = \frac{m}{T} = \frac{m}{\sum_{i=1}^{m} h_i} = \frac{1}{1/m \sum_{i=1}^{m} h_i} = \frac{1}{\bar{h}}$$

So that, the flow is equal to the inverse of the average headway.

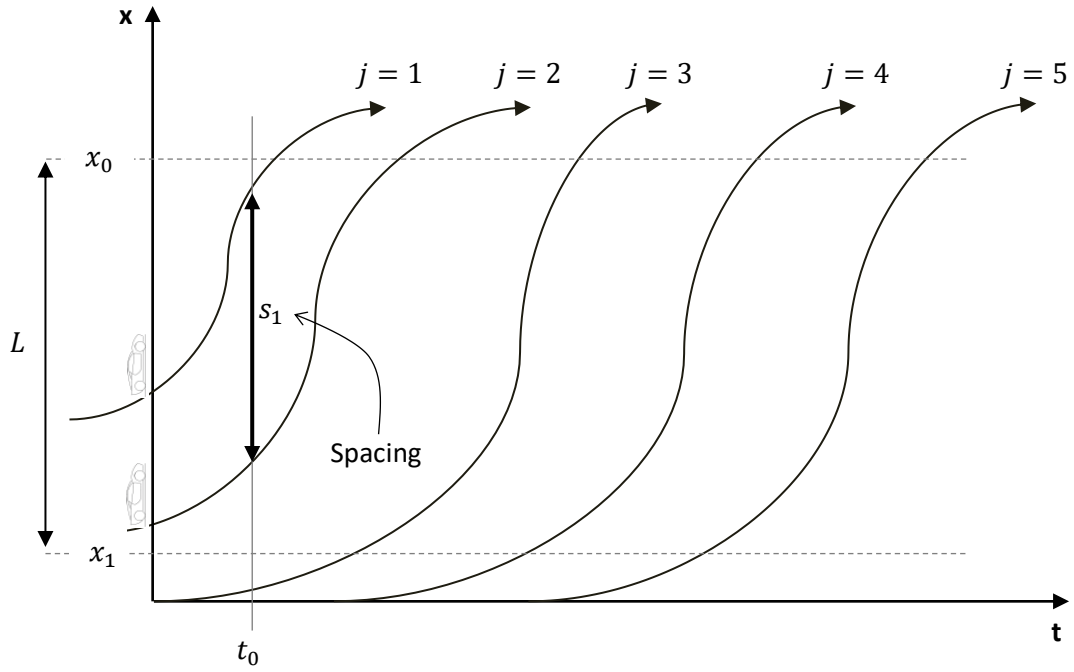


Figure 2. Spacing definition on a time – space diagram.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

The same analysis could be done but considering a spatial measurement region. This is the simultaneous measurement of all the vehicles in a length L of the infrastructure at a particular instant of time, t_0 (see Figure 2). This allows defining the vehicular spacing, s, as the distance between the same point of consecutive vehicles (e.g. the rear bumper in Figure 2). The spacing, is the second fundamental microscopic traffic variable. Note that, as before, s includes the vehicle length plus the empty space gap between vehicles. In Spanish the spacing is named "*espaciado*". The macroscopic variable equivalent to the spacing is the traffic density k, defined as the number of vehicles per unit distance observed at a given instant in a particular infrastructure. The traffic density k can be obtained from the measurement of the existing vehicles, n, on a given infrastructure length, L, at a particular instant of time, t_0. Note that in Figure 2, $n = 2$ (i.e. vehicles $j = 1$ and $j = 2$). Then, traffic density is defined as:

$$k = \frac{n}{L} \quad [veh/distance]$$

Similarly, the spacing (microscopic) and the density (macroscopic) variables are related, as one is the aggregation of the other. Namely:

$$k = \frac{n}{L} = \frac{n}{\sum_{j=1}^{n} s_j} = \frac{1}{1/n \sum_{j=1}^{n} s_j} = \frac{1}{\bar{s}}$$

So that the traffic density can be obtained as the inverse of the average spacing.

The third and last of the fundamental variables is the speed. The microscopic version of the variable is simply the vehicles' individual speed, v_i (i.e. the slope of the vehicles' trajectory at a given point in time and space). However, the macroscopic average, $\bar{v}$, is a bit more problematic. The average speed is defined as the arithmetic average of the individual speed of vehicles, the problem is that this average depends on which vehicles are considered. Amongst all possible selections of the space-time measurement region, the temporal region (e.g. (x_o, T) and the spatial region (L, t_0), define two particular cases, leading to the time-mean speed, $\bar{v}_t$, and to the space-mean speed, $\bar{v}_s$, respectively (see Figure 3). Time-mean and space-mean definitions, could be applied to any property of the travelling vehicles in addition to the travelling speed.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

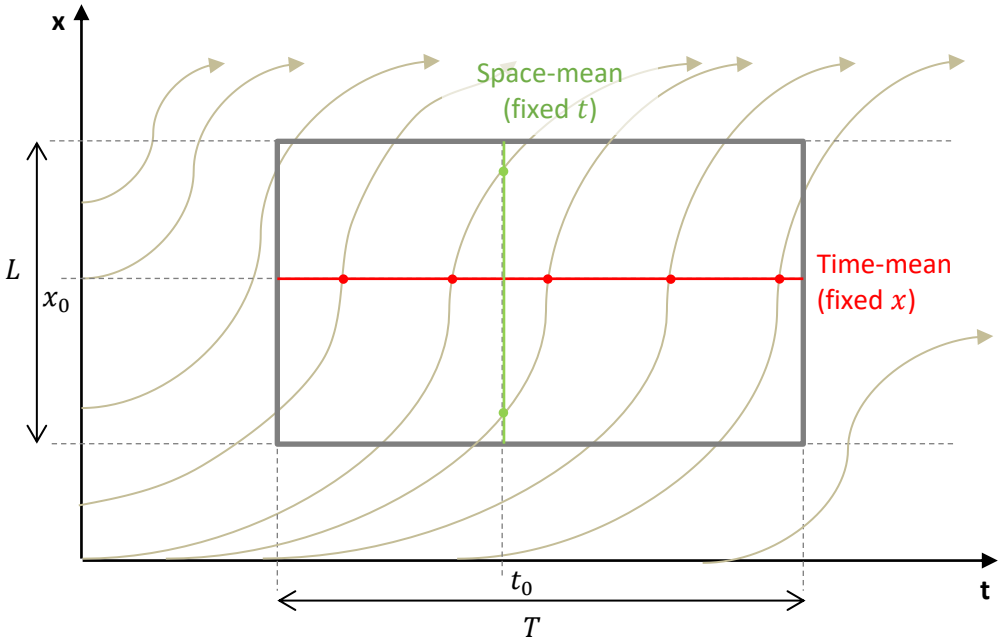


Figure 3. Time vs Space mean averages.

Then, according to the notation in Figure 1 and Figure 2, we define:

$$\bar{v}_t = \frac{\sum_{i=1}^{m} v_i}{m}$$

$$\bar{v}_s = \frac{\sum_{j=1}^{n} v_j}{n}$$

Because faster vehicles are overrepresented when seen by a stationary observer (i.e. in a temporal measurement region (x_o, T)) for any prevailing traffic state, $\bar{v}_t \geq \bar{v}_s$, and they are equal only when the speeds of all vehicles are constant. More precisely:

$$\bar{v}_t = \bar{v}_s + \frac{\sigma_{\bar{v}_s}^2}{\bar{v}_s}$$

where $\sigma_{\bar{v}_s}^2$ is the variance of the individual speeds measured over the spatial measurement region (L, t_0). Recall that the variance is always positive, and only zero when the variable (i.e. the individual speed, v_i) is constant. The previous relationship is named as the Wardrop relationship, in honor of J.G. Wardrop, the distinguished English mathematician and transport analyst who formulated it the first time in 1952 in his famous publication about "some theoretical aspects of road traffic research".

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Time means can be related with flow and space means to density. Note that any traffic state can be thought of being composed of l different families of vehicles (see Figure 4), where within each family traffic is stationary (i.e. constant vehicular speeds, headways and spacings). This is not limiting in any sense, as any non-stationary traffic state can be formulated in these terms (e.g. even a traffic state where all vehicles travel at different speeds, can be thought in terms of "families", where each family is composed of a single vehicle). Then we can define q_l, k_l and v_l, as the flow, density and speed of traffic within family l. Because the flow and density are additive magnitudes, the total flow, q, and the total density, k, can be obtained as:

$$q = \sum_l q_l \,; \quad k = \sum_l k_l$$

Also, we can state that:

$$\bar{v}_t = \frac{\sum_{i=1}^{m} v_i}{m} = \frac{\sum_l m_l v_l}{m} = \frac{(1/T)\sum_l m_l v_l}{(1/T)m} = \frac{\sum_l q_l v_l}{q}$$

where m_l is the number of vehicles of family l in the temporal region (x_o, T). This shows that time-means can be obtained as weighted averages, where the weights are the relative flows of each family of vehicles.

Similarly:

$$\bar{v}_s = \frac{\sum_{j=1}^{n} v_i}{n} = \frac{\sum_l n_l v_l}{n} = \frac{(1/L)\sum_l n_l v_l}{(1/L)n} = \frac{\sum_l k_l v_l}{k}$$

where n_l is the number of vehicles of family l in the spatial region (L, t_o). As before, this shows that space-means can be obtained as weighted averages, where the weights are the relative densities of each family of vehicles.

Figure 4 exemplifies these concepts. Consider three vehicles travelling on a circular track with a length of 30 km at different speeds. This can be seen as 3 families of vehicles:

$$l = 1 \rightarrow v_1 = 90\ km/h$$
$$l = 2 \rightarrow v_2 = 60\ km/h$$
$$l = 3 \rightarrow v_3 = 30\ km/h$$

From the spatial perspective, in the 30 km of the track there are:

1 vehicle @ 90 km/h ($n_1 = 1$)
1 vehicle @ 60 km/h ($n_2 = 1$)
1 vehicle @ 30 km/h ($n_3 = 1$)

So that the space mean speed is simply the arithmetic average of their speeds, $\bar{v}_s = 60\ km/h$, since the relative density of the 3 families is the same (i.e. 1/3).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

In contrast, from the temporal perspective, one observer standing at any fixed location at the side of the track during $T = 1h$, would observe:

3 vehicles @ 90 km/h ($m_1 = 3$)
2 vehicles @ 60 km/h ($m_2 = 2$)
1 vehicle @ 30 km/h ($m_3 = 1$)

Then, the time mean speed computed as a weighted average considering the relative flows is $\bar{v}_t = 70\ km/h$. Note that vehicles travelling at higher speeds are overrepresented with respect to their density.

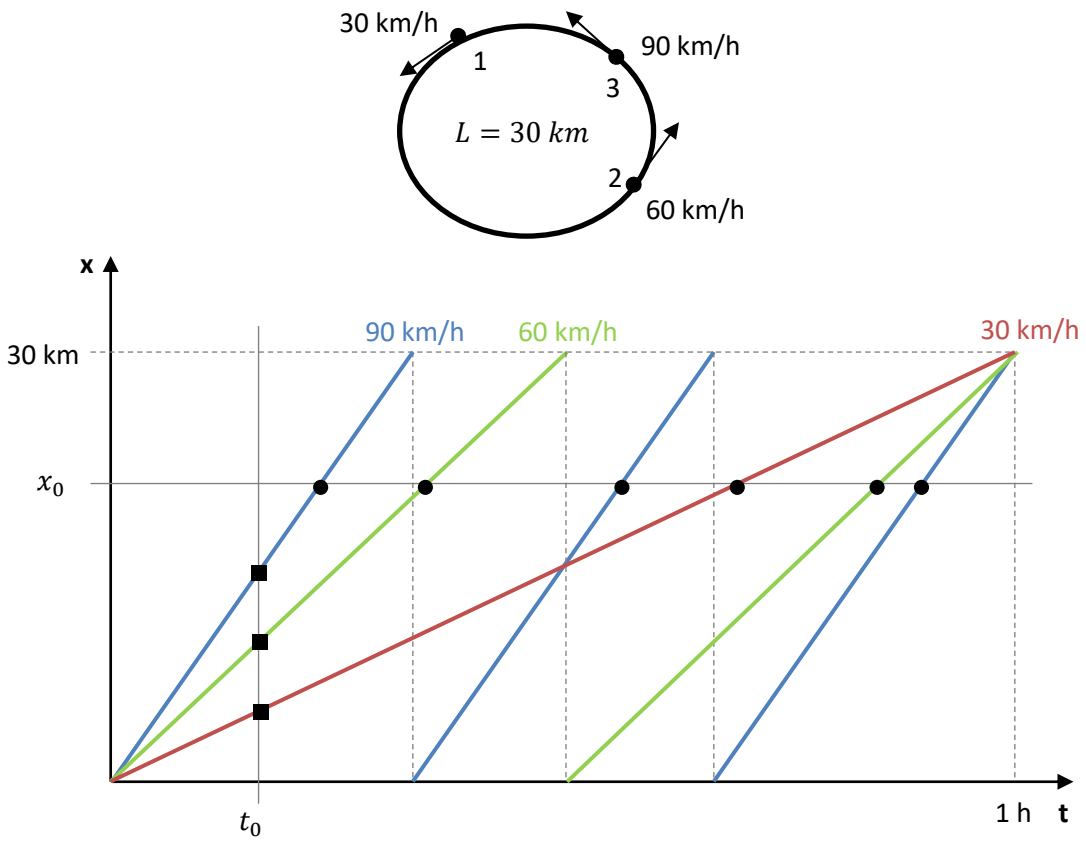


Figure 4. Time vs Space mean speed.

Table 1 provides a summary of what we have seen so far. Additionally, it includes two additional pieces of information. The "Simulators" column shows computer implementations of the different traffic modeling approaches. The Cell Transmission Model (originally developed by Prof. C.F. Daganzo) is a finite difference

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

implementation of the LWR macroscopic model. This, can be coded into any computing language, and several implementations can be found as freeware. On the other hand, microscopic traffic simulators tend to be commercial software under license, and only SUMO represents an open-source alternative developed by the academic community. Table 1 also provides the macro-micro equivalence for a couple of models. For instance, the Greenberg macroscopic $k - v$ is the macroscopic equivalent of the 3rd generation of the General Motors car-following theories. Take this only as informative, as we are not going to discuss the details of these models in this chapter.

Table 1. Macro and micro approaches to traffic modelling – Variables, models and simulators

	Variables			Models	Simulators
MACRO	Flow (q)	Density (k)	Average Speed ($\bar{v}$)	Continuous theories LWR – KW (shock wave theory)	Cell Transmission Model
MICRO	Headway (h)	Spacing (s)	Instantaneous speed (v_i)	Car following theories	e.g. AIMSUN, VISSIM, PARAMICS, SUMO
MACRO-MICRO Relationships	$q = \frac{1}{\bar{h}}$	$k = \frac{1}{\bar{s}}$	$\bar{v} = \frac{\sum_i^n v_i}{n}$	e.g. Greenberg (macro $k - v$) – 3rd General Motors (micro car following) e.g. Newell Triangular diagram (macro $q - k$)– Forbes min. safety distance (micro car followig)	
Aggregation types	(x, T)	(L, t)	$(x, T) \rightarrow \bar{v}_t$ $(L, t) \rightarrow \bar{v}_s$		

3. The fundamental equation of traffic ($q = k\bar{v}$)

The three fundamental variables of traffic (i.e. flow q, density k, and average speed, $\bar{v}$) are related by the so called "Fundamental Equation of Traffic", which states that the flow is equal to the traffic density times the average speed:

$$q = k\bar{v}$$

Obviously, the fundamental equation can also be formulated in terms of the average microscopic variables:

$$\bar{s} = \bar{h}\bar{v}$$

The fundamental equation of traffic directly results from the previous definitions of the variables, meaning that it is true "by definition". This implies that it holds everywhere, for all kind of infrastructures and for all possible traffic states. In spite of this, for stationary traffic (i.e. constant vehicular speeds, headways and spacings) it is

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

particularly easy to prove (see Figure 5). If we carefully define a time-space measurement region (L, T) so that $L/T = \bar{v}$, where $\bar{v} = v_i \ \forall i$ because traffic is stationary, then:

$$q = \frac{m}{T};\ k = \frac{m}{L}$$

$$\frac{q}{k} = \frac{L}{T} = \bar{v}$$

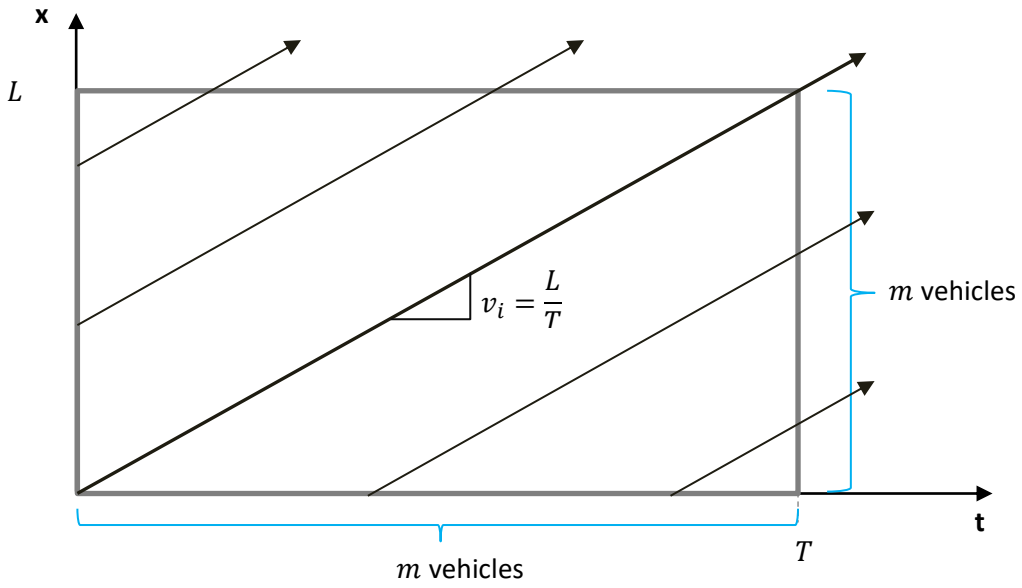


Figure 5. Stationary traffic on a time – space diagram.

In spite of the previous derivation of the fundamental equation of traffic for a stationary traffic state, recall that the equation holds even for non-stationary traffic (i.e. when the vehicular speeds are not constant). Then, an additional issue appears, because the average speed, $\bar{v}$, is not well defined. Should we consider the time-mean speed, $\bar{v}_t$ in the fundamental equation? or should we consider the space-mean speed $\bar{v}_s$? Note that, because they are different in non-stationary traffic, only one of the options can be true.

To give an answer to this question, think again of a non-stationary traffic composed of l different families of vehicles. Within each family, traffic is stationary with q_l, k_l, v_l, and because we have already proved that the fundamental equation holds in stationary traffic, within each family we have:

$$q_l = k_l v_l$$

Then, by simply working with the algebra que can state that:

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$q = \sum_l q_l = \sum_l k_l v_l = k\left[\frac{\sum_l k_l v_l}{k}\right] = k\bar{v}_s$$

Recall that a weighted average of vehicles speeds, where the weights are the relative densities, yields the space-mean speed. So, this proves that in non-stationary traffic, the fundamental equation holds only if the speed is obtained as a space-mean. This is:

$$q = k\bar{v}_s$$

4. The vehicles' conservation equation

One of the postulates of all models of traffic flow is that vehicles are "conserved". This means that they cannot "disappear", which seems a quite reasonable postulate. The formulation of the vehicles' conservation law is equivalent to any other of the existing conservation laws (mass, energy, etc.) and it states that in any closed system the number of vehicles entering minus the number of vehicles exiting, must be equal to the difference in the number of vehicles stored inside the system.

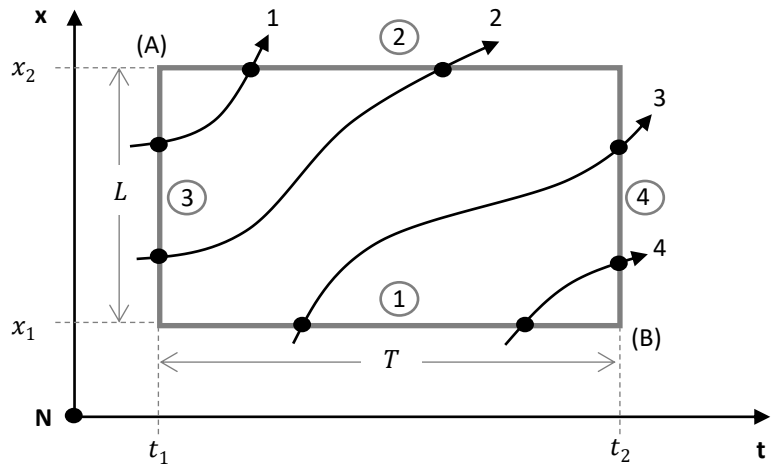


Figure 6. Derivation of the vehicles' conservation equation from a trajectories diagram.

There are several ways of analytically formulating the vehicles' conservation equation, but using a space-time diagram is quite simple and convenient. In the context of traffic flow, a "closed system" means a closed space-time region, where all the entrances and exits to the infrastructure are monitored. Figure 6 shows a closed (x, t) region with dimensions (L, T) and some vehicle trajectories. The infrastructure considered in Figure 6 could be a section of a one-way street so that vehicles can only enter through location x_1 and exit through x_2. This section of the street is observed in the time period between t_1 and t_2. In this context, note that the vehicles entering

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

the section of the street through x_1 during the period T, (i.e. crossing border 1 of the closed space-time region) are vehicles 3 and 4 (i.e. two vehicles). The notation for this variable will be $m_1 = 2$. The number of vehicles exiting through x_2 during the same period (i.e. crossing border 2) are vehicles 1 and 2. In this case $m_2 = 2$. Also, you can see from Figure 6 that, initially (i.e. at t_1), there were two vehicles inside the section of street of length L. These are vehicles 1 and 2. Because these are spatial borders, the notation used in this case for the number of trajectories crossing border 3 is $n_3 = 2$. Finally, the number of vehicles "stored" inside the section of street at the end of the observation period, t_2 (i.e. trajectories crossing border 4) is $n_4 = 2$ (i.e. vehicles 3 and 4). Given these definitions, the conservation of vehicles can be simply formulated by imposing that the number of trajectories entering the closed (x, t) region must be equal to the number of trajectories exiting[1]. This is:

$$m_1 + n_3 = m_2 + n_4$$

Note that the physical meaning of the previous equation is exactly that of the conservation law:

$$entering(m_1) + initial\ accumulaion(n_3) = exiting(m_2) + final\ accumulation(n_4)$$

And working with the algebra, allows obtaining a more familiar form of the vehicles' conservation equation:

$$m_2 - m_1 = n_3 - n_4$$

$$\frac{m_2 - m_1}{(x_2 - x_1)(t_2 - t_1)} = \frac{n_3 - n_4}{(x_2 - x_1)(t_2 - t_1)}$$

$$\frac{m_1 + n_3}{LT} = \frac{m_2 + n_4}{LT}$$

Note that m's divided by T are flows, while n's divided by L are densities. Therefore:

$$\frac{q_2 - q_1}{(x_2 - x_1)} = \frac{k_3 - k_4}{(t_2 - t_1)}$$

$$\frac{\Delta q}{\Delta x} = -\frac{\Delta k}{\Delta t}$$

Which says that the variation of flow with respect to space is equal to the minus variation of the density with respect to time, representing an equivalent formulation of the conservation equation. Considering that typically we deal with many vehicles and that in this case the vehicle count function can be considered continuous, the previous equation can be rewritten as:

[1] This is equivalent to imposing that the change in the cumulative vehicle count, N, between two points in the time-space plane (e.g. points A and B in Figure 7) must be the same independently of the followed path (e.g. Borders 1 and 3 or Borders 2 and 4).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$\frac{\partial q}{\partial x} = -\frac{\partial k}{\partial t}$$

which represents the most typical expression for the vehicles' conservation equation.

4.1. Relative flow seen by a moving observer

As a side note, and as an application of the conservation equation, it is interesting to determine the relative flow seen by an observer that is moving with traffic. The relative flow can be interpreted as the number of overtakings per unit time seen by the moving observer. Note that this depends on the prevailing flow and on speed of the observer (e.g. if the observer travels at the same speed as the average traffic, the relative flow would be zero).

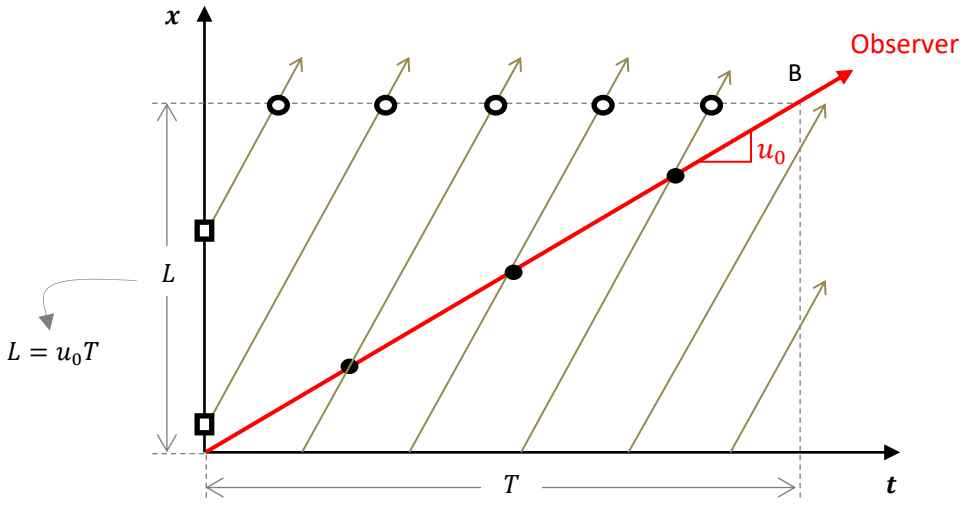


Figure 7. Derivation of the relative flow seen by a moving observer from the conservation equation.

In order to derive an analytical expression for this relative flow, look at Figure 7, where in the time – space plane we define a triangular control area of sizes L and T. By applying the conservation of vehicles in this closed area, we can say:

$$q_0 T + kL = qT$$

Where q_0 is the relative flow seen by an observer moving at speed u_0, and q and k are respectively the flow and density of the prevailing traffic state. Note that $q_0 T$ represents the number of times the observer is overtaken by other vehicles in the duration T (i.e. black full dots in Figure 7), kL is the number of vehicles on L at the beginning of T (i.e. squares in Figure 7), and qT is the total number of vehicles leaving L during T (i.e. empty dots

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

in Figure 7). By realizing that $L = u_0 T$, and substituting in the previous equation, we obtain (after simplyfing for T):

$$q_0 + k u_0 = q$$

And finally:

$$q_0 = q - k u_0$$

The previous equation states that the relative flow seen by a moving observer is equal to the prevailing flow minus the prevailing density time the speed of the observer. Note that, according to the fundamental equation of traffic, $q = kv$, where v is the average speed of traffic. This means that:

$$q_0 = k(v - u_0) = k v_{rel}$$

Where $v_{rel} = v - u_0$ is the relative speed between traffic and the observer. Note that if the relative speed is zero, the relative flow it is also zero. In addition, if the speed of the observer is larger than that of traffic, the relative flow is negative. This means that the observer overtakes the other vehicles in the traffic stream, and seen from the moving observer perspective, traffic is like running backwards (see Figure 8). Finally, consider that this expression can also be applied if the observer travels in the opposite direction of traffic. In such case the speed of the observer is negative, and the relative flow would be larger than the regular flow seen by a static observer.

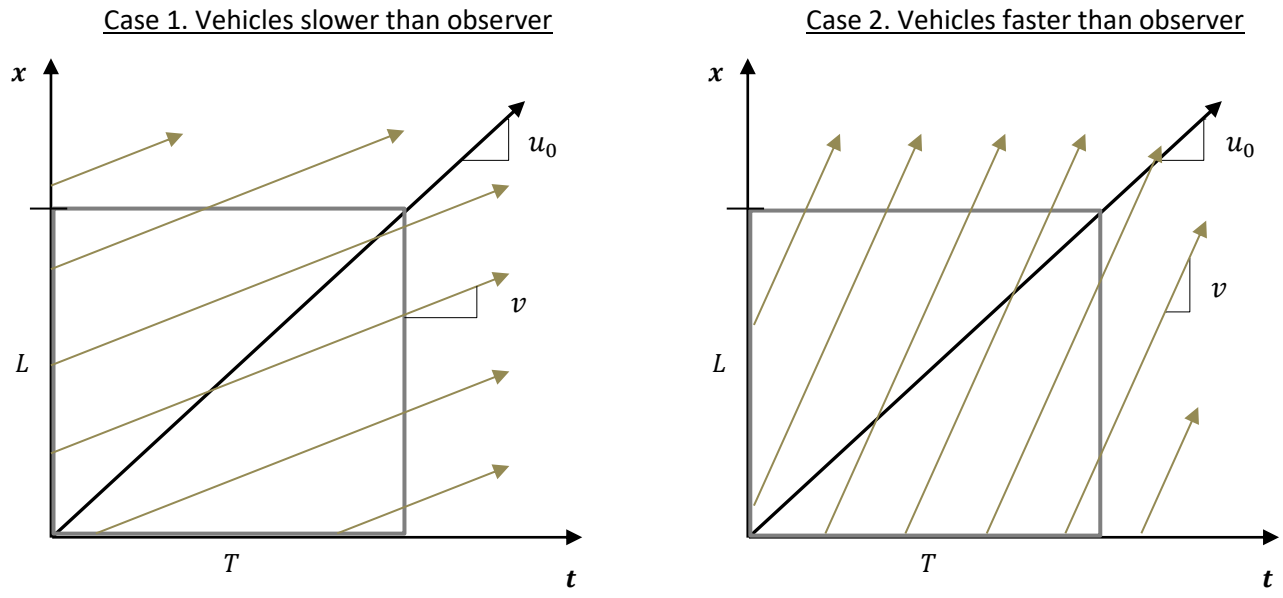


Figure 8. Moving observer on a trajectories diagram.

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

5. Traffic diagrams

Along with the vehicles' conservation equation, the second postulate of macroscopic traffic flow models is the existence of an equation of state. In other words, this means that the state of the system can be univocally determined by applying one equation to a state variable. Typically, this equation of state is plotted in a relevant coordinate axis, defining what is known as a traffic diagram. In the context of traffic flow modelling, there are 3 fundamental variables, which univocally define the traffic state (i.e. q, k, $\bar{v}$). This means that if we want only one degree of freedom (i.e. one state variable), we need two equations between these variables. We already know one, the fundamental equation of traffic (i.e. $q = k\bar{v}$). Still, we need another equation between any pair of the variables.

5.1. Greenshields, Greenberg and Edie's $k - v$ models

The possibility of deriving a relationship between a pair of traffic variables appeared from observation. The first traffic observations were performed by the civil engineer B. D. Greenshields, as early as 1933-35. He carried out experiments to measure traffic density and speed using photographic measurement methods for the first time. Measuring traffic with the equipment available at the 1930's, was not an easy task, so that obtaining a single measuring point was challenging (see http://www.krbalek.cz/For_students/mds/clanky/Greenshields.pdf) for a more detailed description of the Greenshields experiments). In spite of this, Greenshields managed to obtain 6 measurement points, which he represented in a density - speed coordinate axis (see Figure 9). Observing the fairly linear arrangement of the measurement points, Greenshields proposed a linear relationship between speed and density. This linear k-v model was the first proposal for the equation of state, and the first traffic diagram ever.

From this historical note on the appearance of traffic diagrams, two important properties are already apparent:

- Traffic diagrams are empirical. They are obtained from observation (not by definition as in the fundamental equation). This means, that they are valid for the infrastructure and the drivers/vehicles observed. So, care must be taken when extrapolating traffic diagrams to infrastructures and drivers significantly different from where they were measured.

- Traffic diagrams are obtained from regression to data points. This means that they are true on average, but that it is possible that some measurement point lay significantly apart from the average regression.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

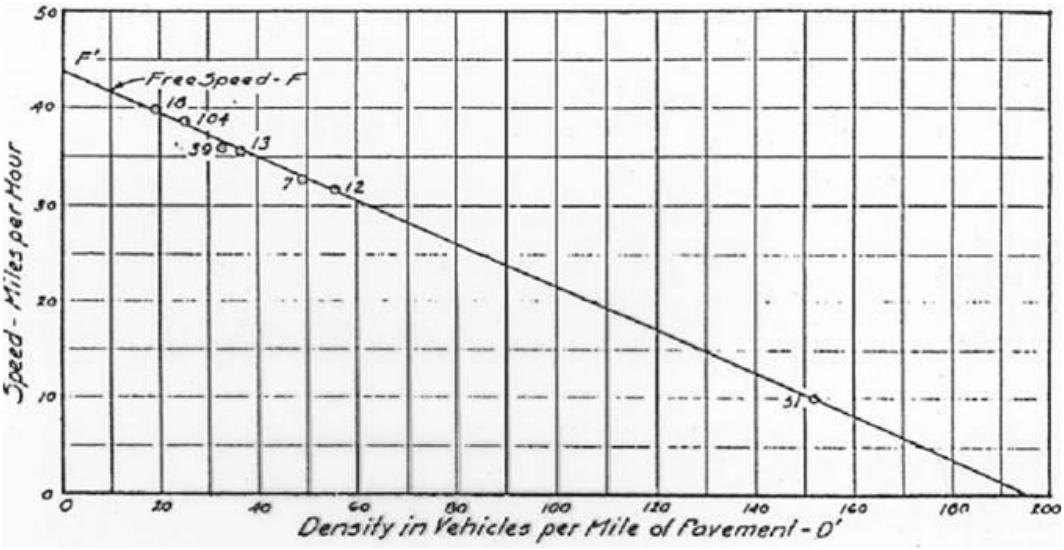


Figure 9. The original Greenshields k-v model

The regression of a functional form to data points for obtaining the traffic diagram (i.e. the calibration of the diagram) yields its parameters. The typical parameters of traffic diagrams are:

- v_f, the free flow speed => This is the maximum average speed that drivers feel safe and comfortable to drive at when the density is very low. A typical value for a freeway lane is $v_f = 100 - 120$ km/h, depending on the physical layout, and although this is influenced by the prevailing speed limit.

- k_j, the jam density => This is the maximum density at the infrastructure, when the vehicles are in a gridlock jam, completely stopped. A typical value for a freeway lane is $k_j = 125 - 150$ veh/km.

- q_{max}, the maximum flow, referred as the capacity => This is the maximum throughput that the infrastructure can hold. In turn, v_0 and k_0 represent the corresponding optimal speed and density, respectively, for which the capacity point is obtained. A typical value for a freeway lane is $q_{max} = 2000 - 2200$ veh/h, $k_0 = 20 - 25$ veh/km and $v_0 = 70 - 90$ km/h.

For instance, the Greenshields linear k-v model, requires the calibration of two parameters, v_f and k_j to obtain the linear functional form as:

$$v = v_f \left(1 - \frac{k}{k_j}\right)$$

Figure 10 shows all the previous parameters in the linear Greenshields k-v diagram.

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

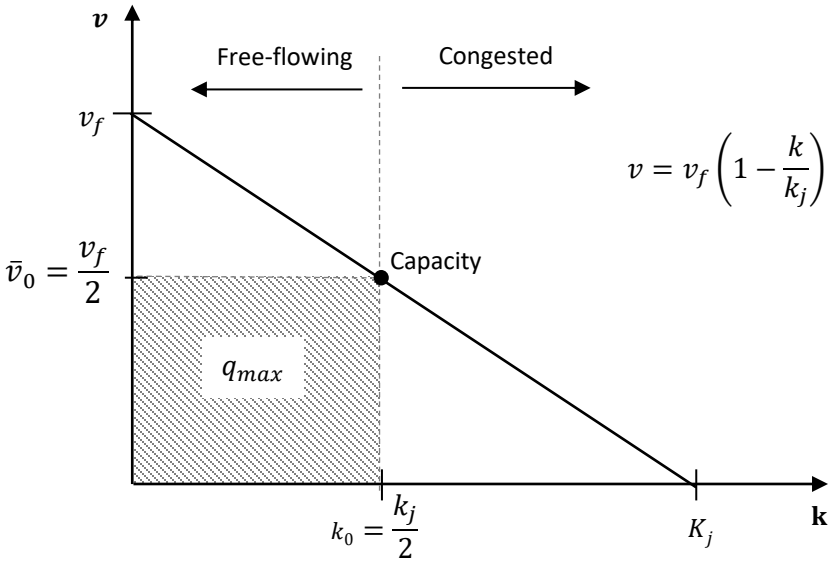


Figure 10. Greenshields $k - v$ linear model (1934).

Note, that the variable not directly represented in the coordinate axis (i.e. the flow in this case) can be obtained by applying the fundamental equation, or figuring out the graphical representation on the diagram. In a k-v diagram the flow is obtained as the area defined by the coordinate points of a given traffic state. The capacity point in the Greenshields diagram illustrates that, although this model is of historical importance and academically used (because of its simplicity), it is not accurate. Note for instance that the optimal speed, v_0 results to be $v_f/2$, while in reality is significantly larger, and $k_0 = k_j/2$, while in reality is much smaller.

The capacity point divides the diagram in two parts. For densities higher than k_0, the flow is reduced with growing densities. This is the most precise definition of congested traffic. So, the right-hand side of the diagram Figure 10 corresponds to congested traffic states. In contrast, the left-hand side corresponds to free-flowing traffic states, where an increase of the density is translated into an increase in the flow.

Since the first linear k-v model proposed by Greenshields, and given his limitations to reproduce the empirical data that became available with the passage of time, many other functional forms were (and still are) proposed to describe traffic diagrams. In the early days, dealing with data scarcity, it could be said that there were as many functional proposals for traffic diagrams as traffic databases. In spite of this, some functional relationships have stood the pass of time, and it is worth referring to them here:

- The Greenberg k-v model, proposed by Harold Greenberg (1959) who considered traffic as continuous compressible fluid (an acceptable assumption for congested traffic states, but not in light traffic conditions) and derived analytically a logarithmic relationship between density and speed, as shown in Figure 11.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Greenberg validated his logarithmic functional form by calibrating the two parameters of his model (i.e. the optimal speed, v_0, and the jam density, k_j) with data measured at the Lincoln Tunnel in New York city and at the Merritt Parkway in Connecticut, which included for the first time a significant number of measurements of congested traffic. The model cannot be applied directly for low densities (i.e. free-flow traffic) as speed tends to infinity when density tends to zero. One possible solution is to truncate the model at the free-flow speed, v_f, adding a third calibration parameter to the model and defining what is called a two-regime diagram (i.e. different functional forms for free-flowing and congested traffic).

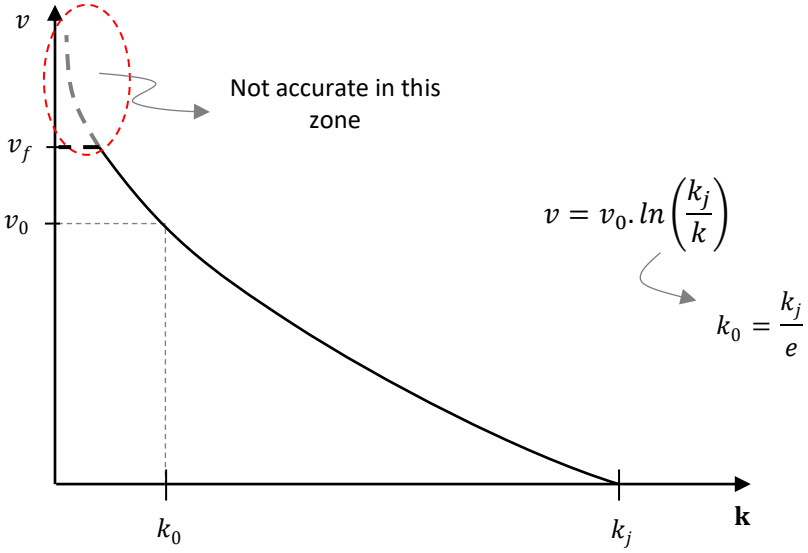


Figure 11. Greenberg model (1959) – logarithmic approach to $k - v$.

- The Underwood k-v model, proposed by Robin T. Underwood (1961) aimed to overcome the limitation of Greenberg's model for free-flowing traffic and put forward an exponential model as shown in Figure 12. Underwood calibrated the two parameters of the model (i.e. the free-flow speed, v_f, and the optimal density, k_0) at two locations: the Merritt Parkway in Connecticut (same as the calibration location for the Greenberg model) and at the Princess Highway in Victoria (south Australia). In Underwood's model, speed becomes zero only when density reaches infinity. Some academics consider this as a drawback of this model, while others postulate that actually, even when traffic density reaches the jam value, k_j, the average speed is not zero as traffic continues moving forward in a stop and go fashion. Hence, using Underwood's model for predicting speeds at high densities is still acceptable.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

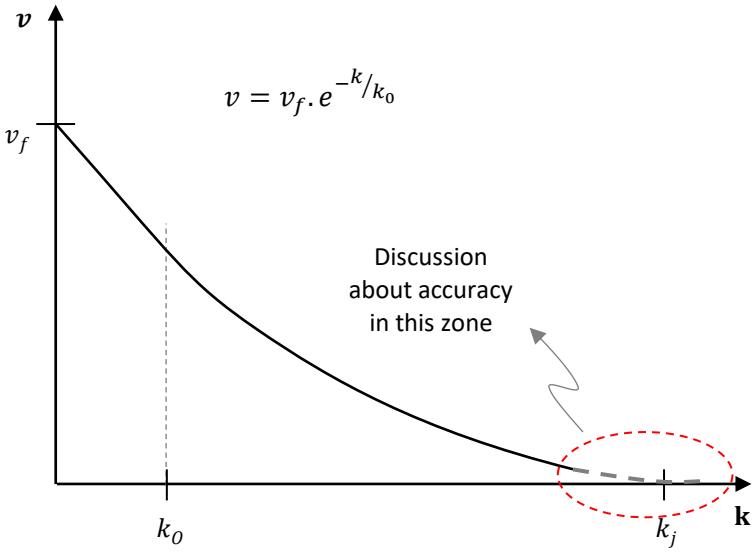


Figure 12. Underwood model (1961) – exponential approach to k-v.

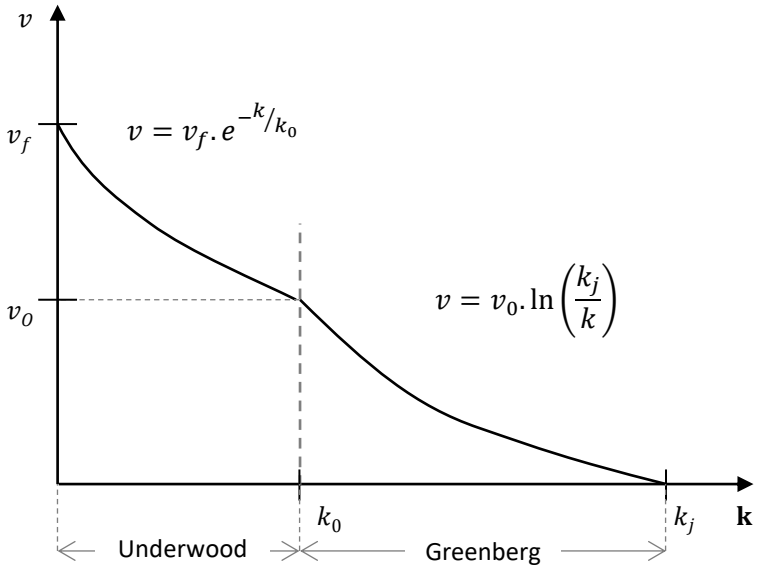


Figure 13. Edie k –v model (1961).

- Leslie C. Edie proposed for the first time a two-regime model, using the Underwood model for free-flow conditions, and the Greenberg model for congested traffic, resulting in a discontinuous exponential form (see Figure 13). Edie's proposal requires the calibration of 4 parameters: the free-flow speed, v_f, the jam

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

density, k_j, and the breakpoint between free-flowing and congested traffic (i.e. the optimal point at capacity, v_0 and k_0). Note that the calibration of "optimal" values at capacity is more cumbersome than those of parameters at jam or light traffic conditions, because establishing that a particular traffic state corresponds to capacity conditions is not evident and easily determined. This is the reason why v_0 and k_0 are usually determined from statistical regression, without much support from its physical interpretation. This can be seen as a drawback (e.g. of Greenberg, Underwood or Edie's models) with respect to other models (e.g. Greenshields) relying only of free-flow or jam parameters.

As an end for this section devoted to traffic diagrams, it should be understood that from the functional form of one diagram (e.g. a k-v model) all the other derived diagrams (i.e. all the possible 2-dimensional plots of different combinations of variables) can be obtained using the fundamental equation of traffic. It is a good exercise for practicing to derive the different type of diagrams that would result, for instance, from the linear Greenshields k-v model. For illustrative purposes, Figure 14 shows the conceptual generic form of several possible diagrams.

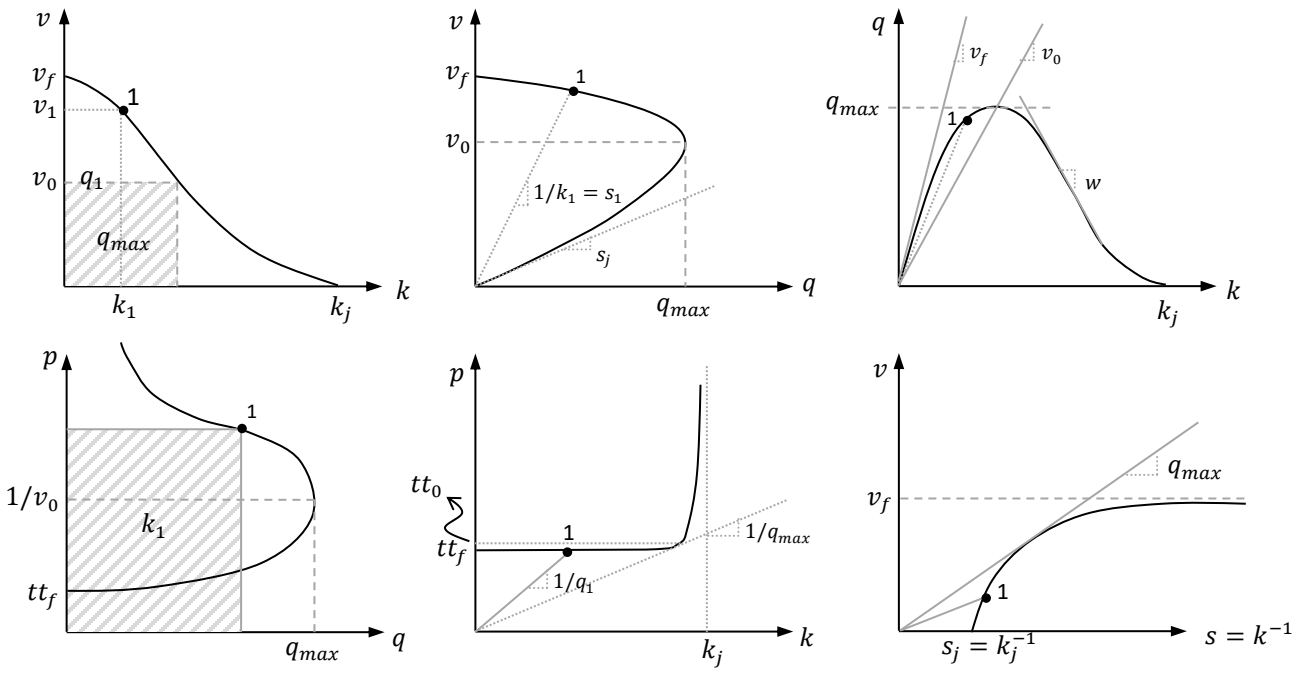


Figure 14. Generic representation of different traffic diagrams
Note: p is defined as the pace, $p = 1/v$, which represents the travel time (tt) per unit distance.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

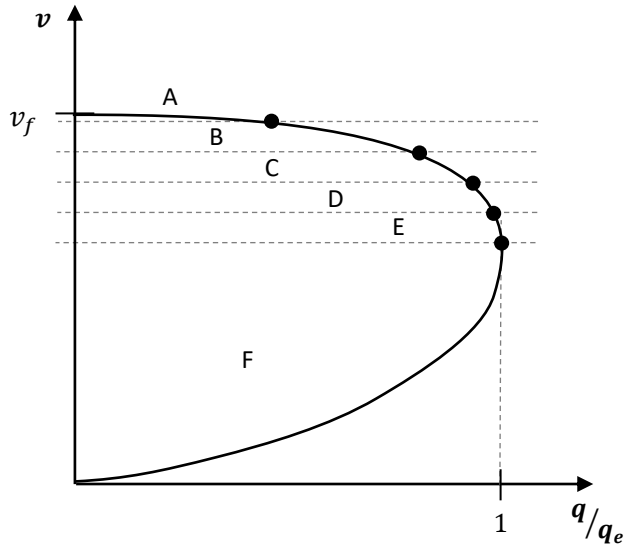
Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

5.2. Traffic engineering manuals: a warning

Manuals on traffic engineering, and specifically the Highway Capacity Manual (HCM), do have a long tradition in the analysis of traffic flows. The first edition of the Highway Capacity Manual was released in 1950 as a result of a collaborative effort between the Transportation Research Board (TRB) and the Bureau of Public Roads (predecessor to the Federal Highway Administration) in the USA. This first edition of the HCM contained 147 pages, presenting concepts, guidelines, and procedures for computing the capacity and quality of service of highway facilities. Today, 70 years and 6 editions later, the HCM is a vast publication (mainly online) which serves as a reference for evaluating the multimodal operation of streets, highways, freeways, arterial roads, roundabouts, signalized and unsignalized intersections, rural highways, and the effects of mass transit, pedestrians, and bicycles on the performance of these systems.

Despite the HCM expansion to multiple types of infrastructures, and the incorporation of computational procedures, the essence of the HCM is still the same as in its origins. It consists on a methodology to define different levels of service (LoS) on "traffic" diagrams measured in the different types of infrastructures (see Figure 15). LoS are categorized between A to F, where A is the best LoS possible, E corresponds to capacity and F to congested states in the infrastructure. The typical application of the HCM consists in the evaluation of the LoS in a given infrastructure for a particular demand input.



q_e= Equivalent flow in ideal conditions

Figure 15. HCM levels of service on a $v-q$ diagram.

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

In line of our previous discussion of traffic diagrams, two warnings should be risen in relation to the usage of traffic engineering manuals:

- Traffic diagrams result from statistical regression of empirical data to some selected functional form. Many errors could exist in this calibration process (e.g. wrong specification of the functional form, low correlation of the function to data, small statistical significance due to the lack of data, …) which are not evident from the utilization of the manual. This implies that the utilization and results of this type of manuals give a false sense of precision.

- Traffic diagrams are valid for the particular infrastructure and driver types for which they were measured. Diagrams may not hold for a particular application where some of the boundary conditions are different.

In conclusion, the warning is that results of applying traffic engineering manuals are not precise and may be unreliable. So, do not make important decisions made only on their outputs, without taking your own measurements and making comparative assessments.

6. Macroscopic modeling of traffic flow: LWR - Kinematic Wave Theory

The objective of macroscopic traffic flow modeling is to predict the evolution of traffic variables in time and space. The adjective "macroscopic" means that the focus is not on individual vehicles, but on the prediction of the aggregate variables (i.e. q, k and $\bar{v}$). This is especially interesting when queued traffic conditions prevail so that the theory could help in answering questions like: when does queue arrive to a point of interest? (e.g. important to prevent blockages), how far does the queue grow? or when does the queue dissipate. Note that traffic flow theory deals with the evolution of traffic states, and it is very well suited to answer questions related to the physical extension of queues. However, there are easier ways of answering questions about delays (e.g. use cumulative count curves).

The first and fundamental macroscopic approach to traffic flow modelling is the LWR (Lighthill-Whitham-Richards) model, also named the continuous model of traffic flow, the kinematic wave theory, or the shock-wave theory. All refer to the same original model by LWR. This is a traffic flow model for heavy traffic (i.e. dense, queued traffic) where an increase of the traffic density implies a reduction in the average travelling speed, and vice versa (see Figure 16).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

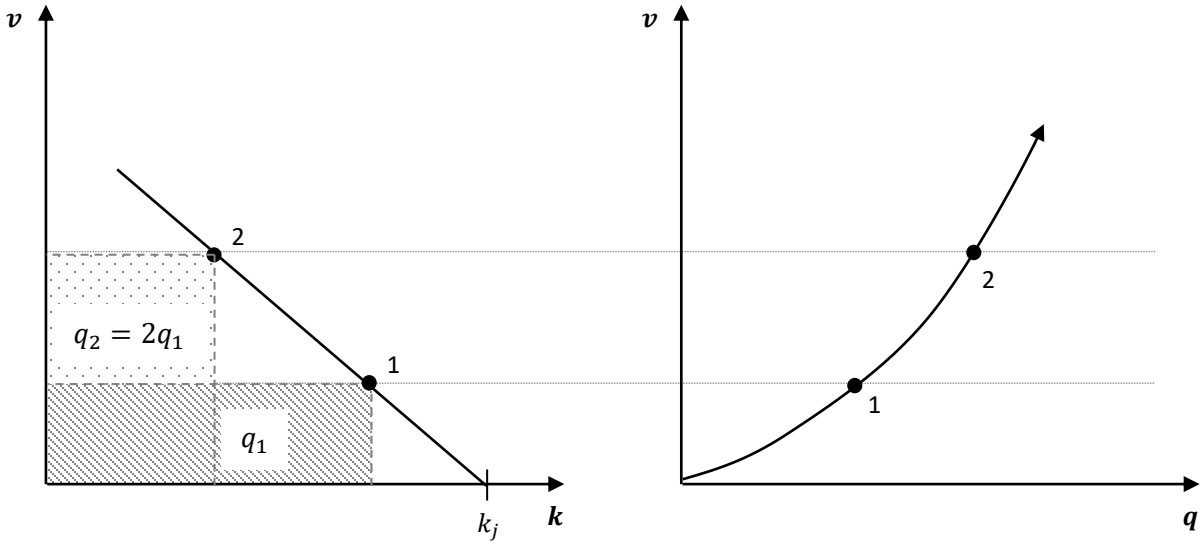


Figure 16. Congested traffic states on a $v - q$ diagram from a $k - v$ model.

Note that the causality for heavy traffic can be twofold. One possibility is that it appears as a result of a flow restriction (Case a; see Figure 17). When the demand is large, approaching or going above the maximum flow that can go through the restriction, the density of vehicles upstream of the bottleneck grows leading to a speed reduction to ensure safety. Another option is that heavy traffic appears as a result of a speed limitation or a moving bottleneck on the infrastructure (Case b; see Figure 18). In such case, drivers need to adapt to the lower restricted travelling speed, accepting a higher density of vehicles.

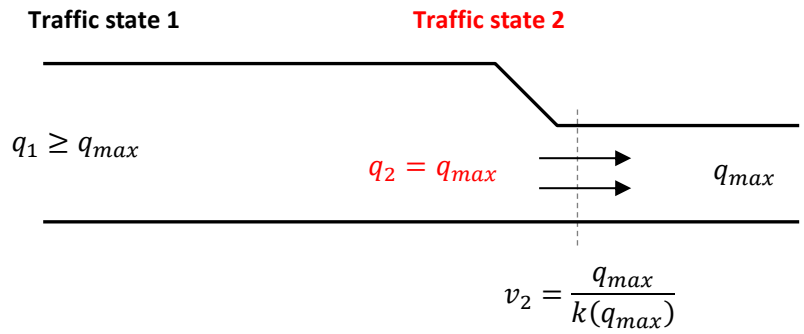


Figure 17. Case (a): Flow restriction

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

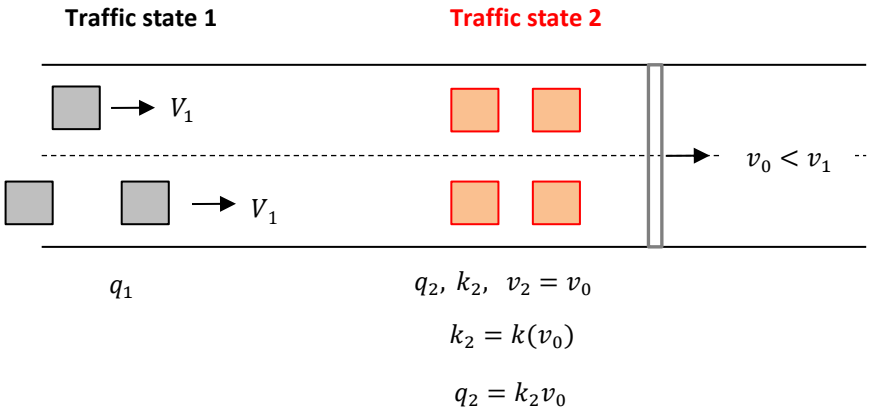


Figure 18. Case (b): moving obstruction.

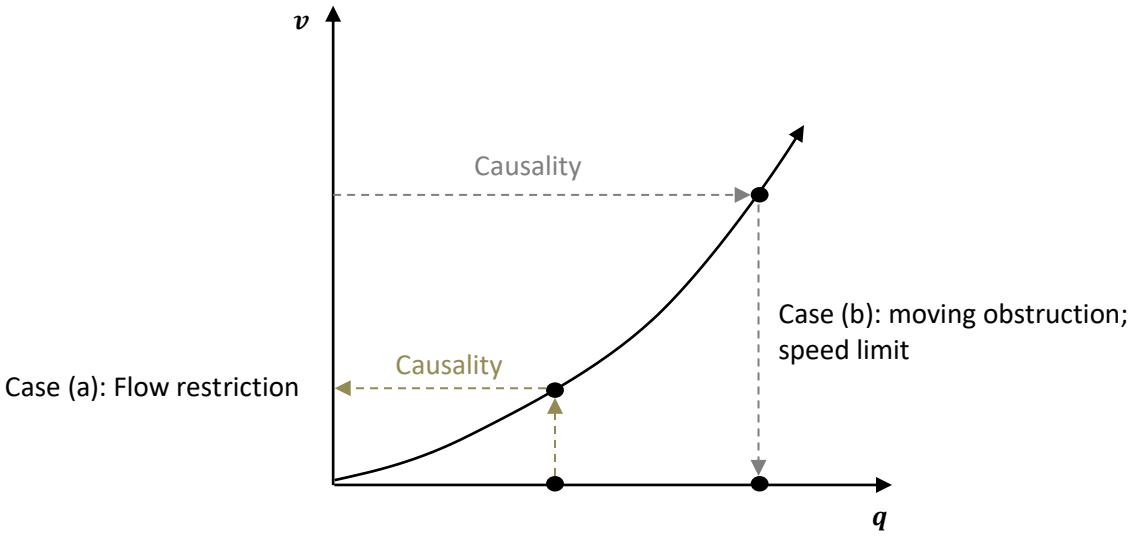


Figure 19. Two possible causes of heavy traffic on a $q - v$ diagram.

In any case, whenever a vehicle joins or leaves a congested traffic state, there is a transition to the new traffic state, which involves a deceleration or an acceleration process. The LWR traffic flow theory for dense traffic aims to predict the evolution of these transitions. The theory is based on the hydrodynamic analogy and considers that traffic state transitions are propagated as waves. Whenever there is a change in the traffic state, a traffic wave is generated. A wave is a transition, an interphase between different traffic conditions (i.e. different q, k or $\bar{v}$). The crossing of a wave informs drivers that they need to adapt to the new traffic conditions. Traffic waves evolve in time and space and they have their own trajectories. It is not the trajectory of any vehicle, but the

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

trajectory of information, the trajectory which locates in time and space the changes in traffic conditions. As an example, Figure 20 shows (in red) the traffic waves generated when vehicles decelerate to stop in a traffic signal for accelerating later on, when the signal turns green. Note that if we could predict the evolution of these waves, we could predict the evolution of traffic states and their transitions, which is precisely the objective of a macroscopic traffic flow model. In summary, LRW theory deals with the prediction of the trajectories of the traffic waves. The following link illustrates quite well the concept of a traffic wave (https://youtu.be/Fn3HMAaEfcQ).

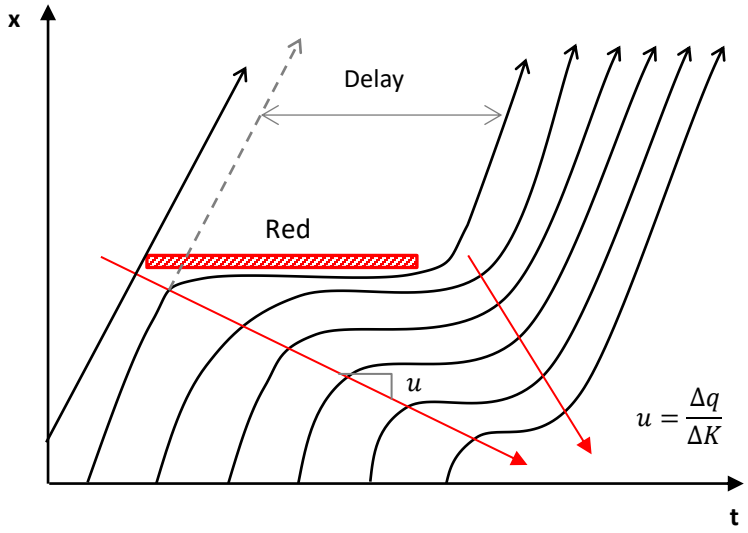


Figure 20. Example of traffic waves (interphases) in a traffic signal.

Given the previous context, the LWR theory assumes two postulates:

- The conservation of vehicles
- The existence of an equation of state (e.g. $q(k)$, the flow as a function of the density, which acts as the state variable). This is the existence of a traffic diagram.

With these two postulates, it is possible to predict the speed of a traffic wave between any two given traffic states.

6.1. Speed of a traffic wave

LWR theory is based on the prediction of the evolution of traffic waves. So, it is fundamental being able to determine the speed of a wave from the knowledge of attributes of the "colliding" traffic states. To this end, consider two different traffic states: $U(q^U, k^U)$ upstream and $D(q^D, k^D)$ downstream. In between there needs to exist a transition, a wave. If traffic states U and D are stationary (i.e. they do not change neither in time nor

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

in space), the trajectory defined by the traffic wave is linear (i.e. constant speed). The objective is to determine the speed, u, of this traffic wave given the traffic variables defining U and D (i.e. (q^U, k^U) and (q^D, k^D).

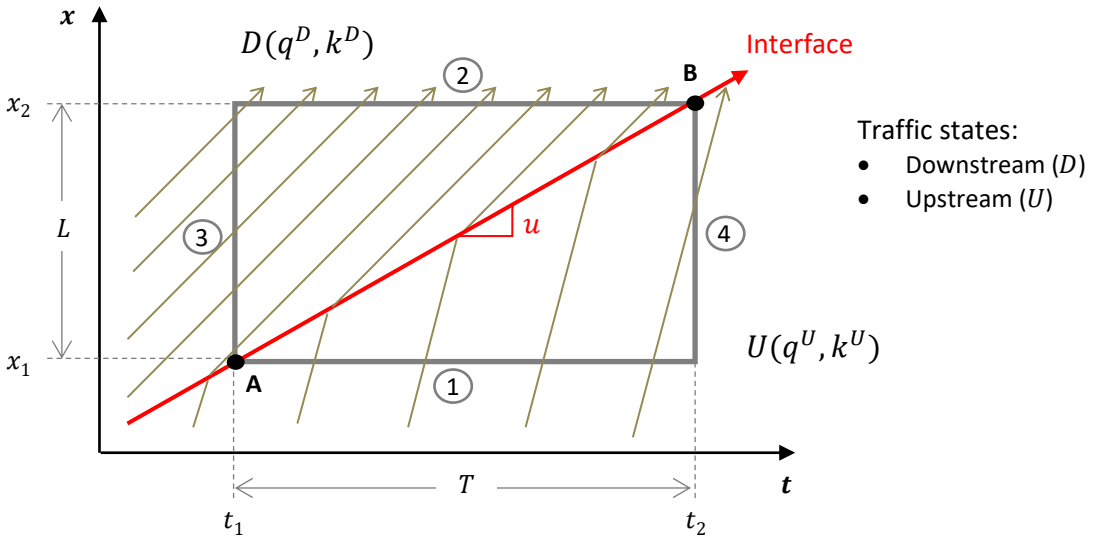


Figure 21. Derivation of the speed of a traffic wave from the conservation equation.

To this end, the conservation equation can be used. Look at Figure 21 and recall the derivation process for the conservation equation. Then, it should be clear that we can formulate the following:

$$m_1 + n_3 = m_2 + n_4$$

Where m and n are the number of vehicle trajectories crossing the borders of the control area in time – space. Subscripts refer to the particular borders considered.

We can rearrange the terms according to if they belong to a border of the region in U or D. This is:

$$m_1 - n_4 = m_2 - n_3$$

Then, express the number of trajectories in terms of flows and densities:

$$q^U T - k^U L = q^D T - k^D L$$

$$q^U - k^U \frac{L}{T} = q^D - k^D \frac{L}{T}$$

And note from Figure 21 that, by definition, $u = L/T$, so that:

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$q^U - k^U u = q^D - k^D u$$

And finally:

$$u = \frac{q^U - q^D}{k^U - k^D} = \frac{\Delta q}{\Delta k}$$

The conclusion is that the speed of the shockwave between two stationary traffic states can be determined by the increase in the flow over the increase in the density of the colliding traffic states. Note that u could be positive (i.e. in the same direction of traffic) or negative (i.e. against traffic), depending on the properties of the colliding traffic states. This solution is of great importance as it allows to predict the evolution of the interphases between different traffic states. Note for instance that if U was a free-flowing traffic state and D a congested one, the interphase between them would be the end of the queue, and the previous expression would allow to predict the evolution of the end of the queue (i.e. the queue extension).

6.2. The fundamental diagram of traffic

Different traffic analysis disciplines prefer to use different traffic diagrams in their studies. For instance, traffic engineers dealing with infrastructural capacity and level of service prefer the $q - v$ diagram, because this is the one appearing in the Highway Capacity Manual. Transportation planners use the $travel\ time - q$ diagram, usually in the form of the BPR function[2]. Traffic researchers working on microscopic modeling of traffic often use the spacing – speed diagram (s-v) because of its direct implications on car-following, and academics working with macroscopic traffic flow modelling prefer the density – flow diagram (k-q), because it results particularly easy to derive some relevant properties of traffic dynamics by using this graphical representation. Possibly, amongst all the possible diagrams, k-q is the most widely used and the one that reveals more information by simple inspection. This is why this diagram is called the Fundamental Diagram of traffic.

[2] The Bureau of Public Roads (BPR) function is a classical relationship between the travel time (tt) in a network link as a function of its travelling flow (q). The traditional expression of the BPR function is:

$$tt = tt_f \left[1 + \beta \left(\frac{q}{q_{max}}\right)^n\right]$$

Where tt_f is the free flow travel time and q_{max} the capacity of the link. β and n are two calibration parameters, whose classical values are $\beta = 0.15$ and $n = 4$, but today often varying by link type. Note that the BPR function implies that tt is continuously increasing with q. While this is valid for free-flowing traffic, it does not represent congested conditions, where tt increase for decreasing q (see the p - q diagram in Figure 14). So, the BPR function should be applied with much caution in oversaturated networks.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

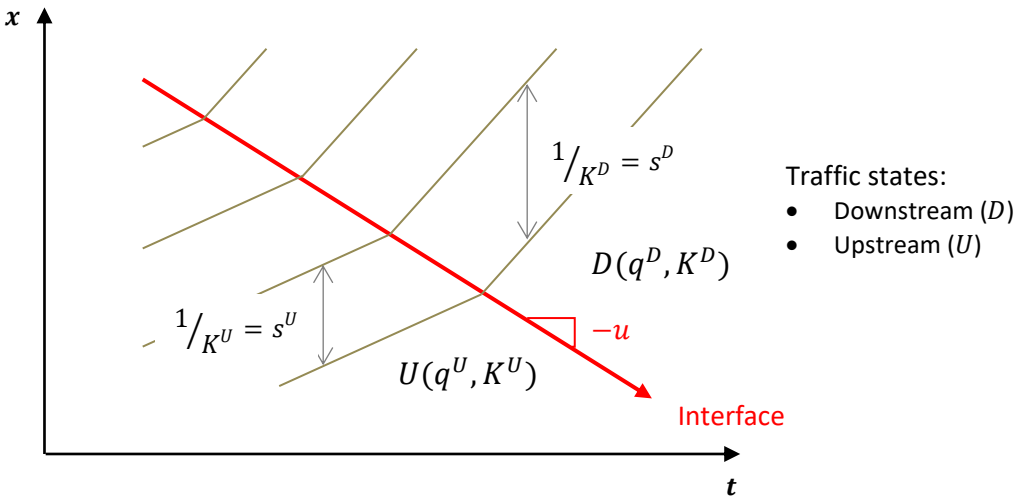


Figure 22. Representation of a traffic wave with negative speed (i.e. against traffic direction).

In particular, on the k-q diagram, it results particularly easy to visualize the speed of a traffic wave between two different traffic states. Take for instance the traffic states U and D shown in Figure 22. Recall that the speed, u, of the traffic wave determining the transition between them is obtained as $\Delta q/\Delta k$, which in the k-q plane represents the slope of the straight line joining traffic states U and D (see Figure 23). Note that all "slopes" on the k-q plane have units of speed, since according to the fundamental equation of traffic, $v = q/k$.

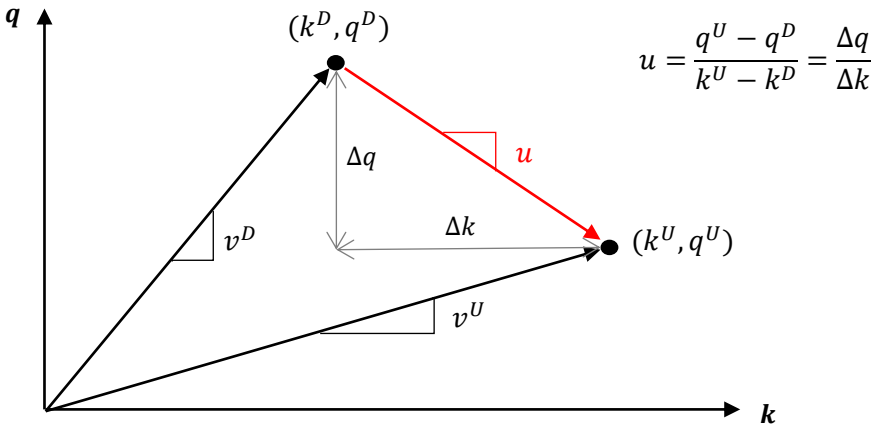


Figure 23. Traffic wave on a flow (q) – density (k) diagram.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Then, in the density – flow diagram represented in Figure 24 we can determine:

- The infrastructure capacity, q_{max}, and the related optimal density, k_0.
- The jam density, k_j.
- The free flow speed, v_f, as the slope of the tangent to the diagram through the origin.
- The speed, u, of the traffic wave between any two traffic states (e.g. U and D) as the slope of the line joining both traffic states.

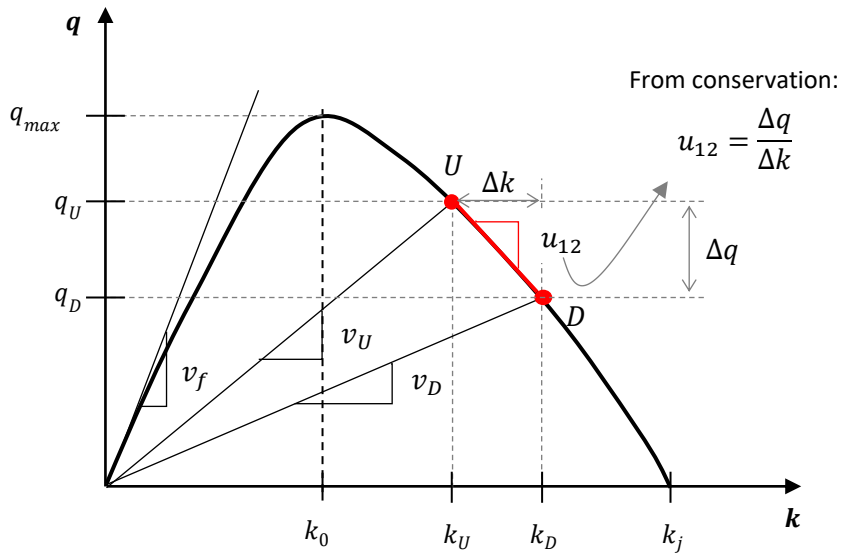


Figure 24. Density – flow diagram: the fundamental diagram

Also, the relative flow, q_0, seen by a moving observer travelling at speed v_0 has a direct representation on the fundamental diagram (see Figure 25). It is just needed to draw an auxiliary straight line with slope v_0 through the origin. Then, for any prevailing traffic state A, the relative flow seen by the observer is identified in the diagram as the vertical separation between the auxiliary line and the diagram. Note that this construction would also hold for traffic states B, where $v_B < v_0$, and that in such case the relative flow is negative, as the distance between the auxiliary line and the diagram follows the opposite direction of the coordinate axis for q.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

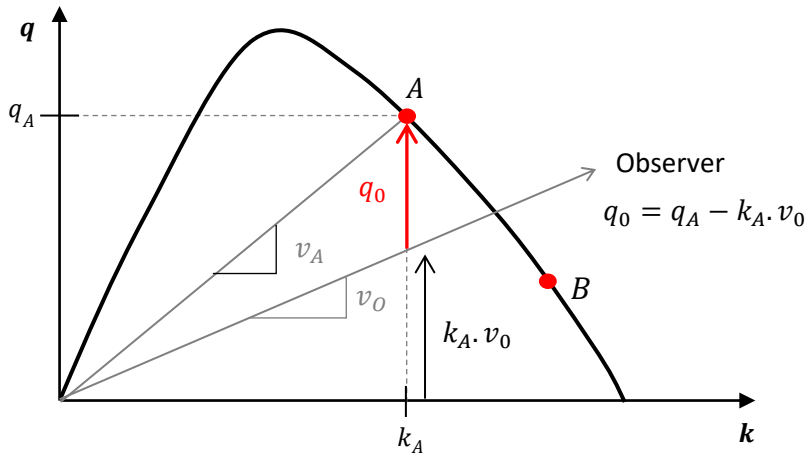


Figure 25. Relative flow seen by a moving observer on a fundamental diagram.

6.3. Shocks and waves

The concept behind LWR traffic flow theory is simple: at every change in the traffic state a traffic wave is generated. Then, it is only needed to track these waves in time and space to predict the traffic evolution. This is the main concept, although its application results more complex as we are going to discuss next.

Only two possible situations can arise in a traffic state change: *i)* an acceleration transition (i.e. vehicles moving to a faster traffic state), or *ii)* a deceleration transition (i.e. vehicles moving to a slower traffic state; joining a queue for instance). First, consider case *i)* and assume a group of vehicles travelling behind a slow truck (i.e. a moving obstruction) until t_1, when the truck exits the road (see Figure 26). Just after t_1, the vehicle immediately after the truck will start accelerating to recover his free-flow speed. In his acceleration, this vehicle will travel at all the intermediate speeds between the speed of the truck, v_B (i.e. the one at the start of the acceleration) and the free-flow speed, v_H (i.e. the one at the end of the acceleration process). In every differential increase of speed, the traffic state changes, and a traffic wave is generated to inform the following vehicles of this transition. When these waves sequentially hit the following vehicles, they start accelerating and modifying their speeds accordingly. Note that the speed of the traffic wave generated by a differential increase in the speed while in state B, is determined by the slope of the line tangent to the diagram at B (see Figure 27). Given the concave shape of the fundamental diagram, it is not difficult to visualize that the infinite number of traffic waves generated by the leading vehicle when accelerating from v_B to v_H always grow in speed (i.e. from very negative, to zero at capacity, and turning positive afterwards), so that, when represented in the x - t plane they create a "fan" of waves. Keep in mind that waves arise from bottlenecks or any other type of disturbance. This is represented in Figure 26 and Figure 27 showing only a selection of the infinite traffic states generated (i.e. traffic states C, D, E, F and G). You can see in Figure 26 how the acceleration process is sharp for the first vehicles after the truck and it is more and more smooth for vehicles further behind Probably you have experienced this phenomenon while driving in a similar context.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

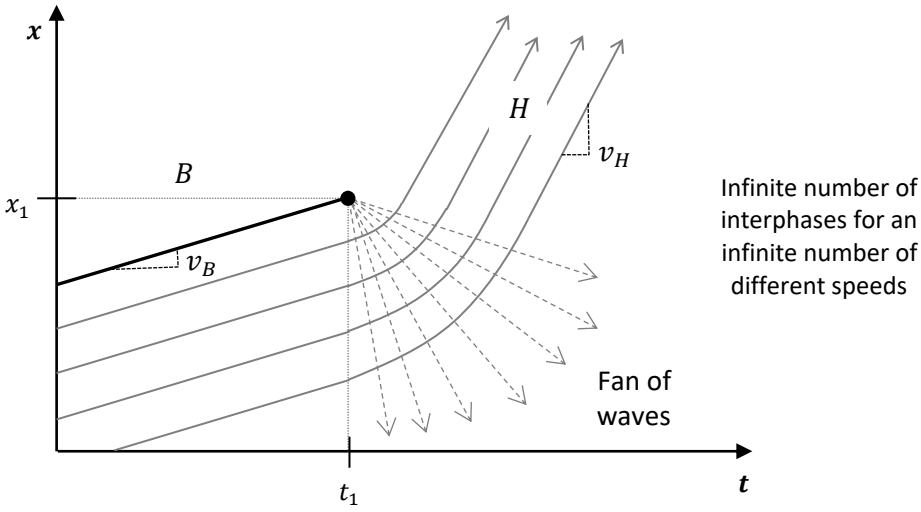


Figure 26. Fan of traffic waves in an acceleration process.

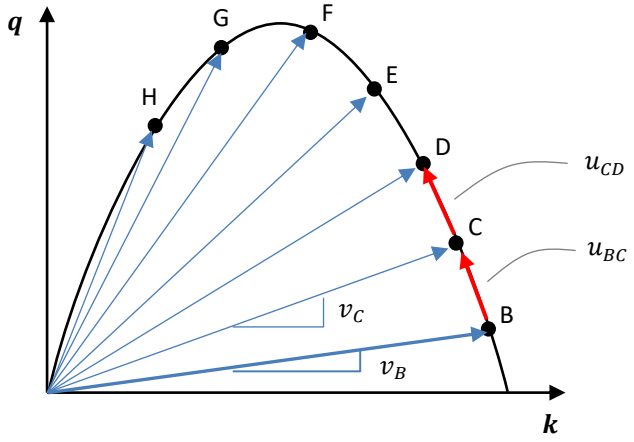


Figure 27. Traffic states in an acceleration process.

The trajectories of traffic waves on the x - t plane define the regions where the different traffic states prevail. In particular, we know the average speed that vehicles will experience at any point in x - t. This accomplishes our traffic flow modeling objective, because we can predict the trajectory of any vehicle given its initial position in the x - t plane (see Figure 28).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

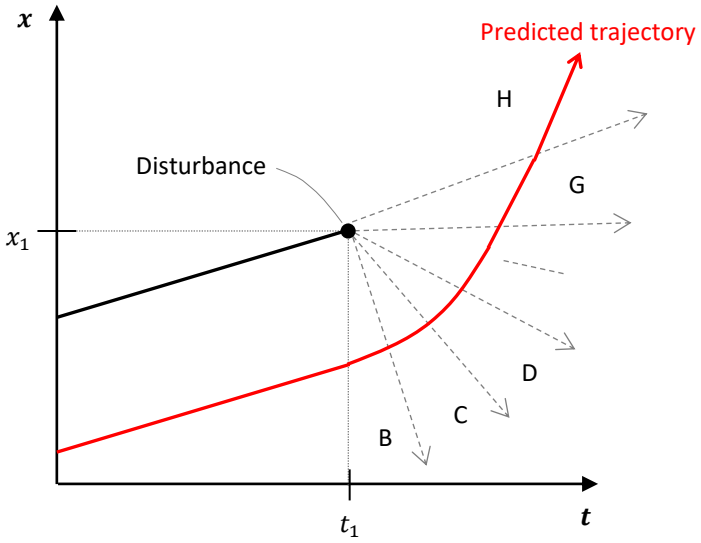


Figure 28. Solution from a map of traffic waves (accelerating).

The second possible type of traffic transition is a deceleration process. Just consider a situation where vehicles need to slow down when joining a queue, from free-flow, v_A, to the slow speed in the queue, v_F. As in the acceleration case, an infinite number of traffic waves are generated at every differential reduction of the travelling speed. The main difference in this case is that the speeds of the generated waves decrease along the deceleration process. This can be seen by looking at the slopes of the tangents to the fundamental diagram when moving from traffic state A to F (see Figure 29). When representing these traffic waves on the x - t plane, they do not fan out as in the acceleration case, but they tend to concentrate and collide (see Figure 30). When two traffic waves collide, the traffic state between them vanishes and a new wave is generated splitting, now, two very different traffic states, because the existing smooth transition between them has vanished (see Figure 31). This new strong wave which results from the collision of two "soft" waves receives the name of a "shock", or more specifically of a "shockwave". So, a shockwave is a strong traffic wave which determines a change of traffic state not passing for all the neighboring states. Still, the speed of a shockwave is determined by $u = \Delta q/\Delta k$, the difference in flow with respect to the difference in density of the colliding traffic states. By looking at Figure 31 you can realize that traffic waves and shockwaves will keep colliding until all the intermediate states vanish. Finally, only one shockwave remains, u_{AF}, splitting traffic states A and F. This means that, while first drivers may experience a soft deceleration process, later ones will find a sharp change of speed when joining a queue. Again, this reflects what we experience in real traffic.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

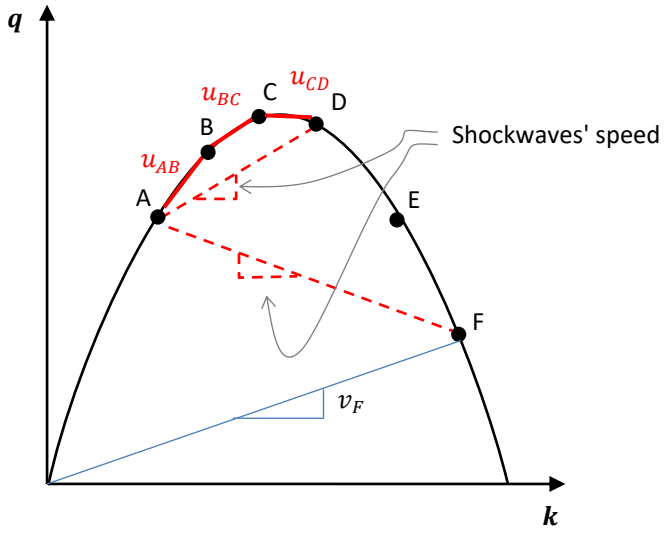


Figure 29. Traffic states in a deceleration process.

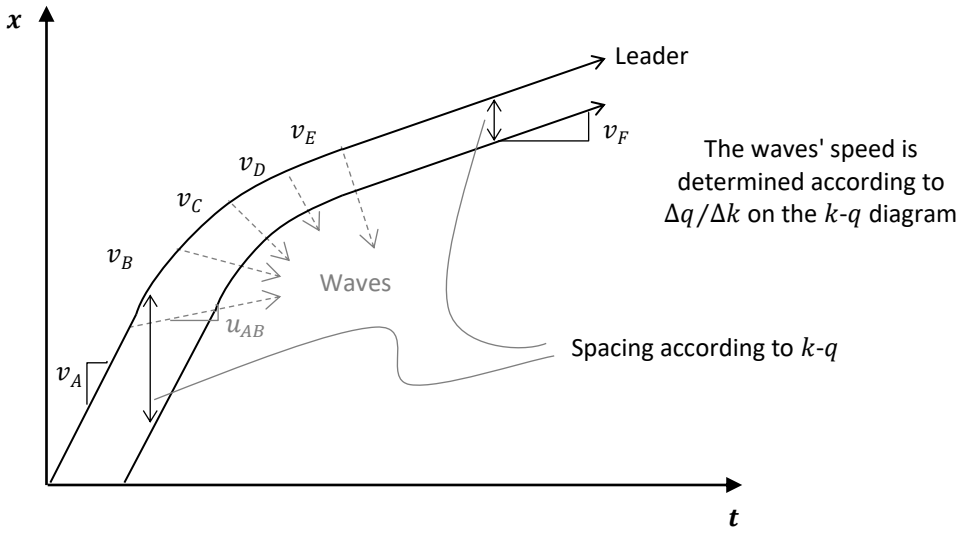


Figure 30. Solution from a map of traffic waves (decelerating).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

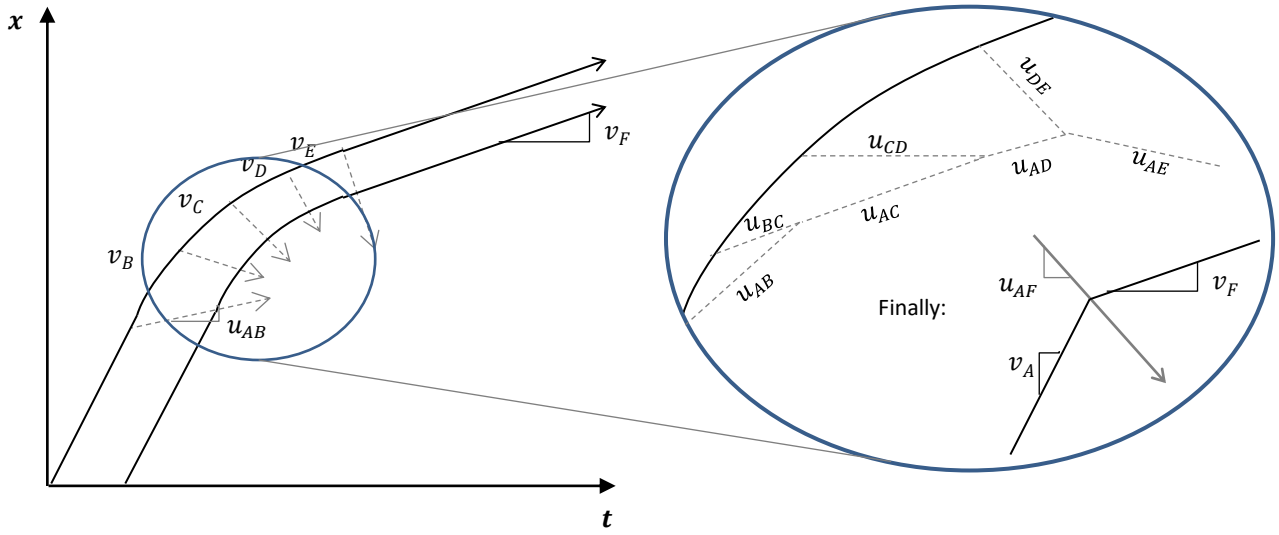


Figure 31. Shockwave propagation.

6.4. Simplifications

It is clear that the LWR theory as it is, with infinite number of shocks and waves, is not convenient for being applied in real practice. To solve these inconveniences two assumptions (or simplifications) are made. These are:

- Consider acceleration / deceleration processes as instantaneous
- Assume a triangular fundamental diagram

Instantaneous accelerations

Assuming instantaneous acceleration / deceleration processes allows simplifying the application of LWR theory, especially in the deceleration processes where an infinite number of waves collide and create new shocks. Note that if the deceleration process is neglected, and therefore vehicles' trajectories are piecewise linear (as in Figure 32), then only one shock is created (i.e. the shock between the initial and the final traffic states). This means that the smooth deceleration transition, experienced by some of the initial drivers after the disturbance, is neglected. Actually, this does not imply a severe drawback, as this transition between multiple traffic states take place only on a very limited x - t region. The evolution of the traffic states that persist in time (e.g. tracking the evolution of queues) is not affected much by this simplification, which in contrast simplifies a lot the application of the theory.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

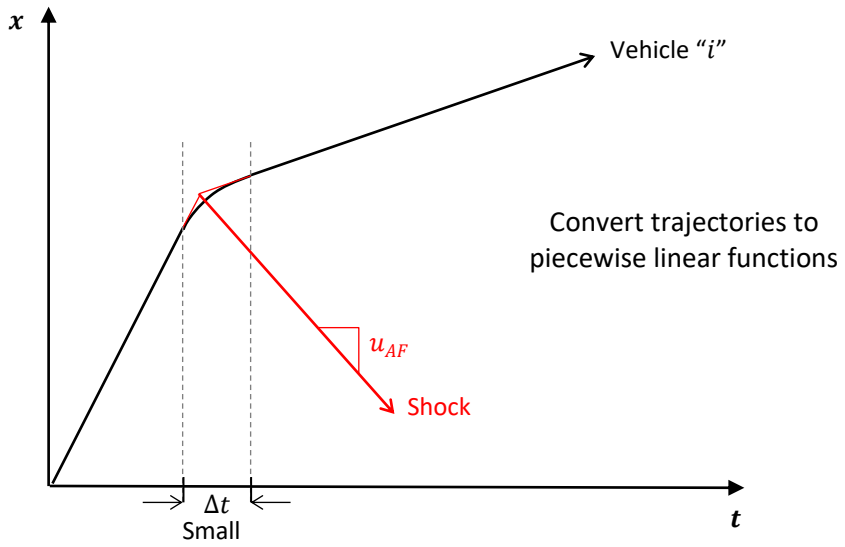


Figure 32. Assumption of piecewise linear trajectories.

Triangular fundamental diagram

The assumption of instantaneous accelerations simplifies a lot the application of LWR theory, especially for decelerations. However, the same solution is not valid for accelerations, because waves do not collide, but fan out covering a vast region in x - t. It would not be realistic to simplify this growing fan into a single shock.

In the mid 1990's, prof. Gordon F. Newell at the University of California, Berkeley proposed to use a triangular diagram in the application of the LWR theory. Newell justified that the triangular diagram is not only easier to work with, but actually it is more realistic to reproduce the behavior of today's real highway traffic. In fact, diagrams that are calibrated today with many available data points, resemble the triangular shape, possibly with the only exception of not having such a sharp vertex at capacity, which typically is smoother.

The benefit of using a triangular fundamental diagram is that between any pair of congested traffic states, the wave speed is the same, w (see Figure 33 and Figure 34). w is a parameter of the triangular fundamental diagram named as the "characteristic wave speed" between congested traffic states. The same happens between any two free-flowing traffic states. The free-flowing branch of the diagram is linear (and with slope v_f), meaning that any traffic state on this branch will travel at the free-flow speed v_f, and actually, any wave between free-flowing traffic states will also travel at v_f. This implies that free-flowing waves do not cross the trajectory of any vehicle, but travel with them (see Figure 35).

In conclusion, with the simplification of instantaneous accelerations and with the assumption of a triangular fundamental diagram, the application of the LWR theory with pen and paper to simple traffic flow problems is straightforward, as we are going to see in the next example.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

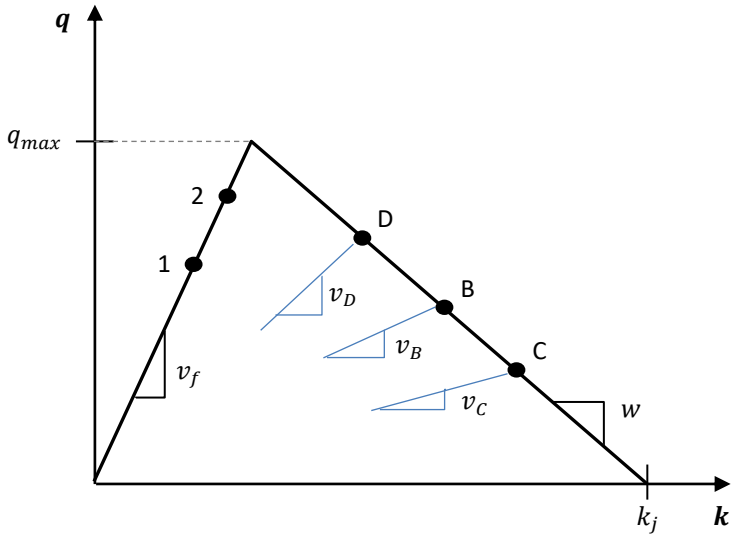


Figure 33. Triangular fundamental diagram.

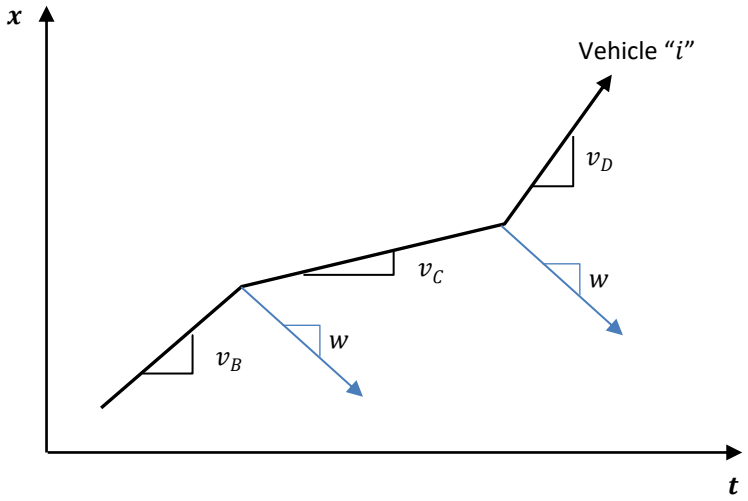


Figure 34. Shockwaves between three different congested traffic states (B, C and D) assuming a triangular fundamental diagram.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

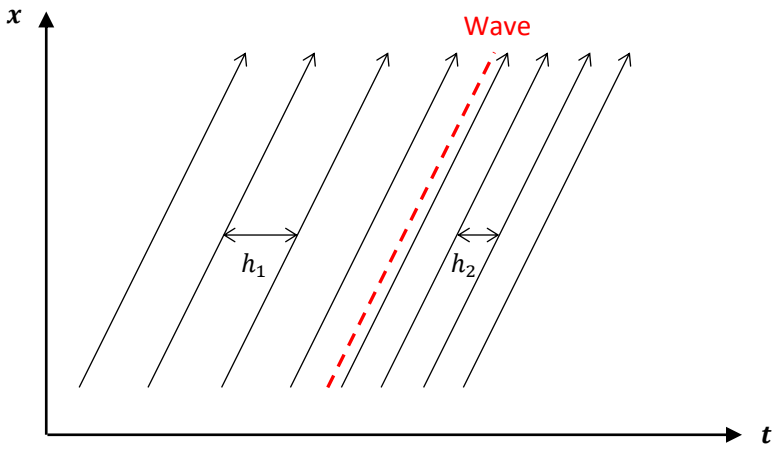


Figure 35. Shockwave between two free flowing traffic states (1 and 2) assuming a triangular fundamental diagram.

7. Example of application of the LWR theory: Incident on a freeway

A traffic flow theory problem is posed by knowing the initial conditions (i.e. the initial traffic states), the contour conditions (i.e. the traffic demand during the analysis period), and the fundamental diagrams for all the sections of the infrastructure under analysis. Given these inputs, the objective is to predict the evolution of shockwaves and traffic states in in x - t, paying special attention to the evolution and physical extension of queues. The procedure of solving a traffic problem with pen and paper using the LWR model is illustrated with an example in this section.

Consider a $q_A = 6000$ veh/h traffic flow travelling on a 3 lane freeway. At time $t = 0$ this traffic flow is disrupted by a vehicle breakdown at the location $x = 0$. This incident blocks one lane during 30 min. (i.e. from $t_0 = 0$ until $t_1 = 30\ min$). Figure 37 represents the flow (q) – density (k) diagram for one lane of the freeway.

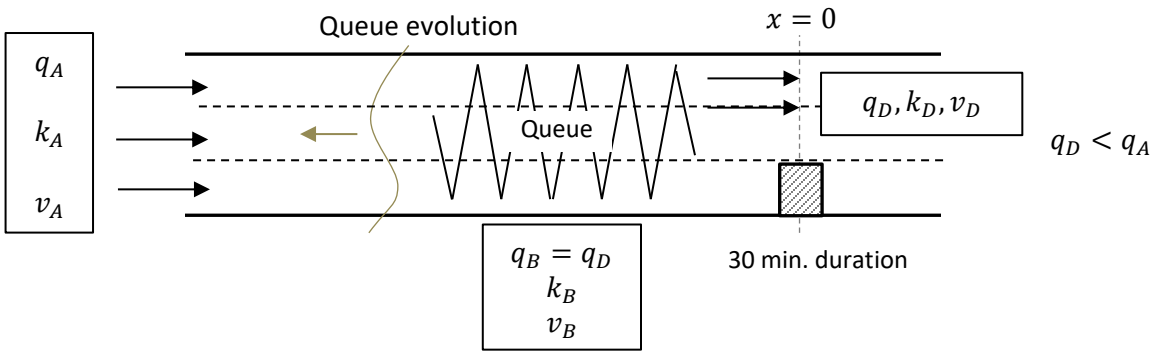


Figure 36. Incident on a three lane freeway.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

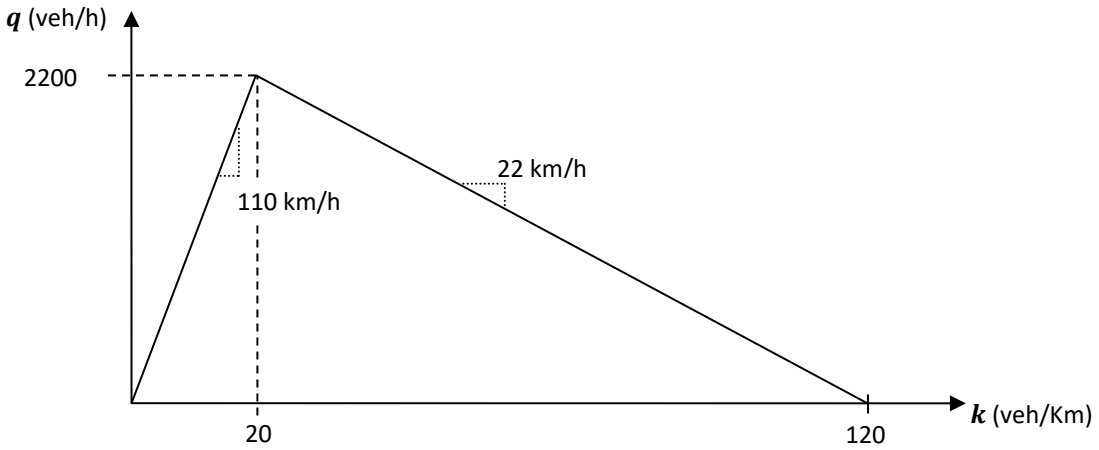


Figure 37. One lane fundamental diagram

And the questions to answer are:

- Does this incident generate queues? Why?
- How far does the queue reach?
- When does the queue dissipate?

In order to answer these questions, the first thing to do is to obtain the fundamental diagrams for all the sections in the problem. Note that we have a 3-lane freeway, except at the location of the incident that we have 2-lanes. So, we need the diagrams for the 3-lane and for the 2-lane sections. From the 1-lane fundamental diagram provided, we can construct the other diagrams assuming that the diagram for a n-lane section exhibits the same speed for n times the density. This is a common assumption that allows us to draw the three and the two-lane sections' diagram on the same plot (see Figure 38).

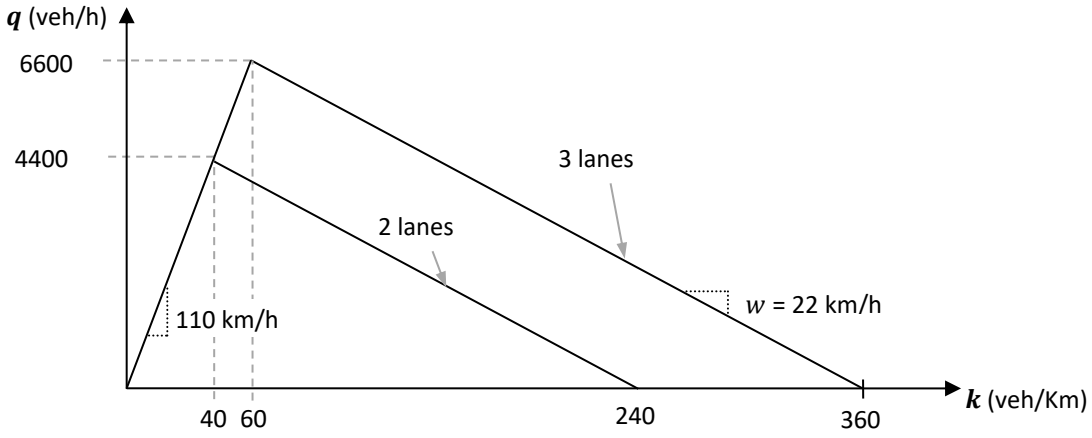


Figure 38. Two and three lane fundamental diagrams

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Note that the problem considers triangular fundamental diagrams. As commented, this is the common simplification in the application of LWR theory, initially proposed by prof. G.F. Newell in the 1990's, that not only simplifies the problem, but also triangular diagrams are in general more accurate that the previously used parabolic shapes.

The solution to the problem consists in predicting the evolution of traffic states in (x, t). It is recommended to locate the $x = 0$ reference at a relevant point in the problem, which allows to visualize upstream and downstream traffic (i.e. somewhere in the middle of the (x, t) diagram, as in Figure 39).

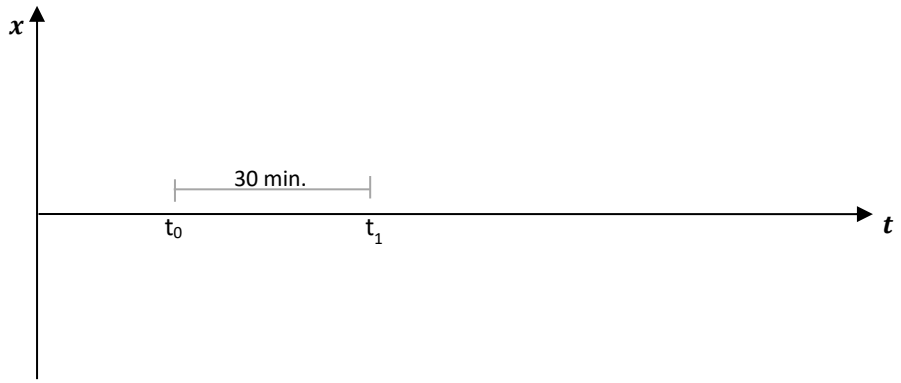


Figure 39. x - t diagram where to draw the solution to the problem.

Next step consists in identifying the initial and contour conditions. This is critical, as it will determine the solution. The identification of the intervening traffic states can be done on the fundamental diagrams, as in Figure 40.

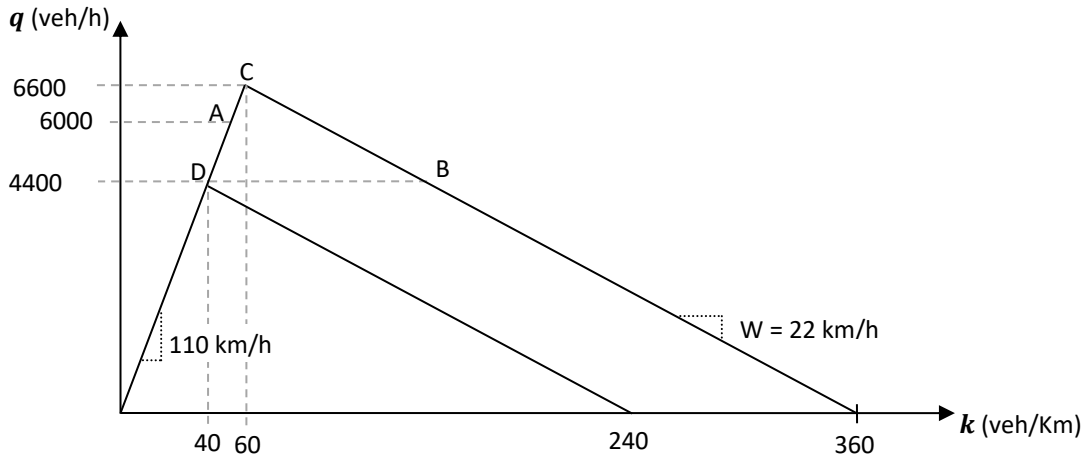


Figure 40. Traffic states identification on the fundamental diagram.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Traffic state A is the traffic demand. This applies before and after the incident for the whole freeway section. Note that the flow of state A is $q_A = 6000$ veh/h, larger than the capacity of 2 lanes, of 4400 veh/h. This implies that the incident will create a queue. The bottleneck at the incident location will discharge the maximum possible flow (i.e. the capacity of 2-lanes, 4400 veh/h). Therefore, the queue that will appear in the 3-lane freeway upstream of the 2-lane bottleneck will be represented by state B in the previous figure. This is a congested traffic state with the flow equal to $q_B = 4400$ veh/h (i.e. same as the capacity of 2-lane bottleneck). Downstream of the bottleneck, the flow still needs to be the same (i.e. conservation), but traffic will be free-flowing on a three lane section. This is represented by state D, with $q_D = 4400$ veh/h. Finally, one additional traffic state needs to be considered. This is the discharge flow when the full capacity of the 3-lane is recovered (i.e. at $t = 30min$). Queues always discharge at the full available capacity, in this case 6600 veh/h. This is represented by traffic state C. Once the intervening traffic states have been identified, they can be located on the previous (x, t) diagram (see Figure 41).

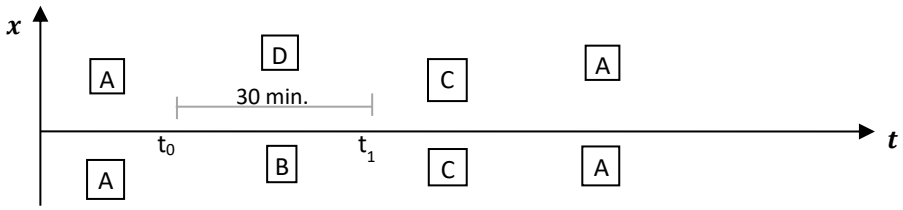
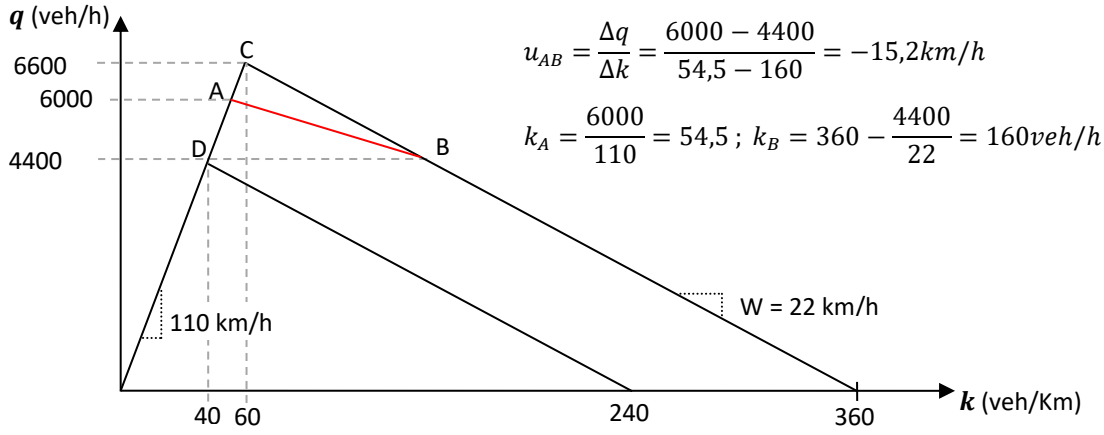


Figure 41. Traffic states location on the x - t diagram.

Finally, it is necessary to identify the shockwaves delimiting each traffic state, which will allow to determine the traffic evolution in the (x, t) diagram. This is done with the help of the fundamental diagram and the graphical derivation of the speeds of the shockwaves (see Figure 42).



$$u_{AB} = \frac{\Delta q}{\Delta k} = \frac{6000 - 4400}{54{,}5 - 160} = -15{,}2 km/h$$

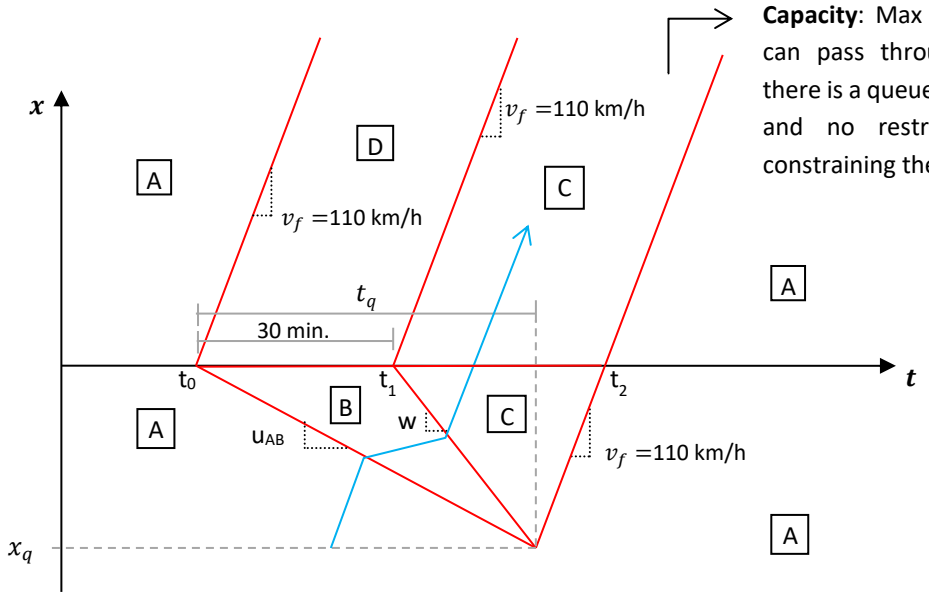
$$k_A = \frac{6000}{110} = 54{,}5 \; ; \; k_B = 360 - \frac{4400}{22} = 160 veh/h$$

Figure 42. Shockwave identification.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >



Capacity: Max sustained flow that can pass through the link when there is a queue upstream feeding it and no restriction downstream constraining the flow.

Figure 43. Solution to the traffic problem on the x - t diagram.

And this is the final solution of the problem, as it allows to predict the evolution of the queued state B. Note that the queue grows against traffic at a speed u_{AB} (i.e. creation shockwave) and it dissolves when the incident is cleared and also from the incident location and against traffic at a speed $u_{BC} = w$ (i.e. dissolving shockwave). When these to waves collide, they determine the end of the queued state B. So, you can observe that the queue has reached the location x_q upstream of the incident and lasts a duration t_q since the start of the blockage. Also note that the queue does not discharge completely until t_2, when the effects of the incident at x_0 finally end. If you wish, you could predict the average trajectory of any vehicle, as you know the prevailing speeds in each traffic state zone. One possible trajectory in blue is shown as an example.

Below, you can find the necessary calculations to determine these times and locations.

$$u_{AB} \cdot t_q = w \cdot (t_q - 0{,}5)$$

$$15.2 \cdot t_q = 22(t_q - 0{,}5)$$

$$15.2 \cdot t_q - 22 \cdot t_q = -11$$

$$t_q = 1.6h$$

$$x_q = 15.2 \cdot 1.6 = 24.5km$$

$$t_0 = 0 \qquad t_1 = 0.5h \qquad t_q = 1.6h \qquad t_2 - t = \frac{24.5}{110} = 0.22 \rightarrow t_2 = 1.8h$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

8. Limitations of the LWR macroscopic traffic flow theory

The LWR macroscopic traffic flow theory is a powerful and very robust theory to model the evolution of traffic states. Note that its calibration requirements are minimal, as it is only necessary to calibrate the 3 parameters of the triangular fundamental diagram. The information provided by the model and its level of accuracy is admirable given its simplicity and few input parameters.

In spite of this, the LWR model has its limitations, which are going to be described next:

1. The LWR model does not consider differences between vehicles. All the vehicles are assumed to have the same properties (e.g. same free-flow speed). Density, for instance, is an absolute variable and it is not stratified by vehicle types with different free-flow speeds. This has some implications in the results, and the theory does not predict, for instance, that when traffic is opened after a road closure (or red signal) downstream of the restriction first are seen the faster vehicles. Also, LWR theory cannot predict overtakings or relative flows. This is a common drawback of macroscopic models, although some attempts have been done to develop macroscopic traffic flow models with different vehicle classes (e.g. the slugs and rabbits model by prof. Carlos F. Daganzo). In summary, LWR theory is a bad model for analyzing light traffic when differential speeds play a role. In contrast, it is good for dense and congested traffic states, when the speeds of all vehicles are approximately the same.

2. Theory neglects "waves" and only works with "shocks". This is the price to pay for simplifying the model by neglecting acceleration and deceleration processes. The implications of this simplification in the results are that shockwaves should not be interpreted as precise trajectories. They should be interpreted as time-space zones where traffic changes state (i.e. a shock wave should not be interpreted as a thin line, but as a wide line or region). You cannot predict the evolution of a queue to the precision of a few meters, but only hundreds of meters. Figure 32 should be interpreted as in Figure 44, where the shock defines a transition zone.

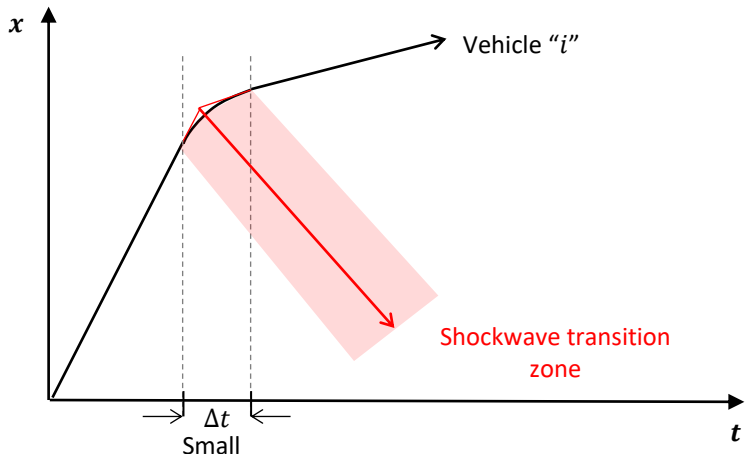


Figure 44. Shockwave interpretation as transition zones.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

3. Traffic is always stable according to LWR. The continuous theory of traffic does not predict stop&go behavior inside congested traffic states. It just predicts the average speed for the congested state, as given by the fundamental diagram. However, in reality this average speed results from periods when the vehicle is stopped and periods when the vehicle travels faster. This unstable behavior is not reproduced in the LWR model. This implies that it is a bad model to analyze anything which is significantly affected by stops and accelerations. For instance, pollution emissions of traffic depend largely on stops and accelerations. If you use LWR to estimate emissions, these would be largely underestimated, because constant speeds, which much smaller emissions, are assumed inside queues instead of stop&go.

In light of these limitations the overall conclusion could be that:

- Macroscopic modeling, and specifically LWR model, is a perfect and robust tool to predict the evolution of congested traffic states to the precision of hundreds of meters.
- However, it is a bad model to be used in light traffic, or to predict the details of what happens inside of the queues (i.e. stop&go).

These limitations give rise to microscopic traffic flow models, were the objective is to predict the trajectories of individual vehicles, with all their details.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

5-MICROSCOPIC TRAFFIC FLOW MODELING

Table of Contents

1. Introduction to microscopic traffic modelling 2
2. Car-following definitions 4
3. Pipes (1953) car-following model 5
4. Forbes (1958) car-following model 6
5. General Motors (1958-1961) car-following models 6
   i. 1st Generation: 6
   ii. 2nd Generation: 7
   iii. 3rd Generation: 7
   iv. 4th Generation: 8
   v. 5th Generation: 8
6. Traffic Stability 9
   i. Local stability of GM car-following models 10
   ii. Asymptotic stability of GM car-following models: 11

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

1. Introduction to microscopic traffic modelling

The vehicle is the main entity in microscopic traffic flow modelling. The objective is to predict the trajectories of individual vehicles in time and space. The movement of the vehicle has two main components: i) the longitudinal movement along the guideway (i.e. the lane), and ii) the lateral movement to change lane. Typically, different models are used to reproduce these movements, being the longitudinal movement of vehicles much more relevant in terms of modelling the traffic stream evolution. In this introductory chapter to microscopic traffic flow modelling we are going to focus on the longitudinal movement of vehicles, not addressing lane-changing models.

The longitudinal movement of vehicles is radically different depending on the prevailing traffic conditions. In light traffic (i.e. with low vehicular densities, nearly free passing, and vehicles travel at free-flow speed, v_f, and where a small increase in density does not affect the average travelling speed; e.g. flow of few hundreds [veh/h/lane] in freeways), vehicles' trajectories are modelled according to their desired free-flow speed. Different vehicle types may exist, with different average v_f's. In turn, each vehicle i has its own particular v_{f_i}, obtained as a realization of some probability distribution whose average value is v_f for the vehicle type. In addition, the appearance of vehicles as inputs to the traffic stream also follows a probability distribution (i.e. typically a Poisson distribution) whose average success rate is the average demand (i.e. the average flow). Light traffic modelling, although interesting and necessary in the modelling of traffic streams, does not apply when traffic analysis is most relevant, like in dense and congested traffic conditions with queues. This is why we are neither going to discuss in detail light traffic models.

The longitudinal movement of vehicles in dense heavy traffic is characterized by the so-called car-following models. This represents how one vehicle (i.e. the follower) follows the vehicle in front (i.e. the leader). In dense traffic, the leader, at the same time is the follower of another leader in front. This means that with a car-following model we can reproduce the behavior of a string of vehicles (i.e. a traffic stream). Different car-following models exist, represented by several analytical specifications. In spite of different specifications, they can always be expressed as an average macroscopic relationship, which can be visualized in terms of the average spacing - flow diagram (see Figure 1).

The vehicular spacing is of great importance in car-following models (and in traffic in general) because it affects two factors:

- Traffic safety
- Infrastructure capacity

Regarding traffic safety, vehicular spacing needs to ensure that drivers can react (i.e. there is a drivers' reaction time) and adapt to the new speeds of vehicles in front without colliding. In terms of traffic safety, the large the spacing the better. In contrast, assuming a constant speed, the smaller the spacing the larger the flow. Recall the fundamental equation of traffic: $q = v/s$. So, in terms of infrastructure capacity, it would be desirable that vehicles circulate at high speeds with small spacings. Clearly, there is a trade-off between safety and capacity.

It is the drivers' perception of risk what manages this trade-off, by reducing the travelling speed when spacing results insufficient to guarantee safety (i.e. this is the definition of heavy traffic: when density is high, the average speed of traffic is reduced with further density increases). In terms of infrastructural capacity, the increase in density which could contribute to larger flows is compensated by a reduction of the speed. While initially this density increase can still represent larger flows (i.e. up to capacity), as traffic density grows, the effect of the

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

reduced speeds is predominant, leading to the large flow reductions characteristic of congested traffic, as seen in Figure 1.

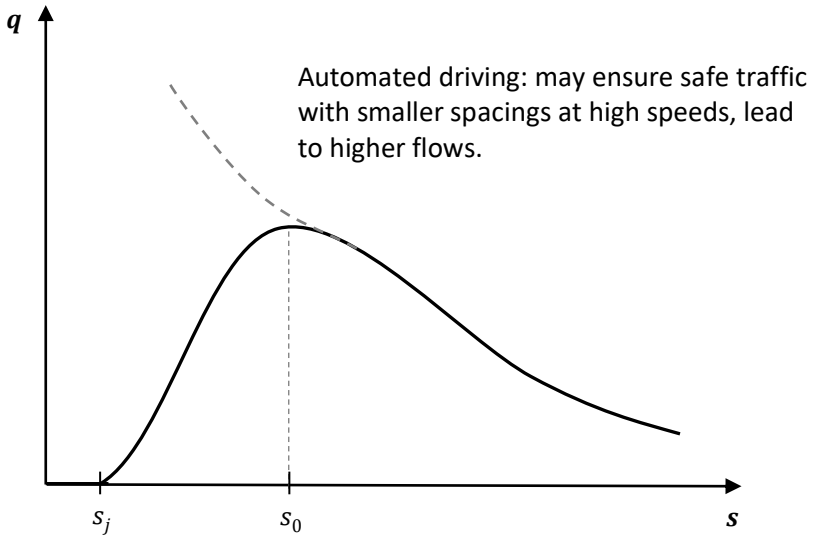


Figure 1. Average spacing - flow diagram

This trade-off between safety and efficiency may be modified with the introduction of automated driving. The benefits of automated driving in order to increase capacity without penalizing safety are:

- Very short reaction times => can be of 1ms with respect to 1s in human driving.
- Extended communication range => vehicles can react altogether almost at the same time. It might take 1ms for a string of "n" vehicles to start mutually braking. This means that traffic shockwaves travel extremely fast. Note that it would have taken "n" seconds to vehicle "n" start braking in human driving.

This means that in the context of automated driving, vehicles can travel with very short spacings at high speeds without compromising safety. This is the concept of automated vehicle platooning. The concept is not new, and already back in 1997, there was an automated vehicle platoon demo on I-15 San Diego (PATH, UC Berkeley) https://www.youtube.com/watch?v=h7tO-4FoKCo. This was a pioneer experiment with old technology but very illustrative. Unfortunately, latest experiments on automated driving seem to have lost the traffic efficiency perspective, focusing only on safety and convenience.

Concluding this introductory section, it is worth knowing that despite the analytical formulation of car-following models dates back from 1950's (i.e. at the same time as macroscopic traffic flow models), their generalized used needed to wait until the popularization of personal computers. Today, with powerful computing capabilities at our desktops, traffic microscopic models are integrated into traffic microsimulators (like this: https://www.youtube.com/watch?v=UWWPUQUrn1U (a microscopic pedestrian simulation example) or this https://www.youtube.com/watch?v=OtYby7QnyAE). Obviously you need many other components in addition to a car-following model to develop a microsimulator (e.g. lane-changing model, controllers behavior, O/D matrixes...) plus an advanced digital graphics animation. You need to know that calibrating these models requires

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

many parameters (e.g. of the order of 30 maybe). This is 10 times more than the 3 parameters we required to calibrate the fundamental diagram in the LWR theory. This means that microscopic models are much less robust with respect to LWR, and all results should be validated with the later.

In the next sections, our analysis of microscopic traffic flow modeling will be restricted to the most basic car-following models.

2. Car-following definitions

Consider the leader vehicle, n, and its follower, $n+1$. The main variables and definitions necessary to formulate a car-following model are shown in Figure 2, where: (subscripts refer to the leader or follower vehicles).

- s is the vehicular spacing (i.e. the distance between the same point of consecutive vehicles; Figure 2 takes the rear bumper as the reference point).
- l is the vehicle length.
- g is the gap (i.e. the empty distance between consecutive vehicles).
- h is the headway (i.e. the time between the passage of the same point of consecutive vehicles; Figure 2 takes the rear bumper as the reference point).
- $\dot{x}$ is the vehicular speed.
- $\ddot{x}$ is the vehicular acceleration.

All the previous variables depend on time, t, except the vehicular length, l.

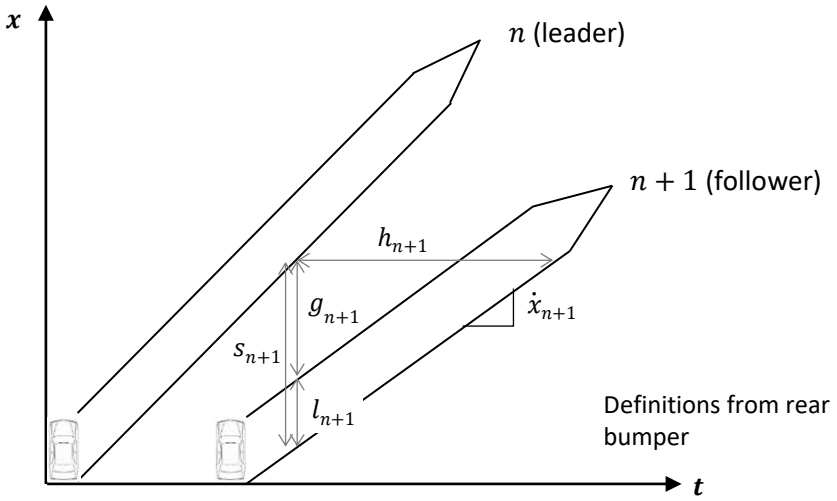


Figure 2. Car following definitions.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

From the previous definitions, recall that:

$$s_{n+1}(t) = l_{n+1} + g_{n+1}(t)$$
$$s_{n+1}(t) = h_{n+1}(t) \cdot \dot{x}_{n+1}(t)$$

3. Pipes (1953) car-following model

It was amongst the first car-following models proposed. It is based on the California driving code, which states that drivers should leave a gap of one vehicle length for every 10mph of travelling speed. This is analytically formulated as:

$$s_{n+1}(t) = l_{n+1} + \frac{\dot{x}_{n+1}(t)}{10mph} l_{n+1}$$

According to this model, the minimum safe distance headway increases linearly with speed. So, the Pipes car-following model represents a linear $s(v)$ model, as shown in Figure 3.

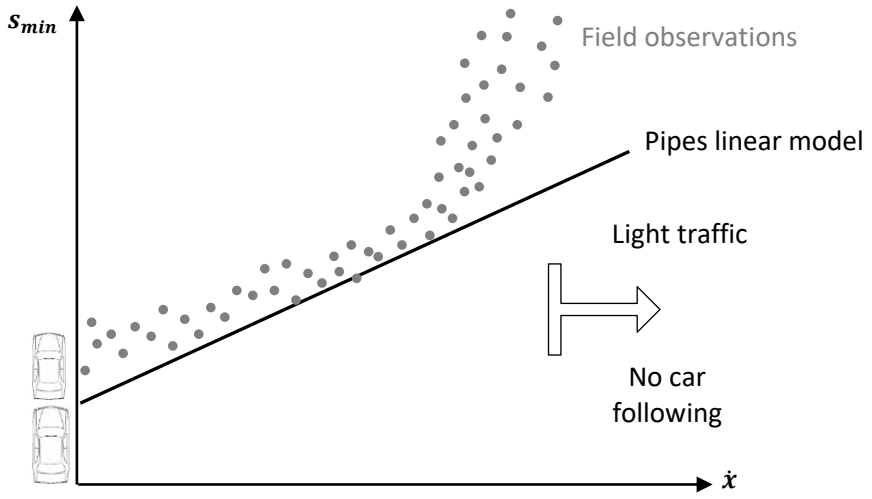


Figure 3. Pipes (1953) car following model.

Despite its simplicity it works reasonably well. However, note two things from Figure 3:

- For low speeds is not accurate and it yields spacings considerably lower than the corresponding field measurements. Note that would predict a zero gap when traffic is completely stopped in a jam.
- Car-following only holds for heavy traffic. In light traffic vehicles do not follow each other, they simply travel at a comfortable desired speed. This is not specific to Pipes car-following, but to all car-following models.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

4. Forbes (1958) car-following model

The main conceptual difference in the Forbes model with respect to that of Pipes was the introduction of the reaction time, Δt. The reaction time is defined as the time needed by a human driver between the stimulus and the reaction (e.g. between realizing that is needed to brake and actually braking). The reaction time varies across drivers, from slightly less than 1s and up to 2s or more.

In order to avoid collision, the time gap between vehicles needs to be at least the reaction time. This is what is imposed in the Forbes car-following model:

$$h_{n+1}(t) = \Delta t + \frac{l_{n+1}}{\dot{x}_{n+1}(t)}$$

Or what is the same:

$$s_{n+1}(t) = l_{n+1} + \Delta t \cdot \dot{x}_{n+1}(t)$$

Note that, again, this is a linear $s(v)$ model, with the particularity that there exists a calibration parameter, Δt, instead of the predefined adherence to an arbitrary driving code in Pipes. So, the functional behavior of the model is the same, but it can be better adjusted to data by calibrating Δt in each particular context. Still, there is a wide difference between modeled and observed values in the minimum spacing at low and high speeds.

5. General Motors (1958-1961) car-following models

The General Motors (GM) car-following models refer to different generations of an analytical car-following model structure which were developed by traffic engineers and physicists hired at the General Motors research lab in Detroit (USA) during the late 1950's and early 1960s. These were the first car-following models to use field data for the specification and calibration of the models. Data were obtained from pioneer field trials at the test track in the General Motors headquarters. Note that in the 1960's was not easy to measure vehicular spacing as a function of the travelling speed, and the field experiments itself were an engineering challenge using wired-linked vehicles.

All the generations of the GM models have the same basic analytical form:

$$Response = function(stimuli, sensitivity)$$

Vehicular $response$ is always an acceleration or deceleration of the follower vehicle, while the $stimuli$ is always the speed difference between the leader and the follower one reaction time before. The various modifications of the $sensitivity$ term led to the different generations of the GM's car-following models.

i. 1st Generation:

Sensitivity, α, is a constant parameter. Then, the 1st generation GM car following model is formulated as:

$$\ddot{x}_{n+1}(t + \Delta t) = \alpha \cdot [\dot{x}_n(t) - \dot{x}_{n+1}(t)]$$

Note that the magnitude of the response, $\ddot{x}_{n+1}(t + \Delta t)$, is directly proportional to the relative velocity between the leader and the follower at the time of observation (i.e. before the reaction time).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

In this model, α and Δt are constant parameters to calibrate. The General Motors research team conducted field experiments to quantify the values for the reaction time and the sensitivity. In the experiment, eight different drivers were used in the instrumented car and were asked to follow the lead vehicle while maintaining a safe distance. This might resemble the GM experiments https://www.youtube.com/watch?v=Suugn-p5C1M. While calibration results for Δt were consistent (e.g. between 1s and 2s), for α they showed a large variability depending on the driving conditions (e.g. from 0.17s^{-1} to 0.74 s^{-1}). This is what lead GM researchers to think that the sensitivity parameter, α, probably was not a constant parameter, and gave rise to the 2nd generation model.

ii. 2nd Generation:

The 2nd generation of the GM model has the same functional expression as in the 1st generation but with two possible values for the sensitivity parameter, α_1 and α_2. α_1 was calibrated for large spacings, when drivers are less attentive to driving conditions. α_2 was calibrated with small spacings, when the sensitivity to the leader changes of speed is larger (see Figure 4). The problem with this discontinuous model was to determine at which spacing to shift from α_1 to α_2. Also, such a sharp change in behavior was not observed in field experiments. This suggested to include in the model a continuous linear variation of the sensitivity parameter, and this led to the 3rd generation.

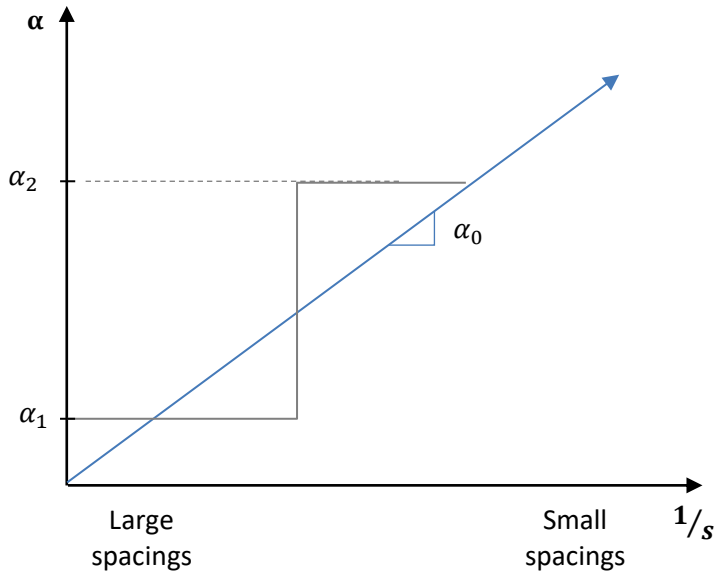


Figure 4. Sensitivity definition in the 2nd and 3rd generations of the General Motors' car following model.

iii. 3rd Generation:

It is assumed that the sensitivity parameter grows linearly with the inverse of the vehicle spacing. Then, the response is inversely proportional to the spacing at the time of observation, which is reasonable. α_0 is the parameter which represents the growing rate (i.e. $\alpha = \frac{\alpha_0}{[x_n(t) - x_{n+1}(t)]}$). Be careful, α_0 is not the sensitivity. It is

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

the rate at which the sensitivity grows with the inverse of spacing (see Figure 4). Note that the units of α_0 are [m/s]. Then, the 3rd generation of the GM car following models is expressed as:

$$\ddot{x}_{n+1}(t+\Delta t) = \frac{\alpha_0}{[x_n(t) - x_{n+1}(t)]} \cdot [\dot{x}_n(t) - \dot{x}_{n+1}(t)]$$

Experiments to calibrate α_0 and Δt were also performed using wired-linked vehicles. These were conducted at the Holland, Queens Mid-town, and Lincoln Tunnels in New York with eleven different drivers, and at the General Motors test track. Calibration results of α_0 demonstrated that the obtained values were similar to the optimal speeds (i.e. the average speed at which capacity takes place) for these infrastructures. This result guided researchers to prove the link between the 3rd GM model generation and the macroscopic Greenberg $k - v$ relationship. The analytical prove of this statement is provided in Appendix 1 to this chapter. Note that by integrating a (microscopic) car-following model you should obtain a traffic diagram (i.e. a macroscopic relationship between two fundamental traffic variables).

iv. 4th Generation:

Further experiments suggested that the sensitivity parameter did not only depend on vehicles' spacing (as in the 3rd generation) but also on the travelling speed. Note that even if the spacing is very small, if speed is virtually zero, drivers are not attentive to traffic. Therefore, sensitivity should grow with the inverse of spacing and with traveling speed. This was incorporated in the 4th generation of the GM car-following, where the sensitivity is expressed as $\alpha = \frac{\alpha' \dot{x}_{n+1}(t)}{[x_n(t) - x_{n+1}(t)]}$. Note that in this case, the parameter α' is dimensionless. This latter artifact does not necessarily agree with the physics of the problem, as fluid resistance and non-constant power curves greatly reduce the acceleration (positive) that a vehicle is capable of at high speeds.

Then, the car-following expression is:

$$\ddot{x}_{n+1}(t+\Delta t) = \frac{\alpha' \dot{x}_{n+1}(t)}{[x_n(t) - x_{n+1}(t)]} \cdot [\dot{x}_n(t) - \dot{x}_{n+1}(t)]$$

v. 5th Generation:

Finally, the objective of formulating a 5th generation was only to establish a generalized form of car-following models developed by the General Motors' research team. All the previous model generations are special cases of this 5th generalized model (see Figure 5). In addition, the 5th generation includes two exponents (i.e. m and l) which represent additional parameters to be calibrated. These exponents were only incorporated to increase the degrees of freedom of the model, which changed from two to four in this 5th generation. The objective was to allow an easier calibration of the model considering the new data sets that started being more and more available at the time.

$$\ddot{x}_{n+1}(t+\Delta t) = \frac{\alpha' [\dot{x}_{n+1}(t)]^m}{[x_n(t) - x_{n+1}(t)]^l} \cdot [\dot{x}_n(t) - \dot{x}_{n+1}(t)]$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

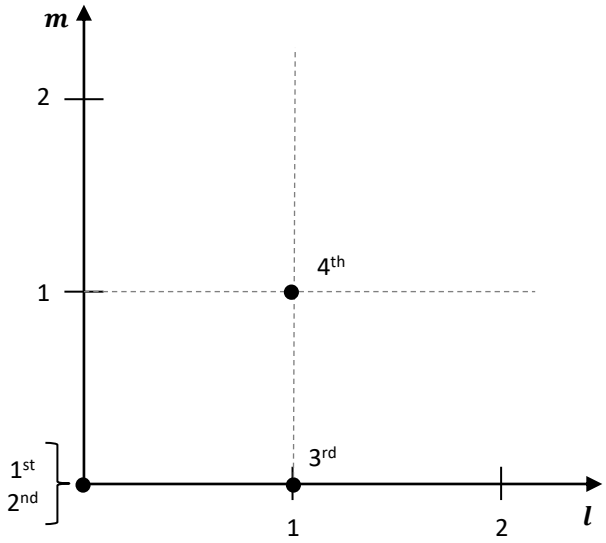


Figure 5. Parameters of the General Motors' car following model in different generations.

6. Traffic Stability

Linear differential equations, like the ones representing the family of GM car-following models are subject to oscillatory behavior. In this context, it is important to address the stability of the models with respect to disturbances.

Two particular types of stability are examined: local stability and asymptotic stability.

- Local Stability is concerned with the response of a following vehicle to a fluctuation in the motion of the vehicle directly in front of it (i.e. it is concerned with the localized behavior between pairs of vehicles).
- Asymptotic Stability is concerned with the manner in which a fluctuation in the motion of any vehicle, say the lead vehicle of a platoon, is propagated through a line of vehicles.

The stability analysis of the car-following equations develops the criteria which characterize the types of possible motion allowed by the model. For a given range of model parameters, a stability analysis determines if the traffic stream (as described by the model) is stable or not, (i.e., whether disturbances are damped, bounded, or unbounded). This is an important determination with respect to understanding the applicability of the model, and it identifies several characteristics with respect to single lane traffic flow, safety, and model validity. If the model is realistic, the stability range should be consistent with measured values of these parameters in any applicable situation where disturbances are known to be stable. It should also be consistent with the fact that following a vehicle is an extremely common experience, and is generally stable.

In order to perform the mathematical stability analysis of the linear differential equation describing the GM car-following models it is useful to rescale time in units of reaction times. This is $t = \tau \cdot \Delta t$. Using this transformation, the GM car-following equation can be written as:

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$\ddot{x}_{n+1}(\tau+1) = C \cdot [\dot{x}_n(\tau) - \dot{x}_{n+1}(\tau)]$$

Where $C = \alpha \cdot \Delta t$.

The stability conditions of this linear differential equation depend on the values of C, which may be referred as the "instability parameter" of the equation.

Determining local stability requires to apply the Laplace transformation to the differential equation. For asymptotic stability the Fourier components of the speed fluctuation of a platoon leader must be analyzed. This mathematical analysis is beyond the objectives of the present course, and only the obtained results will be highlighted here.

i. Local stability of GM car-following models

- If $C \leq e^{-1}$ (≈ 0.368) then motion is non-oscillatory and exponentially damped.
- If $e^{-1} < C < \pi/2$ then the motion is oscillatory with exponential damping.
- if $C = \pi/2$ then the motion is oscillatory with constant amplitude.
- if $C > \pi/2$ then motion is oscillatory with increasing amplitude.

The above establishes criteria for the numerical values of C which characterize the motion of the following vehicle. In particular, it demonstrates that in order for the following vehicle not to over-compensate to a fluctuation, it is necessary that $C \leq e^{-1}$. For values of C that are somewhat greater, oscillations occur but are heavily damped and therefore insignificant. Damping occurs to some extent as long as $C < \pi/2$.

These results concerning the oscillatory and non-oscillatory behavior apply to the speed and acceleration of the following vehicle, as well as to the inter-vehicle spacing. Thus, e.g., if $C \leq e^{-1}$, the inter-vehicle spacing changes in a non-oscillatory manner by the amount Δs, where:

$$\Delta s = \frac{1}{\alpha}(V - U)$$

Δs is the final difference in the inter-vehicle spacing (i.e. between the leader and the follower) after a change of the speed of the leader from speed U to speed V. Note that when the car-following behavior recovers stationarity after the speed change, the speed of the follower has also changed from speed U to speed V.

An important case is when the lead vehicle stops. Then, the final speed, V, is zero, and the total change in inter-vehicle spacing is $-U/\alpha$. This means that in order for a following vehicle to avoid a 'collision' from initiation of a fluctuation in the lead vehicle's speed the inter-vehicle spacing should be at least as large as U/α. On the other hand, in the interests of traffic flow the inter-vehicle spacing should be small by having α as large as possible and yet within the stable limit. Ideally, the best choice is $\alpha = (e\Delta t)^{-1}$.

Figure 6 shows the changes in vehicular spacing when the leader suffers a sudden fluctuation in acceleration This disturbance consists of a slowing down and then a speeding up to the original speed so that the integral of acceleration over time is zero (e.g. a sharp deceleration of -1.8 m/s^2 during 2s followed by an equivalent acceleration to return to the original speed). Results are shown for four different values of C. For the values of

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$C = 0.5$ and 0.8, the spacing is oscillatory and heavily damped. For $C = 1.57$ ($\approx \pi/2$) the spacing oscillates with constant amplitude. For $C = 1.6$ motion is oscillatory with increasing amplitude.

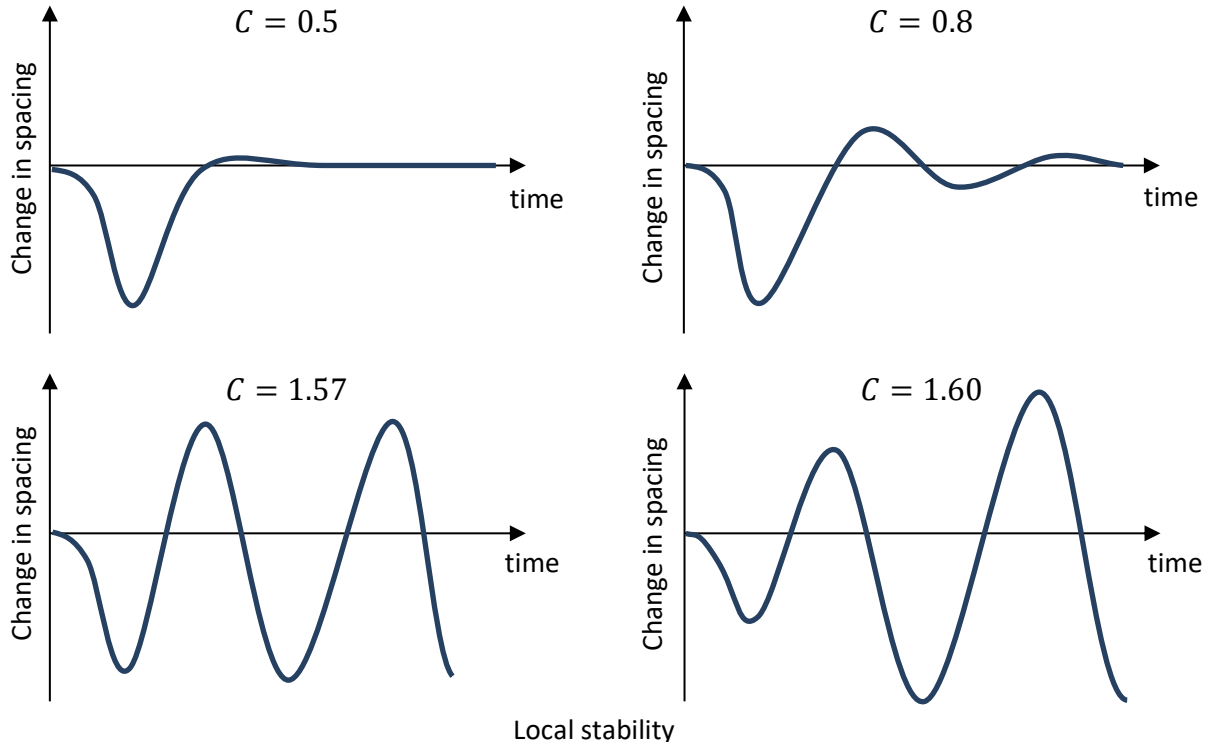


Figure 6. Local stability as a function of the instability parameter "C".

ii. Asymptotic stability of GM car-following models:

In the previous analysis, the behavior of one vehicle following another was considered. Here a platoon of vehicles (except for the platoon leader) follows the vehicle ahead according to the linear car following equation.

In this case, the mathematical analysis of the asymptotic oscillatory behavior of the differential equation shows that the motion of a string of vehicles is stable if $C < 0.5$. This means that the criteria for local stability (namely that no local oscillations occur when $C \leq e^{-1}$) also ensures asymptotic stability.

Also note that while $C < 0.5$ ensures stability, short reaction times increase the range of the sensitivity coefficient, α, that ensures stability. From a practical viewpoint, small reaction times also reduce relatively large responses to a given stimulus, or in contrast, larger response times require relatively large responses to compensate a given stimulus.

Table 1 summarizes the results for local and asymptotic stability of GM car-following models.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Table 1 – Local and asymptotic stability of GM car-following models

C	Stability	
	Local	Asymptotic
0	Not oscillatory	Damped oscillation
0.5	Damped oscillation	
1.0		Growing oscillation
1.5	Growing oscillation	
2.0		

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

APPENDIX 1 - Equivalency between the 3rd generation of the General Motors car following model and the Greenberg $k - v$ macro model

This appendix proves the identity between the sensitivity parameter "α_0" in the 3rd generation of the General Motors car following model and the optimal speed "v_0" in the Greenberg $k - v$ macro model.

Notation:

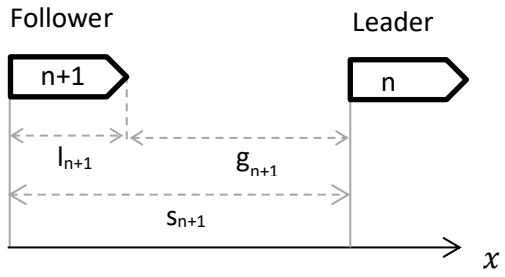
3rd generation of the General Motors car following model:

$$\ddot{x}_{n+1}(t + \Delta t) = \frac{\alpha_0}{[x_n(t) - x_{n+1}(t)]}[\dot{x}_n(t) - \dot{x}_{n+1}(t)]$$

Greenberg $k - v$ macro model:

$$v = v_0 \ln\left(\frac{k_j}{k}\right)$$

Definitions:

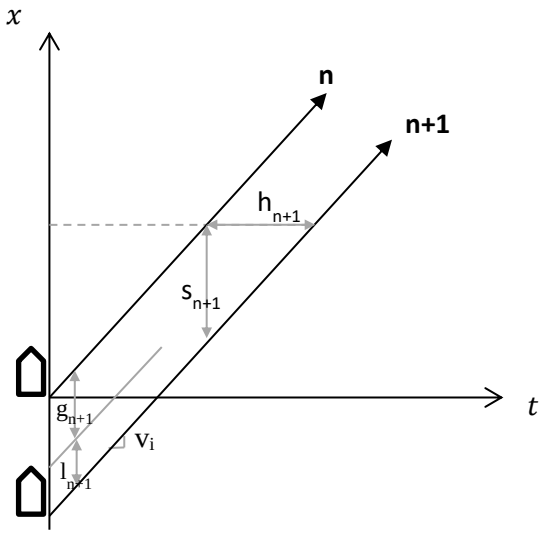


Definitions from the REAR BUMPER

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

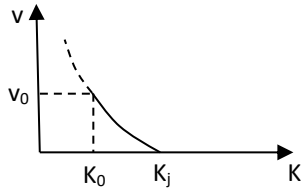


$$s_{n+1}(t) = l_{n+1} + g_{n+1}(t)$$

$$s_n(t) = h_{n+1}(t) \cdot v_{n+1}(t)$$

MICRO-MACRO equivalency

- Greenberg Model:



$$v = v_0 \ln\left(\frac{k_j}{k}\right)$$

v_0 optimal speed (q_{max})

- GM third generation car-following model:

$$\ddot{x}_{n+1}(t + \Delta t) = \frac{\alpha_0}{[x_n(t) - x_{n+1}(t)]}[\dot{x}_n(t) - \dot{x}_{n+1}(t)]$$ MICRO model

Integrating with respect to time "t":

$$\dot{x}_{n+1} = \alpha_0 . log[x_n(t) - x_{n+1}(t)] + c_1$$

Because $\frac{1}{k} = s_{n+1} = [x_n(t) - x_{n+1}(t)]$, then:

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$v = \alpha_0 . log\left(\frac{1}{k}\right) + c_1 = \alpha_0 . log\left(\frac{c_2}{k}\right) \quad with \quad c_2 = e^{\frac{c_1}{\alpha_0}}$$

$$k = k_j \;\rightarrow\; v = 0$$

$$log\left(\frac{c_2}{k_j}\right) = 0 \;\rightarrow\; c_2 = k_j$$

$$v = \alpha_0 . log\left(\frac{k_j}{k}\right)$$

By comparing this expression with the Greenberg macro model, the optimal speed (maximum flow): $v_0 = \alpha_0$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

6-SCHEDULED TRANSPORTATION

Table of Contents

1. Introduction 2
2. Average wait times in scheduled transportation systems 4
2.1. Average wait time in systems with short headways 4
2.2. Average wait time in systems with long headways 8
3. Transfers design 11
4. Model for the vehicles' trip time 13
4.1. Parameters' estimation 15
4.2. Expected number of stops 16
5. Required vehicles to serve a route 19
5.1. Vehicle occupancy in the critical link 19
6. Stochastic effects on service regularity 20
7. Schedule control 22
7.1. Find an expression for $\delta \approx f(k)$ 23
7.2. Select k so that it minimizes the total time a typical passenger is in the system 25
8. Planning of a scheduled transportation system 27
8.1. Determining the optimal stop spacing, s^*. The customer perspective. 28
8.2. Determining the optimal headway, h^*, and the optimal spacing, s^*. The global perspective 31

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

1. Introduction

The objective of this chapter is to analyze how customers interact with the scheduled operation of collective transportation systems.

Scheduled transportation systems are operated based on routes (with predefined stop locations) and schedules (with a predefined frequency). The spatial discreetness of the transportation system is defined by the route spacing, S (i.e. the average distance between routes in the transportation network) and the stop spacing, s (i.e. the average distance between consecutive stops in a given route). The temporal discreetness is defined by the service frequency, F (i.e. number of expeditions per unit time). The inverse of the service frequency is the headway, h (i.e. the average time between consecutive expeditions) which provides equivalent information. Clearly, increasing the service (either increasing the spatial or temporal frequency) implies costs to the operating agencies[1] (infrastructure, vehicles, labor, fuel,...). Part of these costs are transferred to the customers as fares. Note that generally public transportation is subsidized, so that the total revenues paid by the customers only mean part of the total cost of the system. Subsidies are justified because they return to society in terms of reduced externalities of urban mobility (e.g. less congestion, less pollution, less accidents, less consumption of urban space).

Besides, this spatial and temporal discreteness of service means that customers must adapt their travel wishes in time and space. This adaptation involves two customer penalties (or costs): access to the system and wait. This is illustrated in Figure 1, showing the multiple stages in a typical trip using a scheduled collective transportation system.

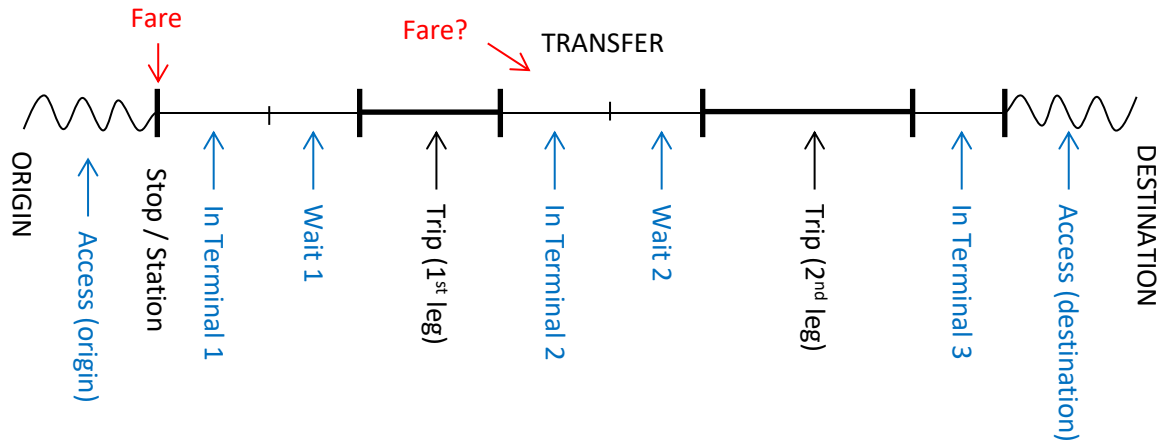


Figure 1. The stages of the scheduled transportation trip

[1] Although in some cities there might be several uncoordinated companies operating the collective urban transportation system, in most developed countries there is a single agency or entity responsible for ensuring an integrated and coordinated service, even if operated by several companies. From now on, we will discuss transport operator or agencies as the company providing public transport services, even if it is operated by different companies.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

First, customers must spend some time traveling from their origins to the nearest access point to the system (e.g. stop, station, terminal...); this represents the access time, A. Note that there is also an "access time" at destination. This is usually called the egress phase of the journey (i.e. from the last stop to the final destination). A includes both, access and egress times, and also the times and activities (if any) needed before reaching the boarding place inside the terminal.

Next, because of the transportation service is only provided every h time units, user must wait for the vehicle to arrive. This waiting time, W, should also be considered as part of the trip. Note that one could also think of a "wait" time at the destination. This is, if the objective of the trip is to attend to some type of scheduled activity (i.e. with a defined start time) it is possible that because of the discreteness of the transportation service, customers cannot arrive just in time. They need to arrive earlier and wait for the start of the activity. This "wait" at the destination should also be considered in W.

As described, customers incur in access and wait times at their origins and destinations. In addition, it is possible that some customers cannot travel their whole trip using a single route. In this case, they need to transfer routes. This may also involve a change of technology (e.g. from rail to bus). In any case, this intermediate stop (i.e. the transfer), implies additional access and wait times. An adequate design of stop areas, together with some type of coordination between services helps in minimizing these customer penalties. In some instances, a transfer might also involve a second fare to the customer. These are avoided in case of "integrated" pricing systems, where the payment of a single fare allows to travel the whole trip within a defined zone, independently of the number of legs of the trip.

Obviously, access and wait penalties decrease with the level of service provided (i.e. spatial coverage and service frequency), defining a clear trade-off with respect to agency costs. This trade-off should be optimized in the strategical planning of scheduled transportation systems.

Finally, the last component of the trip time is the in-vehicle travel time ($IVTT$). This $IVTT$ is used to overcome the distance at the cruising speed of the vehicle, plus some time needed to stop to pick-up and deliver other customers. This last component of the $IVTT$ depends on the number of stops and on the demand for the service, which needs to be carefully analyzed in order to avoid a significant reduction in the commercial speed of the system. Note the difference between the cruising speed (i.e. the average speed at which the vehicle travels, including possible stops due to congestion, traffic signals, etc... but not to pick-up and deliver passengers) and the commercial speed (i.e. the overall speed of the transportation system from the user perspective).

In summary, the total door-to-door travel time when traveling in a scheduled transportation system can be estimated as:

$$T = A + W + IVTT$$

This door-to-door travel time represents the user costs in the system. Note that the fares are not considered in the user costs, as they are only a transference of part of the agency cost to the user. Considering them in the user costs would imply counting these fares twice (i.e. in the users and in the agency costs). Typically, in order to obtain the total costs, Z (i.e. Z_U, users + Z_A, agency costs) we need to monetize the user costs (i.e. change from units of time to monetary units). To that end, the users' value of time needs to be used as the conversion factor. However, the user's perception of time spent in different stages of the transport chain is different. Users often

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

perceive more negatively the time spent in accessing the stop or waiting than the in-vehicle travel time. For this reason, it is common to monetize the perceived travel time, T_p, computed as a weighted sum of A, W and $IVTT$, affected by their corresponding perception weights (see Table 1).

$$T_p = w_A A + w_W W + w_{IVTT} IVTT$$

Often the weight associated with $IVTT$ is considered to be 1, so that the other weights are estimated relative to it. Table 1 shows typical values of these weights caused by different perceptions of time.

Table 1. Typical weights for different perception of times in the various stages of the trip chain.

	w_{IVTT}	w_A	w_W
Average weight	1.0	2.2	2.1

Note: In secondary waits and access at transfers, these weights might be multiplied by 2.

Next in this chapter, we are going to analyze user costs. For instance, we are going to estimate the wait times due to the discreetness of schedules. We will see that in order to minimize wait times it is important to keep the regularity in schedules. Then, the second part of the chapter will deal with the operating agency. We will analyze the vehicle trip times in a multi-stop system, their optimal design, the possible sources of delay, and the snowball effect of such delays. The chapter will end with some possible strategies to control such damaging effects and maintain regular schedules.

2. Average wait times in scheduled transportation systems

Waits in scheduled transportation systems appear as a result of the discreteness of service in time. The objective of this section is to determine the average wait as a function of the service headway, h. Recall that the headway represents the time between consecutive expeditions, and can be obtained as the inverse of the service frequency, F [expeditions/time]. Such analysis will prove that the average wait in scheduled transportation systems depends on the magnitude of the headway and on its regularity.

Before starting the derivations, two contexts must be differentiated. First, services with short headways (e.g. less than 15 min) where users arrive to the stop without reference to the schedule, anticipating a short wait. This is the case in most of the urban collective transportation systems. Second, services with long headways, where customers arrive just before the scheduled departure time, typical in inter-urban transportation systems.

2.1. Average wait time in systems with short headways

Consider the waiting at a single stop and assume:

- Customers arrive to the stop without reference to schedule.
- Customers' arrival times at the stops are uniformly distributed in time, with an average demand rate of $\bar{\lambda}$ [pax/h].
- Vehicles have enough capacity to allow everybody waiting to board.
- The average headway of service is $\bar{h}$ [h], but this is a random variable that fluctuates.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

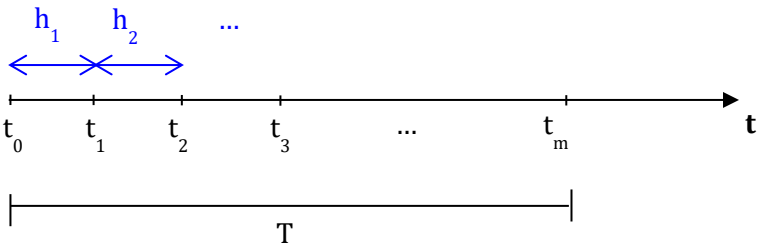
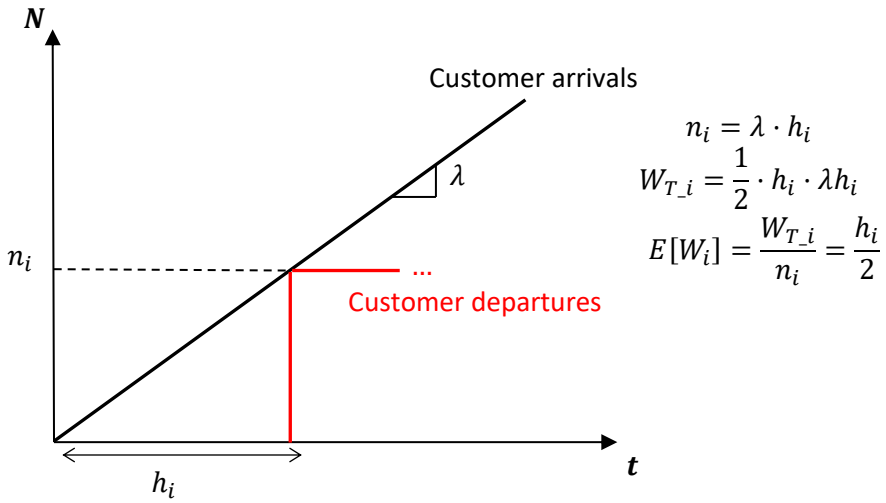


Figure 2. Scheduled departure times

Figure 2 shows the different arrival times of the vehicle to the stop (i.e. t_i , $i = 1 \ldots m$) and the resulting headways (i.e. h_i , $i = 1 \ldots m$). Note that considering a sufficiently long observation period, T, then:

$$\bar{h} = \frac{\sum_{i=1}^{m} h_i}{m}$$

Consider a customer arriving to the stop during the headway h_i. Her maximum wait would be h_i if arriving just after the departure of the previous vehicle. The minimum wait would be 0, if arriving just before the departure of the vehicle. Therefore, assuming uniform arrivals during h_i, the average wait is $E(W_i) = h_i/2$. Figure 3 illustrates this derivation using an (N,t) input-output diagram.



The expected wait time of those passengers arriving during h_i is $\frac{h_i}{2}$

Figure 3. Expected wait time in a given headway

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

However, the expected wait time, $E(W)$, for all those passengers arriving during the observation period T is not $\bar{h}/2$. This is because there is more probability to arrive during the long headways, and therefore to suffer longer waits. More specifically, $E(W)$, can be computed as the weighted sum of the average waits if arriving during h_i, where the weights are the probability of arriving during such headway. If arrivals are uniformly distributed in time, this probability is $\frac{h_i}{T}$. This is:

$$E(W) = \sum_{i=1}^{m} \frac{h_i}{2} \cdot \frac{h_i}{T}$$

Because $T = \sum_{i=1}^{m} h_i$, the previous equation can be rewritten as:

$$E(W) = \frac{\frac{1}{2}\sum_{i=1}^{m}[h_i^2]}{\sum_{i=1}^{m} h_i}$$

And multiplying the numerator and denominator of the previous expression by $\frac{1}{m}$, we obtain:

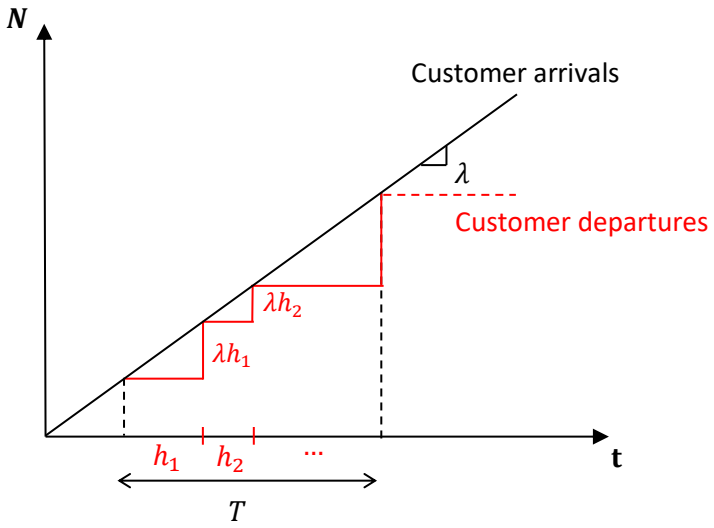
$$E(W) = \frac{\frac{1}{2}\overline{h^2}}{\bar{h}} \geq \frac{\bar{h}}{2}$$

This expression for the expected wait time as a function of the observed headways, can also be obtained by using (N,t) input-output curves, as shown in Figure 4.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >



$$W_{Total} = \sum_i \frac{1}{2} h_i \cdot \lambda \cdot h_i$$

$$N_{Total} = \lambda T = \lambda \sum_i h_i$$

$$E[w] = \frac{W_{Total}}{N_{Total}} = \frac{1/2 \sum_i h_i^2}{\sum_i h_i} = \frac{1/2 \, \overline{h^2}}{\bar{h}} \geq \frac{\bar{h}}{2}$$

Figure 4. Expected wait time in a period with different and equiprobable headways

This derivation proves that the average wait during an extended period of time containing multiple headways, is larger than half the average headway. Specifically, if one recalls that the variance of a random variable (e.g. the headway) can be computed as:

$$Var(h) = E(h^2) - [E(h)]^2$$

Or in terms of their estimations from a sample population:

$$S_h^2 = \overline{h^2} - \left[\bar{h}\right]^2$$

where S_h is the standard deviation of the headways. Then:

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$E(W) = \frac{\frac{1}{2}\left(S_h^2 + \left[\bar{h}\right]^2\right)}{\bar{h}} = \frac{1}{2}\left(\bar{h} + \frac{S_h^2}{\bar{h}}\right)$$

The previous expression shows how the regularity of the service benefits customers by reducing the average wait at stops. Note that if the headway is kept perfectly regular (i.e. $h_i = \bar{h}\ \forall i = 1 \ldots m$) then its variance is null (i.e. $S_h^2 = 0$) and the average wait is minimum, and equal to $\bar{h}/2$. Fluctuations in the headway imply the increase of its variance and result in extended wait times.

2.2. Average wait time in systems with long headways

If the headway at which is operated the system is long (≥15min), passengers arrive at the bus stop as per vehicle schedule. Typically, vehicles suffer random deviations from schedule, as illustrated in Figure 5.

The objective is to show, with a particular example that:

- That $E(W)$ is minimum if passenger arrive at the lowest of the probability distribution of schedule deviations (left hand side).
- That $E(W) = some\ constant \cdot std\ deviation\ of\ bus\ schedule\ deviations$. The $constant$ depends on the particular distribution of schedule deviations, but in any case, this would prove the importance of schedule adherence to minimize wait times.

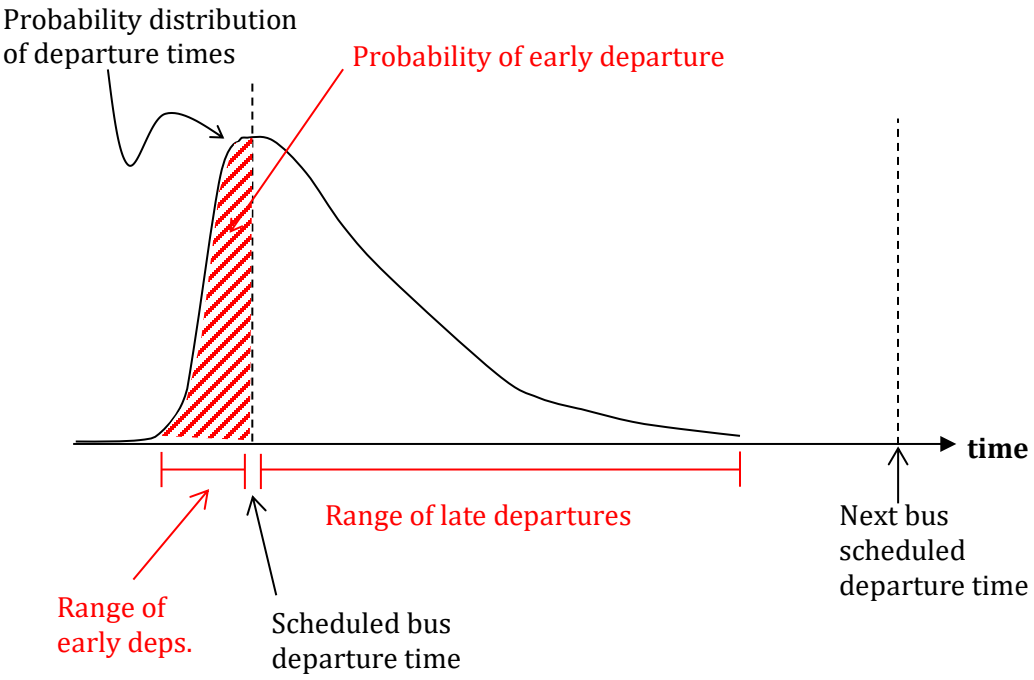


Figure 5. Distribution of vehicle schedule deviations

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

As a simplistic example, consider that the distribution of vehicle's deviations from schedule is uniform, $U(-k,k)$, like illustrated in Figure 6. Note that the scheduled headway is larger than the maximum range of deviations (i.e. non-overlapping distributions), so that if one customer misses the first departure, he will always be able to get on the second one.

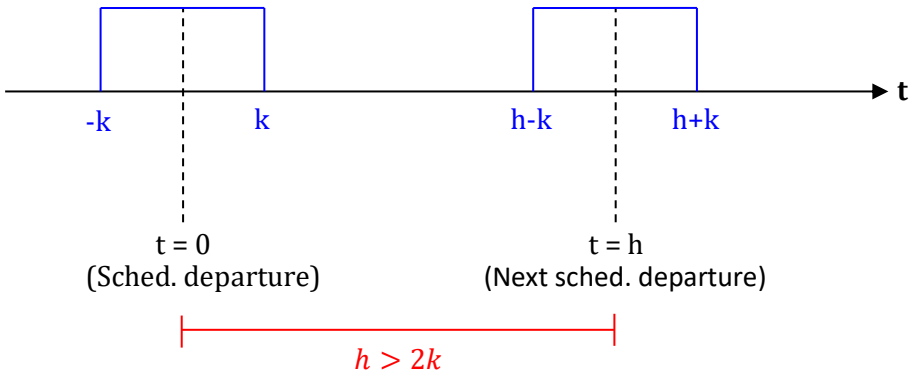


Figure 6. Uniform distribution of schedule deviations $\sim U(-k,k)$

If we call τ as the customer arrival time to the bus stop, and we plot the expected wait conditioned to τ, $E(W|\tau)$ as a function of τ, we would obtain:

$$E(W|\tau) = \begin{cases} -\tau & if\ \tau < -k \\ h - \tau & if\ \tau > k\ and\ \tau < h - k \\ concave\ function & for\ -k \le \tau \le k \end{cases}$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

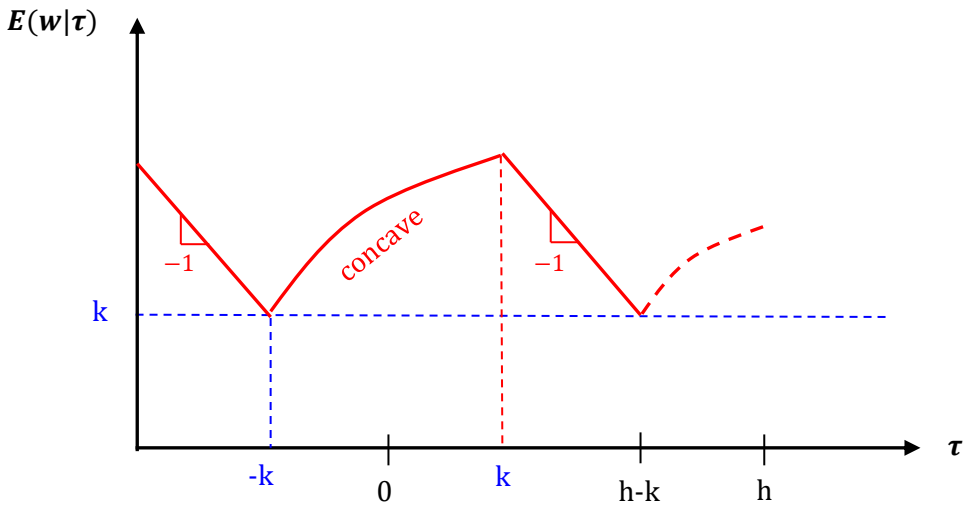


Figure 7. Expected wait as a function of the arrival instant, τ, with respect to the scheduled departure time.

If we show that this is correct, we will have proved that $E(W)$ is minimum if passenger arrive at $-k$ with respect to the scheduled departure (i.e. the left-hand side of the distribution of schedule deviations).

In order to prove the previous expression, note that:

$$E(W|\tau) = \text{Probability}(miss) \cdot \text{Wait}(miss) + \text{Probability}(not\_miss) \cdot \text{Wait}(not\_miss)$$

And for the interesting case where $-k \leq \tau \leq k$ and the probabilities to miss (or not miss) are not zero or one, we obtain:

$$E(W|\tau) = \frac{1}{2k}(\tau + k)(h - \tau) + \left(1 - \frac{\tau + k}{2k}\right)\left(\frac{1}{2}(k - \tau)\right) = \left(\frac{-\tau^2}{2} + h(\tau + k) - 2\tau k + \frac{k^2}{2}\right)/2k$$

You can check the correctness of the previous expression by realizing that for $\tau = -k$, $E(W) = k$. Also, you can realize that the previous expression is concave (i.e. 2nd derivative < 0) because it is the sum of concave and linear components with respect to τ.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

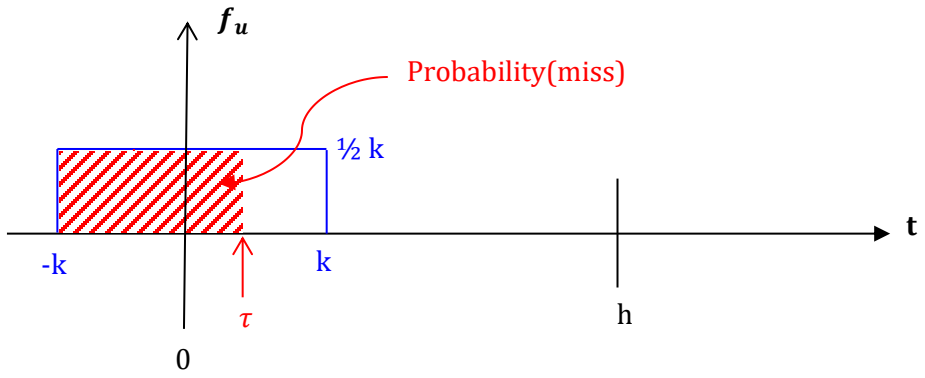


Figure 8. Probability of missing the departure for a given arrival time at the stop, τ.

Finally, note that the variance of a uniform distribution $U(a,b)$ is $Var(U) = (b-a)^2/12$. In our case, $a = -k$ and $b = k$, then:

$$Var(U) = \frac{(k+k)^2}{12} = \frac{k^2}{3}$$

This means that:

$$k^2 = 3 \cdot Var(U)$$

$$k = \sqrt{3} \cdot std\ deviation\ of\ bus\ schedule\ deviations$$

So, we have proved (with an example) that:

- $E(W)$ is minimum if users arrive at the left hand side of the distribution of vehicles' deviations from schedule (e.g. $E(W) = k$).
  In reality, without the knowledge of the distribution of deviations, customers arrive before the scheduled departure time to minimize the risk of missing the departure. The anticipation depends on the penalties of missing the trip (e.g. 5 min for a urban bus, or at least 1h for a flight).
- This minimum $E(W)$ increases with the variability of schedule deviations (i.e. $E(W) = some\ constant \cdot std\ deviation\ of\ bus\ schedule\ deviations$).
  In our example $E(W) = k = \sqrt{3} \cdot std\ deviation\ of\ bus\ schedule\ deviations$.

3. Transfers design

The expected wait at transfer stops follows a different logic than at the origin of the trip. At transfers, we cannot assume that customer arrivals are independent events following a certain probability distribution. At transfers,

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

customers arrive in batches of n_k individuals at discrete time instants a_k. This is illustrated in Figure 9, using an (N,t) input-output diagram.

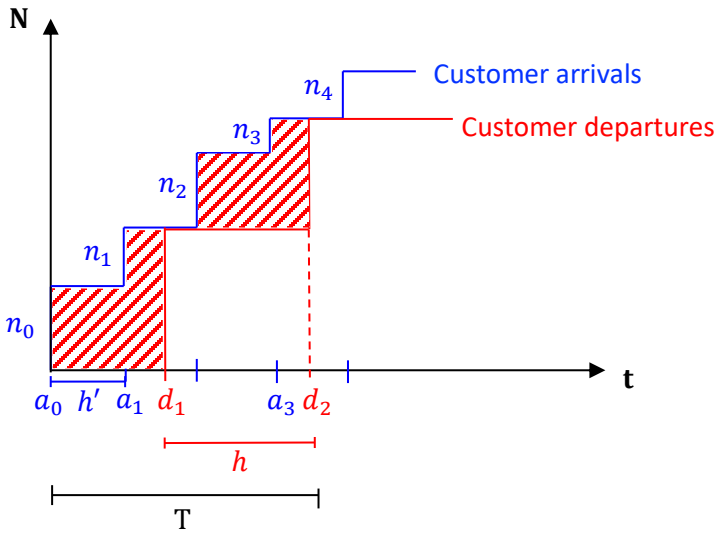


Figure 9. Transfers design

In Figure 9, n_k customers arrive every h' time units (i.e. at a_k). h' is the headway of the arriving services at the transfer stop. In turn, customers depart every h time units (i.e. at d_k). h is the headway of departing services from the transfer stop. Note that it is assumed that departing vehicles have enough capacity to carry all arriving passengers.

The total wait at the transfer stop is represented by the area between the arrival and departure curves in Figure 9. For a particular vehicle k this can be computed as the arriving customers, n_k, multiplied by their wait, w_k. For an extended period of observation T, this is:

$$W_{(T)} = \sum_k n_k \cdot w_k$$

And the expected wait per customer would be:

$$\bar{W}_{(T)} = \frac{\sum_k n_k \cdot w_k}{\sum_k n_k}$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Note that the previous wait could be minimized by synchronizing arrivals and departures, so that the area between arrival and departure curves in the (N,t) diagram was minimum. This synchronization can be achieved if h and h' are natural multiples. In the example of Figure 9, if $h = 2h'$, we could synchronize arrivals and departures so that $w_0 = w_2 = \ldots = h'/2$ and $w_1 = w_3 = \ldots = 0$, so that:

$$\bar{W} = \frac{1}{2}(h - h') = \frac{h}{4}$$

Note that in case of independent customer arrivals, and even considering perfect regularity, the average wait would have been longer (i.e. $h/2$).

In order to ensure that the headway synchronization is possible in case of multiple routes transferring at the same station (possibly with different headways), all the headways need to be natural multiples of the minimum headway, $\tilde{h}$. This is achieved if all the headways are defined from a menu of the form $\tilde{h}2^p$, where $p \in \mathbb{N}$.

4. Model for the vehicles' trip time

So far, we have discussed how schedule adherence benefits customers by minimizing their average wait times. Now, we are going to analyze why it is difficult for operating agencies to maintain regularity. We are going to see that the inevitable random fluctuations in vehicle trip times tend to amplify along the route, defining a positive feedback process. This means that, if nothing is done, schedule adherence would be completely lost as vehicle advances in the route. Therefore, headway control strategies need to be implemented to guarantee a regular service.

The first thing we need to analyze these concepts is a model for the vehicle trip time, T. Figure 10 shows the trajectory of a scheduled transport vehicle between two consecutive stops (e.g. between S_{i-1} and S_i). Being more precise, we can define this time between consecutive stops, T_i, as the time between the opening instant of doors at the given stop (i.e. between o_{i-1} and o_i). We lose nothing if we consider the trajectory as piecewise linear, and we add the lost times due to acceleration and deceleration to the time that the vehicle is stopped. Then, we can formulate:

$$T_i = \frac{D_i}{v_i} + L_i + time\ door\ is\ open\ at\ (i-1)$$

where D_i is the distance, v_i is the cruising speed and L_i is the sum of all lost times, all of them between S_{i-1} and S_i. L_i includes the times lost opening / closing doors, accelerating and decelerating. The time lost accelerating (i.e. from stop to v_i) is $v_i/2a$, where a is the acceleration rate[2]. Note that this is half the time invested accelerating, $t_a = v_i/a$. This allows an easy empirical measurement of times lost accelerating (or decelerating,

[2] The total time accelerating is $t_a = v_i/a$. The total distance covered while accelerating is $x_a = at_a^2/2 = v_i^2/2a$. The time that would be required to cover x_a traveling at v_i in the absence of the stop would have been $x_a/v_i = v_i/2a$. Then, the lime lost accelerating is $t_a - v_i/2a = v_i/2a$. An analogous estimation holds for the lost time decelerating.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

as the process is symmetrical assuming same acceleration / deceleration rates). While traveling on the vehicle (either the underground metro, or a bus, etc.) it can be felt when the vehicle starts braking. If one measures the time from this instant until the full stop, t_a will be obtained. The lost time decelerating would be half of this.

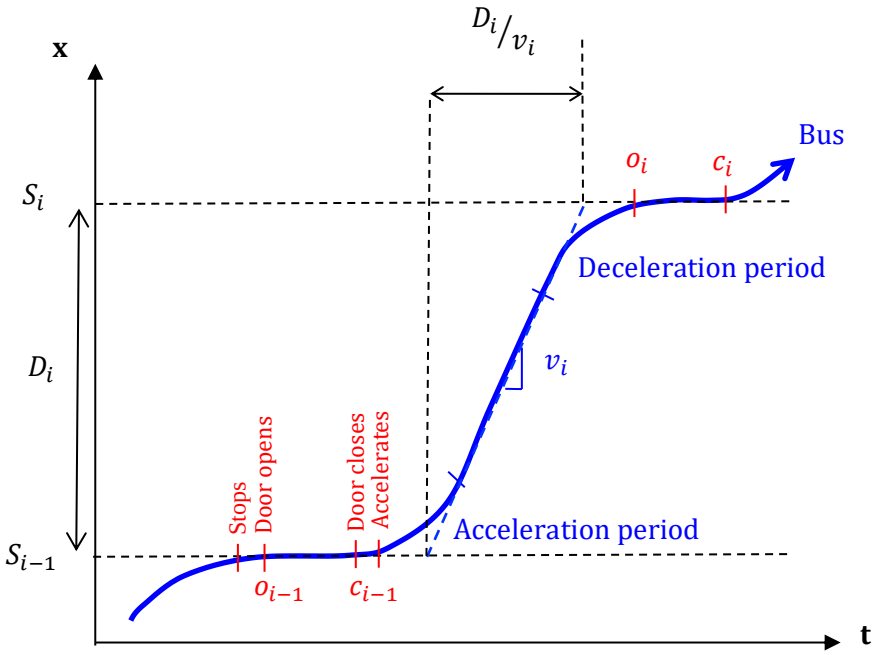


Figure 10. Model of vehicle trip time.

Regarding the time doors are open, this may be fixed and independent of the number of customers boarding and alighting (e.g. typical in rail systems) or dependent on the demand at the stop (e.g. like in urban bus systems). In this last case:

$$Time\ door\ is\ open\ at\ (i-1) = \tau'_{i-1} N_{a(i-1)} + \tau_{i-1} N_{b(i-1)}$$

where $N_{a(i-1)}$ and $N_{b(i-1)}$ are, respectively, the number of customers alighting and boarding at station $i-1$. In turn, τ'_{i-1} and τ_{i-1} are the unitary times for alighting and boarding. Note that this last expression assumes that customers first alight and after board, in sequence (e.g. for instance because there is a single door, or the doors are used for both alighting and boarding). If boarding and alighting happen simultaneously (e.g. because there is a door used only for each of the movements) the previous expression needs to be modified to:

$$Time\ door\ is\ open\ at\ (i-1) = Max\left[\tau'_{i-1} N_{a(i-1)}, \tau_{i-1} N_{b(i-1)}\right]$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

So far, we have derived a model for T_i, the time between consecutive stops. In order to obtain the trip time along the whole route, T, it is only necessary to add up T_i for all i in the route. This is:

$$T = \sum_i \left[\frac{D_i}{v_i} + L_i + \tau'_{i-1} N_{a(i-1)} + \tau_{i-1} N_{b(i-1)}\right]$$

We could assume that v_i, L_i, τ'_{i-1} and τ_{i-1} are independent of i. Then:

$$T = \frac{D}{v} + LN_S + (\tau' + \tau)N$$

Where D is the distance of the whole route, N_S is the total number of stops, and N is the total number of customers boarding (or alighting, they are the same) in the whole trip (i.e. $N = \sum_i N_{a(i)} = \sum_i N_{b(i)}$).

4.1. Parameters' estimation

v and L can be estimated by riding the vehicle, if one records the trip times between c_{i-1} and o_i (i.e. tt_i). Plotting this trip times against D_i, as in Figure 11, and fitting a linear regression, allows to determine v as the inverse of the slope of the linear regression, and L as the intercept with the vertical axis at $D_i = 0$. This L corresponds to the lost time accelerating and decelerating.

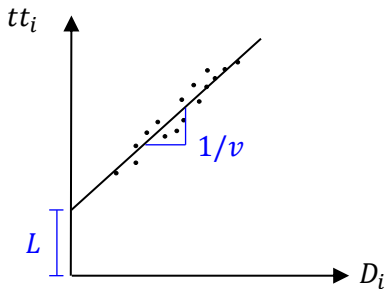


Figure 11. Estimation of "L" (lost times) and "v" average cruising speed of the vehicle.

If the linear relationship in the previous plot is poor, this might imply that there is not a constant cruising speed valid for the whole route. In such circumstance, try dividing the route in different parts where the cruising speed could be more similar.

In turn, τ' (or τ) can be estimated by registering the time taken by all customers alighting (or boarding) at a given stop i, and plotting this total time against the number of alightings (or boardings) (See Figure 12). Typically, these data show heteroscedasticity (i.e. growing variance with the growth of the variable), but this does not invalidate

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

the estimation method. The slope of this linear regression will yield the unitary time to alight, τ'(or to board, τ). This regression can be forced to go through the origin, which would impose that the lost times in opening /closing doors are negligible in front of the lost times accelerating / decelerating. If the interception in the linear regression is significant enough to avoid being neglected, then this lost time should be added to the previous estimation of L.

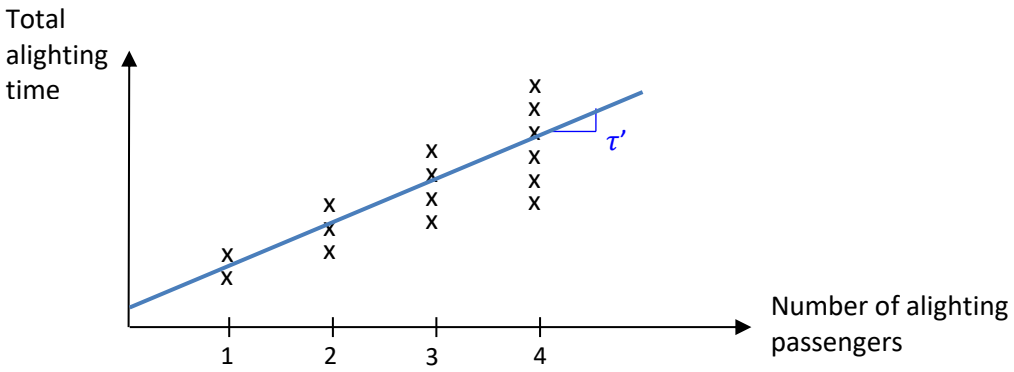


Figure 12. Estimation of unitary alighting time.

In the previous estimation, if the time taken by all customers alighting cannot be discerned from the time taken by boarding passengers (e.g. because of simultaneous movements), then a joint estimation must be performed. This is to perform a multiple linear regression, considering the time doors are open, T_i^o, as the dependent variable and the number of boardings, $N_{b(i-1)}$, and alightinghs, $N_{a(i-1)}$, as the explanatory variables. This is:

$$T_i^o = \alpha + \tau' N_{a(i-1)} + \tau N_{b(i-1)}$$

Note that α would be the time lost opening / closing doors.

4.2. Expected number of stops

So far, we have assumed in the vehicle trip time model that the number of stops is a deterministic fixed variable, N_S. In some scheduled systems (e.g. in rail systems) this is the case, as vehicles stop at every single station. However, in some other systems (e.g. urban bus systems), some stops might be skipped if there is no boarding and alighting demand. In such case, N_S, should be interpreted as a random variable, and it expected value, $E(N_S)$, should be considered in the vehicle trip time model.

In order to estimate $E(N_S)$, we assume that the number of people boarding and alighting at a certain stop i, is a Poisson random variable. This is a common assumption for the demand modeling of transportation systems.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Let's consider that we know the average number of boardings and alightings at every station. This information can be represented in an origin / destination table, like the one represented in Figure 13. Note that q_{ij} represents the demand rate (i.e. [pax/h]) boarding at station i and whose destination is station j.

O/D	1	2	3	…	n	
1	0	q_{12}	q_{13}	…		
2	–	0	q_{23}			
3	–		0	…		$q_i = \sum_{j=i+1}^{n} q_{ij}$
⋮						
n	–			–	0	
			$q'_j = \sum_{i=1}^{j-1} q_{ij}$			

Figure 13. O/D matrix for one route direction

Given the information in the O/D matrix in Figure 15, the average number of passengers boarding one particular vehicle at station i is $N_{b(i)} = q_i h$, where h is the headway between consecutive vehicle expeditions. Analogously, the average number of alighting passengers at station j is $N_{a(j)} = q'_j h$.

Given the Poisson assumption[3], the probability of having x boardings at station i is:

$$P(X = x) = \frac{(q_i h)^x e^{-q_i h}}{x!}$$

[3] The assumption of Poisson customer arrivals at the stop is reasonable if the headway is short and customers arrive without reference to schedule. For long headways, customer arrivals are concentrated some time before the scheduled departure. In such case, the arrival times cannot be assumed Poisson, but the aggregated number of arrivals during h still follow a Poisson distribution.

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Just change q_i by q_i' in order to obtain the probability for alightings.

Then, the probability of having zero boardings (or alightings) at station i is:

$$P(zero\ boardings\ at\ i) = e^{-q_i h}$$
$$P(zero\ alightings\ at\ i) = e^{-q_i' h}$$

Assuming that boardings and alightings at station i are statistically independent events[4], the probability that one vehicle does not stop at station i (i.e. because there are neither boardings nor alightings) is:

$$P(not\ stopping\ at\ i) = e^{-(q_i+q_i')h}$$

And therefore, the complementary probability for stopping is:

$$P(stopping\ at\ i) = 1 - e^{-(q_i+q_i')h}$$

Note that the probability of stopping grows with the demand rate and with the headway, which makes sense.

Finally, while covering a route with n stations, the expected number of stops is obtained as:

$$E(N_S) = \sum_{i=1}^{n} \left[1 - e^{-(q_i+q_i')h}\right]$$

The previous expression is commonly expressed as:

$$E(N_S) = \sum_{i=1}^{n} \left[1 - e^{-(p_i+p_i')qh}\right]$$

[4] This holds for one vehicle single trip. Over the whole day this does not hold, as there are major and minor demanded stations.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

where $p_i = q_i/q$ and $p'_i = q'_i/q$ and q is the total passenger flow over all stations (boardings or alightings, it is the same; i.e. $q = \sum_{i=1}^{n}\sum_{j=i+1}^{n} q_{ij}$). This is because the fractions of flows, p_i and p'_i, hold stationary longer times than the flows (i.e. p's have lower variabilities than q's, and therefore the estimation of $E(N_S)$ is more robust).

And finally, the trip time in this case that the number of stops is not fixed can be expressed as:

$$T = \frac{D}{v} + LE(N_S) + (\tau' + \tau)qh$$

Note that N, the total number of boardings (or alightings) in the whole route is equal to qh.

5. Required vehicles to serve a route

M, the required number of vehicles to serve a route, can be estimated from the total trip time, T, and the service headway, h, as:

$$M = \frac{T + T_0}{h}$$

where T_0 is a terminal layover, necessary to provide the required break time to drivers and to compensate for possible delays during the route. This T_0 should not be shorter that $2\sigma_T$, two times the standard deviation of trip times. This would ensure that the delays of 95% of the trips could be compensated by T_0. Note that T is the total trip tame it takes vehicles to cover the route in one direction. Then, M obtained as in the previous equation is the number of vehicles required to serve one-directional service. If service is provided in two directions, then the previous value needs to be multiplied by two. Also, in real life, the number of vehicles must be a natural number, so that the result of the previous equation should be approximated to the upper integer.

M, obtained as in the previous expression, ensures that one vehicle can be dispatched every h. However, it does not ensure that the vehicles have enough capacity to serve all the demand. In this regard, it needs to be checked that the expected vehicle occupancy in the critical link is smaller than the vehicle capacity.

5.1. Vehicle occupancy in the critical link

O_i, defined as the vehicle occupancy in a particular link i, (see Figure 14), can be obtained from the O/D matrix (see Figure 13) as:

$$O_{i+1} = O_i + N_{b(i)} - N_{a(i)} = O_i + (q_i - q'_i)h$$

Note that the initial occupancy at the terminal can be considered null (i.e. $O_0 = 0$).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

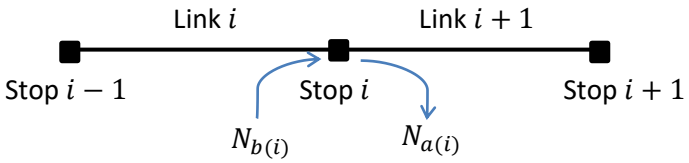


Figure 14. Link definition

Then, the condition to be fulfilled to guarantee enough vehicle capacity is:

$$\max_i O_i \leq C$$

Where C is the vehicle capacity. If this condition is not fulfilled, either vehicles with more capacity must be used, or the headway needs to be reduced. In this last case, the vehicles' capacity would be binding in the selection of h, which would not be determined by the required frequency of service.

There might be situations, where the capacity of the platform at the stops could also be restrictive in order to provide a good level of service. Note that the worst situation would happen if at some instant all the alighting and boarding customers share the platform. Then, it should be ensured that the capacity of the platform of station i is at least $N_{b(i)} + N_{a(i)} = (q_i + q_i')h$. The situation would worsen if the stop is shared by multiple scheduled transportation routes.

6. Stochastic effects on service regularity

As discussed previously, service regularity is an important property in order to ensure the quality of service and minimize user waiting times. However, regularity is rather difficult to maintain, because of the impacts of stochastic effects on scheduled transportation.

Scheduled transportation vehicles suffer random fluctuations in trip times due to dense traffic conditions, red lights at traffic signals, abnormal delays at stops, etc... These create deviations from schedule which affect service regularity and waiting times. However, the most problematic effect with regard to stochastic fluctuations in trip times is the positive feedback effect, which suffer scheduled transportation systems with small headways and a large number of vehicles. This is that trip time perturbations (even small ones), are unstable and tend to amplify as vehicle advances in the route. Then, because small random fluctuations in trip times are unavoidable, vehicles would reach the end of the route with very large deviations from schedule, if nothing is done.

In the context of urban bus systems (i.e. a scheduled transportation system with small headways subject to the positive feedback), this effect is known as the "bus pairing" effect. This phenomenon arises because when a bus suffers a delay, the probability of suffering more delays along the route grows. This is because when the bus falls behind schedule, the headway with the preceding bus grows. This means that at the next stop it is more likely to find more customers to board (i.e. because they have had more time to accumulate at the stop). Recall that the time the vehicle is stopped (i.e. with doors open for boarding and alighting customers) depends on the number of boardings. Therefore, if the vehicle finds more customers than usual at the stop, it will lose some additional time with respect to the average trip time, and it will fall a bit more behind schedule. This means that at each additional stop, the bus will need to invest more and more time to allow for boardings, it will fall more and more behind schedule, its commercial speed will decrease more and more, and the crowded conditions in the bus will

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

deteriorate. In contrast, the bus behind will come closer and closer, because the headway will tend to decrease, as it does the number of customers this following vehicle will find at stops. So, the commercial speed of this bus will tend to increase, and it will travel with very low occupancy. If nothing is done, this feedback process ends with the two buses travelling together, the first one very crowded and the second one almost empty. And what is worst, the headway for the next bus to come is the double of the planned headway, largely affecting waiting times. Figure 15 illustrates this bus pairing phenomenon.

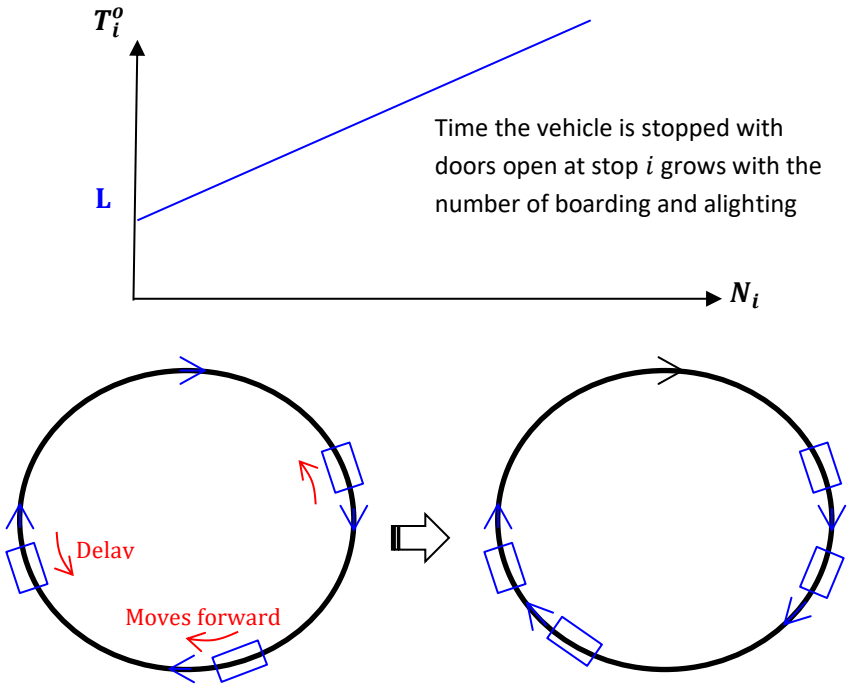


Figure 15. Bus pairing effect

Note that the bus pairing effect will be more important when the demand on the bus route is large, and boarding times represent a significant fraction of the vehicle cycle times. In spite of this, the optimal design of the bus route itself will act as a headway control strategy. Note, that if the demand is large, the optimal headway, h^*, will be shorter, so that the number of passengers boarding and alighting at each stop will be kept bounded. This means that if h^* is implemented, when demand is large and the system is prone to bus pairing, the headways are going to be short so that the fraction of time taken by boardings and alightings will be small with respect to the line-haul and lost times, acting implicitly as a control scheme for the bus headway. In contrast, if for some reason $h \gg h^*$ (e.g. budget limitations implying a limited number of buses on the route, M; consideration of a artificially low passengers' value of time, β, in order to limit the level of service and operational costs of the system; capacity limitation of the corridor and stops which prevent the implementation of short headways, etc.), then, the bus pairing effect will be very important.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

7. Schedule control

Traditionally, bus drivers take their own schedule control measures in order to mitigate bus bunching. For instance, when a driver realizes that he is falling behind schedule and that the following bus is coming close, he typically does not allow boardings at stops. In addition, the following vehicle (which travels ahead of schedule) might try to slow down its pace, by taking more time than strictly necessary at stops, or by not rushing to pass signals in green. With today's technology, many transit agencies have systemized this traditional drivers' control measures. Vehicles are equipped with on-board units (OBUs) equipped with GPS and other communication technologies allowing vehicles to communicate each other and with the infrastructure (e.g. with traffic signals). This technology allows to implement schedule control algorithms which try to maintain a constant headway between buses. This is achieved by trying to speed up vehicles which suffer a random delay (i.e. for instance providing priority at traffic signals by coordinating the green times when this delayed vehicle approaches the intersection) and holding buses which travel ahead of schedule (e.g. by providing information to the driver).

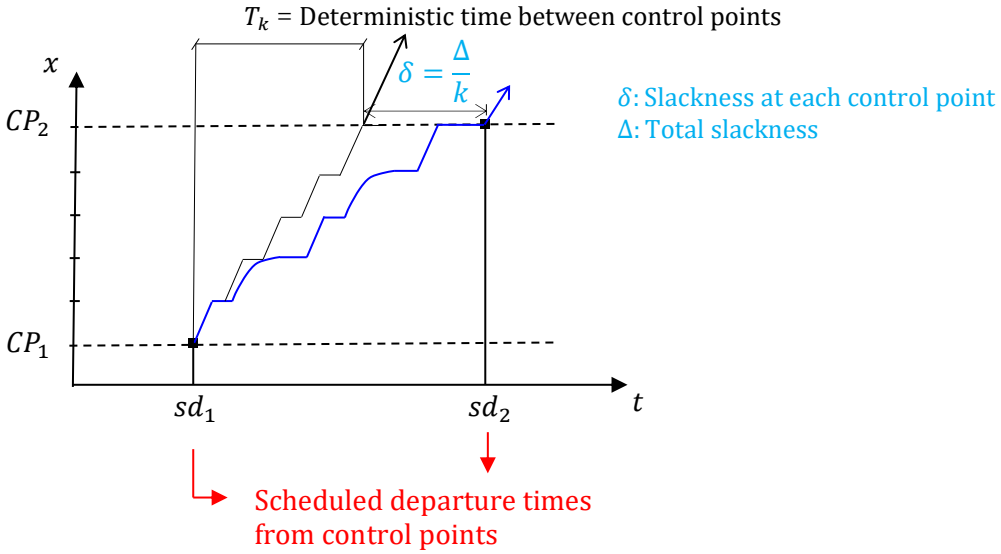


Figure 16. Introduction of slackness between control points.

Furthermore, there exist a traditional schedule control strategy which has been implemented by transit agencies already before the emergence of information and communication technologies and ITS (Intelligent Transportation Systems), in order to avoid the negative effects of stochastic trip time fluctuations and the vehicle pairing phenomenon. This consists in introducing a slack time in the trip times between stops, so that the unavoidable fluctuations could be absorbed without the vehicle falling behind schedule. This is implemented by establishing a route trip time $T_s > T$ (i.e. $T_s - T = \Delta =$ the route slack time) and don't allowing vehicles to depart from certain stops before the scheduled times (to avoid vehicles running ahead of schedule). These stations act as control points to maintain the schedule (see Figure 16). Obviously, the penalty of such strategy is that the trip time increases (i.e. the commercial speed is reduced). We are going to analyze this strategy in more detail.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Consider that the vehicle route is divided into k identical parts between control points. Let T_k be the deterministic (i.e. average travel time without slackness) between two consecutive control points. Then, for selecting δ and k, we will follow the following process:

1. Find an expression for $\delta \approx f(k)$
2. Select k so that it minimizes the total time a typical passenger is in the system (i.e. wait + in vehicle travel time).

7.1. Find an expression for $\delta \approx f(k)$

We want the probability of falling behind schedule to tend to be null:

$$P\{T_k < T_s/k\} \to 1$$

Say with a 95% confidence:

$$E(T_k) + 2\sqrt{Var(T_k)} = T_s/k$$

Operating:

$$T_s - kE(T_k) = 2k\sqrt{Var(T_k)}$$

And because $kE(T_k) = T$ and $T_s - T = \Delta = k\delta$, we obtain:

$$k\delta = 2k\sqrt{Var(T_k)}$$
$$k\delta^2 = 4kVar(T_k)$$

And because $E(T)$ is the sum of k independent components, T_k, then: $kVar(T_k) = Var(T)$. So, we have:

$$k\delta^2 = 4Var(T)$$

And finally:

$$\delta = \sqrt{\frac{4Var(T)}{k}}$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

The variance of the route trip times (i.e. $Var(T)$) can be obtained from observation. Otherwise, it can be estimated from the model of trip times derived previously. Recall that:

$$E(T) = T_f + (2\tau_0)qh$$

where we consider $T_f = \frac{D}{v} + LE(N_S)$ as the traveling time (i.e. with doors closed) and the same unitary boarding and alighting times (i.e. $\tau' = \tau = \tau_0$).

Then, computing the variance of the previous linear expression:

$$Var(T) = \sigma_{T_f}^2 + (2\tau_0)^2 qh$$

Where it is assumed that the total number of customers boarding and alighting during the whole route (i.e. qh) is a Poisson random variable, and therefore the variance is equal to the mean (i.e. this is a property of Poisson distributed random variables).

Expressing the headway as $h = T_s/M$, where M is the number of vehicles operating the route, we obtain:

$$Var(T) = \sigma_{T_f}^2 + (2\tau_0)^2 q \frac{T_s}{M}$$

We do not know T_s. To get rid of this term, we are going to consider that Δ is a small part of T_s, so that we can assume $T_s \approx E(T) \approx E(T_f) + (2\tau_0)q\frac{E(T_f)}{M}$. Then:

$$Var(T) \approx \sigma_{T_f}^2 + 4q\tau_0^2 \left\{\frac{1}{M}\left[E(T_f) + 2q\frac{E(T_f)}{M}\tau_0\right]\right\} =$$

$$= \sigma_{T_f}^2 + 4q\tau_0^2 \frac{E(T_f)}{M}\left\{1 + \frac{1}{M}2q\tau_0\right\} = Constant = \frac{C}{4}$$

So, we have obtained that $Var(T)$ is a constant that we can name $\frac{C}{4}$, so that when plugging this into the previous expression for δ we obtain:

$$\delta = \sqrt{\frac{C}{k}}$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

7.2. Select k so that it minimizes the total time a typical passenger is in the system

We want:

$$Min[Cost\ of\ time\ in\ the\ bus\ +\ Cost\ of\ time\ waiting\ at\ the\ stop]$$

The cost of the additional time in the bus can be seen as the cost of control, while the cost of time waiting at the stop can be seen as the cost of operating without control. This defines a trade-off which allows to determine the optimal level of control which minimizes the sum of both costs.

$$Cost\ of\ time\ in\ the\ bus = C_{T_f} E(T + k\delta) f_T$$

Where C_{T_f} is the unitary monetization factor of the time traveling in the bus and f_T is the fraction of the route travelled by the typical passenger. Considering that the decision variable is δ, this can be expressed as:

$$Cost\ of\ time\ in\ the\ bus = C_{T_f} f_T k\delta + C_{T_f} f_T T = \beta k\delta + constant$$

where $\beta = C_{T_f} f_T$.

In turn, the cost of time waiting at the stop is:

$$Cost\ of\ time\ waiting\ at\ the\ stop = C_W E(W)$$

Recall that we have previously obtained that the variance of the trip time between control points could be expressed as $Var(T_k) = \delta^2/4$. Then, assuming that the passenger demand is uniformly distributed amongst the stops, the average passenger will be waiting at the stop in the middle between control points, so that he will experience half of this variance (see Figure 17).

Therefore, the standard deviation of vehicle lateness, as seen by the typical passenger is $\sqrt{\delta^2/8} \approx \delta/3$

Considering $E(W)$ as one standard deviation of vehicle lateness, we finally obtain:

$$Cost\ of\ time\ waiting\ at\ the\ stop = C_W\ \delta/3 = \alpha\delta$$

where $\alpha = C_W/3$.

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

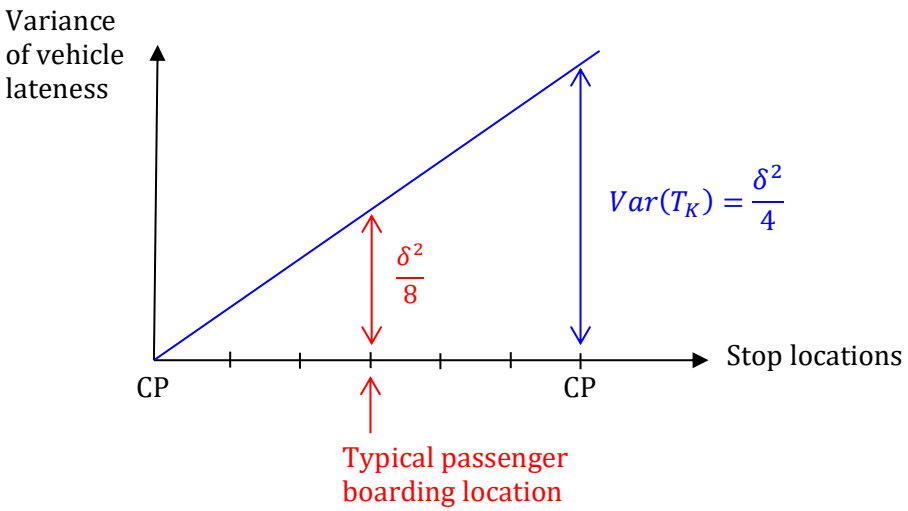


Figure 17. Variance of vehicle lateness.

Then the optimization turns to be:

$$Min[\beta k\delta + \alpha\delta]$$
$$s.t. \quad k\delta^2 = C$$
$$k \leq n$$

Which transforms to:

$$Min\left[\beta\sqrt{Ck} + \alpha\sqrt{\frac{C}{k}}\right]$$

This is a convex function with respect to k and considering the typical parameters, so that the first order condition for the optimization (i.e. derivative with respect to k = 0) yields a minimum. Then the optimal value for k is:

$$k^* = Min\left(\frac{\alpha}{\beta}, n\right) = Min\left(\frac{C_W/3}{C_{T_f} f_T}, n\right)$$

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

and the optimal slack time per control point is $\delta^* = \sqrt{C/k^*}$.

As an example, consider $C_W = 3C_{T_f}$ and $f_T = 1/3$. Then $k^* = 3$ and $\delta^* = \sqrt{C/3}$.

In summary, the concept of the previous headway control strategy is to add slack times along the route in order to absorb the stochastic fluctuations and reduce the variance of the headway, and in consequence the average wait times. Clearly this strategy will increase the average cycle time of the route, meaning that headway control will imply higher operating costs to the agency (i.e. more vehicles and drivers, M, to provide the same average headway, $\bar{h}$). The customers' in-vehicle travel time will also increase, because some customers will need to "wait", inside the vehicle, until the scheduled departure time at control points. This means that it is advisable to set the control points along the route at locations with a low vehicle occupancy. In conclusion, the benefits of the strategy are the reduction of the average wait times, while the penalties are the increase of the in-vehicle trip times and of the agency operating costs. Following from this trade-off, the global benefit of the strategy will be larger if the cost of wait time, w_W, is larger than the cost of in-vehicle travel time w_{IVTT}. Recall that usually $w_W \approx 2w_{IVTT}$. In addition, if the average trip length in the route is short, this will allow stablishing a higher number of control points, as a lower fraction of the demand will need to traverse the control point stop. This will result in a more effective headway control with shorter slack times at control points. Note that these conceptual results are perfectly illustrated by the previous optimization and its result in k^*.

8. Planning of a scheduled transportation system

The planning process of a scheduled transportation system is often articulated in three distinct phases, according to the time horizon to which it refers:

- **Strategic process**: It covers everything related to network design, routes and stops or stations, depending on parameters such as the demand for public transport, technology considered (rail, bus, underground, etc.), their functional characteristics and costs. The planning horizon is often tens of years.
- **Tactical process**: Assuming that the network design is defined in the previous stage, it is based on the determination of the frequencies considering seasonality and variations in demand. The planning horizon is for months or few years.
- **Operational process**: Considering the routes and their corresponding frequencies as fixed, it is mainly about creating a service schedule adjusted to daily demand variations, drivers' labor regulations (shifts, breaks) and other restrictions (maximum mileage, maintenance schedule). It also includes the implementation of strategies to increase the regularity and commercial speed of vehicles.

In this section we will address part of the strategical and tactical planning, by facing the typical problem of determining the optimal stop spacing, s^*, and the optimal headway, h^*, for a given route. Planning consist in finding the combination of these two decision variables that optimize a given objective function. This objective function may present different criteria or goals of the planner: maximizing level of service to customers, minimizing operator costs, maximizing operator benefit, minimizing emissions or a combination of effects to the different stakeholders in the transportation system.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

8.1. Determining the optimal stop spacing, s^*. The customer perspective.

First, let's consider the objective of determining the optimal stop spacing, s, aiming to provide the best service possible to customers. This means that the objective function to minimize will only consider user costs, neglecting the costs incurred by the operating agency in providing such service. To that end, we are going to consider the following context which defines our problem:

- Customers' demand is uniformly distributed along the length of the route. The total demand rate (i.e. per unit time) is q [pax/h].
- Customers access the stops by walking. The customers' walking speed is v_w.
- The distance travelled by the average customer in the route is l (See Figure 18a).
- The cruising speed of public transport vehicles is v. They present a constant deceleration $-a$ in the braking phase when reaching a stop and the same acceleration a (with the opposite sign) when leaving the stop until reaching the cruising speed again.
- The vehicle stops a fixed amount of time at each stop. τ_s is defined as the lost time accelerating / decelerating and opening / closing doors plus the fixed time that doors are open.
- The stop spacing, s, is uniform along the route.

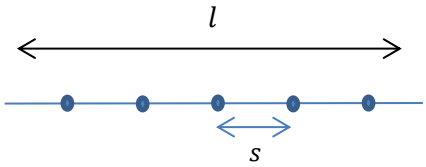


Figure 18. Average trip length l in a uniform stop spacing, s, route.

Average user costs are, Z_U, are defined by the monetization of the average trip time, as:

$$Z_U = \beta_T \cdot T = \beta_T (A + W + IVTT)$$

where β_T is the average value of time considered in a particular society for the planning of transportation systems. Note, that T in the previous expression could be changed by its perceived value, T_p, if one wants to consider the different perceived weights of access, wait and in vehicle travelling times.

In order to minimize Z_U with respect to the decision variable, s, we only need to consider the terms of Z_U which depend on s. All the other terms will play no role in the optimization. Clearly, the access time, A, will depend on on s, and also the in-vehicle travel time, $IVTT$, because it depends on the number of stops which is function of s. In contrast, the wait time W, does not depend on s, and can be neglected in the optimization.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

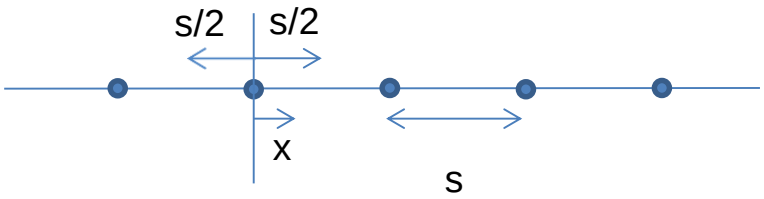


Figure 19. Users' maximum access distance in a route with constant stop spacing, s.

To continue with the optimization process, it is needed to obtain expressions for A and $IVTT$ in terms of s. Considering the previous hypotheses, the maximum walking distance in the access phase from the furthest origin to the stop, is $s/2$ (see Figure 19). The expected value of the access distance $E(x)$ can be calculated using the formula $E(x) = \int_{x_{min}}^{x_{max}} x f(x) dx$, where x is the random variable representing the origin of the trip in the domain $x_{min} \leq x \leq x_{max}$ and $f(x)$ is its probability density function. Considering a uniform distribution of x in the previous domain, $f(x) = \frac{1}{x_{max} - x_{min}}$. In our analytics, $x_{max} = s/2$ and $x_{min} = 0$. Thus, the expected value of the access distance to the nearest stop at the origin of the trip is $E(x) = s/4$. Because A includes the access time (origin) and egress time (destination, which could be estimated analogously), then:

$$A = \frac{2E(x)}{v_w} = 2\left(\frac{s}{4v_w}\right) = \frac{s}{2v_w}$$

Regarding the $IVTT$ for the typical passenger who travels a distance l, can be obtained as:

$$IVTT = \frac{l}{v} + \frac{l}{s}\tau_s$$

Note that the first term of $IVTT$, l/v is the minimum time necessary to cover the distance l without stops (i.e. the linehaul time). This time is constant and can be neglected in the optimization. The second term corresponds to the time lost accelerating / decelerating and boarding / alighting passengers at each stop, τ_s, times the number of stops in the typical trip (i.e. l/s).

Then, the mathematical problem to solve in order to obtain the optimal spacing is:

$$\min_s \left[\left(\frac{s}{2v_w} + \frac{l}{s}\tau_s\right)\right]$$
$$s.t. \;\; s > 0$$

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

It can be seen that the two components in the objective function have opposite effects with respect to the decision variable s: the access cost would increase with s, while the time at stops would decrease. This trade-off in the objective function already unveils an important result. Note that one could think in a policy of reducing the stop spacing to improve user accessibility. But this decision, cannot obviate the detrimental effect to the in-vehicle travel time due to the increase in the number of stops. The overall result could represent an important reduction of the commercial speed leading to an uncompetitive system in terms of the global travel time. Also, recall that we are only considering user costs in the optimization of the stop spacing, and even in this case a trade-off exists. This means that, shortening the stop spacing is not always a good option. Even with all the money of the world to invest in the collective transportation system, a critical spacing exists, so that setting a spacing below this value increases both, users and agency costs[5].

Therefore, the optimal spacing should be selected precisely solving the previous mathematical problem. To find the optimal value of the spacing, s^*, it is needed to take the derivative of the objective function with respect to s and equalize to zero. Because the objective function is convex, solving for s this equation will yield a minimum in the user costs. The solution for the s^* is:

$$s^* = \sqrt{2l\tau_s v_w}$$

As an example, it is proposed to analyze the optimum spacing of a route with the following input parameters: $v = 30\ km/h$, $l = 5\ km$, $v_w = 2.5\ km/h$ and $\tau_s = 40\ s$. Figure 20 shows the two components of the objective function in terms of s. The optimal spacing, s^*, can be identified as the spacing for which the sum of both terms is minimum. The robustness of the optimal solution is seen, since the objective function is practically flat around the optimal value. This means that, even if the planning agency selects a stop spacing slightly higher or lower than the optimal, the increase in the objective function will be negligible. From Figure 20, you can verify that the optimal value of the spacing is within the range of $s = 0.4 - 0.6\ km$, which corresponds to the optimal value obtained applying the previous equation, yielding a value of $s^* = 0.527\ km$.

[5] Note that agency costs are reduced with s, because of less infrastructure costs and reduced route trip times, leading to less vehicles.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

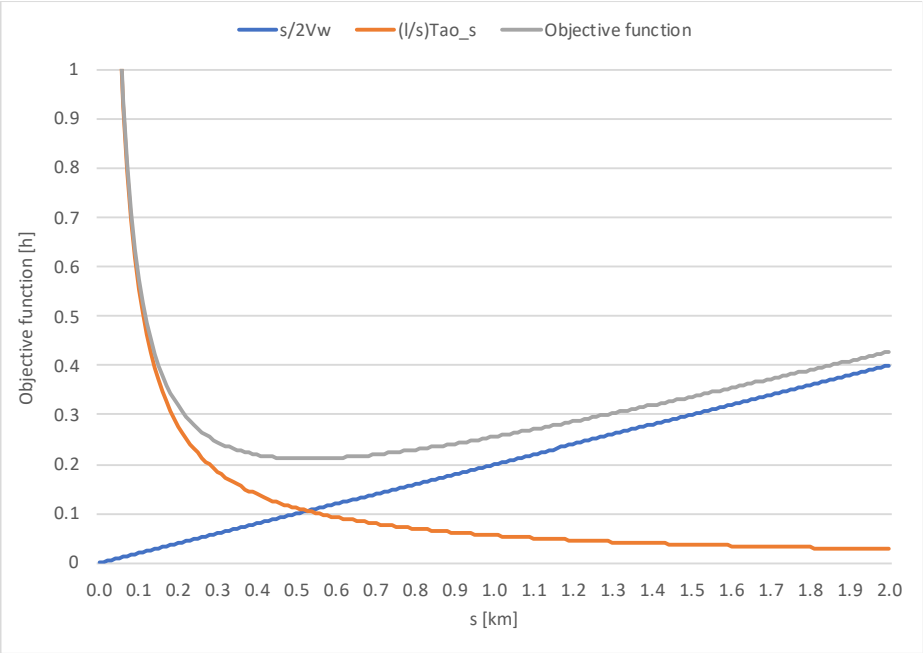


Figure 20. Optimization of the stop spacing, s.

8.2. Determining the optimal headway, h^*, and the optimal spacing, s^*. The global perspective.

In the previous section the optimal spacing has been obtained by minimizing the user's costs. This sets the critical spacing, below which the user's costs grow. This minimum threshold should never be get passed, even when the agency costs of the system do not matter. However, the cost of providing the scheduled transportation service generally matters, and this affects the determination of the optimal design.

In spite of this, the inclusion of agency costs affects little the optimal spacing, which is mainly determined by the trade-off between access and lost times in the user's costs. The effect of including the agency cost in the optimization could imply obtaining somehow larger optimal spacings. In contrast, the optimization of agency costs is essential in the determination of the optimal headway, which is determined by the trade-off between the agency costs (number of vehicles and distance traveled) and user costs (wait).

In the present section, a global perspective is considered in order to determine the optimal headway and spacing. This means that the objective function to minimize will be the total costs of the system, Z, obtained as the sum of the agency, Z_A, and the user, Z_U, costs.

$$\min_{h,s} Z = Z_A + Z_U$$

$$s.t. \quad h \geq 0;\ s \geq 0$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

The agency's operating costs per hour, Z_A, depend on the distance travelled by the vehicle fleet in one hour of service, V [vehicle·km] and on the number of vehicles used to provide the service, M. The unitary monetary costs for these concepts will be $€_V$ [euros/veh·km] and $€_M$ [euros/veh·h]. To continue with the optimization, it is necessary to obtain V and M in terms of the decision variables, h and s.

Because one vehicle needs to arrive to each stop every h, globally, the whole length of the route, $2D$ (i.e. considering bidirectional service), needs to be travelled collectively by all the vehicles every h. Therefore, the total length travelled by all vehicles per unit time can be expressed as:

$$V = \frac{2D}{h}$$

In turn, M is determined by the route trip time[6], $2T$ (i.e. round trip of one vehicle to cover the $2D$ distance), divided by the headway. This is:

$$M = \frac{2T}{h} = \frac{2D}{v_c}\frac{1}{h} = \frac{V}{v_c}$$

where V_c is the commercial speed (i.e. including stops):

$$v_c = \frac{D}{T} = \frac{D}{\frac{D}{v} + \frac{D}{s}\tau_s} = \left[\frac{1}{v} + \frac{\tau_s}{s}\right]^{-1}$$

Not rounding M to the upper integer when estimating the agency operative costs is not critical, as the error made with the continuous approximation is not significant if M is large.

Then, agency costs can be expressed in terms of the decision variables as:

$$Z_A = €_V \frac{2D}{h} + €_M \frac{2D}{v_c h} \quad [€/h]$$

[6] The layover time at the terminals is neglected in order to simplify the calculations.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Regarding the user costs, they are estimated by the monetization of the total trip time. This is

$$Z_U = 2q\beta_T(A + W + IVTT) \quad [€/h]$$

where β_T is the average value of time, and A, W and $IVTT$ are the average access, wait and in-vehicle travel times. Different weighting factors for each component might be used if one wants to consider the different perceptions of each component of the total travel time. Note that the cost for one individual user is multiplied by the route's demand rate (i.e. $2q$, considering both directions) to obtain the total users' cost per unit time.

Then, recall that:

$$A = \frac{s}{2v_w}$$

$$W = \frac{h}{2}$$

$$IVTT = \frac{l}{v} + \frac{l}{s}\tau_s$$

Note that A considers the average access time at the origin and at the destination of the trip. Both access times penalize users in a similar fashion. In contrast, W only considers the wait at the origin, and the egress time at destination is neglected. This is because the perceived penalty of egress times is much lower than the wait time at the origin, and even lower than the perception of $IVTT$ (i.e. weighting factor <1). This is because the egress time at destination can still be used to perform some activity.

Then the total user cost is expressed by the following equation:

$$Z_U = 2q\beta_T\left(\frac{s}{2v_w} + \frac{h}{2} + \frac{l}{v} + \frac{l}{s}\tau_s\right)$$

And the mathematical problem to be solved to find the optimal headway, h^*, and stop spacing, s^*, is:

$$\min_{s,H}\left\{Z = €_V\frac{2D}{h} + €_M 2D\left[\frac{1}{vh} + \frac{\tau_s}{sh}\right] + 2q\beta_T\left[\frac{s}{2v_w} + \frac{h}{2} + \frac{l}{v} + \frac{l}{s}\tau_s\right]\right\}$$

$$s.t. \quad s \geq 0; H \geq 0$$

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

which, neglecting the term τ_s/sh, yields the following solution[7]:

$$h^* = \sqrt{\frac{2D\left(€_V + \frac{€_M}{v}\right)}{q\beta_T}}$$

$$s^* = \sqrt{2l\tau_s v_w}$$

and the minimum cost, $Z^* = Z(h^*, s^*)$, is:

$$Z^* = 2\sqrt{2Dq\beta_T\left(€_V + \frac{€_M}{v}\right)} + 2q\beta_T\left(\sqrt{\frac{2l\tau_s}{v_w}} + \frac{l}{v}\right) + \begin{cases} 0 \quad (lower\ bound) \\ \dfrac{€_M D\tau_s\left(q\beta_T\right)^{1/2}}{\left[l\tau_s v_w D\left(€_V + \frac{€_M}{v}\right)\right]^{1/2}} \quad (upper\ bound) \end{cases}$$

where the upper bound in Z^* is obtained by adding the term neglected in the optimization, τ_s/sh, considering h^* and s^*.

Finally, it is only needed to check that the optimal headway h^*, allows carrying all the demand, given a vehicular capacity C. The average vehicle occupancy and be determined by using the Little Equation in queuing theory, which states that the rate at which customers are served is equal to the total number of customers in the system (i.e. accumulation) divided by the average time in the system. In the context of the present problem, this means that:

$$Total\ customers\ in\ the\ system = 2q\frac{l}{v_c}$$

where $2q$ represents the service rate and l/v_c the average time in the system.

Because this total number of customers are travelling in M vehicles, the average vehicle occupancy is:

$$\bar{O} = \frac{Total\ customers\ in\ the\ system}{M} = \frac{2q\frac{l}{v_c}}{\frac{2D}{v_c h}} = \frac{qlh}{D}$$

[7] Such simplification is necessary in order to obtain a simple explicit analytical solution. If necessary, the complete problem could be solved using numerical optimization methods.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

And the condition to be met to fulfill the vehicle capacity restriction is:

$$C \geq \frac{qlh^*}{D}$$

As an application example, consider the planning of a bus line that should serve a target demand $q = 500\ pax/h$ per direction, with the following input parameters: $v = 30\ km/h$, $l = 5\ km$, $v_w = 2.5\ km/h$, $\tau_s = 40\ s$, $\beta_T = 14\ €/h$, $D = 10\ km$, $€_V = 2\ €/veh \cdot km$, $€_M = 40\ €/veh \cdot h$ and vehicle capacity $C = 75\ pax/veh$ (standard bus)**.** Figure 21 shows the value of the objective function, Z as a function of h and s, and without considering the capacity restriction. It can be identified that the domain of Z between 6000 and 7000 €/h (blue color) contains the optimal solution, which for this problem is $Z^* = 6810\ €/h$, corresponding to a value of $s^* = 550\ m$ and $h^* = 6.5\ min$. The approximate solution using the analytical expressions provided above would have been $s^* = 527\ m$ and $h^* = 5.9\ min$, with Z^* between 6651 and 6823 €/h.

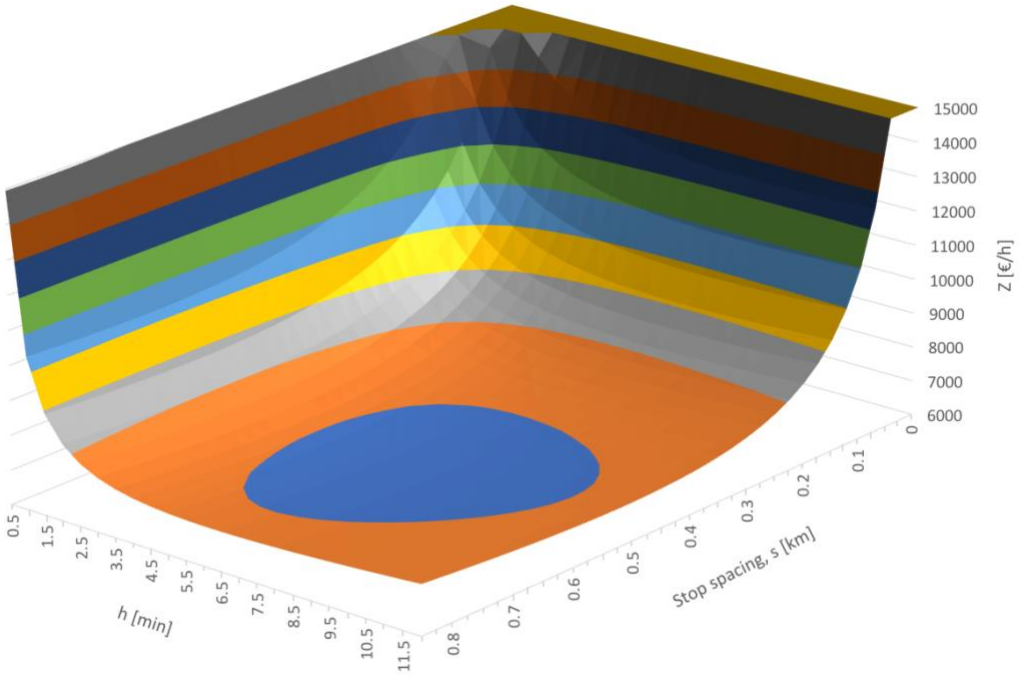


Figure 21. Z as a function of h and s.

Like before, the solution is robust, because around the optimal values, the system exhibits a large domain of (s, h) where the objective function (i.e. the total costs) are practically constant. It can also be seen in Figure 21 that the selection of an adequate stop spacing is probably the most critical decision. This is because if s is too small, the total costs can grow largely. Also, the stop spacing is a strategical decision that, once implemented, it

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

is very difficult to modify, as it involves modifying the whole stop infrastructure. In contrast the sensitivity of total costs to h is lower, unless for very low unfeasible headways.

Finally, it needs to be ensured that the vehicle capacity restriction is met. Considering the previous equation, the maximum headway fulfilling the capacity restriction is $h \leq CD/ql = 18\ min$. Therefore, the optimal solution perfectly complies with the capacity restriction, presenting an average occupancy of $\bar{O} = 27\ pax/veh$.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Frequent asked questions (FAQs)

Q: For the estimation of $E(W|\tau)$, the expected wait time, W, given an arrival time at the stop, τ, in systems with large headways (Figure 7 of the notes), is the transportation vehicle coming at $t = 0$? What is the functional shape of $E(W|\tau)$?

A: $t = 0$ is the scheduled departure time of the vehicle from the stop. But because the actual departure time is somehow random (in the presented example is uniform between $t = -k$ and $t = k$), there is some probability of early or late departure. If you work out the expected wait of a customer arriving at the stop at the instant τ (this is $E(W|\tau)$), given the previous probability distribution of the actual vehicle departure (as it is explained in detail in the lecture notes), you will see that if τ is between $(-k, k)$ the function describing the expected wait is concave with respect to τ. if τ is between $(k, h - k)$ (i.e. $h - k$ is the next scheduled departure, as h refers to the scheduled headway) the expected wait evolves linearly with respect to τ. Just check the expressions obtained in the notes.

Q: Regarding synchronized transfers at stations, I am not sure to understand what synchronizing the headways actually means.

A: If headways are synchronized, when the customer arrives at a transfer stop, the next vehicle is ready to depart, so that there is no additional wait at the transfer. This is achieved by minimizing the red area in Figure 9, as discussed in the notes.

Q: In the equation for the route travel time T, (Page 14 of the lecture notes) how do you know if those variables are independent of the considered stop i?

A: We do not know, but we assume that they are. And the assumption makes sense. Because there is no reason to think that the cruising speed, the lost time accelerating / decelerating and the unitary boarding / alighting times depend on which part of the route you are considering. It is true that the one for the cruising speed is more disputable.

Q: If the number of stops is not fixed you need to estimate the expected number of stops, $E(N_S)$ (formula in Page 18 of the notes) in order to obtain the route trip time. How do you calculate $E(N_S)$?

A: $E(N_S)$ is the expected number of stops, which is not fixed. We assume that the vehicle only stops if there is boarding or alighting demand to serve. This is how we obtain an expression for $E(N_S)$. The calculation process is detailed in the lecture notes.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

7-TRANSPORTATION DEMAND MODELING

Table of Contents

1. The 6 steps of (transportation) planning 2
2. Demand modeling: endogeneity problem & equilibrium solution 3
3. Demand models: specification, estimation and prediction 5
4. The four-step model - Urban Transportation Planning (UTP) procedure 8
5. Utility theory 12
6. Demographics and aggregation 14
   6.1. Aggregation by integration 15
   6.2. Aggregation by classification 16
7. Multinomial choice 17
8. Elasticity of demand 19
9. Utility theory: summary, problems & solutions 20
   9.1. Trick 1: Smart use of socio-economic (SEC) variables 21
   9.2. Trick 2: Random utility models (RUM) 23
10. Random utility models 23
   10.1. The Probit Model 25
   10.2. The Binary Logit Model 26
11. Maximum likelihood estimation in the Logit Model 28
12. Example of application: Estimation of the binary Logit via MLE (from individual choice data) 29
13. Estimation of the binary Logit via OLS (from aggregated choice data) 31
14. Properties of Logit models 33
   14.1. Elasticity of Logit model 35
   14.2. Logit Independence from Irrelevant Alternatives (IIA) 35
15. Nested Logit 37
16. Concluding remarks 39

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

1. The 6 steps of (transportation) planning

If this "Transportation Engineering" course was to be divided in two parts, this would be the break point. With this new topic in the course, we will change the type of questions that we want to answer, and this will also imply a change in the tools and models that we are going to use. These are going to be based in statistics and probability theory, so that we are going to shift from graphical tools to a more mathematically based approach. Because this chapter represents a break point in the course, in this introduction it is important to gain some perspective of the global picture in order to figure out where we are.

The planning process of any transportation infrastructure or service (or in more general terms of any facility) typically goes through different stages until it becomes a reality. These are shown in the simplified flowchart of Figure 1. All the contents of this course can fit into step 4: Technical evaluations. So, you can see that we are only facing a small part of the global problem of planning a transportation infrastructure or service. If you continue your studies on Transportation Engineering, you will find courses addressing each one of these stages.

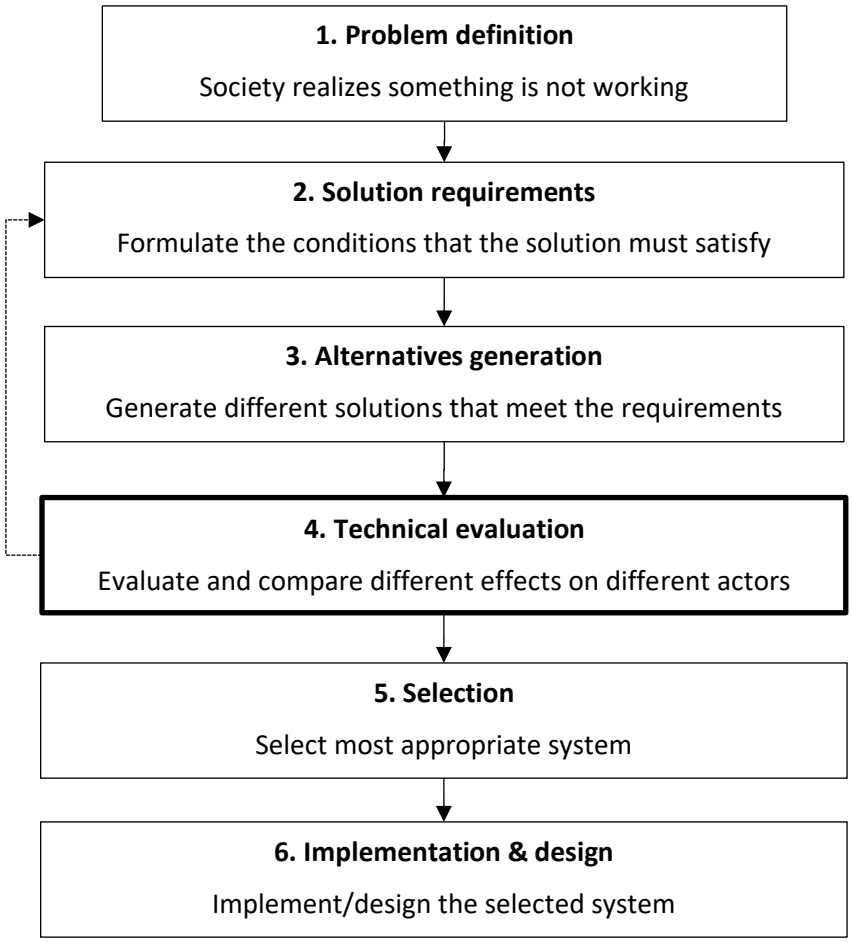


Figure 1. The planning process flow-chart

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

In order to perform technical evaluations, we need to work with models. Models are an abstraction of reality on which we can perform experiments at a lower cost than if we were to perform them in real life. Figure 2 categorizes the different types of models. Full-scale models represent demonstration or trial projects, are expensive and may imply important social costs. Therefore, they should be developed only at the last stage of the planning process. The benefit of demo projects is that there are no model prediction errors, and only (small) measurement errors must be considered.

Idealized models are needed in the early stages of the planning projects. In transportation engineering, physical scaled models are not feasible, because we deal with elements which is not easy to build to scale (e.g. human beings). Therefore, we must deal with mathematical models, either with the traditional analytical approaches, or with the more recent simulation environments resulting from the increase and availability of computational power.

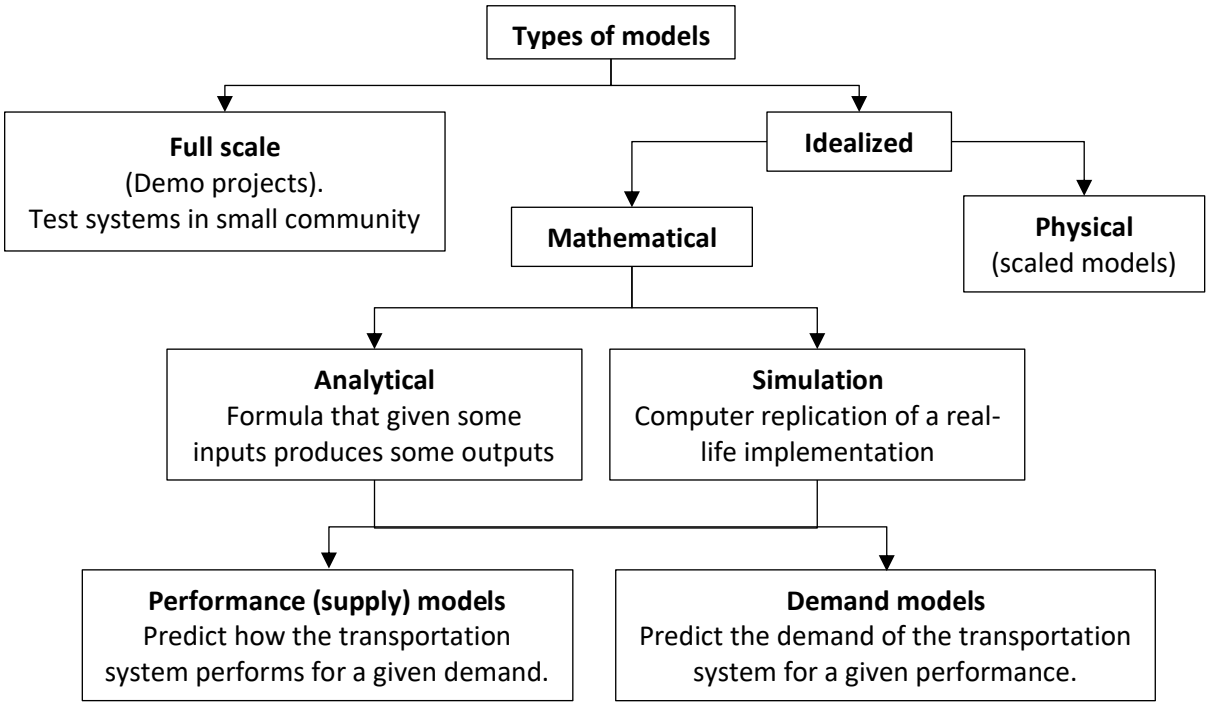


Figure 2. Different types of models

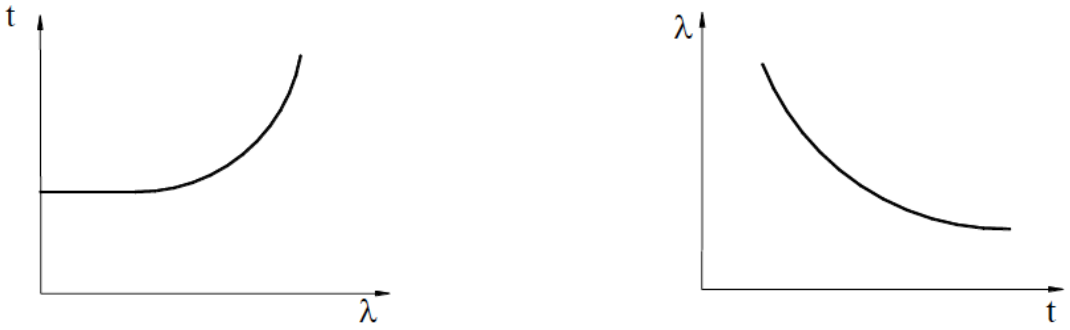
2. Demand modeling: endogeneity problem & equilibrium solution

Mathematical models in the analysis of the transportation system could be divided in two broad categories: *i)* performance (or supply) models, and *ii)* Demand models (see Figure 3).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >



a) Performance model b) Demand model

Figure 3. Performance vs Demand models

On the one hand, Performance Models yield the performance of the transportation system as a function of a number of attributes, including demand (i.e. $Performance = f(demand, \mathbb{Y})$, where $\mathbb{Y}$ is a vector of other attributes). Performance models are used to evaluate the level of service provided by a transportation system (e.g. the average travel time, t, as a function of the demand rate, λ: $t(\lambda)$) and are useful to answer questions like: how long would a trip take? or, how does this change with the load of the system? These are the type of models we have been working on so far in the course. On the other hand, Demand Models yield the demand of the transportation system as a function of its performance and other relevant variables (i.e. $Demand = f(performance, \mathbb{X})$, where $\mathbb{X}$ is a vector of other relevant variables affecting demand). Demand models are used to predict the number of users of a transportation system (e.g. the number of users given the level of service offered, e.g. $\lambda(t)$) and are useful to answer questions like: how much demand will this system have? or, how would different factors (e.g. level of service, socio-economic environment) affect the demand of the system?

In the context of the mathematical modeling of the transportation system, using performance and demand models, one difficulty is the endogeneity of the global problem. This is that the performance of the system depends on the demand (as described by the performance model), and at the same time the demand depends on the performance (as described by the demand model). For example, $t(\lambda)$ and at the same time $\lambda(t)$ (see Figure 4). Endogeneity problems must be addressed recursively, until an equilibrium solution is reached. This is to select an initial solution (e.g. an initial tentative demand) and predict the performance of the system for this demand using the performance model. Then, return to the demand model to see if the demand corresponding to this performance, matches the initial solution. If not, a new demand solution is considered and the whole process repeated. The recursive process ends when an equilibrium solution is reached, matching the demand and performance models at the same time. This recursive process is similar to the choice process that individuals follow when a new (transportation) alternative is available. First, some demand appears (which maybe somehow random, depending on the marketing and appealing of the new alternative). After some days, customers may get an idea of the level of service offered, and depending on this the new transportation system could attract or discourage more demand. The process continues until the equilibrium is reached. This process in real implementations might take weeks or months. That's why the analysis of the performance of new implementations cannon be done during the first operative days.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

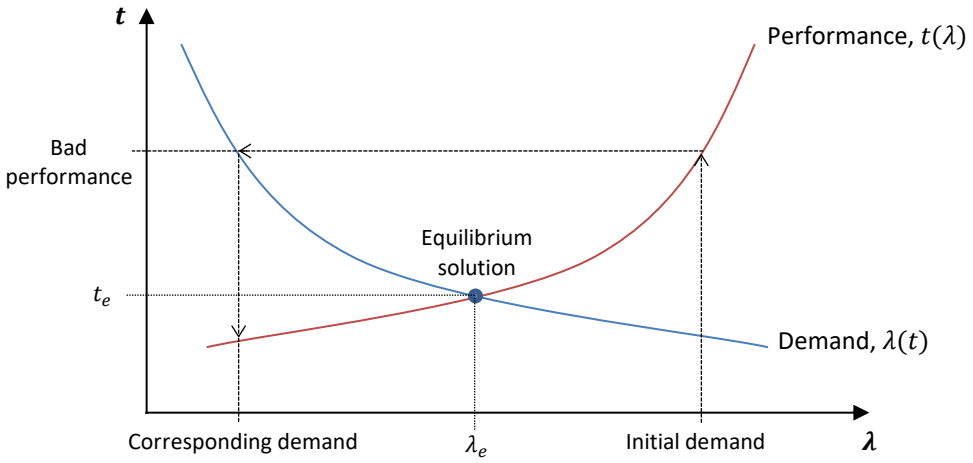


Figure 4. Recursive process and equilibrium solution for an endogenous problem

It needs to be considered that, in general, demand models are much less robust and accurate than performance models. This is because demand models strongly rely on human behavior, which is difficult to predict and requires input data which are difficult to obtain. Large errors (e.g. 20-50%) in good demand modeling approaches should not be strange. Imagine the possible errors if the demand modeling is not taken seriously enough.

3. Demand models: specification, estimation and prediction

The construction of a demand model involves three stages: specification, estimation and prediction.

Specification of the model consists in defining its mathematical form and the variables and parameters involved. Demand models are highly dependent on human behavior and choice, and they may depend on a large number of factors (i.e. variables). The variables affecting the demand for a transportation system can be classified into two broad groups:

- Level of Service (LOS) factors: Travel time, cost (fare), existence of other transportation alternatives, customer service, comfort, ...
- Socio-economic Characteristics (SEC): Total population, number of drivers in the population, employment numbers, age, average income, ...

Both groups of factors affect the demand of a transportation system and are important in the model.

As an example, consider a demand model for a DRT[1] (Demand Responsive Transport) option, which could be specified according to the following functional form:

$$\lambda = (P - H)\frac{\gamma}{1 + \alpha T + \beta F} + H\frac{\gamma'}{1 + \alpha' T + \beta' F}$$

[1] A DRT-Demand Responsive Transport system is a general term used to describe on-demand transportation options, where the user ask for a service (e.g. taxi system).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Where P represents the total population, H the driving population, T the average trip time using the DRT and F the average fare. Note that the demand for the system λ is reduced with the increase of T or F, but these variables might affect differently the driving H, and non-driving $(P-H)$, populations, as described by the different model parameters, (α,β,γ) versus (α',β',γ'). So, the functional relationship of the demand model can be expressed as:

$$\lambda = f(\underbrace{\overbrace{t}^{\text{Explanatory variables}}}_{\text{LOS}}, \underbrace{F}_{\text{SEC}}, \underbrace{P, H}_{\text{SEC}}, \overbrace{\alpha,\beta,\gamma,\alpha',\beta',\gamma'}^{\text{Parameters}})$$

Where:

$t, F \Rightarrow$ LOS variables. Resulting from our design of the transportation system and its demand. F is a policy variable that we can control, and it could be even be decided by contract. t is a somehow endogenous LOS variable (i.e. depends on λ)

$P, H \Rightarrow$ SEC variables. Known as they can be measured.

$\alpha,\beta,\gamma,\alpha',\beta',\gamma' \Rightarrow$ constants without prespecified values (i.e. parameters). They need to be estimated.

Estimation consists in determining the parameters' values. This needs to be done using statistical methods applied to multiple observations of the demand (i.e. the independent variable) together with the explanatory variables.

For the DRT example, parameters $(\alpha,\beta,\gamma,\alpha',\beta',\gamma')$ could be estimated from the following observations in M different cities:

Table 1. Observations for the estimation process

Observation #	Dependent variable (λ)	Explanatory variables
1	$\lambda^{(1)}$	$t^{(1)}, F^{(1)}, \ldots$
2	$\lambda^{(2)}$	$t^{(2)}, F^{(2)}, \ldots$
...	...	...
M	$\lambda^{(M)}$	$t^{(M)}, F^{(M)}, \ldots$

From the database illustrated in Table 1, we seek that the predicted demand, $D^{(m)}$, applying the model to any city m, is approximately equal to the observed demand, $\lambda^{(m)}$. This is:

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$D^{(m)} = D\left(t^{(m)}, F^{(m)}, \dots \alpha, \beta, \gamma, \alpha', \beta', \gamma'\right) \approx \lambda^{(m)} \qquad \forall m$$

If we could achieve this for all m and with small errors, we would have obtained a good model. This sets the condition for the estimation of the parameters, which could be obtained using, for instance, least square error regression. This is to select the parameters $(\alpha, \beta, \gamma, \alpha', \beta', \gamma')$ so that they minimize the sum of the squares of the errors. Specifically:

$$Min \sum_{m=1}^{M} \left(\lambda^{(m)} - D^{(m)}\right)^2$$

This type of estimation using regression is particularly easy if the specification of the model is linear with respect to the explanatory variables. Otherwise, it could be trickier. So, it is a good practice to try to linearize the model before the estimation of parameters. For instance, consider the previous specification of the demand model for the DRT system, assuming $H = 0$:

$$\lambda = \frac{P\gamma}{1 + \alpha T + \beta F}$$

Note that the model is not linear with respect to the explanatory variables, but we could write:

$$\frac{P}{\lambda} = \frac{1}{\gamma} + \left(\frac{\alpha}{\gamma}\right) t + \left(\frac{\beta}{\gamma}\right) F$$

And renaming the parameters as $\theta_0 = 1/\gamma$, $\theta_1 = \alpha/\gamma$ and $\theta_2 = \beta/\gamma$, we obtain a linear model with respect to the explanatory variables:

$$\frac{P}{\lambda} = \theta_0 + \theta_1 t + \theta_2 F$$

Considering this linearized specification of the model and applying multiple linear OLS regression to the observations (i.e. those of Table 1) the estimations of the parameters $\hat{\theta}_0$, $\hat{\theta}_1$ and $\hat{\theta}_2$ can be obtained. Then, it is only needed to undo the linearization to obtain a model for predicting the demand, λ.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$\lambda \approx \frac{P}{\hat{\theta}_0 + \hat{\theta}_1 t + \hat{\theta}_2 F}$$

This model, with the estimated parameters is ready to be applied for demand prediction. This is simply done by applying the model with predicted explanatory variables.

Finally note that it might take hundreds of observations to obtain robust estimations of parameters. The more the parameters, the more the observations needed to achieve robust estimations. This is why it is especially important to keep the number of parameters low.

4. The four-step model - Urban Transportation Planning (UTP) procedure

As discussed in the previous pages, demand for transportation systems depends on multiple factors, including the attributes of the transportation network and available services and the regional socio-economic characteristics. Furthermore, the development and application of demand models require a process of specification, estimation and prediction, which requires a significant amount of people's behavioral data.

In this context, and in an effort to systemize the demand modeling procedure and data requirements in transportation planning, in the 1950's was developed the Urban Transportation Planning (UTP) procedure. This consisted in the sequential application of four steps, resembling the human decision-making procedure when deciding a trip. Since then, the UTP process has become a standard for regional transportation planning, and it is frequently referred to as the "four step model" for transportation demand.

Land-use forecasting and population growth starts the process. Typically, forecasts are made for the region as a whole. Such forecasts provide control totals for the local analysis. Next, the region is divided into zones and by trend or regression analysis, the population, employment and other socio-economic characteristics are determined for each.

Given this zonification of the analysis region, the four steps of the classical UTP model can be applied. These are:

1. Trip generation
2. Trip distribution
3. Mode choice
4. Route assignment

Trip generation determines the frequency of origins or destinations of trips in each zone by trip purpose, as a function of land uses and household demographics, and other socio-economic factors like motorization or the household income level. The matching of Origin - Destination (O-D) pairs is not an objective of this first step.

Trip generation is usually divided into two aspects: the first referring to the trip production generated by an area (origins) and the second referring to its trip attraction (destinations). Also, trip generation and attraction for each zone are usually divided in mandatory trips (e.g. home-based trips whose purpose is travelling to work or study) and optional trips (i.e. non-mandatory, e.g. shopping trips, social or managerial, recreational or leisure trips and others). Note that the modeling of non-mandatory trips, with a lower recurrence, is more complex than the rest. Another component of trip generation is freight distribution, where the typical explanatory variables are the number of companies, employees and sales of a given zone. Detailed trip generation studies may even classify trips according to whether they are generated during peak or off-peak periods.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Traditionally, three types of models have been used for trip generation analysis. The first is the growth factor method, where it is postulated that the number of trips generated in the zone i in the future, G_i^f, will be proportional to the product of the growth factor of the zone, F_i, times the current number of trips generated, G_i^c. This is:

$$G_i^f = F_i \cdot G_i^c$$

Where the growth factor, F_i, is obtained from the quotient of the predicted future values (f) of the considered explanatory variables over their current value (c):

$$F_i = \frac{P_i^f I_i^f H_i^f}{P_i^c I_i^c H_i^c}$$

Where for example, P_i could be the population of zone i, I_i the household average income, and H_i the driving population or motorization.

The second typical model specification of trip generation is the multiple linear regression model, where the trips generated are considered to be a linear function of a series of transportation network attributes and socio-economic variables. These models are calibrated by applying OLS linear regression techniques to available data of trip demand as exemplified previously with the example of the demand model for the DRT system. So, the trip generation model could be expressed as:

$$G_i = a_i + \mathbb{b}_i \mathbb{X}_i$$

Where $\mathbb{X}_i$ is the vector of explanatory variables, and $a_i, \mathbb{b}_i$ a vector of parameters to be estimated.

The regression model can be stablished in aggregated terms (i.e. a single regression for each zone, then i in the previous equation would represent a zone), or in disaggregated terms (typically by household, which then will be represented by i). Disaggregate models are more detailed, could be more precise and could identify intra-zone variations. However, their calibration requires more data, and more detailed (i.e. at the household level). Regression models are also typically used to estimate the demand generation of particular facilities or infrastructures (e.g. hospitals, airports, railway stations, big shopping malls, amusement parks, etc.). Calibrated regressions for multiple types of facilities are provided in the famous Trip Generation Manual, published since the 1980's by the American Institute of Transportation Engineers (ITE) (www.ite.org). In any case, this type of regression models relies in the linear relationship between trip generation and explanatory variables (or a linearization of it), which not always can be established with confidence.

Once the regression equations have been established, the model can be applied by assuming that the model parameters will remain constant in the future. Thus, by knowing the future values of the explanatory variables, the future number of trips can be predicted. Results of disaggregate models can be aggregated for each zone in order to obtain the equivalent aggregated result.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

The last usual method used in the trip generation step is the so called cross-classification technique. In this method households are stratified according to a series of common socio-economic characteristics (e.g. motorization level, average household income, household size, etc.), thus creating homogeneous groups. Then, the average number of trips generated by each household type is estimated from the observed data. It is assumed that for each household category the rate of trip generation does not vary over time, which allows to make predictions in the trip generation for each zone if knowing the composition of the zone in terms of the defined household types. This method avoids the assumptions of linearity and additivity of terms with respect to the explanatory variables, but in contrast it requires a large amount of more detailed data than in previous analyzes, both in the estimation process and in the prediction phase. A large number of household surveys are necessary to obtain reliable results.

The second step in the UTP framework is trip distribution. Trip distribution matches origins with destinations, thus creating O-D pairs and determining the O-D matrix. So, imagine that we are considering a zone i, with a trip generation G_i, with a potential number of destinations, j, for these trips (see Figure 5).

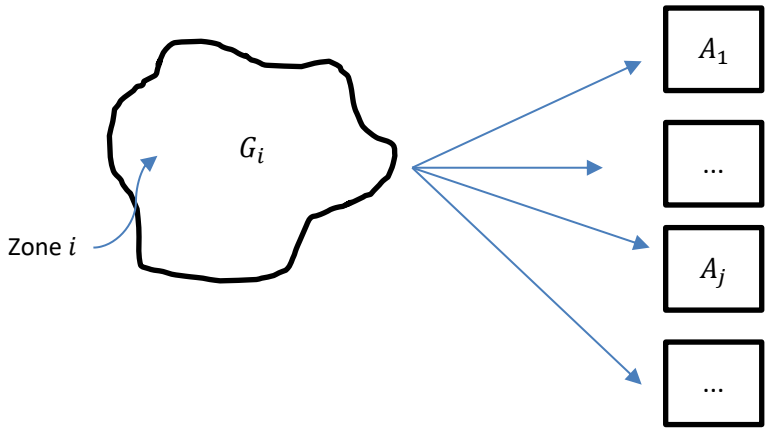


Figure 5. Step 2 - Trip distribution

Each choice j, has a level of attractiveness, A_j, which should be proportional to the number of "opportunities" in zone j, and modeled in terms of some explanatory variables like in the trip generation step. Note that if A_j actually represents the number of attracted trips, this will imply some additional difficulties in the aggregation process, as it is going to be discussed later. Also, each destination j, has an impedance, a cost of traveling to such destination. This cost depends on the configuration of the transportation network and of the available services (i.e. the transportation supply) which steer the travel time and out of pocket monetary cost for traveling to destination j. All these costs (i.e. the impedance) for travelling from i to j, are going to be expressed as t_{ij}.

Then, the trip demand between zones i and j, λ_{ij} is usually expressed as:

$$\lambda_{ij} = \gamma G_i A_j f(t_{ij})$$

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Where f should be a decreasing function often of the Cobb-Douglas type:

$$f(t_{ij}) = t_{ij}^{\alpha} e^{\beta t_{ij}} \quad \alpha, \beta \leq 0$$

One common example for this kind of function is to consider $\alpha = -2$ and $\beta = 0$, not for any particular reason, but because many researchers have found that with this estimation they obtain good results, yielding the following trip distribution model:

$$\lambda_{ij} = \frac{\gamma G_i A_j}{t_{ij}^2}$$

Where the trips between zones i and j are proportional to their respective generation and attraction and inversely proportional to the square of the impedance of the trip. This trip generation model is famously known as the gravitational model, for its resemblance to the Newton's law of universal gravitation.

Because the number of trips generated in the origin zone i is fixed G_i, the demand for one of the destination choices j will affect the demand for the rest.

$$\sum_j \lambda_{ij} = G_i$$

And this determines the value for the proportionality parameter γ:

$$\sum_j \frac{\gamma A_j}{t_{ij}^2} = 1 \implies \gamma = \frac{1}{\sum_j \frac{A_j}{t_{ij}^2}} = \frac{1}{\sum_j A_j f(t_{ij})}$$

This trip distribution model allows to obtain the whole O-D matrix λ_{ij} for the region under analysis. Note that the sum for each row i will yield the total trip generation G_i of the zone. However, if A_j has been estimated as the trip attraction for zone j, the method does not ensure that the sum of the columns in the O-D matrix yields this trip attraction. In such case, it can be applied some kind of matrix rebalancing algorithm to approximate the sum of the columns and rows to total generation and attraction rates while keeping the proportionality between the values in the matrix. Typical iterative proportional fitting algorithms are the Fratar or Furness method.

The third step in the UTP process is mode choice, where it is determined the proportion of trips between each origin and destination that use a particular transportation mode. This is usually modeled using utility theory, which is going to be analyzed in much more detail in the next section.

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Finally, the last fourth step is route assignment, which allocates trips between an origin and destination by a particular mode to a route. Often (for highway route assignment at least) Wardrop's principle of user equilibrium is applied (equivalent to a Nash equilibrium), wherein each driver (or group) chooses the shortest (travel time) path, subject to every other driver doing the same. The difficulty is that travel times are a function of demand, while demand is a function of travel time, the so-called bi-level problem. The route assignment in transportation networks is going to be addressed in the next chapter of the course.

5. Utility theory

Demand for transportation service depends on human decisions. This represents a huge conceptual difference respect many other fields of engineering and science which rely on robust physical laws. Utility theory constitutes a mathematically consistent and convenient theory to model human behavior. It assumes rational human behavior, and despite this is not always true, utility theory is the best we have.

Utility theory is based on three properties of the rational human being. Specifically, it is assumed that humans are:

- Greedy (consumption motivated) => people only care about their own well-being.
- Analytic => people are smart and know how to compare choices
- Well informed => people know all publicly available information

Utility theory is based on the first of the properties, considering that people try to maximize their satisfaction/utility, or equivalently, to minimize their generalized costs/disutility. The last two assumptions, although might not very realistic in complex situations, make possible to apply relatively simple math to rate preferences between different alternatives.

The fundamental idea behind utility theory can be exemplified with a simple example. Imagine that one afternoon you are feeling bad and deciding whether going or not to the doctor's office for a health checkup. You value the health checkup at 100€ per visit, because probably you are going to feel better before. Although you do not need to pay out-of-pocket for the doctor's visit, it takes a one-hour car ride plus a toll of 7€ to reach the doctor's office. Since time is money, you also value an hour of your afternoon at approximately 15€. According to utility theory, you as a rational human would choose to visit the doctor's office for a health checkup, as this implies a net benefit of 78 €. Another day, you consider the idea of going to the park, which is located next to the doctor's office, for recreation. Your value for going to the park is 10€. The decision in this case is not traveling to the park, as the net benefit is negative (i.e. -12€).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

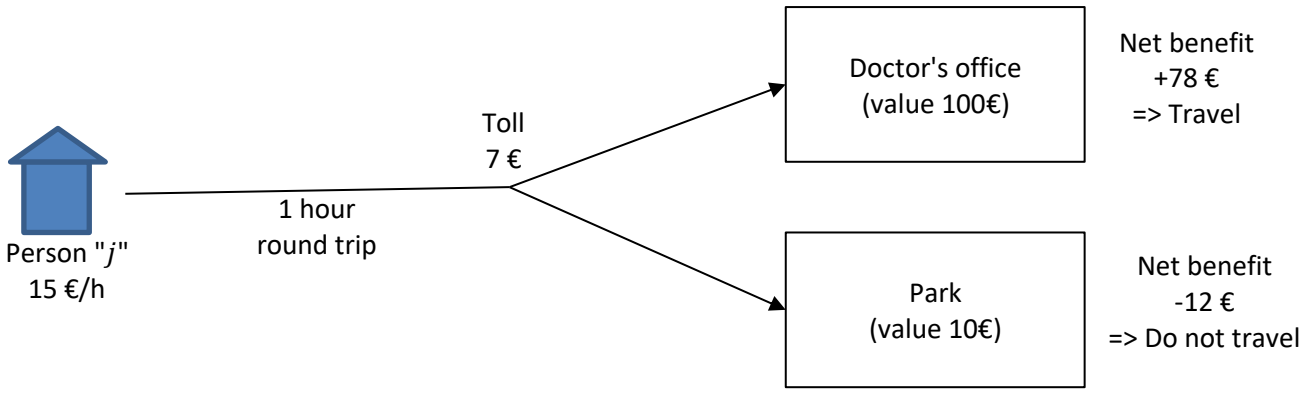


Figure 6. Travel decision according to utility theory

For both choices, the "generalized cost" (defined as the total costs including the monetary costs – gas, transit fare, etc. – and the monetization of non-monetary costs – time, discomfort on the bus, etc.) is the same. In this example, the generalized cost is simply the travel time cost + toll cost. Subtracting this cost from the benefit to be derived from each activity, we find the net benefit or the utility of that activity.

This conceptual framework can be systematized formulating the concept of disaggregate (i.e. individual) utility. Utility is a mathematical expression for preferences. Utility functions describe the likings of individuals and contain explanatory variables (EVs) and parameters (which are sometimes called "taste" constants). EVs are objective, and are available for observation. They can be of two types: *i)* level-of-service (LOS) such as travel time, fare of transit, etc., and *ii)* socio-economic characteristic (SEC) such as income, car ownership rates, etc. On the other hand, parameters are subjective and not observable; hence, they need to be estimated. Since individuals behave and value things very differently, EVs and parameters easily differ from person to person.

Specifically, the notation used in the formulation of the utility function, $\mathcal{U}_j$, for the previous example and for a person j, would be:

$$\mathcal{U}_j = a_j - (b_j t_j + F_j)$$

Where:

a_j = Value of the activity (or value for reaching the destination) – a taste constant

b_j = Value of one hour of time (hourly) – a taste constant

t_j = Total travel time (in hours) – an explanatory variable

F_j = Parking fare (for the whole trip) – an explanatory variable

In the above, remember that $\mathcal{U}_j$ is likely going to be different for a different choice of activity (or destination). To make a final decision, person j compares the utility he derives from different activities, and the activity

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

associated with the greater utility is chosen in the end. For the previous example $a_j = 100€$ for going to the doctor and $a_j = 10€$ for going to the park, $b_j = 15\ €/h$, $t_j = 1\ h$, and $F_j = 7\ €$. Note that the units of the toll are €, the same as the global utility. In this case, a taste constant is not needed, because 1 € is 1€, and it is not needed to monetize the explanatory variable. Also note that, in the previous example, b_j is the same for all options (e.g. going to the doctor or going to the park). This is called a generic specification of the explanatory variable.

In more general cases, explanatory variables other than t_j and F_j may be specified for the utility function. If we denote $\mathbb{X}_j$ as the vector of EVs used to estimate the utility function, so that $\mathbb{X}_j = \{t_j, F_j, other\ EV\ that\ matter\ for\ the\ individual\ j\ decision\}$, then we can write the utility function in more general terms as $\mathcal{U}_j = U_J(\mathbb{X}_j)$. The specification of utility functions, U_J, is usually selected to be linear with the parameters, as in the previous example. This does not mean that it needs to be linear in the attributes. For instance $ln(F_j)$ or F_j/I_j, where I_j is the individual's income, are perfectly valid explanatory variables. Categorical variables could also be used as explanatory variables.

6. Demographics and aggregation

Utility theory is in the foundations of disaggregate demand modelling. Figure 7 illustrates the demand modeling and prediction process using disaggregate models.

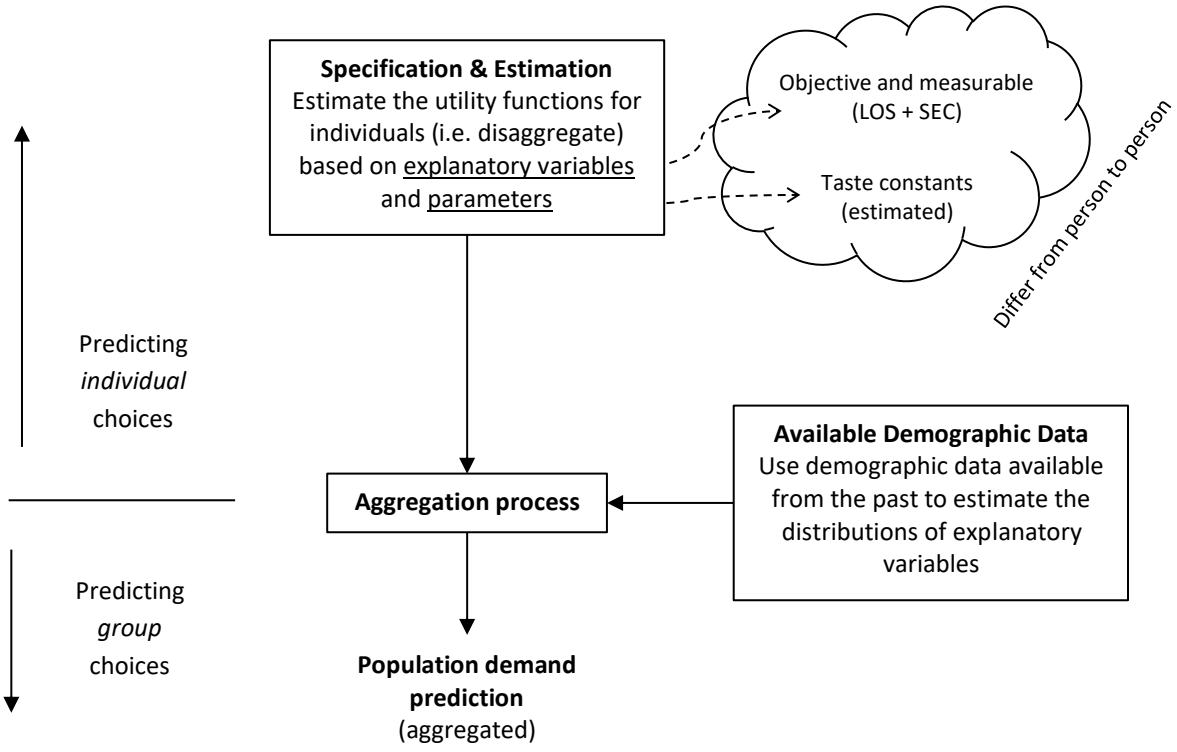


Figure 7. Simplified flow chart describing the demand modeling / prediction process

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Two types of aggregation processes can be applied to disaggregate demand models: *i)* aggregation by integration, *ii)* aggregation by classification.

6.1. Aggregation by integration

Suppose you survey a population of size P in order to determine the value a of an activity for each individual, obtaining $P(a)$, the number of people with a value of the activity larger than a, as: (see Figure 8)

$$P(a) = \begin{cases} P & if\ a < k+1 \\ \dfrac{P}{a-k} & if\ a \geq k+1 \end{cases}$$

Where k is some constant parameter. As a increases the number of people that values the activity more than a decreases as one would expect.

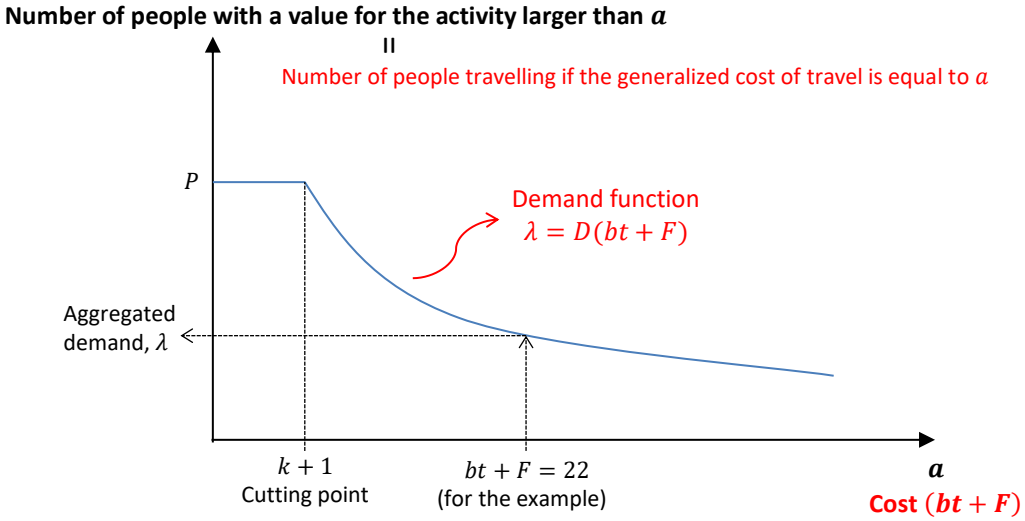


Figure 8. Aggregation by integration

Then according to utility theory $P(a)$ would be equal to the total number of people traveling if the generalized cost of travel was equal to a (i.e. positive utility). This means that the aggregated demand for the activity, λ, function of the generalized cost of the travel, and of the distribution of value a across the population P considered would be:

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$\lambda = D(bt + F) = \frac{P}{a - k} \qquad for\ a = bt + F \geq k + 1$$

and working with the algebra we can obtain:

$$\lambda = \frac{P}{-k + bt + F} = \frac{P/(-k)}{1 + (b/-k)t + (1/-k)F}$$

Which has the same form as the DRT example we saw previously, with $\alpha = (b/-k)$ and $\gamma = (1/-k)$:

$$\lambda = \frac{P\gamma}{1 + \alpha t + \gamma F}$$

6.2. Aggregation by classification

The assumption implicitly made in the previous aggregation by integration was that the population P was nearly homogeneous, as everybody had the same characteristics (e.g. value of time...), and the only thing that varied was the value they obtain from the activity, a.

If there is a deeper knowledge of the demographics and characteristics of society, it is possible to define classes j representing homogeneous groups of people – everybody within one group values the activity almost the same, has similar tastes and is exposed to the same EVs. The number of people in each class is denoted by n_j.

Then it is possible to find a classification curve, as in Figure 9. Note that in this simple example, we just use a_j as the criterion for classification.

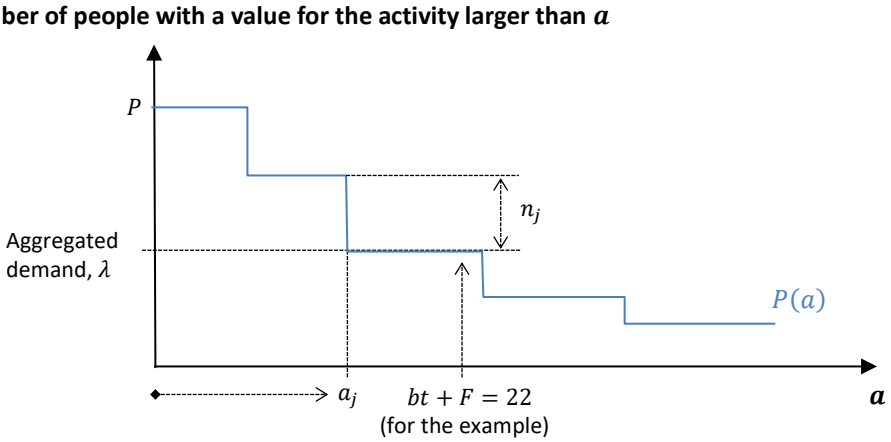


Figure 9. Aggregation by classification

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Then, the total aggregated demand, λ, is obtained as:

$$\lambda = \sum_j n_j \Delta(\mathcal{U}_j)$$

Where Δ is a heaviside function so that:

$$\Delta(\mathcal{U}_j) = \begin{cases} 1 & if\ \mathcal{U}_j \geq 0 \\ 0 & if\ \mathcal{U}_j < 0 \end{cases}$$

7. Multinomial choice

Utility theory is particularly used in multinomial choice. "Multi" refers to more that two, and "nomial" to categorical choices (i.e. discrete choice). As an example consider that groups j of people face 3 alternatives (i.e. $i = 1, 2, 3$). Note that j refers to an individual or group of individuals, and i to a particular choice alternative.

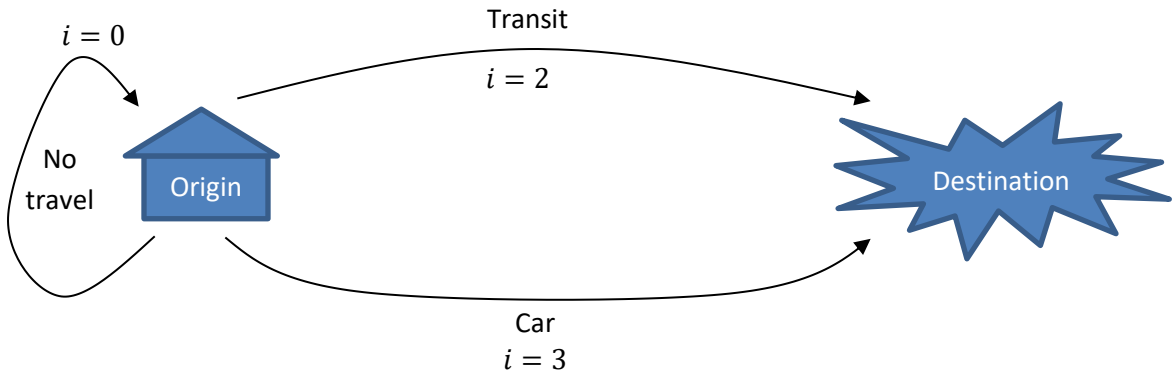


Figure 10. Multinomial choice

The decision to be made is whether to travel or not to an activity and which transportation mode use (car or transit). Then the considered alternatives are: $(i = 1) \implies$ Do not travel; $(i = 2) \implies$ Travel by transit (e.g. bus); $(i = 3) \implies$ Travel by car. And the corresponding utility functions for each alternative are:

$$\mathcal{U}_{1j} = 0$$

$$\mathcal{U}_{2j} = a_j - b_j t_{2j} - F_{2j}$$

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$\mathcal{U}_{3j} = a_j - d_j t_{3j} - F_{3j}$$

Note that people get the same benefit from the activity, a_j, independently of the travel mode used. Also note, that the specification of the model accepts that people value differently their travel time if travelling by car or by transit (i.e. b_j can be different than d_j).

We can write the 3 utilities with a single formula, named the joint utility function, as:

$$\mathcal{U}_{ij} = U(\mathbb{X}_{ij}, \Theta_j)$$

Where $\Theta_j = (a_j, b_j, d_j)$ is the vector of parameters and $\mathbb{X}_{ij} = (\delta_j, t_{2j}, t_{3j}, F_{2j}, F_{3j})$ is the vector of explanatory variables. We need the dummy variable δ_j to construct the joint utility function, where:

$$\delta_j = \begin{cases} 1 & \text{if individual } j \text{ reaches the destination} \\ 0 & \text{otherwise} \end{cases}$$

Note that the taste parameters Θ_j depend only on the persons, j, while alternatives are distinguished by their "attributes", another word for level-of-service explanatory variables, $\mathbb{X}_{ij}$.

Then, for each alternative choice we have:

$$\mathbb{X}_{1j} = (0,0,0,0,0)$$
$$\mathbb{X}_{2j} = (1, t_{2j}, 0, F_{2j}, 0)$$
$$\mathbb{X}_{3j} = (1,0, t_{3j}, 0, F_{3j})$$

and the joint utility function would be:

$$\mathcal{U}_{ij} = U(\mathbb{X}_{ij}, \Theta_j) = a_j \delta_j - b_j t_{2j} - d_j t_{3j} - F_{2j} - F_{3j}$$

Since we assume that everyone is a utility-maximizer and chooses the alternative with the largest utility, we can formulate the general multinomial aggregation formula as:

$$\lambda_i = \sum_j n_j \, \Delta_{(i)[U(\mathbb{X}_{1j},\Theta_j), U(\mathbb{X}_{2j},\Theta_j), U(\mathbb{X}_{3j},\Theta_j)]}$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Where:

$$\Delta_{(i)} = \begin{cases} 1 & \text{if utility of alternative } i \text{ is the largest (i.e. selected)} \\ 0 & \text{otherwise} \end{cases}$$

The previous multinomial demand model allows predicting the demand for each alternative i. As an example, we could plot the demand for each alternative as a function of the transit fare: $\lambda_i(F_2)$ (see Figure 11). You can see that reduced bus fares attracts demand from two sources: *i)* some people that previously was travelling by car (mode shift) *ii)* some people that previously were not travelling and that now (with reduced bus fares) decide to travel. This is the concept of induced demand.

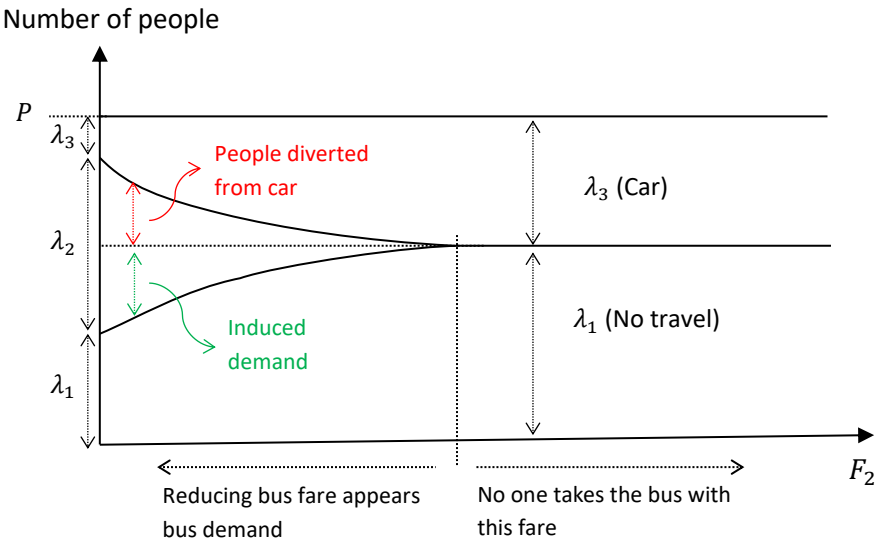


Figure 11. Multinomial demand function

8. Elasticity of demand

The concept of demand elasticity, ε, represents the quantitative effects in demand when we change a policy variable just a little bit. Mathematically, this can be expressed as $\Delta\lambda_i/\Delta X$, or for infinitesimal changes, $\partial\lambda_i/\partial X$, where X represents any explanatory variable. The problem is that this expression might have very weird units. For instance, if we take the fare as the explanatory variable $X = F$, then the units would be $[pax/h€]$. For better comparison we try to get rid of dimensions, and instead of comparing small absolute changes we compare fractional changes. The new question is: what is the percentage change in demand if we change a policy variable by a certain small percentage? The mathematical expression now would be:

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$\varepsilon = \frac{fractional\ change\ in\ \lambda_i}{fractional\ change\ in\ X} = \frac{\Delta\lambda_i/\lambda_i}{\Delta X/X}$$

This concept is called the elasticity of demand with respect to attribute X, and for infinitesimal changes becomes:

$$\varepsilon = \frac{\partial\lambda_i/\lambda_i}{\partial X/X} = \frac{\partial\lambda_i}{\partial X}\frac{X}{\lambda_i} = \frac{\partial ln\lambda_i}{\partial X}X$$

Economists care on how $|\varepsilon|$ compares to 1.

- $|\varepsilon| > 1 \implies$ Elastic demand. Small changes in the attributes produces more than proportional changes in the output (i.e. in the demand).
- $|\varepsilon| < 1 \implies$ Inelastic demand. Small changes in the attributes produces even smaller changes in the output (i.e. in the demand).

Transportation and in particular demand for public transportation tends to be inelastic.

Another interesting concept related to the demand elasticity is the direct or cross-elasticity.

For instance, the direct elasticity (or simply the elasticity) represents the elasticity of the demand of one alternative with respect to some attribute of the same choice alternative. This is:

$$\varepsilon_{(\lambda_i, X_i)} = \frac{\partial\lambda_i}{\partial X_i}\frac{X_i}{\lambda_i}$$

In contrast, the cross-elasticity represents how the demand of one alternative changes as a result of a small change in the attributes of another alternative. For instance how the demand of travelling by car changes due to small changes in the fare of the bus, in the previous example. This cross-elasticity concept is formulated as:

$$\varepsilon_{(\lambda_i, X_j)} = \frac{\partial\lambda_i}{\partial X_{i'}}\frac{X_{i'}}{\lambda_i} \qquad i \neq i'$$

9. Utility theory: summary, problems & solutions

Figure 12 summarizes with a flow chart the processes involved in the construction and application a demand model based on utility theory in order to predict the aggregated demand for a given transportation alternative.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

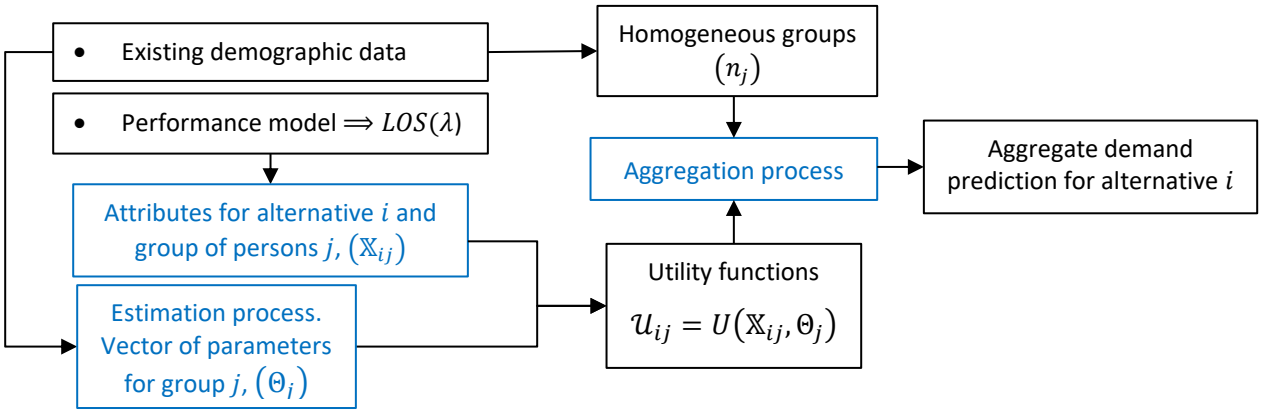


Figure 12. Flow chart describing the demand modeling / prediction process using utility theory

The main caveats in the previous demand modeling framework come from the complexity of human behavior. Specifically, humans are complex in the sense that:

- Inconsistency of decisions: humans do not always make the same decisions given the same choices. And they might be affected by the choices of others (e.g., in fashions).
- Complex choices: people might not be perfectly "rational" in the sense of utility theory if they face complex choices (e.g., job or apartment hunting).
- Diversity: everyone is different.

This last issue is a serious practical problem because it implies that we need to determine a huge number of j taste constants (i.e. one per person.)

In order to face these problems, the specification of utility models have evolved in two ways (tricks):

9.1. Trick 1: Smart use of socio-economic (SEC) variables

The specification of the utility function we have been considering so far is of the form:

$$\mathcal{U}_{ij} = U(\mathbb{X}_{ij}, \Theta_j)$$

where $\mathbb{X}_{ij}$ are level-of-service (LOS) variables of the alternative i if selected by individual (or group of individuals) j, and Θ_j are taste parameters for group j. Instead, we could think of introducing SEC variables in $\mathbb{X}_{ij}$ so that taste parameters could be normalized and applicable to all groups. This is:

$$\mathcal{U}_{ij} = U(\mathbb{X}_{ij}, \Theta)$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

where $\mathbb{X}_{ij}$ include LOS and SEC variables.

Recall the going to the doctor example, with two possible choices:

- Alternative 1: No travel. $\mathcal{U}_{1j} = 0$
- Alternative 2: Travel to the doctor office. $\mathcal{U}_{2j} = a_j - \left(b_j t_j + F_j\right)$

Note that b_j, the value of time, could be different for every individual (or group) j, leading to multiple parameters to be estimated. Instead, we could postulate that b_j is linearly varying with S_j, the average income of individual j (i.e. an observable SEC variable), so that:

$$b_j = b + b' S_j$$

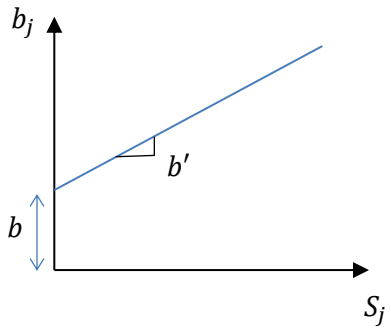


Figure 13. Introducing SEC variables

And then the utility for alternative 2 would be:

$$\mathcal{U}_{2j} = a_j - \left(b t_j + b' t_j S_j + F_j\right)$$

With this new specification of the model we have changed b_j (one parameter for every j) by only two parameters (i.e. b and b'). This can represent a significant improvement in the estimation process and in the reduction of the amount of required data.

Socio-economic variables are less helpful with a_j (the activity value) since this parameter depends on what people want to do at the destination, which is not very related to observable SEC variables. So, as we have seen previously in the aggregation procedure we model a distribution of people's activity values.

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

9.2. Trick 2: Random utility models (RUM)

In order to address the inconsistency and chance of irrationality in human decisions one trick is to introduce a random error in the utility function. The introduction of random effects in utility theory leads to the so called Random Utility Models. In these models, we do not really predict people decisions, but just the chances of each decision (not the actual outcomes, but probabilities). RUM are going to be addressed in more detail in the next session.

10. Random utility models

Random utility models are probabilistic discrete choice models, that yield the probability of one individual (or group of individuals) of making an actual choice. They are based in utility theory, and consist in introducing an error term in the utility function.

Attention: In the following sections I am changing the notation. i and j will represent different choice alternatives, while the decision maker will be represented by subscript n.

So, the utility function, $\mathcal{U}_{in}$, for individual n choosing alternative i in a RUM is represented by:

$$\mathcal{U}_{in} = V_{in} + \varepsilon_{in}$$

Where V_{in} is called the measured utility (or systematic utility, or average utility, all are equivalent names). It represents the part of the utility which is deterministic and observable (e.g. in the previous example $V_{in} = a_n - (bt_i + b't_iS_n + F_i)$), and ε_{in} is the random error term. This random error is the unobservable part of the utility, and it will allow considering not-rational decisions and accepting errors of the model.

We are interested in determining the choice probability. This is determining the probability that a random selected person selects a given alternative, instead of determining actual choices. Let's consider binomial choice (i.e. the selection between two alternatives, i and j). Then, P_{in}, the probability that individual n selects alternative i can be formulated as:

$$P_{in} = P(\mathcal{U}_{in} \geq \mathcal{U}_{jn}) = P(V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}) = P(V_{in} - V_{jn} \geq \varepsilon_{jn} - \varepsilon_{in}) = P(\varepsilon_{jn} - \varepsilon_{in} \leq V_{in} - V_{jn})$$

If we define $\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in}$ and $V_n = V_{in} - V_{jn}$, then:

$$P_{in} = P(\varepsilon_n \leq V_n) = F_\varepsilon(V_n)$$

Where $F_\varepsilon(V_n)$ is the cumulative probability distribution function (cdf) for the random variable ε evaluated at V_n.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

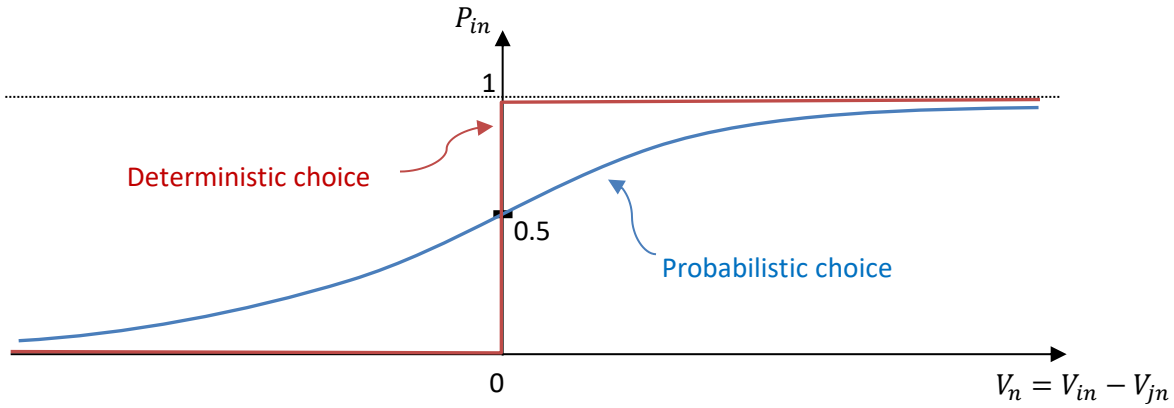


Figure 14. Probabilistic choice

Note (see Figure 14) that the deterministic choice is a particular case where the probability of choosing one alternative is either zero or one, thus not allowing for non-rational decisions. In contrast, probabilistic choice allows for non-zero probabilities for both alternatives. Note that if the difference between systematic utilities of both alternatives is equal to zero (i.e. $V_n = V_{in} - V_{jn} = 0$) then they are equiprobable (i.e. same probability of 0.5). This is because it is generally assumed that the error term, ε, is centered at zero.

Given this framework of random utility models, any particular model will be determined by the probability distribution of the random error ε. Two models are of particular interest. The Probit Model, where it is assumed that ε is normally distributed; and the Logit Model, where the error term follows an extreme value distribution.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

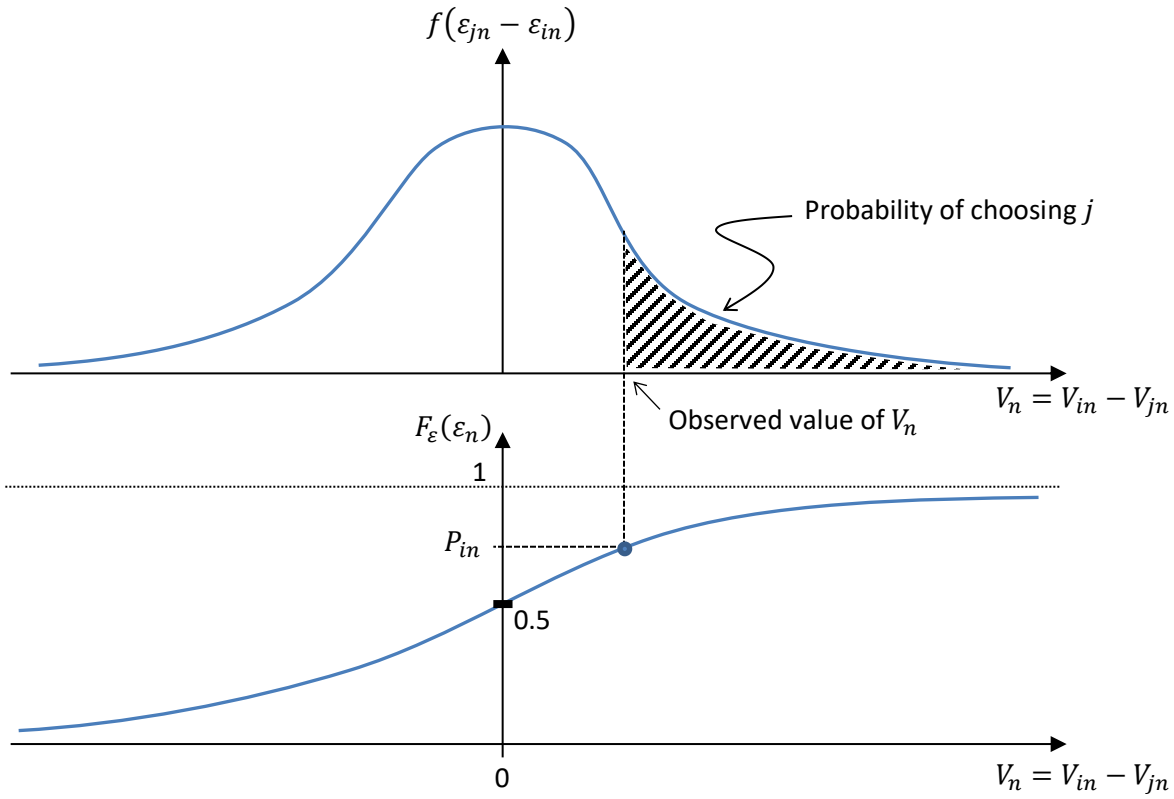


Figure 15. Probabilistic choice and probability distribution of the error term

In Figure 15, note that $V_{in} > V_{jn}$, so the deterministic choice would be always alternative i. However, in the probabilistic choice there is some probability of choosing alternative j, accounting for non-rational decisions, or unobservable utility components.

10.1. The Probit Model

In the Probit model it is assumed that, ε_{in}, the random error in the utility function for one alternative i and individual n, is normally distributed, centered at zero with a variance σ_i^2. So, considering the binomial choice between alternatives i and j:

$$\left.\begin{matrix}\varepsilon_{in} \sim N(0, \sigma_i^2) \\ \varepsilon_{jn} \sim N(0, \sigma_j^2)\end{matrix}\right\} \Rightarrow \varepsilon_{jn} - \varepsilon_{in} = \varepsilon_n \sim N(0, \sigma^2)$$

Where $\sigma^2 = \sigma_i^2 + \sigma_j^2 - 2\sigma_{ij}$ ($\sigma_{ij} \neq 0$ only if there exist correlation between alternatives).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

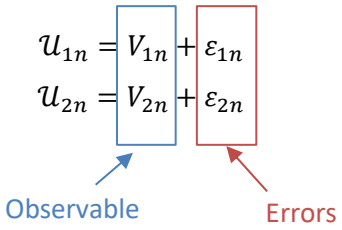
Then:

$$P_{in} = P\left(\varepsilon_{jn} - \varepsilon_{in} \leq V_{in} - V_{jn}\right) = P(\varepsilon \leq V_n) = F_{\varepsilon}(V_n) = \int_{-\infty}^{V_n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\varepsilon}{\sigma}\right)^2} \partial\varepsilon$$

Where F_{ε_n} is the cumulative distribution function for the standard normal probability distribution. The traditional problem in the application of the Probit model has been that there is no closed form solution for the cdf of the standard normal, so that it cannot be derived an analytical form for the model. This implies that the application of the model turns to be somehow computationally complex, although with today's computers it is easily applied, even in combination of other models. In spite of this, this difficulty gave rise to Logit Models, which were formulated with the original objective of obtaining a model similar to the Probit, but with a simple analytical expression. The Logit model turned out to be adequate, mathematically convenient, and have become extremely popular.

10.2. The Binary Logit Model

The Logit model is a particular case of random utility models, so that the utility is defined in terms of a systematic observable part and a random unobservable error. For the binary choice between alternatives 1 and 2, and for individual n, this is:



In the Logit model, the random errors are assumed to follow an extreme value probability distribution:

$$\left.\begin{matrix} \varepsilon_{1n} \sim Extreme\ Value(0,\mu) \\ \varepsilon_{2n} \sim Extreme\ Value(0,\mu) \end{matrix}\right\} \Longrightarrow \varepsilon_{2n} - \varepsilon_{1n} = \varepsilon_n \sim Logistic(0,\mu)$$

The extreme value distribution is a skewed distribution with 2 parameters: Centrality (similar to the mean) and spread (proportional to the inverse of the variance). It is assumed a centrality equal to zero and spread μ. If ε_{in} are independent identically (i.i.d.) extreme value distributed random errors, then the difference between them follows a Logistic distribution with the same parameters (0, μ). The Logistic distribution is very similar to the Normal (with somehow flatter tails), but what is important, with different analytical equations.

The cumulative distribution function for a $Logistic(0,\mu)$ is:

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$F_\varepsilon(\varepsilon_n) = \frac{1}{1 + e^{-\mu\varepsilon_n}}$$

Then the Logit model is formulated as:

$$P_{1n} = P(\varepsilon_{2n} - \varepsilon_{1n} \leq V_{1n} - V_{2n}) = \ P(\varepsilon \leq V_n) = F_\varepsilon(V_n) = \frac{1}{1 + e^{-\mu V_n}}$$

Recall that the parameter μ is inversely proportional to the variance of the distribution. Then, if $\mu \to \infty$, the variance would tend to zero and we would have a deterministic choice model. In contrast, if $\mu = 0$, the vaiance would tend to infinity, and the model would be uninformative as all alternatives would be equally likely. These extreme cases are illustrated in Figure 16.

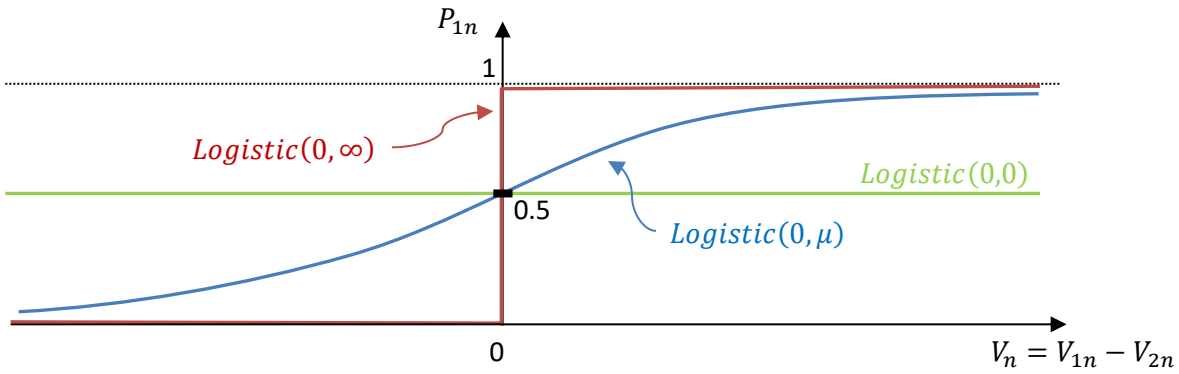


Figure 16. The Logit Model

Usually, $\mu = 1$ is assumed in the Logit model, leading to the following common analytical expression for the model:

$$P_{1n} = \frac{1}{1 + e^{-(V_{1n} - V_{2n})}}$$

Working with the algebra, we could write:

$$P_{1n} = \frac{1}{1 + e^{-(V_{1n} - V_{2n})}} \cdot \frac{e^{V_{1n}}}{e^{V_{1n}}} = \frac{e^{V_{1n}}}{e^{V_{1n}} + e^{V_{2n}}}$$

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Although not proved here, the locust of this expression extends to more than two alternatives, so, a common expression for the multinomial Logit model is:

$$P_{in} = \frac{e^{V_{in}}}{\sum_k e^{V_{kn}}}$$

11. Maximum likelihood estimation in the Logit Model

The remaining question is how to estimate the parameters when using the Logit model. One option could be to try to maximize the joint probability of correct predictions for all individuals, $P_{(i(1),\ldots,i(n))}$, where $i(1),\ldots,i(n)$ represent the selected alternatives for individuals $1,\ldots n$. Because we assume that individual choices are independent, the joint probability is the product of individual probabilities. This is:

$$P_{(i(1),\ldots,i(n))} = \prod_{j=1}^{n} P_{i(j)}(\Theta)$$

In the previous expression $j = 1\ldots n$ represents the different individuals, Θ the vector of parameters we want to estimate, and $P_{i(j)}$ the probability of the correct prediction for individual j. This joint probability of correct predictions is the likelihood function, $L^*_{(\Theta)}$, which using an alternative notation could be expressed as:

$$L^*_{(\Theta)} = P_{(i(1),\ldots,i(n))} = \prod_{j=1}^{n} \left[P_{ij}(\Theta)\right]^{\Delta_{ij}}$$

Where $P_{ij}(\Theta)$ is the probability of individual j selecting alternative i, which is determined by the Logit expression, and Δ_{ij} is:

$$\Delta_{ij} = \begin{cases} 1 & \text{if individual } j \text{ selects alternative } i \\ 0 & otherwise \end{cases}$$

Note that in the likelihood function, only the probabilities of the selected alternatives matter, although to compute them you need the attributes of non-selected alternatives.

The objective of the estimation process is to select Θ in order to maximize $L^*_{(\Theta)}$. The maximum value for $L^*_{(\Theta)}$ is 1, so we are interested in finding values close to 1, although do not expect values like 0.99 or something similar, as there are too much errors in the individual choice model.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

The logarithm function has nice mathematical properties in order to deal with this optimization, this is why in the maximum likelihood estimation (MLE) typically it is maximized the log-likelihood, L, which is equivalent to maximizing the likelihood.

$$L_{(\Theta)} = ln\left(L^*_{(\Theta)}\right) = ln\left\{\prod_{j=1}^{n} P_{i(j)}(\Theta)\right\} = \sum_{j=1}^{n} ln\left(P_{i(j)}(\Theta)\right)$$

The previous expression uses the property of the product of logarithms. The best fit for the log-likelihood, L, would be $ln(1) = 0$, and this expression would be better for dealing with very small probabilities.

Then, the MLE estimation consists in finding $\widehat{\Theta}$, where:

$$\widehat{\Theta} = \arg.\max_{\Theta} L_{(\Theta)} \Longrightarrow \max_{\Theta} \sum_{j=1}^{n} ln\left(P_{i(j)}(\Theta)\right) \Longrightarrow \nabla_L(\Theta) = 0$$

Where $\nabla_L(\Theta)$ is the Lapacian of $L_{(\Theta)}$ (i.e. the vector of partial derivatives of L with respect to Θ). We do not need to check the second derivative, because it is known that the likelihood function for the Logit is convex, and therefore we are going to find a maximum.

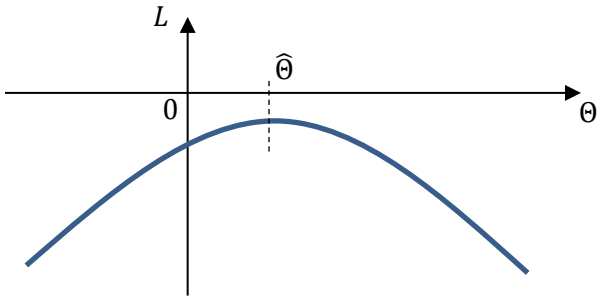


Figure 17. Maximum log-likelihood estimation (MLE)

12. Example of application: Estimation of the binary Logit via MLE (from individual choice data)

Consider to alternative transportation modes: Mode 1 and Mode 2. The specification for the systematic utilities derived from these transportation modes depend on one single generic parameter, θ:

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$V_{1j} = \theta t_{1j}$$
$$V_{2j} = \theta t_{2j}$$

Where t_i represents the travel time in each of the transportation modes.

The following actual choice data is available:

j (individual)	$i(n)$ (selected mode)	t_{1j}	t_{2j}
1	1	5	3
2	1	1	2
3	2	3	4

When individual choice data is available, we can estimate θ using MLE. First, we construct the likelihood function, that in this case is:

$$L^*_{(\theta)} = P_{11}P_{12}P_{23}$$

Recall that we formulate $L^*_{(\theta)}$ as the product of $P_{i(j)}$, the probabilities of actual selected alternatives.

Then, using the Logit expression:

$$P_{1j} = \frac{1}{1 + e^{-(V_{1n} - V_{2n})}}$$
$$L^*_{(\theta)} = \left(1 + e^{-2\theta}\right)^{-1}\left(1 + e^{\theta}\right)^{-1}\left(1 + e^{-\theta}\right)^{-1}$$

And taking logarithms:

$$L_{(\theta)} = ln(L^*_{(\theta)}) = -ln\left(1 + e^{-2\theta}\right) - ln\left(1 + e^{\theta}\right) - ln\left(1 + e^{-\theta}\right)$$

Now you could maximize $L_{(\theta)}$ numerically (solver optimization) or take derivatives with respect to the parameters and make them equal to zero:

$$\max_{\theta} L_{(\theta)} \Longrightarrow \frac{\partial L}{\partial \theta} = 0$$

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$\frac{\partial L}{\partial \theta} = \frac{2e^{-2\theta}}{1+e^{-2\theta}} - \frac{e^{\theta}}{1+e^{\theta}} + \frac{e^{-\theta}}{1+e^{-\theta}} = 0$$

In order to solve this equation, we can change variable $e^{\theta} = x$ and then:

$$\frac{2x^{-2}}{1+x^{-2}} - \frac{x}{1+x} + \frac{x^{-1}}{1+x^{-1}} = 0$$

Which operating leads to:

$$x^3 - x^2 - x - 3 = 0$$

Which using a solver leads to the solution $x^* = 2.13$, and then:

$$\hat{\theta} = ln(x^*) = 0.756$$

Note that the results of this example are somehow strange as we obtain a positive estimation for θ. Given the specification of the utility functions, we would expect a negative value for θ as travel time is considered a negative attribute for any transportation mode. However, results indicate a positive θ, meaning that, for this example, travel time is a positive attribute in the mode selection. This results from the data on actual choices, were individuals tend to select the mode with a longer travel time.

13. Estimation of the binary Logit via OLS (from aggregated choice data)

When individual choice data are not available, but only aggregated data from multiple individuals, ordinary least square regression (OLS) can be used to estimate the parameters in the utility function.

Consider again the example with two alternative transportation modes: Mode 1 and Mode 2, but now the specification for the systematic utilities depend on two parameters, θ_1 and θ_2, where θ_1 is still a generic taste constant for the travel time, and θ_2 is a constant modal preference that only affects to mode 2. Note that now θ_1 is specified as negative in the utility function.

$$V_{1j} = -\theta_1 t_{1j}$$
$$V_{2j} = \theta_2 - \theta_1 t_{2j}$$

Where t_i represents the travel time in each of the transportation modes.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

The available aggregated choice data for different segments of population, j, is presented in the table below:

j (group of individuals)	% selection Mode 2 (i.e. P_{2j})	t_{1j}	t_{2j}
1	0.20	28	24
2	0.49	37	26
3	0.43	28	22

According to the Logit functions the probabilities of selecting each transportation mode within the segment of population, j, are:

$$P_{1j} = \frac{1}{1 + e^{-(V_{1j} - V_{2j})}} = \frac{e^{V_{1j}}}{e^{V_{1j}} + e^{V_{2j}}}$$

$$P_{2j} = \frac{e^{V_{2j}}}{e^{V_{1j}} + e^{V_{2j}}}$$

And in order to apply OLS regression easily, we need to linearize the model. The Logit model can be linearized by applying the so-called Logit transformation. This is:

$$Ln\left(\frac{1 - P_{2j}}{P_{2j}}\right) = Ln\left(\frac{P_{1j}}{P_{2j}}\right) = Ln\left(\frac{e^{V_{1j}}}{e^{V_{2j}}}\right) = V_{1j} - V_{2j} = -\theta_2 - \theta_1(t_{1j} - t_{2j})$$

With the Logit transformation we have a linear model of the form $y = a - bx$, where we have observations of the dependent variable y (I.e. because we have observations of P_{2j}) related to that of the dependent variable x (i.e. the difference of travel times between modes). With this linear model and observations, we simply perform OLS regression, in order to estimate the parameters. In case of more than two parameters, we will need to apply multiple OLS.

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

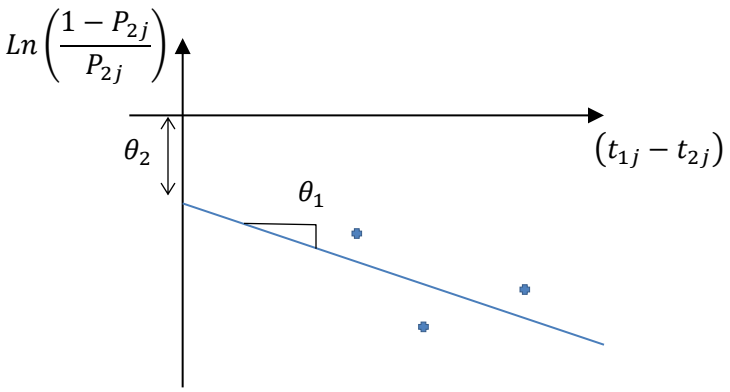


Figure 18. Ordinary least square (OLS) estimation wit aggregated data

14. Properties of Logit models

We want now to analyze how the probability of individual n choosing alternative i (i.e. P_{in}) changes, when the attribute k of alternative i (i.e. x_{ikn}) varies slightly. This is $\partial P_{in}/\partial x_{ikn}$.

Recall that:

$$P_{in} = \frac{e^{V_{in}}}{\sum_j e^{V_{jn}}}$$

Then, working out the derivative, we have:

$$\frac{\partial P_{in}}{\partial x_{ikn}} = \left[\frac{1}{\sum_j e^{V_{jn}}}\right] e^{V_{in}} \cdot \frac{\partial V_{in}}{\partial x_{ikn}} + e^{V_{in}} \left[\frac{-1}{\left(\sum_j e^{V_{jn}}\right)^2}\right] e^{V_{in}} \frac{\partial V_{in}}{\partial x_{ikn}}$$

The previous derivative is obtained realizing that only V_{in}, the systematic utility of alternative i, depends of x_{ikn}. Also, the derivative chain rule is applied (i.e. $\frac{\partial(f(x)g(x))}{\partial x} = g(x)\frac{\partial f(x)}{\partial x} + f(x)\frac{\partial g(x)}{\partial x}$).

Considering that systematic utilities are linear with respect to the attributes:

$$V_{in} = \theta_0 + \sum_k \theta_k x_{ikn}$$

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$\frac{\partial V_{in}}{\partial x_{ikn}} = \theta_k$$

Then:

$$\frac{\partial P_{in}}{\partial x_{ikn}} = P_{in}\theta_k - (P_{in})^2\theta_k = \theta_k P_{in}(1 - P_{in})$$

Yielding the property of the Logit model which states that the sensitivity to the variation of any parameter is maximum when the probabilities between alternatives are similar.

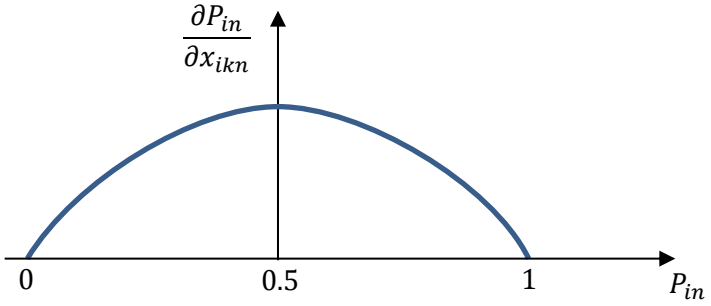


Figure 19. Sensitivity of the Logit model to the direct variation of one attribute

If we perform the same sensitivity analysis as before, but considering the cross-variation, $\partial P_{in}/\partial x_{jkn}$, we have:

$$\frac{\partial P_{in}}{\partial x_{jkn}} = 0 + e^{V_{in}}\left[\frac{-1}{\left(\sum_j e^{V_{jn}}\right)^2}\right]e^{V_{jn}}\frac{\partial V_{jn}}{\partial x_{jkn}}$$

Because $\partial V_{in}/\partial x_{jkn} = 0$. And considering $\partial V_{jn}/\partial x_{jkn} = \theta_k$, we obtain:

$$\frac{\partial P_{in}}{\partial x_{jkn}} = -\theta_k P_{in} P_{jn}$$

This means that if $\theta_k > 0$ (i.e. attribute k is desirable) then $\partial P_{in}/\partial x_{jkn} < 0$, the probability of choosing alternative i decreases, as alternative j becomes more attractive.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

14.1. Elasticity of Logit model

As commented before, the previous magnitudes (i.e. $\partial P_{in}/\partial x_{ikn} = \theta_k P_{in}(1 - P_{in})$; $\partial P_{in}/\partial x_{jkn} = -\theta_k P_{in} P_{jn}$) depend on the units of the attributes. In order to obtain dimensionless relationships, we need to use the definition of elasticity, ε, defined as the fractional change in the demand with respect to fractional change in the attribute.

Then, for the direct elasticity in the Logit model we have:

$$\varepsilon_{x_{ikn}} = \frac{\partial P_{in} / P_{in}}{\partial x_{ikn} / x_{ikn}} = \frac{\partial P_{in} x_{ikn}}{\partial x_{ikn} P_{in}}$$

And using the expressions derived previously, we have:

$$\varepsilon_{x_{ikn}} = \theta_k x_{ikn}(1 - P_{in})$$

Equivalently for the cross-elasticity:

$$\varepsilon_{x_{ikn}} = \frac{\partial P_{in} / P_{in}}{\partial x_{jkn} / x_{jkn}} = \frac{\partial P_{in} x_{jkn}}{\partial x_{jkn} P_{in}} = -\theta_k x_{jkn} P_{jn}$$

Note from this expression of the cross-elasticity of the Logit model, that it is independent of alternative i. This means that if something changes in alternative j, it will imply the same fractional change in all the other alternatives. This is a property of the Logit model called Independence from Irrelevant Alternatives (IIA), which leads to some problems, as described next.

14.2. Logit Independence from Irrelevant Alternatives (IIA)

Imagine we use a Logit model to determine the demand on two equivalent routes from an origin to a destination (i.e. route assignment).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

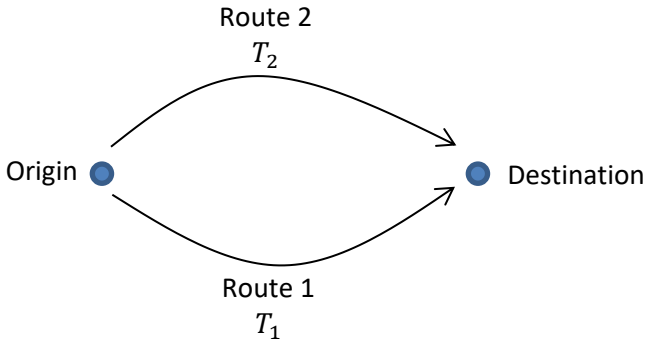


Figure 20. Route assignment discrete choice

Also, consider that the systematic utility on both routes is proportional to their travel time T_i, and that $T_1 = T_2$. This means that the systematic utilities are the same, $V_{1n} = V_{2n}$, and therefore the demand should distribute uniformly between routes (i.e. $P_{1n} = P_{2n} = 1/2$).

Now, consider that we include two possible alternatives (2a and 2b) for those vehicles selecting Route 2, considering if they take the left or the right lane when reaching the destination, as in Figure 21. Note that we changed nothing regarding the main choice between Route 1 and Route 2, so that the selection should not vary. However, if we apply the Logit Model, considering $V_{1n} = V_{2_a n} = V_{2_b n}$ (because $T_{1n} = T_{2_a n} = T_{2_b n}$), we would obtain $P_{1n} = P_{2_a n} = P_{2_b n}) = 1/3$. This means that the actual demands for Routes 1 and 2 have changed so that now we have $1/3$ of the demand through Route 1, and $2/3$ through Route 2, which is clearly wrong. The cause of the error resulting from the IIA property of the Logit is that alternatives 2a and 2b are highly correlated, which is against the assumptions of independent alternatives of the Logit.

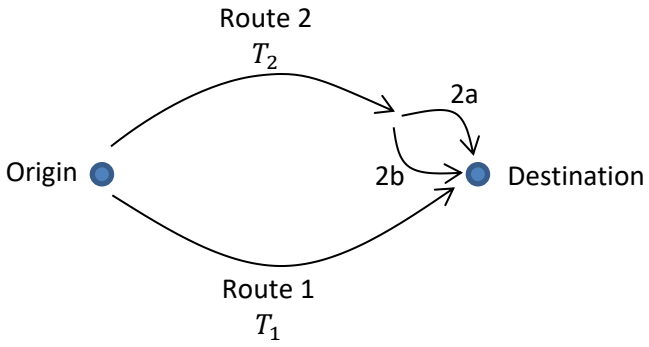


Figure 21. Example of the IIA property of the Logit model

A similar example which is commonly referred to in textbooks is the Red Bus / Blue Bus example. The example is like that: Assume you are using a Logit model to determine the demand for two transportation modes (i.e. auto and bus) for a given origin / destination pair. Assume that the systematic utilities for both alternatives are the same (i.e. $V_{auto} = V_{bus}$), so the probability of choosing either mode is $P_{auto} = P_{bus} = 1/2$. Then, consider that

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

at some point, part of the buses are painted in blue, while they were originally painted in red. If we consider now three transportation alternatives (i.e. auto, blue bus and red bus) with $V_{auto} = V_{red\_bus} = V_{blue\_bus}$, the choice probabilities have changed to $P_{auto} = P_{red\_bus} = P_{blue\_bus} = 1/3$. This means that just by painting the buses, we have changed the probability of the bus from $1/2$ to $2/3$, and an equivalent reduction in the probability of choosing the auto. Obviously, this is wrong result coming from the correlation between both bus alternatives.

In order to address this problem, an extension of the Logit model can be used. This is the nested Logit, which is detailed next.

15. Nested Logit

Consider an example where we want to determine the demand for three possible transportation modes on a route. The considered alternatives are: car, bus and metro. As discussed previously, in order to apply a Logit model, the random errors in the utility of each alternative must be independent identically distributed extreme value random variables (i.e. the assumption of the Logit).

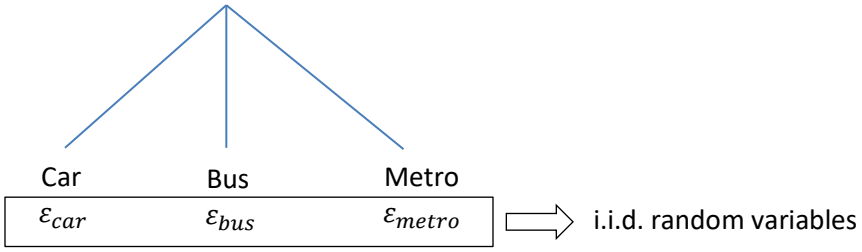


Figure 22. i.i.d. assumption of the Logit model

Given this setting and considering the standard Logit, the probability of choosing the bus option would be:

$$P_{bus} = \frac{e^{V_{bus}}}{e^{V_{car}} + e^{V_{bus}} + e^{V_{metro}}}$$

However, there could be correlation between the two public transportation options, leading to the problems exemplified previously. For example, people may exhibit preference (or aversion) to public transportation. Preference may come because of the externalities of car (e.g. congestion, emissions, etc) or people might have aversion to schedules, waits, access, etc. Note that the effects causing correlation in the errors are those not captured by the systematic utilities, otherwise, the correlation is washed out. So, considering that systematic utilities do not include the taste for public transportation, we might expect some correlation between ε_{bus} and ε_{metro}.

In order to wash out this correlation, we can think of a tree structure of decisions, independent at each level (or nest), as exemplified in Figure 23. This is the concept behind the nested Logit model.

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

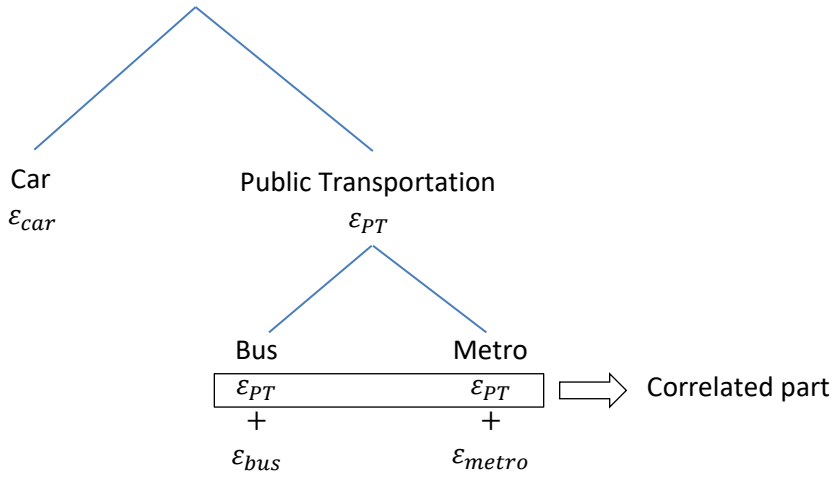


Figure 23. Nested Logit

Note that the correlation part in the errors of one nest will cancel out in the formulation of the Logit, because we only consider the difference between utilities, and therefore the differences between random errors.

In order to formulate the nested Logit model, we need to realize that at the lower levels we have a standard Logit (because correlation cancels out) yielding conditional probabilities. This is:

$$P_{(bus|PT)} = \frac{e^{V_{bus}}}{e^{V_{bus}} + e^{V_{metro}}}$$

$$P_{(metro|PT)} = \frac{e^{V_{metro}}}{e^{V_{bus}} + e^{V_{metro}}}$$

Al the upper level, we also have a standard Logit, because we assume that ε_{car} and ε_{PT} are i.i.d. random variables. In this case, we obtain marginal probabilities:

$$P_{car} = \frac{e^{V_{car}}}{e^{V_{car}} + e^{V_{PT}}}$$

$$P_{PT} = \frac{e^{V_{PT}}}{e^{V_{car}} + e^{V_{PT}}}$$

Then, we can obtain the marginal probabilities of bus and metro as:

$$P_{(bus)} = P_{(bus|PT)} \cdot P_{(PT)}$$

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$P_{(metro)} = P_{(metro|PT)} \cdot P_{(PT)}$$

The only remaining question is how we compute V_{PT} as a function of V_{bus} and V_{metro}, the actual alternatives of the nest. V_{PT} should represent the alternatives below, but these alternatives could be significantly different.

One alternative could be to consider a weighted average of the utilities of the alternatives in the nest. This is:

$$V_{PT} = E[V_{PT}] = P_{(bus|PT)}V_{bus} + P_{(metro|PT)}V_{metro}$$

The problem with this alternative is that if there is a bad alternative in the nest, this affects the probability of all the options in the nest, which should not be the case. To account for this, another alternative could be considered. Think of taking the utility of the nest as the maximum of the included alternatives. This is:

$$V_{PT} = Max[V_{PT}] = Max[V_{bus}, V_{metro}]$$

This seems better, but still there are some drawbacks. For instance, it does not consider that two good alternatives are better than just one.

To conclude, one of the best options, and the one typically used is the so-called expected maximum utility (EMU) of the nest. This is, sometimes, also called the "satisfaction" of the nest, and can be expressed as:

$$V_{PT} = E(Max[V_{PT}]) = E(Max[V_{bus}, V_{metro}])$$

If the errors in the utilities of the alternatives in the nest are extreme value i.i.d. random variables, the expectation of the maximum is computed as:

$$V_{PT} = E(Max[V_{bus}, V_{metro}]) = \theta Ln(e^{V_{bus}} + e^{V_{metro}})$$

Where θ is a correlation parameter that needs to be estimated in this upper level, using the typical methods.

16. Concluding remarks

In this chapter of the course, we have devoted some attention to utility theory to model human behavior and apply it to discrete choice and transportation demand modeling. But we cannot obviate that utility theory is based on strong assumptions, that might be arguable, namely:

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

- Humans are rational? People may not be perfectly "rational" in the sense of utility theory when they face complex decisions.
- Humans act consistently? Do you always make the same decisions given the same choices? Are you affected by choices of others (e.g. fashions, social imitation, or habits)?
- Humans are utility maximizers?
- We can determine the exact utility function? Everyone is different.
- Errors have a certain distribution?

In spite of these, utility theory has a number of advantages:

- We can obtain many details with just a few unknown constants
- We can disaggregate the demand by population group and in this way understand the effects of policy changes in great detail

And disadvantages:

- We need demographic information about the distribution of SEC and LOS variables
- If assumptions are wrong predictions can be wrong

In conclusion, utility theory is appropriate when advantages outweigh disadvantages; when we seek detailed demographic answers to simple choice problems. When these conditions are not satisfied we might be better off using models with fewer assumptions.

As a general rule, always try to use the model that answers the question with fewer assumptions. Complex problems have to be broken into parts in logical ways (e.g. like in the “four-step process” of demand modeling). Of the four steps (i.e. trip generation; trip distribution; mode split and traffic assignment) the third step is the one that can best be addressed with utility theory and disaggregate data. For the other steps simpler models based on aggregate data are often used in practice.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

8-TRAFFIC ASSIGNMENT:

Transportation Demand Modeling in Networks with Route Choice

Table of Contents

1. Introduction and basic ideas 2
2. Wardrop principles of network equilibrium 2
2.1. 1st Wardrop principle of network equilibrium: User Equilibrium (UE) ***2***
2.2. 2nd Wardrop principle of network equilibrium: System Optimal (SO) ***4***
3. The traffic assignment problem 4
3.1. Analytical solution ***7***
3.2. Graphical solution ***8***
3.3. Solution by simulation ***9***
4. Discussion of our traffic assignment analysis: Pros & cons 14
5. Network control and hypersensitivity of UE to input data 15
6. Network control & paradoxes 17
6.1. The Braess' paradox 17
6.2. Myopic network control: Smith's Paradox 19
APPENDIX 1: Inverse cdf method for generating random numbers 27

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

1. Introduction and basic ideas

In this lecture, we are going to analyze the foundations (i.e. the simplest but also most fundamental concepts) of transportation demand modeling in networks with route choice. Because there may exist several possible routes between origin-destination pairs, transportation demand needs to be assigned to a particular route in order to obtain the traffic loads on the network links. This is precisely the 4th step of the UTP (Urban Transportation Planning) process. Traffic assignment consists, precisely, in obtaining traffic flows for the different links of the network using as inputs the origin-destination matrix and the network performance functions (e.g. the links' travel time as a function of the link flow).

In this introductory analysis we are going to assume:

- Undersaturated networks: travel time is an increasing function of the traffic flow on a link. Delays may exist, but there are neither spillback queues between links, nor growing travel times with reducing flows.
- Steady state: time independent demand (i.e. time independent O/D matrix)

Even with these simplifying assumptions, sizing and network control becomes difficult when there exists the possibility of route choice. These difficulties arise because drivers adapt to changes in the network (i.e. control strategies or infrastructure modifications) by re-routing their trips. This re-routing can make the system worst after applying strategies to improve the network performance, even if drivers look for their own benefit and they are individually better off.

Note that if we implement a network control strategy that reduces overall delay considering the current existing conditions, this may not have the same beneficial effects for the conditions that will arise after drivers adjust their routes to the new scenario with control. So, the effects of any change in the network conditions must be evaluated for the conditions it will generate after drivers' adaptation. To achieve this, we need to be able to assess (i.e. to guess) the conditions that will prevail in the long term, and therefore we need a traffic assignment or routing model.

2. Wardrop principles of network equilibrium

As all topics we have covered in the course, the literature regarding traffic assignment models is huge, but the basic concepts that we are interested in are quite simple. The fundamentals of traffic assignment rely on the Wardrop's principles of network equilibrium, originally formulated in the 1950's by the English mathematician and transportation engineer J. G. Wardrop.

In 1952 Wardrop, in his famous publication entitled "*Some Theoretical Aspects of Road Traffic Research*" formulated two principles of network equilibrium, formalizing different behaviors regarding the minimization of travel costs.

2.1. 1st Wardrop principle of network equilibrium: User Equilibrium (UE)

The first Wardrop principle of network equilibrium (sometimes just named as the "Wardrop Equilibrium") postulates that every driver chose the route which minimizes his own travel time (or cost, if other costs than travel time are relevant to the analysis). Because this postulate is based on selfish user decisions, the principle it is also known as the "user equilibrium (UE)". The concept of UE is related to the idea of Nash equilibrium in game theory.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Assuming drivers have perfect information, the traffic flows which satisfy this principle result in equal travel times (or costs) in all used routes, and less than those that would be experienced by a single vehicle on any unused route. As an example, look at the simple example sketched in Figure 1, with three alternative routes for travelling from an origin to a destination. If routes "a" and "b" are used it means that both routes have the same travel time (i.e. $t_a = t_b$), otherwise drivers would have incentives to change routes. And if route "c" is not used, it is because $t_c > t_a = t_b$. Otherwise route "c" would have been used.

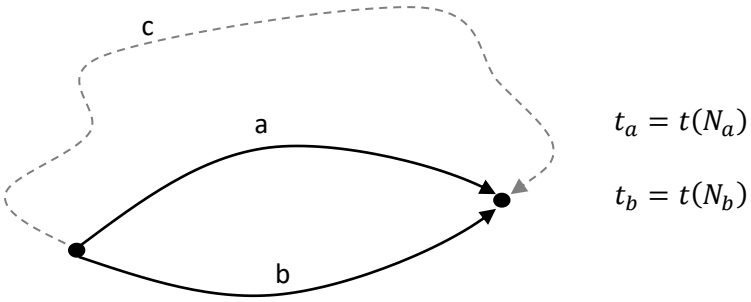


Figure 1. Simple network to illustrate Wardrop's equilibrium.

The User Equilibrium (UE) is reached when no user can lower his transportation cost by unilateral action. Note that, in UE, if one driver decides to change route, he would experience a longer travel time, because the receiving route would have larger than equilibrium demand, and travel time grows with the demand. So, she would return to his previous route choice, recovering the UE. This exemplifies the stability of UE against possible perturbations.

Originally, criticisms to Wardrop's user equilibrium result focused on the assumption of perfect information, which clearly was not true. Drivers' made decisions based on their knowledge of recurrent traffic conditions achieved after commuting multiple days. Actual conditions of the network at the instant of making decisions were unknown. So, it could be argued that UE was only reached when recurrent conditions held, and not in case of incidents which modify the network performance. The fact that recurrent congestion appears day after day at same locations and with similar effects, seems to empirically prove the existence of UE. Also, once some modification in the performance of the network was implemented, it took weeks or even months for drivers (or a fraction of them) to learn the new networks' recurrent traffic performance, after multiple trial and error route choices. This is the reason why the assessment of new network management strategies cannot be done immediately after its implementation. It is needed to wait several weeks to allow for user adaptation and reach a new steady state equilibirum.

In spite of this, today with the generalized use of ICT in the form of on-board navigation devices, the actual state of the transportation network is known by drivers when making decisions, and the assumptions of the first Wardrop principle hold. This means that UE might hold in a broader range of non-recurrent situations and might be achieved much faster than never before.

The principle of UE is based on the same behavioral principles than Utility Theory (i.e. individuals are utility maximizers, analytical, and well-informed), and therefore the same problems appear (i.e. how to model inconsistent or irrational decisions; or how to account for different perceived link costs for different users). The

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

solution is equivalent to that adopted in Random Utility Models, and consists in introducing a random error term in the cost of each link. This random term is non-observable and considers the model errors, individual specific tastes and non-rationality. This is the concept behind Stochastic User Equilibrium (SUE) where drivers choose to minimize their own perceived travel time (or cost), where the perceived travel time is equal to the average (or systematic, or deterministic) observable travel time plus a random error term. In the SUE no driver can unilaterally change routes to improve his perceived, rather than actual, travel times. We are not going to address the details of SUE in this lecture.

2.2. 2nd Wardrop principle of network equilibrium: System Optimal (SO)

In his second principle, Wardrop postulated that there exists a traffic assignment solution for which the average travel time is minimum. Equivalently, the total travel time in the network is also minimal. Wardrop proved that, in general, this SO traffic assignment solution is different than the UE. This means that not all the drivers are individually optimized (like in the UE), and some of they are penalized for the benefit of the whole system. It is difficult to understand how drivers could behave cooperatively in choosing their routes to ensure the most efficient use of the network, without the existence of a central authority, who could command them all which routes to take, or who could implement control strategies establishing penalties and benefits for some links according to its marginal costs, leading to an equilibrium solution approaching the SO. The potential efficiency improvement when moving from UE to SO by means of control is an example of the price of anarchy.

A particular case arises if travel times (or costs) do not depend on link demand (flows). Then, the assignment is called "all or nothing" (AoN) because all the flows are assigned to the minimum spanning tree (i.e. the route of minimum cost between an origin and destination, which is independent of the demand). In such case, UE and SO are equivalent.

3. The traffic assignment problem

Traffic assignment consists in converting the demand inputs given by origin-destination tables to link flows (see Figure 2). This requires to model a set of route choices, according to some behavioral rules (e.g. the Wardrop's principles). The type of questions that the traffic assignement process is able to answer are:

- Given O-D demands, what is the flow on each link? (Traffic Assignment)
- Given these flows, what is the total time on the network?
- Sensitivity analysis: What if O-D demands increase? What if we implement control?

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

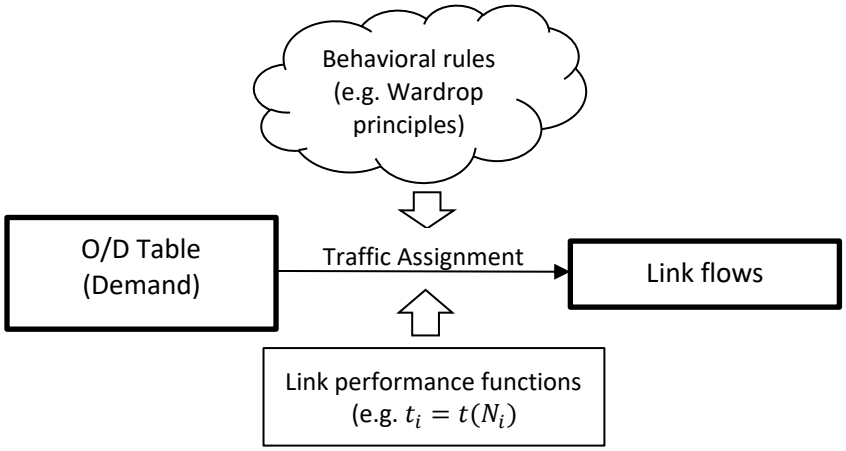


Figure 2. The traffic assignment concept

The solution of the traffic assignment problem can be obtained through a number of methodologies. For instance, a graphical solution procedure is possible in case of very simple problems with just two routes. If the problem is more complex involving a larger network but quite simple link performance functions, the problem might be addressed analytically by solving a system of equations. In spite of these, in realistic problems, with large networks and complex performance functions, the solution needs to be obtained by using simulation.

Below, in Figure 3, we represent a simple example for which the UE assignment will be obtained using the three previous procedures. The example simply consists on some drivers (i.e. Q_{bc}) with two route choices for reaching the Central Business District (CBD). They can either chose to access the freeway and join the preexistent freeway demand (i.e. Q_{ac}) or continue through a sequence of streets and arterials until the city.

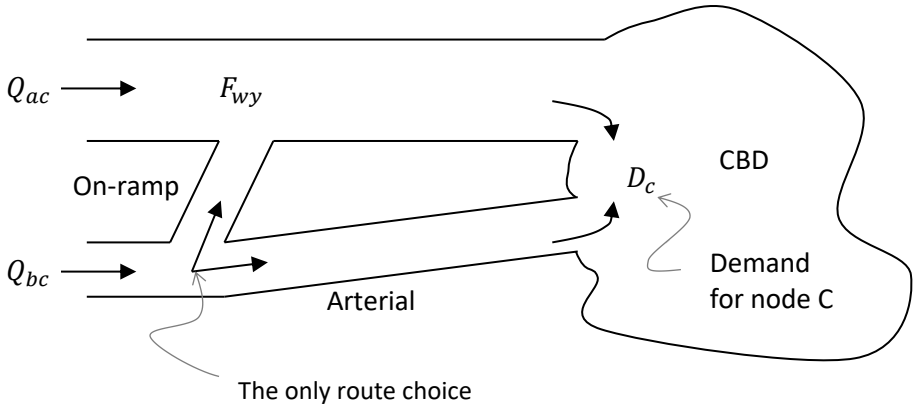


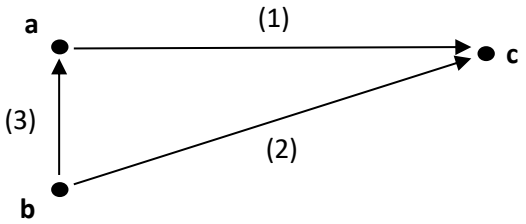
Figure 3. A simple network example.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

The first step is to pose the problem. To that end, it is useful to represent the network as a graph (i.e. nodes and links) and the demand as an O/D matrix, as done in Figure 4.



O \ D	c
a	$Q_{ac} = 80\ {}^{veh}/_{period}$
b	$Q_{bc} = 10\ {}^{veh}/_{period}$

$t_1 = 10 + q_1$
$t_2 = 90 + q_2$ in some units of time
$t_3 = 0\ (ramp)$

$q_1, q_2, q_3 = ?$

Figure 4. Posing the problem.

It is clear from the directed graph in Figure 4 that the only demand with route choice is that from origin "b" bounded to destination "c" (i.e. Q_{bc}) consisting of 10 vehicles in the period of analysis. These vehicles may choose the freeway option (i.e. Link 3 + Link 1) or continue through the arterials (i.e. Link 2). According to the performance functions in Figure 4, the free-flow travel time on the freeway is shorter than that on the arterial (i.e. 10 versus 90 time units). However, these times are affected by the travelling flows on these links, with the same sensitivity (i.e. travel time increases by one time unit for every additional vehicle on the link). Note that linear performance functions are adopted in order to simplify the example, although we know that the diagram relating travel time with flow exhibits strong non-linearity (e.g. typically to the power of 3) as illustrated in Figure 5. In any case, it is quite difficult to obtain accurate performance functions either. Finally, consider that the cost of travelling on the short-on ramp is neglected (i.e. $t_3 = 0$).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

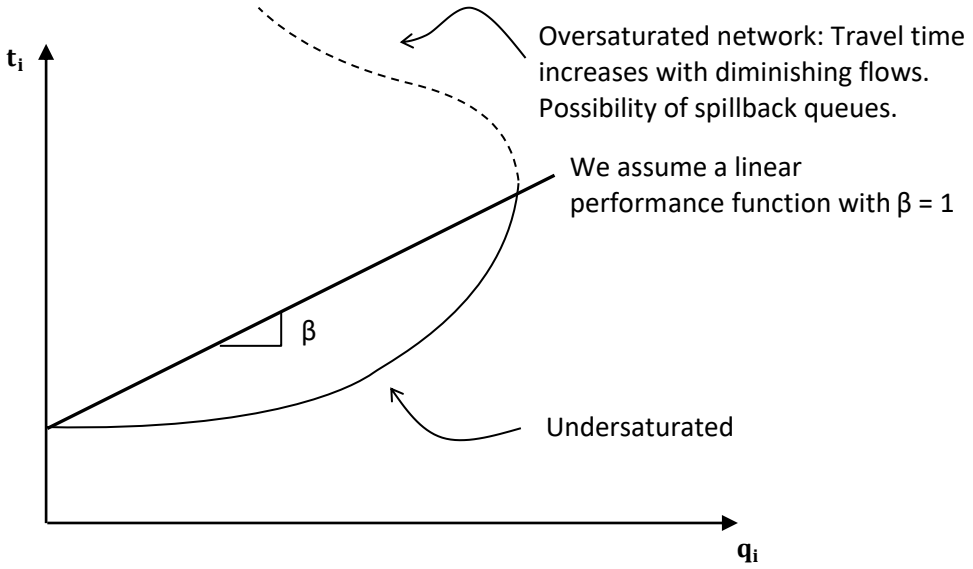


Figure 5. Link performance functions.

3.1. Analytical solution

The analytical solution of the traffic assignment problem is based on the analytical formulation of the vehicles' conservation and equilibrium conditions.

Conservation at nodes: Vehicles conservation must be imposed at every node of the network. There are $n-1$ independent conservation equations. In our example $n-1=3-1=2$ independent equations. These do not depend on the behavioral rules used for the route choice, and represent a connectivity property of the network considered. For our example, we have:

$$Q_{bc} = q_3 + q_2 \quad (at\ node\ \boldsymbol{b})$$
$$Q_{ac} + q_3 = q_1 \quad (at\ node\ \boldsymbol{a})$$

The third node would have only provided a linear combination of the previous equations (e.g. $q_1 + q_2 = Q_{bc} + Q_{ac}$) and can be neglected as it does not provide additional knowledge.

Equilibrium: Equilibrium conditions depend on the behavioral route choice model used. For the Wardrop's UE solution, we need to impose that the travel times on all used routes need to be equal. Considering the cost on Link 3 is null (i.e. $t_3 = 0$), the equilibrium condition is simply $t_1 = t_2$.

Considering together the conservation and equilibrium equations, we have a system of equations of size $n = 3$. In the present case the system of equations is linear, and can be easily solved resulting in the equilibrium flows on each link. Once the UE traffic assignment is solved, is straightforward to compute the equilibrium travel times on each link and the total travel time in the network, as detailed below:

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$\boldsymbol{t_1 = t_2} \rightarrow 2q_3 = 10 \rightarrow q_3 = 5\ {}^{veh}/_{period}$$

$$t_1 = 10 + \overbrace{Q_{ac} + q_3}^{q_1} = 90 + q_3 \rightarrow q_1 = 85\ {}^{veh}/_{period}$$

$$t_2 = 90 + \underbrace{Q_{bc} - q_3}_{q_2} = 100 - q_3 \rightarrow q_2 = 5\ {}^{veh}/_{period}$$

$$t_1 = t_2 = 95\ time\ units$$

$$Total\ travel\ time = \sum_i q_i.t_i = q_1.t_1 + q_2.t_2 = 85.95 + 5.95 = 8550\ {}^{veh.time}/_{period}$$

Remember this total travel time of 8550 veh·time units/period as it will be used as a benchmark for next situations.

3.2. Graphical solution

In very simple situations like the one analyzed here, with only two routing options, the travel times on each route can be plotted (y-axis) as a function of their demand (x-axis), with confronting directions of the x-axes and where the total demand traveling of both routes (known) defines the separation between y-axes, as in Figure 6. Then, the UE solution is simply obtained as the crossing point between the plots of both functions, which represent the equilibrium demand for which travel times on both routes are equal.

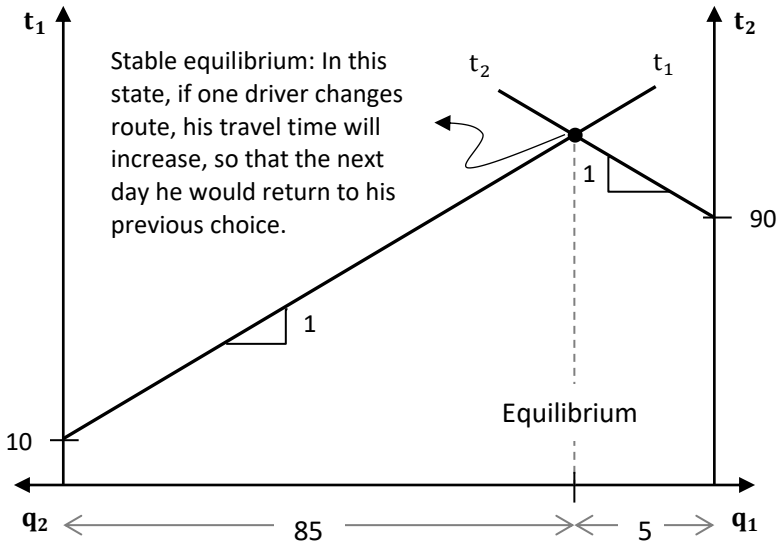


Figure 6. Graphical solution of the assignment problem.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

3.3. Solution by simulation

In case of realistic problems, with large scale networks and complex performance functions, the graphical method is unfeasible, and reaching an analytical solution might be only possible by formulating the traffic assignment as a mathematical program and using numerical methods.

An alternative procedure can be to look for an approximate solution using heuristics and simulation. Heuristics are a set of decision-making rules for individuals. These decisions are then simulated multiple times until equilibrium flows in the network are achieved. This emulates reality, where equilibrium is reached after weeks of drivers (or a fraction of them) making trial and error route switching.

In order to better understand the heuristic's definition process, we are going to review, first, two reasonable models which do not lead to equilibrium (→ Bad models). Afterwards we are going to present a good model leading to the equilibrium solution.

Simulation method 1: Rational human adjustments

The objective of defining heuristics in the traffic assignment problem is to generate solutions (e.g. by computer) to emulate the real decision-making process; in our case how do drivers select routes. According to the Wardrop's principle of UE (our considered behavioral rule), each driver selects the route with the minimum travel time (or cost). So, it is reasonable to think of a simulation model where “rational” human adjustments occur each period. This means that at every new iteration of the simulation (i.e. when a new decision is to be made), all the drivers select the route with the minimum travel time. We are going to see (i.e. with the same example used previously) that this simple and reasonable model is not adequate, as it does not lead to an equilibrium solution.

Consider the same problem as before (see Figure 3 and 4), where $Q_{bc} = 10$ vehicles need to make a route choice decision between Route 1 (i.e. Link 3 + Link 1) and Route 2 (i.e. Link 2) to reach Node c. Given the performance functions on each link (see below) and neglecting the cost on Link 3 (i.e. $t_3 = 0$), we are interested in determining the flows on each route according to UE. This is equivalent to determining the flows on Link 2 and Link 3 (i.e. q_2 and q_3) which need to add up to the 10 vehicles because of conservation.

$$\begin{matrix} t_1 = 90 + q_3 \\ t_2 = 100 - q_3 \end{matrix} \rightarrow \Delta t = t_2 - t_1 = 10 - 2q_3$$

According to our heuristics simulating rational human adjustment, every new iteration of the algorithm (e.g. every new "day" if considering the resemblance to real decision-making) the flow on Link i, q_i^{new}, would be equal to the "rational" flow on the link, q_i^{rat}, where this rational flow is the corresponding flow on Link i if all the drivers take the route with minimum cost. This is formulated as:

$$q_i^{new} = q_i^{old} + \underbrace{\left(q_i^{rat} - q_i^{old}\right)}$$

These drivers want to change route because they were not on the route with minimum cost

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Note that the previous expression simply states that $q_i^{new} = q_i^{rat}$, but in a recursive form to better compare to the next simulation methods.

Table 1 summarizes the results of applying the "rational human adjustment" simulation process. The simulation needs to start with a "seed". The seed is just a feasible first approach solution. The only requirement for the seed is to be a feasible solution, but the more the seed approaches the actual equilibrium solution, the smaller number of iterations would need the algorithm to converge. In Table 1, $q_3 = 2$ and $q_2 = 8$ is used as the seed. Note that this is a feasible solution, because $q_3 + q_2 = Q_{bc} = 10$. Given these flows, the travel times on both routes can be obtained, in this case $t_1 = 92$ and $t_2 = 98$. Because $t_1 < t_2$ (i.e. $\Delta t = 6$) drivers on Route 1 are better off. So that the rational flows (as we have defined them; all demand through the route with minimum cost) would be $q_3^{rat} = 10$ and $q_2^{rat} = 0$. With this information, and using the recurrent equation for the rational human adjustment algorithm, we can obtain the new flows, q_i^{new}, for iteration 2 (i.e. Day 2). The process can be repeated for this iteration. You can see from the results of Table 1 that, after a couple of iterations, the solution enters into an unstable flip – flop behavior, not reaching an equilibrium solution. The conclusion is that this is not an adequate simulation model for traffic assignment.

Table 1 – 1st simulation method: "Rational" human adjustment.

Days (n)	q_3	q_2	Δt	t_1	t_2	q_3^{rat}	q_2^{rat}
Day 1 (Random seed)	2	8	6	92	98	10	0
Day 2	10	0	-10	100	90	0	10
Day 3	0	10	10	90	100	10	0
… flip-flop behavior							

Simulation method 2: Staggered decisions

One could think that the problem in the rational adjustment method is that drivers over-react to travel time differences between routes. There might be a number of drivers captive of one route, or unaware of travel time differences, so that they do not change route at each new iteration. This is the concept behind the "staggered decisions" simulation method. Here it is assumed that only a fraction, α, of the drivers are subject to route switch at each new iteration. This is:

$$q_i^{new} = q_i^{old} + \alpha \cdot \left(q_i^{rat} - q_i^{old}\right)$$

Table 2 provides the results of the staggered decisions simulation method by considering $\alpha = 0.5$. Note that from the 8 drivers that were on a bad route selection on Day 1, only half of them (i.e. 4) change route, so that the flow on Link 3 on Day 2 is $q_3 = 2 + 4 = 6$. The simulation can be continued for successive iterations, and one would realize that equilibrium is not reached either. So, still this is a bad simulation model.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Table 2 – 2nd simulation method: Staggered decisions.

Days (n)	q_3	q_2	Δt	t_1	t_2	q_3^{rat}	q_2^{rat}
Day 1 (Random seed)	2	8	6	92	98	10	0
Day 2	6	4	-2	96	94	0	10
Day 3	3	7	4	93	97	10	0
Day 4	6.5	3.5	…	…	…		
... No Equilibrium							

Simulation method 3: Adaptive simulation

Adaptive simulation is a good simulation model to emulate route choice decisions and reach an equilibrium solution. The concept behind adaptive simulation is that the fraction of drivers subject to change route diminishes throughout the adaptation process (i.e. fewer travelers adjust every day as the adaptation process evolves). This means that α, which was a constant in the staggered decisions simulation model, is substituted by P_n, a decreasing function with the iteration number, n.

$$q_i^{new} = q_i^{old} + P_n \cdot \left(q_i^{rat} - q_i^{old}\right)$$

Portion of travelers who adapt on day "n"

Any decreasing function with n could be a valid option for P_n, and in particular $P_n = \frac{1}{n+1}$ works well. Table 3 shows the results of the adaptive simulation model to our example problem. You can see that after some iterations the solution would converge to the equilibrium solution of $q_2 = q_3 = 5$. In large scale realistic problems, it would take a very large number of iterations to reach the exact equilibrium solution, or it may not even happen. So, in general an approximate solution suffices. This is achieved by imposing a tolerance in the solution, so that if $q_i^{new} \approx q_i^{old}$ the simulation ends.

Table 3 – 3rd simulation method: Adaptive simulation with $P_n = 1/(n+1)$

Days (n)	q_3	q_2	Δt	t_1	t_2	P_n	q_3^{rat}	q_2^{rat}
Day 1 (Random seed)	2	8	6	92	98	1/2	10	0
Day 2	6	4	-2	96	94	1/3	0	10
Day 3	4	6	2	94	96	1/4	10	0
Day 4	5.5	4.5	-1	95.5	94.5	1/5	0	10
... convergence towards equilibrium solution.								

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

It can be proved that if link cost (i.e. performance, C) functions $C_i(q_i)$ are increasing functions for $q_i \geq 0;\ \forall i \in L$, where L is the set of links in the network, then: *i)* There is a unique set of equilibrium link flows and costs; and ii) adaptive simulation method reaches equilibrium for some iteration.

Be careful because the simulation converges provided that the performance function is continuously increasing, while the constant cost is not an increasing function. So, performance functions must be defined continuously increasing, even in the light traffic region, where travel times are insensitive to travel flows.

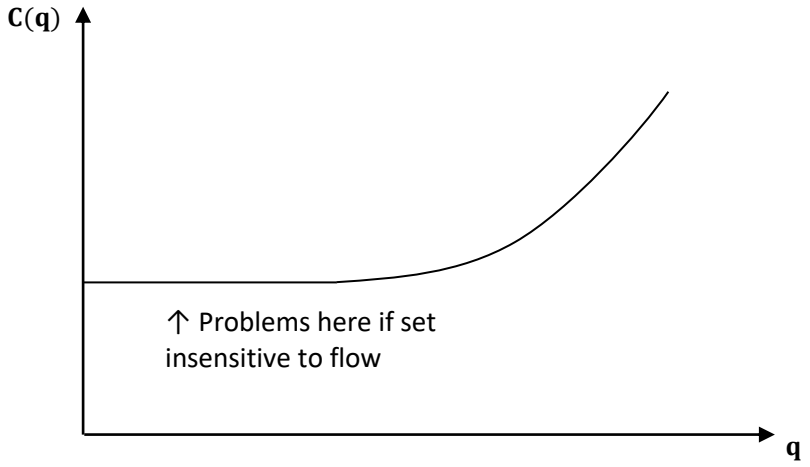


Figure 7. Example of a non-strictly increasing performance function.

As a conclusion of this section exploring the solution of the traffic assignment problem using simulation, Figure 8 provides a flow chart of the main steps of the process. At this point is worth highlighting the most challenging steps in a large-scale realistic problem. The first step, consisting in finding the initial solution to be used as a seed for the simulation is challenging. In large and complex networks, it might be difficult even to find a feasible solution. The help of some algorithm dealing with the connectivity of the network might be necessary. From the initial solution, computing link costs is straightforward, as only it is needed to apply the link cost functions. The next step is to identify best routes on a network for all O-D pairs given the link costs. In complex networks, this is also a challenging task and the use of a shortest path algorithm[1] is necessary. Rational link flows can be obtained by simply assigning all the demand to the shortest path routes, and adding up all the flows traversing each link. Next, new link flows are obtained by applying the recursive formula of the adaptive simulation, and the simulation ends or starts again with a new iteration.

As a side note, note that the intuitive process for the last step of the simulation would be: *i)* to compute new route flows according to the adaptive simulation recursive algorithm; and *ii)* find new link flows adding up all the demand traversing each link. However, this procedure is equivalent to directly applying the recursive formula to link flows. Note that if r is a particular route and i is a particular link:

[1] Dijkstra's shortest path algorithm, is one of the first and most well-known algorithms for finding the shortest paths between nodes in a graph It is due to the computer scientist E. W. Dijkstra in 1956.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$q_r^{new} = q_r^{old} + P_n\left(q_r^{rat} - q_r^{old}\right)$$

$$q_i^{new} = \sum_{\forall r} \delta_{ri} q_r^{new}$$

where:

$$\delta_{ri} = \begin{cases} 1 & \text{if route "}r\text{ " uses link "}i\text{"} \\ 0 & \text{otherwise} \end{cases}$$

Then:

$$q_i^{new} = \sum_{\forall r} \delta_{ri} q_r^{old} + \sum_{\forall r} \delta_{ri} P_n\left(q_r^{rat} - q_r^{old}\right) = q_i^{old} + P_n\left(q_i^{rat} - q_i^{old}\right)$$

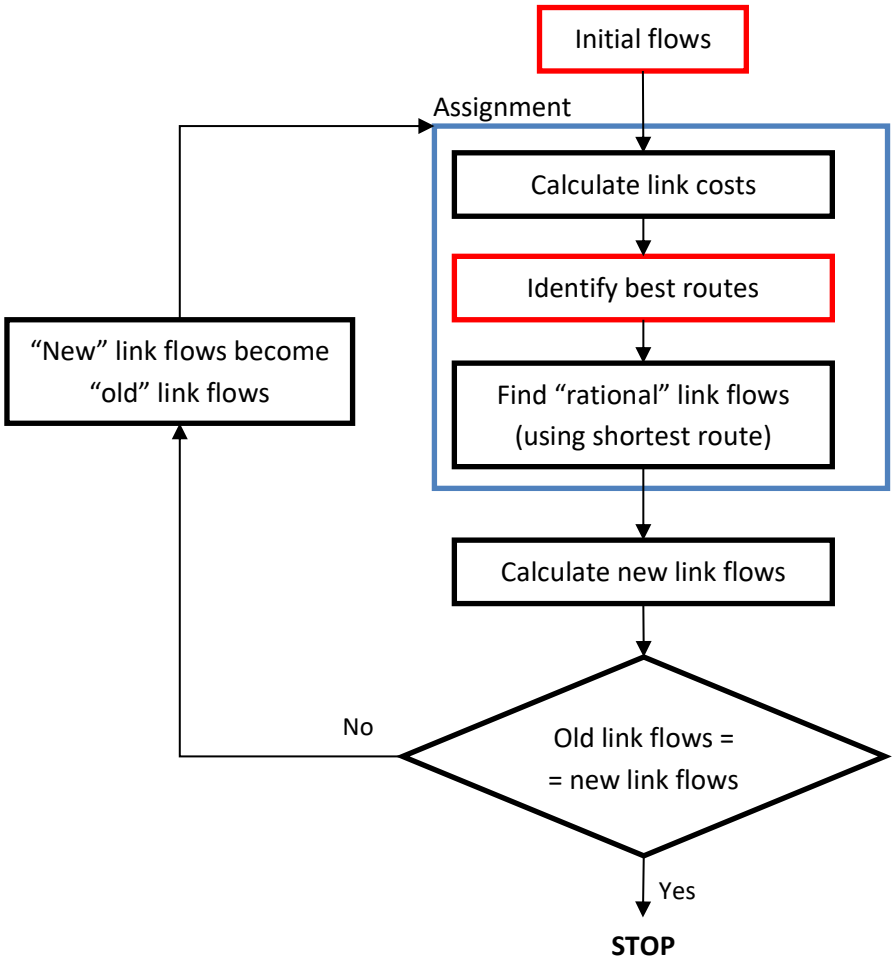


Figure 8. Adaptive simulation (Algorithm flow chart)
(Note: Challenging steps are in red)

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

4. Discussion of our traffic assignment analysis: Pros & cons

Our solution of the traffic assignment problem through adaptive simulation has some strengths but also some limitations. These are going to be discussed in this section.

Strengths

1. *Computationally efficient*. The computation time of the simulation algorithm (in units of elementary operations) is of the order of $n \cdot l \cdot ln(l)$, where n is the number of nodes and l is the number of links in the network. This avoids exponential computing times which are harmful for the computational efficiency. The reason behind the efficiency of the algorithm is that it is not needed to enumerate all routes (a number which might be astronomical) when computing the "rational" route flows. It is only needed to identify the best route for each O/D pair and allocate the demands directly to the composing links.
2. *Flexible and general framework*. The simulation framework presented can be used not only for the traffic assignment problem, but also for solving the other steps of the UTP process, by representing the tree structure of decisions as a graph. Then, all decisions can be modeled in this "virtual" network as a traffic assignment problem. Look for instance at Figure 9. The network represents the decisions involved for an individual in deciding whether travelling to a destination D_1 or to another destination D_2 (i.e. trip distribution) and selecting the transportation mode for each part of the transportation chain (i.e. mode choice). By solving the traffic assignment in this network, trip distribution and modal choices would be obtained.

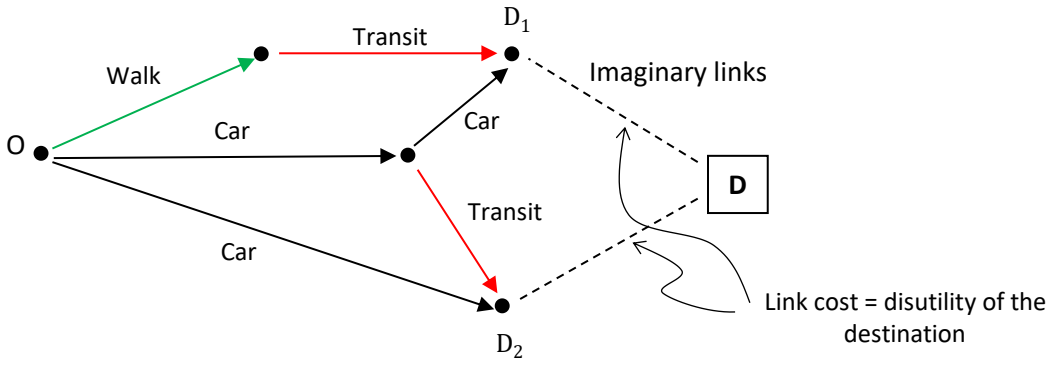


Figure 9. Multimodal framework of traffic assignment problem.

Limitations of our network analysis

1. *Static*. We have assumed time independence of demand (i.e. O-D matrixes) and of performance functions. Both may vary during the course of the day, specially demands if strong peak periods exist. The traffic assignment considering time-varying O-D matrixes and performance functions is known as the Dynamic Traffic Assignment (DTA) problem and implies and important conceptual difference to our static analysis. The main difference is that users' decisions are not made considering the current traffic state of the network,

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

but by trying to anticipate the forecasted state at the time of crossing the different links, which depends on time. Drivers use their experience on the network to anticipate traffic conditions, and try to guess other drivers' decisions. So, decisions are based on double guessing, trying to anticipate other drivers' decisions, who at the same time are trying to anticipate your decisions (i.e. like playing poker). Game theory can be used to try to predict the outcomes of these double guessing situations.

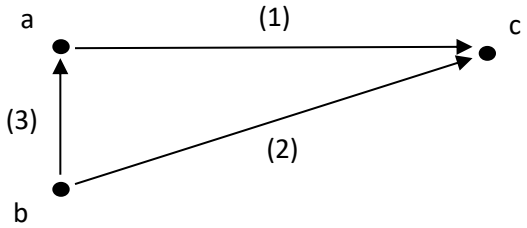
2. *Undersaturated networks*. We have assumed that queues do not spillback from one link to another. This allows considering the cost on each link independent from other links. If oversaturated networks and queue spillbacks between links are to be considered, the costs on each link might depend on the vehicles' accumulation in the neighboring links in past iterations. This requires "system memory" and adds additional complexity in the process of finding the solution. In addition, and as it is going to be exemplified in the next section, traffic assignment on networks (and especially on oversaturated networks) is hypersensitive to the inputs. So that a minor perturbation in the input data (e.g. O-D) might lead to under-saturation or to complete gridlock. This opposite behavior depending of very small differences in the inputs (which are uncertain by nature) is highly detrimental to the robustness of traffic assignment in over-saturated networks.
3. *Complexity of human decisions.* As discussed several times during the course, human decisions are complex. Their modelling according to the assumptions of utility theory might be too simple in many situations. In any case, this is the best we have.

5. Network control and hypersensitivity of UE to input data

One of the main limitations of UE equilibrium analysis in transportation networks is its high sensitivity to the input data. This means that small perturbations in the input data may lead to large variations in the equilibrium solution. The solution is said to be hypersensitive to the inputs, which is the opposite of a robust solution, and a very detrimental property. The problem is aggravated by the fact that inputs (i.e. performance functions and O/D matrixes) are uncertain by nature, so that it is not possible to have precise inputs.

In this section we are going to show, returning to our previous example, that the solution for the UE is not scalable (which would be a very desirable property). This means that if we multiply by two the demand, for instance, the results of the link flows in UE are not multiplied by two. Instead, the solution can change dramatically.

Back to our previous example, we had:



O \ D	c
a	$Q_{ac} = 80\ ^{veh}/_{period}$
b	$Q_{bc} = 10\ ^{veh}/_{period}$

Figure 10. Initial problem – network structure and OD demands.

With the following performance functions:

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$t_1 = 10 + q_1 = 10 + Q_{ac} + q_3$$
$$t_2 = 90 + q_2 = 90 + Q_{bc} + q_3$$
$$t_3 = 0\ (ramp)$$

For the original demand, the UE solution was:

$$\begin{matrix} Q_{ac} = 80 \text{ veh/period} \\ Q_{bc} = 10\ v\text{eh/period} \end{matrix} \rightarrow \text{User equilibrium solution} \rightarrow q_3 = 5 \text{ veh/period}$$

Now, if we double the demand:

$$\begin{matrix} Q_{ac} = 160 \text{ veh/period} \\ Q_{bc} = 20\ v\text{eh/period} \end{matrix} \rightarrow \text{UE solution? Scalable problem? (i.e. double link flows)} \rightarrow \text{the answer is NO}$$

We can solve the problem with the new data using the graphical method, and we would obtain the result as in Figure 11. Note that the solution is $q_3 = 0$, and no one will use the on-ramp, as it is always better to use the arterial. This illustrates the dramatic and problematic change of behavior of UE with respect to the inputs.

$$t_1 = 170 + q_3$$
$$t_2 = 110 - q_3$$

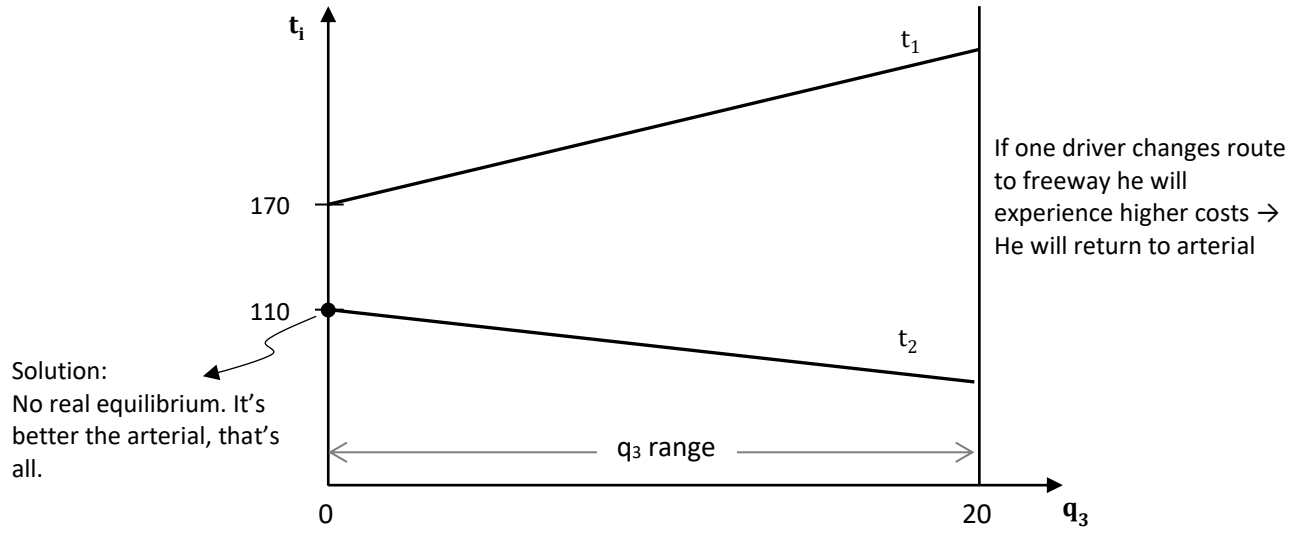


Figure 11. Graphical solution in double demand scenario.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

6. Network control & paradoxes

So far in the analysis of the UE traffic assignment solution, we have seen that the equilibrium is unique in many situations of practical interest (which is good news), but the equilibrium solution depends on input data in an unpleasant way (bad news), which results in additional difficulties to control flows on traffic networks.

The concept behind network control is to limit (or apply benefits - penalties) to some of the drivers' decisions in order to modify the user equilibrium improving the network performance and moving towards a system optimal situation.

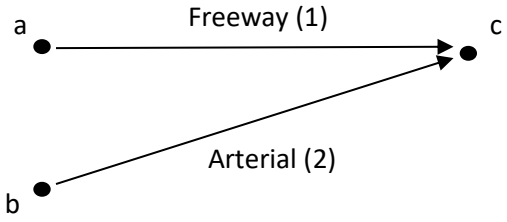
Network control is difficult due two main reasons: First, drivers adapt route choice to the implemented control strategies, modifying the baseline situation for which control was designed; Second, the resulting equilibrium flows are hypersensitive to the inputs, so that in reality the outcomes can vary in a quite random way. This is why in some cases, the results of applying control are counterintuitive and lead to outcomes being the opposite as expected. There are two famous paradoxes of network control which are illustrative, and help understanding some concepts of network control. These are the Braess' and the Smith's paradoxes.

6.1. The Braess' paradox

The Braess' paradox illustrates the fact that reducing some link costs on a network (e.g. increase capacity, reduce travel times, ...) which seems good, sometimes can get performance of the network worst (higher overall travel times).

So, the Braess' Paradox is summarized as: Reduce link cost (may) increase network cost. And viceversa: increase link cost (may) improve network performance.

We can use the same example we have been working on during the lecture. Just imagine that we remove the on-ramp link to the freeway (see Figure 12). This represents removing network capacity, and eliminating travel options. One might think that this would inevitably lead to worse network performance. But the result is the opposite: overall travel time on the network improves.



O \ D	c
a	$Q_{ac} = 80\ {}^{veh}/_{period}$
b	$Q_{bc} = 10\ {}^{veh}/_{period}$

Figure 12. Original problem – Removed on-ramp.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Recall that the considered performance functions were:

$$t_1 = 10 + q_1$$
$$t_2 = 90 + q_2$$

And because in this case (without on-ramp) there is no route choice, the solution is direct:

$$q_1 = Q_{ac} \quad q_2 = Q_{bc}$$
$$t_1 = 90 \quad t_2 = 100$$

Then, the total time spent on the network is:

$$Total\ travel\ time = \sum_i q_i . t_i = q_1 . t_1 + q_2 . t_2 = 90.80 + 100.10 = 8220\ {}^{\text{veh. time}}/_{\text{period}}$$

Recall that with the on ramp, the total time spent was:

$$Total\ travel\ time = 8550\ {}^{\text{veh. time}}/_{\text{period}}$$

So, this proves that building the on-ramp (i.e. adding capacity), which one may assert that it cannot be bad, it is NOT TRUE. This happens because, with the on-ramp, some of the drivers enter the freeway (because they will be better off; note that the people from "b" are better off with the on-ramp, 95 vs 100 time units) but with their decisions they will penalize a lot of people from "a".

The opposite is also true, we could increase link costs and improve network performance. *Ramp metering*[2] on freeways is a clear example. The concept is to increase costs on ramps, so that some drivers divert to surface streets (losers), avoiding crowded delays at freeway (winners). Note that ramp metering benefits drivers already on the freeway with respect to those trying to access through the on-ramps. In some scenarios, (e.g. metropolitan freeways accessing big cities) the equity issue of ramp metering could be justified arguing that drivers coming from farther apart (i.e. those already on the freeway who are benefited from the control strategy)

[2] Ramp metering is a well-known freeway management strategy to reduces congestion at on-ramps. It is very commonly applied in the USA. A more detailed description can be found at https://www.youtube.com/watch?v=QDMYODIgLcs. A nice pedestrians' experiment to illustrate ramp metering benefits was performed at the 2010 TRAIL Conference: https://www.youtube.com/watch?v=6ODvNZsXEvs (with no control); https://www.youtube.com/watch?v=EJo7O8f0JiY (with control).

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

have less alternative transportation options than the private vehicle, while those closer to the city (who need to use the ramps and are panelized) do have public transportation alternatives.

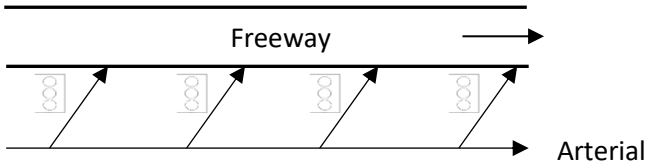


Figure 13. Network structure of ramp metering.

In conclusion, modifying the UE by means of control implies:

- There will be some losers and some winners (e.g. eliminating the on-ramp, "b" are losers and "a" are winers).
- We want that overall, the effect in winners overcome the effect on losers ("b" are few; "a" are many – few people should not penalize many people).
- Network control and modifying user equilibrium opens up the discussion on equity. It is fair to penalize some drivers? even if we benefit many other drivers? the answer should include many other policy and sociological inputs.

6.2. Myopic network control: Smith's Paradox

The Smith's paradox is an example to prevent against myopic control on networks. It illustrates how an optimal control strategy which focuses only on an isolated part of the network, might turn to be detrimental when considering the whole network. The example is formulated like that: Consider the residential area near the CBD sketched in Figure 14 below. The residents of this area have two possible routes to reach the main freeway to the CBD: Link 1 and Link 2. Capacity of Link 1 is $\mu_1 = 13$ veh/min, while capacity of Link 2 is $\mu_2 = 26$ veh/min. During the rush period, lasting a duration of 12.5 min, trips are generated in the residential area at a rate of 10 trips/min, so that a total of 125 trips are generated during the rush period. The maximum travel time from any point of the residential area to the intersection between Link 1 and Link 2 is 0.2 minutes.

Then, assuming a uniform distribution of households over the residential area, the travel time to the intersection is a random variable uniformly distributed between 0 and 0.2 minutes. This is:

$$T_{1j} \sim U[0, 0.2]$$
$$T_{2j} \sim U[0, 0.2]$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Where T_{1j} and T_{2j} are the travel times to the intersection of individual j if taking Link 1 or Link 2, respectively. Considering that travelers chose the closest link, they would take Link 1 if $T_{2j} > T_{1j}$, or Link 2 if $T_{1j} > T_{2j}$. Alternatively, we could define ΔT_j as:

$$\Delta T_j = T_{1j} - T_{2j} \sim U[-0.2\,, 0.2]$$

So that if $\Delta T_j > 0$, traveler j would take Link 2, and if $\Delta T_j < 0$, he would take Link 1.

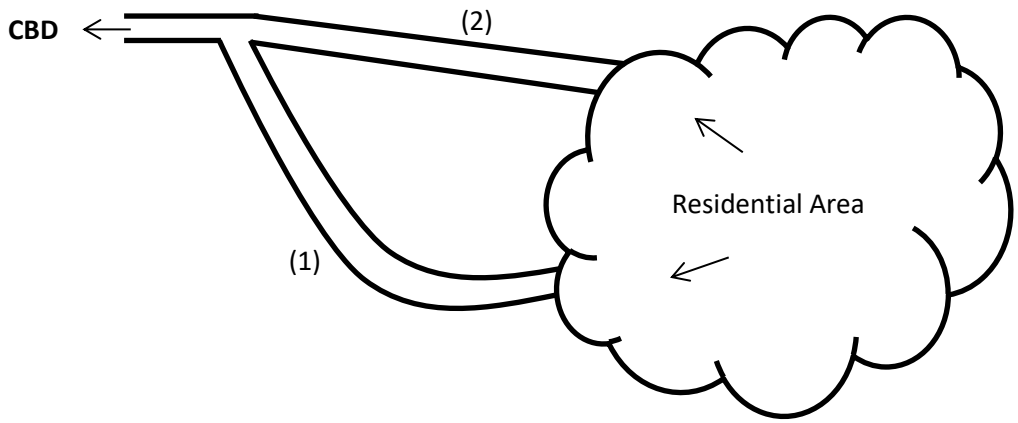


Figure 14. Network structure in the Smith's paradox.

In order to determine the flows through both links on a particular day (e.g. Day 1), we can compute one realization of the travel times of the 125 drivers to the intersection by generating 125 random numbers[3] following the $U[-0.2\,, 0.2]$ distribution. Let's say that, from this random number generation, we obtain 72 observations with $\Delta T_j < 0$ (i.e. a 57.6%), and 53 observations with $\Delta T_j > 0$ (i.e. a 42.4%). Then the resulting flows on links 1 and 2 are the ones shown in Table 4. Note that the expected delays at the intersection are null, because capacities are much larger than demands.

Table 4 – Resulting flows on Day 1

	Link 1	**Link 2**
Users	72 (57.6%)	53 (42.4%)
$\boldsymbol{q_j}$	5.76/min	4.24/min
$\boldsymbol{W_{avg}}$	0	0

[3] The inverse cdf method (see Appendix 1) is useful for generating random numbers following any distribution with the input of a $U[0, 1]$ random number generator.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Imagine now that some residents argue that the uncontrolled intersection is not safe, and an actuated traffic signal is installed. An actuated traffic signal is an adaptive flow control device, which assigns the green time for each approach (i.e. Link 1 and Link 2) proportionally to the flow ratio on the approach (i.e. the flow ratio is the demand over the capacity ratio before the installation of the signal). The demand inputs to the signal controller are provided by traffic detectors located on each approach. Consider a signal cycle $C = 1.2$ min (big, but not enormous) with a lost time $L = 0.2$ min. The signal cycle is the duration of the green plus the red plus the lost times in the signalized intersection.

This situation is common in the real word. In a street grid with multiple intersections (e.g. like the one represented in Figure 15) actuated traffic signals are installed with a fixed common cycle C (or integer multipliers) and variable green-red phase duration at each signal as a function of the demand on each approach. It seems reasonable to keep adjusting a node (e.g. traffic signal) in response to time-varying demand, but we are going to see (i.e. through the Smith's paradox, which is a simple illustration of this context) that the myopic optimization of isolated nodes without worrying about system wide effects, may lead to the deterioration of system performance with gradually worsening delays.

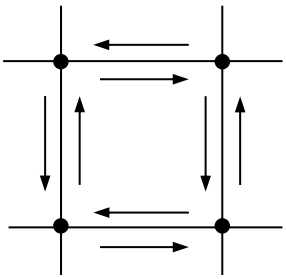


Figure 15. A street grid with actuated traffic signals.

Once the actuated traffic signal has been installed, the green time for each approach, G_i, as a function of its demand, q_i, and capacity, μ_i, for approaches $i = 1, 2$ can be determined as:

$$\frac{G_1}{G_2} = \frac{q_1/\mu_1}{q_2/\mu_2} \stackrel{\text{def}}{=} \frac{y_1}{y_2}$$

Where y_i is defined as the flow ratio for approach i. Then we have:

$$\frac{G_1}{G_2} = \frac{q_1}{q_2} \cdot r \quad \text{where} \quad r = \frac{\mu_2}{\mu_1} = 2$$

So, we can construct a system of 2 equations with 2 unknowns that we can solve.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$\begin{cases} G_1 = G_2 \cdot \dfrac{q_1}{q_2} \cdot r \\ G_1 + G_2 = C - L = 1\ min \end{cases}$$

The solution is:

$$G_2 = \frac{q_2}{q_2 + q_1 r}$$
$$G_1 = \frac{q_1 r}{q_2 + q_1 r}$$

The result is reasonable as more green time is assigned proportionally to more flow, so that the deterministic delay remains constant.

Although for this example the cycle time, C, is fixed, as a side note, Figure 16 details the procedure for the selection of C, so that the traffic signal results undersaturated in both approaches.

Side note: Determine "C"

$$\lambda_i \cdot C \leq \mu_i \cdot G_i \quad \rightarrow \quad \frac{\lambda_i}{\mu_i} \leq \frac{G_i}{C}$$

$$\forall i \rightarrow \quad \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} + \frac{L}{C} \leq \frac{G_1}{C} + \frac{G_2}{C} + \frac{L}{C}$$

$$y_1 + y_2 + \frac{L}{C} \leq 1$$

$$C \geq \frac{L}{1 - y_1 - y_2}$$

Figure 16. Cycle duration selection in a traffic signal.

Recall from the chapter devoted to queuing theory that, even if the average demand rate is smaller than capacity at the signal (i.e. an on-off server), stochastic delays may exist. The average excess accumulation due to stochastic effects on an under-saturated traffic signal can be roughly computed as:

$$\bar{Q}_j = \frac{\Delta/2}{1 - \rho_j}$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Where the parameter Δ depends on the variability of arrival and service processes. Usually $\Delta \approx 1$, which assumes independent Poisson arrivals. However, in a grid of signalized streets, where vehicles arrive in batches, we can consider $\Delta \approx 0.6$ which would be a better estimation.

In turn, ρ_j, is the traffic intensity, defined as the demand to capacity ratio (i.e. another name for the flow ratio defined previously, but here it needs to consider that the signal has already been installed). Note that once the signal is installed, the capacity is defined by the capacity of the approach, μ_j, times the fraction of the green time over the cycle duration. This is:

$$\rho_j = traffic\ intensity = \frac{demand}{capacity} = \frac{q_j}{\mu_j \frac{G_j}{C}}$$

Finally, the stochastic delay for the unsaturated signal is obtained from the excess of accumulation, determined previously, and using the Little equation:

$$\bar{W}_j = \frac{\bar{Q}_j}{q_j}$$

Rearranging and plugging in all the terms, we have:

$$\bar{W}_1 = \frac{\Delta/_2}{q_1\left(1 - \frac{q_1 C}{\mu_1 G_1}\right)}$$

$$\bar{W}_2 = \frac{\Delta/_2}{q_2\left(1 - \frac{q_2 C}{\mu_2 G_2}\right)}$$

Note that the delays are function of the green times. So, we can substitute in the previous equation the expression for the green times in the actuated signal we derived previously. We would obtain:

$$\bar{W}_1 = \frac{\Delta/_2}{q_1 - \frac{q_1 C}{\mu_1 r}(q_2 + q_1 r)}$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

$$\bar{W}_2 = \frac{\Delta/_2}{q_2 - \frac{q_2 C}{\mu_2}(q_2 + q_1 r)}$$

Substituting the numerical values in the example we obtain:

$$\bar{W}_1 = \frac{0.3}{q_1 - \frac{q_1 \cdot 1.2}{13 \cdot 2}(q_2 + 2q_1)}$$

$$\bar{W}_2 = \frac{0.3}{q_2 - \frac{q_2 \cdot 1.2}{26}(q_2 + 2q_1)}$$

To continue our discussion on the Smith's paradox, we are going to assume that the following day after the installation of the traffic signal (i.e. Day 2) the demand is unaware of the signal, and therefore is invariant with respect to Day 1 (i.e. same flows on each approach as in Table 4). With these demands, the total stochastic delay per cycle collectively experimented by drivers on both approaches can be obtained from the equations above, and the results would be those shown on Table 5.

Table 5 – Resulting flows on Day 2

	Link 1	Link 2	
Users	72 (57.6%)	53 (42.4%)	← Same as day 1 (drivers don't know the signal)
$\boldsymbol{q_j}$	5.76/min	4.24/min	
$\boldsymbol{W_{avg}}$	0.19	0.26	← From the equations we have derived

Then, the total delay collectively incurred by all drivers is:

$Total\ \ delay/cycle = 5.76 \cdot 1.2 \cdot 0.19 + 4.24 \cdot 1.2 \cdot 0.26 = 3.6$ min/cycle (recall 1.2 min is the cycle duration)

The next day (i.e. Day 3) given the experienced delays on Day 2, drivers will adapt to minimize their travel times. Note that delays on Link 1 were smaller than on Link 2, so that some drives who took Link 2 will switch to Link 1. The drivers who will reroute are those whose difference in travel time to switch to Link 1 (in relation to the closer Link 2) is less than $\Delta\bar{W}$, the difference in delay at both approaches.

$$\Delta\bar{W} = \bar{W}_1 - \bar{W}_2 = -0.07$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

So, only the drivers with $\Delta T_j = T_{1j} - T_{2j} > 0.07$ will take Link 2. Let's say that they count 31 in the random number list of 125 drivers[4]. Then, the obtained traffic assignment and delays for Day 3 is that of Table 6.

Table 6 – Resulting flows on Day 3

	Link 1	Link 2
Users	94 (75.2%)	31 (24.8%)
$\boldsymbol{q_j}$	7.52/min	2.48/min
$\boldsymbol{W_{avg}}$	0.21	0.63

With these new route decisions, the total delay per cycle would be:

$$Total\ \ delay/cycle = 7.52 \cdot 1.2 \cdot 0.21 + 2.48 \cdot 1.2 \cdot 0.63 = 3.8\, min/cycle$$

So that the delay difference between approaches is:

$$\Delta\bar{W} = \bar{W}_1 - \bar{W}_2 = -0.42$$

With these results, you can see that, as a result of rerouting and adaptive control at the intersection, the global delay has increased. And the process does not end here, because the next day (Day 4), all drivers would switch to Link 1, resulting in a total delay of 4.7 min/cycle[5]. On Day 5, observing that Link 2 is empty, all drivers would select Link 2 resulting in the unstable flip-flop behavior of the rational human adjustments simulation method (i.e. the route choice logic used in the Smiths' paradox). Smith's paradox assumes drivers have perfect information from the previous day and systematic rational human choice decisions. As discussed previously in the lecture, in reality not everybody is prone to change routes every day, and there is an adaptation, anticipation and double guessing decision-making process leading to somehow random results.

In any case, the conclusion is that the system degrades by controlling. The example is an artifact, but illustrates.

One remaining question that one could formulates is what would be the best for the system, if unsignalized intersection is not acceptable? The answer might depend on many conditions but one option could be to close

[4] It is highly recommended, in order to fully understand the example, to repeat the exercise with your own list of 125 $U[-0.2\ , 0.2]$ random numbers.

[5] This total delay is obtained if we consider that there's still an actuated signal and although all demand uses Link 1, still a very small green is to be assigned to Link 2 (e.g. to serve vehicles not measured by the detector), so that the lost time still exist, and we could assume $G_1 = C - L$, $G_2 = 0$.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Link 1, meaning that no signal is necessary, because there is no intersection. The total delay in this case would be:

$$10 \cdot 1.2 \cdot 0.1 = 1.2\ min/cycle$$

Average time to Link 2 of all travelers
Cycle duration
Demand rate

With this solution, there are some losers, but overall all win. This is another example of the Braess' paradox.

Smith's paradox illustrates the fact that when dealing with the control of networks with route choice, an overall perspective is needed. Detailed optimal control measures applied to some nodes of the network maybe detrimental when considering the network wide effects. These concepts could also be applied to signal coordination in a street grid, like the one in Figure 17.

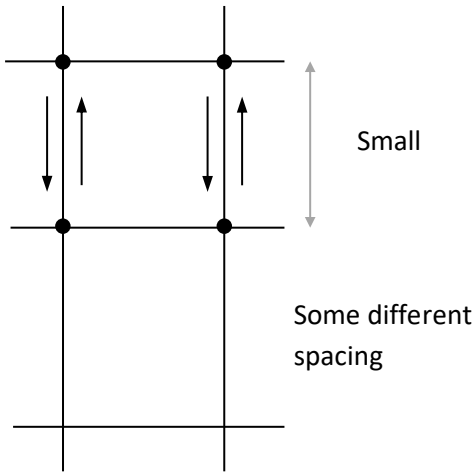


Figure 17. Signal coordination in a non-uniform street grid

Assume heavy demand in both (vertical) directions, and a non-uniform spacing between horizontal streets. In this context, obtaining effective green waves for all directions (verticals) and streets it is impossible[6]. In even more complex networks, there exist fancy software packages that say they that help. The reality is that signal coordination in these situations is non-effective, and myopic actuated controls results in a complete mess.

In such situations, the traffic engineers' solution would be to change two-way streets to one-way streets, so that it is easier to construct effective green waves. However, it is very plausible that merchants, neighbors and other stakeholders argue, because of the loose of accessibility at mid-block locations in case of one-way streets. Note

[6] You can illustrate this fact by using trajectories analysis.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

that you need to make a detour to reach mid-blocks from one direction, meaning less accessibility and less attractive locations.

A smart solution would be to allow traffic in both directions, but coordinate signals so that there is only a green wave for one. Then you would inform through traffic of the recommended routes with green waves and they will switch.

To conclude this network control discussion, remember to avoid control schemes based on myopic input data in highly demanded networks, and always look for system wide effects of control strategies.

APPENDIX 1: Inverse cdf method for generating random numbers

The inverse cdf method allows generating random numbers following a specific distribution from a $U[0,1]$ random number generator. This can be useful in case the software package used does not have implemented a random number generator with the probability distribution of interest.

The concept is simple. The cdf (i.e. the cumulative probability distribution function) represents the cumulative probability of the target distribution. The cdf is bounded between 0 and 1, so that if we apply the inverse of the cdf to a random $U[0,1]$ value, we would obtain a random value according to the desired distribution.

Below (see Figure 18) we present an example for the generation of $U[-0.2,0.2]$ random numbers as needed in the Smith's paradox example. $f_{\Delta t}(\Delta t)$ is the probability density function (i.e. the pdf) of the $U[-0.2,0.2]$ distribution for the random variable Δt. $F_{\Delta t}(\Delta t)$ is the cdf, which is obtained as the integral of the pdf.

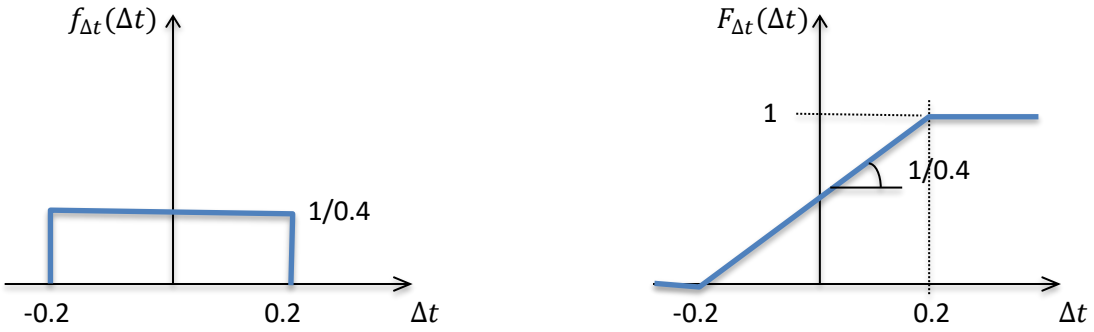


Figure 18. $U[-0.2, 0.2]$ probability density & cumulative probability functions

The analytical expression for the $U[-0.2,0.2]$ cdf is:

$$F_{\Delta t}(\Delta t) = \frac{1}{0.4}(\Delta t + 0.2) \quad \text{for} \quad \Delta t \in (-0.2, 0.2)$$
$$0 \quad \text{otherwise}$$

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Civil and Environmental Engineering

Francesc Soriguera
Associate Professor

Jordi Girona 1-3, Office B1-114
08034 Barcelona
Tel. (+34) 93 401 72 66
< francesc.soriguera@upc.edu >
< fsoriguera.com >

Then, the inverse cdf is[7]:

$$\Delta t = 0.4 \cdot F_{\Delta t}(\Delta t) - 0.2$$

So, that the random number generator for the $U[-0.2,0.2]$ distribution is:

$$\Delta t = 0.4[rand(0,1)] - 0.2$$

Where $rand(0,1)$ is a random number generated with a $U[0,1]$ distribution.

[7] Determining the inverse of the cdf function is the critical part of the method. It might be difficult or even impossible to find an explicit analytical solution for some distributions.