Introduction to Transportation Systems
Lecture Notes (draft)

Michel Bierlaire

April 30, 2025

Contents

Preface

These lecture notes accompany the course *Introduction to Transportation Systems*, offered to bachelor students in civil engineering at EPFL.

The course begins by situating transportation systems within a broader societal and economic context. It examines how costs and externalities influence transportation decisions and outcomes. The discussion then moves to engineering aspects, introducing performance indicators used to assess system design, operational efficiency, and maintenance effectiveness. These ideas are illustrated through a series of case studies focused on vertical transportation, road maintenance, and shuttle operations.

The next part introduces fundamental economic principles relevant to transportation. Concepts such as equilibrium, elasticities, and consumer surplus are presented to explain how demand responds to changes in price and service levels. Different behavioral assumptions are discussed, providing a basis for understanding how individuals make transportation-related choices.

Discrete choice theory is introduced as a method for representing decision-making among alternatives, with particular attention given to the notion of value of time. This framework is essential for modeling traveler preferences and evaluating transport policies.

The course then presents mathematical modeling approaches, explaining how models are constructed and interpreted. The treatment of causality is followed by an overview of model development strategies for different types of variables — discrete, continuous, and mixed.

A general overview of travel demand is provided, introducing the concepts of trip production and attraction. Methods for constructing origin-destination matrices are described, including challenges such as under-determination and incompatibility. The structure of travel behavior is further detailed through the analysis of mode and route choice, as well as the components of the four-step model. These include trip generation, distribution, mode choice, and assignment.

Transportation networks are examined across various modes, including road systems, public transport, pedestrian infrastructure, and multimodal

configurations. The representation of paths within these networks is discussed, along with methods for analyzing accessibility and connectivity.

The four-step model is further developed, with attention to data requirements and calibration. Techniques for collecting and interpreting survey data and traffic counts are outlined. The gravity model is introduced as an approach to model trip distribution. Modal split is examined in more detail through logit models, including treatment of choice data and behavioral heterogeneity.

Traffic assignment models are explored next, starting with all-or-nothing assignment and continuing with user equilibrium formulations. Beckmann's model is presented along with its algorithmic solution and its equivalence with equilibrium conditions. Examples illustrate the computational process. Conceptual frameworks such as Braess's paradox and the prisoner's dilemma are used to highlight tensions between individual decisions and collective outcomes. The idea of system optimum is introduced as a target for network performance.

The course concludes with an introduction to congestion pricing, showing how pricing mechanisms can be used to steer demand towards more efficient outcomes. The theoretical underpinnings are linked to practical implementation challenges.

A final component introduces freight transportation. Topics include facility location problems, inventory management under fixed and variable consumption scenarios, and vehicle routing. These are presented through numerical examples and scenario analysis, emphasizing the role of spatial and temporal constraints in freight logistics.

Chapter 1

Introduction

Transportation systems play a fundamental role in modern society, serving as the backbone of economic, social, and environmental well-being. They enable the efficient movement of people and goods, fostering productivity, accessibility, and sustainability. The primary objectives of transportation systems revolve around enhancing mobility and promoting accessibility, each contributing to broader societal benefits.

Mobility is a key goal of transportation systems, as saving travel time is a critical resource for individuals and businesses alike. Increased productivity is one of the major benefits of efficient transportation networks. When commuting times are reduced, individuals can allocate more time to work, education, or personal development. For instance, well-functioning metro systems in major cities allow workers to reach their offices more quickly, leading to higher economic output. Economic growth is another significant impact, as businesses benefit from faster deliveries and reduced transportation costs. High-speed rail networks, such as France's TGV, have drastically shortened travel times between major cities, fostering business interactions and regional economic integration.

Beyond economic benefits, efficient transportation can contribute to environmental sustainability. Shorter travel times translate to lower fuel consumption and fewer emissions. Investments in express bus lanes or optimized traffic signals help to alleviate congestion and reduce urban air pollution. Additionally, improvements in transportation systems enhance work-life balance, as employees with shorter commutes have more time for family, leisure, and rest, ultimately improving their mental and physical well-being. Cities that promote flexible transit[1] options provide a better quality of life for their residents. Another important aspect of mobility is its role in emergency re-

[1] "transit" is synonymous with "public transportation systems."

sponse efficiency. Well-designed transportation networks ensure that ambulances, fire trucks, and law enforcement agencies can reach their destinations faster, which can be critical in life-threatening situations.

Accessibility is another fundamental objective of transportation systems, as it promotes travel and thereby contributes to social and economic development. Social inclusion is enhanced when public transport is made accessible to all individuals, including those with disabilities, the elderly, and low-income populations. Vienna, for example, has implemented fully accessible trams and buses, significantly improving social equity (Emberger et al., 2013). Economic opportunities are also expanded through well-developed transportation networks. By connecting individuals to better job prospects, particularly those in low-income or suburban areas, mobility contributes to reducing economic disparities. Bogotá's TransMilenio Bus Rapid Transit (BRT) system has successfully improved job accessibility for residents in peripheral areas (Hidalgo et al., 2013).

Transportation infrastructure also plays an important role in public health. Cities that prioritize pedestrian-friendly environments and cycling infrastructure, such as Amsterdam, encourage active modes of transportation, reducing obesity and related health risks. Additionally, well-integrated transport networks strengthen urban-rural connections, providing rural populations with better access to healthcare, education, and markets for agricultural goods. Efficient transportation systems also have a direct impact on tourism and local businesses. Well-connected cities attract more visitors and customers, thereby stimulating local economies and enhancing cultural exchange.

To summarize, transportation systems serve as a vital component of modern society, with far-reaching implications for productivity, economic growth, environmental sustainability, social inclusion, and public health. By prioritizing both mobility and accessibility, policymakers and planners can ensure that transportation networks contribute to a more efficient, equitable, and sustainable society.

Transportation has evolved significantly over millennia. A simplified timeline is reported in Table 1.1.

From a technological point of view, while the airplane, introduced in 1904, was a major breakthrough, subsequent advances have been largely incremental. No fundamentally new mode of transportation has emerged in recent times. On the contrary, there is an emerging trend toward promoting slower, sustainable modes (e.g., walking). This is surprising, compared to the huge technological progresses in computers and communication technologies.

In order to understand why, let's explore an interesting technology: the *magnetic levitation (maglev) train* that operates in Shanghai. This train connects Shanghai Pudong International Airport to Longyang Road station

4000 BC	Horses
3500 BC	Wheel, river boats
2000 BC	Chariots
312 BC	Paved roads (Romans)
1662	Horse-drawn public bus
1783	Hot air balloon
1801	Steam road locomotive
1814	Steam-powered railway train
1816	Bicycle
1900	Airship (Zeppelin)
1904	Airplane
1908	Ford car

Source: www.twinkl.ae/teaching-wiki/transportation

Table 1.1: Historical milestones in transportation.

over a distance of *30 km*. It is designed to reach a maximum speed of *430 km/h*, meaning that if it traveled at full speed for the entire journey, the travel time would be approximately *4 minutes* (source: Wikipedia).

However, in reality, the trip takes *8 minutes*. Why is that? The train needs time to accelerate and decelerate, meaning it cannot maintain its top speed for the entire journey. Because of this, the *average speed* over the 30 km trip is *225 km/h*.

We also need to consider another important factor: the *headway*. The headway is the time between two consecutive trains. In this case, the maglev runs every *15 minutes*, meaning that the waiting time for passengers is *7.5 minutes* on average. This waiting time affects the *effective travel time*, which now becomes:

$$8 \text{ minutes (train ride)} + 7.5 \text{ minutes (average waiting time)} = 15.5 \text{ minutes} \tag{1.1}$$

From this, we can compute the *overall average speed*, considering both the travel and waiting time:

$$\frac{30 \text{ km}}{15.5 \text{ min}} \approx 116 \text{ km/h}. \tag{1.2}$$

This is *much lower* than the potential speed of 430 km/h! Even though the maglev train is an impressive technology, its *advantages are underutilized* because of the short distance and waiting time.

Would it be more beneficial if it were used for longer distances, where it could maintain high speeds for a greater portion of the trip? Given that the construction cost of the Shanghai maglev is reported to be *40 million USD per kilometer*, we might ask: Is such an expensive technology justified for a short route?

In Chapter 11, we explore *cost-benefit analysis*, a tool used to assess projects like this one. This method helps determine whether the costs of a project are justified by the benefits it brings, considering factors like travel time savings, passenger demand, and alternative transportation options.

An additional aspect to consider is the *travel demand*. Indeed, most passengers do not have Longyang Road station as their final destination. Instead, they must continue their journey using another mode of transportation, such as a bus, taxi, or metro, to reach their actual destination. This additional transfer increases the total travel time and reduces the overall convenience of the trip.

In contrast, a taxi ride from the airport, which takes about 30 minutes and brings passengers directly to their final destination, may be preferable for many travelers due to the door-to-door convenience and flexibility it offers. This highlights an important consideration in transportation planning: speed alone does not determine the attractiveness of a mode of transport. While the maglev train represents a significant technological achievement, its competitiveness depends on how well it integrates into the broader transportation network and meets passengers' needs in terms of accessibility and overall travel efficiency.

1.1 Costs and Externalities

Transportation systems involve a variety of costs, which can be distributed among different stakeholders. These costs are typically covered by three main groups:

- *Travelers*, who bear expenses such as fares.
- *Transport operators*, who manage the infrastructure and provide services, incurring costs for maintenance, labor, and operations such as vehicle acquisition, wages and administrative expenses.
- *Governments (i.e., taxpayers)*, who finance public infrastructure projects, subsidies, and policy enforcement.

Recognizing how these costs are allocated is important in decision-making, particularly in *cost-benefit analysis*. A key challenge in transportation plan-

ning is that those who *pay* for a transportation system are not always those who *benefit* the most from it.

For instance, if a city plans to build a *new metro line*. The financial implications for each stakeholder include:

- *Travelers*: Pay for tickets but benefit from shorter commute times.
- *Operators*: Invest in trains, staff, and maintenance.
- *Government*: Provides funding but expects long-term benefits like reduced congestion and pollution.

If the *taxpayer-funded costs* exceed the benefits to society, the project may not be justified unless external benefits (e.g., sustainability) outweigh the direct financial losses.

Transportation systems generate effects beyond direct monetary costs. These additional consequences, known as *externalities*, represent impacts on society that are not directly accounted for in the pricing of transport services.

An *externality* is a side effect or consequence of an activity that affects other parties without being reflected in the market price. Externalities can be either *positive* or *negative*, influencing various aspects of society and the environment.

Positive externalities generate benefits for society without direct compensation. Examples include:

Social Growth: Improved accessibility fosters social interactions and community engagement.

Economic Growth: Efficient transport networks stimulate business development, job creation, and market expansion.

Equity and Accessibility: Public transportation enhances mobility for disadvantaged populations, reducing social inequalities.

Negative externalities impose costs on society without corresponding payment. Examples include:

Pollution: Vehicle emissions contribute to air and water pollution, leading to health and environmental damages.

Energy and Space Consumption: Road congestion and land use for infrastructure reduce urban efficiency.

Noise and Safety Concerns: High traffic volumes increase noise pollution and the risk of accidents.

Inequity: Disparities in transportation accessibility can reinforce economic and social inequalities.

While externalities are often not priced into transport systems, they play a critical role in policy-making. Governments use measures such as *taxes, subsidies, and regulations* to mitigate negative externalities and encourage positive ones.

As an example, consider space consumption. Indeed, transportation infrastructure occupies a significant portion of land, impacting urban development, environmental sustainability, and land availability for other uses. In Switzerland, transport infrastructure covers approximately 800 km^2, representing about 2% of the country's total territory.

The space required for transport infrastructure competes with other essential land uses. One-third of Switzerland's surface is dedicated to housing and infrastructure, highlighting the challenge of balancing mobility needs with environmental and urban planning considerations. The country has an extensive road network of 84,000 km and a railway system spanning 5,200 km. While rail transport is generally more space-efficient than roads, both contribute to landscape fragmentation, ecological disruption, and urban sprawl. Managing these externalities requires policies that promote land-efficient mobility solutions, such as multimodal transport integration, compact city planning, and investment in high-capacity public transit systems.

1.2 Engineering Challenges and Performance Indicators

Engineers play an important role in transportation systems, ensuring their efficiency, safety, and sustainability. Their work spans multiple stages, from the initial design and planning to the continuous operation and maintenance of infrastructure. Each stage presents unique challenges that require specialized knowledge and innovative solutions.

In the *design* phase, engineers focus on long-term planning and construction. This involves designing road networks, rail systems, and public transportation infrastructure to meet growing mobility needs while considering environmental impact, land use, and economic feasibility. Engineers use advanced modeling techniques to optimize traffic flow, reduce congestion, and improve accessibility, ensuring that transportation systems are resilient and adaptable to evolving demands.

Beyond the initial construction, engineers are responsible for the *maintenance* of transportation infrastructure. Roads, bridges, tunnels, and railway

tracks require regular inspections and repairs to ensure safety and longevity. Well-maintained infrastructure reduces the risk of accidents, minimizes economic disruptions, and extends the lifespan of costly investments.

In the *operations* phase, engineers oversee the day-to-day functioning of transportation systems. This includes traffic management, railway signaling, and optimizing public transit schedules to maximize efficiency. Engineers leverage real-time data to monitor network performance, adjust capacity during peak hours, and integrate emerging technologies such as autonomous vehicles and intelligent traffic systems. Their expertise ensures that transportation networks remain reliable, responsive, and aligned with user needs.

1.2.1 Performance Indicators

To ensure efficient and sustainable transportation systems, engineers must develop objective performance indicators that help evaluate their effectiveness. These indicators provide a quantitative basis for decision-making, allowing engineers to assess trade-offs between different aspects of system performance. A well-functioning transportation network should balance multiple, sometimes conflicting, criteria to meet user needs, minimize costs, and reduce societal impacts.

One key measure is the *level of service*, which encompasses travel time, comfort, convenience, and flexibility. Engineers analyze factors such as congestion levels, frequency of public transit, and overall accessibility to ensure that transportation systems provide a reliable and user-friendly experience. At the same time, *costs* must be carefully managed, including those associated with infrastructure design, maintenance, and daily operations. Engineers work to optimize efficiency while ensuring that investments are financially sustainable. Beyond direct costs and service quality, transportation systems also generate *externalities*, which can be either positive, such as economic development and improved connectivity, or negative, such as environmental pollution and noise. By incorporating these factors into their evaluations, engineers can design systems that are not only functional but also equitable and environmentally responsible.

1.3 Design case study: vertical transportation

A common design challenge faced by engineers is ensuring efficient mobility and accessibility in high-rise buildings. Consider the case of a 40-story building with no internal transportation system, such as elevators. Without an

efficient means to move between floors, significant mobility and accessibility issues arise, directly affecting the building's usability and appeal.

From a mobility perspective, reaching the upper floors solely by stairs is highly impractical. If an individual were to walk at a fast pace, taking approximately 15 seconds per floor, reaching the top floor would require around 10 minutes. At a slower pace of 22 seconds per floor, the total time would increase to approximately 15 minutes. This prolonged travel time makes daily movement within the building inconvenient, particularly for individuals who need to access higher floors multiple times a day. The lack of a fast and efficient transportation solution could significantly reduce the building's functionality for both residential and commercial use.

Beyond mobility concerns, accessibility also becomes a major issue. If reaching the top floors is time-consuming and physically demanding, these spaces may become undesirable for tenants, leading to economic inefficiencies. Prospective residents and businesses are unlikely to rent or purchase units on higher floors if they are deemed too difficult to access. This creates a fundamental design problem: without an effective transportation system, a significant portion of the building may remain underutilized, ultimately diminishing its economic viability.

To address this challenge, engineers must design a transportation system that ensures both mobility efficiency and universal accessibility. Elevators become an essential component, but their placement, capacity, and speed must be carefully optimized to prevent congestion and long wait times.

One possible solution to address the mobility and accessibility challenges in a 40-story building is to install an elevator system that minimizes travel time. A simple yet extreme approach would be to provide one dedicated direct elevator for each floor. Assuming that an elevator takes approximately five seconds per floor, a trip from the ground floor to the top floor would take about 200 seconds, or roughly three minutes. This significantly reduces the time required for vertical movement compared to using stairs, making the building far more functional and accessible.

However, the practicality of such a design raises several important considerations. If each floor were to have its own dedicated elevator, the total number of elevators would be 40, which would require an enormous amount of space and entail prohibitively high construction and maintenance costs. On the other hand, having only one shared elevator for the entire building may create excessively long waiting times, especially during peak hours, leading to congestion and inefficiencies. Engineers must therefore strike a balance between the level of service provided and the financial and spatial constraints of the system.

Another key factor in the design of the elevator system is understanding

demand patterns. Not all floors may have the same level of importance or usage. For instance, if there is a secondary entrance at the fourth floor, a significant number of users may not require transportation from the ground level. Similarly, if the building includes an observation deck at the top floor, there may be higher demand for direct access to that level. These factors highlight the necessity of a well-planned elevator network that considers user behavior, travel patterns, and operational efficiency to optimize both cost and service quality.

The same considerations that apply to the design of vertical transportation systems in buildings also extend to horizontal transportation networks in cities and regions. Just as elevators must balance travel time, cost, and capacity, road and public transit systems must be designed to efficiently manage demand, minimize congestion, and optimize infrastructure use. A citywide transportation system that builds a separate road for every possible route would be unrealistic due to excessive space consumption, high construction costs, and environmental impact. Similarly, having only one road or transit line serving an entire city would result in severe congestion and long travel times. Engineers must therefore strike a balance, designing an optimal number of roads, rail lines, and public transit routes that efficiently serve demand while minimizing costs and negative externalities. Additionally, just as some floors in a building experience more traffic than others, certain areas in a city—such as business districts, commercial centers, or major transit hubs—require more transportation capacity than low-density residential areas. Understanding travel demand patterns, implementing multimodal solutions, and leveraging smart traffic management technologies are important for ensuring that horizontal transportation systems remain efficient, accessible, and sustainable.

1.4 Maintenance case-study: road maintenance

One fundamental aspect of transportation system management is maintenance, which ensures infrastructure remains functional, safe, and efficient over time. Consider the example of road maintenance, where the quality of a road surface degrades at a constant rate over time due to traffic loads, weather conditions, and material wear. Assume that the road quality is characterized by a "pavement condition index" (Setyawan et al., 2015), ranging from g_0 to $\mathsf{g}_{\max}$, where $\mathsf{g}_{\max}$ represents a newly paved road and g_0 indicates a failed pavement.

If we define the initial road quality as $\mathsf{g}_0 = 0$ and assume that degradation occurs at a constant rate τ, then the quality of the road will progressively

decline unless repairs are carried out. The cost of repairing the road at a given time t consists of two components: a fixed cost c_f that is independent of when maintenance is performed, and a variable cost $c_v \tau t$ that increases with the level of deterioration.

Given these dynamics, an important question arises: how frequently should maintenance be performed to minimize overall costs while keeping the road in an acceptable condition? If repairs are conducted too frequently, expenses accumulate due to repeated fixed costs, and resources may be wasted on roads that are still in good condition. Conversely, if maintenance is delayed too long, degradation becomes severe, leading to higher variable costs and potentially requiring extensive reconstruction rather than simple resurfacing. Engineers must therefore determine an optimal maintenance schedule that balances cost efficiency with infrastructure longevity, ensuring that roads remain safe and operational while minimizing financial and logistical burdens.

A road maintenance model relies on several key components to determine the optimal strategy for preserving pavement quality while minimizing costs. One example of such a model is represented in Figure 1.1, where the x-axis represents time, and the y-axis represents the pavement condition index.

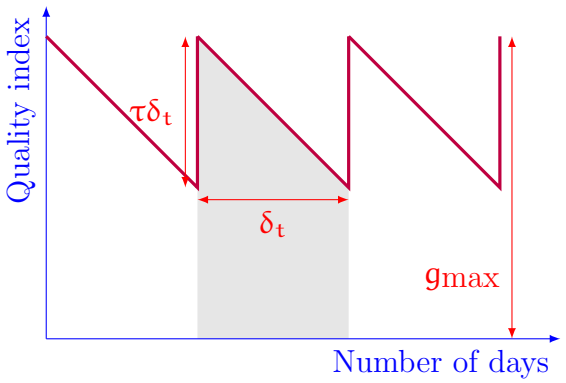


Figure 1.1: A model for road maintenance

The deterioration rate, denoted as τ, represents the rate at which the road surface degrades over time due to factors such as traffic load, weather conditions, and material wear. Engineers must decide on the appropriate maintenance interval, δ_t, which defines how frequently maintenance interventions should be performed to maintain an acceptable road quality level. The planning horizon, t_H, represents the total period over which maintenance decisions are evaluated.

The interval δ_t must be chosen such that the pavement condition does not deteriorate beyond $g_0 = 0$, that is

$$\delta_t \leqslant g_{max}/\tau.$$

Given a fixed time horizon, the total number of maintenance interventions required is determined by t_H/δ_t. During each interval, road quality follows a predictable pattern, starting at the maximum level, g_{max}, and gradually declining due to wear and tear. The average quality over one interval is given by the surface under the curve, that is calculated as

$$g_{max}\delta_t - \frac{\tau\delta_t^2}{2}.$$

Summing over the full horizon results in a total quality measure of

$$t_H(g_{max} - \frac{\tau\delta_t}{2}).$$

Each intervention consists of a fixed cost, c_f and a variable cost, $c_V\tau\delta_t$. Over the entire planning horizon, the total cost of maintenance is expressed as

$$t_H(c_f/\delta_t + c_V\tau).$$

This formulation illustrates the fundamental trade-off in road maintenance planning: frequent but lower-cost interventions versus less frequent yet more expensive repairs. If maintenance is performed often, each intervention may require only minor repairs, reducing variable costs per intervention but increasing the number of interventions over the planning horizon. On the other hand, if maintenance is postponed for longer periods, the extent of road deterioration before each repair is greater, leading to higher costs per intervention, even though the total number of interventions is lower.

An important observation is that the contribution of variable costs remains independent of the chosen maintenance interval, δ_t. Regardless of how frequently maintenance is carried out, the total degradation accumulated over the entire time horizon must eventually be repaired. This means that while the timing of interventions affects the distribution of costs over time, it does not change the overall variable cost associated with repairing the accumulated wear and tear. The key decision, therefore, is to determine the optimal balance between the timing of interventions and the associated fixed costs, ensuring that the road remains in good condition while minimizing long-term expenses.

Figure 1.2 presents a numerical example that illustrates the relationship between the total quality index and the total cost as a function of the interval

between two maintenance interventions. The total quality index, represented on the left axis, decreases linearly as the interval increases. This reflects the fact that the longer the period between interventions, the more the road deteriorates before being repaired, leading to a steady decline in overall quality. In contrast, the total cost, shown on the right axis, follows an inverse relationship with the interval length. Initially, frequent interventions lead to high costs due to recurring fixed expenses, but as the interval increases, the cost per unit of time decreases significantly.

A key observation from the figure is that beyond approximately 100 days, extending the interval further has little effect on total cost, as the marginal reduction in cost becomes negligible. However, road quality continues to decline at a constant rate, indicating that longer intervals compromise road conditions without substantial financial benefit. This type of visualization provides valuable insights for engineers responsible for maintenance planning. By analyzing the trade-off between cost and quality, they can determine the most appropriate maintenance frequency that balances infrastructure longevity with economic efficiency.

$\tau = 1$, $g_{max} = 100$, $t_H = 365$, $c_f = 100$, $c_v = 5$

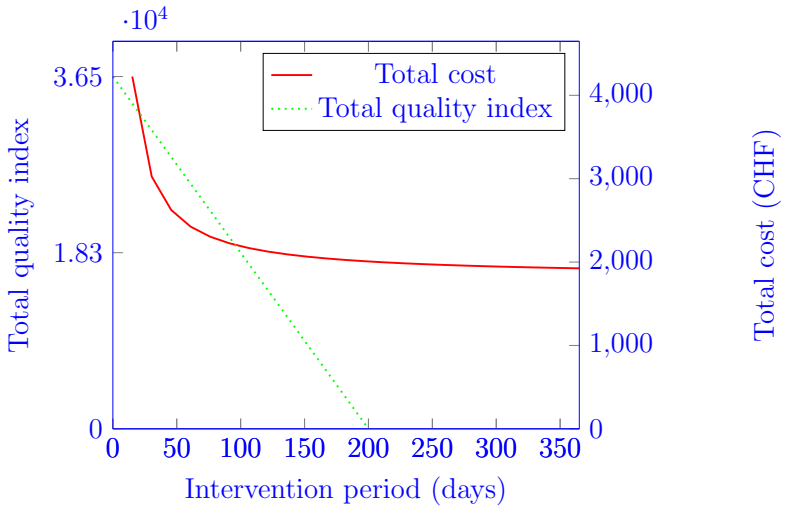


Figure 1.2: Trade-off between maintenance costs and quality index

1.5 Operations case-study: shuttle service

Consider the operation of a shuttle service designed to transport spectators from a stadium to a parking area at the end of a football game. Once the game concludes, a large number of people exit the stadium and need transportation. To accommodate this demand efficiently, a shuttle service is implemented, running at regular intervals to transfer spectators in a timely manner.

The key operational challenge is determining the optimal frequency at which the shuttles should be deployed. The flow of people leaving the stadium is assumed to be uniform, with a rate of f persons per minute. This means that every minute, a fixed number of spectators arrive at the shuttle boarding area, awaiting transport. If the service frequency is too low, long queues will form, increasing waiting times and causing dissatisfaction among passengers. Conversely, if the service frequency is too high, shuttles may operate with many empty seats, leading to inefficient resource utilization and unnecessarily high operating costs.

Each shuttle incurs a fixed cost of operation, denoted as c. Running more frequent shuttles increases total costs, while running fewer shuttles reduces costs but may compromise service quality by extending passenger waiting times. The goal is to strike a balance between cost efficiency and passenger convenience.

Figure 1.3 provides a graphical representation of the shuttle operations, illustrating how passengers are transported over time. The horizontal axis represents time, while the vertical axis corresponds to the cumulative number of travelers.

The red line in the figure represents the demand, indicating the total number of passengers requesting transportation. Since spectators leave the stadium at a uniform rate, this line increases steadily over time. The purple curve, on the other hand, represents the actual number of passengers transported. This curve exhibits two distinct behaviors: horizontal segments, which correspond to the periods when the shuttle is stationed at the stadium and passengers are boarding, and vertical segments, which represent the actual journey to the parking area. For the sake of simplicity, the travel time is assumed to be instantaneous in this illustration, meaning that once a shuttle departs, it immediately reaches its destination.

Each point on the vertical axis corresponds to a specific passenger, and the horizontal gap between the red demand curve and the purple transport curve indicates the waiting time experienced by that passenger. The larger this horizontal gap, the longer the individual waits before being accommodated on a shuttle. As a result, the grey triangular area between the two

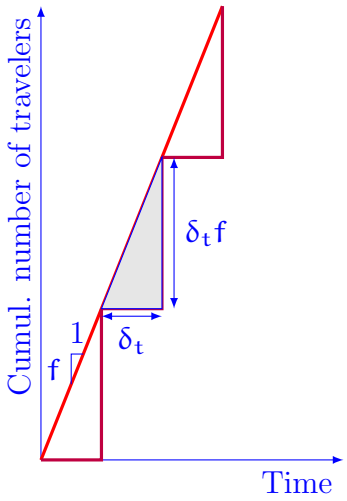


Figure 1.3: Illustration of the operations of the shuttle service

curves represents the total waiting time accumulated by all passengers within a single shuttle departure cycle. This waiting time is a function of the shuttle headway, denoted as δ_t, which defines the time interval between two consecutive shuttle departures.

The shuttle operation can be modeled mathematically to analyze the trade-off between operational costs and passenger waiting times. The model assumes that travelers arrive at a constant rate of f passengers per minute, and the entire transportation process takes place over a planning horizon of t_H minutes. Given this arrival rate, the total number of travelers requiring transportation during the entire period is ft_H.

The shuttle headway, δ_t directly influences both the frequency of shuttle trips and the service quality experienced by passengers. The total number of shuttle trips required over the horizon is t_H/δ_t, ensuring that all passengers are transported within the specified period. Each shuttle trip accommodates $\delta_t f$ passengers, meaning that the number of passengers per vehicle is determined by how frequently the shuttles operate.

From a cost perspective, each shuttle trip incurs a fixed operational cost of c CHF. Consequently, the total cost of operating the shuttle service over the entire horizon is given by $t_H c/\delta_t$, indicating that as the headway increases, the total cost decreases, since fewer trips are required.

Since travelers arrive continuously at a uniform rate, the waiting time for a given passenger depends on their arrival time relative to the next available shuttle departure. As explained above, the total waiting time per trip is the area of the grey triangle, that is $\delta_t^2 f/2$ passenger-minutes. Aggregating over

the full planning horizon, the total waiting time for all passengers is given by $t_H \delta_t f/2$ passenger-minutes. This formulation highlights a fundamental trade-off: a larger headway reduces the total operational cost but increases total waiting time, whereas a smaller headway improves passenger service but raises costs.

Figure 1.4 provides a visual representation of the trade-off between operational cost and passenger waiting time in the context of the shuttle service. The horizontal axis represents the headway, measured in minutes, which corresponds to the time interval between consecutive shuttle departures. The vertical axis on the left displays the total operational cost, while the vertical axis on the right shows the total accumulated passenger waiting time. This figure illustrates a specific scenario in which the planning horizon is set to $t_H = 60$ minutes, with an arrival rate of $f = 60$ passengers per minute, and a fixed cost of $c = 200$ CHF per shuttle trip.

As the headway increases, the total cost follows a decreasing trend, reflecting the fact that fewer shuttles are needed when departures are spaced farther apart. Since each trip incurs a fixed operational cost, reducing the number of trips results in lower overall expenses. However, this reduction in cost comes at the expense of increased passenger waiting time. With a longer headway, passengers must wait longer before boarding a shuttle, leading to a proportional increase in total waiting time. This relationship is depicted on the right vertical axis, where the total waiting time grows as the headway increases.

The figure clearly demonstrates the fundamental trade-off in shuttle operations: reducing costs by increasing the headway leads to a decline in service quality, while minimizing waiting time requires running more frequent trips, thereby increasing costs. This type of graphical representation is particularly useful for transportation planners, as it provides a concrete means of evaluating different operational strategies.

1.6 Summary

The development and management of transportation systems present a range of engineering challenges that require careful consideration at every stage. Engineers are responsible for designing and constructing infrastructure that meets current and future needs, ensuring that roads, rail networks, and public transport systems are both functional and sustainable. Once built, these systems must be maintained to preserve their quality and reliability, requiring strategic planning to minimize degradation while optimizing costs. In addition, the daily operation of transportation services involves complex co-

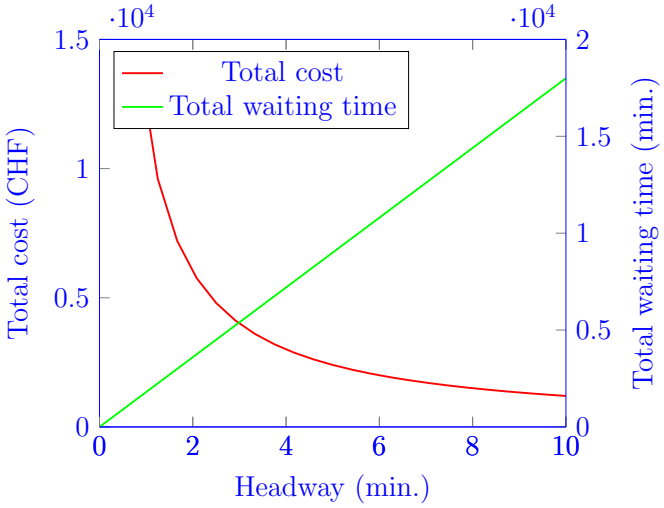


Figure 1.4: Trade-off between costs and level of service

ordination to ensure efficiency, safety, and accessibility for all users.

To assess the performance of a transportation system, engineers rely on various indicators that provide valuable insights into its effectiveness. One of the most critical measures is the level of service, which evaluates the quality of transportation from the user's perspective, including factors such as travel time, reliability, and comfort. Costs are another essential consideration, encompassing both direct expenses, such as construction and operational costs, and long-term maintenance expenditures. Beyond these financial aspects, externalities must also be taken into account, including both positive and negative effects on society, such as environmental impact, congestion, and social equity. Since these indicators are measured in different units—time, money, environmental impact, and social factors—transportation planning involves complex trade-offs that require a multidimensional approach to decision-making.

Transportation systems also involve multiple stakeholders, each with different interests and priorities. Travelers seek affordable, convenient, and efficient mobility solutions that minimize their travel time and maximize accessibility. Transport operators are responsible for delivering services efficiently while managing costs and maintaining infrastructure. Governments and taxpayers play an important role in funding transportation projects, ensuring that investments serve the broader public interest and contribute to economic growth and sustainability.

Understanding the role of demand is fundamental in transportation plan-

ning. Identifying who needs the service, who will benefit from it, and who is willing to pay for it helps shape investment decisions and operational strategies. The design of transportation systems must align with actual user demand to ensure that resources are allocated effectively and that services are both financially viable and socially beneficial. By integrating engineering principles with a thorough analysis of demand and stakeholder interests, transportation planners can create systems that are efficient, sustainable, and responsive to the evolving needs of society.

Chapter 2

Fundamentals

In this chapter, we review the fundamental concepts that underpin the analysis of transportation systems, drawing on principles from microeconomics. Transportation systems are complex, involving interactions between infrastructure, users, and service providers, and their analysis requires a rigorous framework to understand how travel choices emerge and how transport services operate. By leveraging economic theories and behavioral models, we can develop tools to evaluate system performance, predict demand, and assess the impact of policy interventions.

A central concept in transportation analysis is *equilibrium*, which describes a state in which no traveler has an incentive to unilaterally change their behavior. Understanding equilibrium dynamics is important for designing efficient and fair transportation policies, whether in road networks, public transit systems, or multimodal mobility solutions.

A central aspect of transportation system analysis emphasized in this course is understanding travel demand, as it serves as the foundation for evaluating system performance, planning infrastructure investments, and designing effective policies. Travel demand reflects the choices individuals make regarding when, where, and how to travel, based on factors such as cost, travel time, convenience, and personal preferences.

To quantify and analyze travel demand, we introduce two key economic indicators: demand elasticities and consumer surplus. Demand elasticities measure the sensitivity of travel demand to changes in key variables, such as travel cost or travel time. For example, price elasticity of demand captures how a change in transport fares influences ridership levels, while time elasticity reflects how variations in travel time affect mode choice.

Another fundamental concept is consumer surplus, which represents the economic benefit that travelers receive from using a transportation service beyond what they actually pay for it. In essence, consumer surplus quan-

tifies the difference between the maximum amount a traveler is willing to pay for a trip and the actual cost incurred. This measure is particularly useful for evaluating the social benefits of transportation projects, as it helps policymakers assess whether an investment improves overall welfare. For instance, a new public transit line that reduces travel costs and travel time for commuters generates additional consumer surplus by making transportation more affordable and efficient.

Finally, we explore the behavioral foundations of travel demand, which help explain how individuals make decisions about when, where, and how to travel. Travel choices depend on a variety of factors, including economic constraints, time availability, personal preferences, and social influences. By incorporating behavioral models into transportation analysis, we can better predict demand patterns, assess the impact of policy measures, and design systems that align with user needs. The integration of these concepts—equilibrium theory, quantitative performance indicators, and behavioral modeling—forms the foundation for a systematic and comprehensive approach to transportation system analysis.

2.1 Equilibrium

Consider[1] a flight from Geneva (GVA) to Tenerife (TFS), where ticket pricing is influenced by the principles of supply and demand. The airline uses a dynamic pricing strategy in which the price of a seat depends on the level of demand, represented by the number of passengers willing to book the flight. Specifically, the ticket price follows the function:

$$\mathsf{p} = 200 + 0.02\mathsf{q}.$$

This equation indicates that the base price of a ticket is 200 CHF, and for each additional passenger willing to purchase a seat, the price increases by 0.02 CHF. This reflects a common airline pricing mechanism, where higher demand leads to higher fares, as fewer seats remain available.

At the same time, the popularity of the flight, measured by the number of passengers q willing to buy a ticket, depends on the ticket price. As prices rise, fewer travelers choose to book the flight. This relationship is modeled by the demand function:

$$\mathsf{q} = 5000 - 20\mathsf{p}.$$

[1] Example inspired by Khisty and Lall, 2003.

This equation suggests that if the ticket price were to rise, fewer passengers would be interested in flying, while lower prices would attract more travelers. The parameters in the equation indicate that at a price of 0 CHF (hypothetically), 5000 passengers would be interested in the flight, while for every 1 CHF increase in price, the demand drops by 20 passengers.

To determine the equilibrium price and the number of passengers, we solve the system of equations by substituting the first equation into the second:

$$\mathsf{q} = 5000 - 20(200 + 0.02\mathsf{q}).$$

Expanding and solving for q:

$$\mathsf{q} = 5000 - 4000 - 0.4\mathsf{q}.$$

$$\mathsf{q} + 0.4\mathsf{q} = 1000.$$

$$1.4\mathsf{q} = 1000.$$

$$\mathsf{q} = \frac{1000}{1.4} \approx 714.$$

Now, substituting this value into the price equation:

$$\mathsf{p} = 200 + 0.02 \times 714.$$

$$\mathsf{p} = 200 + 14.28 = 214.28.$$

Thus, the equilibrium ticket price is approximately 214.28 CHF, and the number of passengers who will fly on this route is around 714. This example illustrates the fundamental economic interaction between price and demand, showing how airlines adjust fares dynamically based on booking levels. By understanding these principles, transportation planners and airline operators can optimize pricing strategies to balance profitability and passenger demand.

To compare the demand and the supply functions directly, it is necessary to rewrite the demand function in a form that expresses p as a function of q. This process, known as *inverting the demand function*, allows it to be plotted on the same set of axes as the supply function (Figure 2.1). Solving for p in terms of q, we rearrange the equation:

$$\mathsf{q} = 5000 - 20\mathsf{p}.$$

$$20p = 5000 - q.$$

$$p = 250 - \frac{q}{20}.$$

This inverse demand function now expresses ticket price as a function of passenger quantity, making it directly comparable to the supply function. By plotting both equations on the same graph, we can visually identify the equilibrium point where the airline's pricing strategy intersects with traveler demand. This equilibrium determines both the final ticket price and the number of passengers flying on the route.

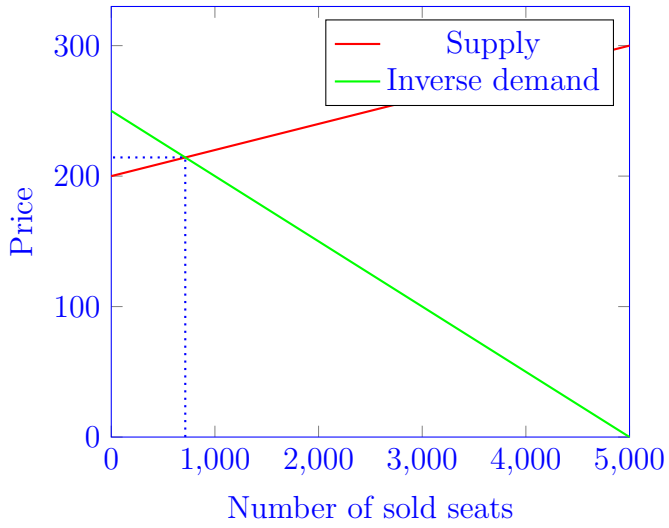


Figure 2.1: Supply and demand: airline example

To further illustrate the concept of equilibrium in transportation systems, we consider a case where there is no monetary price involved: traffic flow on a highway. Specifically, we analyze the relationship between traffic volume and travel time on Highway A1 between Morges and Rolle. Unlike pricing in airline tickets, where demand and supply interact through financial incentives, highway traffic operates under a different dynamic—one where congestion itself regulates demand.

In this example, travel time is influenced by the number of vehicles on the road. As more vehicles enter the highway, congestion builds up, leading to longer travel times. This relationship is captured by the following equation:

$$t = 15 + 0.02x,$$

where t represents the travel time in minutes, and x is the number of vehicles per hour. This equation indicates that in free-flow conditions, when there is no congestion, the travel time is $t_0 = 15$ minutes. However, as traffic increases, the additional congestion adds 0.02 minutes (or 1.2 seconds) per additional vehicle per hour.

At the same time, the number of vehicles using the highway depends on the perceived travel time. When travel time is low, more drivers choose to take the highway, as it offers a convenient and time-efficient route. However, as congestion builds up and delays increase, some drivers may opt for alternative routes, adjust their departure times, or even switch to other modes of transportation. This dynamic relationship between traffic volume and travel time is represented by the following equation:

$$x = 4000 - 120(t_0 + \Delta t) = 4000 - 120t = 2200 - 120\Delta t.$$

Here, t represents the total travel time, which can be decomposed into two components: the *free-flow travel time* t_0, which corresponds to the minimum possible travel time when there is no congestion, and the *congestion delay* Δt, which accounts for additional delays caused by traffic. In this example, we assume a free-flow travel time of $t_0 = 15$ minutes. As congestion increases, the total travel time t grows beyond this minimum value, with Δt capturing the excess time spent due to vehicle interactions, lane-changing maneuvers, and speed reductions.

Rewriting the equation in terms of free-flow and congested conditions, we see that when the highway operates at free-flow speed ($\Delta t = 0$), the maximum demand is 2200 vehicles per hour. This reflects the number of travelers who would choose the highway if it offered an optimal, uncongested travel experience. However, as congestion builds and Δt increases, the number of vehicles willing to use the highway decreases at a rate of 120 vehicles per additional minute of travel time. This means that for each extra minute spent in congestion, 120 drivers opt out of using the highway, choosing alternative routes or modes of transport instead.

To determine the equilibrium, we solve for the values of t and x that satisfy both equations simultaneously. Substituting the travel time equation into the traffic demand equation:

$$x = 4000 - 120(15 + 0.02x).$$

Expanding and rearranging,

$$x = 4000 - 1800 - 2.4x,$$

$$x + 2.4x = 2200,$$

$$3.4x = 2200,$$

$$x = 647 \text{ vehicles per hour.}$$

Substituting this value back into the travel time equation,

$$t = 15 + 0.02 \times 647,$$

$$t = 15 + 12.94 = 27.94 \text{ minutes.}$$

Thus, the equilibrium state of the highway is a traffic flow of 647 vehicles per hour, with each vehicle experiencing a travel time of approximately 27.94 minutes. This equilibrium represents the point where the congestion level naturally regulates itself: if travel times were lower, more drivers would enter the highway, increasing congestion and pushing the travel time back up; if travel times were higher, fewer drivers would choose the route, reducing congestion and bringing travel times down.

To facilitate a direct comparison between traffic supply and demand, we rewrite the demand function in its inverse form, expressing travel time as a function of traffic volume:

$$t = \frac{100}{3} - \frac{x}{120}.$$

By plotting both the travel time function and this inverse demand function on the same graph, we can visualize how equilibrium is reached at their intersection (see Figure 2.2). This example highlights the fundamental economic principles at play in transportation systems, where congestion serves as an implicit pricing mechanism, regulating demand in the absence of monetary costs.

One common strategy to improve highway performance is to expand its capacity by adding an additional lane. This modification increases the road's ability to accommodate more vehicles while reducing congestion and travel delays. To illustrate this, we analyze the impact of adding a new lane to Highway A1 between Morges and Rolle, a key corridor experiencing regular congestion.

Before the improvement, the travel time on the highway followed the equation:

$$t = 15 + 0.02x,$$

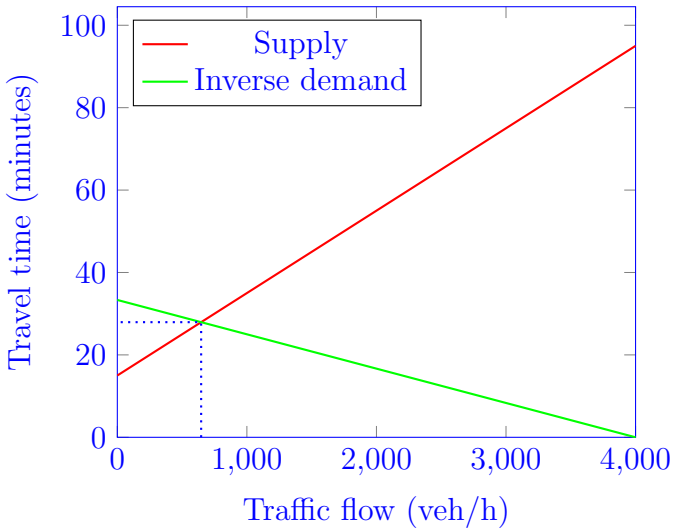


Figure 2.2: Supply and demand: highway example

where t represents the travel time in minutes and x is the number of vehicles per hour using the highway. In equilibrium, this relationship resulted in an average travel time of 27.94 minutes with a traffic volume of 647 vehicles per hour.

With the addition of a new lane, the supply of road space increases, effectively doubling the capacity available to vehicles. This improvement reduces the congestion effect per vehicle, modifying the travel time equation to:

$$t = 15 + 0.01x.$$

Here, the congestion impact per vehicle is reduced from 0.02 minutes per additional vehicle to 0.01 minutes, reflecting the added capacity. At first glance, one might assume that simply inserting the previous traffic volume ($x = 647$) into the new travel time equation would yield an improved travel time of:

$$t = 15 + 0.01 \times 647 = 21.5 \text{ minutes}.$$

However, this approach, illustrated in Figure 2.3, is incorrect because traffic volume is not constant—it reacts dynamically to changes in travel conditions. When travel time improves, more drivers are encouraged to use the highway, increasing demand. The number of vehicles on the road adjusts until a new equilibrium is reached.

To find this new equilibrium, we combine the updated travel time function

with the demand equation:

$$\mathfrak{x} = 4000 - 120\mathfrak{t}.$$

Substituting the new travel time equation into this demand function:

$$\mathfrak{x} = 4000 - 120(15 + 0.01\mathfrak{x}).$$

Expanding and solving for $\mathfrak{x}$,

$$\mathfrak{x} = 4000 - 1800 - 1.2\mathfrak{x},$$

$$\mathfrak{x} + 1.2\mathfrak{x} = 2200,$$

$$2.2\mathfrak{x} = 2200,$$

$$\mathfrak{x} = 1000 \text{ vehicles per hour.}$$

With this new traffic volume, the corresponding travel time is:

$$\mathfrak{t} = 15 + 0.01 \times 1000 = 25 \text{ minutes.}$$

Thus, after adding a lane, the highway experiences both an increase in traffic volume, rising from 647 to 1000 vehicles per hour, and an improvement in travel time, decreasing from 27.94 minutes to 25 minutes, as illustrated in Figure 2.4. While the travel time reduction is not as large as initially expected, the highway now serves significantly more travelers while still providing a slightly faster journey.

This example highlights a fundamental characteristic of transportation systems: *induced demand*. When road capacity is increased, travel becomes more attractive, leading to an increase in usage. This self-regulating effect means that infrastructure improvements do not always lead to proportional reductions in congestion but rather a redistribution of travel behavior. Understanding this dynamic is essential for making informed decisions about road expansion projects and evaluating their true long-term benefits.

As we have seen, the interaction between supply and demand plays an important role in transportation systems, shaping how infrastructure is utilized and how travelers make decisions. Engineers have the ability to modify the *supply function* by introducing capacity enhancements, optimizing traffic management, or implementing new technologies to improve system performance. However, any change in supply inevitably affects demand, as travelers adjust their behavior in response to improved or degraded travel conditions.

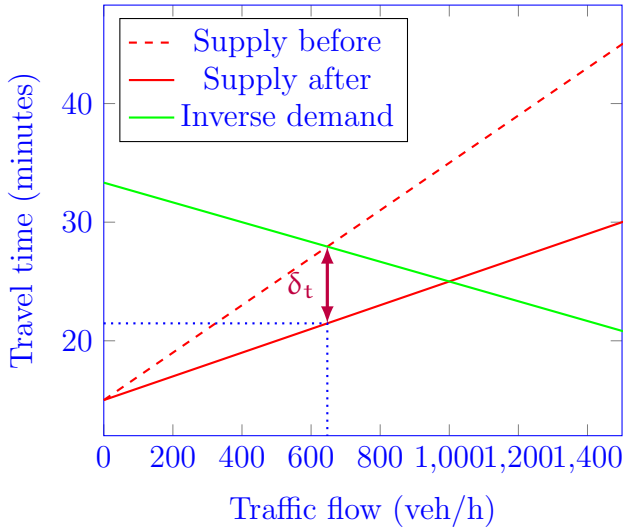


Figure 2.3: Travel time improvement after capacity increase: wrong analysis

Understanding this interdependence is essential for designing effective and sustainable transportation solutions.

One of the most important lessons in transportation analysis is that *demand is not fixed* — it responds dynamically to changes in supply. When road capacity is expanded, congestion effects may initially be reduced, but over time, increased attractiveness can lead to a rise in demand, partially offsetting the benefits of the capacity increase. Conversely, measures such as congestion pricing or improved public transport options can shift demand away from road networks, leading to a more balanced distribution of traffic. Engineers must therefore not only focus on modifying infrastructure but also anticipate how users will respond to these changes.

Note that, in real-world applications, supply and demand functions are *nonlinear*. Unlike the simplified linear models used for illustrative purposes, actual travel time and demand relationships exhibit more complex behavior. A more realistic supply function accounts for the fact that congestion effects grow exponentially as traffic approaches capacity, while a more sophisticated demand function reflects behavioral patterns where sensitivity to travel time varies across users.

For instance, a more realistic supply function could be expressed as:

$$t = 10\left(1 + 0.15\left(\frac{x}{2000}\right)^4\right),$$

indicating that as traffic volume increases, travel time rises sharply due

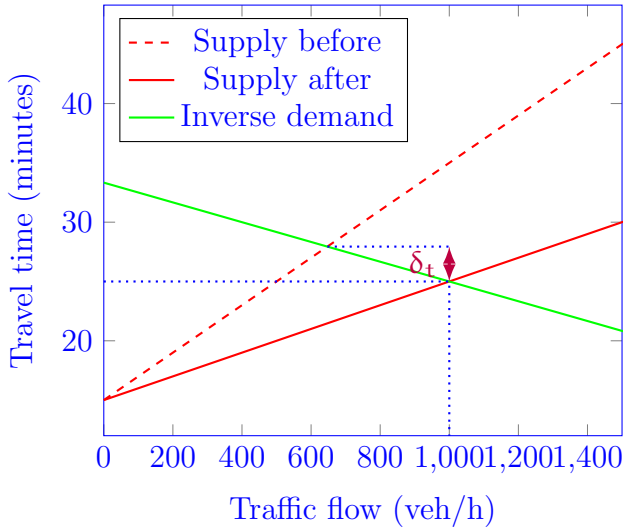


Figure 2.4: Travel time improvement after capacity increase: correct analysis

to congestion effects. Meanwhile, a more refined demand function might take the form:

$$x = 8000 \frac{1}{1 + e^{0.06t}},$$

which represents a demand curve where travel time reductions attract more users but at a diminishing rate.

As for the linear examples, the graphical representation of these functions (Figure 2.5) highlights the *equilibrium point*, where supply and demand intersect, determining both the actual travel time and traffic volume. Engineers and policymakers must carefully evaluate these nonlinear dynamics when proposing interventions, ensuring that transportation improvements lead to long-term benefits rather than unintended consequences such as induced demand.

2.2 Elasticities

The concept of *elasticity* plays a fundamental role in understanding how travel demand responds to changes in external conditions, such as travel time or cost. Elasticities provide a measure of the sensitivity of demand to these changes, helping transportation planners anticipate traveler behavior and assess the potential impact of policy interventions.

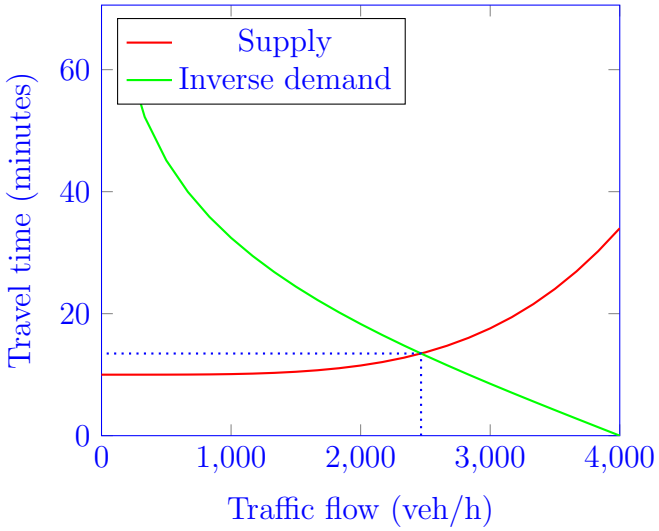


Figure 2.5: Supply and demand: nonlinear functions

To illustrate this concept, consider the demand function for a highway:

$$x = 4000 - 120t.$$

This equation indicates that as travel time t increases, the number of vehicles x using the highway decreases, following a linear relationship. To assess how sensitive demand is to travel time, we analyze what happens when t is increased by 1%.

For an initial travel time of $t = 27.94$ minutes and traffic volume $x = 647.2$ vehicles per hour, a 1% increase in t results in a new travel time of $t = 28.2194$ minutes. The corresponding traffic volume drops to $x = 613.672$ vehicles per hour, representing a relative change of approximately -5.18%. Similarly, for an initial travel time of $t = 25$ minutes with traffic volume $x = 1000$ vehicles per hour, a 1% increase in t leads to a decrease in x to 970, corresponding to a relative change of -3%.

These observations suggest that demand sensitivity varies depending on the initial conditions. To formalize this, we introduce the concept of *point elasticity*, which quantifies the percentage change in demand relative to the percentage change in travel time:

$$e_t = \frac{dx/x}{dt/t} = \frac{dx}{dt}\frac{t}{x}.$$

Applying this formula to our example, we differentiate the demand function:

$$\frac{dx}{dt} = -120.$$

Substituting into the elasticity formula,

$$e_t = -120\frac{t}{x} = -120\frac{4000 - x}{120x} = 1 - \frac{4000}{x}.$$

Evaluating this at $x = 1000$ gives:

$$e_t = 1 - \frac{4000}{1000} = -3,$$

that corresponds to the value calculated above. Note that this result indicates that demand is highly sensitive to changes in travel time in this scenario.

Elasticities are commonly classified into two categories based on their magnitude. If the absolute value of the elasticity is greater than 1, meaning that demand changes proportionally more than the change in travel time, the demand is considered *elastic*. As the elasticity in negative, this happens when

$$e_t < -1.$$

In contrast, if the absolute value is less than 1, meaning that demand is less responsive to changes in travel time, it is considered *inelastic*:

$$e_t > -1.$$

Understanding whether demand is elastic or inelastic is important for transportation policy and pricing strategies. For example, if demand is elastic, small increases in congestion or toll pricing can lead to significant reductions in traffic volume, which may be desirable for congestion management. Conversely, if demand is inelastic, even substantial increases in travel time may not significantly reduce the number of travelers, indicating that alternative policies such as infrastructure expansion or improved public transit may be needed to influence travel behavior effectively.

Figure 2.6 provides a graphical representation of the demand elasticity with respect to travel time in the highway example discussed earlier. This visualization helps illustrate how the sensitivity of travelers to congestion varies depending on traffic conditions. The horizontal axis represents traffic flow, measured in vehicles per hour, while the left vertical axis corresponds to travel time in minutes. Additionally, the right vertical axis indicates the demand elasticity.

One key insight from this figure is that the demand elasticity is not constant across different levels of traffic, although the demand function is linear.

When traffic flow is below 2000 vehicles per hour, demand is *elastic*, meaning that small increases in travel time lead to significant reductions in the number of vehicles using the highway.

However, as traffic flow exceeds 2000 vehicles per hour, demand becomes *inelastic*, indicating that further increases in travel time have a relatively smaller effect on the number of vehicles on the road.

Note that the minimum possible travel time is 15 minutes, which represents the free-flow condition where there is no congestion, and the maximum flow is 2200 vehicles per hour. Therefore, the part of the graph beyond 2200 does not represent any real situation, and we observe that *demand is almost always elastic* in this example, meaning that a small percentage increase in travel time leads to a proportionally larger percentage decrease in the number of vehicles using the road. This implies that travelers are highly responsive to congestion: when travel times worsen, many users opt out of using the highway. Such elasticity suggests that policies aimed at reducing congestion—such as road pricing, high-occupancy vehicle (HOV) lanes, or improved public transportation—can have a significant impact on travel behavior.

The high elasticity of demand in this scenario also means that even minor improvements in travel time can lead to substantial increases in road usage. For instance, infrastructure projects that reduce travel times by just a few minutes can significantly boost demand, potentially leading to induced traffic effects. This highlights the need for careful planning when implementing road expansions, as increased capacity may initially reduce congestion but could later attract additional travelers, partially offsetting the benefits.

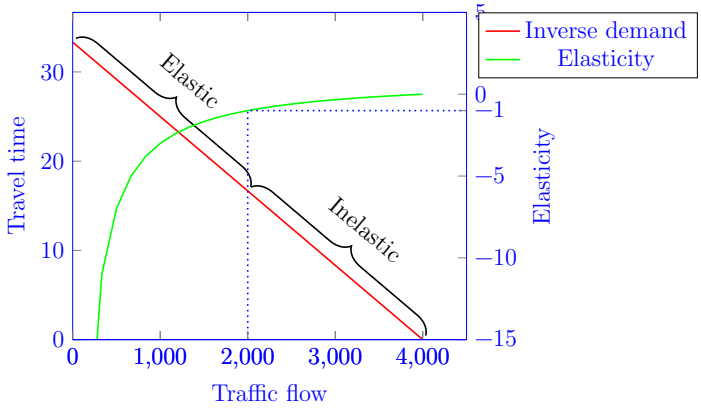


Figure 2.6: Point elasticity

While point elasticity evaluates the sensitivity at a specific travel time,

arc elasticity offers a more general measure by comparing demand changes over a finite interval. The arc elasticity of demand with respect to travel time is defined as:

$$e_{\Delta t} = \frac{\Delta x/x}{\Delta t/t} = \frac{\Delta x}{\Delta t}\frac{t}{x}.$$

This formula expresses the relative change in traffic flow (x) over a given change in travel time (t), providing an average elasticity value over a specific range rather than at an infinitesimal point.

In practical applications, the quantities Δx and Δt are typically computed from two distinct scenarios: a *before* and an *after* situation, such as the implementation of a new policy or infrastructure change. For instance, x_{before} and t_{before} represent the observed demand and travel time prior to the change, while x_{after} and t_{after} are the corresponding values following the change. The differences $\Delta x = x_{\text{after}} - x_{\text{before}}$ and $\Delta t = t_{\text{after}} - t_{\text{before}}$ capture the overall impact. This approach may be more meaningful for engineers and decision-makers, as it reflects tangible, scenario-based changes rather than theoretical sensitivity at a single point.

A key distinction arises when comparing linear and nonlinear demand functions. Indeed, for a *linear* demand function, the arc elasticity and point elasticity are equal:

$$e_t = e_{\Delta t}.$$

However, for a *nonlinear* demand function, elasticity varies across different travel times. In such cases, arc elasticity serves as an approximation of elasticity over a range, while point elasticity corresponds to the instantaneous sensitivity of demand at a particular travel time. Mathematically, point elasticity is obtained by taking the limit of arc elasticity as the interval shrinks to zero:

$$e_t = \lim_{\Delta_t \to 0} e_{\Delta t}.$$

Figure 2.7 illustrates the difference between arc elasticity and point elasticity for a nonlinear demand function. The horizontal axis represents travel time, while the vertical axis represents traffic flow. In this graph, *point elasticity* is associated with the slope of the tangent line at a specific travel time, capturing the local responsiveness of demand to small changes in travel time. In contrast, *arc elasticity* corresponds to the slope of the secant line between two points on the demand curve, representing an average elasticity over a broader interval.

The distinction between arc and point elasticity is particularly relevant for transportation policy and planning. When demand functions are nonlinear, using a single point elasticity may not accurately capture variations in traveler sensitivity across different traffic conditions. Arc elasticity provides a more practical approach when evaluating the impact of large-scale changes, such as infrastructure expansions or pricing adjustments, allowing planners to estimate the overall effect of a policy rather than just its local impact.

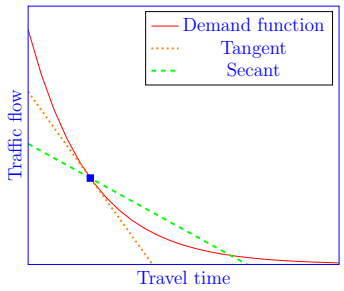


Figure 2.7: Arc elasticity for nonlinear demand functions

2.3 Consumer surplus

The concept of *consumer surplus* is a fundamental principle in economics and transportation analysis, providing a measure of the benefits that consumers receive from a service beyond what they actually pay for it. In the context of transportation systems, consumer surplus helps quantify the value that travelers derive from access to infrastructure and mobility services.

Consumer surplus is formally defined as the *difference between what consumers are willing to pay for a service and the price they actually pay.* In transportation, this means that if a traveler is willing to pay a high price for a trip but only needs to pay a lower fare, the difference represents a surplus or benefit to the traveler. This concept is particularly useful in assessing the economic impact of transportation policies, pricing strategies, and infrastructure investments.

Figure 2.8 provides a graphical illustration of *consumer surplus* for the airline example discussed in Section 2.1. The horizontal axis represents the number of sold seats, while the vertical axis corresponds to the price of a ticket. The figure includes the supply function and the inverse demand function.

Consumer surplus is the area between the demand curve and the equilibrium price. It represents the total monetary benefit received by passengers

who were willing to pay more than the actual ticket price. For example, some travelers might have been willing to pay 250 CHF for a ticket, but because of the market equilibrium, they only pay 214.3 CHF, thus gaining a surplus of 35.7 CHF per ticket. The sum of such differences across all passengers forms the total consumer surplus, shown as the triangular region above the equilibrium price and below the demand curve.

The formula for consumer surplus is given by the area of a triangle. In this context, the height of the triangle corresponds to the difference between the maximum willingness to pay (the highest price on the demand curve) and the equilibrium price. From the demand function, we see that the highest willingness to pay is 250 CHF, while the equilibrium price is 214.3 CHF. Thus, the height is:

$$250 - 214.3 = 35.7 \text{ CHF}.$$

The base of the triangle corresponds to the number of tickets sold at equilibrium, which is 714 seats. Therefore,

$$\text{Consumer Surplus} = \frac{714 \times 35.7}{2} = 12744.9 \text{ CHF}.$$

Thus, the total consumer surplus for this airline pricing scenario is 12,744.9 CHF. This value represents the total economic benefit passengers receive due to the pricing strategy, as many travelers are paying less than their maximum willingness to pay.

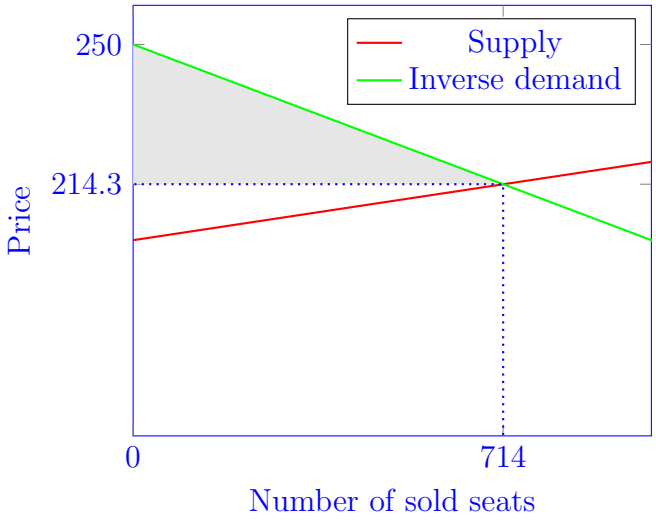


Figure 2.8: Consumer surplus: airline example

Consumer surplus can also be calculated for the highway example. Figures 2.9 and 2.10 illustrate the increase in consumer surplus following an improvement in highway capacity.

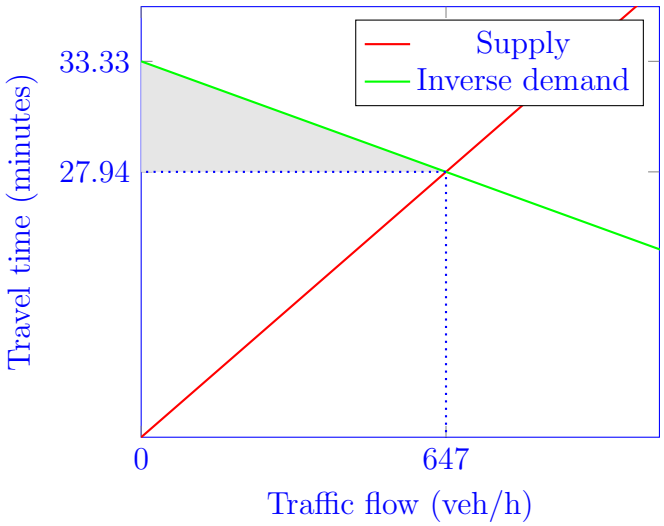


Figure 2.9: Consumer surplus: highway example

In the initial situation, before the capacity increase, the highway operated at an equilibrium where 647 vehicles per hour were using the road, with an average travel time of 27.94 minutes. The consumer surplus, represented by the gray area between the demand curve and the equilibrium travel time, is calculated as:

$$\text{Consumer surplus} = \frac{(33.33 - 27.94) \times 647}{2} = 1744.5 \text{ minutes.}$$

After adding an extra lane, the highway's capacity increased, effectively shifting the supply curve downward. The new equilibrium occurs at a higher traffic volume of 1000 vehicles per hour, with a reduced travel time of 25 minutes. With this improvement, the new consumer surplus increases to:

$$\text{Consumer surplus} = \frac{(33.33 - 25) \times 1000}{2} = 4165 \text{ minutes.}$$

This represents a significant gain in consumer surplus, indicating that travelers experience a net benefit due to reduced congestion and lower travel times.

The additional consumer surplus consists of two components:

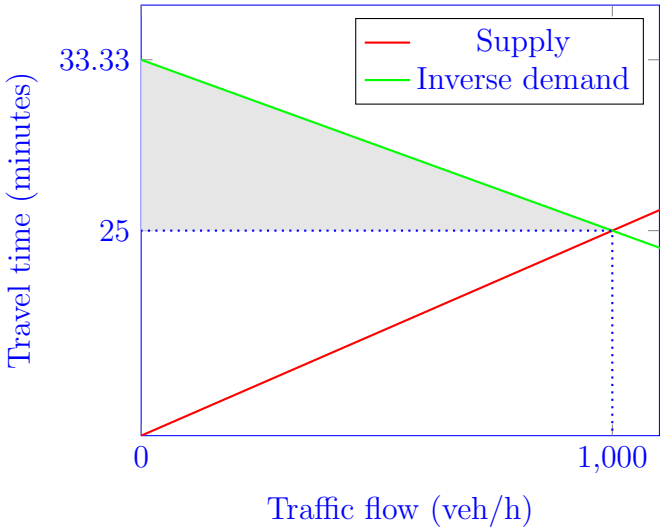


Figure 2.10: Consumer surplus: highway example with one more lane

1. The travelers who were already using the highway before the capacity improvement now experience shorter travel times, leading to an increase in their individual consumer surplus. This is represented by the rectangular gray area in Figure 2.11 between the old and new travel times for the original 647 users.

2. The reduced travel time attracts additional travelers who previously avoided the highway due to excessive delays. These new users contribute to an additional gain in consumer surplus, represented by the triangular region formed between the old and new equilibrium points.

When the supply and demand functions are linear, the calculation of the additional consumer surplus follows the *rule of half*, which states that the total increase in consumer surplus is given by:

$$\frac{1}{2}(x_1 + x_2)(t_1 - t_2),$$

where x_1 and x_2 represent the traffic volumes before and after the capacity expansion, and t_1 and t_2 denote the corresponding travel times (see Figure 2.12).

From a broader perspective, consumer surplus serves as a key indicator of *social welfare*, reflecting the overall benefits that a transportation system provides to society. A well-designed transportation network — whether it includes highways, public transit, or multimodal options — can generate

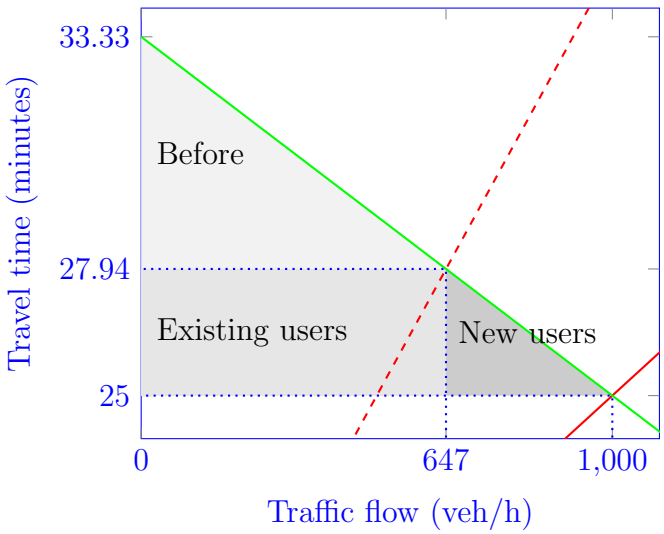


Figure 2.11: Additional consumer surplus for the highway example

significant consumer surplus by offering affordable and efficient mobility options. When a new facility, such as a highway or railway line, is introduced, the increased accessibility and reduced travel times often enhance consumer surplus, as travelers experience greater convenience without necessarily incurring higher costs.

Consumer surplus is not only influenced by changes in the supply function, such as increasing road capacity, but also by modifications in the demand function. As described above, demand reflects how travelers respond to travel conditions, and various policies or interventions can shift this relationship. For example, improving public transportation, promoting flexible work hours, or introducing pricing mechanisms can all alter the demand function by changing how travelers perceive and experience congestion.

Initially, the demand function for the highway follows:

$$x = 4000 - 120t.$$

This equation indicates that as travel time increases, fewer drivers choose to use the highway, leading to a decrease in traffic volume. However, suppose an intervention modifies the demand function to:

$$x = 4000 - 90t.$$

This new demand function suggests that travelers are now less sensitive to increases in travel time, meaning that a higher number of vehicles will

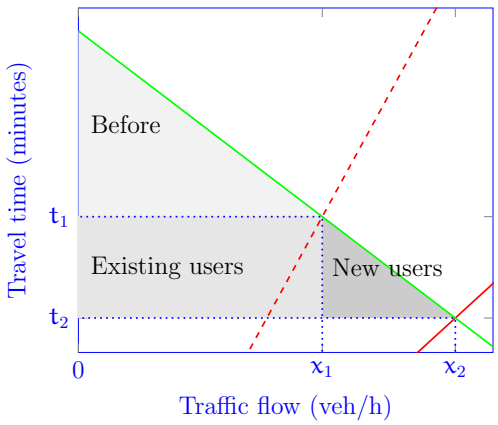


Figure 2.12: Additional consumer surplus: rule of half

continue using the highway even as congestion worsens. Such a shift might result from improved road reliability, better infrastructure design, or behavioral changes among commuters.

Figure 2.13 illustrates how this change affects the equilibrium. Before the demand shift, the highway operated at an equilibrium of 647 vehicles per hour, with an average travel time of 27.94 minutes. After modifying the demand function, the new equilibrium occurs at a higher traffic flow of 946 vehicles per hour, with a corresponding travel time determined by the intersection of the new demand function with the supply curve.

This demand shift leads to a significant increase in consumer surplus. Previously, the consumer surplus was:

$$\frac{(33.33 - 27.94) \times 647}{2} = 1744.5 \text{ minutes.}$$

After the demand modification, the consumer surplus rises to:

$$\frac{(33.33 - 25) \times 946}{2} = 4976.3 \text{ minutes.}$$

This increase in consumer surplus indicates that travelers are now experiencing greater overall benefits from the highway system, because their perception, their behavioral response, has changed.

Figure 2.14 highlights this increase in consumer surplus. The gray area before the demand shift represents the initial consumer surplus, while the expanded gray region after the shift accounts for the additional benefits gained. The larger consumer surplus suggests that policies or infrastructure improvements that alter travel behavior can have a profound impact on user benefits,

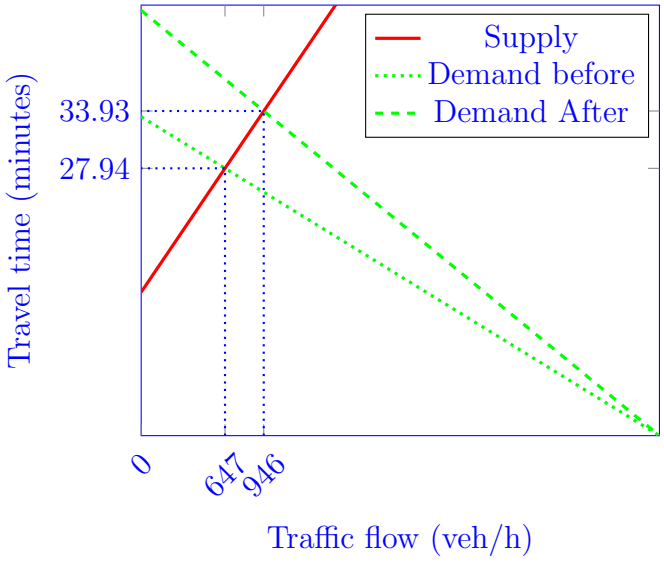


Figure 2.13: Modified demand function: new equilibrium

even without directly modifying road capacity.

In summary, engineers manage mobility by influencing both the supply and demand functions. Traditionally, engineering efforts have focused on modifying the *supply function*, which involves designing, building, and improving infrastructure, as well as introducing new transportation services. By increasing capacity, optimizing traffic flow, or enhancing public transit networks, engineers can directly impact travel conditions and system performance.

However, another equally important approach is to modify the *demand function*. Instead of increasing supply, demand-side strategies aim to influence traveler behavior and perceptions. This can be achieved through incentives, such as discounted transit fares or carpooling benefits, or penalties, such as congestion pricing or restricted access to certain areas during peak hours. By adjusting how travelers make decisions, these interventions can help reduce congestion, promote sustainable mobility choices, and improve overall efficiency without necessarily expanding infrastructure.

This leads to a fundamental question: *Where does the demand function come from?* Understanding travel demand requires insights into human decision-making, behavioral responses, and economic factors that influence mobility choices. To answer this question, we now turn to the behavioral foundations of travel demand, exploring how individuals make transportation decisions and how these choices can be modeled and predicted.

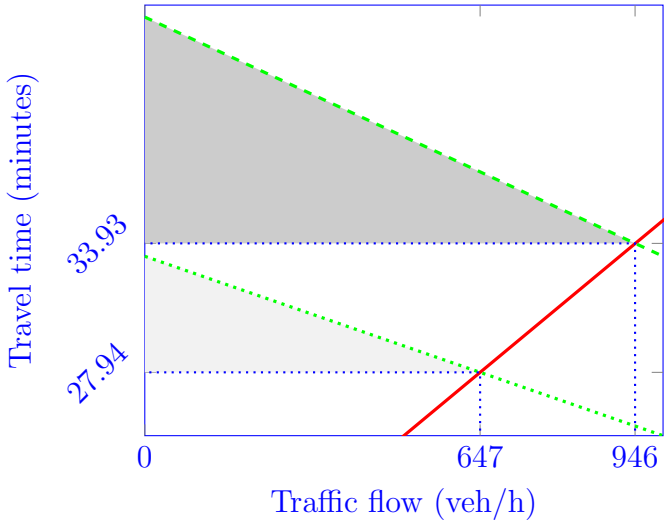


Figure 2.14: Modifed demand function: consumer surplus

2.4 Behavioral assumptions

In microeconomic theory, decision-makers are often faced with choices regarding the consumption of multiple goods or services. These choices are represented as a *consumption bundle*, where the individual selects quantities of different goods, taking into account their prices and a budget constraint.

In this framework, the decision-maker's objective is to determine the optimal quantity of each good to consume. The consumption bundle is represented as a vector:

$$q = \begin{pmatrix} q_1 \\ \vdots \\ q_K \end{pmatrix},$$

where each q_k represents the quantity of a specific good k. The prices of these goods are also given as a vector:

$$p = \begin{pmatrix} p_1 \\ \vdots \\ p_K \end{pmatrix}.$$

The key constraint in this decision process is the *budget constraint*, which ensures that the total expenditure does not exceed the available budget b. Mathematically, this is expressed as:

$$p^T q = \sum_{k=1}^{K} p_k q_k = b.$$

This equation states that the total cost of purchasing the chosen quantities of goods, considering their respective prices, must be equal to the budget.

The budget constraint, illustrated in Figure 2.15, defines a hyperplane in the space of goods. The feasible set of consumption bundles lies within this constraint, meaning that the decision-maker must allocate their resources efficiently to maximize utility while remaining within their budget.

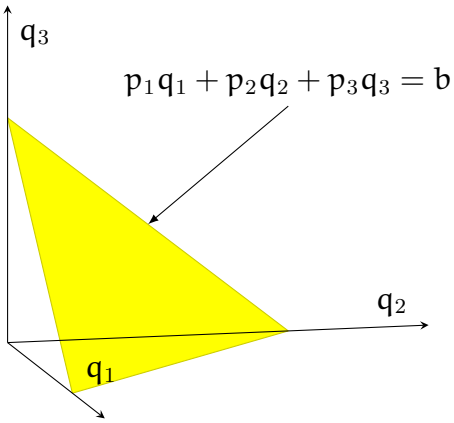


Figure 2.15: Budget constraint

A fundamental assumption in microeconomics is the concept of *Homo economicus*, which describes an idealized decision-maker who behaves in a rational and self-interested manner when making choices. This assumption provides a foundation for many economic models, including those used in transportation analysis.

The decision-maker is assumed to be *consistently rational*, meaning that they evaluate available options logically and systematically, making choices that maximize their personal benefit. This rationality implies that individuals process information without cognitive biases and always make decisions that align with their best interests.

Additionally, the decision-maker is considered to be *narrowly self-interested*. This means that their decisions are guided primarily by personal gain, rather than broader social, ethical, or altruistic considerations. In the context of transportation, this assumption suggests that travelers choose routes, modes, and departure times that minimize their own travel costs and time, without directly accounting for the impact on other users.

Finally, the decision-maker is assumed to *optimize her outcome*, meaning that given all available choices and constraints, she selects the option that provides the highest possible utility. In transportation, this translates to travelers selecting the fastest, cheapest, or most convenient mode of transport based on their preferences and constraints.

In decision theory and microeconomics, preferences describe how a decision-maker ranks different alternatives based on their desirability. The *preference-indifference operator* provides a formal way to compare these alternatives and establish a consistent ranking system.

The notation $q^k \succ q^\ell$ indicates that the decision-maker *strictly prefers* bundle q^k to bundle q^ℓ, meaning that given a choice between the two, the decision-maker would always select q^k. If the decision-maker perceives both q^k and q^ℓ as equally desirable, the relationship is expressed as $q^k \sim q^\ell$, signifying *indifference* between the two bundles. The relation $q^k \succsim q^\ell$ states that q^k is *at least as preferred* as q^ℓ, meaning that either q^k is strictly preferred or the decision-maker is indifferent between the two.

For preferences to be *rational*, they must satisfy key properties. The first is *completeness*, which ensures that for any two bundles q^k and q^ℓ, the decision-maker is able to compare them in some way, meaning that one of the three relations $q^k \succ q^\ell$, $q^k \prec q^\ell$, or $q^k \sim q^\ell$ must always hold. This guarantees that there are no situations where the decision-maker is unable to express a preference or indifference.

The second property is *transitivity*, which ensures logical consistency in rankings. If the decision-maker prefers q^k to q^ℓ and also prefers q^ℓ to q^m, then rationality requires that q^k must also be preferred to q^m. Formally, if $q^k \succsim q^\ell$ and $q^\ell \succsim q^m$, then it must follow that $q^k \succsim q^m$. This property prevents circular preferences, ensuring that choices remain logically structured.

The third property is *continuity*, which states that if q^k is preferred to q^ℓ, then any bundle q^c that is arbitrarily close to q^k must also be preferred to q^ℓ. This condition ensures that small variations in a bundle do not lead to abrupt or irrational shifts in preferences.

The concept of a *utility function* provides a formal mathematical representation of preferences, allowing for the quantification and comparison of different consumption bundles. Instead of expressing choices directly through preference relations, the utility function assigns a numerical value to each bundle, capturing the level of satisfaction or benefit derived from it.

The utility function is parameterized and written as:

$$\widetilde{u} = \widetilde{u}(q_1, \ldots, q_K; \theta) = \widetilde{u}(q; \theta),$$

where $q = (q_1, \ldots, q_K)$ represents the consumption bundle, and θ includes any parameters that may influence individual preferences. This function is

designed to be *consistent with the preference-indifference operator*, meaning that the ranking of consumption bundles based on utility values must align with the original preference relations.

Formally, if a decision-maker considers bundle $\mathbf{q}^k$ at least as preferred as bundle $\mathbf{q}^\ell$, this relationship must hold in terms of utility values:

$$\mathbf{q}^k \succsim \mathbf{q}^\ell \iff \widetilde{\mathfrak{u}}(\mathbf{q}^k; \theta) \geqslant \widetilde{\mathfrak{u}}(\mathbf{q}^\ell; \theta).$$

This equivalence ensures that the utility function correctly captures the decision-maker's underlying preferences. If a bundle has a higher utility value, it is preferred over other bundles with lower utility values.

Any preference indicator that is complete, transitive and continuous can be associated with a consistent utility function. Moreover, this function is *unique up to an order-preserving transformation*. This means that different mathematical transformations of the utility function, as long as they maintain the ranking of choices, result in equivalent representations of preferences. For example, applying a strictly increasing transformation such as the exponential function preserves the preference order:

$$\mathbf{q}^k \succsim \mathbf{q}^\ell \iff \widetilde{\mathfrak{u}}(\mathbf{q}^k; \theta) \geqslant \widetilde{\mathfrak{u}}(\mathbf{q}^\ell; \theta) \iff \exp \widetilde{\mathfrak{u}}(\mathbf{q}^k; \theta) \geqslant \exp \widetilde{\mathfrak{u}}(\mathbf{q}^\ell; \theta).$$

While utility values themselves may not have direct numerical meaning, their relative comparisons are what matter for decision-making. The utility function serves as a mathematical representation of the structure of preferences of ecomonic actors.

An example from microeconomics is the *Cobb-Douglas utility function*, which provides a simple yet powerful way to model consumer preferences. Illustrated in Figure 2.16, this function takes the form:

$$\widetilde{\mathfrak{u}}(\mathbf{q}) = \mathbf{q}_1^{\theta_1} \mathbf{q}_2^{\theta_2} \cdots \mathbf{q}_K^{\theta_K},$$

where $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_K)$ represents the quantities of different goods, and $\theta_1, \theta_2, \ldots, \theta_K$ are positive parameters that reflect the relative importance of each good in the consumer's preferences.

The Cobb-Douglas utility function has several key properties that make it useful for economic analysis. First, it exhibits *positive marginal utility*, meaning that consuming more of any good increases the overall utility, but at a decreasing rate. This property aligns with the intuitive idea that while additional consumption brings satisfaction, each extra unit of a good contributes less than the previous one.

Second, the Cobb-Douglas function exhibits *constant elasticity of substitution*, meaning that the way consumers trade off one good for another

remains stable across different consumption levels. The parameters θ_k determine the relative preference for each good. If one good has a higher θ_k, it means that the consumer derives more benefit from it compared to the others.

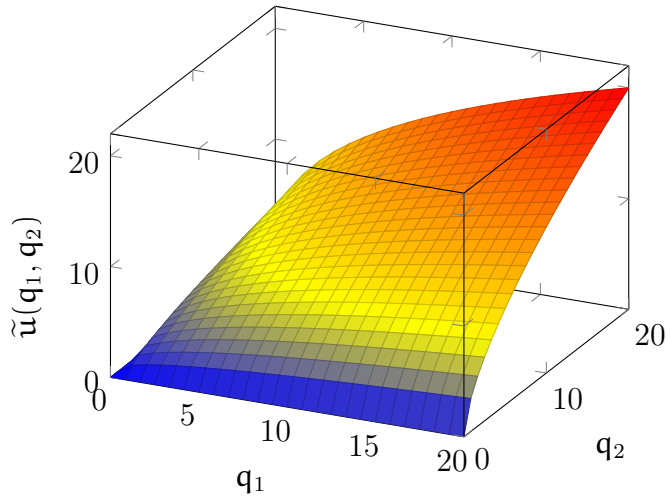


Figure 2.16: Example of a utility function: Cobb-Douglas

Figure 2.17 illustrates the *indifference curves*, which correspond to the level curves of the Cobb-Douglas utility function. These curves represent sets of consumption bundles that provide the same level of satisfaction to the decision-maker. If two bundles, denoted as A and B, lie on the same indifference curve, this means that they yield the same utility value. As a result, the decision-maker perceives no difference in preference between these two bundles and is indifferent when choosing between them.

Indifference curves capture the fundamental idea that individuals make trade-offs when consuming multiple goods. For example, if a person consumes less of one good, they may need more of another to maintain the same level of utility. This trade-off is reflected in the shape of the curves. In the case of the Cobb-Douglas utility function, indifference curves exhibit a smooth, convex shape, indicating that the decision-maker is willing to substitute one good for another, but at a diminishing rate.

The further an indifference curve is from the origin, the higher the level of utility it represents. This means that any bundle located on a higher indifference curve provides greater overall satisfaction compared to bundles on a lower curve. However, movement along the same curve does not change the utility level, as the decision-maker remains equally satisfied regardless of which bundle along the curve is chosen.

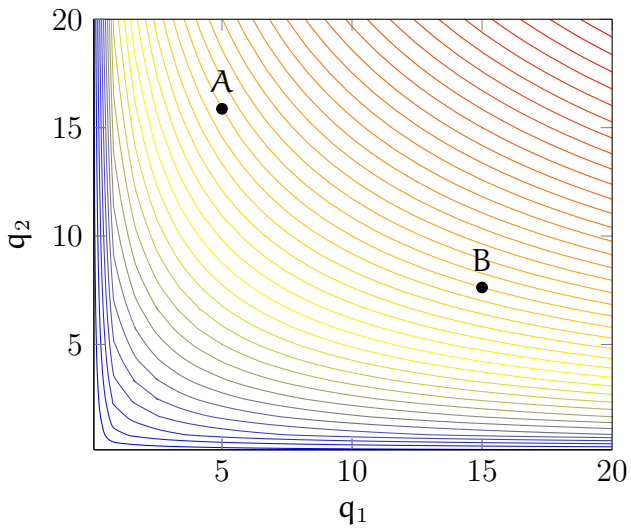


Figure 2.17: Cobb-Douglas utility function: indifference curve

Thanks to the representation of the preferences by a utility function, the behavior can be formally described as an *optimization problem*, where individuals seek to achieve the highest possible level of satisfaction or benefit given their available resources.

The decision-maker's objective is to maximize their *utility function*, which represents their preferences over different consumption bundles. The problem is expressed as:

$$\max_{q} \widetilde{u}(q; \theta),$$

where q is the vector of quantities of different goods or services consumed, and θ represents parameters influencing preferences, such as individual tastes or external conditions.

This optimization is subject to the budget constraint, which ensures that total expenditure does not exceed available resources:

$$p^T q = b.$$

Here, p is the vector of prices for each good, and b is the total budget available to the decision-maker. This equation ensures that the total cost of purchasing the chosen quantities remains within the financial means of the individual.

Additionally, there is a *non-negativity constraint*:

$$q \geqslant 0.$$

Figure 2.18 provides a visual representation of how a decision-maker selects an optimal consumption bundle by combining *indifference curves* with the *budget constraint*. This figure illustrates the fundamental principle of utility maximization, which states that individuals make choices to achieve the highest possible satisfaction while staying within their financial limitations.

The budget constraint is represented as a straight line, reflecting the total expenditure available for different combinations of goods. The slope of this line is determined by the relative prices of the goods, indicating how much of one good must be sacrificed to afford more of the other. The constraint ensures that the decision-maker does not exceed their available financial resources.

The optimal consumption bundle, denoted as q^*, corresponds to the point where the highest possible indifference curve is tangential to the budget constraint. At this point, the decision-maker achieves the maximum attainable utility given their financial limitations.

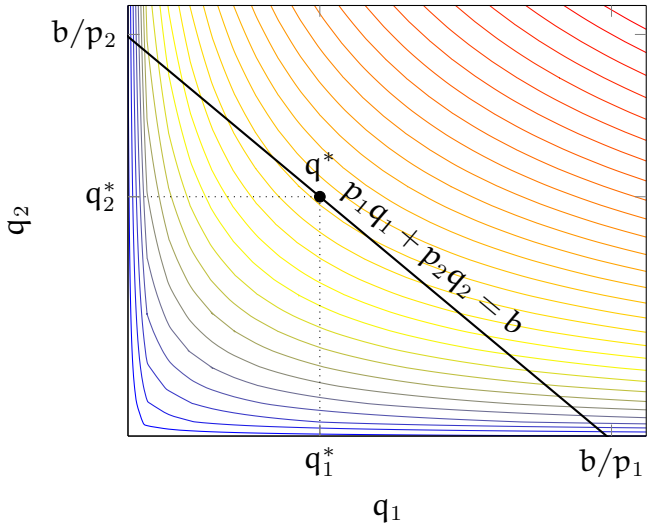


Figure 2.18: Cobb-Douglas utility function and budget constraint

To derive the demand functions from the utility maximization problem, we use the Karush-Kuhn-Tucker (KKT) optimality conditions. The optimization problem consists of maximizing a Cobb-Douglas utility function, which takes the form:

$$\max_{q_1,q_2} \widetilde{u}(q_1, q_2; \theta_0, \theta_1, \theta_2) = \theta_0 q_1^{\theta_1} q_2^{\theta_2}$$

subject to the budget constraint:

$$p_1 q_1 + p_2 q_2 = b.$$

Since the logarithm is a monotonic transformation, we can equivalently maximize the log-utility function:

$$\max_{q_1, q_2} \ln \widetilde{u}(q_1, q_2; \theta_0, \theta_1, \theta_2) = \theta_1 \ln q_1 + \theta_2 \ln q_2$$

under the same budget constraint. This transformation simplifies the optimization process while preserving the optimal solution.

To solve this constrained optimization problem, we construct the Lagrangian function:

$$L(q_1, q_2, \lambda) = \theta_1 \ln q_1 + \theta_2 \ln q_2 + \lambda(b - p_1 q_1 - p_2 q_2).$$

The necessary first-order conditions for optimality require that the gradient of the Lagrangian function be equal to zero:

$$\frac{\partial L}{\partial q_1} = \frac{\theta_1}{q_1} - \lambda p_1 = 0.$$

$$\frac{\partial L}{\partial q_2} = \frac{\theta_2}{q_2} - \lambda p_2 = 0.$$

$$\frac{\partial L}{\partial \lambda} = b - p_1 q_1 - p_2 q_2 = 0.$$

Rearranging the first two equations, we express λ in terms of the optimal quantities:

$$\theta_1 = \lambda p_1 q_1, \quad \theta_2 = \lambda p_2 q_2.$$

Summing these equations and using the budget constraint:

$$\lambda p_1 q_1 + \lambda p_2 q_2 = \theta_1 + \theta_2.$$

Since $p_1 q_1 + p_2 q_2 = b$, solving for λ gives:

$$\lambda = \frac{\theta_1 + \theta_2}{b}.$$

Substituting this back into the expression for q_1:

$$q_1 = \frac{\theta_1}{\lambda p_1} = \frac{b\theta_1}{p_1(\theta_1 + \theta_2)}.$$

Similarly, for q_2:

$$q_2 = \frac{b\theta_2}{p_2(\theta_1 + \theta_2)}.$$

These expressions represent the demand functions, which describe how the optimal quantities of each good depend on income b, prices p_1, p_2, and the parameters θ_1, θ_2, which capture the consumer's relative preferences for each good.

The final result shows that the fraction of the budget allocated to each good is proportional to the preference parameter θ_k. This means that as income increases, the demand for each good scales proportionally, maintaining the same expenditure shares. These demand functions are fundamental in consumer theory and provide key insights into how individuals allocate resources in response to price and income changes.

2.5 Summary

This chapter introduced the fundamental principles of *supply and demand* in transportation systems, exploring how they interact to determine equilibrium conditions. The supply function, represented as $p = f_s(q)$, characterizes how the system responds to different levels of demand, while the demand function, $q = f_d(p)$, reflects how users make choices based on prices or generalized costs. The equilibrium price p^* is the fixed point that satisfies $p^* = f_s(f_d(p^*))$, balancing supply and demand. Understanding these functions is important, as any modification to the system — whether through infrastructure improvements, policy changes, or behavioral interventions — affects both supply and demand dynamics.

The concept of *demand elasticity* was introduced to quantify how sensitive demand is to changes in price. Elasticity measures the percentage change in quantity demanded in response to a price variation, helping to evaluate the effectiveness of pricing policies, congestion charges, or transit fare adjustments. Both *point elasticities*, which measure sensitivity at a specific price, and *arc elasticities*, which capture the effect over a range of prices, provide valuable insights into user behavior.

Consumer surplus was presented as a key economic indicator, representing the difference between what consumers are willing to pay for a service and what they actually pay. This surplus serves as a measure of social welfare, as higher consumer surplus indicates that users receive greater value from the transportation system. Improvements in infrastructure, pricing strategies, and demand management policies can all influence consumer surplus,

making it a critical tool for evaluating transportation investments.

The chapter also covered *demand functions*, which are derived from behavioral assumptions using the principle of *utility maximization*. By modeling individual choices mathematically, demand functions can be obtained by solving optimization problems under budget constraints.

Finally, the discussion highlighted two primary ways to influence transportation systems: modifying the *supply function* and modifying the *demand function*. Supply-side changes involve infrastructure expansions or the introduction of new services to improve system performance. Demand-side interventions, on the other hand, focus on shaping user behavior through incentives, penalties, or informational campaigns. Both approaches are essential in transportation planning, and their effects must be carefully analyzed to ensure efficient and equitable mobility solutions.

Chapter 3

Discrete choice and value of time

In the previous chapter, we introduced the concept of *utility*, which allowed us to derive *demand functions* based on the principle of utility maximization. However, the framework we developed was primarily suited for *continuous choices*, where individuals select quantities of various goods while adhering to a budget constraint. This formulation, relying on the Karush-Kuhn-Tucker optimality conditions, applies naturally to decisions such as how much fuel to purchase, how many kilometers to drive, or how much money to allocate to different travel options.

3.1 Discrete choice

In transportation, however, many fundamental decisions are inherently *discrete*, not continuous. Travelers must choose between a *finite set* of mutually exclusive alternatives, such as selecting a mode of transport (private car, public transportation, cycling, or ride-hailing), deciding on a travel destination (shopping in the city center versus a suburban mall), determining an itinerary (using a highway or taking local roads), or even deciding whether to commute to the office or work from home. These choices do not involve gradual adjustments in quantities but rather involve selecting one option over others.

Clearly, the analysis based on the first-order optimality conditions of continuous optimization does not directly apply in this *discrete choice* setting. In particular, the classical demand functions derived in the previous chapter—where quantities vary smoothly in response to price and income—are not valid when individuals are faced with discrete alternatives. Instead, a

different analytical framework is required, one that models the *probability* of choosing each alternative based on the underlying utility associated with each option.

A comprehensive theory of discrete choice goes beyond the scope of this introduction course. However, in this chapter, we extend the concept of utility to the discrete choice context and introduce models that explain how individuals make such categorical decisions. A key concept that emerges from this analysis is the *value of time*, an essential measure in transportation economics. The value of time quantifies how individuals trade off travel time against monetary cost when making travel decisions. It plays a crucial role in evaluating transportation policies, infrastructure investments, pricing strategies, and congestion management measures. By understanding how travelers value their time, we can develop models that predict travel behavior and inform decision-making at both the individual and policy levels.

In the context of discrete choice, individuals select one option from a finite set of alternatives based on their perceived utility. This example illustrates how travelers decide between two transportation options: *public transportation (PT)* and *not using public transportation* (which could include driving, walking, or cycling). The decision is influenced by key attributes associated with each alternative, namely *travel time* and *travel cost*.

Each alternative is characterized by a specific travel time and cost, denoted as t_1, c_1 for public transportation and t_2, c_2 for the other option. Travelers assess these attributes when making their choice, trading off time and cost in a way that reflects their personal preferences.

To formally represent this decision process, we introduce utility functions, which assign a numerical value to each alternative based on its attributes. The utility function for each option is given by:

$$u_1 = -\theta_t t_1 - \theta_c c_1, \quad u_2 = -\theta_t t_2 - \theta_c c_2.$$

In this formulation, the parameters $\theta_t > 0$ and $\theta_c > 0$ capture the traveler's sensitivity to travel time and cost, respectively. The negative sign indicates that both time and cost reduce the perceived attractiveness of an option, meaning that travelers prefer shorter and cheaper trips.

Consistently with utility theory in the continuous case, the choice between alternatives depends on a comparison of these utility values. If $u_1 > u_2$, the traveler selects public transportation; otherwise, they choose the other mode. This framework provides a systematic way to model travel decisions and understand how different factors—such as fare changes, travel time improvements, or personal preferences—affect mode choice.

Clearly, this modeling approach extends naturally to situations with more

than two alternatives. Unlike traditional consumer choice models, where a budget constraint explicitly limits spending, no such constraint is imposed here. Instead, if an option is unaffordable for the decision-maker, it is effectively eliminated from consideration. In practice, this means that its utility function will not be evaluated, as individuals inherently disregard alternatives that are beyond their financial means.

This example highlights a fundamental principle in transportation modeling: individuals make choices by evaluating the trade-offs between time and cost. By estimating the parameters θ_t and θ_c, researchers can quantify the relative importance of travel time and cost in decision-making. This, in turn, allows policymakers to predict the impact of interventions such as fare reductions, service improvements, or congestion pricing on traveler behavior.

Figure 3.1 illustrates the trade-off between *travel time* and *travel cost* in the discrete choice framework. The horizontal axis represents the difference in travel time between the two alternatives, $t_1 - t_2$, while the vertical axis represents the difference in travel cost, $c_1 - c_2$. Each point in this space corresponds to a specific comparison between the two alternatives, based on their relative time and cost attributes.

The diagonal line in the figure represents the *indifference condition*, where both alternatives provide the same level of utility:

$$-\theta_t t_1 - \theta_c c_1 = -\theta_t t_2 - \theta_c c_2.$$

Along this line, the decision-maker is indifferent between the two options, meaning that the trade-off between time and cost is exactly balanced. The slope of this line is given by θ_t/θ_c, which represents the ratio of the sensitivity to travel time relative to the sensitivity to travel cost. This slope quantifies how much additional cost a traveler is willing to accept in exchange for saving one unit of travel time.

The figure is divided into four quadrants, each representing a different decision scenario. In the upper-right quadrant, where both $t_1 > t_2$ and $c_1 > c_2$, alternative 2 dominates, as it is both faster and cheaper than alternative 1. Conversely, in the lower-left quadrant, where $t_1 < t_2$ and $c_1 < c_2$, alternative 1 dominates, as it is superior in both dimensions. In these two quadrants, there is no trade-off; one alternative is strictly better than the other.

In the remaining two quadrants, a trade-off occurs. In the upper-left quadrant, alternative 1 has a shorter travel time but a higher cost, while in the lower-right quadrant, alternative 1 has a lower cost but a longer travel time. In these cases, the decision depends on the traveler's relative valuation of time versus money. If the point lies above the indifference line, alternative 2 is preferred, as the additional cost of alternative 1 outweighs its time

advantage. If the point lies below the line, alternative 1 is preferred, as the time savings justify the extra cost.

This figure provides an intuitive graphical representation of how discrete choices are made based on time and cost attributes. By estimating the parameters θ_t and θ_c, we can quantify the *value of time*, which is a key concept in transportation economics and policy analysis.

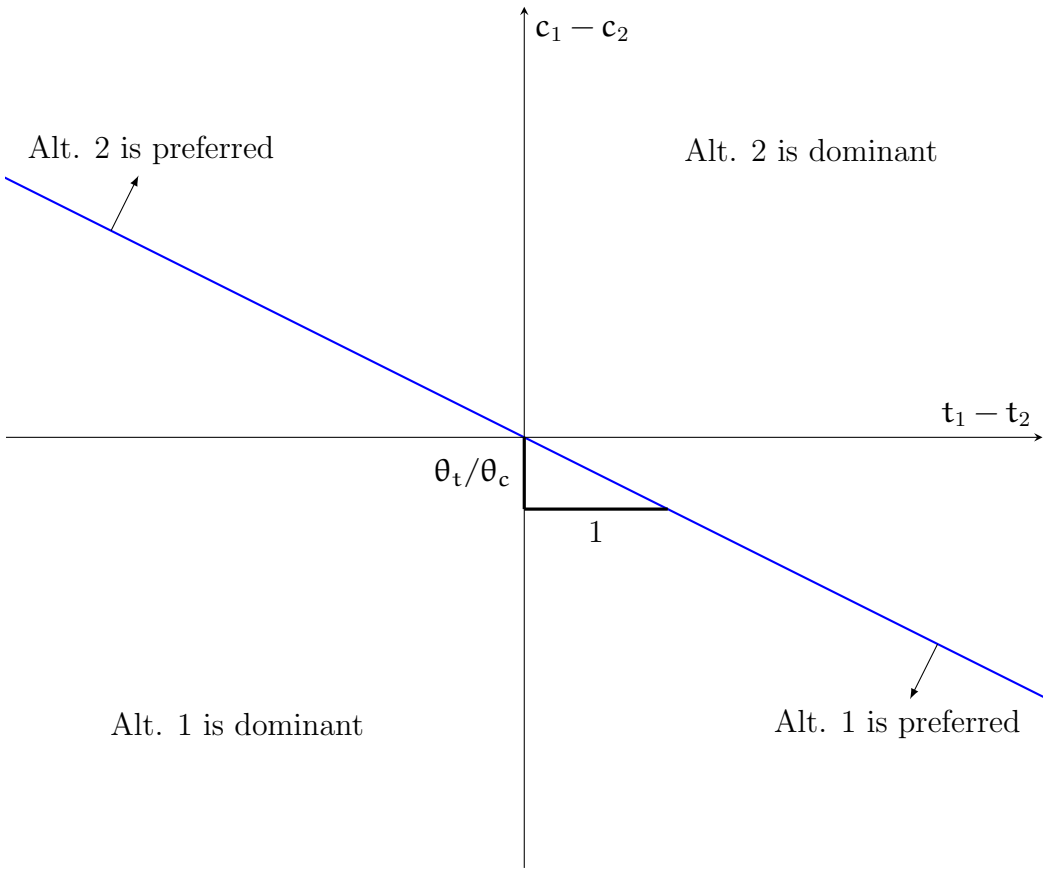


Figure 3.1: Trade-off between cost and time

While the simplest models consider only travel time and cost, additional attributes can be incorporated to better capture real-world decision-making. These additional attributes, denoted as x_{in}, may include variables such as waiting time, number of transfers, comfort level, or even weather conditions, depending on the context of the choice being modeled.

The general form of the utility function is given by:

$$u_{in} = -\theta_t t_{in} - \theta_c c_{in} - \theta_x x_{in},$$

where t_{in} represents travel time, c_{in} represents travel cost, and x_{in} captures another relevant attribute influencing the decision. The coefficients $\theta_t, \theta_c, \theta_x$ are parameters that measure the sensitivity of the decision-maker to each attribute. Since utility itself has no intrinsic unit, these coefficients serve to convert each attribute into the same unit of measurement, ensuring comparability across different factors.

Consequently, each coefficient has a unit that is the reciprocal of the corresponding attribute's unit. For instance, if travel time is measured in minutes, then θ_t has units of 1/minute, ensuring that $\theta_t t_{in}$ is dimensionless. Similarly, if cost is measured in Swiss Francs (CHF), then θ_c has units of 1/CHF. This allows the model to combine attributes with different units into a single numerical value of utility.

A key property of utility functions is that multiplying or dividing the entire function by a positive constant does not affect the ranking of alternatives. This means that utility can be rescaled without changing the relative preference between options. For example, dividing the utility function by θ_x results in:

$$u_{in}^x = \frac{u_{in}}{\theta_x} = -\frac{\theta_t}{\theta_x} t_{in} - \frac{\theta_c}{\theta_x} c_{in} - x_{in}.$$

Since utility comparisons remain unchanged under such transformations, this operation effectively expresses utility in terms of the units of x_{in}. In particular, dividing by θ_c allows utility to be expressed in monetary terms.

3.2 Value of time

A key concept that emerges from this framework is the *value of time*, which represents the price a traveler is willing to pay to reduce their travel time. This concept plays a central role in transportation economics, influencing policy decisions, infrastructure investments, and pricing strategies.

To understand the value of time, we start with the general utility function:

$$u_{in} = -\theta_t t_{in} - \theta_c c_{in} - \theta_x x_{in}.$$

We then express utility in terms of monetary units by dividing the function by θ_c, the coefficient associated with cost:

$$u_{in}^c = \frac{u_{in}}{\theta_c} = -\frac{\theta_t}{\theta_c} t_{in} - c_{in} - \frac{\theta_x}{\theta_c} x_{in}.$$

This transformation ensures that all terms in the equation are expressed in the same unit, namely *monetary value* (e.g., Swiss Francs, Euros, or Dollars). The coefficient θ_t/θ_c then has units of *currency per unit of time* (e.g.,

CHF/minute), representing the amount of money an individual is willing to pay to save one minute of travel time. This quantity is formally defined as the *value of time*:

$$\text{Value of Time} = \frac{\theta_t}{\theta_c}.$$

Geometrically, this corresponds to the slope of the indifference line in the cost-time trade-off diagram in Figure 3.1, indicating the rate at which a traveler is willing to exchange travel time for money.

The value of time reflects a fundamental economic trade-off: given the choice between a cheaper but slower option and a more expensive but faster option, how much is an individual willing to pay to reduce their travel time? The answer varies depending on personal income, trip purpose, urgency, and external conditions. For instance, business travelers or commuters on tight schedules may have a higher value of time than leisure travelers who are less time-sensitive.

More generally, the concept of *willingness to pay* extends beyond time to other attributes in the model. Any coefficient ratio of the form θ_x/θ_c represents the monetary equivalent of a given attribute, measuring how much a traveler is willing to pay for a specific improvement in travel conditions, such as increased comfort, reduced waiting time, or fewer transfers.

After introducing the value of time in a simple context, we now present a more general definition that applies to any utility-based discrete choice model. The value of time quantifies how much additional cost an individual is willing to pay to reduce their travel time while maintaining the same level of utility.

Consider a decision-maker choosing between alternatives, where c_{in} represents the cost of alternative i for individual n, and t_{in} represents the corresponding travel time. The utility associated with this alternative is given by a function $u(c_{in}, t_{in})$, which encapsulates the individual's preferences.

Now, suppose there is an improvement in travel conditions that reduces the travel time by δ^t_{in}, leading to a new travel time:

$$t'_{in} = t_{in} - \delta^t_{in}.$$

However, reducing travel time may come at an additional monetary cost. We denote this additional cost by δ^c_{in}, which is the amount that must be added to the original cost so that the overall utility remains unchanged:

$$u(c_{in} + \delta^c_{in}, t_{in} - \delta^t_{in}) = u(c_{in}, t_{in}).$$

The value of time is then defined as the *marginal rate of substitution* between cost and time:

$$\frac{\delta_{in}^{c}}{\delta_{in}^{t}}.$$

This quantity expresses the additional cost per unit of saved time, indicating the monetary value an individual assigns to reducing their travel time.

To formally derive the value of time, we apply *Taylor's theorem* to approximate the change in utility:

$$u(c_{in}, t_{in}) \approx u(c_{in}, t_{in}) + \delta_{in}^{c} \frac{\partial u}{\partial c_{in}}(c_{in}, t_{in}) - \delta_{in}^{t} \frac{\partial u}{\partial t_{in}}(c_{in}, t_{in}).$$

Since utility remains constant, setting the change in utility to zero gives:

$$\frac{\delta_{in}^{c}}{\delta_{in}^{t}} = \frac{\frac{\partial u}{\partial t_{in}}}{\frac{\partial u}{\partial c_{in}}}.$$

This ratio captures how changes in cost and travel time compensate for each other in the utility function. It provides an economic interpretation of how travelers perceive the trade-off between time and money.

For instance, in the commonly used linear utility function:

$$u_{in} = -\theta_t t_{in} - \theta_c c_{in} - \theta_x x_{in},$$

the value of time simplifies to:

$$\frac{\delta_{in}^{c}}{\delta_{in}^{t}} = \frac{\theta_t}{\theta_c}.$$

This result confirms our earlier finding: when the utility function is linear in time and cost, the value of time is the ratio of the sensitivity to time (θ_t) and the sensitivity to cost (θ_c).

The concept of *willingness to pay* extends beyond travel time and can be applied to any attribute that influences an individual's choice. In transportation, travelers may be willing to pay for various improvements in service quality, such as fewer transfers, reduced waiting times, better seat availability, or access to amenities like WiFi. Understanding willingness to pay for these attributes helps policymakers and transit operators design services that align with users' preferences.

For continuous attributes, such as travel time or waiting time, willingness to pay is derived using Taylor's theorem as described above. However, when

the attribute is discrete, such as whether WiFi is available on a bus, Taylor's theorem does not apply, since the attribute changes in a discontinuous manner.

To illustrate, consider a simple utility function that includes travel cost c_{in} and a binary variable w_{in}, which equals 1 if WiFi is available and 0 otherwise:

$$u_{in} = -\theta_c c_{in} + \theta_w w_{in}.$$

If a traveler compares two scenarios — one where WiFi is unavailable ($w_{in} = 0$) and one where it is available ($w_{in} = 1$) — the difference in utility can be expressed as follows:

Without WiFi, at current cost:

$$u_{in} = -\theta_c c_{in}.$$

With WiFi, but with an additional cost:

$$u_{in} = -\theta_c (c_{in} + \delta^c_{in}) + \theta_w.$$

Setting these two utility expressions equal (since willingness to pay ensures the traveler is indifferent between the two cases):

$$-\theta_c c_{in} = -\theta_c (c_{in} + \delta^c_{in}) + \theta_w.$$

Rearranging,

$$0 = -\theta_c \delta^c_{in} + \theta_w.$$

Solving for δ^c_{in}, we obtain:

$$\delta^c_{in} = \frac{\theta_w}{\theta_c}.$$

This result shows that, even in the case of a discrete variable, willingness to pay can still be expressed as a ratio of coefficients. In this example, θ_w/θ_c represents the monetary value that an individual assigns to having WiFi available on their trip. This same approach can be applied to other discrete attributes, such as the presence of air conditioning, priority boarding, or a direct route without transfers.

By quantifying willingness to pay for different service features, transit agencies can evaluate whether the benefits of investing in such improvements outweigh the costs. This approach is particularly useful in pricing strategies, where service enhancements can be offered at a premium price to users who value them most.

3.3 Summary

This chapter extended the concept of *utility theory* to *discrete choice models*, where individuals select one alternative from a set of mutually exclusive options. The fundamental assumption in this framework is that travelers choose the alternative that provides the highest utility, which is influenced by attributes such as travel time, cost, and service quality.

A key property of utility functions is that they have no inherent unit. However, by normalizing with respect to a specific variable, such as cost, we can express utility in meaningful units. This transformation allows utility to be interpreted in monetary terms, leading to the concept of *generalized cost*, which represents the total perceived burden of travel in a common unit, such as Swiss Francs.

The chapter also introduced the concept of *willingness to pay*, which quantifies how much a traveler is prepared to spend to improve a particular travel attribute. A primary application is the *value of time*, which measures the trade-off between time savings and cost. This principle extends to other factors, such as waiting time reductions, fewer transfers, or improved service features like WiFi availability.

By understanding how travelers evaluate trade-offs between different attributes, these models provide a foundation for predicting behavior and designing transportation policies. Applications range from pricing strategies and infrastructure planning to service improvements that align with traveler preferences. The insights gained from discrete choice models help in assessing the impact of policy decisions and optimizing transportation systems to better meet user needs.

Chapter 4

Mathematical modeling

Before exploring the operational models used for the planning and management of transportation systems, this chapter provides a refresher on fundamental mathematical concepts that serve as essential tools for modeling complex systems. The concepts introduced here are not limited to transportation applications; they are widely applicable across various domains where mathematical modeling plays a significant role.

Mathematical modeling is a structured approach to representing real-world systems through variables, equations, functions, and logical relationships. It allows us to describe interactions, analyze dependencies, and make informed predictions. In the context of transportation, these models support decision-making by capturing the dynamics of travel demand, network performance, and user behavior.

This chapter introduces key elements of mathematical modeling, beginning with the definition of *variables* and *random variables*, which form the building blocks of any model. A distinction is made between deterministic models, where outcomes are fully determined by input variables, and probabilistic models, which incorporate uncertainty. The concept of *causality* is also discussed, highlighting the importance of distinguishing correlation from causal relationships when developing models.

The process of *model development* is then presented, focusing on how model parameters can be estimated from collected data to ensure that the mathematical representation accurately reflects observed reality. We introduce the concept of *maximum likelihood estimation*, a fundamental method for determining parameter values that maximize the probability of reproducing the observed data. Initially applied in the context of contingency tables, where both independent and dependent variables are discrete, this approach is then extended to cases where both variables are continuous, leading to the introduction of *linear regression*. Finally, the chapter broadens these meth-

ods to encompass models with discrete dependent variables and independent variables of any type, including continuous. These models are particularly relevant in transportation modeling, where they play a key role in analyzing travel behavior and decision-making processes.

4.1 Mathematical models

A *mathematical model* is a structured representation of a system using mathematical concepts and notation. It provides a way to describe complex relationships, dependencies, and behaviors in a formalized manner, allowing for systematic analysis and interpretation.

Mathematical models serve several important roles. First, they help to *understand* a system by identifying key variables and their interactions. Through abstraction, models simplify real-world complexities while retaining essential features, making it easier to analyze how different components influence each other.

Second, models enable *prediction*. By incorporating observed data and logical relationships, they can forecast future states of a system under varying conditions. This predictive capability is particularly useful in decision-making, as it allows for the evaluation of potential outcomes before implementing changes in practice.

Finally, mathematical models support *optimization*. By formulating objectives and constraints, they provide a framework for finding the best possible solutions to problems, whether it be minimizing costs, maximizing efficiency, or balancing competing factors. Optimization techniques help improve system performance and guide policy decisions.

In mathematical modeling, a *variable* is a symbol used to represent a quantity that can take different values. Variables allow models to describe how different components of a system interact and change over time or under different conditions. They serve as the foundation for formulating relationships and expressing dependencies between elements within a system.

Variables play multiple roles in mathematical models, depending on the aspect of the system being represented. One important function is to *capture the state of the system*. For example, in transportation models, variables may represent traffic flow, indicating the number of vehicles passing through a specific point on a roadway over time.

Another key role of variables is to *capture the decisions made by engineers and planners*. Design choices, such as the number of lanes on a road or the frequency of public transportation services, can be represented as variables that influence system performance.

Variables also help to *measure system performance*. Metrics such as travel time, congestion levels, or energy consumption can be expressed mathematically to assess the efficiency and effectiveness of a transportation network.

Finally, models must account for *external factors* that influence the system but are not directly controlled by decision-makers. These include weather conditions, economic fluctuations, or unexpected events such as accidents, all of which can be represented using variables.

Variables can take different forms depending on the nature of the data they represent.

A *continuous variable* is one that can take any real value within a given range. These variables are often associated with a unit of measurement, making them useful for representing quantities that can be measured with precision. Examples include travel time, which can be expressed in minutes or seconds, and distance, which can be measured in kilometers or miles.

A *qualitative discrete variable* takes values from a predefined set of categories or labels, rather than a numerical range. These variables are used to represent attributes that do not have a natural numerical interpretation. For example, the mode of transportation a traveler chooses — such as driving, taking the bus, or cycling — can be represented as a qualitative discrete variable. Similarly, subjective measures like comfort level (e.g., very comfortable, comfortable, rather comfortable, not comfortable) can also be modeled using categorical variables. These variables often require special handling in mathematical models, as traditional arithmetic operations do not apply.

A *binary variable* is a special case of a discrete variable that can take only two possible values, typically represented as 0 or 1. These variables are frequently used to model decision-making situations, where an option is either selected or not. For example, in transportation planning, a binary variable could represent whether to open a new lane on a highway (1 for yes, 0 for no). Binary variables are fundamental in optimization problems, particularly in decision-making models that involve yes/no choices.

A *counting discrete variable* takes values from the set of natural numbers ($\mathbb{N}$), representing quantities that can only be whole numbers. Examples include the number of people in a household, the number of buses operating on a route, or the number of trips a person makes in a day. Although counting variables are technically discrete, they are often treated as continuous in mathematical models when the numbers involved are large enough that the distinction becomes negligible.

Each type of variable plays a specific role in modeling complex systems. Continuous variables capture measurable quantities, discrete variables categorize attributes, binary variables represent decisions, and counting variables quantify distinct elements.

A *random variable* is a function that assigns numerical values to outcomes of a random process. Formally, a random variable X maps an element from a sample space Ω, which represents all possible events, to the real numbers:

$$X : \Omega \to \mathbb{R}.$$

Random variables are used to quantify uncertainty in models, allowing for the representation of unpredictable elements in real-world systems. For example, in a transportation study, where individuals are randomly selected for a survey, an event $\omega \in \Omega$ can the fact that a specific individual is selected in the sample, and the corresponding value is her income or the number of cars in her household.

When we write $X = x$, we are actually describing the set of outcomes in Ω that lead to the value x:

$$X = x \iff \{\omega : \omega \in \Omega \text{ and } X(\omega) = x\}.$$

This notation means that $X = x$ defines a subset of Ω, representing the collection of outcomes where the random variable takes on a specific value.

Similarly, when we write $X \leqslant x$, we are describing another event: the set of all outcomes ω that lead to values of X that are less than or equal to x:

$$X \leqslant x \iff \{\omega : \omega \in \Omega \text{ and } X(\omega) \leqslant x\}.$$

This means that the event $X \leqslant x$ includes all possible realizations of the random variable where its value does not exceed x.

By viewing expressions like $X = x$ and $X \leqslant x$ as sets of possible outcomes, we connect random variables to probability theory. The probability of an event, such as $\Pr(X = x)$ or $\Pr(X \leqslant x)$, is then computed by summing or integrating over these subsets of Ω, depending on whether X is discrete or continuous.

A key aspect of random variables is the set of values they can take, known as their *range*. A random variable is *discrete* if it takes values from a finite or countably infinite set, such as the number of vehicles a person owns. In contrast, it is *continuous* if it can take any value within an interval, such as the travel time for a trip.

The behavior of a random variable is described by its *cumulative distribution function* (CDF), which gives the probability that the variable takes a value less than or equal to a given number:

$$F_X(x) = \Pr(X \leqslant x).$$

This function has some mathematical properties that help describe the behavior of probability distributions. First, the function is *monotonic*, meaning that as x increases, $F_X(x)$ does not decrease. More formally:

$$x < y \implies F_X(x) \leqslant F_X(y).$$

This property makes intuitive sense because if we increase the threshold x, the probability of X being less than or equal to x can either stay the same or grow larger, but it can never decrease. In other words, as we expand the range of possible values, the probability accumulates.

Additionally, the cumulative distribution function has well-defined limits over any bounded or unbounded interval $[a, b]$, which defines the range of possible values for the random variable X.

If X takes values in the interval $[a, b]$, then the cumulative distribution function satisfies:

$$\lim_{x \to a} F_X(x) = 0, \quad \lim_{x \to b} F_X(x) = 1.$$

For an unbounded support, such as the real line $(-\infty, +\infty)$, the cumulative distribution function satisfies:

$$\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to +\infty} F_X(x) = 1.$$

Thus, the CDF always increases from 0 to 1 over the support of the distribution, regardless of whether the interval is finite or infinite.

For discrete variables, probabilities are assigned to specific values using the *probability mass function* (PMF), defined as:

$$p_X(x) = \Pr(X = x),$$

where the sum of all probabilities must equal one. For continuous variables, probabilities are described by the *probability density function* (PDF), which is the derivative of the cumulative distribution function:

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

Unlike discrete variables, for which each value may have a nonzero probability, a strictly continuous variable has $\Pr(X = x) = 0$ for any specific value x; instead, probabilities are assigned to intervals:

$$\Pr(x < X \leqslant x + dx) = F_X(x + dx) - F_X(x) = \int_x^{x+dx} f_X(t)dt.$$

This means that $f_X(x)$ itself does not represent a probability, but rather a *density*, describing the relative likelihood of different values of X.

Despite this, in many modeling applications, it is useful to think of the pdf as a "probability" in an informal sense. It may simplify the analysis and enhance the intuition behind probabilistic modeling.

Two fundamental properties of random variables are *expectation* and *variance*. The expectation, or mean, represents the average value of the variable:

$$\mathrm{E}[X] = \sum_{x \in \mathcal{A}} x \, p_X(x) \quad (X \text{ discrete}),$$

$$\mathrm{E}[X] = \int_{x \in \mathcal{A}} x \, f(x) dx \quad (X \text{ continuous}).$$

The variance measures the dispersion of the values around the mean:

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2.$$

When working with random variables in mathematical models, it is often convenient to use an abuse of notation and treat them as if they were regular numerical variables. This simplifies the formulation of equations and allows for a more intuitive representation of probabilistic relationships.

Strictly speaking, a random variable is a function that maps outcomes from a sample space Ω to real numbers, meaning that statements like $X \in \mathbb{R}$ are not entirely precise. However, in practice, we often write:

$$X \in \mathbb{R}$$

to indicate that the possible values taken by X belong to the real number set. Similarly, when dealing with multiple random variables, we may define a vector:

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \in \mathbb{R}^n.$$

This notation allows us to work with vectors of random variables as if they were elements of a standard n-dimensional space.

A common example of this notation appears in linear operations involving random variables. We often write expressions such as:

$$\alpha X + \beta Y$$

where α and β are deterministic scalars, and X and Y are random variables. While the true mathematical meaning involves transformations of probability distributions, this notation simplifies analysis and computation.

Despite this simplification, it is important to remember that working with random variables requires keeping track of their underlying probability distributions. For instance, performing transformations on random variables affects their probability density functions (pdf) or probability mass functions (pmf), which must be considered when analyzing results.

For a discrete random variable X, if we define a new variable Y as a linear transformation of X, given by $Y = \alpha X + \beta$, then the probability mass function of Y is directly related to the probability mass function of X. Specifically, the probability of Y taking a particular value y is equal to the probability that X takes the corresponding value $x = (y - \beta)/\alpha$. This relationship is expressed as:

$$p_Y(y) = p_X\left(\frac{y - \beta}{\alpha}\right).$$

This equation shows that while the transformation shifts and scales the values of X, the probability masses remain unchanged. The new pmf is simply a re-indexing of the original one.

For a continuous random variable X, a similar transformation applies. If we define $Y = \alpha X + \beta$, then the probability density function of Y, denoted as $f_Y(y)$, is given by:

$$f_Y(y) = \frac{1}{|\alpha|} f_X\left(\frac{y - \beta}{\alpha}\right).$$

Unlike the discrete case, the transformation does not just shift values but also modifies the density of the distribution. The presence of the term $\frac{1}{|\alpha|}$ ensures that the total probability remains normalized to one. This factor accounts for the fact that scaling X by α either stretches or compresses the distribution, affecting the density accordingly. The absolute value in the denominator is necessary because a negative α would reverse the order of values, but probability densities must remain positive.

The shift parameter β simply moves the entire distribution to the left or right without affecting its shape. The scaling parameter α changes the spread of the distribution. If $|\alpha| > 1$, the values are stretched apart, making the distribution more spread out and lowering the density. If $|\alpha| < 1$, the values are compressed, increasing the density.

A mathematical model is designed to describe relationships between different variables in a system. Its purpose is to explain or predict the behavior of one or more variables based on the values of others.

Formally, we consider a situation where a random variable Y depends on another variable X. Given a specific value $X = x$, we analyze the corresponding distribution of Y, denoted as:

$$Y|X = x.$$

The variable Y is referred to as the *dependent*, *endogenous*, or *explained* variable. This means that its value is determined by or influenced by the value of X. The goal of the model is to describe how Y behaves based on different values of X.

On the other hand, X is called the *independent*, *exogenous*, or *explanatory* variable. This implies that X is known or chosen, and it serves as an input to the model.

Consider the example where X represents the travel time on a stretch of highway, and Y represents the traffic flow on that highway. In this case, a model seeks to understand how travel time affects the number of vehicles passing through. For instance, as travel time increases due to congestion, one might expect a reduction in traffic flow, reflecting the impact of delays on overall road capacity.

Another example examines household characteristics. Suppose X is the number of persons in a household, and Y is the number of cars owned by the household. Here, a model captures the tendency for larger households to own more vehicles. However, constraints such as parking availability and financial considerations may influence the shape of this relationship.

A third example considers the effect of weather conditions on transportation choices. Let X represent weather conditions, quantified through temperature, precipitation, or a categorical rating, and let Y denote the number of bike trips recorded in a given area. The model would capture the fact that, on rainy or cold days, the number of bike trips may decrease, while in pleasant weather, cycling activity might increase.

4.2 Causality

In mathematical modeling, one of the key objectives is to capture *causal effects*, meaning that changes in one variable X lead to changes in another variable Y. A necessary condition for a causal relationship is that the two variables must be *correlated*, but correlation alone does not imply causation.

Mathematically, correlation measures the strength and direction of the linear relationship between two variables. The correlation coefficient between X and Y, denoted as ρ_{XY}, is defined in terms of their covariance:

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

The covariance between two random variables X and Y, denoted as $\mathrm{Cov}(X, Y)$, measures the degree to which they vary together and is defined as:

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])].$$

Here, σ_X and σ_Y represent the standard deviations of X and Y, respectively:

$$\sigma_X = \sqrt{\mathrm{Var}(X)}, \quad \sigma_Y = \sqrt{\mathrm{Var}(Y)}.$$

This coefficient quantifies how closely changes in X correspond to changes in Y. If $\rho_{XY} > 0$, there is a positive correlation, meaning that when X increases, Y tends to increase as well. If $\rho_{XY} < 0$, there is a negative correlation, indicating that when X increases, Y tends to decrease. If $\rho_{XY} = 0$, there is no linear relationship between the two variables.

While causality always implies correlation, correlation alone does not imply causality. Just because two variables X and Y are correlated does not mean that X causes Y, or Y causes X.

Figure 4.1 presents data from the Swiss Microcensus 2015, illustrating the relationship between household monthly income and daily distance traveled using different modes of transport. The figure shows that as household income increases, the total daily distance traveled also rises. This pattern is observed consistently across different transport modes, including slow modes, public transport, and private cars.

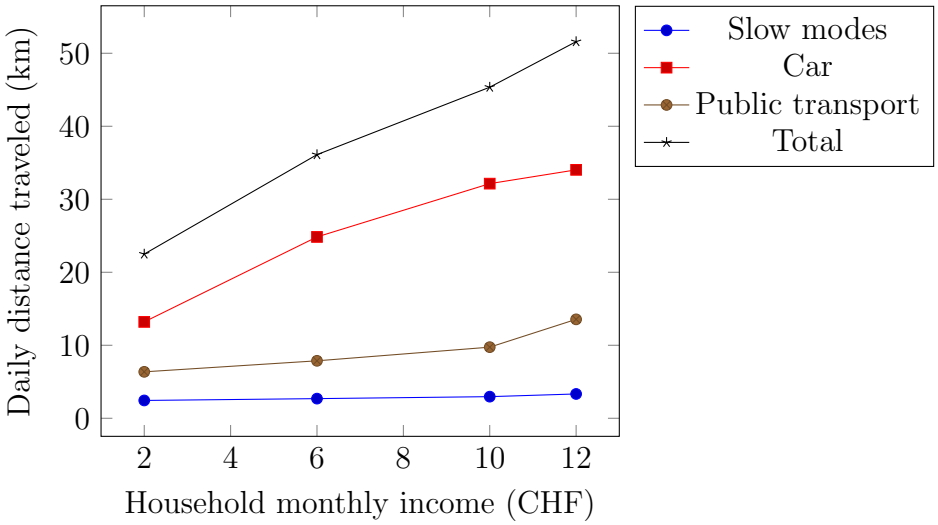


Figure 4.1: Swiss Microcensus 2015. Source: ARE

In this case, it is reasonable to assume that the correlation between income and travel distance reflects an underlying causal relationship. One explanation is that higher-income households have greater financial resources

to afford mobility, enabling them to own and use private vehicles more frequently. As a result, they can travel longer distances for work, leisure, or other activities. Additionally, individuals with higher incomes may have jobs that require commuting over greater distances, particularly in urban regions where housing prices push higher-income households toward suburban areas.

Although the relationship between income and travel distance is likely to be causal, other factors could also play a role. For instance, urban versus rural living conditions, workplace locations, and lifestyle preferences may influence both income and mobility patterns. Nonetheless, given the economic constraints associated with transportation and the clear financial implications of travel choices, the assumption of a causal link in this case is well-founded.

Figure 4.2 presents an intriguing relationship between chocolate consumption and the number of Nobel laureates per 10 million inhabitants across various countries (Messerli, 2012). The $\mathtt{x}$-axis represents the average chocolate consumption per capita in kilograms per year, while the $\mathtt{y}$-axis measures the number of Nobel laureates per 10 million people. Each country is represented by its flag, plotting its position based on these two variables.

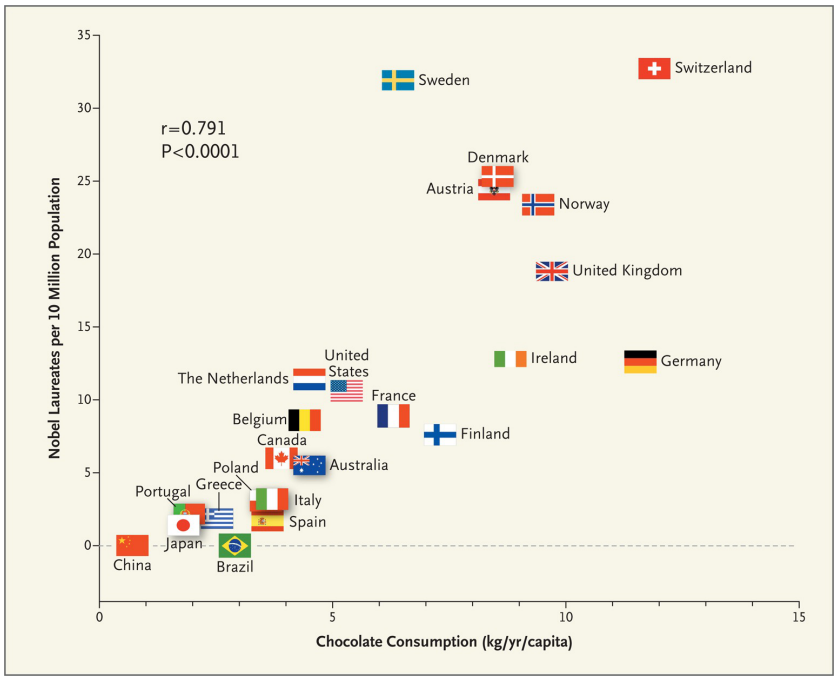


Figure 4.2: Nobels laureates and chocolate

The data reveal a strong correlation between these two variables, with

a correlation coefficient of $\mathtt{r} = 0.791$. The high value of $\mathtt{r}$ suggests that countries with higher chocolate consumption also tend to have more Nobel laureates per capita.

However, correlation does not imply causation. There is no logical reason to believe that eating more chocolate directly leads to an increased likelihood of winning a Nobel Prize, nor that having more Nobel laureates causes a country to consume more chocolate. The observed relationship may instead be explained by a third factor, such as the overall wealth of a country. Wealthier nations tend to have better-funded education and research institutions, which could contribute to a higher number of Nobel laureates. At the same time, wealthier populations may also have greater access to luxury goods such as chocolate.

Thus, instead of a direct causal link between chocolate consumption and Nobel laureates, it is plausible that both variables are influenced by national wealth. This scenario can be represented as:

$$\text{Nobel laureates} \mid \text{Wealth} \quad \text{and} \quad \text{Chocolate consumption} \mid \text{Wealth}.$$

The concept of spurious correlation is well illustrated by various amusing but misleading statistical relationships collected on the website

`www.tylervigen.com`.

These examples highlight cases where two variables appear strongly correlated, yet there is no logical causal connection between them. One such case is the observed correlation between *U.S. spending on science, space, and technology* and *suicides by hanging, strangulation, and suffocation*. Another example shows a strong correlation between the *divorce rate in Maine* and *per capita consumption of margarine*. Similarly, the number of *civil engineering doctorates awarded* appears to be correlated with *per capita consumption of mozzarella cheese*, but it is highly unlikely that cheese consumption influences academic achievements in engineering.

Understanding causality in mathematical modeling requires careful analysis of whether changes in one variable directly influence another. In some cases, causality can be bidirectional, while in others, it only makes sense in one direction.

Consider the relationship between travel time on a highway (X) and traffic flow (Y). There are two potential causal directions. First, if we consider $\mathsf{Y}|\mathsf{X}$, we are examining how traffic flow responds to a given travel time. This perspective aligns with demand functions, as travelers make behavioral choices based on perceived travel times. On the other hand, if we consider $\mathsf{X}|\mathsf{Y}$, we

are looking at how travel time is affected by traffic flow, which corresponds to the supply function and system performance. As congestion increases, travel times generally rise due to reduced speeds and bottlenecks. In this case, causality works in both directions: behavior affects the system, and the system in turn affects behavior.

A different example is the relationship between income (X) and distance traveled (Y). It is reasonable to assume that higher-income individuals travel longer distances, either because they can afford to commute farther for better housing, take more leisure trips, or own private vehicles that enable long-distance travel. Thus, $Y|X$ makes sense as a causal direction. However, the reverse relationship, $X|Y$, does not hold in the same way. Simply traveling a greater distance does not cause an individual's income to increase, making this direction of causality implausible.

Another scenario involves bus fares (X) and the number of riders (Y). Here, causality can again work in both directions. The number of riders is influenced by fare prices, meaning that $Y|X$ captures demand functions, as travelers decide whether to use public transport based on cost. However, transit operators may also adjust bus fares in response to demand levels, which corresponds to $X|Y$. If ridership declines, an operator may lower fares to attract more passengers, demonstrating how supply-side decisions influence pricing.

Finally, the relationship between weather (X) and the number of bike trips (Y) provides an example where causality is unidirectional. It is clear that weather conditions affect biking activity: rainy or cold weather discourages cycling, while warm and sunny weather increases ridership. Thus, $Y|X$ is a valid causal direction. However, the reverse relationship, $X|Y$, does not hold. The number of people riding bicycles does not influence the weather, making this direction nonsensical.

These examples illustrate the importance of distinguishing between correlation and causation. While two variables may be statistically related, it is essential to determine whether the relationship is truly causal and, if so, in which direction it operates. This distinction is fundamental for developing reliable predictive models and making informed policy decisions.

Causality is inherently context-dependent, meaning that the same variable can be considered exogenous in one setting and endogenous in another. A clear example of this is the distinction between supply and demand functions, where the roles of variables shift depending on the perspective taken. Theoretical assumptions play a fundamental role in defining causal relationships, as they provide the necessary structure for interpreting data and making meaningful predictions. Without a strong theoretical foundation, correlations may be misinterpreted, leading to incorrect conclusions about cause

and effect.

A well-constructed model is always grounded in theory. For example, in transportation analysis, utility theory provides a structured approach to understanding behavioral choices. Theoretical models are particularly essential when making predictions or extrapolating beyond observed data. Unlike purely data-driven approaches such as machine learning, which primarily identify patterns without necessarily understanding causal mechanisms, theory-based models assume that causal relationships remain stable over time and across different configurations of the system. This stability allows for more robust predictions and policy analysis, reinforcing the need for carefully considering causality in mathematical modeling.

4.3 Model development

Model development is an iterative process that involves several interconnected steps: specification, estimation, prediction, analysis, and decision-making. Each of these stages plays a fundamental role in ensuring that the model accurately represents reality and serves its intended purpose. This process is illustrated in Figure 4.3.

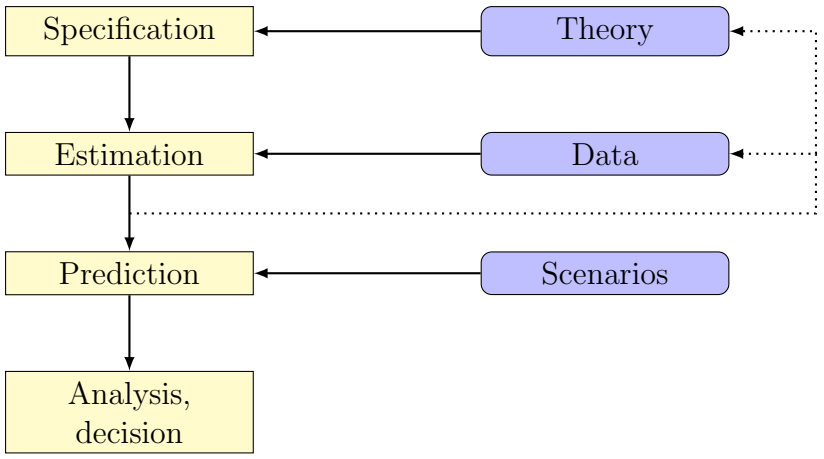


Figure 4.3: Model development

The first stage, *specification*, involves defining the mathematical structure of the model based on an underlying theory. This theory provides the logical foundation for identifying the key variables, their relationships, and the assumptions governing the model. The choice of specification determines how well the model can explain the system being studied.

Once the model is specified, the next step is *estimation*, which relies on empirical data to calibrate the model parameters. Using statistical techniques, the model parameters are adjusted to ensure the best possible fit to the observed data. If the model does not perform adequately, it may indicate the need to collect additional data or refine the theoretical assumptions.

Following estimation, the model can be used for *prediction*. In this stage, the model is applied to new input scenarios to forecast possible outcomes. These scenarios can reflect policy changes, infrastructure developments, or external factors such as economic conditions. The predictive power of the model is essential for assessing potential future states of the system.

After generating predictions, the results undergo *analysis*, leading to informed *decisions*. This phase involves evaluating the model's output to support decision-making processes. If the predictions suggest unexpected or unreliable outcomes, it may be necessary to revisit the specification or collect additional data to refine the model.

The model development process is rarely linear. As shown in Figure 4.3, the insights gained from estimation and prediction can highlight gaps in the existing data or weaknesses in the theoretical framework. In such cases, researchers may decide to gather more data or refine the theoretical model before proceeding with further predictions and decisions. This iterative refinement ensures that the model remains relevant and improves in accuracy over time.

4.3.1 The case of discrete variables

In transportation research, one possible hypothesis is that the choice of transportation mode depends on the purpose of the trip. This hypothesis suggests that travelers select their mode of transport based on the nature of their journey, rather than making random or habitual choices.

The reasoning behind this hypothesis stems from factors such as convenience, flexibility, and comfort. For example, individuals commuting to work may prioritize reliability and travel time, making public transport or private vehicles preferable. In contrast, shopping trips may favor transportation modes that allow for carrying goods easily, such as private cars or ride-hailing services. Similarly, recreational trips may involve preferences for active modes like walking or cycling, where enjoyment and health benefits outweigh time constraints.

To investigate the hypothesis that transportation mode choice depends on trip purpose, a data collection campaign is designed to gather empirical evidence. The study involves surveying a representative sample of travelers and recording their travel behavior. This approach allows researchers to quantify

the relationship between trip purpose and the use of public transportation.

The data collection campaign selects a random sample of 2000 individuals who qualify as travelers, meaning they made at least one trip the previous day. By ensuring that the sample is randomly chosen, the survey minimizes bias and enhances the reliability of the findings.

Each participant is asked a set of simple yet informative questions about one of their trips from the previous day. The first question identifies the purpose of the trip, distinguishing between work-related travel, leisure activities, and other purposes. The second question determines whether public transportation was used for that trip.

A *contingency table* is a type of table used in statistics to summarize the frequency distribution of categorical variables. It provides a structured way to analyze the relationship between two categorical variables by displaying their joint distribution in a matrix format. In the context of this study, the contingency table in Table 4.1 captures the responses from the data collection campaign and organizes them to highlight the connection between trip purpose and public transportation usage.

	Work	Leisure	Others
PT	172	191	150
Not PT	345	648	494

Table 4.1: Synthetic data generated from Microcensus 2015.

The table consists of two dimensions. The rows represent the transportation mode choice, distinguishing between travelers who used public transportation (PT) and those who did not (Not PT). The columns represent the purpose of the trip, categorized into work, leisure, and other activities. Each cell in the table indicates the number of travelers who fall into the corresponding category, effectively summarizing the entire dataset.

With the data collected and summarized in the contingency table, we can now specify a mathematical model that formalizes our hypothesis. The goal of the model is to describe and quantify the relationship between trip purpose and transportation mode choice.

In this case, the dependent variable, denoted as Y, represents the transportation mode chosen by the traveler. It is a qualitative variable that takes values from the set $\mathcal{A} = \{\text{public transport, others}\}$, meaning that a traveler either uses public transportation or another mode. The explanatory variable, denoted as X, represents the purpose of the trip, also a qualitative variable, with possible values in the set $\mathcal{A} = \{\text{work, leisure, others}\}$.

The model is specified as $Y|X$, indicating that the transportation mode choice Y is modeled as a function of the trip purpose X. In probabilistic terms, the objective is to determine the probability distribution of Y given X, expressed as:

$$\mathbb{P}(Y = \text{public transport} \mid X = x), \quad x \in \{\text{work, leisure, others}\}.$$

The mathematical model we have specified involves unknown parameters that need to be estimated from the data. These parameters represent the probabilities of choosing public transport for different trip purposes.

For trips made for work purposes, we define the parameter θ_1 as the probability that a traveler chooses public transport:

$$\theta_1 = \mathbb{P}(Y = \text{PT} \mid X = \text{work}).$$

Since the traveler must either use public transport or another mode, the probability of not using public transport is simply:

$$\mathbb{P}(Y = \text{not PT} \mid X = \text{work}) = 1 - \theta_1.$$

Similarly, for trips made for leisure, we define another parameter θ_2, which represents the probability of choosing public transport for leisure trips:

$$\theta_2 = \mathbb{P}(Y = \text{PT} \mid X = \text{leisure}).$$

Again, the probability of using another mode is:

$$\mathbb{P}(Y = \text{not PT} \mid X = \text{leisure}) = 1 - \theta_2.$$

For trips classified as "other," a third parameter θ_3 is introduced, representing the probability of choosing public transport:

$$\theta_3 = \mathbb{P}(Y = \text{PT} \mid X = \text{others}).$$

And the probability of using another mode is:

$$\mathbb{P}(Y = \text{not PT} \mid X = \text{others}) = 1 - \theta_3.$$

These parameters θ_1, θ_2, and θ_3 are unknown and must be estimated from the data collected in the survey. To perform this task, we introduce the concept of *likelihood function*.

The likelihood function quantifies the probability that a given model correctly predicts the observations in the dataset. It measures how well the model parameters align with the actual data.

To begin, consider a single observation. Suppose a traveler makes a trip for work and chooses public transport. According to our model, the probability of this event occurring is given by:

$$\mathbb{P}(\mathsf{Y} = \text{PT} \mid \mathsf{X} = \text{work}) = \theta_1.$$

If another traveler in the sample also makes the same choice, the probability of correctly predicting both observations would be the product of the probabilities, assuming independence.

Now, consider all travelers in the sample who make work-related trips and use public transport. There are 172 such travelers. The probability of correctly predicting all of them is:

$$\theta_1^{172}.$$

Similarly, the probability of correctly predicting all 345 travelers who travel for work but do not use public transport is:

$$(1 - \theta_1)^{345}.$$

Extending this to all categories of trip purposes in our contingency table, the overall probability that our model correctly predicts the entire dataset is given by:

$$\theta_1^{172}(1 - \theta_1)^{345}\theta_2^{191}(1 - \theta_2)^{648}\theta_3^{150}(1 - \theta_3)^{494}.$$

This expression defines the likelihood function, denoted as:

$$\mathcal{L}^*(\theta_1, \theta_2, \theta_3) = \theta_1^{172}(1 - \theta_1)^{345}\theta_2^{191}(1 - \theta_2)^{648}\theta_3^{150}(1 - \theta_3)^{494}.$$

Since likelihood values are often very small due to the multiplication of many probabilities, it is common to work with the logarithm of the likelihood function. This transformation, called the log-likelihood function, simplifies computations and converts the product into a sum:

$$\begin{aligned}\mathcal{L}(\theta_1, \theta_2, \theta_3) = {} & 172 \log \theta_1 + 345 \log(1 - \theta_1) + \\ & 191 \log \theta_2 + 648 \log(1 - \theta_2) + \\ & 150 \log \theta_3 + 494 \log(1 - \theta_3).\end{aligned}$$

This expression decomposes into three independent terms:

$$\mathcal{L}_1(\theta_1) + \mathcal{L}_2(\theta_2) + \mathcal{L}_3(\theta_3),$$

represented in Figure 4.4, where the $\mathfrak{x}$-axis represent the value of the unknown parameter, and the $\mathfrak{y}$-axis the corresponding log-likelihood.

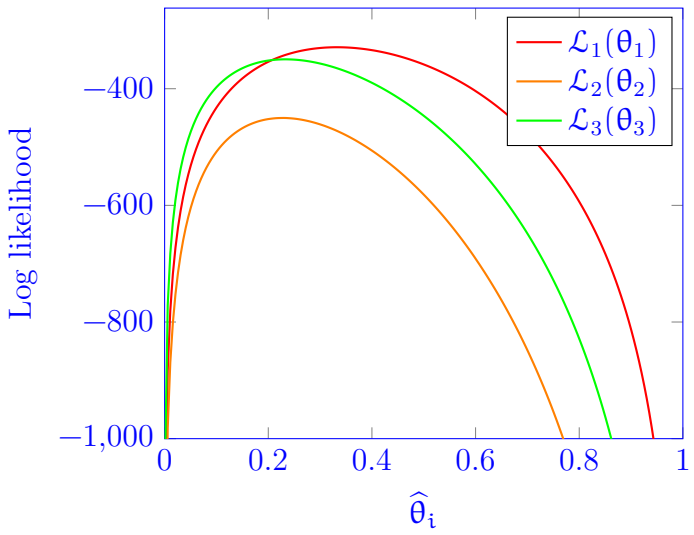


Figure 4.4: Three terms of the log-likelihood function

Figure 4.5 illustrates the maximum likelihood estimates (MLE) for the model parameters. For each parameter, we identify the point where the function reaches its peak. This corresponds to the maximum likelihood estimate, which represents the most probable value given the observed data.

In the figure, the maximum likelihood estimates of the probabilities associated with choosing public transport for each trip purpose are:

$$\widehat{\theta}_1 = 0.333$$

for work-related trips (marked in red),

$$\widehat{\theta}_2 = 0.228$$

for leisure trips (marked in orange), and

$$\widehat{\theta}_3 = 0.233$$

for other trips (marked in green).

It is important to note that the maximum likelihood estimates of the parameters correspond exactly to the observed frequencies in the contingency table. Since the likelihood function is maximized when the model accurately reflects the observed data, the estimated probabilities align with the relative frequencies of each category.

For each trip purpose, the probability of choosing public transport is given by the proportion of travelers who reported using public transport

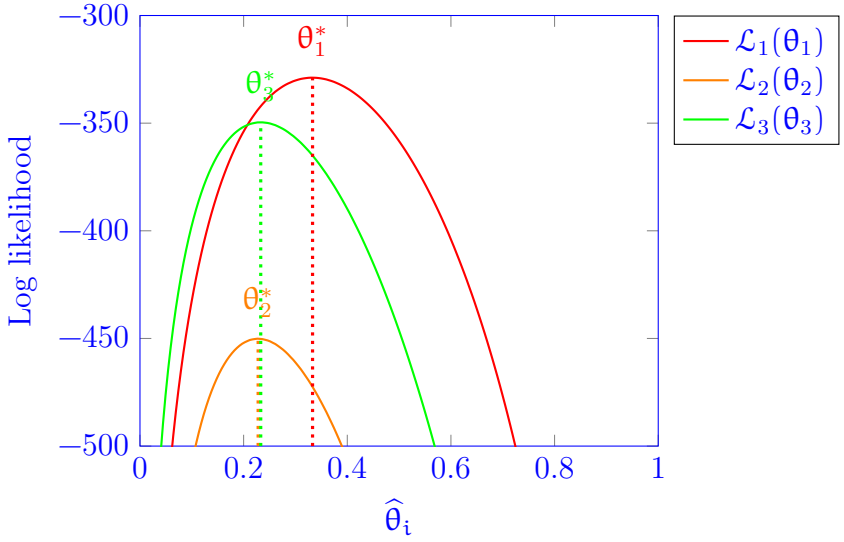


Figure 4.5: Maximum likelihood estimates of the parameters

within that category. Mathematically, this is expressed as:

$$\widehat{\theta}_1 = \frac{172}{517} = 0.333, \quad \widehat{\theta}_2 = \frac{191}{839} = 0.228, \quad \widehat{\theta}_3 = \frac{150}{644} = 0.233.$$

When analyzing data, it is important to recognize that different samples can lead to different estimates of the same underlying parameters. Suppose a colleague conducted an identical data collection process (reported in Table 4.2) and performed the same analysis. However, their results differ from ours. In our study, we estimated that the probability of choosing public transport for work-related trips is:

$$\widehat{\theta}_1 = \frac{172}{517} = 33.3\%.$$

In contrast, our colleague's data produced a slightly different estimate:

$$\widehat{\theta}_1 = \frac{168}{485} = 34.6\%.$$

This discrepancy arises because the sample of travelers surveyed is different. Even though the same methodology was applied, the random nature of data collection means that each sample provides a different realization of the estimator.

This introduces the concept of an estimator as a *random variable*. The estimated parameters are not fixed values but vary depending on the specific

	Work	Leisure	Others
PT	168	207	140
Not PT	317	677	491

Table 4.2: Data collected by another analyst.

sample drawn from the population. If another researcher repeated the study with a new random sample, they would likely obtain yet another slightly different estimate.

This variation highlights a fundamental challenge in statistical inference: how to draw reliable conclusions about the underlying population from a single sample. Since different samples yield different results, it becomes necessary to quantify the uncertainty associated with our estimates. This leads to further questions, such as how to measure the variability of an estimator and how to construct confidence intervals to assess the precision of our estimates.

The variation in the estimation of $\widehat{\theta}_1$ across different samples is illustrated in Figure 4.6. The histogram represents the distribution of $\widehat{\theta}_1$ computed from 1000 different random samples, each drawn from the same underlying population. The blue curve represents a probability density function that approximates this distribution.

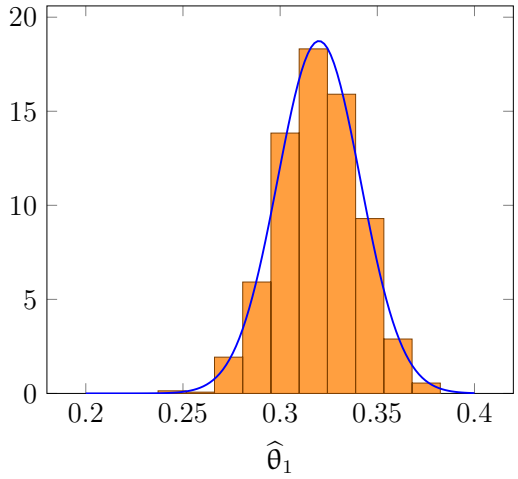


Figure 4.6: Distribution of $\widehat{\theta}_1$ across 1000 different samples

The key insight from this figure is that $\widehat{\theta}_1$ is not a fixed value but instead varies depending on the specific sample used for estimation. The variation

in $\widehat{\theta}_1$ is due to the randomness inherent in data collection. Even though each sample follows the same methodology, the individuals surveyed differ slightly, leading to fluctuations in the estimated parameter.

This variability suggests that our estimator $\widehat{\theta}_1$ is itself a *random variable* with a probability distribution. The shape of this distribution provides useful information about the reliability of our estimate. A narrower distribution (smaller spread) implies that $\widehat{\theta}_1$ is more stable across samples, meaning the estimate is more precise. Conversely, a wider distribution (larger spread) suggests greater uncertainty in the estimation.

Figure 4.6 illustrates the concept that estimators are random variables. The blue curve represents the probability density function of $\widehat{\theta}_1$, the estimator of θ_1. The true value of the parameter is $\theta_1 = 0.320$, shown by the green vertical line. However, different samples yield different estimates due to the randomness in data collection.

Two specific estimates are marked in orange: my estimate ($\widehat{\theta}_1 = 0.333$) and my colleague's estimate ($\widehat{\theta}_1 = 0.346$). These values are different from the true parameter because each sample contains a different set of travelers, leading to variations in the observed proportions. However, if we repeated this estimation process across many samples, the estimates would be distributed around the true value, forming the distribution seen in the figure.

Maximum likelihood estimation (MLE) ensures that, on average, the estimator is *unbiased*, meaning that the mean of this distribution coincides with the true value θ_1. This property implies that while any individual sample may produce an estimate that deviates from the true value, the overall estimation method is systematically correct in the long run.

The figure visually demonstrates how the estimator varies across samples. Some estimates will be below the true value, while others will be above, but the distribution is centered around θ_1. The spread of this distribution, determined by its standard deviation (standard error), quantifies the uncertainty associated with the estimator. A smaller standard error would result in a more concentrated distribution, leading to more precise estimates.

Figure 4.8 illustrates the relationship between sample size and the variance of an estimator. The blue curve represents the probability density function of the estimator $\widehat{\theta}_1$ when the sample size for work trips is $N_{\text{work}} = 2000$, while the orange curve corresponds to the case where the sample size is doubled to $N_{\text{work}} = 4000$. The true value of $\theta_1 = 0.32$ is marked by the green vertical line.

The figure demonstrates that as the sample size increases, the variance of the estimator decreases. This is reflected in the fact that the orange curve is more concentrated around θ_1, meaning that estimates are more precise when

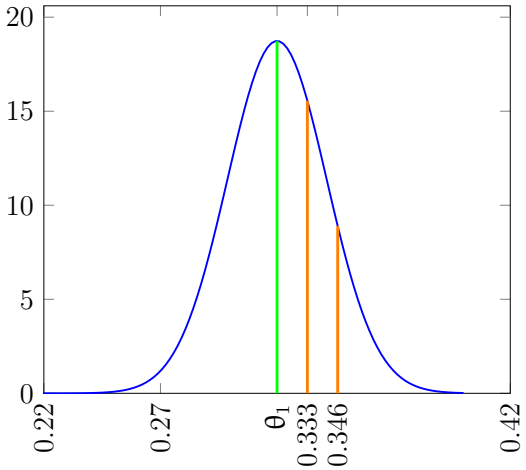


Figure 4.7: Estimator as a random variable

more data is available. Mathematically, this phenomenon is explained by the fact that the standard error of an estimator decreases with the square root of the sample size:

$$\mathrm{SE}(\widehat{\theta}_1) \propto \frac{1}{\sqrt{N}}.$$

Thus, when the sample size is quadrupled, the standard error is halved, leading to a narrower distribution of possible estimates.

In summary, maximum likelihood estimation (MLE) is a fundamental method for estimating parameters in statistical models. The key idea behind MLE is to find the parameter values that maximize the probability of observing the given data.

Formally, the likelihood function represents the probability that the model correctly predicts all observed values in the dataset. Given N observations (x_n, y_n), the likelihood function is defined as:

$$\begin{aligned}\mathcal{L}^*(\theta) &= \prod_{n=1}^{N} \Pr(Y = y_n, X = x_n; \theta) \\ &= \prod_{n=1}^{N} \Pr(Y = y_n | X = x_n; \theta) \Pr(X = x_n).\end{aligned}$$

In practice, working with the product of many probabilities can be computationally challenging due to numerical underflow. To simplify calculations,

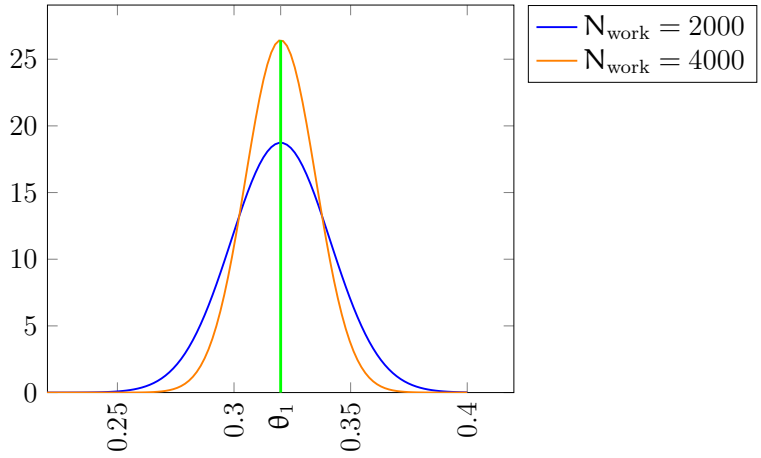


Figure 4.8: Larger sample, lower variance

the log-likelihood function is used:

$$\mathcal{L}(\theta) = \log \mathcal{L}^*(\theta) = \sum_{n=1}^{N} \log \Pr(Y = y_n | X = x_n; \theta) + \log \Pr(X = x_n).$$

Since the second term does not depend on θ, it is omitted.

The goal of MLE is to find the parameter θ that maximizes this log-likelihood function:

$$\widehat{\theta} = \operatorname{argmax}_\theta \mathcal{L}(\theta).$$

This results in the set of parameter values $\widehat{\theta}$ that make the observed data most probable under the model.

MLE has several desirable properties. It is *consistent*, meaning asympotically unbiased. Additionally, among all consistent estimators, MLE has the lowest possible variance, making it *efficient*. Finally, under standard regularity conditions, the MLE estimator is approximately normally distributed for large samples.

In summary, MLE provides a systematic way to estimate model parameters by maximizing the probability of the observed data. Its strong theoretical properties make it a widely used technique in statistical modeling and data analysis.

4.3.2 The case of continuous variables

We now introduce another example that involves continuous rather than discrete variables. This example is motivated by the data represented in

Figure 4.1 and Table 4.3, which originates from the Swiss Microcensus 2015 and provides insights into travel behavior.

The hypothesis in this case is that the daily distance traveled by individuals depends on their household income. The rationale behind this hypothesis is that individuals with different income levels engage in different socio-professional activities, which may influence their travel patterns. Additionally, income can affect access to various modes of mobility, such as private vehicles or long-distance public transport, further shaping travel behavior.

Monthly income (KCHF)	Daily distance (km)
2	22.49
6	36.11
10	45.35
12	51.59

Table 4.3: Collected data

To model the relationship between daily distance traveled and household income, we introduce a mathematical framework that describes how one continuous variable depends on another. Specifically, we define a model where the dependent variable Y represents the daily distance traveled (in kilometers), and the explanatory variable X corresponds to the household's monthly income (in thousands of Swiss francs). Given the nature of both variables, we seek to establish a functional relationship that captures how Y varies as a function of X.

A commonly used approach for modeling such relationships is *linear regression*, which assumes that the expected value of the dependent variable Y given X follows a linear function of X. Mathematically, we express this as:

$$Y|(X = x_n) = \theta_1 x_n + \xi_n, \quad \text{where} \quad \xi_n \sim N(\theta_0, \theta_2^2).$$

In this formulation, the parameter θ_1 represents the effect of income on the expected daily travel distance, while ξ_n captures random deviations from this relationship. These deviations are assumed to follow a normal distribution with mean θ_0 and variance θ_2^2. This accounts for the fact that while income may influence travel behavior, other unobserved factors contribute to variations in daily distance traveled.

Equivalently, the model can be rewritten as:

$$Y|(X = x_n) = \theta_1 x_n + \theta_0 + \theta_2 \xi_n,$$

where $\xi_n \sim N(0,1)$ represents an independent standard normal random variable across observations. The parameters θ_0, θ_1, and θ_2 are unknown and need to be estimated from the data.

Since we assume that the deviations from the linear relationship follow a normal distribution, the conditional distribution of Y given $X = x_n$ is also normally distributed. This allows us to write the probability density function as:

$$f_{Y|x_n}(z; \theta_0, \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2^2}} \exp\left(-\frac{1}{2}\left(\frac{z - \theta_1 x_n - \theta_0}{\theta_2}\right)^2\right).$$

This expression follows from the standard normal density function, where the mean of the distribution is given by the linear regression equation $\theta_1 x_n + \theta_0$, and the standard deviation is θ_2.

In the discrete case, the likelihood function represents the probability of observing the data given the model parameters. However, in the continuous case, the probability of any specific observation occurring is technically zero, since a continuous variable can take infinitely many values. Instead of using probability directly, we rely on the probability density function (pdf), which plays a similar role in expressing the likelihood of observing the given data.

To illustrate this, consider a single observation where $x_1 = 2$ and $y_1 = 22.49$. In the discrete case, we would compute the probability that our model predicts this observation correctly. However, since we are dealing with a continuous variable, the probability of observing exactly $y_1 = 22.49$ is zero. Instead, we use the pdf:

$$f_{Y|x_1}(y_1; \theta_0, \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2^2}} \exp\left(-\frac{1}{2}\left(\frac{y_1 - \theta_1 x_1 - \theta_0}{\theta_2}\right)^2\right).$$

The log-likelihood function then follows naturally by taking the logarithm of the pdf. For a single observation:

$$\log f_{Y|x_1}(y_1; \theta_0, \theta_1, \theta_2) = -\frac{1}{2}\log(2\pi) - \log(\theta_2) - \frac{1}{2\theta_2^2}(y_1 - \theta_1 x_1 - \theta_0)^2.$$

By summing over all observations in the dataset, we obtain the total log-likelihood function:

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \log f_{Y|x_n}(y_n; \theta) = -N\log(\theta_2) - \frac{1}{2\theta_2^2}\sum_{n=1}^{N}(y_n - \theta_1 x_n - \theta_0)^2.$$

To estimate the parameters θ_0, θ_1, and θ_2, we use the maximum likelihood estimation (MLE) approach, which consists of maximizing the log-likelihood

function:

$$\max_{\theta} \mathcal{L}(\theta) = -N \log(\theta_2) - \frac{1}{2\theta_2^2} \sum_{n=1}^{N} (y_n - \theta_1 x_n - \theta_0)^2.$$

The least-squares method provides a practical way to estimate the parameters of a linear regression model. Since the logarithm of the variance term, θ_2, is independent of θ_0 and θ_1, we begin by fixing $\theta_2 = \sigma$ and solving for these parameters.

$$\mathcal{L}(\theta) = -N \log(\sigma) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \theta_1 x_n - \theta_0)^2.$$

As the first term does not depend on θ, and the second term is associated with a negative sign, this transforms the optimization problem into minimizing the sum of squared residuals:

$$\min_{\theta} \frac{1}{2} \sum_{n=1}^{N} (y_n - \theta_1 x_n - \theta_0)^2.$$

The solution to this minimization problem provides the estimates $\widehat{\theta}_0$ and $\widehat{\theta}_1$.

Once these parameters are estimated, we turn to the estimation of σ. The residual sum of squares is given by:

$$z = \sum_{n=1}^{N} (y_n - \widehat{\theta}_1 x_n - \widehat{\theta}_0)^2.$$

Maximizing the log-likelihood function with respect to σ, we obtain:

$$\max_{\sigma} -N \log(\sigma) - \frac{1}{2\sigma^2} z.$$

Taking the derivative and solving for σ, we find:

$$\widehat{\sigma}^2 = \frac{z}{N}.$$

It is important to note that the two-step procedure of first estimating θ_0 and θ_1 using least squares and then estimating σ^2 separately does not provide the exact maximum likelihood estimates of the original model formulation. The primary reason for this discrepancy lies in the estimation of the variance σ^2.

This estimator is known to be biased. Specifically, it systematically underestimates the true variance because it does not account for the degrees of freedom lost due to estimating θ_0 and θ_1.

To correct for this bias, an unbiased estimator for the variance is given by:

$$\widehat{\sigma}^2_{\text{unbiased}} = \frac{1}{N-K}\sum_{n=1}^{N}(y_n - \widehat{\theta}_1 x_n - \widehat{\theta}_0)^2,$$

where $K = 2$ is the number of estimated parameters in the regression model. The adjustment from N to $N-K$ in the denominator accounts for the fact that the residuals are computed using estimated parameters, thereby reducing the available degrees of freedom.

While this correction provides an unbiased estimator for the variance, it also highlights that the two-step procedure is an approximation rather than the exact MLE solution. Despite this, the least squares approach remains widely used due to its simplicity and desirable properties, particularly when N is large, as the bias in $\widehat{\sigma}^2$ becomes negligible in such cases, as illustrated in Figure 4.9.

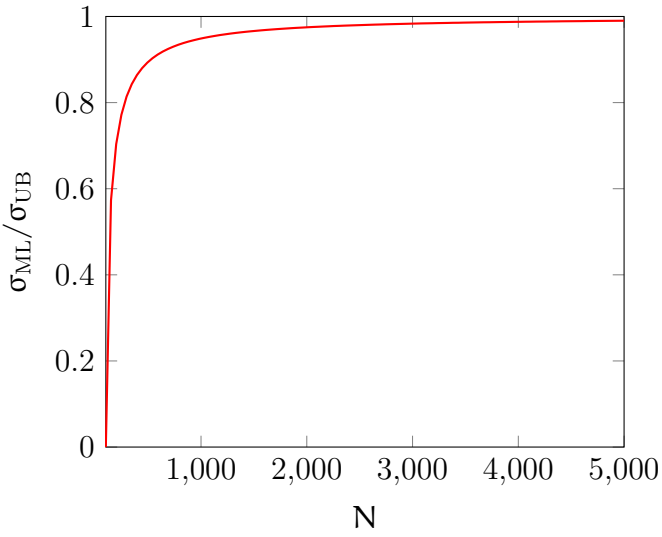


Figure 4.9: Ratio between two-step and unbiased estimate of σ, for $K = 100$

To summarize, linear regression is a statistical method used to model the relationship between a dependent variable Y and one or more explanatory variables X. The general formulation of a linear regression model assumes

that Y, given X, follows a linear function plus an error term:

$$Y|(X = x) = \sum_{k=1}^{K-1} \theta_k x_k + \theta_0 + \sigma\varepsilon.$$

Here, the parameters $\theta_0, \theta_1, \ldots, \theta_{K-1}$ define the linear relationship, while $\sigma\varepsilon$ represents a random error term that accounts for variations not explained by the model.

The expected value of Y given X, also known as the regression line, expresses the deterministic part of the relationship:

$$E[Y|X = x] = \sum_{k=1}^{K-1} \theta_k x_k + \theta_0.$$

In summary, the goal of estimation is to find the values of $\theta_0, \theta_1, \ldots, \theta_{K-1}$ that best fit the observed data. This is achieved through the least squares method, which minimizes the sum of squared residuals:

$$\widehat{\theta} = \text{argmin}_\theta \frac{1}{2} \sum_{n=1}^{N} (y_n - \sum_{k=1}^{K-1} \theta_k X_k - \theta_0)^2.$$

Once the parameters are estimated, the variance σ^2 of the residuals can also be estimated. Two commonly used estimators for the variance are:

$$\widehat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^{N} (y_n - \sum_{k=1}^{K-1} \widehat{\theta}_k X_k - \widehat{\theta}_0)^2$$

or, to obtain an unbiased estimator:

$$\widehat{\sigma}^2 = \frac{1}{N-K} \sum_{n=1}^{N} (y_n - \sum_{k=1}^{K-1} \widehat{\theta}_k X_k - \widehat{\theta}_0)^2.$$

The unbiased version accounts for the fact that K parameters have been estimated from the data, reducing the degrees of freedom. This adjustment ensures that the variance estimate is not systematically underestimated.

Figures 4.10 and 4.11 illustrate the optimization process for estimating the parameters in the least squares regression model. The first figure represents the objective function of the least squares problem in three dimensions, where the horizontal axes correspond to the regression parameters θ_0 and θ_1, and the vertical dimension represents the sum of squared residuals.

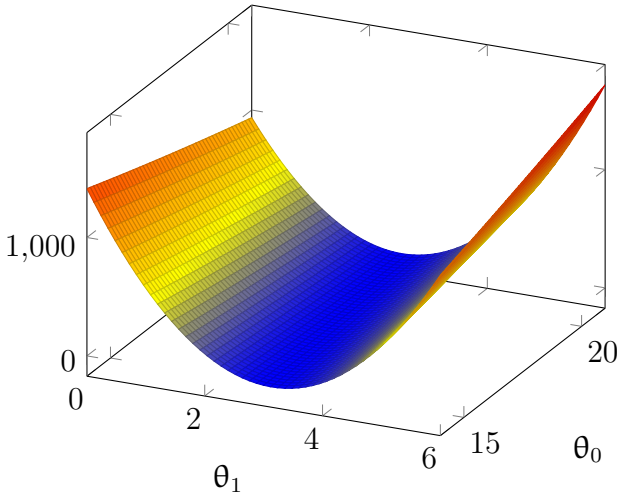


Figure 4.10: Objective function of the least-squares method

The second figure, Figure 4.11, presents the same objective function in the form of level curves, viewed from above. Each contour line represents a set of parameter values that yield the same sum of squared residuals. The contours become denser as they approach the minimum, indicating the region where the optimal parameter values are located. The estimated values, $\widehat{\theta}_0 = 17.6$ and $\widehat{\theta}_1 = 2.84$, are marked at the point where the function reaches its minimum.

These estimates suggest that the expected daily distance traveled by an individual is approximately 17.6 km when household monthly income is zero, and for every additional 1,000 CHF in income, the predicted travel distance increases by approximately 2.84 km. Additionally, the estimated standard deviation of the residuals, $\widehat{\sigma}$, is either 0.896 or 1.27, depending on whether the biased or unbiased estimator is used.

Figure 4.12 illustrates the estimated regression line along with a 99% confidence interval. The regression line, shown in red, represents the predicted relationship between household monthly income (in thousand CHF) and daily distance traveled (in km). It is given by the equation:

$$\widehat{Y} = 2.84X + 17.6.$$

The shaded region around the regression line represents the 99% confidence interval, which accounts for the uncertainty in our estimates. The confidence interval is computed using the estimated standard deviation of the residuals, $\widehat{\sigma} = 1.27$, and the critical value from the standard normal dis-

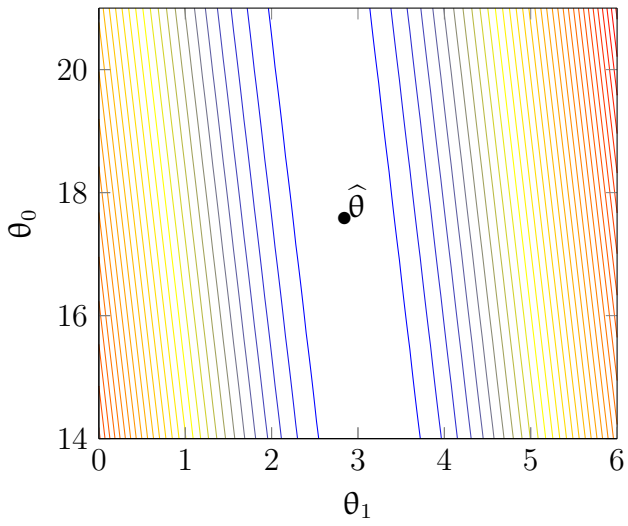


Figure 4.11: Objective function of the least-squares method: level curves

tribution corresponding to a 99% confidence level, which is approximately 2.576. The upper and lower bounds of the confidence interval are given by:

$$\widehat{Y} \pm 2.576 \cdot \widehat{\sigma}.$$

This means that for a given income level X, the true mean value of Y is expected to fall within this shaded region 99% of the time. The confidence interval reflects the variability in the data and the uncertainty in the estimated regression parameters. A wider confidence interval suggests greater uncertainty in predictions, while a narrower interval indicates more precise estimates.

Linear regression can be conveniently expressed in matrix form, which allows for a more compact representation of the problem and facilitates computational implementation. The general formulation of a linear regression model with K explanatory variables and N observations is given by:

$$y = x\theta + \sigma\varepsilon$$

where y is an N-dimensional vector representing the dependent variable, x is an $N \times K$ matrix containing the independent variables, θ is a K-dimensional vector of parameters to be estimated, and $\sigma\varepsilon$ represents the error term, where $\varepsilon \sim N(0, I)$ is a standard normal error term.

For our specific example with $K = 2$, we consider the relationship between

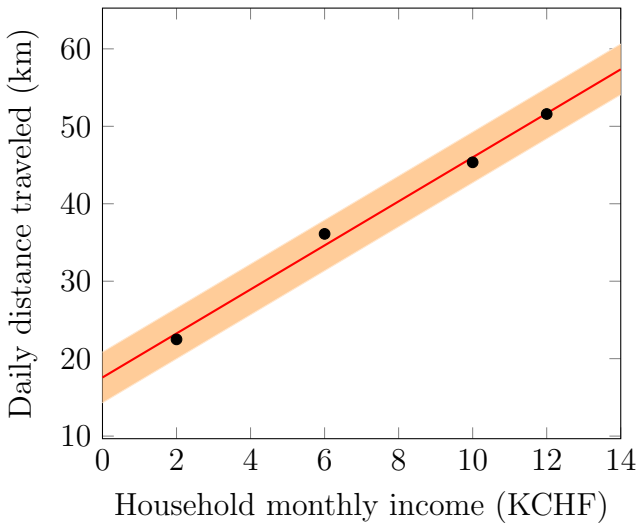


Figure 4.12: Regression line with confidence interval

household income and daily travel distance. The model can be written as:

$$\begin{bmatrix} 22.49 \\ 36.11 \\ 45.35 \\ 51.59 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 6 \\ 1 & 10 \\ 1 & 12 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} + \sigma \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

where the first column of x corresponds to the intercept term (a column of ones), and the second column contains the observed values of income.

To estimate the parameters, we solve the normal equations:

$$x^T x\widehat{\theta} = x^T y$$

which leads to the closed-form solution:

$$\widehat{\theta} = (x^T x)^{-1} x^T y.$$

For our example, computing $x^T x$ and $x^T y$, we obtain:

$$\begin{bmatrix} 4 & 30 \\ 30 & 284 \end{bmatrix} \begin{bmatrix} \widehat{\theta}_0 \\ \widehat{\theta}_1 \end{bmatrix} = \begin{bmatrix} 155.74 \\ 1334.22 \end{bmatrix}.$$

Solving for $\widehat{\theta}$, we find:

$$\widehat{\theta}_0 = 17.6, \quad \widehat{\theta}_1 = 2.84.$$

The residuals, representing the differences between observed and predicted values, are given by:

$$y - x\widehat{\theta} = \begin{bmatrix} -0.7647 \\ 1.4878 \\ -0.6397 \\ -0.0834 \end{bmatrix}.$$

The sum of squared residuals is:

$$(y - x\widehat{\theta})^{\mathsf{T}}(y - x\widehat{\theta}) = 3.2145.$$

From this, we estimate the standard deviation of the residuals. The maximum likelihood estimator of σ is:

$$\widehat{\sigma} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - x_n^{\mathsf{T}}\widehat{\theta})^2} = \sqrt{\frac{3.2145}{4}} = 0.896.$$

The unbiased estimator of σ, which accounts for the degrees of freedom, is:

$$\widehat{\sigma} = \sqrt{\frac{1}{N-K}\sum_{n=1}^{N}(y_n - x_n^{\mathsf{T}}\widehat{\theta})^2} = \sqrt{\frac{3.2145}{4-2}} = 1.27.$$

This formulation in matrix notation generalizes naturally to multiple explanatory variables and provides an efficient way to estimate regression parameters.

4.3.3 The case of both discrete and continuous variables

So far, we have explored different approaches to parameter estimation, depending on the nature of the dependent variable Y and the explanatory variable X. When both X and Y are discrete, we used contingency tables to estimate probabilities and analyze relationships. In contrast, when both variables are continuous, we applied linear regression, deriving estimates using the least-squares method. We now investigate the case when some variables are discrete, and some are continuous.

A straightforward case arises when the dependent variable Y is continuous, while the explanatory variable X is discrete. In order to be used as an explanatory variable in a regression model, it must first be encoded in a way that allows it to be incorporated into a mathematical framework. This is typically achieved by introducing binary (or dummy) variables that represent the different categories of the qualitative variable.

For example, consider the categorical variable X representing the level of comfort, which has four possible values: "very comfortable," "comfortable," "rather comfortable," and "not comfortable." Since these categories are not naturally numerical, we define four binary variables:

$$z_{\text{vc}}, \quad z_{\text{c}}, \quad z_{\text{rc}}, \quad z_{\text{nc}}.$$

Each of these variables takes a value of 1 if the observation belongs to the corresponding category and 0 otherwise. This encoding is summarized in the following table:

	z_{vc}	z_{c}	z_{rc}	z_{nc}
very comfortable	1	0	0	0
comfortable	0	1	0	0
rather comfortable	0	0	1	0
not comfortable	0	0	0	1

Once the categorical variable is represented in this way, it can be incorporated into a regression model. The dependent variable Y, which may represent a continuous outcome such as user satisfaction or willingness to pay, can be expressed as a function of these binary variables:

$$\mathsf{Y} = \cdots + \theta_1 z_{\text{vc}} + \theta_2 z_{\text{c}} + \theta_3 z_{\text{rc}} + \theta_4 z_{\text{nc}} + \sigma\varepsilon.$$

This formulation allows us to estimate the impact of each comfort level on Y using standard linear regression techniques. The coefficients $\theta_1, \theta_2, \theta_3$, and θ_4 represent the expected value of Y for each category. The model can be estimated using least squares, following the same principles as in a standard regression with continuous variables.

A more complex situation occurs when Y is discrete and X is continuous. This setup requires different modeling techniques since neither contingency tables nor standard regression are directly applicable.

Indeed, predicting a discrete outcome from continuous explanatory variables presents challenges that do not arise in standard regression models. Consider a traveler's decision to use public transportation or an alternative mode of transport. This choice can be represented by a qualitative variable Y that takes two possible values: "public transport" or "other." The decision is influenced by factors such as travel time X_1 and travel cost X_2, both of which are continuous variables.

A natural first attempt might be to apply a linear regression model:

$$\mathsf{Y} = \theta_1 \mathsf{X}_1 + \theta_2 \mathsf{X}_2 + \sigma\varepsilon.$$

However, this approach is inappropriate because the left-hand side of the equation, Y, represents a categorical choice, while the right-hand side is a continuous function. The model would predict values of Y that are not constrained to discrete outcomes, making interpretation difficult and leading to meaningless results.

A better approach is to return to utility theory, which models the decision-making process by assuming that each traveler associates a level of utility with each available choice. The traveler selects the option that provides the highest utility.

Consider again the example of a choice between "public transportation" and "not public transportation" described in Section 3.1, illustrated in Figure 3.1. Suppose we have collected data on travelers' choices. For each traveler, we observe the travel time for both alternatives, the travel cost for both alternatives, and the mode of transportation that was actually chosen.

In Figure 4.13, each point represents an individual traveler's data. The x-coordinate of each point corresponds to the observed difference in travel times between the two alternatives, $t_1 - t_2$, while the y-coordinate represents the observed difference in travel costs, $c_1 - c_2$. The shape of each point indicates the traveler's actual choice: circles represent travelers who chose alternative 1, while squares represent those who selected alternative 2.

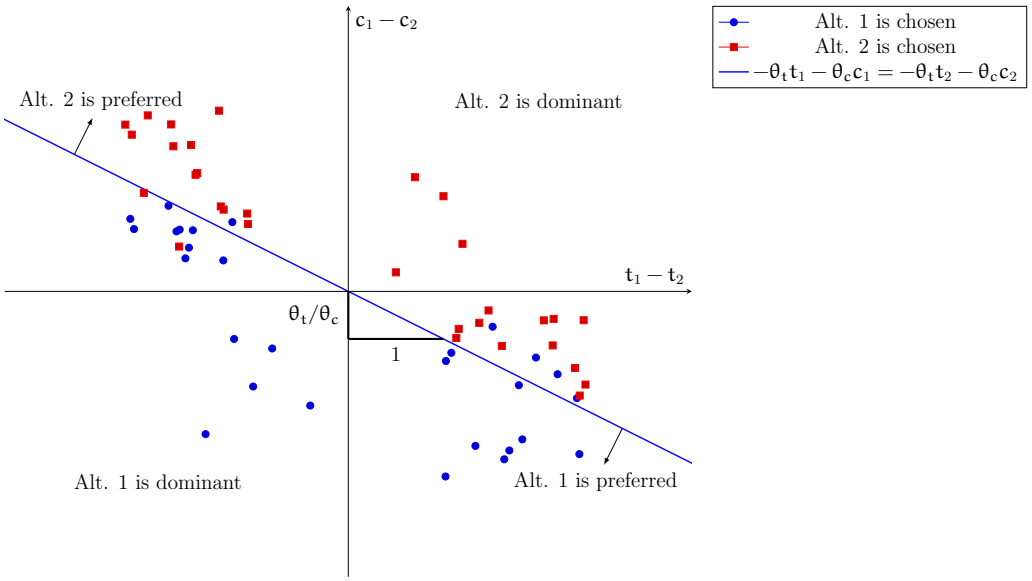


Figure 4.13: Utility model and observed data

Intuitively, the slope of the line, defined by the ratio of parameters θ_t/θ_c (that is, the value of time) should be estimated in such a way that the in-

difference line acts as an optimal separator between the observed choices, completely distinguishing the circles (representing travelers who chose alternative 1) from the squares (representing travelers who chose alternative 2).

This is because the indifference line, defined by the equation

$$-\theta_t(t_1 - t_2) - \theta_c(c_1 - c_2) = 0,$$

represents the threshold at which a traveler is equally likely to choose either alternative, as discussed in Section 3.1. If the estimated parameters accurately reflect real-world decision-making, then all observations corresponding to alternative 1 should ideally fall on one side of the indifference line, while all observations corresponding to alternative 2 should lie on the other side.

However, it is clearly impossible to find a single slope that perfectly separates the two sets of observations. In other words, no choice of the parameter ratio θ_t/θ_c can define an indifference line that completely divides the circles (representing travelers who chose alternative 1) from the squares (representing travelers who chose alternative 2).

The reason for this is that real-world choice data is inherently noisy and influenced by many unobserved factors. Travelers do not make perfectly deterministic decisions based solely on travel time and cost. Instead, their choices are affected by personal preferences, habitual behaviors, comfort, reliability, accessibility, and other latent[1] variables that are not explicitly captured in the observed data.

Mathematically, this means that there will always be some travelers whose choices appear inconsistent with a strict deterministic model. For example, some travelers may opt for public transportation even when it is slower and more expensive, perhaps due to factors like convenience or the ability to work during the trip. Conversely, others may prefer driving despite a longer and costlier journey due to personal comfort or flexibility. As a result, the data points corresponding to different choices are mixed in the $(t_1 - t_2, c_1 - c_2)$ plane, with no clear linear boundary that can separate them entirely.

This observation suggests that a purely deterministic approach to modeling discrete choices is insufficient. Instead, we must adopt a probabilistic framework that acknowledges the inherent variability in human decision-making.

In practice, the analyst does not have direct access to the true preference structure of the decision-maker and, consequently, to the exact utility function. Approximation arises from various sources, such as missing variables, an incorrect functional form, or measurement errors in observed variables.

[1] A latent variable is actually an unobserved variable.

To account for these unobserved factors, the concept of random utility is introduced. In this framework, the utility of an alternative is represented as a continuous random variable, composed of a deterministic component and an unobserved random component.

Mathematically, this can be expressed as

$$U_i = u_i + \varepsilon_i = -\theta_t t_i - \theta_c c_i + \varepsilon_i,$$

where u_i represents the deterministic component, which depends on observable attributes such as travel time t_i and cost c_i. The term ε_i is a random component that captures unobserved factors, including individual preferences, past experiences, and contextual influences.

This equation closely resembles a linear regression model. However, a fundamental difference is that the dependent variable U_i is not directly observed. As a result, the least-squares estimation method described in Section 4.3.2 cannot be applied in this context.

Since individuals seek to maximize their utility, they choose the alternative that provides the highest utility. From the point of view of the analyst, this leads to the probability of selecting an alternative i over another alternative j being given by

$$\Pr(Y = i) = \Pr(U_i \geqslant U_j).$$

The probability of making a particular choice depends not only on the deterministic utility difference but also on the distribution of the random components ε_i and ε_j. Because these random components introduce uncertainty, the model predicts the probability of each alternative being chosen rather than determining choices with absolute certainty.

The structure of the model follows a causal relationship where the observed choice Y is determined by the underlying latent utilities U_i, which in turn depend on the observed explanatory variables X. This relationship can be written as

$$Y|U|X.$$

This notation emphasizes that the explanatory variables X influence the choice indirectly through the latent utilities. Unlike standard regression models, where both explanatory and dependent variables are observed, random utility models introduce an additional layer of complexity by incorporating latent variables. The true utilities U_i are never observed directly; instead, only the final choice Y is recorded.

The *logit* model is a widely used discrete choice model that describes the probability of selecting an alternative from a given set of choices. Consider a decision-maker faced with a set of alternatives, denoted by C. Each

alternative $i \in \mathcal{C}$ is associated with a utility U_i, which is composed of a deterministic component u_i and an unobserved random component ε_i, such that $U_i = u_i + \varepsilon_i$.

If the unobserved components ε_i are independently and identically distributed (i.i.d.) following the Extreme Value distribution with parameters $(0, \mu)$, then the probability that alternative i is chosen is given by:

$$\Pr(Y = i) = \Pr(U_i \geqslant U_j, \forall j \in \mathcal{C}) = \frac{e^{\mu u_i}}{\sum_{j \in \mathcal{C}} e^{\mu u_j}}.$$

The parameter μ scales the sensitivity of the choice probabilities to differences in utility, with higher values of μ indicating more deterministic choices and lower values introducing greater randomness in decision-making.

In random utility models, the probabilities of choosing an alternative depend only on differences in utility rather than their absolute values. This property leads to two fundamental invariances: shift invariance and scale invariance.

Shift invariance means that adding a constant K to all utility values does not change the choice probabilities. Formally, if each utility function is shifted by the same constant, the probability of selecting an alternative remains the same:

$$\Pr(Y = i) = \Pr(U_i + K \geqslant U_j + K, \forall j \in \mathcal{C}), \quad \forall K \in \mathbb{R}.$$

This implies that only differences in utility matter, not their absolute levels. As a consequence, one of the utility functions can be normalized by setting one intercept to zero without loss of generality.

Scale invariance refers to the property that multiplying all utilities by a positive constant μ does not affect the choice probabilities:

$$\Pr(Y = i) = \Pr(\mu U_i \geqslant \mu U_j, \forall j \in \mathcal{C}), \quad \forall \mu > 0.$$

Since utility is a latent construct without a natural unit of measurement, the scale of the utility function is arbitrary. This means that the model is identified only up to scale, and for estimation purposes, the scale parameter is typically normalized to one.

These two properties imply that when estimating discrete choice models, a normalization is required. One intercept is often fixed at zero to account for shift invariance, and the scale of the utility function is standardized by setting a parameter (often μ in logit models) to one.

Back to the example, Figure 4.14 illustrates the probability of choosing public transportation (PT) or not (not PT) as a function of the difference in utility between these two alternatives. The x-axis represents the difference in

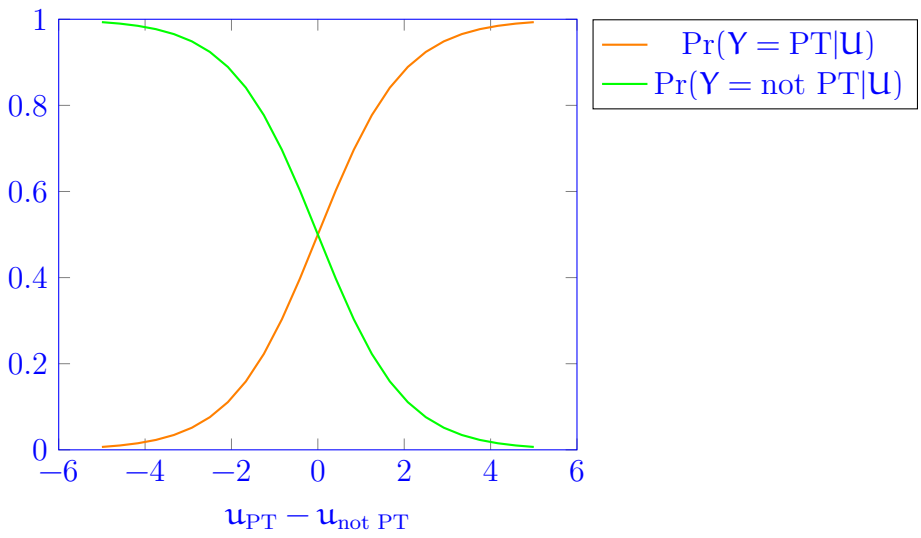


Figure 4.14: Example of a logit model

deterministic utility, defined as $u_{PT} - u_{\text{not PT}}$. This difference quantifies the relative attractiveness of public transportation compared to the alternative.

The y-axis represents the probability of selecting each alternative. The orange curve corresponds to $\Pr(Y = PT|U)$, which is the probability that a traveler chooses public transportation given the utility values. The green curve corresponds to $\Pr(Y = \text{not PT}|U)$, representing the probability of selecting the alternative option.

The logit function governs the shape of these curves. When $u_{PT} - u_{\text{not PT}} = 0$, meaning both alternatives provide the same deterministic utility, the probability of choosing either option is 50%. As $u_{PT} - u_{\text{not PT}}$ increases, public transportation becomes more attractive, and its probability approaches 1. Conversely, as $u_{PT} - u_{\text{not PT}}$ decreases, the probability of choosing not PT approaches 1, meaning the alternative mode is preferred.

To estimate the parameters of a choice model, we need data, that is, observation of real choices. Table 4.4 presents an example of choice data collected from individual travelers. Each row in the table corresponds to a separate observation, representing the travel decision of a single individual. The first column, labeled '#', is an index that uniquely identifies each observation. The second and third columns provide information on the travel times associated with two alternative modes: the time taken by car ('Time car') and the time taken by public transportation ('Time PT'). These values are measured in consistent units, typically minutes, and reflect the travel conditions faced by each individual at the time of decision-making.

The fourth column, labeled 'Choice', records the mode of transportation that was actually chosen by each traveler. The notation 'T' indicates that the traveler opted for public transportation, while 'C' indicates that the traveler chose to travel by car. Since each individual has only one chosen alternative per observation, this column provides the outcome variable for estimating the choice model.

#	Time car	Time PT	Choice
1	52.9	4.4	T
2	4.1	28.5	T
3	4.1	86.9	C
4	56.2	31.6	T
5	51.8	20.2	T
6	0.2	91.2	C
7	27.6	79.7	C
8	89.9	2.2	T
9	41.5	24.5	T
10	95.0	43.5	T
11	99.1	8.4	T
12	18.5	84.0	C
13	82.0	38.0	C
14	8.6	1.6	T
15	22.5	74.1	C
16	51.4	83.8	C
17	81.0	19.2	T
18	51.0	85.0	C
19	62.2	90.1	C
20	95.1	22.2	T
21	41.6	91.5	C

Table 4.4: Example of choice data

Assume that the utility of traveling by car, denoted as u_{C1}, is given by:

$$u_{C1} = \theta_1 t_{C1}.$$

Similarly, the utility of traveling by public transportation, denoted as u_{T1}, incorporates an additional parameter θ_T:

$$u_{T1} = \theta_1 t_{T1} + \theta_T.$$

For the sake of illustration, consider the following parameter values $\theta_T = 0.5$ and $\theta_1 = -0.1$. Using the travel times observed for the first individual

($t_{C1} = 52.9$ minutes and $t_{T1} = 4.4$ minutes), we can compute the specific utility values:

$$u_{C1} = (-0.1) \times 52.9 = -5.29,$$
$$u_{T1} = (-0.1) \times 4.4 + 0.5 = 0.06.$$

Since the first individual chose public transportation (PT), the probability assigned by the model to this choice must be computed using the logit formula, where we have normalized $\mu = 1$:

$$P_1(\text{PT}) = \frac{e^{u_{T1}}}{e^{u_{T1}} + e^{u_{C1}}}.$$

Substituting the computed utility values:

$$P_1(\text{PT}) = \frac{e^{0.06}}{e^{0.06} + e^{-5.29}}.$$

Since $e^{-5.29}$ is a very small number compared to $e^{0.06}$, the denominator is approximately equal to $e^{0.06}$, leading to a probability close to 1. This result indicates that, given the model parameters, the choice of public transportation for this individual is almost fully explained by the estimated utility function.

For the second individual, we use the given parameter values $\theta_T = 0.5$ and $\theta_1 = -0.1$, along with the observed travel times: $t_{C2} = 4.1$ minutes and $t_{T2} = 28.5$ minutes. Substituting these values into the utility equations:

$$u_{C2} = (-0.1) \times 4.1 = -0.41,$$
$$u_{T2} = (-0.1) \times 28.5 + 0.5 = -2.35.$$

Since the second individual also chose public transportation (PT), the probability assigned by the model to this choice is computed using the logit formula:

$$P_2(\text{PT}) = \frac{e^{u_{T2}}}{e^{u_{T2}} + e^{u_{C2}}}$$

Substituting the computed utility values:

$$P_2(\text{PT}) = \frac{e^{-2.35}}{e^{-2.35} + e^{-0.41}}$$

The probability of choosing public transportation for this individual is approximately 0.13.

If we consider the two first individuals, the probability that the model correctly predicts both choices is:

$$P_1(\text{PT})P_2(\text{PT}) = 0.13.$$

This quantity is the likelihood, introduced in Section 4.3.1. It is a function of the unknown parameters. The value obtained above assummes $\theta_T = 0.5$ and $\theta_1 = -0.1$. When extended to all individuals in the dataset, the likelihood function becomes:

$$\mathcal{L}^*(\theta) = \prod_n (P_n(\text{car}|\theta)^{y_{\text{car},n}} P_n(\text{PT}|\theta)^{y_{\text{PT},n}})$$

where $y_{j,n}$ is an indicator variable that takes the value 1 if individual n has chosen alternative j and 0 otherwise. This ensures that only the probability corresponding to the chosen alternative contributes to the likelihood.

Since the likelihood function involves the product of many probabilities, it is often more convenient to work with the logarithm of the likelihood, known as the log-likelihood function:

$$\begin{aligned}\mathcal{L}(\theta) &= \log \mathcal{L}^*(\theta) \\ &= \sum_n \left(y_{\text{car},n} \log P_n(\text{car}|\theta) + y_{\text{PT},n} \log P_n(\text{PT}|\theta)\right).\end{aligned}$$

Table 4.5 presents the values of the likelihood function for different parameter values in the choice model. The table includes four rows, each corresponding to a different combination of the parameters θ_T and θ_1, and the resulting value of the likelihood function $\mathcal{L}^*$.

θ_T	θ_1	$\mathcal{L}^*$
0	0	$4.57\ 10^{-07}$
0	-1	$1.97\ 10^{-30}$
0	-0.1	$4.1\ 10^{-04}$
0.5	-0.1	$4.62\ 10^{-04}$

Table 4.5: Value of th elikelihood for some values of the parameters

The first column represents θ_T, the alternative-specific constant associated with public transportation. The second column represents θ_1, the coefficient associated with travel time. The third column provides the corresponding likelihood value $\mathcal{L}^*$.

Examining the values in the table, we observe that the likelihood is 4.57×10^{-7}. When θ_1 is set to -1, the likelihood drops significantly to 1.97×10^{-30}, meaning that a very strong sensitivity to travel time leads to poorer predictions.

Setting $\theta_1 = -0.1$, the likelihood improves considerably, reaching 4.1×10^{-4}. When θ_T is also adjusted to 0.5 while keeping $\theta_1 = -0.1$, the likelihood

slightly increases to 4.62×10^{-4}. This suggests that incorporating a small alternative-specific preference for public transportation improves the model's predictive power.

Overall, this table highlights the importance of selecting appropriate parameter values to maximize the likelihood function. The goal of estimation is to identify the values of θ_T and θ_1 that yield the highest likelihood, ensuring that the model best represents the observed choices.

Figure 4.15 represents the log-likelihood function for the example choice model. The plot visualizes how the log-likelihood varies as a function of the two parameters θ_1 and θ_T, which are estimated to best explain the observed choices.

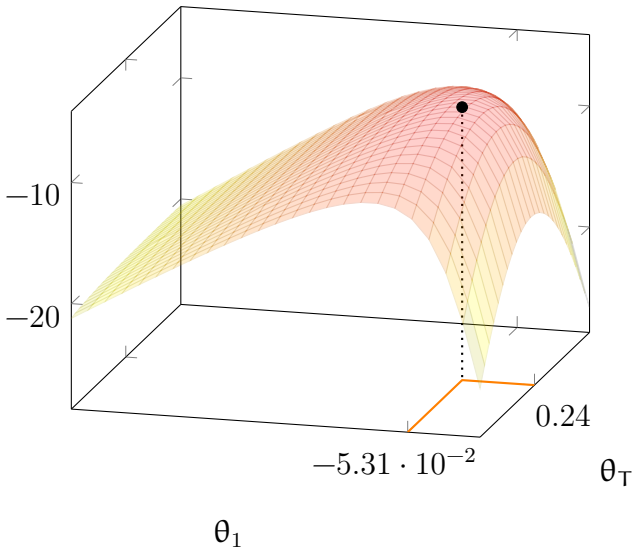


Figure 4.15: Log-likelihood function for the example

The horizontal axes represent the parameters of the model. The x-axis corresponds to θ_1, the coefficient associated with travel time, and the y-axis corresponds to θ_T, the alternative-specific constant for public transportation. The vertical axis represents the value of the log-likelihood function, which measures how well a given pair of parameters explains the observed choices.

The surface plot illustrates how different parameter values affect the log-likelihood. The maximum of the log-likelihood function is indicated by the orange lines and the black dot, which corresponds to the optimal parameter estimates. The maximum occurs at approximately $\theta_1 = -0.0531$ and $\theta_T = 0.2376$. These values maximize the probability of correctly predicting the choices observed in the dataset.

Since the log-likelihood function is concave, it ensures that the estimation process leads to a unique maximum. The contour of the surface suggests how sensitive the likelihood is to changes in the parameters: a steep slope indicates that small changes in the parameters significantly affect the likelihood, while a flatter region indicates that small changes have a minimal impact.

Figure 4.16 illustrates how the probability of choosing public transportation or car varies as a function of travel time by public transportation.

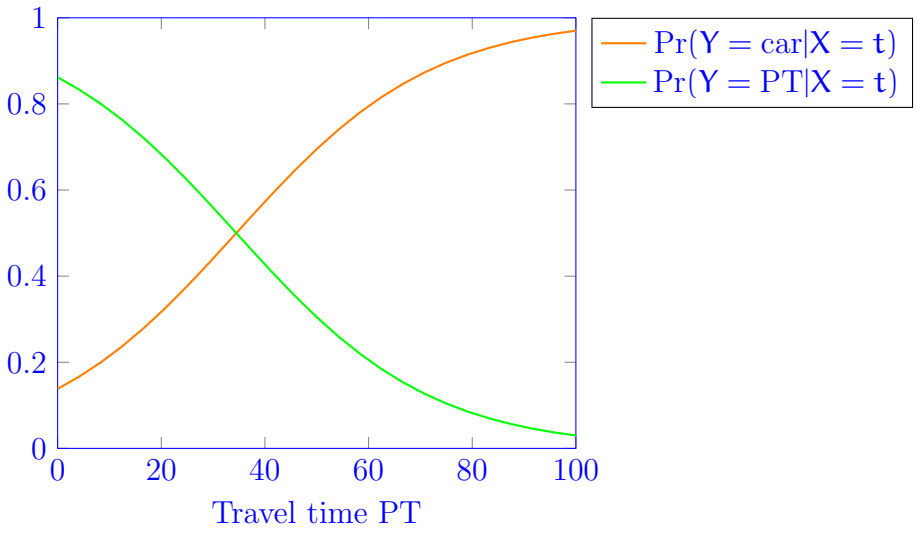


Figure 4.16: Choice probability of public transportation, assuming travel time by car is 30 minutes

The x-axis represents the travel time by public transportation, measured in minutes. The y-axis represents the probability of choosing each transportation mode. The two curves correspond to the probabilities predicted by the estimated choice model, assuming that the travel time by car is fixed at 30 minutes.

The orange curve represents $\Pr(Y = \text{car}|X = t)$, the probability that an individual chooses to travel by car given that the travel time by public transportation is t. The green curve represents $\Pr(Y = \text{PT}|X = t)$, the probability of choosing public transportation.

The model predicts that when public transportation is significantly faster than 30 minutes, the probability of choosing it is high, and the probability of choosing the car is low. Conversely, when the travel time by public transportation is much longer than 30 minutes, the probability of choosing the car increases, while the probability of choosing public transportation decreases.

The two probabilities sum to one for any given value of $\mathtt{t}$, as an individual must choose exactly one of the two alternatives.

4.3.4 Back to the contingency table

We now revisit the treatment of discrete variables described in Section 4.3.1. Consider the contingency table shown in Table 4.1, that summarizes the observed choices of individuals based on their trip purpose. The rows correspond to whether an individual chose public transportation (PT) or not, while the columns correspond to the purpose of the trip: work, leisure, or other. The goal is to model the probability of choosing PT as a function of the trip purpose.

To achieve this, we introduce a utility-based framework similar to the logit model used in discrete choice analysis. We define a utility function for the public transportation alternative:

$$u_{\text{PT}} = \theta_1 z_{\text{work}} + \theta_2 z_{\text{leis}} + \theta_3 z_{\text{others}}$$

where z_{work}, z_{leis}, and z_{others} are binary variables indicating the trip purpose. For example, if an individual is traveling for work, then $z_{\text{work}} = 1$ and the other two variables are zero. Similarly, if the purpose is leisure, then $z_{\text{leis}} = 1$, and for other purposes, $z_{\text{others}} = 1$. The utility of the "not PT" alternative is normalized to zero:

$$u_{\text{not PT}} = 0.$$

Following the logit model formulation, the probability of choosing PT is given by:

$$\Pr(\text{PT}) = \frac{e^{u_{\text{PT}}}}{e^{u_{\text{PT}}} + e^{u_{\text{not PT}}}} = \frac{e^{u_{\text{PT}}}}{1 + e^{u_{\text{PT}}}}.$$

By estimating the parameters using maximum likelihood estimation, we obtain:

	Work	Leisure	Others
θ_i^*	-0.696	-1.22	-1.19
u_{PT}	-0.696	-1.22	-1.19
$\Pr(\text{PT})$	0.333	0.228	0.233

These results show that the probabilities predicted by the logit model exactly match the relative frequencies observed in the contingency table. This confirms that any discrete outcome model, including those derived from contingency tables, can be formulated as a logit model.

In conclusion, the logit model provides a general framework for modeling discrete choices, whether they come from structured behavioral models (such as those based on travel time and cost) or from observed categorical data

(such as trip purpose). The flexibility of the logit model allows it to be applied in a wide range of settings where the dependent variable is discrete.

4.4 Summary

This chapter provided a structured introduction to mathematical modeling, focusing on different types of variables, the distinction between correlation and causality, and the estimation of model parameters.

The first part of the chapter explored the nature of variables used in models. Variables can be continuous, such as travel time or income, qualitative discrete, such as transportation mode choice, or random, incorporating uncertainty into the model.

A key concept discussed was causality, which is distinct from correlation. While correlation measures the association between two variables, it does not imply a cause-and-effect relationship. Causality is context-dependent and must be justified through theoretical assumptions and hypotheses. This distinction is fundamental when developing models that aim to explain and predict behavior.

The chapter then introduced two primary modeling frameworks based on the nature of the dependent variable. When the dependent variable Y is continuous, linear regression provides a suitable modeling approach, expressed as:

$$Y|(X = x_n) = \sum_{k=1}^{K-1} \theta_k x_{nk} + \theta_0 + \sigma\varepsilon.$$

For cases where the dependent variable is discrete, the random utility model forms the basis of choice modeling. The logit model was introduced as a method to estimate the probability of selecting an alternative i, given by:

$$\Pr(Y = i|X = x_n) = \frac{e^{u_i(x_n)}}{\sum_{j \in \mathcal{C}} e^{u_j(x_n)}},$$

where the utility of each alternative is modeled as:

$$u_i(x_n) = \sum_{k=1}^{K-1} \theta_k x_{ink} + \theta_0.$$

Another important aspect covered was the treatment of discrete independent variables. When an explanatory variable is categorical, it can be modeled using binary variables to represent each possible category. This allows discrete attributes to be seamlessly incorporated into models that traditionally use continuous predictors.

Parameter estimation was discussed in detail, emphasizing the use of maximum likelihood estimation. This method ensures that model parameters are estimated in a way that maximizes the probability of reproducing observed data. Maximum likelihood estimation was applied to both regression models for continuous outcomes and discrete choice models like the logit model.

In summary, this chapter provided foundational tools for mathematical modeling, covering different types of variables, model specification, and estimation techniques. The concepts introduced here serve as a basis for developing models that can explain and predict behaviors in various domains.

Chapter 5

Travel demand: an introduction

People rarely travel for the sole purpose of moving from one place to another. Instead, travel demand is a derived demand: it arises because individuals need to reach different locations to engage in activities such as work, education, shopping, or social interactions. The true demand is not for travel itself but for participation in these activities, which are essential to daily life.

Since activities are spread across space and time, individuals must travel to access them. This spatial and temporal dispersion shapes mobility patterns, influencing how, when, and why people move. Travel choices depend on the availability of transportation options, the location of homes and workplaces, personal schedules, and constraints such as time and cost. The demand for transportation services is therefore directly linked to the distribution and scheduling of activities.

Human activities can be broadly categorized into primary and secondary activities, each playing a distinct role in shaping travel demand.

Primary activities are essential to daily life and often have fixed locations and schedules. These include activities that take place at home, such as rest and personal care, as well as work and education. Work commitments generally follow regular hours and require commuting to specific locations, while education, whether at schools or universities, follows structured timetables that influence mobility patterns. Since these activities are fundamental, they strongly determine how people organize their time and where they need to travel.

Secondary activities, on the other hand, are more flexible in nature and include leisure, shopping, escorting others, and business-related travel. Leisure activities encompass a wide range of pursuits such as entertainment, sports, and social gatherings, often undertaken at discretionary times and locations. Shopping can range from routine grocery purchases to more occasional trips for other goods and services. Escort trips involve accompanying others, such

as taking children to school or picking up a family member. Business-related travel extends beyond commuting to work and includes meetings, site visits, and other professional engagements that require movement across different locations. These secondary activities contribute significantly to travel patterns, often leading to additional trips beyond the primary commuting journey.

The distinction between primary and secondary activities is important for understanding mobility needs. While primary activities create a baseline structure for travel, secondary activities introduce additional complexity by adding variability in trip timing, frequency, and destination choice.

Travel demand is the result of a combination of choices made by different actors at various levels. It is shaped by decisions made by public authorities, as well as by households and individuals, each contributing to how, when, and where people travel.

Public authorities play a central role in shaping travel demand through a series of decisions made at different time horizons. These choices influence the structure of cities, the availability of transportation options, and the way people move in response to daily needs and unexpected events.

Some of these decisions have long-term consequences, particularly in the areas of urban planning and land use. The way cities are designed, where housing developments are located, and how commercial and industrial zones are distributed all affect travel patterns for decades. Infrastructure investments, such as the construction of a new metro line, not only improve accessibility but also influence real estate values and economic activities in surrounding areas. These large-scale projects define the overall mobility landscape and shape the choices available to individuals and businesses.

Other decisions are made on a mid-term basis and focus on regulations that govern transportation systems and urban activities. Policies such as traffic restrictions, environmental regulations, and parking rules influence how people navigate the city. Additionally, the scheduling of public events such as concerts, sports games, and festivals requires careful planning to manage crowds and minimize disruptions. By coordinating opening hours and event logistics, public authorities can reduce congestion and optimize transportation networks.

Some choices must be made in the short term to respond to immediate challenges. Crisis management is designed to ensure safety and continuity in times of disruption. Natural disasters, extreme weather events, and public health emergencies require rapid decisions such as closing schools during storms or implementing lockdown measures to control the spread of disease. These short-term actions, while temporary, can have significant impacts on travel behavior and mobility patterns.

At the same time, households and individuals continuously make decisions that shape their travel demand. These choices also occur at different time horizons, ranging from long-term commitments to short-term adjustments, and reflect personal preferences, constraints, and external conditions.

Long-term decisions define an individual's lifestyle and mobility habits. Choices such as the type of job, whether to live in a house or an apartment, and the number of children in the household influence daily routines and travel needs. Mobility-related decisions, such as purchasing a public transportation pass or owning one or more cars, determine the available travel options. Regular activities, including participation in sports, theater, or other hobbies, also shape recurring travel patterns, contributing to a stable framework of mobility demands.

Mid-term decisions focus on planning activities and travel over weeks or months. Scheduling work hours, social events, or leisure activities requires coordination, and these choices directly impact when and where travel occurs. Individuals must also decide on travel-related aspects, such as selecting a mode of transport, evaluating travel time, and balancing convenience and cost. These decisions may vary based on external factors such as weather conditions, temporary disruptions, or financial considerations.

Short-term choices involve adapting plans in response to immediate circumstances. Rescheduling activities due to unexpected constraints, such as last-minute work obligations or personal commitments, can lead to modifications in travel behavior. Additionally, individuals make use of real-time travel information, adjusting departure times or selecting alternative routes to optimize their trips. These short-term adaptations highlight the dynamic nature of travel decisions, as people continuously respond to changing conditions.

Travel demand also evolves over different time horizons. Some choices, like infrastructure development or residential location, are long-term and set the foundation for mobility patterns over decades. Other decisions, such as choosing a travel mode for a specific trip, are made on a short-term basis and can change frequently.

The different choices made by public authorities and individuals, as well as their interactions, are illustrated in Figure 5.1. This figure visually represents the role of decisions at different time horizons and how they shape travel demand.

On the left side of the figure, public authorities' decisions are enclosed within an orange frame. These decisions operate at three levels: long-term, mid-term, and short-term. On the right side, household and individual choices are enclosed within a blue frame. Similarly, these choices span different time horizons.

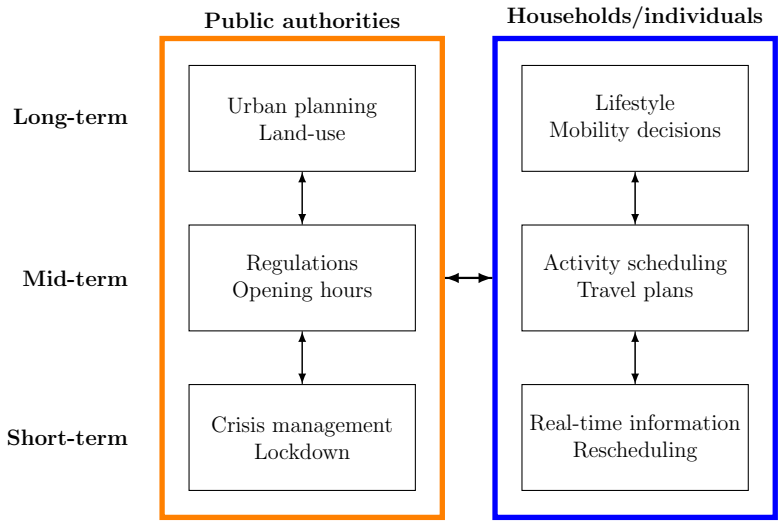


Figure 5.1: Choices and decisions

The figure also highlights the dynamic interactions between these two levels of decision-making. Thick arrows between the public authorities and households/individuals emphasize the reciprocal influence between government policies and personal choices. Policies set by public authorities directly impact individuals' travel behavior, while citizens' mobility needs and behaviors, in turn, shape policy responses. Additional arrows between different time horizons indicate the connections between long-term, mid-term, and short-term choices, illustrating how decisions made at one level influence the others.

Modeling travel demand requires representing the choices made by all actors involved, including households and individuals. Since every person makes decisions about their daily activities, the number of possible travel patterns becomes extraordinarily large, making the problem computationally complex.

An example helps illustrate this complexity. Consider a population of seven million individuals, each engaging in approximately ten activities per day. The possible ways in which these activities can be arranged over the course of a day quickly lead to an enormous number of potential scenarios. The sequencing of ten activities alone results in 10! possible orderings, already a significant number.

Beyond sequencing, each activity involves additional choices. Individuals must decide when to perform each activity, with multiple possible time slots available throughout the day. They must also choose a location, selecting

from a variety of destinations that offer the necessary services or opportunities. Furthermore, each activity requires a decision on the mode of transport, such as walking, cycling, driving, or using public transportation, and for each mode, there are multiple possible routes to reach the destination efficiently.

When all these factors are combined, the total number of possible travel plans becomes astronomically large. The example calculation shows that the number of possible combinations reaches the order of 10^{23}. This illustrates the immense complexity involved in modeling travel demand, as capturing every possible choice for every individual would be computationally infeasible.

Given this complexity, models must rely on simplifications, approximations, and behavioral principles to make the problem tractable.

The level of complexity required in a travel demand model depends on the specific question being addressed. Some situations require only basic information about people's movements, while others demand a detailed understanding of individual behavior and activity patterns. Selecting the appropriate level of complexity ensures that the model is both computationally efficient and capable of providing meaningful insights.

For simple systems like an elevator, there is no need to model the entire daily schedule of each individual. What matters is knowing when and where people press the button to request the elevator. The focus is on local interactions rather than the broader travel behavior of individuals. A similar level of complexity applies to metro systems, where the most critical information is where people board and alight. Just as an elevator moves people between floors, the metro transports passengers between stations, and the model can often abstract away the details of individual travel plans.

More complex systems, such as park-and-ride facilities, require additional considerations. Since travelers use multiple modes of transport in a single journey, it is necessary to account for car availability throughout the entire tour. A person driving to a park-and-ride lot must be able to retrieve their vehicle later, so the model must ensure consistency between different legs of the journey. This introduces interdependencies between choices, increasing the complexity of the modeling process.

At the highest level of complexity, understanding the impact of lockdown measures requires a detailed representation of individual activities. In this case, it is not enough to know where and when people travel; the model must capture the reasons for their trips and how restrictions on movement affect daily schedules. Since location is a critical factor in determining whether activities can still take place under lockdown conditions, the model must include a fine-grained representation of activity locations and their accessibility.

These examples highlight how the level of model complexity should be carefully adjusted to match the needs of the analysis. A simple model may be sufficient for studying metro ridership, but a detailed, activity-based model is essential for evaluating the effects of a major disruption like a lockdown.

The complexity of a travel demand model is determined by several key factors that define the level of detail included in the representation of travel behavior. Adjusting these factors allows for a balance between computational efficiency and the accuracy needed to answer specific research questions.

One of the main ingredients is granularity, which includes both time resolution and spatial discretization. A model with fine temporal resolution can capture rapid changes in travel behavior, such as rush hour dynamics or short-term disruptions, while a coarser resolution may be sufficient for long-term planning. Similarly, spatial discretization determines whether travel is represented at the level of entire cities, neighborhoods, or even individual intersections. A finer spatial scale provides more detailed insights but requires significantly more data and computation.

Another fundamental aspect is the level of aggregation. A model can be disaggregate, tracking the decisions of each individual separately, or aggregate, focusing on overall flows of travelers between different locations. Disaggregate models provide a richer representation of behavior but require more computational resources. Aggregate models, on the other hand, simplify the analysis by grouping individuals into categories or by modeling flows rather than individual movements. The choice between these approaches depends on the type of insights required and the available data.

The representation of travel patterns is also a important factor in defining model complexity. The most detailed models simulate full activity schedules, capturing how individuals organize their daily lives and how different activities influence their mobility choices. A more simplified approach focuses on tours, representing sequences of trips that begin and end at the same location (typically, home). The least complex models consider individual trips in isolation, without explicitly linking them to other activities. The choice between these representations affects how well the model captures dependencies between different trips and activities.

Figure 5.2 illustrates those three different ways of representing travel behavior, all based on the same underlying pattern of activities. In each case, the individual follows the same daily schedule, visiting home, work, shops, and dining locations at the same times and in the same sequence. What differs across the representations is how the interdependence between trips is considered.

The schedule-based representation on the left captures the full sequence of activities, preserving all dependencies between trips. It accounts for how

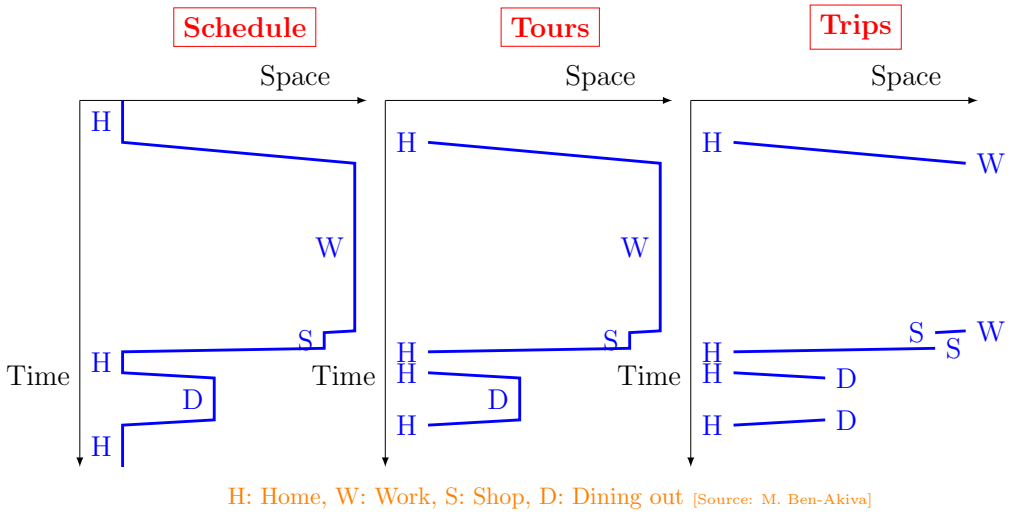


Figure 5.2: Trip-based, tour-based and schedule-based representations

earlier activities influence later ones, such as the need to return home before heading out again for dining. By modeling the complete schedule, this approach allows for a detailed understanding of how individuals structure their day and how their travel choices are interconnected.

In the middle, the tour-based approach simplifies the representation by grouping trips that begin and end at home into coherent sequences. Instead of modeling each trip separately, it recognizes that a person's journey to work and their stop at a shop on the way back are linked as part of a larger tour. While this approach reduces complexity, it still maintains important relationships between trips and captures dependencies within each tour.

On the right, the trip-based representation treats each trip in isolation, ignoring how they relate to one another. A trip from home to work is modeled separately from a shopping trip later in the day, even if both trips are made by the same individual as part of an overarching schedule. This approach is computationally simpler but overlooks behavioral connections, such as how a person's decision to shop may depend on their prior work schedule or travel constraints.

All three representations describe the exact same travel behavior, but at different levels of abstraction. The choice of representation depends on the research question: a full schedule is necessary to capture the interdependence of activities, a tour-based approach provides a structured yet simplified alternative, and a trip-based approach is useful when focusing only on individual movements without considering broader patterns.

5.1 Production and attraction

In order to illustrate the various steps of travel modeling, we consider an example where we wan to analyze how crowded an elevator becomes during peak hours. Unlike large-scale transportation systems, where individual differences and detailed activity schedules may be important, the focus in this case is on understanding the overall flow of people using the elevator.

Since the objective is to estimate congestion in the elevators, the granularity of the model can be relatively coarse. Instead of tracking movements throughout the entire day, the analysis can focus on the morning peak hour when most people arrive at work or school. Similarly, spatial resolution does not require mapping entire building layouts in fine detail; a simple list of floors, representing where passengers enter and exit the elevator, is sufficient.

To keep the model efficient, an aggregate approach can be used. Rather than distinguishing between individual travelers, the model considers total flows of people moving between floors. This means that all elevator users are treated as identical, without accounting for variations in their personal schedules or specific behaviors. This level of simplification is appropriate because the primary interest is in estimating the overall demand for elevator capacity, not in modeling individual preferences.

Finally, travel patterns are represented at the trip level, where each elevator ride is treated as an independent movement from one floor to another. There is no need to track entire activity schedules or link trips into tours, as in more complex transportation models. This simplification is justified because an elevator journey is typically a short, direct trip with no intermediate decisions, making it reasonable to ignore dependencies between trips.

Before analyzing elevator congestion, it is essential to understand how many trips are generated and where they are directed. This requires addressing two fundamental aspects: trip production and trip attraction.

Trip production refers to where people start their journeys. In the morning peak hour, most trips originate from the floors where people live. Residents leave their apartments to go to work, school, or other activities, and the elevator becomes their primary means of reaching the ground floor. An important question is determining how many people will leave during this peak period, as this directly impacts elevator demand.

Trip attraction, on the other hand, focuses on where these trips are headed. In this case, the ground floor plays a crucial role, as it serves as the main exit point from the building. However, not all trips necessarily end at the ground floor. In buildings with mixed uses, some people may travel between floors to reach workplaces, offices, or shared spaces within the same structure.

The starting point of travel demand analysis is census data, which provides fundamental demographic information about the population. This data serves as the foundation for understanding how, when, and why people travel by capturing key characteristics such as household locations, job locations, and socioeconomic factors.

Census data records where individuals live, which helps identify residential areas and estimate the number of people who may need to travel daily. It also includes information about workplaces, enabling an understanding of commuting patterns and employment hubs. By linking home and job locations, analysts can infer major flows of travelers and anticipate demand for transportation infrastructure.

Going back to our elevator example, Table 5.1 presents synthetic demographic data describing the distribution of residents and workers across different floors of a building. This data provides essential input for understanding elevator demand, as it helps estimate the number of people who will use the elevators during peak hours.

Floor	Residents	Workers
0	0	0
1	12	2
2	5	70
3	17	5
4	20	0

Table 5.1: Demographic data for the elevator example

Each row represents a floor, while the two columns provide information on the number of residents and the number of workers located on that floor. The ground floor (floor 0) does not have any residents or workers, indicating that it likely serves as an entrance and exit point rather than a living or working space. Floors 1, 3, and 4 primarily accommodate residents, with 12, 17, and 20 people living on these floors, respectively. These individuals are likely to leave their apartments in the morning, contributing to elevator trips toward the ground floor.

In contrast, floor 2 is primarily a workplace, with 70 workers but only 5 residents. This suggests that a significant number of people will use the elevator to reach this floor in the morning, creating an inbound flow distinct from the residential floors where most trips originate. Floor 1 also hosts a small number of workers, meaning that some elevator trips may be directed to this level as well. Floor 3 has a minor mix of both residents and workers, while floor 4 is entirely residential.

In travel demand analysis, the concepts of *trip production* and *trip attraction* describe the movement of individuals within a system. Trip production refers to the number of trips that begin at a given location, while trip attraction corresponds to the number of trips that end at that location. Since every trip has both a starting point and a destination, the total number of trips produced must always be equal to the total number of trips attracted.

Table 5.2 presents synthetic data on trip production and attraction within a building during the morning peak hour. The first three columns describe the number of residents and workers per floor, while the last two columns indicate how many trips are produced and attracted on each floor.

Floor	Residents	Workers	Prod.	Attr.
0	0	0	51	26
1	12	2	10	0
2	5	70	3	51
3	17	5	7	2
4	20	0	8	0
			79	79

Table 5.2: Production and attraction for the elevator example

The total number of trips produced across all floors is 79, which matches the total number of trips attracted. However, at first glance, some apparent inconsistencies emerge. For instance, the total number of trips attracted to the ground floor is 26, but the sum of trips produced by the other floors does not exactly match this number ($10 + 3 + 7 + 8 \neq 26$). This discrepancy arises because some residents do not travel directly to the ground floor. For instance, a person working in an office located on another floor may take the elevator from their apartment to their workplace without ever reaching the ground floor. These internal trips within the building create minor imbalances when considering only floor-level statistics.

Estimating trip production and attraction requires data from multiple sources. Census records and surveys provide information about where people live and work, while direct observations, such as counts taken at the ground floor, help validate estimated flows.

5.2 Origin-destination tables

An *origin-destination (OD) table* is a structured representation of trips that captures both where they begin and where they end. Unlike simple trip

production and attraction counts, which summarize how many trips start or end at each location, an OD table distributes these trips across specific destinations and origins, providing a detailed picture of travel patterns.

Table 5.3 presents the OD table for the elevator example, showing how trips are distributed among floors during the morning peak hour. Each row represents an origin floor, while each column corresponds to a destination floor. The values in the table indicate how many trips are made from each origin to each destination. The rightmost column sums the total trips produced at each floor, while the bottom row sums the total trips attracted to each floor. Since every trip has both a starting point and an endpoint, the total number of trips produced (79) equals the total number of trips attracted.

	0	1	2	3	4	
0		0	50	1	0	51
1	10		0	0	0	10
2	3	0		0	0	3
3	6	0	1		0	7
4	7	0	0	1		8
	26	0	51	2	0	79

Table 5.3: Origin-destination table for the elevator example

This table illustrates how the trips produced at a given floor are not directed toward a single destination but are distributed across multiple floors. For example, on floor 3, seven trips are produced: six people travel to the ground floor, and one person travels to floor 2. Similarly, the trips attracted to a particular floor come from different origins. Floor 2 attracts 51 trips, primarily from the ground floor but also from floor 3. These distributions are what distinguish an OD table from simple trip production and attraction totals, as they reveal how movement is structured within the system.

Beyond the elevator example, OD tables are widely used in transportation planning. They play an important role in modeling commuting patterns, assessing transit demand, and designing infrastructure improvements.

It is important to realize that constructing an OD table from trip production and attraction data does not necessarily lead to a unique solution. Table 5.4a and Table 5.4b illustrate this concept. Both OD tables correspond to the same production and attraction totals, shown in the last row and last column of each table. Each row represents an origin floor, and each column represents a destination floor. The values indicate how many trips travel between each pair of floors.

	0	1	2	3	4	
0		0	50	1	0	51
1	10		0	0	0	10
2	3	0		0	0	3
3	6	0	1		0	7
4	7	0	0	1		8
	26	0	51	2	0	79

(a) Table 1

	0	1	2	3	4	
0		0	49	2	0	51
1	9		1	0	0	10
2	3	0		0	0	3
3	7	0	0		0	7
4	7	0	1	0		8
	26	0	51	2	0	79

(b) Table 2

Table 5.4: Two OD tables with the same production and attraction

5.2.1 Under-determination

While the total number of trips originating from and arriving at each floor remains the same in both tables, the way trips are allocated among destinations differs. For example, in Table 5.4a, 50 trips travel from the ground floor to floor 2, while in Table 5.4b, this number is reduced to 49, with an additional trip instead going to floor 3. Similarly, the distribution of trips originating from floors 1, 3, and 4 also varies slightly between the two tables.

These differences highlight the fact that production and attraction data alone do not determine a unique OD table. This is due to a fundamental mathematical issue: the problem is *under-determined*. The number of unknowns in an OD matrix grows quadratically with the number of locations, while the number of available constraints from production and attraction totals grows only linearly. Specifically, for a system with M locations, the number of unknown OD flows is $\mathsf{M}^2 - \mathsf{M}$ (excluding trips from a location to itself), while the number of equations provided by production and attraction data is only $2\mathsf{M}$.

This discrepancy quickly becomes significant. For $\mathsf{M} = 5$, there are 20 unknown OD flows but only 10 equations. When $\mathsf{M} = 40$, the number of unknowns explodes to 1560, while the number of constraints remains just 80. Since the number of unknowns grows much faster than the number of available constraints, there are infinitely many OD tables that satisfy the same production and attraction data.

To resolve this under-determination problem, additional information is required. There are two main approaches to achieving this:

One approach is to introduce more *theoretical assumptions* about trip distribution. For example, one may assume that trips are distributed in a way that minimizes overall travel cost, or that people tend to travel more frequently to nearby destinations than distant ones. Such assumptions allow us to impose additional constraints on the problem, reducing the number of

feasible solutions.

The second approach is to collect more *data*. Direct observations, such as traffic counts at intermediate locations, surveys, or smart card data in public transport systems, provide additional equations that can help uniquely determine the OD table. The more data that is available, the more accurately we can infer the actual distribution of trips.

In practice, a combination of both methods is often used. Theoretical assumptions provide a framework for estimating missing values, while real-world data helps refine and validate these estimates. This balance between assumptions and observations is crucial for generating OD tables that accurately reflect travel behavior and can be used for transportation planning and policy decisions.

The elevator example demonstrates this challenge in a controlled environment, but the same issue arises in large-scale urban transportation models. Understanding that multiple OD tables can satisfy the same production and attraction constraints is essential for interpreting travel demand data correctly and making informed planning decisions.

5.2.2 Incompatibility

In constructing an *origin-destination (OD) table*, an important condition must hold: the total number of trips produced must equal the total number of trips attracted. If this balance is not satisfied, then no OD table can be constructed that satisfies the given constraints, meaning that the problem has no solution. This situation arises when production and attraction data are inconsistent, which may occur due to data collection errors, survey biases, or missing observations.

Table 5.5 illustrates an example of such an inconsistency. The last column represents the total number of trips produced at each origin, while the last row represents the total number of trips attracted to each destination. However, the sum of trips produced is $8 + 3 = 11$, while the sum of trips attracted is $1 + 9 = 10$. Since these totals do not match, it is mathematically impossible to construct a valid OD table that satisfies both constraints simultaneously.

t_{11}	t_{12}	8
t_{21}	t_{22}	3
1	9	

Table 5.5: Example of incompatible production and attraction data

Such inconsistencies pose a significant challenge in travel demand modeling. If left uncorrected, they can lead to errors in transportation planning and decision-making. To address this issue, one possible solution is to treat the trip values as *random variables* rather than fixed numbers. Instead of requiring exact equality between trip productions and attractions, a probabilistic approach allows small discrepancies while maintaining an overall balance through statistical adjustments.

In real-world applications, production and attraction data are often estimated from different sources, such as household travel surveys, traffic counts, and census data. Discrepancies may arise due to incomplete data collection or errors in estimation.

5.3 Mode choice

Once the *origin-destination (OD) table* is established, it must be further refined by distributing trips across different modes of transportation. Not all individuals use the same travel mode to complete their trips. In the elevator example, some people may choose to take the stairs instead of using the elevator, depending on factors such as the number of floors they need to travel or the level of congestion in the elevators.

Tables 5.6a and 5.6b illustrate this division by mode. The total number of trips remains the same, but they are now allocated to two different transportation options: stairs and elevators.

	0	1	2	3	4	
0	0	0	20	0	0	20
1	7	0	0	0	0	7
2	1	0	0	0	0	1
3	0	0	0	0	0	0
4	0	0	0	0	0	0
	8	0	20	0	0	28

(a) Trips taken using the stairs

	0	1	2	3	4	
0	0	0	30	1	0	31
1	3	0	0	0	0	3
2	2	0	0	0	0	2
3	6	0	1	0	0	7
4	7	0	0	1	0	8
	18	0	31	2	0	51

(b) Trips taken using the elevator

Table 5.6: Splitting OD trips by transportation mode

The decision to take the stairs or the elevator depends on multiple factors. One key determinant is the number of floors a person needs to travel. Shorter trips, particularly those spanning only one or two floors, are more likely to be completed via stairs, while longer trips are more commonly taken using the elevator. Another important factor is congestion: if the elevator is overcrowded or slow, individuals may opt to take the stairs instead.

Table 5.6a shows that most stair users are traveling to or from lower floors, as shorter trips are more practical to complete on foot. For instance, 20 trips from the ground floor to floor 2 are taken using the stairs, while no trips are recorded beyond floor 2. Meanwhile, Table 5.6b includes longer trips that span multiple floors, such as those from the ground floor to floors 3 and 4.

5.3.1 Route choice

When multiple elevators are available, travelers face an additional level of decision-making: which elevator to take. This *route choice* problem arises because not all elevators provide the same experience. Some people may prefer a specific elevator due to its location, expected waiting time, or perceived speed, while others may simply take the first elevator that arrives. However, elevators can also be full, forcing individuals to wait for the next available option.

This choice process creates a feedback loop between individual decisions and system-wide conditions. The number of people choosing a particular elevator influences how crowded it becomes, while the level of crowdedness in turn affects future choices. If an elevator is consistently overused and reaches capacity quickly, travelers may start choosing alternative elevators, shifting the demand distribution across the system.

The concept of *equilibrium*, as discussed in Section 2.1. As discussed later, this type of equilibrium is analogous to route choice in transportation networks. Just as commuters adjust their travel routes based on congestion levels, elevator users adjust their choices based on real-time conditions. The equilibrium that emerges depends on several factors, including the frequency of elevator arrivals, the capacity of each elevator, and individual preferences regarding waiting time versus comfort.

5.3.2 The four step approach

The concepts introduced in the elevator example — such as production, attraction, OD tables, and mode choice — are directly applicable to the analysis of public transportation systems. In a city like Lausanne, the same fundamental principles can be used to understand how people travel across the network and how transportation demand can be modeled and managed.

A clear analogy can be drawn between floors in a building and stops in a public transportation network. Just as individuals travel between floors using elevators or stairs, public transport users travel between bus stops and metro stations using different modes of transportation. The ground floor in

the elevator example serves as a central access point, much like train stations function as key transfer hubs in Lausanne"s transportation system.

The concepts of *trip production* and *trip attraction* also remain relevant. In the case of public transportation, production and attraction zones correspond to different areas of the city where people live and work. Just as individuals leaving their apartments in the morning generate trips to the ground floor, residents of suburban areas generate trips to transit stops as they begin their daily commutes. Similarly, attraction zones represent workplaces, schools, or commercial centers where people are traveling. Trip counts at bus stops and metro stations serve as real-world equivalents to the production and attraction data observed in the elevator example.

When constructing an *origin-destination (OD) table* for a public transport network, we face the same challenges of under-determination and potential inconsistencies. Without additional data, multiple OD tables can satisfy the same aggregate production and attraction totals, making it necessary to use additional observations or behavioral assumptions to infer the most likely travel patterns. Incompatibilities can also arise when data sources are incomplete or inconsistent, requiring adjustments to balance the total number of trips.

Mode choice in a public transport system mirrors the decision of whether to take the elevator or the stairs. Some travelers may choose public transportation, while others opt for private cars, cycling, or walking, depending on factors such as travel time, cost, and convenience. Understanding these mode choices is essential for transportation planning, as it helps predict how policy changes or infrastructure investments might influence traveler behavior.

Finally, just as travelers in a building must decide which elevator to take, public transport users must choose an itinerary from multiple available routes. This *route choice* process determines congestion levels on different transit lines and helps planners anticipate demand for specific services. The concept of equilibrium, which describes how individuals adjust their decisions based on system-wide conditions, applies to both cases: when one route becomes too crowded, travelers naturally seek alternatives, balancing demand across the network.

Motivated by this analogy, we now introduce the *four-step approach*, a widely used framework for modeling transportation systems. This approach generalizes the decisions illustrated in the elevator example — trip generation, trip distribution, mode choice, and route choice — to urban transport networks.

5.4 Introduction to the four step model

Before introducing the four-step model, it is important to define the fundamental assumptions regarding time and space, as these determine how trips are categorized and analyzed. These assumptions help simplify the representation of travel behavior while maintaining a meaningful level of detail for transportation planning.

Time assumptions specify the interval of interest over which trips are considered. Rather than analyzing all trips throughout an entire day, the focus is typically on a specific period that captures critical travel patterns. A common example is the *morning peak hour*, during which travel demand is highest due to commuting to work and school. The analysis includes all trips that start and end within this time window, ensuring that the most relevant travel activity is captured.

Space assumptions define the geographical scope of the study. The study area is partitioned into zones that serve as units of analysis, simplifying travel behavior representation. These zones are often based on statistical regions or census units, reflecting meaningful divisions of urban space. Within each zone, individual travel movements are not explicitly considered; instead, trips are only modeled when they cross from one zone to another. This aggregation reduces the complexity of the model while still preserving key travel dynamics at a regional level.

These time and space definitions provide the foundation for the four-step approach, ensuring that the model captures travel demand patterns in a way that is both manageable and useful for transportation planning.

The four-step model breaks down travel behavior into four distinct stages, each corresponding to a decision that travelers make when planning their trips. These steps ensure a structured approach to analyzing and forecasting mobility patterns.

The first step, *trip generation*, determines the number of trips that will be made by individuals or households within a given time period. This step is closely related to activity choices, as trips are made to participate in activities such as work, school, shopping, or leisure. Additionally, trip frequency depends on factors like household composition, employment status, and personal preferences.

The second step, *trip distribution*, identifies where trips will be made by assigning each generated trip to a destination. This reflects the traveler's choice of activity location. The likelihood of selecting a particular destination depends on factors such as distance, travel time, accessibility, and the availability of desired activities at different locations.

The third step, *modal split*, determines which mode of transportation

travelers will use for their trips. This step captures the decision-making process between options such as private cars, public transportation, walking, or cycling. The choice of transportation mode depends on variables such as cost, travel time, convenience, and individual preferences.

The fourth and final step, *assignment*, allocates trips to specific routes and transportation networks. This step corresponds to the traveler's route or itinerary choice, as individuals seek the most efficient or preferred path to their destination. Route assignment considers factors such as congestion, transit schedules, road conditions, and travel time reliability.

5.4.1 Trip generation

Figure 5.3 illustrates the concept of trip purposes by depicting a series of trips undertaken by an individual over a given period (this is the same example as in Figure 5.2). The horizontal axis represents space, while the vertical axis represents time. Each segment in the diagram corresponds to a specific trip between two locations, with labels indicating the purpose of the trip.

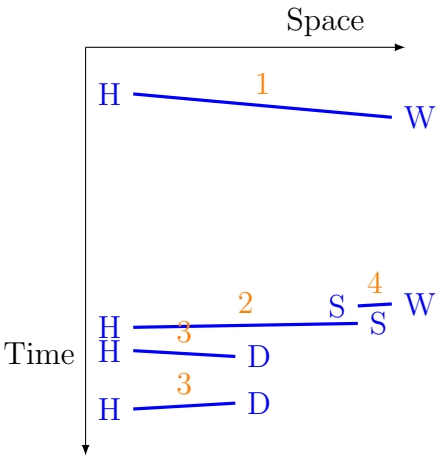


Figure 5.3: Trip purposes

The starting and ending points of the trips are marked with different locations: H (Home), W (Work), S (Shop), and D (Other Destination). The numbered segments indicate different types of trip purposes. The first trip (1) is from home to work, representing a home-based work trip. The second trip (2) is from home to a shopping location, categorized as a home-based shopping trip. The third trip (3) is from home to another type of destination, such as a recreational or social activity, labeled as a home-based other

trip. The fourth trip (4) is from one non-home location to another, such as traveling from work to a store, which is classified as a non-home-based trip.

Trip purposes structure daily mobility patterns. Different types of trips contribute to overall travel behavior. Therefore, different trip production and trip attraction models are usually defined for each trip purpose.

Trip generation models aim to estimate the number of trips produced and attracted by different zones in a study area. These models are typically based on linear regression, where the number of trips is expressed as a function of explanatory variables.

The trip production model predicts the number of trips originating from a given zone r, denoted as O_r. The dependent variable in this model is the number of trips leaving the zone. The independent variables capture characteristics that influence an individual's need to travel. These include

- individual and household characteristics such as age, income, or family size,
- mobility tools, such as car ownership or public transport subscriptions,
- characteristics of the home location, such as urban density or housing type,
- accessibility, characterizing how easily people can reach opportunities and services, affecting their likelihood of making trips.

The trip attraction model estimates the number of trips arriving at a given zone r, represented by D_r. The dependent variable in this case is the number of trips reaching the zone. The independent variables reflect the factors that make a location attractive for different activities. They include

- land use characteristics, such as the presence of industrial, commercial, or service areas,
- employment levels, as workplaces generate commuter traffic,
- accessibility, characterizing how easily people can reach the zone.

As an example, the Swiss model (Danalet et al., 2021) incorporates a comprehensive set of variables to predict trip production and attraction. The model considers various aspects of individuals, households, mobility tools, home location characteristics, and accessibility to provide accurate estimations of travel behavior.

Individual characteristics include the level of education, sex, age, and nationality, which can influence daily activity choices. Work-related factors

such as whether a person does some home office, their function in the company, work percentage, and business sector also affect travel demand. Additionally, studying status and language spoken may impact trip frequency and destinations.

Household characteristics are equally important, as family composition shapes mobility behavior. The model considers household structure and the number of children across different age groups (0-6, 6-15, and 15+). Specific categories such as couples without children under the age of 30 or between 30 and 49 help refine the understanding of trip production patterns. Household income is also a key factor, as financial resources influence travel options and mode choices.

Mobility tools capture the modes available for travel: car availability, number of cars in the household, ownership of public transport travel cards, such as the GA card, are included in the model.

Home location characteristics influence trip production by defining the context in which individuals live. The urban-rural typology categorizes locations as urban, intermediate, or rural, affecting accessibility and travel choices. The region of the place of living also plays a role, as different areas may have varying infrastructure, economic activities, and public transport services.

Accessibility measures complete the model by assessing how easily individuals can reach opportunities and services. The quality of public transport connections at the place of living impacts travel decisions, while the home-work crow-fly distance provides an estimate of commuting needs.

Before moving to the next step of trip distribution, it is essential to introduce the concept of transportation networks. Indeed, they define how trips can be made and how travel costs, such as time and distance, vary between different locations. The structure and performance of the network directly affect the choices travelers make when selecting destinations, modes of transport, and routes.

Chapter 6

Transportation networks

Each mode of transportation operates within a specific network that defines how trips can be made. These networks consist of interconnected links[1] and nodes that facilitate movement and determine the accessibility of different locations.

The road network consists of streets, highways, and intersections that enable vehicle travel. It is characterized by road hierarchies, traffic regulations, and congestion levels that influence travel times and route choices. Figure 6.1 represents the network representation of the city center of Lausanne as provided by Google Maps.

Figure 6.1: Road network

Public transportation networks in urban areas include bus, tram, and metro systems. These networks are structured around stops and lines, ensuring connectivity between residential, commercial, and service areas. Figure 6.2 represents the public transportation network operated by the main operator in Lausanne, TL (Transports publics de la région lausannoise).

[1] In transportation, the terms "link" and "arcs" are synonyms, and are used interchangeably throughout this document.

Figure 6.2: Urban public transportation network

Railway networks support intercity and regional travel, offering high-capacity, efficient, and sustainable mobility. Stations serve as key nodes, and rail lines connect cities and economic hubs. Figure 6.3 represents the network of Inter-City trains operated by the Swiss Railways.

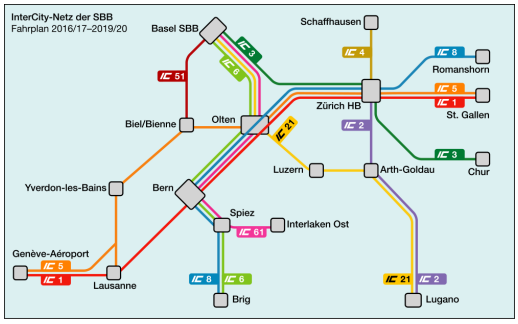


Figure 6.3: Railway network

Waterborne public transportation networks enable travel movement over water. They provide alternative mobility solutions, such as the ferry services on the Lake of Geneva, as represented by Figure 6.4.

Airline transportation networks connect cities and countries through airports, facilitating long-distance travel. These networks rely on hub-and-spoke models, where major airports act as central hubs. Figure 6.5 represents a portion of the network for the flights of Swiss International Air Lines (SWISS), involving flights from Geneva and Zürich airports.

Maritime freight transportation networks handle the global movement of goods using shipping routes and container terminals. These networks are essential for international trade and logistics. Figure 6.6 represents a portion of the network of maritime shipping routes operated by MSC Mediterranean Shipping Company.

Pedestrian walkway networks provide infrastructure for walking, ensuring

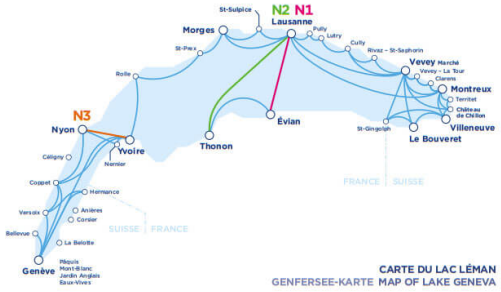


Figure 6.4: Maritime transportation network (public transportation on Lake Léman)

Figure 6.5: Airline transportation network

safe and accessible paths in urban and natural environments. They play a key role in sustainable mobility and active transport. In the context of leisure activites, Figure 6.7 represents the network of walkways for hiking in the mountains around Zinal.

Ski slope networks enable winter sports by connecting ski areas with lifts and designated trails. These networks ensure accessibility to different levels of difficulty and facilitate movement across resorts. Figure 6.8 represents the network of ski slopes in the "4 vallées" site.

A transportation network can be mathematically represented as a directed graph consisting of a set of nodes $\mathcal{N}$ and a set of arcs $\mathcal{A}$. Nodes represent locations, such as intersections, stations, or zones, while arcs define the possible connections between these locations, representing roads, railway tracks, or transit lines.

The network topology is characterized by an incidence function $\phi : \mathcal{A} \to \mathcal{N} \times \mathcal{N}$, which maps an arc to an ordered pair of nodes, indicating its direction. This function defines which nodes are connected and in which direction travel is permitted.

Additionally, networks often include quantities defined on nodes and arcs. Node-related quantities can represent aspects such as demand, capacity, or

Figure 6.6: Maritime transportation network (routes for ships with container terminals)

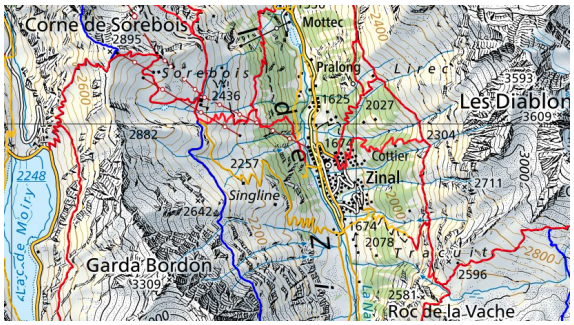


Figure 6.7: Pedestrian walkway network

supply at a given location. Arc-related quantities may include travel costs, distances, or flow values.

Figure 6.9 illustrates this abstract representation. The network consists of four nodes, labeled o, a, b, and c, with associated values indicating supply/demand data associated with the nodes. The arcs between the nodes are oriented, with numerical values representing attributes such as travel cost or distance. The incidence function determines the connectivity structure, specifying which nodes are linked and in which direction.

6.1 Road networks

In the specific case of road networks, nodes are usually geo-coded, meaning they are assigned precise geographical coordinates that correspond to real-world locations. Two main types of nodes exist in this representation. Centroids, shown in blue in Figure 6.10, are associated with geographical zones and serve as the interface between demand models and the network representation. Intersection nodes, represented in orange, correspond to locations where roads meet, such as intersections, merging points, or areas

Figure 6.8: Ski slopes network

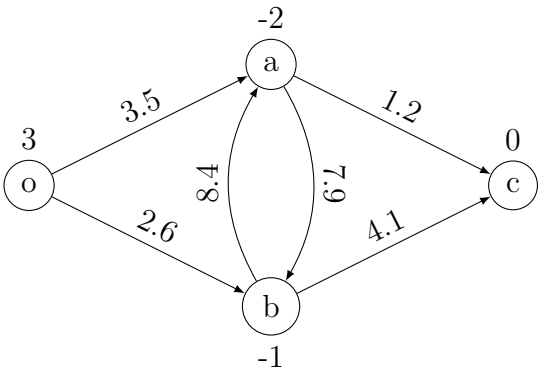


Figure 6.9: Mathematical representation of a network

where road capacity changes.

Links represent the connections between nodes and define how movement occurs within the network. Two types of links exist in the road network representation. Centroid connectors, shown in orange in Figure 6.1, link centroids to the network and are used to model access points where trips enter or exit the system. The other type consists of homogeneous road segments, which represent sections of road with uniform characteristics such as speed limits, number of lanes, or capacity.

Modeling intersections in road networks involves a trade-off between *level of detail* and *computational complexity*. Figure 6.11 presents two different approaches: a *simplified model* (left) and a *detailed model* (right).

In the *simplified model* (Figure 6.11a), the entire intersection is represented by a single node. This abstraction is computationally efficient and easy to implement, making it particularly useful for large-scale simulations where the primary focus is on network-wide traffic flow. However, this representation does not distinguish between different movements within the in-

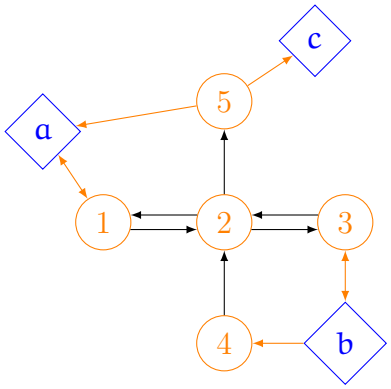
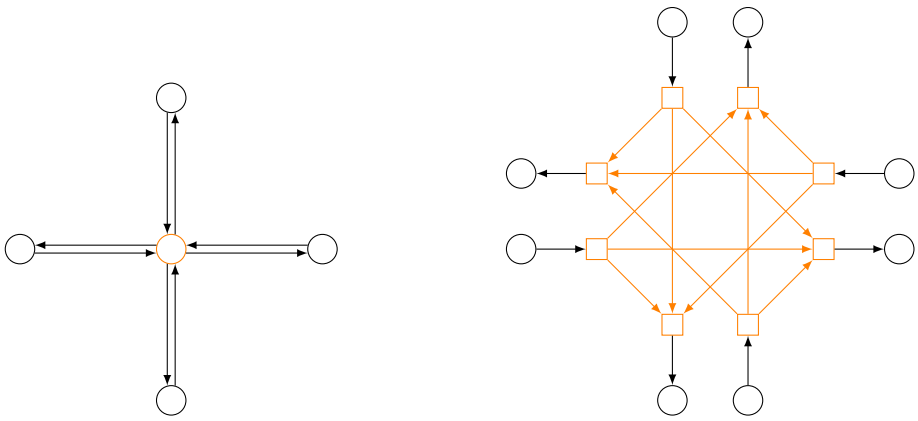


Figure 6.10: Representation of a road network



(a) Simplified intersection model

(b) Detailed intersection model

Figure 6.11: Comparison of simple and detailed intersection models

tersection. As a result, it cannot account for *movement-specific restrictions*, such as special traffic light phases, dedicated lanes for turning, or prohibited maneuvers. All vehicles passing through the intersection are treated uniformly, limiting the model's ability to reflect realistic traffic conditions.

In contrast, the *detailed model* (Figure 6.11b) provides a finer representation of intersection movements by associating each possible maneuver with a dedicated link. This allows for a much more precise representation of traffic control and infrastructure constraints. For instance, specific movements can be *blocked*, reflecting turn restrictions or road closures. Traffic lights can be assigned different *signal cycles* depending on the movement, improving the accuracy of delay estimations. Moreover, *dedicated lanes* for certain directions can be modeled explicitly, allowing for a more realistic simulation of

traffic behavior at the intersection.

While the detailed model provides a richer and more flexible representation, it comes at the cost of increased *computational complexity*. The number of nodes and links in the network grows significantly, requiring more memory and processing power. This makes it less suitable for large-scale applications but ideal for studies focusing on localized traffic operations or traffic signal optimization.

We illustrate another example in the context of modeling highways. One key consideration is whether variations in road capacity should be explicitly represented in the model. Figure 6.1 illustrates a highway section where the number of lanes changes: it starts with three lanes, narrows to two lanes in the middle, and then expands back to three lanes.

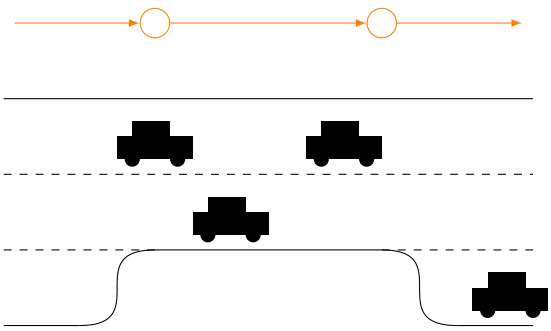


Table 6.1: Highway segment with capacity changes

If the entire stretch of highway is represented as a *single link* in the network model, it would not be possible to differentiate between sections with different capacities. This simplification may be sufficient for certain macroscopic analyses, but it can lead to inaccuracies when modeling congestion, lane restrictions, or capacity reductions.

To account for these variations in capacity, it is necessary to insert *nodes* at points where the number of lanes changes. In the example shown in Figure 6.1, two additional nodes are introduced where the transition occurs. As a result, the highway segment is now represented by *three distinct links*, each associated with a specific capacity.

This approach increases the model's accuracy but also introduces additional complexity. More nodes and links mean a larger network, which can impact computation time and memory usage, especially in large-scale simulations. The choice between a simplified and a detailed representation depends here as well on the requirements of the analysis. If congestion effects or traffic flow variations due to capacity changes are important, a more detailed representation is necessary. On the other hand, if the study focuses on broader

travel patterns where local bottlenecks are less relevant, a simpler approach may be sufficient.

Road networks associate relevant data with both *nodes* and *links*. The primary data associated with nodes include:

- *In-flow*: The number of vehicles entering the network at a given node over a specific period.
- *Out-flow*: The number of vehicles leaving the network from a node.

These values help define travel demand and traffic generation patterns within a network. They are typically derived from travel demand models, real-world traffic counts, or simulations. The most important data associated with links include:

- *Flow* and *capacity* (measured in vehicles per minute), which indicate how much traffic the link currently carries compared to its maximum potential.
- *Travel time* and *free-flow travel time* (measured in minutes), capturing both the actual travel conditions and the ideal conditions when there is no congestion.
- *Travel cost* and *toll* (in CHF or €), representing monetary costs associated with using a specific road segment.
- *Density* and *jam density* (measured in vehicles per kilometer), which quantify how closely vehicles are spaced on the link and the level of congestion.

Another quantity associated with the link is a *link performance function*. A link performance function maps the traffic flow to the corresponding travel time. There are two main types. One *without an upper limit*, which is defined for all values of the traffic flow. This is convenient in algorithmic applications, where evaluations may occur at unknown levels of demand. One *with an upper limit*, which is more realistic since roads have a physical capacity. However, it is not defined for flows beyond this capacity.

A commonly used function *without an upper limit* is the Bureau of Public Roads (BPR) (Bureau of Public Roads, 1964), given by:

$$t(x) = t_0 \left(1 + \alpha \left(\frac{x}{\ell}\right)^{\beta}\right) \tag{6.1}$$

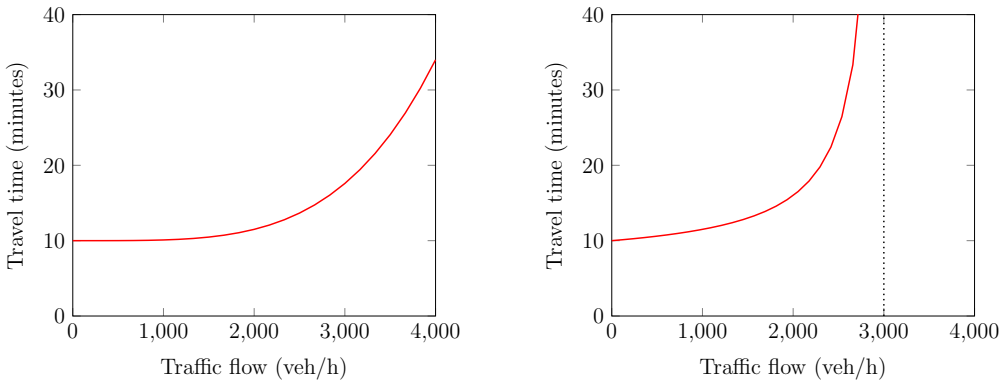
where t_0 is the free-flow travel time, x is the traffic flow, ℓ is the reference capacity, and α, β are calibration parameters. The advantage of this function

is its smoothness and differentiability, which make it useful in equilibrium traffic assignment problems.

In contrast, a function *with an upper limit* accounts for the physical constraints of the road. A simple example is Davidson's function (Davidson, 1966, Akçelik, 1991):

$$t(x) = t_0 \left(1 + \alpha \left(\frac{x}{\ell - x}\right)\right) \tag{6.2}$$

where travel time approaches infinity as flow nears capacity ($x \to \ell$). This formulation reflects congestion effects more realistically and is used in dynamic traffic models.



(a) Link performance function *without* an upper limit

(b) Link performance function *with* an upper limit

Figure 6.12: Comparison of link performance functions

6.2 Public transportation networks

A public transportation network can also be represented as a network, where *nodes* and *links* capture key elements of the system. Nodes in the network are *geo-coded* and can be classified into two main types:

Centroids (in blue in Figure 6.13): These are associated with geographical zones and serve as origins and destinations for demand.

Stations or stops (in orange in Figure 6.13): These are locations where passengers board, alight, or transfer between services.

Links in the network represent movements between nodes and can take different forms:

Line segments These correspond to segments of a transit route where vehicles travel between stops.

Walking links (W in Figure 6.13): These represent pedestrian movement between locations.

Transfer links (T in Figure 6.13): These capture the movement required to switch from one transit line to another.

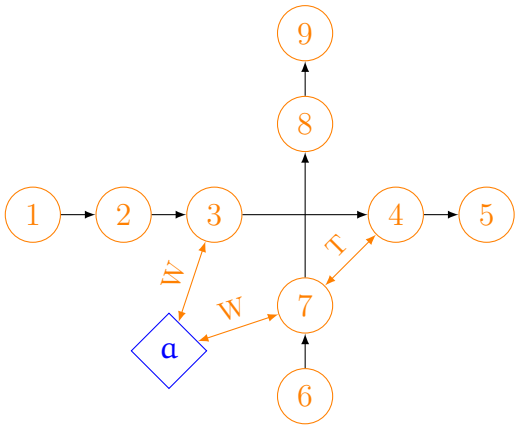


Figure 6.13: Representation of a public transportation network

Each component of the network carries data for operations and planning:

Nodes The key attribute is demand, measured as *in-flow* (arrivals) and *out-flow* (departures).

Links The following attributes are relevant:

Frequency (vehicles per hour) and *headway* (minutes between consecutive departures).

Travel time (minutes) between stops.

Walking time (minutes) for pedestrian links.

Waiting time (minutes) at stations.

Transfer time (minutes) when switching between lines.

Capacity representing the number of available seats per vehicle.

The above representation of a public transportation network captures the spatial structure of the system but does not explicitly consider the *schedule*, or timetable, of public transport services. In reality, public transportation

operates according to a predefined schedule, meaning that transit services are not only defined in terms of locations but also in terms of their departure and arrival times. To account for this, a *scheduled public transportation network* must be introduced.

In a scheduled representation, nodes and links are extended into the *time dimension*. The network structure consists of:

Nodes These now represent both *space and time*, rather than just locations. Each stop is replicated at different time instances to account for scheduled departures and arrivals. In addition to centroids and stations, the nodes are linked to a *timetable* that determines when a vehicle departs or arrives at a stop.

Links Similar to the previous representation, but now incorporating *departure time choices*. These links define not only movement between locations but also movement across time. A link connects a departure event at a stop to the arrival event at the next stop at a specific scheduled time.

This concept is illustrated in Figure 6.14, where:

The x-axis represents *time*. Each vertical dotted line corresponds to a specific time instant, such as 14:20, 14:30, etc.

The y-axis represents *space*. Each horizontal dotted line corresponds to a specific location, such as different stations along a transit route.

In this diagram, the orange and green paths correspond to different scheduled transit services. Each dot represents a specific transit event (e.g., a bus arriving or departing from a station at a given time). The arrows indicate the movement of transit vehicles, following the scheduled departures and arrivals at each stop.

The data structure of a scheduled public transportation network extends the attributes of the basic representation:

Nodes Capture *demand* as in-flow and out-flow, but now at specific time points.

Links Contain additional attributes that define transit operations:

- **Travel time** (min) between stops at specific scheduled intervals.
- **Walking time** (min) for pedestrian transfers between scheduled services.

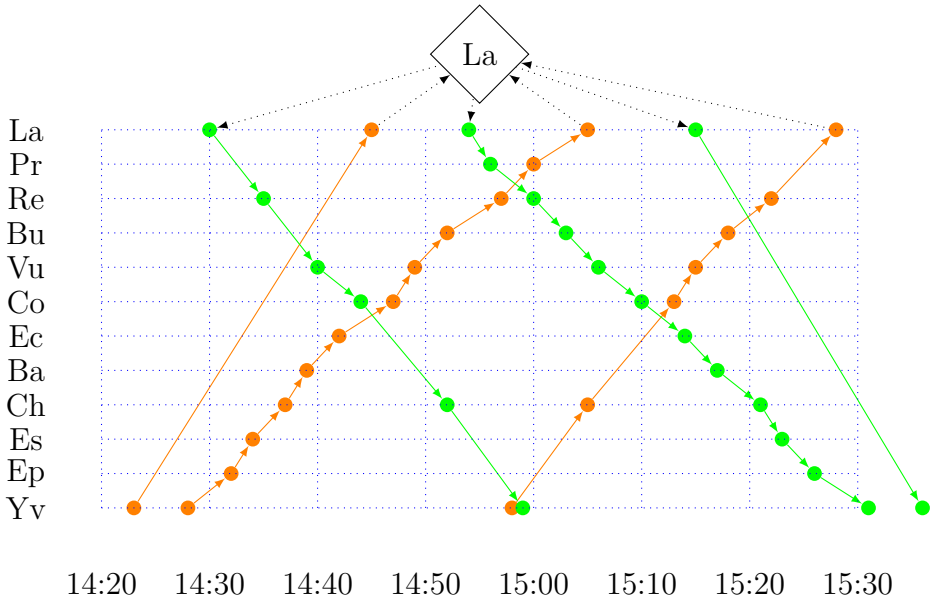


Figure 6.14: Scheduled public transportation network: time on the x-axis, space on the y-axis

Waiting time (min) at stations, determined by the schedule.

Transfer time (min) between scheduled services.

Capacity (number of seats) for each scheduled departure.

6.3 Pedestrian networks

Representing pedestrian traffic mathematically poses significant challenges due to the unique characteristics of pedestrian movement. Unlike vehicular traffic, where roads and lanes provide a well-defined network, pedestrian mobility is more flexible and less constrained by infrastructure.

One key challenge is the *level of granularity*. Pedestrian trips often follow a *door-to-door* pattern, meaning that every possible origin and destination must be accounted for. Unlike transit or road networks with predefined stops and links, pedestrian movements occur on a highly detailed and dynamic scale, making it difficult to discretize the network efficiently.

Another difficulty arises from the absence of a *physical network*. While vehicles follow lanes and predefined paths, pedestrians can move freely across open spaces, sidewalks, plazas, and even cut through buildings or parks. Traditional graph-based models struggle to capture this flexibility, as the network

structure is not explicitly defined in the same way as for other transportation modes.

Additionally, pedestrian traffic is strongly influenced by *interactions with other modes*. Walking is often a *feeding mode*, meaning that pedestrians interact closely with public transportation, cycling, and even road traffic. This intermodal dependency complicates mathematical modeling, as walking trips are not always independent but are instead integrated into broader multimodal travel patterns.

6.4 Multi-modal networks

Transportation networks can be integrated to form a *multi-modal network*, where different modes of transportation are combined into a single system. This integration allows travelers to move seamlessly between modes, enhancing connectivity and efficiency.

A major challenge in constructing a multi-modal network is the *superposition of networks* for each mode. Each transportation mode has its own structure: roads for cars, tracks for trains, air routes for planes, and pedestrian pathways. These networks must be accurately represented while maintaining a coherent structure that allows for efficient transfers between modes.

Another complexity is modeling *all possible transfers*. Travelers often switch between modes at various locations such as park-and-ride facilities, train stations, bus terminals, airports, and ferry docks. These transfer points must be explicitly incorporated into the network, ensuring that passengers can navigate from one mode to another with minimal friction. The timing of transfers, waiting times, and synchronization of schedules further complicate the representation.

To address these challenges, multi-modal transportation models use layered network structures, where each mode is represented as a separate layer, and transfer links connect corresponding nodes across layers. This enables efficient modeling of intermodal journeys, optimizing route selection while accounting for transfer times, accessibility, and capacity constraints.

6.5 Paths

A *path* in a network is a sequence of links that connects an origin to a destination, ensuring a continuous and valid movement through the network. Paths represent possible routes that travelers or flows can take within the network.

Figure 6.15 illustrates an example of a transportation network, where nodes represent locations, and directed links indicate possible movements between them.

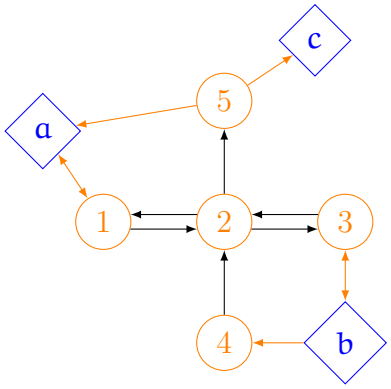


Figure 6.15: Example of a transportation network

The paths in this network, which represent possible routes between origins and destinations, are listed in Table 6.2.

Path description
$a \to 1 \to 2 \to 3 \to b$
$a \to 1 \to 2 \to 5 \to c$
$b \to 3 \to 2 \to 5 \to a$
$b \to 3 \to 2 \to 1 \to a$
$b \to 4 \to 2 \to 5 \to a$
$b \to 4 \to 2 \to 1 \to a$
$b \to 3 \to 2 \to 5 \to c$
$b \to 4 \to 2 \to 5 \to c$

Table 6.2: List of paths in the network

To mathematically represent the relationship between paths and links, we use a *link-path incidence matrix*, shown in Table 6.3. This matrix has rows corresponding to network links and columns corresponding to paths. Each entry is *1* if the link is used in the corresponding path and *0* otherwise.

The performance of a path in a transportation network can be expressed in various units, depending on the criterion used to evaluate travel quality. These units include *travel time*, *distance*, *travel cost*, and *generalized costs*, each providing different insights into the efficiency of a given route.

A fundamental requirement for path performance measures is that they must be *link additive*. This means that the total performance of a path

	1	2	3	4	5	6	7	8
$(\mathtt{a}, 1)$	1	1	0	0	0	0	0	0
$(\mathtt{b}, 3)$	0	0	1	1	0	0	1	0
$(\mathtt{b}, 4)$	0	0	0	0	1	1	0	1
$(1, \mathtt{a})$	0	0	0	1	0	1	0	0
$(1, 2)$	1	1	0	0	0	0	0	0
$(2, 1)$	0	0	0	1	0	1	0	0
$(2, 3)$	1	0	0	0	0	0	0	0
$(2, 5)$	0	1	1	0	1	0	1	1
$(3, \mathtt{b})$	1	0	0	0	0	0	0	0
$(3, 2)$	0	0	1	1	0	0	1	0
$(4, 2)$	0	0	0	0	1	1	0	1
$(5, \mathtt{a})$	0	0	1	0	1	0	0	0
$(5, \mathtt{c})$	0	1	0	0	0	0	1	1

Table 6.3: Link-path incidence matrix

must be obtained by summing up the performance values of each link that composes the path. Mathematically, if P is the link-path incidence matrix, if t_ℓ is the performance of link ℓ, then the total performance of a path p is given by:

$$\mathsf{c}_\mathsf{p} = \sum_\ell \mathsf{P}_{\ell\mathsf{p}} \mathsf{t}_\ell. \tag{6.3}$$

This property ensures that the overall path performance is well-defined and can be consistently evaluated by decomposing the path into its constituent links.

Several performance measures satisfy the *link additivity* property:

Travel time The total travel time of a path is simply the sum of the travel times of each link along the path.

Distance The total distance traveled along a path is the sum of the distances of the links.

Travel cost If each link has an associated monetary cost, the total cost of the path is the sum of individual link costs.

Generalized costs This combines multiple factors such as travel time, distance, and monetary cost into a single performance metric, and remains additive over links.

However, some measures are *not* link additive and cannot be used directly to evaluate path performance:

Speed Speed is not additive because the average speed of a path is not the sum of the speeds of its links. Instead, it depends on the total travel time and distance.

Flow Flow represents the number of users or vehicles on a link, but the flow of a path is not simply the sum of link flows. Instead, it depends on network-wide equilibrium and routing constraints.

6.6 Summary

Networks play a fundamental role in modeling and analyzing transportation systems. They represent the *transportation supply*, providing the infrastructure through which travelers (and goods) move. A transportation network also serves as the *interface with the demand*, connecting users with the available travel options and enabling the study of congestion, route choices, and accessibility.

A transportation network is structured using *network models*, which consist of *nodes* and *links*. Nodes represent specific locations such as intersections, stations, or bus stops, while links define the connections between these locations, corresponding to roads, railway lines, or air routes. The model is further enriched with *data* associated with these components, including travel times, distances, capacities, and costs, allowing for a quantitative analysis of network performance.

One of the key challenges in transportation network modeling is complexity. The level of detail must be consistent with the needs of the analysis. A model that is too simple may fail to capture essential dynamics and provide misleading results, making it useless for decision-making. Conversely, a model that is too complex may include excessive details, leading to computational intractability and impractical solutions. Striking the right balance between simplicity and accuracy is important to ensuring that the model remains both useful and computationally feasible.

Chapter 7

Travel demand: the four step model

In this chapter, we continue our investigation of the four-step model, a fundamental approach in trip-based transportation modeling. It consists of four sequential stages: trip generation, trip distribution, modal split, and assignment. The first step, *trip generation*, has already been detailed in Chapter 5. This stage estimates the number of trips produced and attracted by each zone based on various explanatory variables, typically using linear regression models. At this point, we have access to the key data necessary for further steps.

Specifically, we have information on *trip production* and *trip attraction*. The number of trips originating from each zone, denoted as $\mathsf{O}_\mathtt{r}$, and the number of trips destined for each zone, denoted as $\mathsf{D}_\mathtt{r}$, have been determined. These quantities are treated as random variables, as they result from statistical models.

Additionally, we have access to *transportation networks*. As described in Chapter 6, there is a separate network for each mode of transport, capturing network performance and associated costs. The cost of traveling between an origin $\mathtt{r}$ and a destination $\mathtt{s}$ by mode $\mathtt{i}$ is represented as $\mathsf{c}^\mathtt{i}_{\mathtt{rs}}$. These costs are assumed to be deterministic. To facilitate comparisons across modes, we define a generalized cost $\mathsf{c}_{\mathtt{rs}} = \min_\mathtt{i} \mathsf{c}^\mathtt{i}_{\mathtt{rs}}$, which represents the lowest travel cost among all available modes for a given origin-destination pair.

With this information in hand — trip production, trip attraction, and transportation networks — we are now ready to proceed to the next stages of the four-step model: *trip distribution*, described in Section 7.1, and *modal split*, described in Section 7.2. The last step is covered in Chapter 8.

7.1 Trip distribution

The objective of the *trip distribution* step is to determine how many trips occur between each origin-destination (OD) pair. This results in the construction of an *origin-destination table*, which specifies the expected number of trips, denoted as $\mathtt{f}_{\mathtt{rs}}$, between each origin $\mathtt{r}$ and destination $\mathtt{s}$. Ideally, the total number of trips produced in each zone should match the expected productions, and the total number of trips attracted to each zone should match the expected attractions.

There are two important challenges associated with trip distribution. The first is the issue of *incompatibility*, which arises when the total number of expected trip productions does not equal the total number of expected trip attractions. This inconsistency prevents the direct balancing of flows across the network. To address this, one possible approach is to represent trip productions and attractions using random variables, allowing for a probabilistic formulation that accounts for uncertainty. Another strategy is to relax the strict equality constraints and instead ensure that the expected sum of distributed trips is approximately equal to the expected productions and attractions.

A second major issue is *under-determination*, meaning that there are infinitely many possible solutions that satisfy the production and attraction constraints. In other words, without additional information, multiple different OD tables could be generated, making it difficult to determine a unique and reasonable solution. To overcome this problem, two strategies can be employed. The first is to incorporate *more data*, such as historical trip observations or survey data, to refine the estimates. The second is to introduce *additional assumptions* which can help guide the model toward a more realistic solution.

We now explore various methods to address these challenges, ensuring that the trip distribution model produces accurate and meaningful results.

7.1.1 Data collection: surveys

Several types of data sources can provide valuable insights into travel patterns. *Surveys* are a direct method of collecting travel information from individuals. Roadside interviews, for instance, involve stopping vehicles at specific locations to ask drivers about their trips, including their origin, destination, and purpose. Another approach is *license plate mail-out surveys*, where license plate numbers are recorded at checkpoints, and drivers are later contacted to provide details about their trips. More modern techniques include the use of *GPS data*, which can track vehicle movements and provide

highly detailed origin-destination information over time.

Figure 7.1 illustrates a strategy for screening vehicles in order to collect travel data. The map represents the Canton of Vaud, with an orange line dividing the region into two parts: a northern and a southern section. Every vehicle, or a sample of them, that crosses this line is subject to an interview where drivers are asked to report the origin and destination of their trips.

This approach enables the systematic collection of information about travel flows across the region. By focusing on a specific screening line, it is possible to gather data on major travel patterns without requiring exhaustive surveys throughout the entire network. The collected data can then be used to infer origin-destination flows, helping to construct more accurate trip distribution models.

Figure 7.1: Example of a screening strategy

Once data is collected through such a screening strategy, it provides valuable informarion into the movement of vehicles across different regions. The type of data obtained from a screening survey is illustrated in Figure 7.2.

The collected data provides information on the number of vehicles traveling from the north to the south and vice versa. The survey is most effective for capturing trips that cross the screening line. However, it provides less information about trips that begin and end within the same region (e.g., entirely in the north or entirely in the south). Therefore, the screening strategy

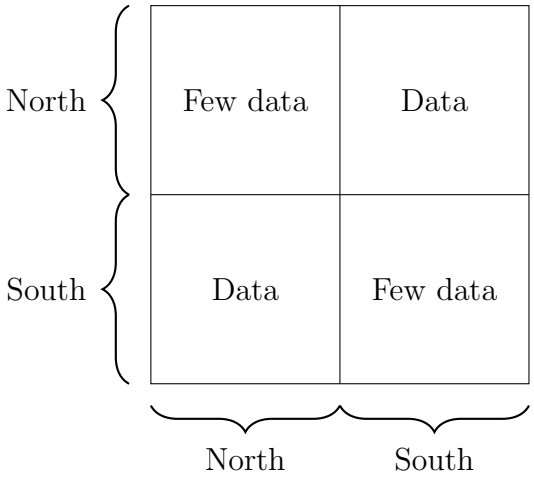


Figure 7.2: Data collected from a screening survey

must be designed consistently with the data needs.

An alternative screening strategy is illustrated in Figure 7.3. In this approach, specific cities — Lausanne, Yverdon-les-Bains, Vevey, and Nyon — are encircled by a screen line to capture all trips that either originate from or arrive at these strategic locations in the canton. In addition to the primary screen line capturing north-south traffic, two additional lines are introduced: one capturing traffic between the west and the center, and another capturing traffic between the east and the center.

This multi-line screening strategy provides a more comprehensive dataset, ensuring that not only inter-regional movements but also cross-directional travel patterns are captured effectively.

Roadside interviews come with several challenges that must be carefully considered. One major issue is the *sampling rate*. Conducting roadside interviews is an expensive process, as it requires personnel, equipment, and logistical coordination. Given a fixed budget, the number of interviews that can be conducted is limited. For instance, if resources allow for 1000 interviews, all of these could be collected at a single screen line, providing a relatively dense dataset for that particular location. However, if the same budget is distributed across seven different screen lines, the number of interviews per location drops significantly to just 143 per line. This reduction in sample size per screen line can make it more difficult to obtain statistically reliable data for each individual crossing point.

Another challenge concerns *logistics*. Roadside interviews often require stopping vehicles, which can disrupt traffic flow and may necessitate police

Figure 7.3: Alternative screening strategy focusing on key cities and additional traffic corridors

intervention to ensure safety. This can make the data collection process complex and time-consuming. In the case of public transportation, in-vehicle interviews are sometimes conducted instead. While this approach avoids interrupting traffic, it presents its own set of difficulties, such as coordinating with transit operators and ensuring that interviewers can effectively engage with passengers during their journeys.

Finally, roadside interviews introduce potential *biases* in the collected data. In-vehicle interviews, for example, are more likely to capture travelers making long-distance trips, since these individuals remain in the vehicle for extended periods and are thus more likely to be available for questioning. Additionally, some travelers may cross multiple screen lines, which can lead to overrepresentation of certain trip patterns in the dataset. This can distort the overall picture of mobility patterns and requires careful handling when interpreting the results.

To illustrate the previous discussion, we consider a simple example of a building with an elevator. The goal is to construct the OD table that describes the movement of individuals between floors. For the sake of this example, we assume that we have access to the *true* OD table, which provides the exact number of trips occurring between each pair of floors on a given

day. This table is presented in Table 7.1.

	0	1	2	3	4	
0		0.0	500.0	10.0	0.0	510.0
1	100.0		0.0	0.0	0.0	100.0
2	30.0	0.0		0.0	0.0	30.0
3	60.0	0.0	10.0		0.0	70.0
4	70.0	0.0	0.0	10.0		80.0
	260.0	0.0	510.0	20.0	0.0	

Table 7.1: True origin-destination table for the elevator example.

This table provides a complete description of the trips occurring in the building. Each row represents an *origin floor*, and each column represents a *destination floor*. The values in the table indicate the number of trips from one floor to another. For example, the value 500 in row 0 and column 2 means that 500 trips originate from floor 0 and have floor 2 as their destination. The last column in the table represents the total number of trips originating from each floor, while the last row shows the total number of trips arriving at each floor.

In practice, however, we rarely have direct access to such a detailed OD table. Instead, we often rely on aggregated data from the *trip generation* step, which provides information on the number of trips produced by and attracted to each floor. For each origin,

Floor	O_r	D_r
0	515.5	248.8
1	98.9	0.0
2	16.4	506.4
3	51.3	9.6
4	96.2	0.0
Total	778.3	764.8

Table 7.2: Total trips originating from and arriving at each floor.

In Table 7.2, the column O_r represents the total number of trips that originate from each floor, while the column D_r represents the total number of trips arriving at each floor. The last row provides the sum of all originating and arriving trips.

It is important to remember that the values in this table do not correspond to direct measurements but rather to the output of a model. As such, they contain errors that must be accounted for when using them to estimate the full OD matrix. A key observation is that the total number of trips generated (778.3) does not exactly match the total number of trips attracted (764.8). This discrepancy highlights a common issue in real-world data collection, where inconsistencies arise due to estimation errors, missing data, or variations in model assumptions.

In order to illustrate the concept of road-side interviews, assume now that we conduct interviews on the ground floor. Individuals entering the building are asked about their destination, while those exiting are asked about their origin.

	0	1	2	3	4
0		0.0	501.9	9.6	0.0
1	100.7				
2	29.7				
3	59.5				
4	70.9				

Table 7.3: Data collected from interviews at the ground floor.

Table 7.3 captures partial origin-destination information based on those direct surveys conducted at the entrance and exit of the building. The row indices represent the origin floors of individuals exiting the building, while the column indices correspond to the destination floors of those entering.

The empty cells correspond to the absence of data. Indeed, while this dataset provides useful insights into elevator usage, it remains incomplete. It does not account for trips occurring entirely between upper floors. Moreover, it may contain biases due to sampling limitations. It can be seen by comparing the values with the "true" values in Table 7.1.

To estimate the entries of the OD table, the first idea that comes to mind is to formulate the problem as a linear regression model. Since we treat the number of trips generated O_r, the number of trips attracted D_r, and the partially observed flows $\widehat{f}_{rs}$ as random variables, we define a least-squares estimation approach to determine the most likely values of f_{rs}. The objective function to minimize is given by:

$$\min_{f} \sum_{r} \left(O_r - \sum_{s} f_{rs} \right)^2 + \sum_{s} \left(D_s - \sum_{r} f_{rs} \right)^2 + \sum_{rs} \left(\widehat{f}_{rs} - f_{rs} \right)^2.$$

Each term in this equation plays a specific role in ensuring that the estimated OD table remains consistent with the available data:

- The first term, $\sum_r (O_r - \sum_s f_{rs})^2$, ensures that the estimated number of trips originating from each floor aligns as closely as possible with the observed trip generation values O_r. Since the estimated flows f_{rs} must sum to the total number of trips generated from each floor r, any deviation is penalized in the objective function.

- The second term, $\sum_s (D_s - \sum_r f_{rs})^2$, enforces a similar constraint on trip attractions. It ensures that the total number of trips arriving at each floor s matches the observed trip attraction values D_s as closely as possible. Again, discrepancies are penalized to encourage consistency.

- The third term, $\sum_{rs} \left(\widehat{f}_{rs} - f_{rs}\right)^2$, incorporates the partially observed flows obtained from survey data. Since these measurements are subject to errors and sampling limitations, the estimation process does not enforce strict equality but instead minimizes the deviation between the estimated values and the observed samples $\widehat{f}_{rs}$.

When we solve the least-squares problem, we obtain the estimated origin-destination table shown in Table 7.4. Ideally, this solution should provide a reasonable approximation of the true OD flows based on the available data. However, the results reveal several issues that must be addressed.

	0	1	2	3	4	
0		-18.0	483.9	-8.4	-18.0	439.4
1	55.8		-154.1	-76.2	197.0	22.5
2	-15.2	36.3		39.3	-120.4	-60.0
3	14.6	74.3	-141.5		27.5	-25.1
4	26.0	-16.1	-111.9	121.8		19.8
	81.3	76.4	76.4	76.4	86.0	396.5

Table 7.4: Estimated OD table obtained from the least-squares solution.

One aspect of the solution that is *not* problematic is the fact that the estimated values are not integers. While it might seem intuitive to expect integer values since trips are made by individual people, an OD table does not represent the count of individual travelers. Instead, it describes *flows*, which correspond to the number of persons traveling per unit of time (e.g., per hour or per day). These flow values are continuous and are typically represented

by real numbers. As such, obtaining non-integer values is expected and does not indicate an issue with the model itself.

However, the most problematic aspect of the solution is the presence of *negative entries*. In an OD table, each entry represents a number of trips between an origin and a destination, which must necessarily be non-negative. However, the results show several negative values, such as -18.0 trips from floor 0 to floor 1 and -154.1 trips from floor 1 to floor 2. These values are clearly nonsensical in a real-world setting, as the number of trips cannot be negative.

The issue of negative entries arises because the least-squares approach does not inherently enforce non-negativity constraints. The model attempts to minimize discrepancies between the estimated flows and the available data, but in doing so, it allows for solutions that may not be meaningful in the context of human mobility. This is a fundamental limitation of using an unconstrained least-squares formulation for OD estimation.

This issue can be understood in the context of *maximum likelihood estimation*. Indeed, the least-squares estimator is the maximum likelihood estimator (MLE) for a linear regression model under the assumption of normally distributed errors (see Section 4.3.2). This assumption plays a key role in explaining why negative values can arise.

Indeed, a fundamental property of the normal distribution is that it has *infinite support*, meaning that any value, including negative ones, has a nonzero probability of occurring. When the true value of a parameter is close to zero, the likelihood of obtaining a negative estimate becomes significant due to the spread of the normal distribution.

To illustrate this phenomenon, consider the probability density functions in Figures 7.4 and 7.5, which show two different normal distributions centered at different values.

In Figure 7.4, the normal distribution has a mean of 0.2 with a standard deviation of 0.3. Because the mean is close to zero, a substantial portion of the probability mass falls into the negative region, leading to a high probability of obtaining negative estimates.

In contrast, Figure 7.5 shows a normal distribution with a mean of 1 and the same standard deviation of 0.3. In this case, most of the probability mass remains in the positive domain, meaning that negative estimates are much less likely to occur.

The same reasoning applies to the estimated OD flows. If the true value of a flow is small, the normal variation around that value makes it likely that some estimates will be negative. Since the least-squares approach does not impose constraints on the sign of the estimates, the method freely assigns negative values when the optimization process dictates it. This explains

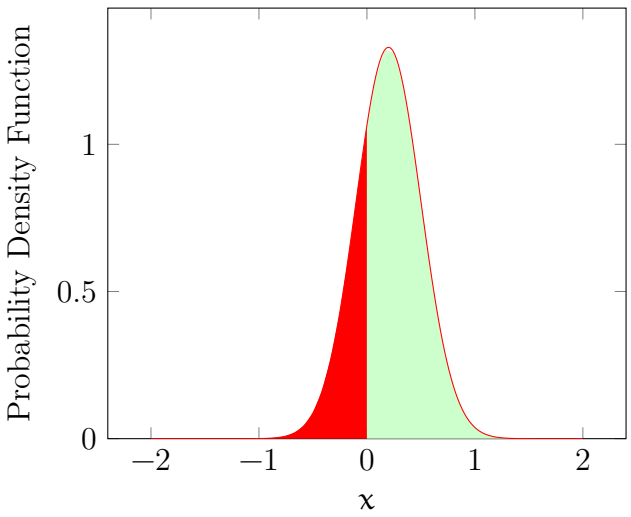


Figure 7.4: Probability density function of a normal distribution with mean close to zero. A significant portion of the distribution lies in the negative region.

why some of the estimated OD flows in our table are negative, even though negative trip counts are not meaningful in reality.

A natural way to address the issue of negative values in the estimated OD table is to enforce a strict non-negativity constraint. One way to achieve this is to reparameterize the OD flows using an exponential transformation. Instead of estimating f_{rs} directly, we define:

$$f_{rs} = \exp(\tau_{rs}), \quad \tau_{rs} \in \mathbb{R}.$$

This transformation ensures that the estimated flow values f_{rs} are always positive, regardless of the values taken by the underlying parameters τ_{rs}. Since the exponential function only produces positive outputs, this approach inherently avoids the problem of negative estimates that arise in the standard least-squares formulation.

In this new formulation, the regression is performed on the transformed variable τ_{rs} instead of f_{rs}. Because least-squares estimation corresponds to maximum likelihood estimation under the assumption of normally distributed errors, the estimated values of τ_{rs} remain normally distributed:

$$\hat{\tau}_{rs} \sim \mathcal{N}(\mu_{rs}, \sigma^2_{rs}).$$

Since f_{rs} is defined as the exponential of τ_{rs}, the estimated values of f_{rs} follow a *log-normal distribution* instead of a normal distribution. This is

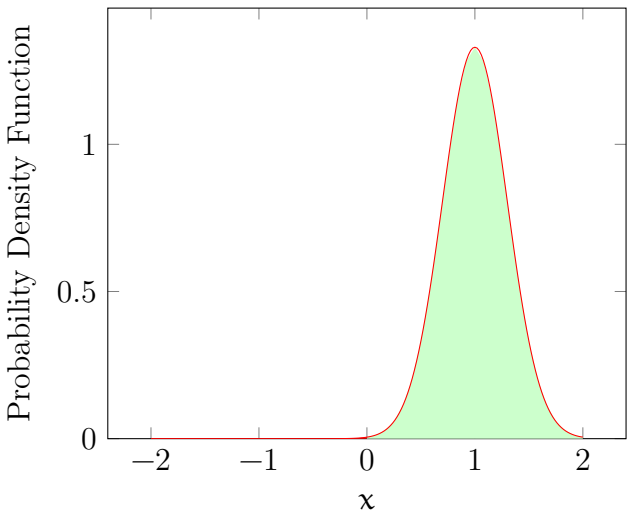


Figure 7.5: Probability density function of a normal distribution with mean far from zero. The probability of negative values is much lower.

an important distinction because log-normal distributions are strictly positive, which aligns naturally with the requirement that OD flows cannot be negative.

This approach offers a key advantage: it allows us to maintain the principles of maximum likelihood estimation while ensuring that all estimated flows remain non-negative. However, the main drawback is that the introduction of an exponential function makes the model *nonlinear*. As a result, solving the estimation problem requires nonlinear optimization techniques, which are computationally more complex than standard linear regression. Despite this added complexity, the use of a log-normal model provides a more robust and theoretically sound framework for estimating OD flows while addressing the fundamental issue of negative values.

Using this reformulation, we define the nonlinear least-squares problem as:

$$\min_{\tau} \sum_{r} \left(O_r - \sum_{s} e^{\tau_{rs}} \right)^2 + \sum_{s} \left(D_s - \sum_{r} e^{\tau_{rs}} \right)^2 + \sum_{rs} \left(\bar{f}_{rs} - e^{\tau_{rs}} \right)^2 .$$

Similary to the previous specification, each term in this objective function ensures consistency between the estimated flows and the available data:

- The first term, $\sum_r \left(O_r - \sum_s e^{\tau_{rs}}\right)^2$, enforces that the total number of

trips originating from each floor matches the observed trip generation values O_r as closely as possible.

- The second term, $\sum_s \left(D_s - \sum_r e^{\tau_{rs}}\right)^2$, ensures that the total number of trips arriving at each floor remains consistent with the observed trip attraction values D_s.
- The third term, $\sum_{rs} \left(\bar{f}_{rs} - e^{\tau_{rs}}\right)^2$, incorporates prior information from observed partial flows, minimizing the deviation between the estimated and observed values.

By solving this nonlinear optimization problem, we obtain the estimated OD table shown in Table 7.5.

	0	1	2	3	4	
0		57.2	252.1	65.0	65.0	439.2
1	44.1		0.0	0.0	0.0	44.1
2	0.0	0.0		0.0	0.0	0.0
3	0.0	0.0	0.0		0.0	0.0
4	26.2	6.4	0.0	0.0		32.6
	70.2	63.6	252.1	65.0	65.0	515.9

Table 7.5: Estimated OD table using the nonlinear least-squares approach.

By construction, all values in Table 7.5 are non-negative, addressing the primary issue encountered in the standard least-squares formulation.

When comparing Table 7.5 with the true OD table in Table 7.1, it becomes evident that the estimated values remain far from the actual ones. This discrepancy is particularly noticeable for the entries corresponding to the ground floor. This is surprising, given that we have invested additional resources in collecting survey data specifically for this floor, as shown in Table 7.3.

One key issue is that the current estimation method treats all data sources equally. However, the production-attraction data (O_r and D_r) are not directly observed values but rather outputs of a model from the first phase of the estimation process. In contrast, the survey data $\bar{f}_{rs}$ consists of direct observations, providing valuable disaggregate information about trips between specific origins and destinations. This disaggregate nature is important, and we do not want to distort it by treating these data points the same way as the aggregated production-attraction data.

To address this issue, we introduce *weighted least squares*, in which different components of the objective function are assigned different weights

according to their reliability and importance. The new optimization problem is formulated as:

$$\min_{\tau} w_o^2 \sum_r \left(O_r - \sum_s e^{\tau_{rs}} \right)^2 + w_d^2 \sum_s \left(D_s - \sum_r e^{\tau_{rs}} \right)^2 + w_f^2 \sum_{rs} \left(\bar{f}_{rs} - e^{\tau_{rs}} \right)^2 .$$

Here, w_o, w_d, and w_f are weights that determine the relative importance of the trip generation, trip attraction, and survey data, respectively. By setting w_f to a larger value, we put greater emphasis on ensuring that the estimated OD table remains close to the directly observed survey data, reducing discrepancies for the ground floor.

The results obtained using $w_o = w_d = 1$ and $w_f = 100$ are presented in Table 7.6.

	0	1	2	3	4	
0		0.0	500.8	9.6	0.0	510.4
1	100.2		0.0	0.0	4.1	104.3
2	29.2	0.0		0.0	0.0	29.2
3	58.9	0.0	0.0		0.2	59.1
4	70.3	11.8	0.0	2.3		84.4
	258.6	11.8	500.8	11.9	4.3	787.4

Table 7.6: Estimated OD table using weighted least squares with increased emphasis on survey data.

By increasing the weight of the survey data, the estimated values now align much more closely with the true OD table. The discrepancies for the ground floor have been significantly reduced, ensuring that the additional information obtained from interviews is effectively incorporated into the estimation process. This highlights the importance of assigning appropriate weights to different data sources in order to maximize the accuracy and reliability of the OD estimation.

To formalize the modeling assumptions, we introduce the sets, data, and regression equations that underpin the estimation process.

The model considers a set of *centroids*, indexed by $r = 1, \ldots, N$, which represent distinct locations within the study area. Additionally, we define *survey zones*, indexed by $p = 1, \ldots, P$, where direct survey data is available. Each survey zone p consists of a subset of centroids, denoted as S_p.

The data used in the model consists of three key components:

- *Production data* (O_r), representing the total number of trips originating from each centroid r.
- *Attraction data* (D_s), representing the total number of trips arriving at each centroid s.
- *Survey data* ($\bar{f}_{rs}$), which provides direct observations of flows between centroids belonging to a survey zone ($r \in \mathcal{S}_p$) and centroids outside the survey zone ($s \notin \mathcal{S}_p$).

To estimate the unknown OD flows, we define a system of regression equations:

$$O_r = \sum_{s=1}^{N} e^{\tau_{rs}} + \sigma_o \varepsilon_r^o,$$

$$D_s = \sum_{r=1}^{N} e^{\tau_{rs}} + \sigma_d \varepsilon_s^d,$$

$$\ln \bar{f}_{rs} = \tau_{rs} + \sigma_{rs} \varepsilon_{rs}.$$

These equations describe the relationship between observed data and the underlying travel flows. The first equation ensures that the estimated flows are consistent with the observed trip production values, while the second equation enforces consistency with the trip attraction data. The third equation models the survey data in a way that preserves its disaggregate nature.

However, an important issue arises: the variance of the error terms σ_o, σ_d, and σ_{rs} is not the same across all observations. This violates the fundamental assumption of ordinary least squares regression, which assumes homoscedastic[1] errors. As a result, we must account for differences in data reliability by introducing *weights* into the estimation process.

Weighted least squares addresses this issue by assigning a weight to each observation, ensuring that more reliable data has a stronger influence on the final estimates. Specifically, we introduce different weights for each data type:

- w_o for trip production data.
- w_d for trip attraction data.
- w_f for survey data.

[1] It is a complicated way to say that they have the same variance.

The weights must be defined beforehand and should reflect the precision of the data. More precise observations receive larger weights, effectively reducing their variance in the regression model. Typically, the survey data is assigned a much higher weight than the production and attraction data ($w_f \geqslant w_o \approx w_d$) since it represents direct observations rather than modeled outputs.

The weighted least squares optimization problem is then formulated as:

$$\min_{\tau} w_o^2 \sum_r \left(O_r - \sum_s e^{\tau_{rs}} \right)^2 + w_d^2 \sum_s \left(D_s - \sum_r e^{\tau_{rs}} \right)^2 + w_f^2 \sum_{rs} \left(\bar{f}_{rs} - e^{\tau_{rs}} \right)^2 .$$

By incorporating these weights, we account for variations in data reliability and improve the accuracy of the OD table estimation. The corresponding regression equations are adjusted as follows:

$$O_r = \sum_{s=1}^{N} e^{\tau_{rs}} + \frac{\sigma}{w_o} \varepsilon_r^o,$$

$$D_s = \sum_{r=1}^{N} e^{\tau_{rs}} + \frac{\sigma}{w_d} \varepsilon_s^d,$$

$$\bar{f}_{rs} = e^{\tau_{rs}} + \frac{\sigma}{w_f} \varepsilon_{rs}.$$

Larger weights correspond to smaller variances, meaning that observations with higher reliability contribute more to the final solution. This methodology ensures that the estimation process effectively balances the different sources of information while maintaining statistical rigor.

7.1.2 Data collection: traffic counts

Another important data source is *traffic counts*, which provide indirect information about travel flows. These can be obtained using *loop detectors*, which are embedded in road surfaces to count passing vehicles. Similarly, *pneumatic road tubes* are temporary installations that record vehicle counts based on air pressure changes when vehicles pass over them. *Magnetic sensors* are another option, detecting vehicles through changes in the magnetic field.

The measured flow on a given link ℓ is denoted as $\bar{x}_\ell$, and while this data is available for some links, it does not provide a direct observation of the

complete OD flows. Instead, it reflects aggregated traffic volumes resulting from multiple OD pairs using the same link.

To integrate traffic count data into OD estimation, we rely on the *assignment matrix*, a fundamental tool that describes how OD flows translate into link flows. This matrix, denoted as Q, encodes the relationship between the number of trips between each OD pair and the resulting traffic on network links.

The assignment matrix has the following properties:

- It transforms OD flows into link flows, capturing the distribution of traffic across the network.
- It consists of a number of rows equal to the number of links in the network and a number of columns equal to the number of OD pairs.
- It is only available after the *assignment phase*, during which trips are routed through the network according to a traffic assignment model (see Chapter 8).

The mathematical relationship between OD flows and link flows is expressed as:

$$\mathsf{x} = \mathsf{Q}\mathsf{f}, \quad \mathsf{x}_\ell = \sum_{\mathsf{q}} \mathsf{Q}_{\ell\mathsf{q}}\mathsf{f}_\mathsf{q}, \quad \text{where } \mathsf{q} = (\mathsf{r}, \mathsf{s}).$$

Here, x represents the vector of link flows, f represents the vector of OD flows, and Q is the assignment matrix. The entry $\mathsf{Q}_{\ell\mathsf{q}}$ indicates the proportion of flow from OD pair q that traverses link ℓ. By summing over all OD pairs, we obtain the total flow on each link.

To better understand the structure of the assignment matrix, we decompose it into two separate matrices: one capturing the *network topology* and the other describing *route choice behavior*.

The first component, the *link-path incidence matrix* P, introduced in Section 6.5, encodes the physical structure of the network. This matrix has dimensions equal to the number of links by the number of paths, where each entry $\mathsf{P}_{\ell\mathsf{p}}$ takes the value 1 if link ℓ is part of path p, and 0 otherwise. The link-path incidence matrix of the network represented in Figure 7.6 is reported in Table 7.7.

The second component, the *OD-path matrix* R, captures travelers' route choices. This matrix has dimensions equal to the number of paths by the number of OD pairs, and each entry $\mathsf{R}_{\mathsf{p}\mathsf{q}}$ represents the proportion of OD flow q that uses path p. Table 7.8 provides an example of the OD-path matrix

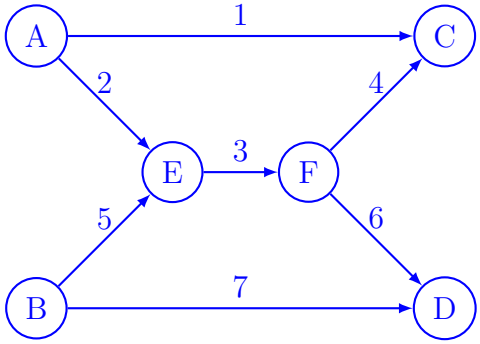


Figure 7.6: Example of network topology

	A-C	A-E-F-C	A-E-F-D	B-E-F-C	B-E-F-D	B-D
1	1	0	0	0	0	0
2	0	1	1	0	0	0
3	0	1	1	1	1	0
4	0	1	0	1	0	0
5	0	0	0	1	1	0
6	0	0	1	0	1	0
7	0	0	0	0	0	1

Table 7.7: Path-link matrix encoding network topology.

for the network represented in Figure 7.6, where there are multiple possible paths for different OD pairs. For instance, to go from A to C, the path $\mathsf{A}-\mathsf{C}$ or the path $\mathsf{A}-\mathsf{E}-\mathsf{F}-\mathsf{C}$ can be used, each of them with probability 0.5.

By multiplying these two matrices together, we obtain the *assignment matrix* Q, which directly relates OD flows to link flows:

$$\mathsf{Q} = \mathsf{PR}.$$

Each entry $\mathsf{Q}_{\ell\mathsf{q}}$ in this matrix represents the proportion of OD flow q that uses link ℓ. Table 7.9 shows an example of the assignment matrix derived from the network topology and route choices.

To illustrate how OD flows are transformed into link flows using the assignment matrix, we consider an example where OD demand is assigned to the network from Figure 7.6. The assignment process is governed by the relationship:

	A-C	A-D	B-C	B-D
A-C	0.5	0	0	0
A-E-F-C	0.5	0	0	0
A-E-F-D	0	1	0	0
B-E-F-C	0	0	1	0
B-E-F-D	0	0	0	0.5
B-D	0	0	0	0.5

Table 7.8: OD-path matrix capturing route choice proportions.

	A-C	A-D	B-C	B-D
1	0.5	0	0	0
2	0.5	1	0	0
3	0.5	1	1	0.5
4	0.5	0	1	0
5	0	0	1	0.5
6	0	1	0	0.5
7	0	0	0	0.5

Table 7.9: Assignment matrix Q derived from path-link and OD-path matrices.

$$x = Qf,$$

where x represents the vector of link flows, Q is the assignment matrix, and f is the vector of OD flows.

The assignment matrix, OD flows, and resulting link flows are presented in Table 7.10. Each row in the matrix corresponds to a network link, while each column corresponds to an OD pair. The multiplication of the assignment matrix with the OD flow vector yields the total flow on each link.

The network representation in Figure 7.7 provides a visual interpretation of this transformation. Solid arrows indicate link flows, while dotted red arrows represent the OD flows.

Thanks to the assignment matrix, we can now incorporate traffic count data into the OD estimation process. This is achieved through the following weighted least-squares formulation:

$$\min_{\tau} w_o^2 \sum_r \left(O_r - \sum_s e^{\tau_{rs}} \right)^2 + w_d^2 \sum_s \left(D_s - \sum_r e^{\tau_{rs}} \right)^2 + w_\ell^2 \sum_\ell \left(\bar{x}_\ell - \sum_q Q_{\ell q} e^{\tau_q} \right)^2 .$$

$$\underbrace{\begin{pmatrix} 0.5 & 0 & 0 & 0 \\ 0.5 & 1 & 0 & 0 \\ 0.5 & 1 & 1 & 0.5 \\ 0.5 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0.5 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & 0 & 0.5 \end{pmatrix}}_{\text{Assignment matrix}} \quad \underbrace{\begin{bmatrix} 100 \\ 200 \\ 300 \\ 400 \end{bmatrix}}_{\text{OD flows}} = \underbrace{\begin{bmatrix} 50 \\ 250 \\ 750 \\ 350 \\ 500 \\ 400 \\ 200 \end{bmatrix}}_{\text{Link flows}}$$

Table 7.10: Example of assignment matrix application: transformation of OD flows into link flows.

Each term in this objective function corresponds to a different source of data, ensuring that the estimated OD table remains consistent with available information. As before, the first term, $\sum_r \left(O_r - \sum_s e^{\tau_{rs}}\right)^2$, enforces consistency between the estimated OD flows and the observed trip production values. Similarly, the second term, $\sum_s \left(D_s - \sum_r e^{\tau_{rs}}\right)^2$, ensures alignment with observed trip attractions.

The new term introduced in this formulation, $\sum_\ell \left(\bar{x}_\ell - \sum_q Q_{\ell q} e^{\tau_q}\right)^2$, integrates traffic count data into the estimation process. Here, $\bar{x}_\ell$ represents the observed flow on link ℓ, while $\sum_q Q_{\ell q} e^{\tau_q}$ corresponds to the estimated flow on that link, obtained by summing the contributions from all OD pairs weighted by the assignment matrix Q. This term ensures that the estimated OD flows are compatible with the observed link flows, further refining the accuracy of the OD table.

The role of the weights w_o, w_d, w_ℓ remains the same as discussed previously. Each weight reflects the reliability and importance of the corresponding data source. A larger weight increases the influence of that term in the objective function, giving higher priority to reducing discrepancies for that particular type of data. Typically, traffic count data is assigned a weight w_ℓ that reflects the precision of the measurements, ensuring that the estimation method appropriately balances information from different sources.

7.1.3 More assumptions: the gravity model

Another way to resolve the under-determination issue is to introduce additional assumptions to guide the estimation process. One widely used approach is the *gravity model*, which provides a systematic way to estimate OD flows based on intuitive behavioral principles.

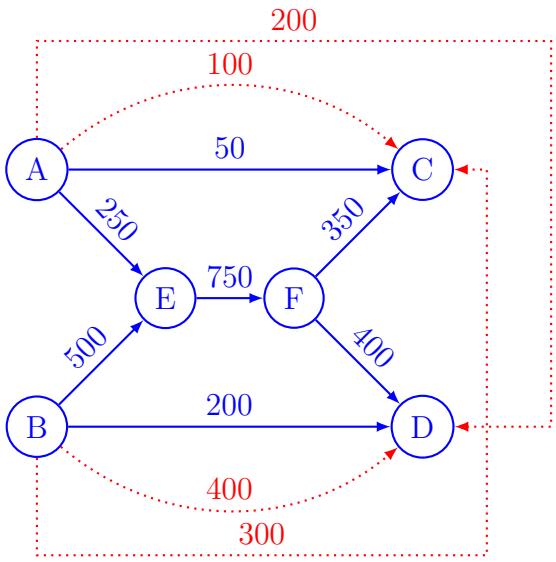


Figure 7.7: Network representation of OD and link flows. Solid arrows represent link flows, while dotted red arrows indicate OD flows.

The gravity model is inspired by Newton's law of universal gravitation, which states that the force between two objects is proportional to their masses and inversely proportional to the square of the distance between them. By analogy, the gravity model assumes that the flow f_{rs} between an origin r and a destination s is:

- Proportional to the *trip production* O_r at the origin, meaning that locations with higher trip generation will send more trips.
- Proportional to the *trip attraction* D_s at the destination, meaning that locations with high demand will receive more trips.
- Decreasing as the *generalized cost* c_{rs} increases, reflecting the fact that longer or more expensive trips are less likely to occur.

The simplest mathematical formulation of the gravity model, which directly follows the analogy with Newton's law, is given by:

$$f_{rs} \approx \frac{\alpha_r O_r \beta_s D_s}{c_{rs}^2}.$$

This equation states that the OD flow is inversely proportional to the square of the generalized cost. However, while this formulation gives the model its

name, it is not necessarily the most appropriate for travel demand modeling, as trip-making behavior does not always follow an inverse-square law.

To provide more flexibility, alternative mathematical formulations are commonly used in practice. One alternative expresses the deterrence effect of travel cost using an exponential decay function:

$$f_{rs} \approx \alpha_r O_r \beta_s D_s e^{-\gamma c_{rs}}.$$

Here, the decay is controlled by the parameter γ, which determines how quickly trip probability decreases with increasing cost.

More generally, practitioners may use any function $h(c_{rs})$ that satisfies $h'(c_{rs}) < 0$, ensuring that the OD flow decreases with increasing cost while allowing for different deterrence effects:

$$f_{rs} \approx \alpha_r O_r \beta_s D_s h(c_{rs}), \quad h'(c_{rs}) < 0.$$

The choice of $h(c_{rs})$ depends on empirical observations and model calibration, ensuring that the estimated OD flows align with real-world travel behavior.

The gravity model can be integrated into the weighted least-squares estimation process by modifying the objective function to include an additional term enforcing the gravity-based structure. The resulting optimization problem is formulated as:

$$\begin{aligned} \min_{\tau,\alpha,\beta,\gamma} \quad & w_o^2 \sum_r \left(O_r - \sum_s e^{\tau_{rs}} \right)^2 \\ & + w_d^2 \sum_s \left(D_s - \sum_r e^{\tau_{rs}} \right)^2 \\ & + w_g^2 \sum_{rs} \left(\alpha_r O_r \beta_s D_s e^{-\gamma c_{rs}} - e^{\tau_{rs}} \right)^2. \end{aligned}$$

The first two terms remain exactly as before:

- The first term, $\sum_r (O_r - \sum_s e^{\tau_{rs}})^2$, ensures that the estimated OD table aligns with observed trip productions.
- The second term, $\sum_s (D_s - \sum_r e^{\tau_{rs}})^2$, enforces consistency with trip attractions.

The third term introduces the gravity model constraint. It ensures that the estimated OD flows $e^{\tau_{rs}}$ remain close to the form suggested by the gravity model:

$$\alpha_r O_r \beta_s D_s e^{-\gamma c_{rs}}.$$

Here, α_r and β_s are balancing factors that adjust productions and attractions, while γ determines the sensitivity of flows to generalized cost. By minimizing the squared difference between this expression and the estimated OD flows, the model enforces the gravity assumption while still allowing for deviations where necessary.

Regarding the weights, we expect that $w_g \leqslant w_o \approx w_d$. This is because the gravity model assumption is a structural hypothesis rather than direct empirical data. Unlike productions and attractions, which are directly measured or estimated from other models, the gravity model is a theoretical formulation that may not perfectly fit all scenarios. A lower weight w_g allows flexibility, ensuring that the model can deviate from the gravity assumption when necessary to better match observed data.

It is important to note that the gravity model is *not* appropriate for all OD estimation problems. In particular, it is not suitable for the *elevator example*. The gravity model assumes that flows are influenced by production, attraction, and travel cost in a way that mirrors long-distance trip behavior. However, in the case of an elevator, trip-making decisions are not governed by distance deterrence in the same manner. The cost of traveling between floors is nearly identical, and people do not choose destinations based on minimizing effort in the same way they would in a large-scale transportation network. As a result, alternative methods — such as direct estimation from survey data — are more appropriate for modeling OD flows in an elevator system.

All sources of information can be integrated into a single model formulation, ensuring that the estimated OD table is consistent with multiple constraints simultaneously. The optimization problem is formulated as follows:

$$\begin{aligned}\min_{\tau,\alpha,\beta,\gamma} \; & w_o^2 \sum_r \left(O_r - \sum_s e^{\tau_{rs}} \right)^2 \\ & + w_d^2 \sum_s \left(D_s - \sum_r e^{\tau_{rs}} \right)^2 \\ & + w_g^2 \sum_{rs} \left(\alpha_r O_r \beta_s D_s e^{-\gamma c_{rs}} - e^{\tau_{rs}} \right)^2 \\ & + w_f^2 \sum_{rs} \left(\bar{f}_{rs} - e^{\tau_{rs}} \right)^2 \\ & + w_\ell^2 \sum_\ell \left(\bar{x}_\ell - \sum_q Q_{\ell q} e^{\tau_q} \right)^2 .\end{aligned}$$

This objective function combines multiple data sources, each contributing

to refining the estimated OD flows. The different terms correspond to:

- *Trip production*: Ensures that the total number of trips originating from each location matches observed productions.
- *Trip attraction*: Ensures that the total number of trips arriving at each location matches observed attractions.
- *Gravity model*: Encourages OD flows to follow the gravity model, maintaining consistency with travel behavior assumptions.
- *Survey data*: Ensures that estimated OD flows align with directly observed OD pairs from surveys.
- *Traffic count*: Ensures that estimated OD flows, when assigned to the network, match observed link flows.

Again, the weights $w_o, w_d, w_g, w_f, w_\ell$ determine the relative importance of each constraint. Higher weights give greater priority to reducing discrepancies for that particular data source. Typically, survey data and traffic counts are given higher weights when available, as they represent direct observations, whereas the gravity model provides a theoretical structure. This integrated approach ensures that the estimated OD table balances all available information in a statistically sound manner.

7.2 Modal split

Having completed the first two steps of the four-step model — *trip generation* and *trip distribution* — we now have a comprehensive origin-destination (OD) table. This table, with entries f_{rs}, represents the number of trips between each pair of zones or centroids (r, s). At this stage, we have a detailed understanding of how many trips are expected to occur between different locations but without any information about the travel mode that will be used. This brings us to the next step: *modal split*.

The modal split step aims to determine how these trips are distributed across different transportation modes. To do so, we must consider the characteristics of each mode, particularly their associated *generalized costs*. The generalized cost, denoted as c^i_{rs}, represents the perceived burden of traveling from r to s using mode i. This cost accounts for various factors, such as travel time, monetary expenses, comfort, and reliability.

At this point in the process, we recognize that:

- The OD table provides a detailed representation of the travel demand, but without mode-specific information.
- Each transportation mode i has its own network and associated generalized cost c_{rs}^{i}, which influences travelers' mode choices.

The next step in our investigation involves understanding how travelers select a mode based on these factors.

7.2.1 The logit model

To formally introduce the modal split problem, we define a choice model that determines how trips between an origin r and a destination s are distributed among the available transportation modes. For each OD pair (r, s), travelers can select a mode from the set $\mathcal{C}_{rs}$, which contains all feasible transportation options for that specific trip.

The most commonly used model for mode choice is the *logit model*, introduced in Section 4.3.3, which assigns a probability π_i^{rs} to each mode i in $\mathcal{C}_{rs}$. The probability of selecting mode i is given by:

$$\pi_i^{rs} = \frac{e^{-\theta c_{rs}^{i}}}{\sum_{j \in \mathcal{C}_{rs}} e^{-\theta c_{rs}^{j}}}, \quad \theta \geqslant 0.$$

This model is based on the principle that travelers are more likely to choose modes with lower generalized costs. The parameter θ controls the sensitivity of travelers to differences in cost:

- When θ is close to zero, all modes are chosen with nearly equal probability, meaning that travelers are not very sensitive to cost differences.
- When θ is large, travelers overwhelmingly prefer the mode with the lowest cost, making the choice process more deterministic.

The value of θ should be estimated using real observations.

We illustrate now the modal split in the context of the elevator example, to determine the proportion of travelers who choose to take the stairs versus the elevator. The set of available modes for each OD pair (r, s) consists of: $\mathcal{C}_{rs} = \{$ elevator, stairs $\}$.

As the floors are numbered in a consecutive way, the number of floors between origin r and destination s is denoted as $d_{rs} = |r - s|$. The logit model is used to estimate the probability of taking each mode. The utility functions are specified as:

$$\begin{aligned} u_{\text{elevator}} &= 0, \\ u_{\text{stairs}} &= -\theta d_{rs}, \quad \theta = 1.1. \end{aligned}$$

Note that, as discussed in Section 4.3.3, only differences in utility matter, not their absolute levels. As a consequence, one of the utility functions is normalized to zero without loss of generality. Here, the utility function of the elevator has been selected arbitrarily.

From these utilities, the probability of taking the stairs is computed as:

$$\pi^{rs}_{stairs} = \frac{e^{-1.1 d_{rs}}}{1 + e^{-1.1 d_{rs}}}.$$

The probability of choosing the stairs decreases as the number of floors increases, reflecting the increasing effort required to walk up multiple floors. The computed probabilities are presented in Table 7.11.

Number of Floors d_{rs}	π^{rs}_{stairs}
1	0.250
2	0.100
3	0.0356
4	0.0121

Table 7.11: Probability of taking the stairs as a function of the number of floors.

Using these probabilities, the original OD table is now split into two separate tables: one for elevator users and one for stair users. The complete OD table, including both modes, is shown in Table 7.6. Applying the modal split, the OD flows are now divided into two separate tables. The OD table for elevator users is shown in Table 7.12.

	0	1	2	3	4	
0		0.0	450.9	9.2	0.0	460.1
1	75.2		0.0	0.0	4.0	79.2
2	26.3	0.0		0.0	0.0	26.3
3	56.8	0.0	0.0		0.1	56.9
4	69.5	11.3	0.0	1.7		82.5
	227.7	11.4	450.9	11.0	4.1	705.0

Table 7.12: OD table for elevator users.

Similarly, the OD table for stair users is presented in Table 7.13.

The parameter θ plays an important role in the choice model, as it determines how sensitive travelers are to differences in generalized cost. To

	0	1	2	3	4	
0		0.0	50.0	0.3	0.0	50.3
1	25.0		0.0	0.0	0.1	25.2
2	2.9	0.0		0.0	0.0	2.9
3	2.1	0.0	0.0		0.0	2.1
4	0.9	0.4	0.0	0.6		1.9
	30.9	0.4	50.0	0.9	0.2	82.4

Table 7.13: OD table for stair users.

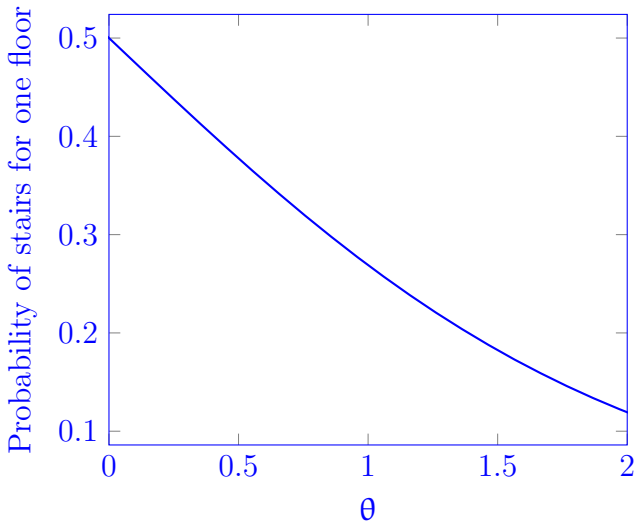


Figure 7.8: Impact of the parameter θ on the probability of choosing the stairs for a one-floor trip.

illustrate its effect, Figure 7.8 shows how the probability of choosing the stairs for a one-floor trip changes as θ varies.

The horizontal axis represents the value of θ, while the vertical axis shows the probability of choosing the stairs for a one-floor trip. When $\theta = 0$, both the elevator and stairs have equal probability, meaning travelers do not differentiate between the two options. As θ increases, the probability of taking the stairs decreases, reflecting a greater reluctance to walk up even a single floor.

The value of θ is not chosen arbitrarily but must be estimated from real-world data. As discussed in Section 4.3.3, this is typically done using maximum likelihood estimation, which finds the value of θ that best explains

observed mode choices. By fitting the model to empirical data, we ensure that the predicted probabilities align with actual travel behavior.

7.2.2 Choice data

To estimate choice models, we rely on different types of choice data. Two primary sources of choice data are *revealed preferences* (RP) and *stated preferences* (SP), each offering unique insights and limitations.

Revealed preference (RP) data consists of observations of actual choices made by travelers in real-world settings. These choices reflect genuine behavior, making RP data essential for reproducing observed modal shares. To utilize RP data effectively, researchers must also collect data on explanatory variables, including attributes associated with each mode such as travel time and cost. However, RP data comes with several limitations. It is restricted to existing transportation options, attributes, and attribute levels, making it difficult to evaluate hypothetical or future scenarios. Some attributes exhibit little variability, limiting the ability to estimate their effect accurately. Additionally, high correlation among variables complicates the identification of independent effects. Data collection is costly and often requires extensive travel surveys or sensor-based monitoring. Another significant challenge is that information on unchosen alternatives is often incomplete or missing, as researchers may only observe the mode that was selected but not those that were considered.

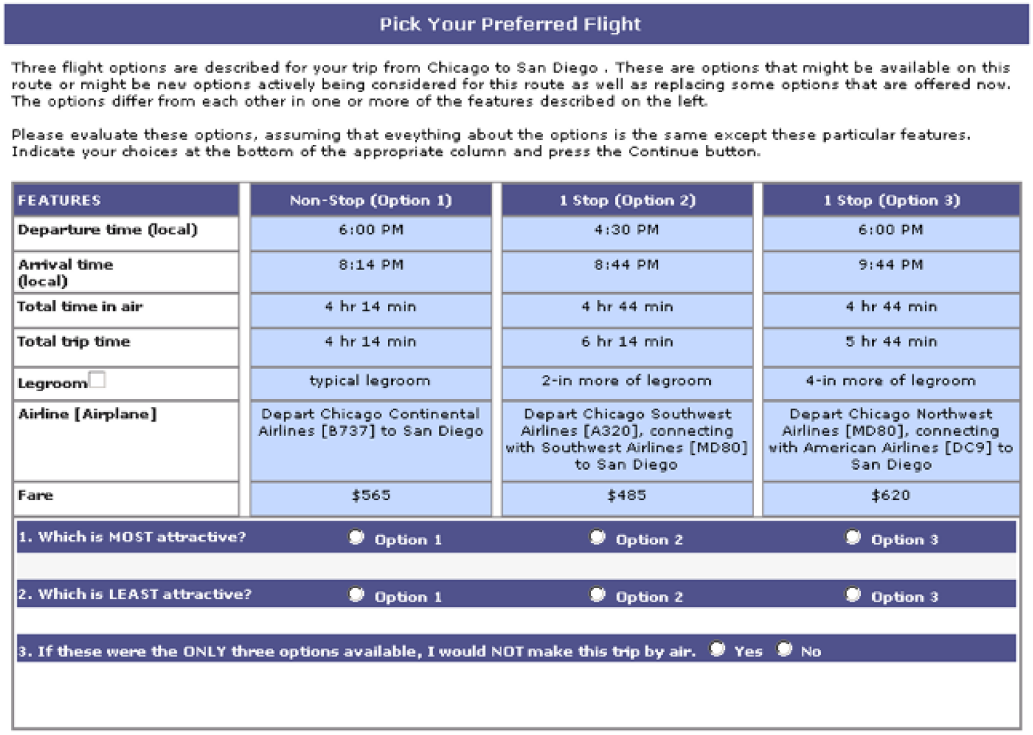
To overcome some of these limitations, researchers often rely on stated preference (SP) data, which is collected through surveys and interviews. In SP surveys, individuals are presented with hypothetical scenarios and asked what they would choose under specified conditions. This approach allows analysts to define the choice context, controlling for the attributes and their variability. By systematically modifying the presented alternatives, researchers can study factors that are difficult to isolate in RP data. SP data enables the exploration of new transportation modes and scenarios that do not yet exist. Researchers can control attribute variability, ensuring sufficient variation for model estimation. All available alternatives can be explicitly presented, overcoming the issue of missing data on unchosen options. Correlation between variables can be managed, avoiding multicollinearity issues, and a single respondent can answer multiple choice questions, increasing the quantity of collected data.

Figure 7.9 presents an example of a stated preference scenario designed to evaluate passengers' choices when selecting among different flight alternatives. It originates from an Internet choice survey conducted by Boeing Commercial Airplanes in 2004 and 2005. The survey is described by Gar-

row et al. (2006). Boeing aimed to better understand the sensitivity of air passengers to various airline itinerary attributes, including *fare*, *travel time*, *transfers*, *legroom*, and *aircraft type*.

The survey was conducted by intercepting users of an Internet airline booking service specializing in low-cost travel deals. While waiting for the search engine to return the real itineraries matching their specific travel request, randomly selected customers were invited to complete a survey tailored to their chosen origin and destination. This methodology ensured that respondents were actively engaged in making travel decisions, thereby increasing the relevance and reliability of the collected data.

The scenario in Figure 7.9 considers a journey from Chicago to San Diego, offering three distinct flight options with varying characteristics.



Pick Your Preferred Flight

Three flight options are described for your trip from Chicago to San Diego . These are options that might be available on this route or might be new options actively being considered for this route as well as replacing some options that are offered now. The options differ from each other in one or more of the features described on the left.

Please evaluate these options, assuming that everything about the options is the same except these particular features. Indicate your choices at the bottom of the appropriate column and press the Continue button.

FEATURES	Non-Stop (Option 1)	1 Stop (Option 2)	1 Stop (Option 3)
Departure time (local)	6:00 PM	4:30 PM	6:00 PM
Arrival time (local)	8:14 PM	8:44 PM	9:44 PM
Total time in air	4 hr 14 min	4 hr 44 min	4 hr 44 min
Total trip time	4 hr 14 min	6 hr 14 min	5 hr 44 min
Legroom	typical legroom	2-in more of legroom	4-in more of legroom
Airline [Airplane]	Depart Chicago Continental Airlines [B737] to San Diego	Depart Chicago Southwest Airlines [A320], connecting with Southwest Airlines [MD80] to San Diego	Depart Chicago Northwest Airlines [MD80], connecting with American Airlines [DC9] to San Diego
Fare	$565	$485	$620

1. Which is MOST attractive? Option 1 Option 2 Option 3

2. Which is LEAST attractive? Option 1 Option 2 Option 3

3. If these were the ONLY three options available, I would NOT make this trip by air. Yes No

Figure 7.9: Example of a stated preference scenario

The first alternative is a *non-stop flight*, providing the convenience of direct travel without layovers. The second alternative involves *one stop on the same airline*, requiring a connection but maintaining consistency in service. The third alternative also involves *one stop*, but with *multiple airlines*, which may introduce variations in service quality, scheduling, and ticketing conditions.

For each of these alternatives, several key attributes are presented to the respondent, allowing for a detailed comparison. These attributes include *departure time*, *arrival time*, *total time in air*, and *total trip time*, which together provide insights into the overall travel duration. Comfort-related aspects such as *legroom*, as well as operational details like *airline* and *aircraft type*, are also specified. Finally, the *fare* is displayed, representing an important factor in the decision-making process.

Participants in the survey are asked three questions. First, they must identify the *most attractive* flight option, reflecting their preferred choice based on the presented attributes. Next, they indicate the *least attractive* option, highlighting the alternative they find least desirable. Finally, they answer a broader question: *If these were the only options available, would you travel by air?* This last question helps assess the extent to which the given options meet the minimum acceptable requirements for air travel or if respondents would consider alternative transportation modes.

Despite these benefits, SP data also presents notable challenges. Because it is based on hypothetical scenarios, responses may not fully reflect actual behavior. It cannot be used to directly infer market shares, as the values of the explanatory variables presented in a survey do not necessarily translate to real-world scenarios. Respondents may not fully consider the consequences of their choices, as they do not bear real costs or constraints. The credibility of the scenarios is important — respondents may react differently if they perceive a scenario as unrealistic. Results are valid only within the range of the experimental design, limiting their generalizability. Policy bias can occur, where respondents state preferences that align with social expectations rather than their actual behavior, for example, stating that others should take the bus while continuing to drive themselves. Justification bias or inertia may lead respondents to choose familiar options rather than considering new ones. The way questions are phrased, known as framing effects, can influence responses. Respondents may anchor their decisions on a single variable, distorting the results. Finally, repeated questioning can lead to fatigue effects, where respondents provide less thoughtful answers over time.

Given these complementary strengths and weaknesses, both RP and SP data play an important role in estimating choice models. RP data is essential for capturing actual market behavior, while SP data is valuable for investigating new policies, services, or infrastructure changes. The two data sources are often combined to enhance model robustness. Estimation is typically conducted using maximum likelihood estimation, ensuring that model parameters best fit the observed choices.

7.2.3 Behavioral heterogeneity

In reality, individuals exhibit diverse behaviors and preferences when making travel choices. No two travelers are identical, as their decisions are influenced by various personal and contextual factors. Choice models aim to capture this *heterogeneity* by considering differences across individuals. Instead of treating the population as a uniform entity, these models are *disaggregate*, meaning that they account for individual-level variations rather than relying on aggregate trends.

To represent this heterogeneity, the population is divided into segments based on socio-economic characteristics. These segments allow models to reflect variations in travel behavior across different groups. A typical segmentation includes factors such as trip purpose, which distinguishes between work commutes, leisure trips, and shopping activities, as different trip types may lead to different mode choices. Gender can influence preferences, with research showing differences in risk perception, safety concerns, and travel patterns between men and women. Income plays an important role, as individuals with higher financial resources may have greater access to private vehicles and be less sensitive to fare changes in public transportation. Age is another key factor, with younger individuals possibly favoring active modes of transport, while older individuals may prioritize comfort and accessibility. Employment status affects travel behavior, as employed individuals often have structured schedules and constraints that influence their mode selection. The availability of mobility tools, such as driver's licenses, public transport passes, or access to bicycles, further differentiates individuals' travel behavior.

By incorporating these socio-economic characteristics into choice models, we can better understand the diverse decision-making processes of travelers.

Although the four-step model primarily deals with aggregate flows, it can still incorporate disaggregate choice models. This is possible because the results of a disaggregate model can be aggregated in a straightforward manner to align with the aggregate nature of the four-step model.

To show this, we assume that the population is divided into N distinct segments based on socio-economic characteristics. For a given OD pair (r, s), the proportion of individuals belonging to segment n is denoted as π_n^{rs}. Within each segment, individuals make mode choices according to their specific preferences and constraints. The probability that an individual in segment n chooses mode i for the OD pair (r, s) is denoted as π_{in}^{rs}.

To obtain an aggregate mode choice probability, we compute the weighted sum of the segment-specific probabilities, where each segment's contribution is weighted by its proportion in the population. This aggregation is expressed

mathematically as:

$$\pi_i^{rs} = \sum_n \pi_{in}^{rs} \pi_n^{rs}.$$

This equation ensures that the overall mode choice probabilities reflect the diversity of individual choices while remaining compatible with the aggregate framework of the four-step model.

In the following example, the modal split for the elevator scenario is further refined by incorporating population heterogeneity. Specifically, we assume that individuals differ in their reluctance to take the stairs based on their age. The population is divided into two segments: *young* and *old*, each with different sensitivity to the number of floors they need to climb.

For a given OD pair (r, s), the probability of taking the stairs is modeled separately for young and old individuals. The disaggregate choice model assumes that young individuals have a lower sensitivity parameter, $\theta_{\text{young}} = 1.1$, meaning they are more willing to take the stairs compared to older individuals, whose sensitivity parameter is higher, $\theta_{\text{old}} = 2.1$. The probability of choosing the stairs decreases with the number of floors, but it decreases at a much faster rate for older individuals.

The proportions of individuals choosing the stairs in each segment are shown in Table 7.14.

Number of Floors d_{rs}	$\pi^{rs}_{\text{stairs, young}}$	$\pi^{rs}_{\text{stairs, old}}$
1	0.250	0.110
2	0.100	0.0148
3	0.0356	0.00183
4	0.0121	0.000225

Table 7.14: Probability of taking the stairs by age segment.

To obtain the aggregate modal split for the total population, we compute a weighted sum of the segment-specific probabilities, where each segment's contribution is weighted by its proportion in the population. In this case, we assume that the population consists of 25% young individuals and 75% old individuals. The overall probability of taking the stairs for a given OD pair is therefore computed as:

$$\pi^{rs}_{\text{stairs}} = 0.25 \cdot \pi^{rs}_{\text{stairs, young}} + 0.75 \cdot \pi^{rs}_{\text{stairs, old}}.$$

Applying this formula, the aggregate probabilities are obtained as shown in Table 7.15.

Number of Floors $\mathtt{d}_{\mathtt{rs}}$	$\pi^{\mathrm{rs}}_{\mathrm{stairs}}$
1	0.144
2	0.0360
3	0.0103
4	0.003

Table 7.15: Aggregate probability of taking the stairs.

These probabilities are then applied to the original OD table to divide the trips between those taking the elevator and those taking the stairs, as explained above.

7.3 Summary

This chapter has explored two steps of the four-step model: trip distribution and modal split. Through a detailed examination of OD table estimation and mode choice modeling, we have developed a framework for understanding how trips are distributed across a network and how individuals select their preferred mode of transportation.

The estimation of OD tables relies on multiple sources of information, including production and attraction data for different zones. To enhance the accuracy of these estimates, additional data sources such as roadside interviews and traffic counts can be incorporated. Since observed data is often incomplete or inconsistent, further assumptions are necessary to structure the estimation process. One common assumption is the *gravity model*, which expresses trip flows as a function of trip productions, attractions, and generalized cost. The estimation problem is then formulated as a *weighted least squares* problem, ensuring that different data sources are accounted for with appropriate emphasis. Moreover, special attention must be given to the issue of non-negativity, as OD flows must remain non-negative, requiring adjustments to the standard least squares approach.

Once the OD table is established, the modal split step determines how these trips are distributed among available transportation modes. This process requires choice models that capture individual decision-making. The data used for mode choice modeling comes from two primary sources: *revealed preferences* (RP), which reflect actual observed choices, and *stated preferences* (SP), which derive from surveys presenting hypothetical scenarios. The mode choice model itself is typically based on a probabilistic framework such as the *logit model*, where the probability of selecting a given mode

depends on its perceived utility. Because choice models operate at the individual level, an aggregation step is necessary to reconcile disaggregate choices with the aggregate structure of the four-step model.

At this stage, three of the four steps in the model have been addressed: trip generation, trip distribution, and modal split. The final step, *traffic assignment*, remains to be explored. In the next chapter, we will examine how trips are assigned to the transportation network, determining the routes taken by travelers and the resulting network flows. This final step is important for understanding congestion effects, travel times, and overall system performance.

Chapter 8

Traffic assignment

The four-step model consists of four sequential phases: trip generation, trip distribution, modal split, and finally, traffic assignment. The first three steps, which we have already covered, aim to estimate the number of trips originating from and destined for different zones, allocate these trips between zones, and determine the mode of transportation chosen by travelers. The final step, traffic assignment, builds upon these results to determine how trips are distributed across the transportation network.

The objective of the traffic assignment phase is to determine *link flows*, meaning the number of vehicles traveling on each road segment in the network. This information is important for evaluating congestion, travel times, and the overall performance of the transportation system. Traffic assignment models rely on principles of route choice behavior, typically assuming that travelers choose paths that minimize travel costs, such as time or distance.

The context of this analysis is limited to *single-mode* transportation, specifically private car traffic. This assumption simplifies the assignment problem by focusing solely on road networks without considering interactions between different transport modes. While multimodal assignment methods exist, the fundamental principles of traffic assignment are best understood within the framework of a single-mode system.

At this stage of the modeling process, two key datasets are available. First, the *origin-destination (OD) table* provides trip demand information between different zones or centroids. This OD table specifies the number of trips, denoted as $\mathtt{f}_{\mathtt{rs}}$, between each pair of zones $(\mathtt{r}, \mathtt{s})$.

The second dataset consists of the *transportation network*, which includes information about road links and their characteristics. Each link has an associated *link performance function*, $\mathtt{t}_\ell = \mathtt{t}(\mathtt{x}_\ell)$, which expresses travel time as a function of traffic flow. These functions capture congestion effects by reflecting how travel times increase with higher traffic volumes. Additionally,

the network is represented using the *link-path incidence matrix*, denoted as P, which encodes the relationship between links and paths. The elements of this matrix, $\mathsf{P}_{\ell p}$, indicate whether a given link ℓ belongs to a particular path p.

8.1 All-or-nothing assignment

A fundamental assumption in traffic assignment is that travelers choose routes based on the principle of *utility maximization*. This means that each individual seeks to minimize their perceived travel cost, often interpreted as choosing the *shortest* or *best* path available. In the simplest case, this best path is defined by travel time: travelers are assumed to select the route that minimizes their time spent on the road.

Figure 8.1 illustrates this concept using a simple two-link network connecting an origin node, r, to a destination node, s. Travelers have two route options, with associated travel times t_1 and t_2. If we assume that all travelers strictly minimize their travel time and that $t_1 < t_2$, then all traffic will be assigned to the first route.

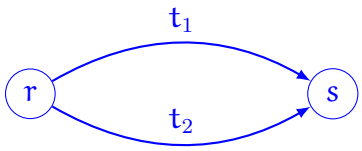


Figure 8.1: Illustration of route choice behavior in a simple two-link network.

An important consideration is that not all travelers behave identically. Differences in preferences, risk tolerance, and perceived travel costs lead to variations in route choice behavior. Some individuals may prioritize reliability over speed, while others may consider factors such as road comfort or toll costs. Addressing these aspects is out of the scope of this document.

Another important limitation of this naive route choice assumption is that it does not account for *congestion effects*. In reality, as more travelers choose the shortest route, traffic density increases, leading to a rise in travel time. If congestion is severe enough, the second route may become a more attractive option, as its initially higher travel time remains stable while the first route deteriorates.

The concept of *all-or-nothing assignment* assumes that all travelers select the single fastest route available, without considering congestion effects. Under this assumption, every unit of flow is assigned to the path with the

lowest travel time in free-flow conditions. However, this approach oversimplifies real-world traffic behavior and leads to unrealistic outcomes.

Figure 8.2 illustrates an example where three units of flow travel from an origin node r to a destination node s. Two possible routes exist: the first route has an initial travel time of 2, while the second has an initial travel time of 4. Since the first route is the fastest in free-flow conditions, all three units of flow are assigned to it, as dictated by the all-or-nothing principle.

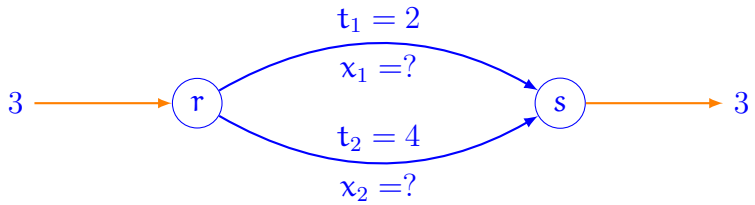


Figure 8.2: Illustration of all-or-nothing assignment in a simple two-link network.

However, this naive approach neglects the fact that travel time is not static but depends on the volume of traffic on each link. To better represent congestion effects, we introduce *link performance functions*, which describe how travel time changes with increasing flow. For this example, the travel time functions are given by:

$$t_1(x) = 2 + x_1^2,$$
$$t_2(x) = 4 + 2x_2^2.$$

These equations indicate that travel time increases as flow increases, reflecting the fundamental relationship between congestion and travel conditions.

At the start, under free-flow conditions, we have

$$t_1(0) = 2,$$
$$t_2(0) = 4.$$

The first route being the fastest, all travelers select it: $x_1 = 3$, $x_2 = 0$. However, as congestion builds, the travel time on link 1 increases, eventually reaching a point where the second route, despite its higher initial travel time, becomes the faster option.

$$t_1(3) = 11,$$
$$t_2(0) = 4.$$

At this stage, all traffic shifts to link 2, and $x_1 = 0$ and $x_2 = 3$. But the same congestion effect occurs: the increasing flow raises travel time, making

the first route more attractive again.

$$\mathsf{t}_1(0) = 2,$$
$$\mathsf{t}_2(3) = 22.$$

This back-and-forth redistribution of flow highlights the fundamental flaw of all-or-nothing assignment — it does not lead to a stable state.

A more sophisticated assignment model is needed, one that acknowledges that travelers respond to congestion in a way that balances the system. Instead of assuming that all traffic blindly follows a single path, we must develop models that distribute flow in a way that accounts for travel time dynamics and traveler behavior.

8.2 User equilibrium

A more refined approach to traffic assignment involves loading the flow onto the network incrementally, updating travel times at each step. This method provides insight into how congestion builds up and how travelers dynamically adjust their route choices in response to increasing travel times.

Figure 8.3 presents the process of loading flow one unit at a time, recalculating travel times after each step. Initially, when the network is empty, both links have their free-flow travel times: link 1 has a travel time of 2, while link 2 has a travel time of 4. Since link 1 is the faster option, the first unit of flow is assigned to it.

	x_1	t_1	x_2	t_2	Choice
Empty network	0	2	0	4	$\ell = 1$
First unit	1	3	0	4	$\ell = 1$
Second unit	2	6	0	4	$\ell = 2$
Third unit	2	6	1	6	Equilibrium

Figure 8.3: Incremental loading of flow and resulting travel times.

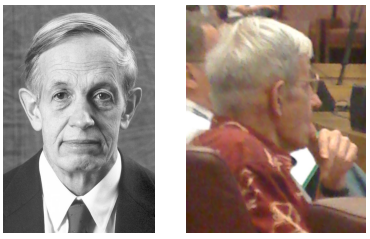
Once the first unit of flow is loaded onto link 1, congestion increases, and the travel time on this link rises from 2 to 3. The second unit of flow is then added, further increasing the travel time on link 1 to 6, while link 2 remains at 4. At this point, link 2 becomes the preferable route, and the third unit of flow chooses this alternative instead. After assigning one unit to link 2, its travel time also increases, reaching 6. Now, both links have equal travel

times, meaning no traveler has an incentive to switch routes. This condition represents *equilibrium*, where travel demand is balanced, and no individual traveler can improve their travel time by unilaterally changing routes.

The concept of equilibrium in traffic assignment has strong theoretical foundations in game theory. The idea that no traveler can improve their travel time by unilaterally changing routes was formalized by the mathematician John Forbes Nash Jr., whose contributions to the field of game theory earned him the Nobel Prize in Economics in 1994 . His groundbreaking work (Nash, 1950a) laid the foundation for the study of strategic interactions, not only in economics but also in transportation and many other disciplines.

Nash was born in 1928 and made significant contributions to mathematics early in his career. His doctoral thesis (Nash, 1950b), completed in 1950 at Princeton University, introduced the concept of *Nash equilibrium*, a fundamental principle in game theory. Remarkably, his thesis on non-cooperative games was only 28 pages long, yet it revolutionized the understanding of competitive behavior in strategic environments.

Figure 8.4 presents images of John Nash, including one taken by the author at a conference in Lisbon in 2010. His work provided a rigorous framework for analyzing situations in which multiple decision-makers interact, each seeking to optimize their own outcome given the choices of others. In the context of traffic assignment, Nash equilibrium describes a state where no individual traveler can reduce their travel time by switching routes, assuming that all other travelers maintain their current choices.



nobelprize.org Lisbon, 2010

Figure 8.4: John Forbes Nash Jr., Nobel Laureate in 1994.

The Nash equilibrium concept is particularly relevant in transportation modeling because it accounts for the fact that travelers make independent route choices while responding to congestion effects. Unlike the all-or-nothing approach, which unrealistically assumes that all travelers choose the same route, the equilibrium concept allows for a more stable and self-regulating distribution of traffic.

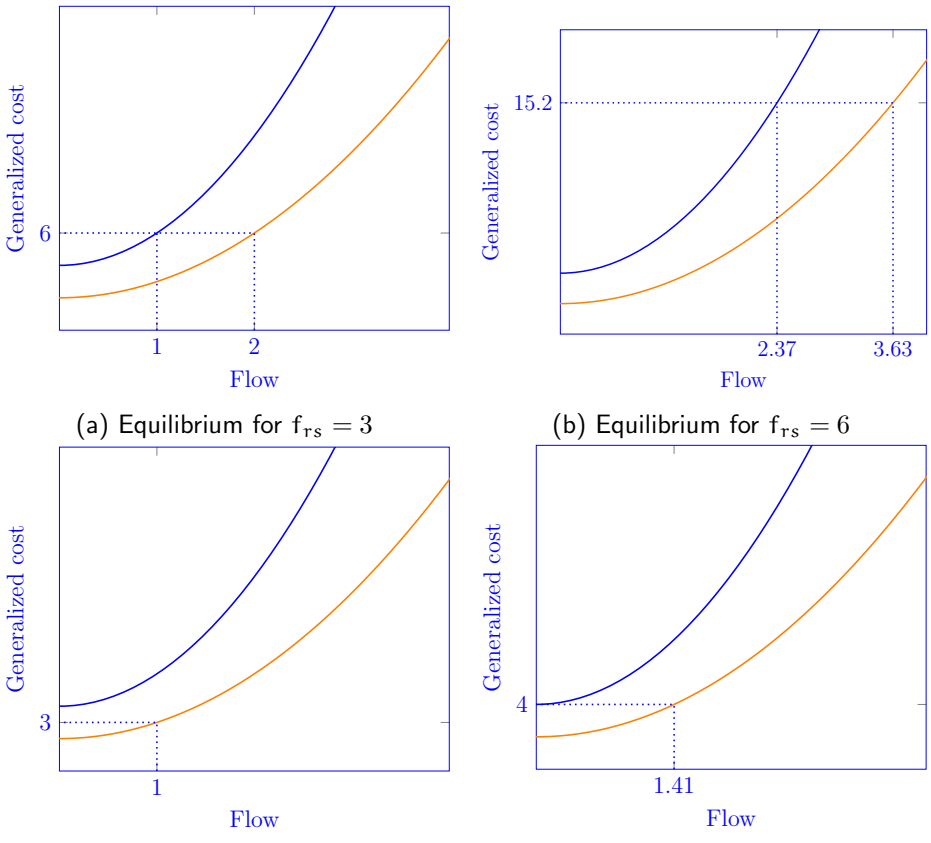
At equilibrium, no traveler can improve their travel time by switching

routes, provided all others maintain their choices. This condition leads to what is commonly referred to as *user equilibrium*, where all used routes have the same generalized cost, and any unused route has a higher cost.

Figure 8.5 illustrates how equilibrium is reached for different levels of total demand f_{rs}. The travel time functions governing the two available links are given by:

$$t_1(x) = 2 + x^2,$$
$$t_2(x) = 4 + 2x^2.$$

As demand increases, congestion builds up, leading to different equilibrium conditions.



(a) Equilibrium for $f_{rs} = 3$

(b) Equilibrium for $f_{rs} = 6$

(c) Equilibrium for $f_{rs} = 1$

(d) Equilibrium for $f_{rs} = \sqrt{2}$

Figure 8.5: Nash equilibrium for different levels of demand.

For $f_{rs} = 3$, as shown in Figure 8.5a, congestion causes link 1's travel time to rise above the free-flow travel time of link 2. As a result, travelers begin using both links, distributing the flow so that travel time equalizes on both routes.

For $f_{rs} = 6$, as shown in Figure 8.5b, a higher level of demand results in even greater congestion, further balancing the flow between the two links. The equilibrium principle ensures that both links maintain the same generalized cost.

For $f_{rs} = 1$, as shown in Figure 8.5c, the total flow is low enough that link 1 remains the fastest option. Since its travel time does not exceed the free-flow travel time of link 2, all travelers continue to choose link 1.

For $f_{rs} = \sqrt{2}$, as shown in Figure 8.5d, this is the highest demand level at which all travelers still select link 1. Beyond this threshold, congestion increases enough that link 2 starts attracting some of the flow.

To illustrate the concept of traffic assignment in a more realistic setting, we now consider a small network example, as shown in Figure 8.6. This network consists of six nodes, including two centroids that represent origins and destinations, and two regular nodes that serve as intersections where traffic flow can be redistributed.

The network contains seven directed arcs, each associated with a *link performance function* that describes how travel time depends on congestion. These functions range from constant travel times, which are insensitive to congestion, to those that increase significantly as flow grows, reflecting different road characteristics.

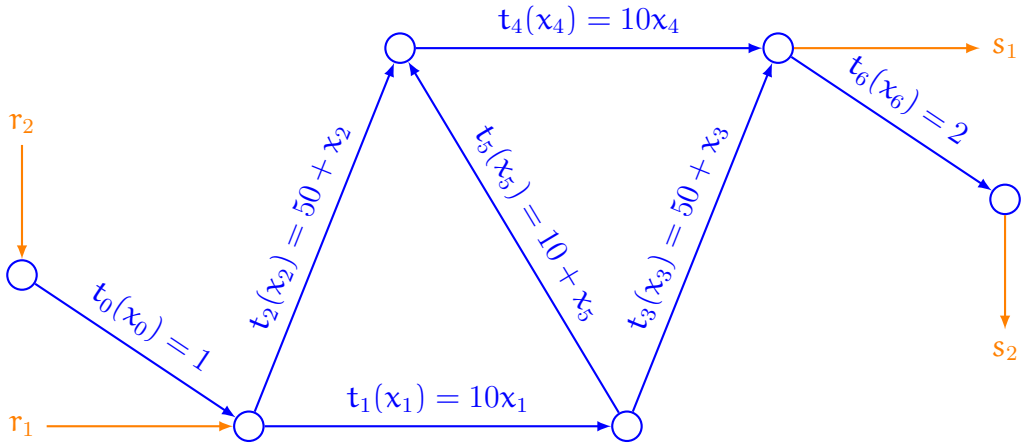


Figure 8.6: A small network example with congestion-sensitive and insensitive arcs.

Traffic enters the network from the left, where two units of flow originate from node r_2 and one unit from node r_1, resulting in a total of three entering units. The flow exits on the right, with two units reaching node s_1 and one unit reaching node s_2. This demand pattern is summarized in Table 8.1.

The network consists of different types of arcs with varying sensitivity to congestion. *Arcs 0 and 6* have a constant travel time, meaning they are

	s_1	s_2
r_1	3	1
r_2	2	0

Table 8.1: Origin-destination demand table.

not affected by congestion. *Arcs 2 and 3* have a relatively long free-flow travel time, but their sensitivity to congestion is low, resembling highways with multiple lanes that bypass the city center. *Arc 5* has an intermediate free-flow travel time and a low sensitivity to congestion, similar to a tunnel under the city with as many lanes as the highways. Finally, *arcs 1 and 4* have negligible free-flow travel times (actually, zero, in this simple example) but are highly sensitive to congestion, representing shortcuts through narrow streets that quickly become congested.

Each origin-destination pair has three available paths, depicted schematically in Figure 8.7. The first path follows the northern detour, which corresponds to the long highways bypassing congestion. The second path follows the southern detour, taking an alternative long route. The third path goes through the tunnel under the city, which offers a shorter distance.

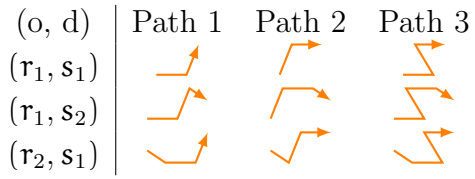


Figure 8.7: List of paths in the simple network

Table 8.2 summarizes the characteristics of the different paths available for each origin-destination (OD) pair in the network when it is empty, meaning that no vehicles are initially using it. The first column of the table lists the different paths, while the second column indicates the path flow. Since the network is empty, all path flows are equal to zero.

The next group of columns represents the link flows x_i, where each column corresponds to a specific arc in the network. These values indicate the number of vehicles traveling on each link when the given path is used. Since the path flows are zero, the link flows are also zero.

Following the link flows, the table presents the link travel times t_i, which are computed using the link performance functions. As the network is empty, the travel times correspond to the free-flow travel times, meaning the minimum time required to traverse each link when there is no congestion.

p	flow	x_0	x_1	x_2	x_3	x_4	x_5	x_6	t_0	t_1	t_2	t_3	t_4	t_5	t_6	cost
r_1, s_1:	$f_{rs} = 3$															
	0		0		0					0		50				50
	0			0		0					50		0			50
	0		0			0	0			0			0	10		10
r_1, s_2:	$f_{rs} = 1$															
	0		0		0			0		0		50			2	52
	0			0		0		0			50		0		2	52
	0		0			0	0	0		0			0	10	2	12
r_2, s_1:	$f_{rs} = 2$															
	0	0	0		0				1	0		50				51
	0	0		0		0			1		50		0			51
	0	0	0			0	0		1	0			0	10		11

Table 8.2: Path flows, link flows, link travel times, and path costs for each OD and each path in an empty network.

Finally, the last column provides the total path cost for each path, which is obtained by summing the travel times of all links composing the path. The values in this column indicate the total time required to travel from the origin to the destination via each path under free-flow conditions.

From the results, it is evident that for each OD pair, the fastest available path is always the third option (paths  ,  , and ), which takes advantage of the tunnel under the city. This path consistently offers the shortest travel time compared to the alternative routes.

Table 8.3 summarizes the network conditions when all travelers select their routes based on the fastest option available in free-flow conditions. The structure of the table remains the same as in the previous case.

One important observation is that the link flows are computed as the cumulative effect of all origin-destination (OD) pairs. Notably, for each link, the values are identical across all rows where the link is used. This repetition in the table simplifies the computation of link and path travel times by making it clear which links contribute to each path's total cost.

A consequence of this behavior is the significant congestion that emerges in the streets leading to the tunnel. Although the tunnel itself remains the fastest option in free-flow conditions, its accessibility becomes a bottleneck when demand increases. As a result, the total travel time along the paths that use the tunnel ( ,  , and ) is now considerably higher than in the previous scenario. The congestion on the feeder roads slows

p	flow	x_0	x_1	x_2	x_3	x_4	x_5	x_6	t_0	t_1	t_2	t_3	t_4	t_5	t_6	cost
r_1, s_1:	$f_{rs} = 3$															
	0		6		0					60		50				110
	0			0		6					50		60			110
	3		6			6	6			60			60	16		136
r_1, s_2:	$f_{rs} = 1$															
	0		6		0			1		60		50			2	112
	0			0		6		1			50		60		2	112
	1		6			6	6	1		60			60	16	2	138
r_2, s_1:	$f_{rs} = 2$															
	0	2	6		0				1	60		50				111
	0	2		0		6			1		50		60			111
	2	2	6			6	6		1	60			60	16		137

Table 8.3: Path flows, link flows, link travel times, and path costs for each OD and each path when all travelers follow the fastest route based on free-flow travel time.

down travelers before they even reach the tunnel, effectively diminishing its advantage. This highlights a fundamental concept in transportation: the shortest route in free-flow conditions is not necessarily the most efficient once congestion is considered.

At equilibrium, as presented in Table 8.8, the system reaches a state where no traveler can unilaterally reduce their travel time by switching to another route. All used paths connecting an origin-destination pair must have equal and minimal travel costs. If any path had a strictly lower cost, travelers would shift toward it until the costs equilibrate.

Table 8.9 presents a second equilibrium state, highlighting an essential property of network equilibria: while the distribution of travelers among paths can vary, the resulting link flows remain unchanged. This observation demonstrates that the equilibrium conditions do not uniquely determine the path flows but (under some conditions) may uniquely determine the link flows and travel times.

8.3 Modeling

To develop a rigorous mathematical model for traffic assignment, we first establish a set of notations that will be used throughout the formulation.

The network consists of a set of links, each representing a segment of infrastructure such as a road or a highway section. The total number of links

p	flow	x_0	x_1	x_2	x_3	x_4	x_5	x_6	t_0	t_1	t_2	t_3	t_4	t_5	t_6	cost
r_1, s_1:	$f_{rs} = 3$															
	1		4		2					40		52				92
	1			2		4					52		40			92
	1		4			4	2			40			40	12		92
r_1, s_2:	$f_{rs} = 1$															
	1		4		2			1		40		52			2	94
	0			2		4		1			52		40		2	94
	0		4			4	2	1		40			40	12	2	94
r_2, s_1:	$f_{rs} = 2$															
	0	2	4		2				1	40		52				93
	1	2		2		4			1		52		40			93
	1	2	4			4	2		1	40			40	12		93

Figure 8.8: Path flows, link flows, link travel times, and path costs at equilibrium.

in the network is denoted by K^{ℓ}. Travelers move between different origin-destination (OD) pairs using predefined paths, where the total number of paths is given by K^{p}. Since multiple OD pairs exist in a transportation network, we also introduce K^{rs} to denote the number of OD pairs considered in the model. For a specific OD pair q, the set of available paths is represented by $\mathcal{P}_q$.

Each link in the network carries a certain traffic volume, which we define as the *link flow* and denote by x. This represents the number of vehicles or travelers using a specific link over a given period. Similarly, each path carries a flow of travelers, referred to as the *path flow* and denoted by y.

In addition to flows, each link has an associated *link cost*, denoted by t. This cost generally represents the travel time experienced by users on that particular link, which may depend on the level of congestion. Likewise, each path has an associated *path cost*, denoted by c, which corresponds to the sum of the link costs along the path.

A key component of the traffic assignment model is the link-path incidence matrix, denoted by P, which establishes the relationship between the network's links and the available paths (see Section 6.5). This matrix, of dimension $K^{\ell} \times K^{p}$, is defined as a binary matrix where each entry P_{lp} takes the value 1 if link l is part of path p, and 0 otherwise.

Another essential element in the formulation is the route choice matrix, denoted by R, which captures the distribution of travelers across available paths for each origin-destination (OD) pair. The matrix R has dimensions

p	flow	x_0	x_1	x_2	x_3	x_4	x_5	x_6	t_0	t_1	t_2	t_3	t_4	t_5	t_6	cost
r_1, s_1:	$f_{rs} = 3$															
	1		4		2					40		52				92
	0			2		4					52		40			92
	2		4			4	2			40			40	12		92
r_1, s_2:	$f_{rs} = 1$															
	1		4		2			1		40		52			2	94
	0			2		4		1			52		40		2	94
	0		4			4	2	1		40			40	12	2	94
r_2, s_1:	$f_{rs} = 2$															
	0	2	4		2				1	40		52				93
	2	2		2		4			1		52		40			93
	0	2	4			4	2		1	40			40	12		93

Figure 8.9: Alternative equilibrium illustrating different path flows but identical link flows.

$K^p \times K^{rs}$, where each entry R_{pq} represents the proportion of travelers associated with OD pair q who choose path p. The sum of all elements in a given column of R must equal 1, ensuring that all travelers for a given OD pair are assigned to one of the available paths.

The route choice matrix plays an important role in translating demand into network flows. Since each path is uniquely associated with a single OD pair, the presence of a zero in R_{pq} signifies that path p is not an option for OD pair q. For certain assignment models, such as all-or-nothing assignment, the entries of R are restricted to binary values, meaning that all travelers from an OD pair use the same path. In more general cases, R contains continuous values representing the proportion of travelers choosing each path, as is the case in stochastic user equilibrium models where travelers distribute themselves probabilistically based on perceived costs.

An origin-destination (OD) demand table can be transformed into path flows, which are then used to determine network congestion and travel times. This transformation is achieved using the route choice matrix R, which specifies the proportion of travelers for each OD pair who choose a given path. Given the demand vector f, where each entry f_q represents the number of travelers for OD pair q, the path flow vector y is computed as $y = Rf$.

The example provided illustrates this transformation by applying the

route choice matrix R to the OD demand vector f:

$$\mathsf{y} = \mathsf{R}\mathsf{f}: \quad \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/3 & 0 & 0 \\ 1/3 & 0 & 0 \\ 1/3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1/2 \\ 0 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}$$

$$\mathsf{y}_\mathsf{p} = \sum_\mathsf{q} \mathsf{R}_{\mathsf{p}\mathsf{q}} \mathsf{f}_\mathsf{q}, \forall \mathsf{p}.$$

The resulting path flow vector y captures the number of travelers assigned to each path. Each entry in y is obtained by summing the contributions from all OD pairs using that path, weighted by the corresponding entry in R. Mathematically, this relationship is expressed as

$$\mathsf{y}_\mathsf{p} = \sum_\mathsf{q} \mathsf{R}_{\mathsf{p}\mathsf{q}} \mathsf{f}_\mathsf{q}, \forall \mathsf{p},$$

ensuring that the flow along each path correctly reflects the demand distribution.

Having determined the path flows, the next step is to compute the link flows. This transformation is achieved using the link-path incidence matrix P. Given the path flow vector y, the link flow vector x is computed as $\mathsf{x} = \mathsf{P}\mathsf{y}$, where each entry x_ℓ represents the total number of travelers using link ℓ.

The example provided illustrates this transformation by applying the link-path incidence matrix P to the path flow vector y:

$$\mathsf{x} = \mathsf{P}\mathsf{y}: \quad \begin{pmatrix} 2 \\ 4 \\ 2 \\ 2 \\ 4 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

Each entry in x is obtained by summing the contributions from all paths that use link ℓ, as determined by the corresponding row in P:

$$\mathsf{x}_\ell = \sum_p \mathsf{P}_{\ell p} \mathsf{y}_p, \forall \ell.$$

A key simplification in traffic assignment modeling is the use of the assignment matrix, which is obtained as the product of the link-path incidence matrix P and the route choice matrix R. This transformation allows for a direct computation of link flows from OD demand without explicitly handling the intermediate path flows. Given the OD demand vector f, the link flow vector x is computed as $\mathsf{x} = \mathsf{PRf}$, or, equivalently,

$$\mathsf{x}_\ell = \sum_p \sum_q \mathsf{P}_{\ell p} \mathsf{R}_{pq} \mathsf{f}_q, \forall \ell.$$

The example provided illustrates this transformation by applying the assignment matrix $\mathsf{Q} = \mathsf{PR}$ to the OD demand vector f:

$$\mathsf{x} = \mathsf{PRf}: \begin{pmatrix} 2 \\ 4 \\ 2 \\ 2 \\ 4 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 2/3 & 1 & 1/2 \\ 1/3 & 0 & 1/2 \\ 1/3 & 1 & 0 \\ 2/3 & 0 & 1 \\ 1/3 & 0 & 1/2 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}$$

An important advantage of the assignment matrix Q is that, unlike P and R, it does not involve the number of paths K^p as one of its dimensions. Instead, it directly maps OD demand to link flows, reducing computational complexity. In real networks, the number of possible paths between each OD pair is extremely large, making it impractical to enumerate or store all paths explicitly. As a result, matrices P and R are typically not manipulated directly; instead, computations rely on the assignment matrix Q, which provides a more scalable representation of the system.

The link-path indicence matrix is useds also to transform link costs into the corresponding path costs. As discussed in Section 6.5, the total cost of a path is simply the sum of the costs of the links it traverses. This transformation is efficiently expressed using the transposed link-path incidence matrix, denoted as P^T. Given the vector of link costs t, the path cost vector c is computed as $\mathsf{c} = \mathsf{P}^\mathsf{T}\mathsf{t}$.

The example below illustrates this transformation by applying the transposed link-path incidence matrix to the link cost vector t:

$$\mathsf{c} = \mathsf{P}^\mathsf{T}\mathsf{t}: \begin{pmatrix} 92 \\ 92 \\ 92 \\ 94 \\ 94 \\ 94 \\ 93 \\ 93 \\ 93 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 40 \\ 52 \\ 52 \\ 40 \\ 12 \\ 2 \end{pmatrix}$$

Each entry in the path cost vector c is obtained by summing the costs of all links that belong to the corresponding path:

$$\mathsf{c}_\mathsf{p} = \sum_\ell \mathsf{P}_{\ell\mathsf{p}}\mathsf{t}_\ell.$$

As discussed above, the number of available paths can be huge, making direct manipulation of full matrices computationally inefficient. Therefore, it is often useful to extract portions of the link-path incidence and route choice matrices that correspond to a specific origin-destination (OD) pair. This allows for focused computations and facilitates a more structured analysis of how demand is distributed within individual OD pairs.

To achieve this, we introduce OD-specific submatrices. The OD-specific link-path incidence matrix, denoted by P^q, consists of the columns of the full link-path incidence matrix P that correspond to the paths available for OD pair q. This extraction isolates the relevant connections between links and paths for that particular OD pair. For example, in the case of OD pair 1 and OD pair 3, we have:

$$\mathsf{P}_1 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathsf{P}_3 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Similarly, we define the OD-specific route choice vector, denoted as R^q. This vector consists of the subset of rows and the single column of the full route choice matrix R that corresponds to OD pair q. This provides a direct

representation of how travelers for a given OD pair distribute their demand across available paths. For OD pairs 1 and 3, we have:

$$\mathsf{R}_1 = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}, \quad \mathsf{R}_3 = \begin{pmatrix} 0 \\ 1/2 \\ 1/2 \end{pmatrix}.$$

Once the OD-specific link-path incidence and route choice matrices have been defined, they can be used to compute both the path flows and the path costs for a given OD pair. These computations allow us to analyze how demand is distributed among available routes and to determine the travel costs associated with each route.

Path flows for a given OD pair q are computed using the OD-specific route choice vector R^{q}. Given the OD demand f_{q}, the path flow vector y^{q} is obtained as $\mathsf{y}^{\mathsf{q}} = \mathsf{R}^{\mathsf{q}}\mathsf{f}_{\mathsf{q}}$. This operation distributes the total OD demand among the available paths based on the route choice proportions. For example, the path flows for OD pairs 1 and 3 are computed as follows:

$$\mathsf{y}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} 3, \quad \mathsf{y}_3 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1/2 \\ 1/2 \end{pmatrix} 2.$$

This formulation ensures that the sum of the path flows matches the total OD demand while adhering to the route choice probabilities.

Similarly, path costs for OD pair q are computed using the transposed OD-specific link-path incidence matrix $(\mathsf{P}^{\mathsf{q}})^{\mathsf{T}}$. Given the link cost vector t, the path cost vector c^{q} is obtained as $\mathsf{c}^{\mathsf{q}} = (\mathsf{P}^{\mathsf{q}})^{\mathsf{T}}\mathsf{t}$. This operation aggregates the link costs along each path, providing the total cost incurred by travelers using that path. For OD pair 1, this computation is illustrated as follows:

$$\mathsf{c}_1 = \begin{pmatrix} 92 \\ 92 \\ 92 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 40 \\ 52 \\ 52 \\ 40 \\ 12 \\ 2 \end{pmatrix}.$$

A fundamental behavioral assumption in traffic assignment is that each traveler seeks to minimize their travel cost. This assumption implies that, when faced with multiple available paths between an origin-destination (OD) pair, travelers will choose the one with the lowest perceived cost.

To formalize this idea, we define the minimum cost for OD pair q as:

$$c_q^* = \min_p c_p^q, \quad \forall q.$$

To provide a structured summary of the key notations introduced so far, we present them in Table 8.4. This table organizes the different quantities related to flow and cost across three levels of the transportation network: individual links, complete paths, and origin-destination (OD) pairs.

	Links	Paths	OD pair
Flow	x_ℓ	y_p	f_q
Cost	t_ℓ	c_p	c_q^*

Table 8.4: Summary of flow and cost notations at different levels of the network.

Individual links are characterized by their flow x_ℓ, representing the total number of travelers using link ℓ, and their cost t_ℓ, which typically corresponds to the travel time on the link. At the path level, we define the path flow y_p, which represents the number of travelers using path p, and the corresponding path cost c_p, computed as the sum of the costs of the links composing the path. Finally, at the OD level, we introduce the OD demand f_q, which indicates the total number of travelers wishing to move between origin r_q and destination s_q, as well as the minimum OD cost c_q^*, representing the lowest path cost available for OD pair q.

The concept of traffic equilibrium is formalized through a set of mathematical conditions that govern how travelers distribute themselves across available paths in a network. These conditions ensure that no individual traveler can reduce their travel cost by unilaterally changing their route choice, leading to a stable and self-consistent assignment of traffic flows.

The first equilibrium condition states that for each OD pair q, the cost of traveling on any path p must be at least as high as the minimum cost c_q^* associated with that OD pair:

$$c_p^q \geqslant c_q^*, \quad \forall q.$$

The second condition enforces the idea that travelers only use paths that have the minimum cost. If a path p is used by travelers in OD pair q (i.e., $y_p^q > 0$), then its cost must be equal to c_q^*. If the path has a strictly higher cost, then it must carry no flow. This condition is expressed mathematically as:

$$y_p^q(c_p^q - c_q^*) = 0, \quad \forall p, q.$$

This condition guarantees that used paths have the lowest possible cost and that no traveler would benefit from switching to another route.

The third condition ensures that the total demand for each OD pair is fully assigned to the available paths:

$$\sum_{p} y_p^q = f_q, \quad \forall q.$$

This ensures that all travelers are accounted for and that there is no unallocated demand in the system.

Finally, we impose the natural requirement that path flows cannot be negative:

$$y_p^q \geqslant 0, \quad \forall p, q.$$

Since flow represents the number of travelers on a path, it must always be a non-negative quantity.

Together, these four conditions define a traffic equilibrium state in which travelers distribute themselves optimally according to the lowest available travel cost.

8.4 Beckmann's model

An alternative way to define traffic equilibrium is through an optimization framework known as Beckmann's model. This model formulates the equilibrium conditions as the solution to a convex optimization problem.

The objective function of Beckmann's model is given by:

$$\min_{y} \sum_{\ell} \int_0^{x_\ell} t_\ell(z) dz,$$

where the function $t_\ell(x_\ell)$ is the link performance function, that provides the travel cost on link ℓ as a function of its flow x_ℓ.

The optimization problem is subject to the following constraints:

$$\sum_{p} y_p^q = f_q, \quad \forall q, \quad \text{(demand conservation)}$$

which ensures that all OD demand is assigned to available paths, and:

$$y_p^q \geqslant 0, \quad \forall p, q, \quad \text{(non-negativity of flows)}$$

which guarantees that flows remain physically meaningful.

Additionally, the link flows x_ℓ are defined in terms of the path flows through the link-path incidence matrix P:

$$x_\ell = \sum_p P_{\ell p} y_p^q, \quad \forall \ell, q.$$

For this optimization model to be valid, it must satisfy two key assumptions:

$$\frac{\partial t_\ell(x_\ell)}{\partial x_\ell} > 0, \quad \forall \ell, \quad \text{(increasing cost function)}$$

which ensures that travel costs increase as congestion rises, and:

$$\frac{\partial t_\ell(x_\ell)}{\partial x_{\ell'}} = 0, \quad \forall \ell \neq \ell'.$$

This second assumption states that the cost on one link depends only on its own flow and not on the flow of other links, meaning that links operate independently in terms of congestion effects.

If these assumptions hold, then the optimal solution to the Beckmann optimization problem corresponds to a user equilibrium. This result is significant because it provides a mathematical foundation for traffic equilibrium analysis, demonstrating that the equilibrium conditions can be derived from a well-defined optimization problem. The convexity of the problem ensures that an optimal solution exists and can be efficiently computed using numerical methods, making it a powerful tool for transportation planning and network analysis.

8.4.1 Example

To illustrate Beckmann's model in a simple yet insightful way, we consider a network with two parallel links connecting the same origin and destination. The objective is to determine the equilibrium distribution of flow between these two links, given their respective travel time functions.

The link performance functions for links 1 and 2 are given by:

$$t_1(x) = 2 + x_1^2, \quad t_2(x) = 4 + 2x_2^2.$$

These functions indicate that the cost of travel increases as more users choose a particular link, reflecting congestion effects.

Using Beckmann's formulation, we define the objective function by integrating the travel time functions:

$$\int_0^{x_1} t_1(z)dz = 2x_1 + \frac{1}{3}x_1^3,$$

$$\int_0^{x_2} t_2(z)dz = 4x_2 + \frac{2}{3}x_2^3.$$

The optimization problem is then formulated as:

$$\min_{x_1,x_2} 2x_1 + \frac{1}{3}x_1^3 + 4x_2 + \frac{2}{3}x_2^3,$$

subject to the constraints:

$$x_1 + x_2 = 3, \quad x_1, x_2 \geqslant 0.$$

To solve this problem, we express x_2 in terms of x_1 as $x_2 = 3 - x_1$, substituting it into the objective function:

$$f(x_1) = 2x_1 + \frac{1}{3}x_1^3 + 4(3 - x_1) + \frac{2}{3}(3 - x_1)^3.$$

The first derivative is computed to find critical points:

$$f'(x_1) = 2 + x_1^2 - 4 - 2(3 - x_1)^2 = -x_1^2 + 12x_1 - 20.$$

Setting $f'(x_1) = 0$ leads to solving the quadratic equation:

$$-x_1^2 + 12x_1 - 20 = 0.$$

Solving for x_1 gives two potential solutions, $x_1 = 2$ and $x_1 = 10$. Since $x_1 = 10$ is infeasible given the constraint $x_1 + x_2 = 3$, the optimal solution is:

$$x_1 = 2, \quad x_2 = 3 - x_1 = 1.$$

To determine that it corresponds to a minimum, we compute the second derivative:

$$f''(x_1) = -2x_1 + 12.$$

Evaluating this at $x_1 = 2$,

$$f''(2) = 8 > 0.$$

Since $f''(2) > 0$, $x_1 = 2$ is a local minimum.

This result shows that, at equilibrium, 2 travelers choose link 1 while 1 traveler chooses link 2.

To further illustrate the optimization problem, Figure 8.10 presents a plot of the objective function as a function of x_1.

The x-axis represents the flow on link 1, denoted as x_1, while the y-axis represents the value of the objective function. The function is plotted over a broader range of values for x_1, but only the interval $0 \leqslant x_1 \leqslant 3$ is feasible, as

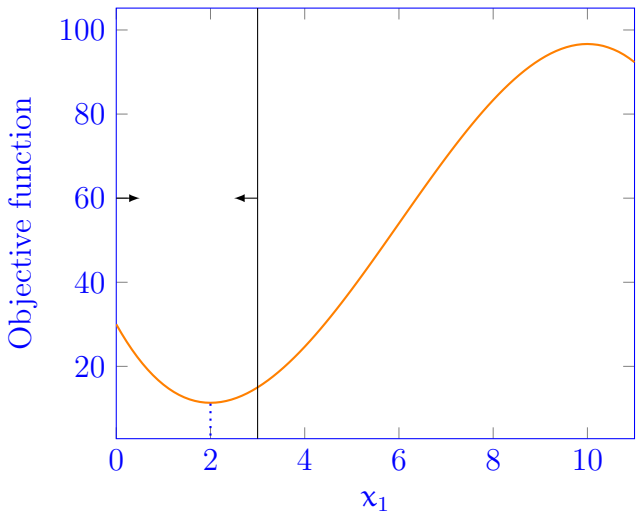


Figure 8.10: Plot of the objective function as a function of x_1. The feasible region is restricted to $0 \leqslant x_1 \leqslant 3$, where the function is convex. The optimal solution is found at $x_1 = 2$.

dictated by the constraint $x_1 + x_2 = 3$ with non-negative flows. This feasible region is indicated on the plot.

The function exhibits convexity over the feasible range, meaning that it has a single minimum. This property ensures that the optimization problem has a unique optimal solution, which corresponds to the equilibrium condition. The vertical dotted line marks the optimal value $x_1 = 2$, confirming that this is the point where the objective function is minimized.

8.4.2 Equivalence with equilibrium

To establish the equivalence between Beckmann's optimization model and the user equilibrium conditions, we begin by formulating the Lagrangian function associated with the optimization problem. It is given by:

$$L(y; \lambda, \mu) = \sum_{\ell} \int_0^{x_\ell} t_\ell(z)\,dz + \sum_{q'} \lambda_{q'}(f_{q'} - \sum_{p'} y_{p'}^{q'}) - \sum_{p'} \sum_{q'} \mu_{p'q'} y_{p'}^{q'}.$$

Here, $\lambda_{q'}$ are the Lagrange multipliers associated with the demand conservation constraints, while $\mu_{p'q'}$ are the multipliers associated with the non-negativity constraints on path flows.

The necessary optimality conditions require that:

- **Stationarity condition:** The gradient of the Lagrangian with respect

to each path flow variable y_p^q must be zero:

$$\frac{\partial L}{\partial y_p^q} = 0, \quad \forall p, q.$$

- **Inequality constraints:** The Lagrange multipliers associated with non-negativity must satisfy:

$$\mu_{pq} \geqslant 0, \quad \forall p, q.$$

- **Complementarity slackness:** If a path is used, its associated Lagrange multiplier must be zero:

$$\mu_{pq} y_p^q = 0, \quad \forall p, q.$$

Next, we compute the derivatives of the objective function. The total cost function is given by:

$$f(y) = \sum_\ell \int_0^{x_\ell} t_\ell(z) dz,$$

where the link flow x_ℓ is expressed in terms of the path flows:

$$x_\ell = \sum_p P_{\ell p} y_p^q, \quad \forall \ell, q.$$

Taking the derivative of the objective function with respect to a path flow y_p^q gives:

$$\frac{\partial f}{\partial y_p^q} = \sum_\ell \frac{\partial f}{\partial x_\ell} \frac{\partial x_\ell}{\partial y_{pq}}.$$

Since $\frac{\partial f}{\partial x_\ell} = t_\ell(x_\ell)$ and $\frac{\partial x_\ell}{\partial y_{pq}} = P_{\ell p}$, we obtain:

$$\frac{\partial f}{\partial y_p^q} = \sum_\ell P_{\ell p} t_\ell(x_\ell).$$

By definition, the path cost c_p^q is the sum of the link costs weighted by the incidence matrix:

$$c_p^q = \sum_\ell P_{\ell p} t_\ell(x_\ell).$$

Thus, the derivative of the objective function with respect to each path flow corresponds exactly to the path cost c_p^q. This result plays a central role

in proving the equivalence between the optimality conditions of Beckmann's model and the user equilibrium conditions.

Next, we compute the derivatives of the Lagrangian function with respect to the path flows y_p^q. The Lagrangian function is given by:

$$L(y; \lambda, \mu) = \sum_{\ell} \int_0^{x_\ell} t_\ell(z) dz + \sum_{q'} \lambda_{q'} (f_{q'} - \sum_{p'} y_{p'}^{q'}) - \sum_{p'} \sum_{q'} \mu_{p'q'} y_{p'}^{q'}.$$

Taking the derivative with respect to y_p^q, we obtain:

$$\frac{\partial L}{\partial y_p^q} = c_p^q - \lambda_q - \mu_{pq}.$$

The necessary optimality conditions require that the Lagrange multipliers satisfy the non-negativity constraint:

$$\mu_{pq} = c_p^q - \lambda_q \geqslant 0, \quad \forall p, q.$$

This implies that the cost of using a path must be at least as large as the equilibrium cost λ_q.

Finally, the complementarity slackness condition ensures that any path carrying flow must have the minimum cost:

$$y_p^q (c_p^q - \lambda_q) = 0, \quad \forall p, q.$$

This means that if a path p carries positive flow, then its cost must be exactly equal to λ_q. Conversely, if a path is not used ($y_p^q = 0$), then it must have a strictly higher cost than λ_q.

To establish the equivalence between the optimality conditions derived from Beckmann's model and the user equilibrium conditions, we simply need to interpret the Lagrange multipliers in the optimization problem. Specifically, the Lagrange multiplier λ_q associated with the demand conservation constraint for each OD pair q represents the minimum path cost at equilibrium. This interpretation allows us to directly relate the optimality conditions of the optimization problem to the conditions required for user equilibrium.

The first condition states that the cost of using any path p within an OD pair q cannot be lower than the minimum cost c_q^* across all paths for that OD pair. Mathematically, this is expressed as:

$$c_p^q \geqslant c_q^*, \quad \forall q.$$

This ensures that no path has a cost lower than the equilibrium cost and follows naturally from the necessary optimality conditions of the optimization problem.

The second condition reinforces that only the paths with the minimum cost can carry flow. If a path is used, meaning $y_p^q > 0$, then its cost must be exactly equal to c_q^*. Conversely, if a path is not used, its cost must be strictly greater than the minimum cost. This is captured by the complementarity slackness condition:

$$y_p^q(c_p^q - c_q^*) = 0, \quad \forall p, q.$$

The third condition ensures that the total demand for each OD pair is fully assigned among the available paths:

$$\sum_p y_p^q = f_q, \quad \forall q.$$

This follows directly from the demand conservation constraint in the optimization problem and guarantees that all travelers in the system are accounted for.

Finally, the non-negativity condition states that all path flows must be non-negative:

$$y_p^q \geqslant 0, \quad \forall p, q.$$

This ensures that the solution remains physically meaningful, as negative flows do not make sense in a traffic assignment context.

These conditions are precisely the equilibrium conditions introduced earlier, demonstrating that the optimal solution to Beckmann's optimization model corresponds exactly to a user equilibrium. This theoretical result provides a powerful justification for using optimization-based approaches to solve equilibrium traffic assignment problems, as finding an optimal solution to the mathematical program automatically yields a valid equilibrium distribution of traffic flows.

A fundamental property of Beckmann's optimization model is the uniqueness of its solution in terms of link flows. This result follows from the strict convexity of the objective function with respect to link flows. The function to be minimized consists of integrals of the link cost functions, which, under the standard assumption that link costs increase monotonically with flow, ensures strict convexity. This property guarantees that the optimization problem has a unique global minimum.

As a consequence, the equilibrium solution in terms of link flows is unique. This means that, regardless of the specific path flows used to assign travelers to the network, the total number of vehicles using each link will always be the same at equilibrium.

However, the same uniqueness property does not extend to path flows. Because multiple sets of path flows can result in the same link flow distribution, there may be multiple valid equilibrium solutions in terms of route

choice. That is, different groups of travelers may distribute themselves among alternative routes in different ways while still satisfying the equilibrium conditions.

8.5 Algorithm

One of the key challenges in solving the traffic assignment problem is the complexity associated with the path-based formulation. In theory, one could define the problem in terms of path flows, where each traveler is explicitly assigned to a specific route. However, this approach becomes computationally infeasible for real-world networks due to the sheer number of possible paths. The number of paths between an origin and a destination grows exponentially with the number of nodes in the network, making it impossible to enumerate and store all potential routes explicitly.

This issue is reminiscent of the computational challenges encountered in shortest path problems. In large transportation networks, listing all possible routes between nodes is impractical, which is why efficient algorithms such as Dijkstra's algorithm are used to determine the shortest path dynamically rather than relying on precomputed route lists. The same principle applies to the traffic assignment problem: instead of working with an exhaustive enumeration of paths, we employ algorithms that dynamically compute paths as needed, ensuring computational feasibility.

In practice, solution methods for the traffic assignment problem rely on iterative procedures that update path flows based on travel times, progressively adjusting the assignment until an equilibrium is reached. These algorithms typically incorporate shortest path computations to identify the most efficient routes at each iteration.

The algorithm for solving the traffic assignment problem is structured as follows:

Initialization Start with an empty network.

- Set initial link costs to their free-flow values: $t_\ell(0)$.
- Compute initial link flows using an all-or-nothing assignment: x^0.
- Set iteration counter $k = 0$.

Step 1 Update link costs based on current flows:

$$t_\ell^k = t_\ell(x^k), \quad \forall \ell.$$

Step 2 Compute new link flows using an all-or-nothing assignment:

$$\tilde{x}^k = \text{All-or-Nothing}(t^k).$$

Step 3 Perform a line search to determine the step size:

$$x^{k+1} = x^k + \alpha(\tilde{x}^k - x^k), \quad 0 \leqslant \alpha \leqslant 1,$$

where α is obtained by solving:

$$\min_{\alpha} \sum_{\ell} \int_0^{x_\ell^{k+1}} t_\ell(z)\,dz.$$

Step 4 Check convergence. If not converged, return to Step 1.

The algorithm starts with an *initialization* step, where all link flows are initially set to zero, and the travel costs are assigned their free-flow values. At this stage, an *all-or-nothing assignment* is performed, meaning that all travelers select the shortest path based on the free-flow link costs, without considering congestion. This produces an initial set of link flows, denoted as x^0, and the iteration counter is set to $k = 0$.

In *Step 1*, the algorithm updates the travel costs based on the current traffic conditions. Since travel times typically increase with congestion, this step ensures that the costs used in subsequent iterations reflect the congestion levels at iteration k. The updated costs are computed as $t_\ell^k = t_\ell(x^k)$ for each link ℓ.

Step 2 involves reassigning traffic using an *all-or-nothing assignment* based on the updated link costs. Each traveler selects the shortest available route under the current cost conditions, generating a new set of link flows, denoted as $\tilde{x}^k$. This step models how travelers respond to changing traffic conditions by choosing the best available paths.

Step 3 introduces a *line search* to ensure smooth convergence. Instead of fully adopting the newly computed flow vector $\tilde{x}^k$, the algorithm updates the flows as a weighted combination of the previous and new flow estimates. The parameter α, which determines how much of the new assignment is incorporated, is chosen to minimize the total system travel cost, ensuring numerical stability and efficient convergence.

Finally, in *Step 4*, the algorithm checks whether the solution has stabilized. Convergence is typically assessed by measuring the difference between successive iterations of link flows. If the changes are small enough, the process terminates, indicating that an equilibrium assignment has been reached. If convergence is not achieved, the algorithm returns to Step 1 and continues iterating until equilibrium conditions are met.

8.5.1 Example

To illustrate the first iteration of the algorithm, we consider an example of an all-or-nothing assignment on an empty network. The initial flow assignment is determined based on the shortest paths when no congestion is present. Figure 8.11 provides a visual representation of this process.

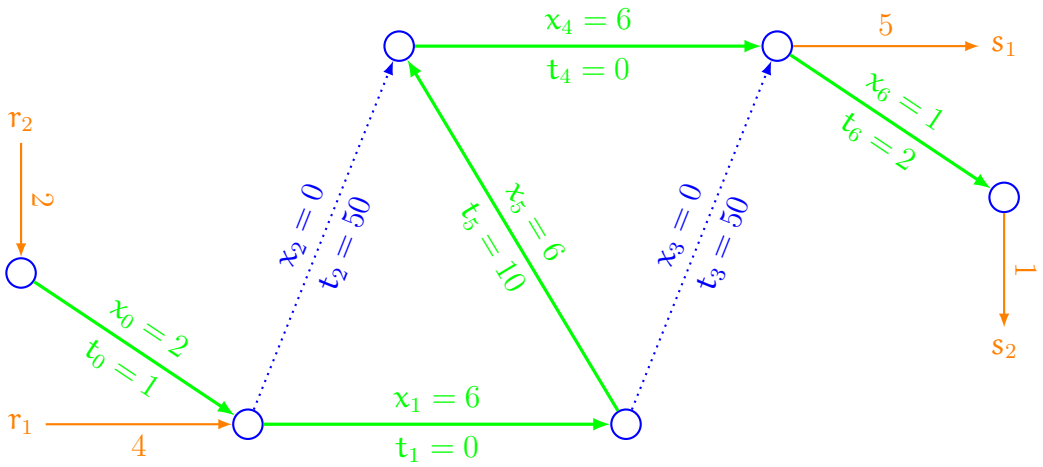


Figure 8.11: All-or-nothing assignment on an empty network.

In this representation, the orange arrows indicate where the demand enters and exits the network. These correspond to the origins and destinations of the OD pairs. Each link in the network is associated with a flow value x_ℓ and a corresponding travel cost t_ℓ, which initially reflects free-flow conditions.

The all-or-nothing assignment results in flows only on the shortest paths, shown as thick green arcs in the figure. Since no congestion effects are considered in this initial step, all travelers choose the shortest available routes based solely on the free-flow travel times. Dotted arcs represent alternative links that are not used in this assignment, as they correspond to longer paths under the current travel cost conditions.

From the flow assignment in Figure 8.11, we can derive the path costs for each OD pair, which correspond to the travel time on the used paths:

$$\mathrm{c}^*_{11} = 10, \quad \mathrm{c}^*_{12} = 12, \quad \mathrm{c}^*_{21} = 11.$$

These values indicate the minimum travel times experienced by travelers in each OD pair before any congestion effects are introduced.

After performing the initial all-or-nothing assignment, the next step in the iterative algorithm updates the link costs based on the newly assigned flows. As congestion begins to develop, these updated costs reflect the increased

travel times on links that experience higher traffic volumes. Figure 8.12 illustrates the network with these revised link costs.

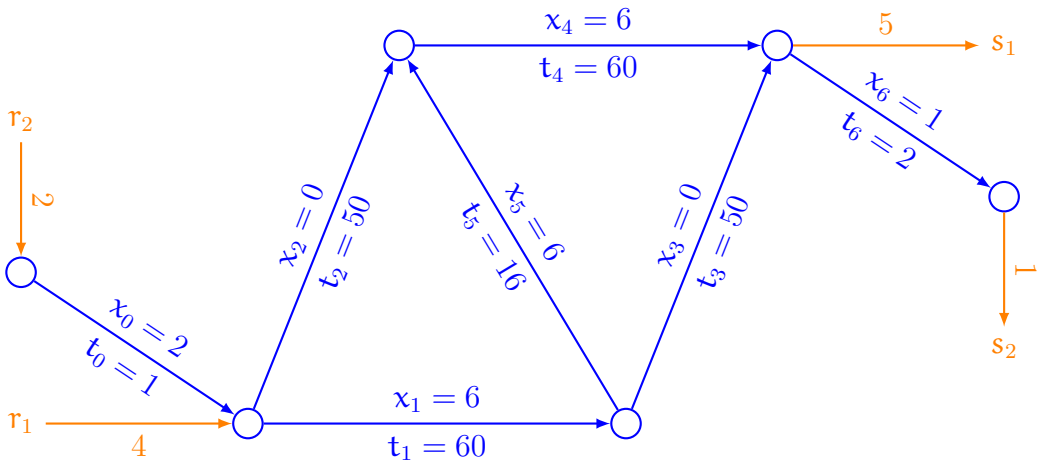


Figure 8.12: Updated link costs after the first iteration.

In this updated network representation, the structure remains the same, but the travel times on certain links have changed due to increased congestion. The link costs are now computed based on the assigned flow values. For instance, the cost on link x_1 has increased significantly, reaching $t_1 = 60$. These updated costs will influence the next iteration of the assignment process, as travelers will now reassess their route choices based on these revised values.

A new all-or-nothing assignment is performed based on the shortest paths with the revised travel times. Figure 8.13 illustrates the new flow assignment.

In this new assignment, the updated shortest paths are determined by the revised travel costs. The most significant change is the avoidance of link x_1, which now has a high cost of $t_1 = 60$. Instead, travelers who previously used this link have rerouted through alternative paths.

The new flow assignment results in a redistribution of traffic across the network. The shortest path calculations now lead to the selection of paths through the north, with flows reallocated accordingly. The dotted links in the figure represent those that are no longer used. The resulting path costs for each OD pair are:

$$c_{11}^* = 110, \quad c_{12}^* = 112, \quad c_{21}^* = 111.$$

To illustrate the line search step in the iterative optimization process, Table 8.5 presents the flow evolution for each arc.

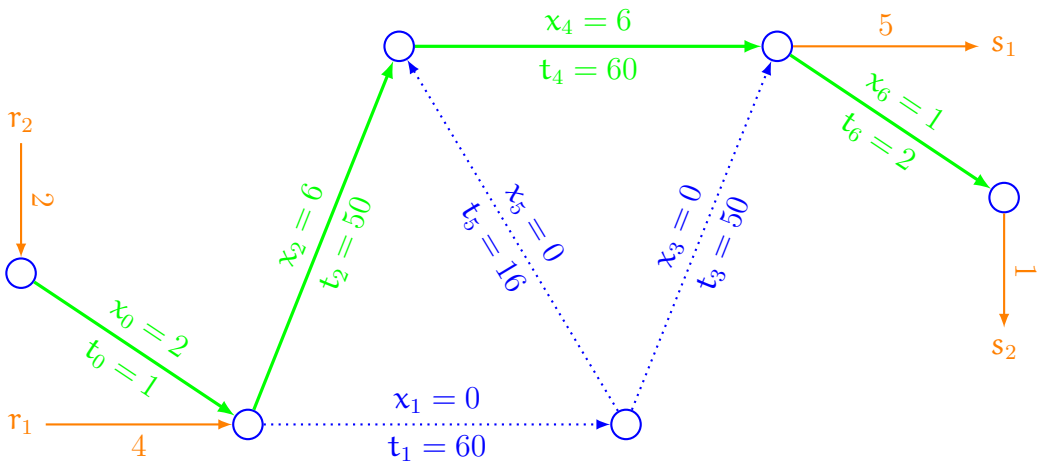


Figure 8.13: All-or-nothing assignment with updated costs.

Arc	First flow	Second flow	Convex combination	
0	2	2	2 + α (2-2)	= 2
1	6	0	6 + α (0-6)	= 6 - 6α
2	0	6	0 + α(6-0)	= 6α
3	0	0	0 + α(0-0)	= 0
4	6	6	6 + α(6-6)	= 6
5	6	0	6 + α (0-6)	= 6 - 6α
6	1	1	1 + α (1-1)	= 1

Table 8.5: Convex combination of two successive all-or-nothing assignments.

The first column represents the arc number, which identifies each link in the network. The second column shows the flow assigned to each arc after the first all-or-nothing assignment, which follows the initial shortest path selection. The third column indicates the flow assigned to each arc after the second all-or-nothing assignment, reflecting the updated shortest paths based on the revised link costs.

The last column expresses the convex combination of the two flow assignments as a function of the step-size parameter α. This parameter determines the weight given to each of the two flow configurations. For arcs where the two assignments are identical (such as arcs 0, 3, 4, and 6), the convex combination remains constant. However, for arcs where the flow has changed between the two iterations (such as arcs 1, 2, and 5), the convex combination interpolates between the two values, adjusting the flow assignment in a controlled manner.

To further analyze the impact of the step-size parameter α on the opti-

mization process, we illustrate the value of the objective function as a function of α in Figure 8.14. The x-axis represents α, which determines the weight given to the new flow assignment in the convex combination. The y-axis represents the value of the objective function, measuring the total travel cost in the network.

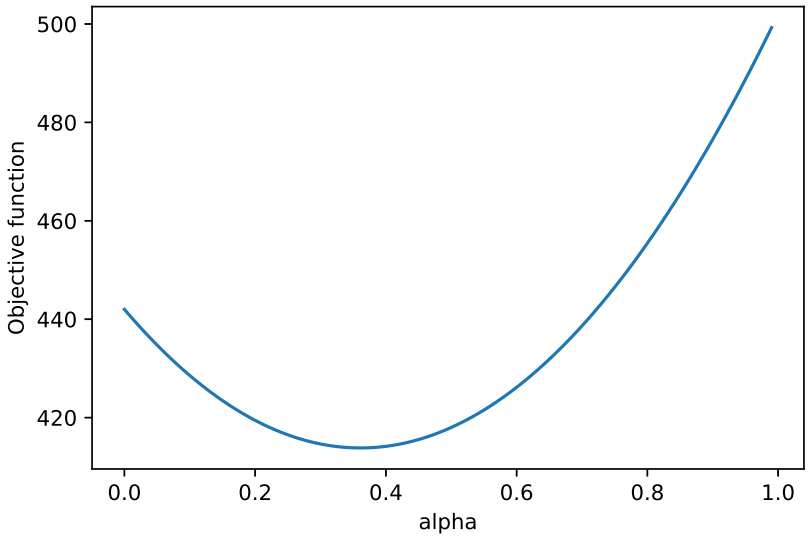


Figure 8.14: Value of the objective function as a function of α.

As α varies, the objective function exhibits a convex behavior, indicating that there exists a unique optimal value that minimizes the total cost. Since α represents a convex combination of the previous and updated flow assignments, its feasible range is restricted between 0 and 1. A value of $\alpha = 0$ corresponds to retaining the previous assignment, while $\alpha = 1$ fully adopts the new assignment computed from the all-or-nothing step.

The optimal step size, denoted as α^*, is determined by identifying the value that minimizes the objective function. In this case, the optimal value is found to be $\alpha^* = 0.361$.

The updated link flows, computed using the optimal step size $\alpha^* = 0.361$, are illustrated in Figure 8.15. This figure represents the network after applying the convex combination step, incorporating a weighted adjustment of the previous and newly computed flows.

The updated flows are computed using the convex combination formula:

$$x_\ell^{k+1} = x_\ell^k + \alpha(\tilde{x}_\ell^k - x_\ell^k),$$

where x_ℓ^k represents the link flows from the previous iteration, and $\tilde{x}_\ell^k$ represents the new all-or-nothing assignment. For example, considering link 5,

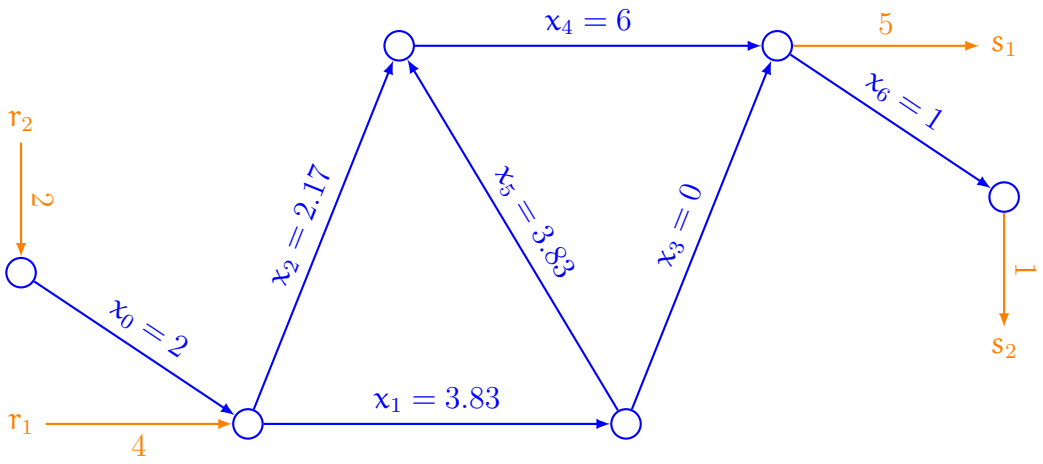


Figure 8.15: Updated link flows after applying the convex combination step.

the initial flow from the first all-or-nothing assignment was 6, while the second assignment yielded a flow of 0. Applying the convex combination with $\alpha^* = 0.361$ results in the updated flow:

$$x_5 = 6 + 0.361(0 - 6) = 6 - 2.166 = 3.83.$$

To illustrate the convergence of the algorithm, we report in Table 8.6 the values of the step size α at each iteration, along with the corresponding value of the objective function.

Iteration	α	Objective function
0		442.00
1	0.361	413.83
2	0.309	391.72
3	0.0885	390.67
4	0.0538	390.31
5	0.0358	390.15
6	0.0249	390.08
7	0.0179	390.04
8	0.0131	390.02
9	0.00967	390.01
10	0.00722	390.01
11	0.00544	390.00

Table 8.6: Values of α and the objective function at each iteration.

As observed in the table, the initial step size is relatively large, allowing for a significant reduction in the objective function. However, as the iterations

progress, the step size decreases, reflecting the refinement of the solution as the algorithm converges toward the optimal flow pattern. By iteration 11, the objective function has stabilized at 390.00. The iterative process has reached equilibrium.

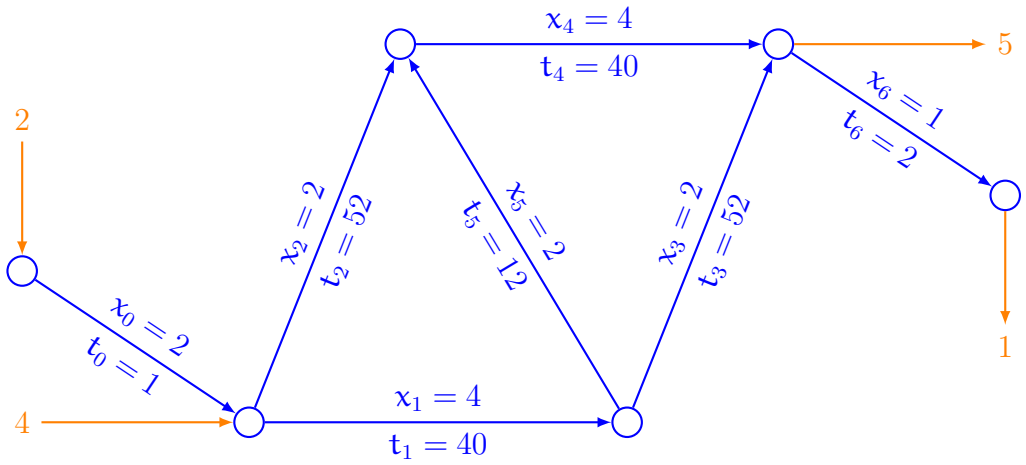


Figure 8.16: Link flows at equilibrium and corresponding travel times.

Figure 8.16 illustrates the flow on each link along with the corresponding travel times at equilibrium.

The *level of service* is commonly assessed through the average travel time experienced by all travelers in the system. It represents the expected time a traveler would spend on the network, taking into account both the individual path costs and the number of travelers using each path.

The average travel time is computed as a weighted mean of the path costs, where the weights correspond to the demand assigned to each origin-destination pair. It is given by:

$$\bar{c} = \frac{1}{\sum_q f_q} \sum_q f_q c_q^*$$

where f_q represents the total demand between origin-destination pair q, and c_q^* is the minimum travel cost experienced by travelers for that OD pair. Using the values from the figure, we compute:

$$\bar{c} = \frac{1}{6}(92 \cdot 3 + 94 \cdot 1 + 93 \cdot 2) = 92.7.$$

This value reflects the overall efficiency of the network and the congestion effects induced by the distribution of flows. A lower average travel time typically indicates a better-performing network, whereas higher values suggest congestion and inefficiencies.

8.6 Braess paradox

We now investigate a scenario where the tunnel under the city, represented by link 5, is removed. The expectation is that this would reduce network capacity and thus degrade the *level of service*, leading to an increase in average travel times. However, the computed equilibrium results reveal an unexpected outcome: the average travel time has actually *decreased*.

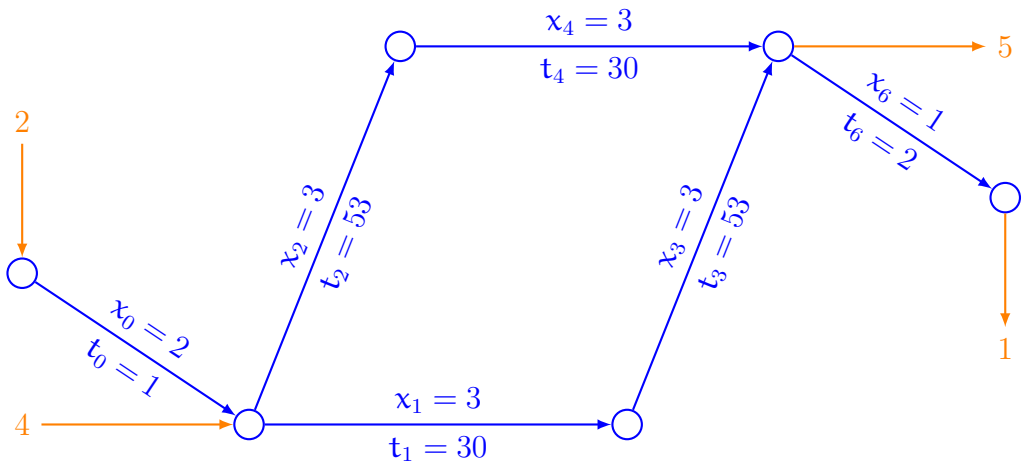


Figure 8.17: Equilibrium flows and travel times after removing link 5.

Before the removal of link 5, the average travel time was 92.7 minutes. After its removal, the recalculated mean travel time is:

$$\bar{c} = \frac{1}{6}(83 \cdot 3 + 85 \cdot 1 + 84 \cdot 2) = 83.7.$$

This paradoxical result suggests that the presence of the removed link may have induced suboptimal route choices, leading to congestion and inefficiencies. By eliminating link 5, travelers are forced to redistribute across the remaining network, and the new flow patterns result in a more balanced utilization of available routes.

This phenomenon is an illustration of *Braess's paradox*, a counterintuitive phenomenon in transportation networks where adding extra capacity to a network, such as building a new road, can lead to increased overall travel times rather than reducing congestion. Conversely, removing a road from the network can sometimes *improve* overall travel times for all users, like in our example.

At first glance, this seems paradoxical. One would naturally expect that increasing the number of available routes would provide more choices for

travelers, thereby reducing congestion and improving efficiency. However, under certain conditions, individual users acting in their own self-interest — each choosing the shortest perceived travel time — can lead to a new equilibrium where everyone is worse off.

From a practical standpoint, Braess's paradox is not merely a theoretical curiosity. It has been observed in real-world transportation networks, where new road constructions or capacity expansions have, in some cases, resulted in increased congestion. Similarly, traffic restrictions, such as pedestrianizing certain streets or implementing road closures, have sometimes led to *improved* traffic conditions, as drivers are forced to redistribute their routes more efficiently.

One real-world occurrences of Braess's paradox took place in Stuttgart in 1968 (Knödel, 1969). The event unfolded in the central area of Schlossplatz, where a new road network was introduced with the intention of improving traffic flow and reducing congestion. However, instead of alleviating the situation, the network expansion led to severe disruptions and unexpected traffic chaos.

The increased road capacity initially seemed like a logical improvement, as it provided additional routing options for drivers. However, as travelers adjusted their routes based on their individual preferences — each seeking to minimize personal travel time — the overall equilibrium of the system shifted in an unintended way. The redistribution of traffic flow actually increased congestion, resulting in longer travel times for most users.

In response to this unexpected deterioration in traffic conditions, the city of Stuttgart took an unconventional yet effective measure: it decided to close Königstrasse, a key road in the network. Counterintuitively, rather than making congestion worse, this intervention *improved* traffic conditions. The closure effectively removed an inefficient routing option, forcing vehicles to redistribute more optimally across the remaining network. As a result, travel times decreased, and the overall performance of the system improved.

Another interesting real-world occurrence of Braess's paradox took place in New York City in 1990 (Kolata, 1990). The event was triggered by the celebration of Earth Day on April 22, during which authorities decided to close 42nd Street, a major thoroughfare in Manhattan. Given the high traffic volume typically observed in this area, many expected the closure to cause severe congestion throughout the surrounding network. Some even referred to the day as *doomsday* for New York's traffic.

At first glance, the logic seemed straightforward: removing such an important road from the network would force more vehicles onto the remaining streets, leading to bottlenecks and longer travel times. Concerns were so high that transportation experts publicly speculated about potential gridlock. A

widely cited comment at the time emphasized this expectation: *"You didn't need to be a rocket scientist or have a sophisticated computer queuing model to see that this could have been a major problem."*

However, contrary to all predictions, the traffic situation actually *improved* following the closure. Observers noted that travel times on alternative routes were lower than expected, and congestion did not worsen significantly. In fact, many commuters experienced smoother flows than on a typical day. The paradoxical effect can be explained by the fact that closing 42nd Street altered the way drivers made route choices, eliminating inefficient paths and leading to a more balanced and effective use of the overall network.

The Cheonggyecheon restoration project in Seoul, initiated in 2003, provides a third real-world example of the Braess Paradox (Baker, 2009). The project involved dismantling an elevated six-lane highway that had been constructed over the Cheonggyecheon stream in the 1970s. This highway was originally built to accommodate increasing car traffic in the rapidly growing city. However, over time, congestion worsened, and concerns about environmental degradation and urban livability became more pressing.

The decision to remove the highway and restore the stream was met with considerable skepticism. Many transportation experts and local residents feared that eliminating such a critical infrastructure element would result in severe congestion and disrupt mobility in central Seoul. The prevailing assumption was that reducing road capacity would inevitably increase travel times, leading to gridlock.

Contrary to these expectations, the removal of the highway did not create traffic chaos. Instead, the overall traffic conditions in the city improved. Several factors contributed to this surprising outcome. First, the city government implemented complementary policies, including improvements to public transportation and traffic management strategies, which encouraged commuters to shift from private cars to buses and subways. Second, the redistribution of traffic across the network led to a more efficient use of existing roads, as drivers adapted their routes in response to the changes. Finally, the reduction in road capacity altered travel behavior, discouraging unnecessary car trips and promoting alternative modes of transport.

8.7 The prisonner dilemma

A key insight from the Braess Paradox is that traffic inefficiencies arise because travelers make decisions based on their own individual travel times rather than considering the impact of their choices on the entire network. In a non-cooperative setting, each traveler selects what appears to be the

shortest route for themselves, often leading to suboptimal outcomes for the system as a whole. If travelers could be convinced to make route choices that benefit the entire network, rather than just minimizing their personal travel time, the overall performance of the network could improve.

Figure 8.18 illustrates a hypothetical scenario where travelers are encouraged to avoid using the tunnel, which was previously a major contributor to network congestion. In this scenario, no vehicles use the tunnel, and instead, traffic is redistributed among the remaining routes. The network's overall level of service improves. The total travel time experienced by travelers is reduced, even though a major road remains available but unused.

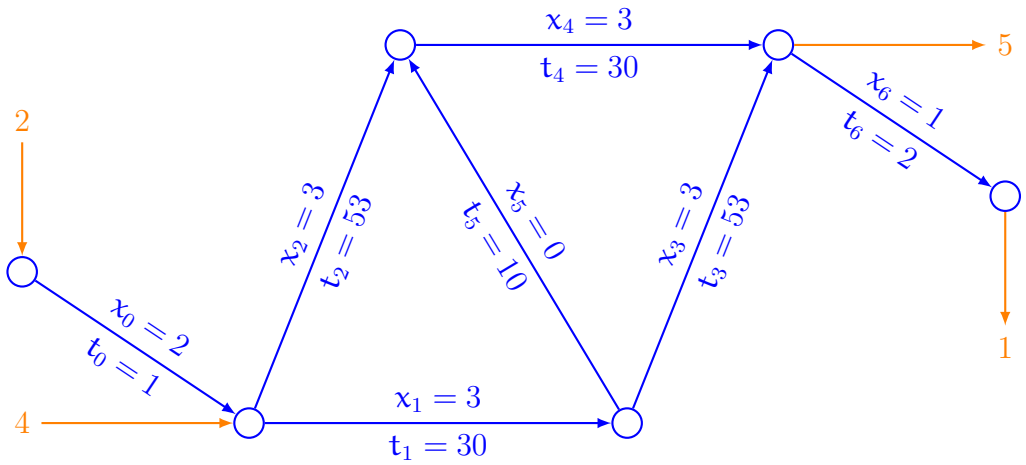


Figure 8.18: Hypothetical scenario where travelers are convinced to avoid the tunnel, leading to an overall improvement in network performance.

In this scenario, if travelers could collectively agree to avoid the tunnel, the overall travel time in the network would decrease, benefiting everyone. However, despite the clear advantage of cooperation, individuals tend not to coordinate spontaneously. This phenomenon can be explained through the well-known *Prisoner's Dilemma*, a fundamental concept in game theory.

The *Prisoner's Dilemma* describes a situation in which two individuals, Joe and Averell, have been arrested. They are suspected of committing a major robbery, but the authorities lack sufficient evidence to convict them. However, they do have evidence of a minor robbery, for which both would receive a light sentence of one year in prison. The two prisoners are separated and cannot communicate with each other. Each is presented with the same choice: they can either remain silent or betray the other by confessing to the major robbery.

The consequences of their decisions are as follows:

- If both remain silent, they will each serve only 1 year in prison, as the authorities can convict them only for the minor crime.
- If both betray each other, they each receive a sentence of 2 years, as their confessions provide sufficient evidence for the major crime.
- If one prisoner betrays the other while the other remains silent, the betrayer is set free, while the betrayed prisoner receives the maximum sentence of 3 years.

The optimal strategy for the prisoners can be understood by analyzing all possible outcomes of their decisions. Table 8.7 presents the different scenarios that can arise based on whether Joe and Averell choose to remain silent or betray the other.

Table 8.7: Possible outcomes in the Prisoner's Dilemma

Decision		Penalty		
Joe	Averell	Joe	Averell	Total penalty
Silent	Silent	1	1	2
Silent	Betray	3	0	3
Betray	Silent	0	3	3
Betray	Betray	2	2	4

The table lists all four possible combinations of decisions and the corresponding penalties for each prisoner. The key observation is that when both prisoners remain silent, they each receive only one year in prison, resulting in a total penalty of 2 years. This is the lowest possible total penalty across all scenarios, making it the *globally optimal* outcome.

However, from an individual perspective, betraying the other always seems like the best option. To fully understand why, we must examine the problem from the perspective of each individual. Each prisoner must make their decision in isolation, without knowing what the other will choose. Their reasoning follows a logical analysis of the possible outcomes.

Consider first Joe's perspective. He knows that Averell can either stay silent or betray him. If Averell stays silent, Joe has two choices: he can also stay silent, in which case he will receive a 1-year prison sentence, or he can betray Averell and go free. Clearly, betraying is the better option in this scenario. Now, suppose instead that Averell betrays Joe. In this case, if Joe stays silent, he will receive the maximum penalty of 3 years in prison. However, if he also betrays Averell, his sentence is reduced to 2 years. Once again, betraying is the better choice. The important observation is that,

regardless of what Averell chooses to do, Joe's best strategy is always to betray.

Now, let us examine the problem from Averell's perspective. His reasoning follows exactly the same logic. If Joe stays silent, Averell can either remain silent and serve 1 year in prison or betray Joe and go free. The best choice, from his perspective, is to betray. If Joe betrays him, Averell faces a choice between remaining silent, which results in a 3-year sentence, or betraying Joe and serving only 2 years. Again, betrayal is the better option. Just like Joe, Averell is always better off betraying, *regardless of what Joe does*.

This is the essence of the Prisoner's Dilemma: when each individual acts in their own self-interest, they both end up in a situation that is worse than if they had cooperated. Even though both prisoners would have been better off by remaining silent, rational decision-making from an individual standpoint leads them to betray. This non-cooperative behavior is what ultimately results in the worst collective outcome.

The same logic applies to travelers choosing their routes in the network. Each traveler aims to minimize their own travel time without considering the impact on others. If avoiding the tunnel would improve travel times for everyone, the optimal outcome would be for all travelers to avoid using it. However, each traveler sees an individual advantage in taking the tunnel, hoping that others will choose an alternative route. If all travelers think this way, congestion increases, and the network performs worse than if cooperation had taken place.

This analogy illustrates why spontaneous cooperation rarely occurs in transportation systems: individuals act in their self-interest, leading to suboptimal outcomes for the group. This provides an important justification for policy interventions, such as tolls, access restrictions, or incentives for alternative transportation modes, to guide users toward more efficient network usage.

8.8 System optimum

In the context of traffic assignment, the distinction between *user equilibrium* and *system optimum* is fundamental in understanding how individual decision-making differs from centrally coordinated traffic management. Both concepts arise from the same network structure and demand conditions but lead to different flow distributions and overall system performance.

The *user equilibrium* corresponds to the situation in which every traveler selects their route independently, seeking to minimize their own travel

time. As we have seen in Section 8.4, this is captured mathematically by the following optimization problem:

$$y^* = \text{argmin}_y \sum_\ell \int_0^{x_\ell} t_\ell(z)dz$$

subject to the constraints:

$$\sum_p y_p^q = f_q, \qquad \forall q,$$
$$y_p^q \geqslant 0, \quad \forall p, q.$$

This formulation ensures that all users make their choices based on personal optimization, leading to a *Nash equilibrium*, where no traveler can unilaterally switch to a different path and achieve a lower travel time. If x^* denotes the vector of link flows corresponding to the user equilibrium path flow y^*, the total cost for all users is

$$c^* = \sum_\ell x_\ell^* t_\ell(x_\ell^*)$$

In contrast, the *system optimum* represents the best possible traffic assignment from the perspective of minimizing total system-wide travel time. Instead of individual travel time minimization, the goal is to minimize the overall cost of congestion by optimizing flows across all links. The corresponding mathematical formulation is:

$$\tilde{y}^* = \text{argmin}_y \sum_\ell x_\ell t_\ell(x_\ell)$$

subject to the same constraints:

$$\sum_p y_p^q = f_q, \qquad \forall q,$$
$$y_p^q \geqslant 0, \quad \forall p, q.$$

Here, rather than integrating the travel time function, the objective function directly sums the total cost experienced on each link, weighted by its flow. This results in a flow pattern that minimizes total system congestion, rather than ensuring that each individual has no incentive to change their route. If $\tilde{x}^*$ denotes the vector of link flows corresponding to the system optimum path flow $\tilde{y}^*$, the total cost for all users is

$$\tilde{c}^* = \sum_\ell \tilde{x}_\ell^* t_\ell(\tilde{x}_\ell^*).$$

The ratio between the total costs associated with these two solutions, represented by $\mathfrak{c}^*/\tilde{\mathfrak{c}}^*$, is known in game theory as the *price of anarchy* (Papadimitriou, 2001). From an engineering point of view, the difference between the two, $\mathfrak{c}^* - \tilde{\mathfrak{c}}^*$ quantifies the inefficiency introduced by decentralized, selfish decision-making in contrast to an optimally coordinated traffic assignment. Indeed, *user equilibrium leads to higher total travel costs than the system optimum*, because travelers do not take into account the externalities their choices impose on others. This gap highlights the potential benefits of traffic management policies, such as congestion pricing or routing incentives, which aim to reduce the inefficiencies inherent in self-interested decision-making.

From an engineering perspective, traffic management is not only about observing travel patterns but actively shaping the transportation system to achieve better efficiency and reliability. Engineers's role is the *design*, *maintenance*, and *operation* of networks to ensure their functionality under increasing demand. Their primary objective is to *minimize the price of anarchy*, reducing the inefficiencies that arise when individual decisions do !not align with system-wide optimal performance. The benchmark for assessing these efforts is the *system optimum*, which represents the best overall state of traffic flow in terms of minimizing total travel costs.

To move towards the system optimum, engineers rely on two main types of interventions: *supply-based* and *demand-based* approaches. *Supply-based* measures include direct control strategies such as traffic lights, speed limits, and other regulations enforced by law. These measures are designed to shape traffic flow and prevent congestion in an orderly manner. Advanced control strategies, such as those studied in traffic flow theory, can further optimize network performance by dynamically adjusting traffic signals or implementing adaptive speed regulations.

On the other hand, *demand-based* approaches aim to influence traveler behavior through information, incentives, and pricing mechanisms. Unlike supply-based measures, compliance with these strategies is not always guaranteed, as travelers may choose not to follow recommendations or may react unpredictably to incentives. Congestion pricing is a well-known example of a demand-based intervention, where travelers are charged based on road usage to encourage a more efficient distribution of trips across time and space.

While system optimum represents a desirable state for the average traveler, its implementation must also consider broader societal factors. In some cases, achieving the system optimum may mean that a subset of travelers experiences a *worse* outcome, even if the majority benefits. As a result, policymakers and engineers must also consider concepts like *equity* and *minimum level of service*, ensuring that solutions do not disproportionately disadvan-

tage certain groups.

Thus, the engineering approach to traffic management requires balancing efficiency with fairness. While the goal is to align individual decision-making with system-wide benefits, the challenge lies in implementing strategies that optimize network performance while maintaining an acceptable level of service for all users.

8.9 Summary

The study of *traffic assignment* provides fundamental insights into how travelers distribute themselves across a transportation network and the consequences of their individual decisions. One of the central concepts in this analysis is *user equilibrium*, a condition where no traveler can unilaterally improve their travel time by choosing an alternative route. At equilibrium, all used paths between a given origin-destination pair have the same travel cost, ensuring that no individual traveler has an incentive to switch routes. Conversely, paths that are not used have higher costs, making them unattractive to travelers. This condition can be formulated as an optimization problem, which allows for systematic analysis and computation of equilibrium states.

A particularly surprising result in traffic assignment is known as *Braess paradox*, which demonstrates that increasing the capacity of a network does not always lead to better overall performance. In some cases, adding a new road link can induce travelers to behave in a way that increases congestion, ultimately deteriorating the level of service for all users. Conversely, *removing* a link from the network may lead to an improved distribution of traffic, reducing total travel time. This paradox highlights the complex and often counterintuitive nature of network dynamics, emphasizing that more infrastructure does not necessarily translate into better traffic conditions.

Beyond individual decision-making, another critical concept in traffic management is the *system optimum*, which represents the most efficient use of the network from a societal perspective. Achieving this state requires cooperation among travelers, as opposed to each traveler acting purely in their self-interest. However, such cooperation does not arise naturally in decentralized systems. This challenge can be understood through the lens of the *prisoner's dilemma*, where individuals acting in their own best interest produce an outcome that is suboptimal for everyone. In the context of traffic, reaching the system optimum often requires external intervention, such as congestion pricing, incentives, or regulatory measures, to align individual decisions with the collective good.

From an engineering perspective, achieving system-optimal conditions is a

primary objective. Engineers and policymakers must carefully design strategies that mitigate the inefficiencies of user equilibrium while ensuring fair and effective traffic management. By understanding the interplay between individual choices and network-wide performance, traffic assignment models provide a foundation for developing solutions that enhance mobility, reduce congestion, and improve overall transportation efficiency.

8.10 The four step model: a summary

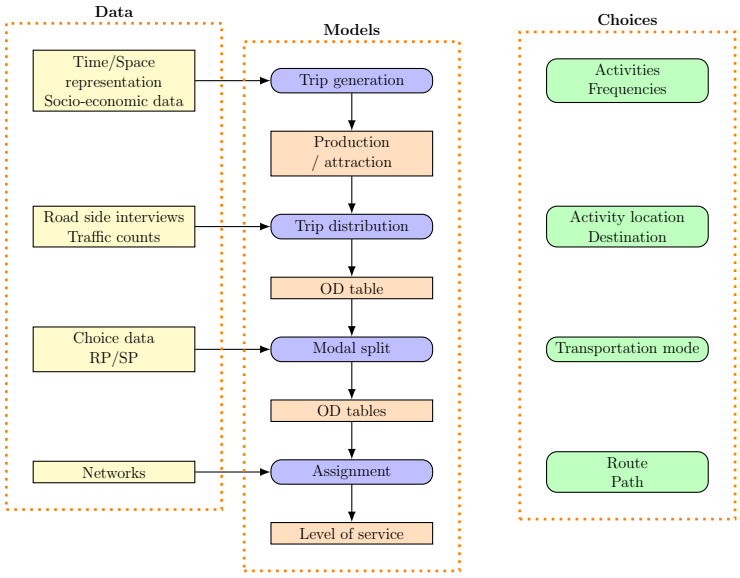


Figure 8.19: The four-step model structure: data, models, and behavioral choices.

The *four-step model* is a traditional framework used for predicting travel demand and analyzing transportation systems. It is structured around a sequence of interdependent steps, each representing a different decision that travelers make. The process is data-driven, relies on mathematical models, and aims to capture behavioral choices at various levels.

Figure 8.19 illustrates the structure of the four-step model. The figure is divided into three main sections. On the left, it displays the *data* used as inputs for each step. In the middle, it presents the *models* that describe traveler behavior. On the right, it outlines the *choices* travelers make at each step.

The first step, *trip generation*, estimates the number of trips produced and attracted by different zones based on socio-economic characteristics, and

land use. This step relies on data such as demographic distributions and land use statistics.

The second step, *trip distribution*, determines the destinations of trips. Given the number of trips generated in each zone, it predicts where these trips will be directed, forming an origin-destination (OD) table. This step incorporates observed travel patterns from surveys and roadside interviews and traffic counts to ensure consistency with real-world behavior.

The third step, *mode choice*, assigns each trip to a specific transportation mode, such as car, public transit, walking, or cycling. This stage is influenced by factors such as cost, travel time, convenience, and personal preferences. Choice models, often estimated from revealed and stated preference surveys, are used to capture the decision-making process.

The fourth step, *traffic assignment*, determines the specific routes taken by travelers based on the network conditions and level of service. It simulates how trips distribute across the network and computes travel times on individual links, considering congestion effects.

A fundamental feature of the four-step model is its *iterative nature*. The final outcome, the *level of service* experienced on the network, influences each step in the process. If congestion increases, it alters trip distribution, mode choice, and even trip generation, requiring recalibration of earlier steps. This feedback loop ensures that the model remains consistent with real-world travel behavior and allows for the evaluation of different transportation policies and infrastructure scenarios.

Chapter 9

Congestion pricing

Congestion is a complex social problem that affects the quality of life in urban areas around the world. It leads to longer travel times, increased fuel consumption, elevated pollution levels, and overall inefficiency in the transportation system. Traditional planning approaches, such as building more roads or expanding infrastructure, have often proven insufficient or even counterproductive due to induced demand.

An alternative approach, rooted in economic principles, is to create incentives that influence individual behavior without prescribing specific actions. The idea is not to dictate how people should travel, but rather to modify the context in which they make travel decisions, allowing them to adapt in ways that work best for them individually. This approach forms the philosophical foundation of *congestion pricing*.

Congestion pricing refers to a set of pricing mechanisms designed to charge users of public infrastructure — particularly roads — for the negative externalities they generate during peak periods, when demand exceeds supply. In simple terms, it aims to internalize the social cost of congestion by making travelers face a monetary cost that reflects the impact of their travel decisions on others. By doing so, it discourages unnecessary trips, encourages the use of alternative modes or off-peak travel, and ultimately leads to a more efficient use of the transport system.

This concept has been implemented in several cities worldwide, beginning with Singapore in 1975. London introduced a congestion charge in 2003, followed by Stockholm in 2006 and Milan in 2008, among others.

The city of Stockholm provides an instructive and well-documented example of congestion pricing in practice. It demonstrates not only the effectiveness of such a policy in reducing traffic congestion, but also the political, technical, and social challenges involved in its implementation.

The congestion pricing scheme in Stockholm was first introduced on a trial

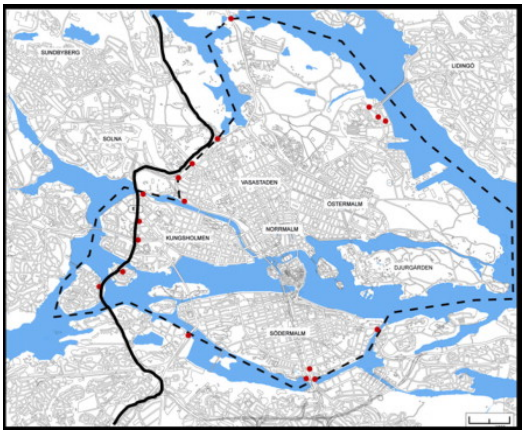


Figure 9.1: Cordon around Stockholm

basis between January 3 and July 31, 2006. This pilot phase was designed to test the practical aspects of the system and to evaluate its effects on traffic patterns, public transport usage, environmental impact, and public opinion. The trial was monitored closely and accompanied by extensive data collection and analysis.

Following the trial period, political opponents of the congestion charges insisted on a referendum to determine whether the scheme should become permanent. The vote was held in August 2006. The results highlighted a clear geographical divide in public opinion: within the city of Stockholm itself, a majority of voters supported the charges, whereas voters in the surrounding municipalities were predominantly opposed. Despite the mixed results, the decision was ultimately made to reinstate the charges on a permanent basis. This reintroduction took place in August 2007, marking the beginning of Stockholm's full-scale, long-term congestion pricing policy (see Börjesson et al., 2012).

The structure of the Stockholm system is relatively straightforward. A toll cordon surrounds the inner city (see Figure 9.1), and vehicles are charged each time they cross this boundary. The pricing is time-dependent, reflecting the intensity of congestion throughout the day. The system sas installed with the following configuration. During peak hours—defined as 7:30 to 8:30 in the morning and 16:00 to 17:30 in the afternoon—the charge is set at 2 euros. In the periods just before and after the peaks, the fee is slightly reduced to 1.5 euros. During the rest of the day, the charge drops to 1 euro, while no toll is applied at all between 18:30 in the evening and 6:30 in the morning.

To ensure fairness and avoid excessive charges for frequent travelers, the system includes a cap on daily payments. Regardless of the number of entries

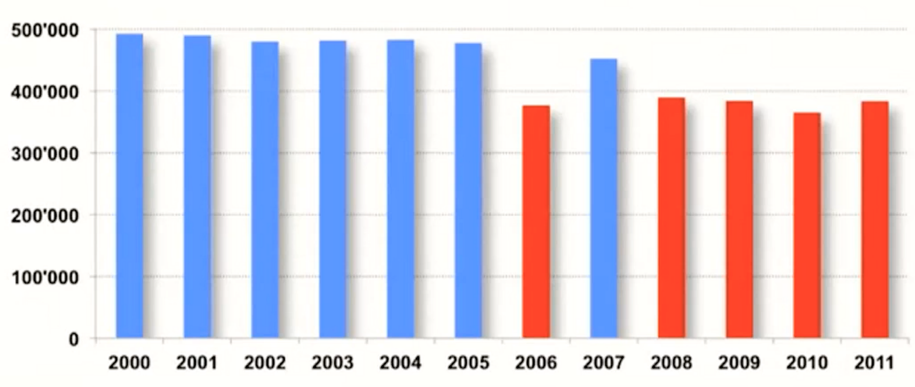


Figure 9.2: Average Daily Traffic volumes

and exits across the cordon, no driver is required to pay more than 6 euros per day. This helps mitigate concerns about the financial burden on commuters while preserving the incentive structure of the policy.

The introduction of congestion pricing in Stockholm had an immediate and measurable impact on traffic patterns within the city. Shortly after implementation, the number of vehicles entering the inner city dropped by approximately 20% (see Figure 9.2, extracted from Eliasson, 2012). This substantial reduction in car traffic occurred almost overnight, demonstrating the effectiveness of price signals in influencing travel behavior. The decrease was not the result of any major infrastructural changes or coercive restrictions, but simply of introducing a cost to what had previously been a free resource: access to the congested city center during peak hours.

This raises the natural question: where did all those cars go? Some travelers chose to cancel non-essential trips altogether, while others modified their travel behavior in different ways. A portion of drivers changed their destination, opting for locations outside the toll cordon. Others shifted to alternative modes of transport, such as public transit, cycling, or walking. Still others adjusted their departure times to avoid the most expensive time slots. Each of these behavioral adaptations contributed to the overall reduction in congestion, and importantly, they emerged organically in response to the pricing structure, without requiring direct intervention or micromanagement from authorities.

Public perception of the congestion pricing scheme also evolved significantly over time. At the outset of the trial, support for the policy was relatively low. Approximately 30% of the population viewed it favorably, while the majority remained skeptical or outright opposed. However, as the

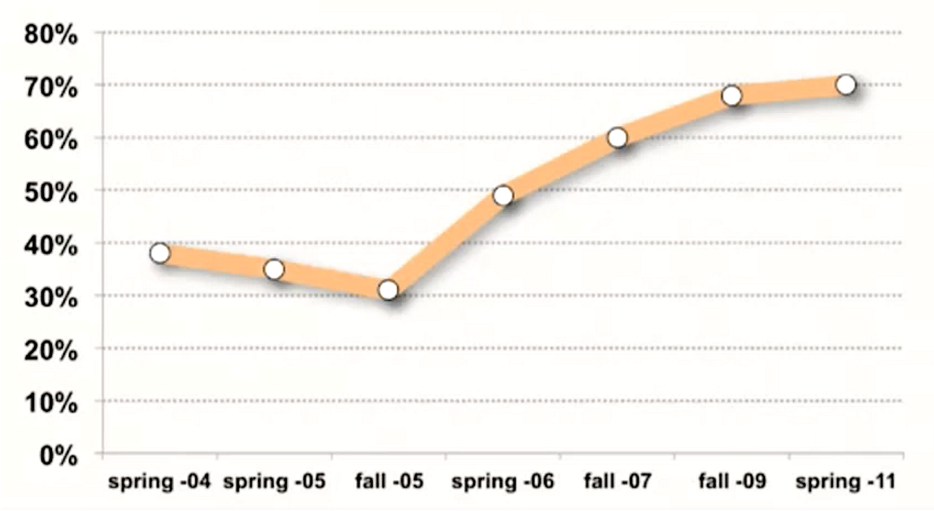


Figure 9.3: Evolution of public support

benefits became more apparent — less traffic, shorter travel times, and improved air quality — public opinion shifted dramatically. By the time the system was reintroduced permanently in 2007, support had effectively reversed, with around 70% of residents in favor of the charges (see Figure 9.3). This reversal is a compelling illustration of how experiential evidence can transform attitudes toward controversial policies.

Despite its success, the congestion pricing policy in Stockholm was not without its critics. Opponents expressed concern that the scheme amounted to yet another tax imposed on citizens. Others raised issues of social equity, arguing that the charges could disproportionately affect low-income individuals who might not be able to afford the tolls. There were also doubts about the long-term effectiveness of the policy, as well as questions regarding the use of the revenue generated. Some feared that the policy would merely displace congestion to surrounding areas rather than eliminate it.

Over time, however, the evidence has pointed toward sustained positive effects. The initial reduction in traffic not only persisted but even increased slightly in the years following implementation. This suggests that the changes in behavior were not merely temporary reactions but more lasting adjustments. Additionally, the congestion pricing policy stimulated a notable increase in the adoption of environmentally friendly vehicles. Until the end of 2008, cars powered by alternative fuels were exempted from the charge, which provided a clear incentive for cleaner vehicle choices and contributed to broader environmental objectives.

In summary, the Stockholm example offers a powerful case study of how a carefully designed congestion pricing policy can reduce traffic congestion, shift travel behavior, and even reshape public opinion. It also highlights the importance of monitoring, transparency, and adaptability in the face of legitimate concerns about equity and effectiveness.

9.1 System optimum

Behind the practice of congestion pricing lies an economic theory grounded in the concept of externalities. Engineers and scientists play a vital role in translating this theory into actionable policies. While much attention is understandably given to the practical aspects of implementation — such as data collection, tolling technology, and enforcement infrastructure — one of the most intellectually demanding tasks is determining the appropriate price to charge. This is not merely a technical detail; it is the very heart of the congestion pricing strategy.

To understand how to calculate the "right" price, we must first consider the decision-making process of an individual traveler. Take, for instance, Pat Trafficson, who is deciding whether to drive into the city during peak hours. From her perspective, the cost of the trip consists of two components: the time she expects to spend in traffic and the monetary costs associated with the journey, such as fuel, parking, and any tolls. Naturally, she chooses the option that minimizes her personal cost, balancing time and money in a way that reflects her own preferences.

However, Pat's travel decision does not occur in isolation. Her presence on the road has an impact on other travelers. By joining the stream of traffic, she contributes to congestion, thereby increasing travel times for everyone else. This effect is not accounted for in her personal cost calculation, yet it imposes a real cost on society. As introduced in Section 1.1, economists refer to this as a *negative externality* — a situation in which the actions of one individual impose unpriced costs on others.

Congestion pricing is designed to internalize this externality. In other words, Pat should face not only the direct costs of her trip, but also the additional costs that her trip imposes on others. This aligns her private incentives with the broader public good. The theoretical foundation for this idea is often referred to as the *polluter pays principle*, which holds that those who generate external costs should bear the responsibility for them.

But how much should Pat be charged? The answer lies in estimating the value of the deterioration in service quality that she causes by entering the congested network. More formally, the "right" price is equal to the

marginal external cost — the additional delay or inconvenience inflicted on all other users, expressed in monetary terms. Calculating this requires data and models capable of capturing the relationship between traffic volume and travel time, as well as information about how travelers value their time.

As we have seen in Chapter 8, the *user equilibrium* reflects a decentralized decision-making framework. In this setting, each traveler chooses the route that minimizes their own travel time or cost, without considering the broader impact of their choice on others. This situation is often described as a *selfish routing* scenario. As explained in Section 8.4, the user equilibrium can be mathematically formulated as the solution to the following optimization problem:

$$y^* = \operatorname{argmin}_y \sum_\ell \int_0^{x_\ell} t_\ell(z)\, dz$$

subject to the constraints:

$$\sum_p y_p^q = f_q \quad \forall q, \qquad y_p^q \geqslant 0 \quad \forall p, q.$$

Here, the variable y_p^q represents the flow of travelers with origin-destination pair q using path p, and f_q is the total demand between those points. The total flow on link ℓ is denoted x_ℓ, and $t_\ell(x_\ell)$ is the travel time on that link as a function of the flow.

In contrast, the *system optimum* represents a centralized approach where the objective is to minimize the total travel time experienced by all users in the system, treating them as a coordinated whole. Instead of acting selfishly, travelers are assumed to select routes in a way that optimizes the collective outcome. The corresponding optimization problem is:

$$\tilde{y}^* = \operatorname{argmin}_y \sum_\ell x_\ell\, t_\ell(x_\ell)$$

subject to the same demand conservation and non-negativity constraints:

$$\sum_p y_p^q = f_q \quad \forall q, \qquad y_p^q \geqslant 0 \quad \forall p, q.$$

In this formulation, the objective function directly computes the total travel time over all links, since $x_\ell t_\ell(x_\ell)$ represents the travel time incurred by all users traversing link ℓ. The solution to this problem leads to a flow pattern where the overall efficiency of the network is maximized, potentially at the expense of some individual users experiencing longer routes than they would under a purely selfish strategy.

The difference between these two outcomes is not merely theoretical. The user equilibrium, while stable under individual decision-making, is generally inefficient from a global perspective. This inefficiency is captured by the concept known as the *price of anarchy*, defined as the ratio in total cost between the user equilibrium and the system optimum:

$$\sum_{\ell} x_\ell^* t_\ell(x_\ell^*) / \sum_{\ell} \tilde{x}_\ell^* t_\ell(\tilde{x}_\ell^*) \geqslant 1.$$

For engineering applications, it is useful to look at the difference instead of the ratio,

$$\sum_{\ell} x_\ell^* t_\ell(x_\ell^*) - \sum_{\ell} \tilde{x}_\ell^* t_\ell(\tilde{x}_\ell^*) \geqslant 0,$$

where x^* (resp. $\tilde{x}^*$) denotes the vector of link flows corresponding to the user equilibrium path flow y^* (resp. $\tilde{y}^* x$).

This difference quantifies the loss in system performance due to the lack of coordination among users. It provides a rigorous way to evaluate how much worse a decentralized, self-optimizing system performs compared to a centrally coordinated one.

To gain a deeper understanding of the behavioral assumptions behind the user equilibrium and system optimum formulations, it is instructive to examine the marginal costs associated with each approach. These marginal costs can be obtained by computing the derivative of the objective function with respect to the path flows y_p^q, which represent the number of travelers between an origin-destination (OD) pair q using path p.

Let us first consider the case of the user equilibrium. The objective function in this formulation is given by the sum of integrals:

$$\sum_{\ell} \int_0^{x_\ell} t_\ell(z)\, dz,$$

where x_ℓ is the total flow on link ℓ, and $t_\ell(x_\ell)$ is the travel time on that link. The integral reflects the cumulative cost experienced by all users on link ℓ, accounting for the fact that congestion increases travel time. Since the total flow on each link x_ℓ depends on the OD path flows y_p^q, we apply the chain rule to compute the derivative:

$$\frac{\partial}{\partial y_p^q} \left[\sum_{\ell'} \int_0^{x_{\ell'}} t_{\ell'}(z)\, dz \right] = \sum_{\ell} \frac{\partial x_\ell}{\partial y_p^q} \cdot t_\ell(x_\ell).$$

The term $\frac{\partial x_\ell}{\partial y_p^q}$ is equal to 1 if path p uses link ℓ, and 0 otherwise. This is captured using the entry $P_{\ell p}$ of the link-path incidence matrix, which equals

1 when link ℓ belongs to path p. The resulting marginal cost of assigning an additional unit of flow to path p is:

$$\sum_{\ell} P_{\ell p}\, t_\ell(x_\ell) = c_p^q,$$

which corresponds to the actual travel cost experienced by a user on path p. This confirms that, under user equilibrium, travelers choose routes based on their own perceived travel time, aiming to minimize their individual cost.

In contrast, the system optimum seeks to minimize the total travel cost for all users in the network. The corresponding objective function is:

$$\sum_{\ell} x_\ell\, t_\ell(x_\ell),$$

which directly represents the aggregate travel time on each link. Applying the chain rule again, we obtain:

$$\frac{\partial}{\partial y_p^q}\left[\sum_{\ell'} x_{\ell'} t_{\ell'}(x_{\ell'})\right] = \sum_{\ell} \frac{\partial x_\ell}{\partial y_p^q} \cdot \left(t_\ell(x_\ell) + x_\ell \frac{\partial t_\ell(x_\ell)}{\partial x_\ell}\right).$$

Here again, the derivative $\frac{\partial x_\ell}{\partial y_p^q}$ is represented by $P_{\ell p}$, and the resulting marginal cost becomes:

$$\sum_{\ell} P_{\ell p} \left(t_\ell(x_\ell) + x_\ell \frac{\partial t_\ell(x_\ell)}{\partial x_\ell}\right).$$

This expression highlights an important distinction. In addition to the direct travel time $t_\ell(x_\ell)$, the system-optimal cost includes a second term $x_\ell \frac{\partial t_\ell(x_\ell)}{\partial x_\ell}$, which represents the marginal delay imposed on other users by increasing the flow on link ℓ. This term captures the externality associated with congestion: when a new traveler enters the network, they not only experience congestion, but also make the situation worse for everyone else.

To simplify notation, we can define the marginal social cost function:

$$\tilde{t}_\ell(x_\ell) = t_\ell(x_\ell) + x_\ell \frac{\partial t_\ell(x_\ell)}{\partial x_\ell}.$$

This allows us to express the system-optimal marginal cost on a path as a weighted sum of the marginal social costs on the links used by that path. The key insight is that, under system optimality, users should be routed not according to the cost they themselves experience, but according to the total cost they impose on the system. This discrepancy between individual and

social costs lies at the heart of the inefficiency observed in user equilibrium and is precisely what congestion pricing seeks to correct.

A powerful insight emerges when we reinterpret the user equilibrium through the lens of the system optimum. Suppose that instead of experiencing only the direct travel time $\mathsf{t}_\ell(\mathsf{x}_\ell)$, each user is faced with a modified cost function that includes the marginal impact of their decision on the rest of the system. This adjusted cost function is defined as

$$\tilde{\mathsf{t}}_\ell(\mathsf{x}_\ell) = \mathsf{t}_\ell(\mathsf{x}_\ell) + \mathsf{x}_\ell \frac{\partial \mathsf{t}_\ell(\mathsf{x}_\ell)}{\partial \mathsf{x}_\ell}.$$

This quantity represents the *marginal social cost* on link ℓ: it includes not only the travel time experienced by the user, but also the additional delay their presence causes to all other users on the link. Remarkably, if each traveler were to minimize this marginal social cost instead of their own private cost, the resulting traffic pattern would coincide with the system optimum. In other words, the user equilibrium under the marginal social cost function is equivalent to the system optimum under the standard cost function.

This observation forms the theoretical foundation of congestion pricing. To align individual behavior with the social optimum, it is sufficient to modify the cost structure faced by travelers so that it reflects the full consequences of their decisions. The required adjustment is straightforward: each user should be charged an additional cost equal to

$$\mathsf{x}_\ell \frac{\partial \mathsf{t}_\ell(\mathsf{x}_\ell)}{\partial \mathsf{x}_\ell},$$

which represents the delay they impose on all other users by contributing to the flow on link ℓ. This term captures the *external cost* of congestion, and charging it to the user ensures that their routing decisions reflect not only personal preferences but also societal impacts.

This pricing principle is analogous to the well-known "polluter pays" concept in environmental economics: those who impose costs on others should be held financially responsible. By applying this idea to traffic networks, congestion pricing becomes a mechanism for converting inefficient, self-interested behavior into socially optimal outcomes—without requiring central planning or coercive regulation.

Let us now return to the example introduced in Section 8.5.1. The cost functions for each link ℓ are denoted $\mathsf{t}_\ell(\mathsf{x}_\ell)$, where x_ℓ is the flow on that link. For this example, the cost functions are given as follows:

$$
\begin{aligned}
t_1(x_1) &= 10x_1, & \tilde{t}_1(x_1) &= 10x_1 + x_1 \cdot 10 = 20x_1,\\
t_2(x_2) &= 50 + x_2, & \tilde{t}_2(x_2) &= 50 + x_2 + x_2 = 50 + 2x_2,\\
t_3(x_3) &= 50 + x_3, & \tilde{t}_3(x_3) &= 50 + x_3 + x_3 = 50 + 2x_3,\\
t_4(x_4) &= 10x_4, & \tilde{t}_4(x_4) &= 10x_4 + x_4 \cdot 10 = 20x_4,\\
t_5(x_5) &= 10 + x_5, & \tilde{t}_5(x_5) &= 10 + x_5 + x_5 = 10 + 2x_5.
\end{aligned}
$$

The first column in each pair defines the *private cost function*, which represents the travel time experienced by an individual user on that link. These are the functions used by travelers when making route choices under the user equilibrium assumption.

The second column gives the corresponding *marginal social cost function* $\tilde{t}_\ell(x_\ell)$. As discussed earlier, this function accounts not only for the travel time experienced by the user, but also for the additional delay imposed on other users due to increased congestion. It is computed by adding to the private cost the product of the flow x_ℓ and the derivative of the private cost function with respect to flow:

$$\tilde{t}_\ell(x_\ell) = t_\ell(x_\ell) + x_\ell \cdot \frac{dt_\ell}{dx_\ell}(x_\ell).$$

For example, on link 1, the travel time increases linearly with flow, $t_1(x_1) = 10x_1$, and its derivative is constant: $\frac{dt_1}{dx_1} = 10$. Thus, the marginal social cost is $\tilde{t}_1(x_1) = 10x_1 + 10x_1 = 20x_1$. On links 2 and 3, which represent facilities with a base travel time and a linear congestion term, the derivative is 1, leading to marginal costs that are simply the original function plus an additional x_ℓ.

To illustrate the effect of using marginal social cost functions in a traffic assignment problem, consider the flow pattern depicted in Figure 9.4. This flow configuration corresponds to a user equilibrium computed with respect to the marginal costs $\tilde{t}_\ell(x_\ell)$, rather than the private costs $t_\ell(x_\ell)$. As established earlier, when users respond to marginal costs, the resulting equilibrium flow pattern coincides with the system optimum.

In this example, there are two origin nodes, labeled with orange values on the left: origin 1 with a demand of 4 units, and origin 2 with a demand of 2 units. The destinations are shown on the right, with demand values 5 and 1, respectively. Each directed link is annotated with its flow x_ℓ and corresponding marginal cost $\tilde{t}_\ell(x_\ell)$, both of which have been determined by solving the user equilibrium using the marginal cost functions.

The resulting flow pattern satisfies all demand constraints and route choices are consistent with the assumption that each traveler selects a path

minimizing their perceived marginal cost. Notably, link 5 remains unused, reflecting the fact that although a link may be physically available, it may not be efficient from the system's point of view when users are charged the full cost of their decisions.

The total marginal costs experienced by travelers for each OD pair are indicated at the bottom of the figure. Specifically, the system-optimal path cost from origin 1 to destination 1 is $\tilde{c}^*_{11} = 116$, from origin 1 to destination 2 is $\tilde{c}^*_{12} = 118$, and from origin 2 to destination 1 is $\tilde{c}^*_{21} = 117$. These values represent the total costs of the chosen routes under the system-optimal assignment, and can be used to evaluate or compare the efficiency of different flow patterns.

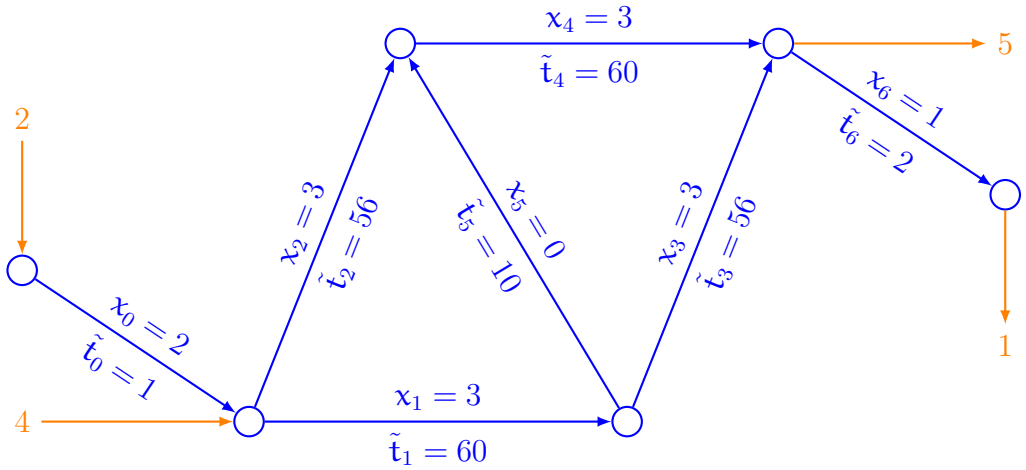


Figure 9.4: System-optimal user equilibrium with marginal cost functions. The flows and marginal costs on each link reflect the outcome when users respond to full social costs rather than private costs. The total perceived costs for each OD pair are $\tilde{c}^*_{11} = 116$, $\tilde{c}^*_{12} = 118$, and $\tilde{c}^*_{21} = 117$.

The equilibrium flow pattern obtained using the updated performance (cost) functions is reported in Table 9.1. The structure of the table is the same as Table 8.2 and the others presented in Section 8.2.

9.2 From theory to practice

While the theory behind congestion pricing is conceptually sound and mathematically elegant, translating it into practice presents a number of significant challenges. The theoretical framework relies on the idea that each traveler should be charged a toll equal to the external cost they impose on others.

p	flow	x_0	x_1	x_2	x_3	x_4	x_5	x_6	$\tilde{t}_0$	$\tilde{t}_1$	$\tilde{t}_2$	$\tilde{t}_3$	$\tilde{t}_4$	$\tilde{t}_5$	$\tilde{t}_6$	cost
r_1, s_1: $f_{rs} = 3$																
	1		3		3					60		56				116
	2			3		3					56		60			116
	0		3			3	0			60			60	10		130
r_1, s_2: $f_{rs} = 1$																
	1		3		3			1		60		56			2	118
	0			3		3		1			56		60		2	118
	0		3			3	0	1		60			60	10	2	132
r_2, s_1: $f_{rs} = 2$																
	1	2	3		3				1	60		56				117
	1	2		3		3			1		56		60			117
	0	2	3			3	0		1	60			60	10		131

Table 9.1: User equilibrium computed with marginal cost functions. Same structure as Table 8.2 and others in Section 8.2.

This cost is derived from the marginal impact of their presence on the link and is expressed as

$$x_\ell \frac{\partial t_\ell(x_\ell)}{\partial x_\ell},$$

where x_ℓ is the flow on link ℓ, and $t_\ell(x_\ell)$ is the corresponding travel time. This term represents the additional delay inflicted on other users by increasing the traffic volume on that link.

In theory, this value should be charged to each user in monetary form. If the cost is expressed in units of time, it must first be converted into a monetary value using a value-of-time conversion factor (see Chapter 3). The idea is to reflect the fact that time has economic value, and that different travelers may perceive this value differently depending on income, trip purpose, or urgency.

However, the practical implementation of this principle is far from straightforward. One major obstacle is the technical and financial difficulty of equipping all links in a transportation network with tolling infrastructure. While it is feasible to install gantries or sensors on key arterial roads or bridges, doing so across an entire network is often prohibitively expensive and logistically complex.

A second issue relates to the timing of information. For congestion pricing to be effective, travelers must know the tolls before making their route

choices. This is particularly challenging in systems that rely on dynamic pricing, where tolls vary in real-time according to traffic conditions. If toll information is updated too frequently or is communicated too late, travelers may be unable to incorporate it into their decisions, defeating the purpose of the pricing mechanism.

Finally, while the value-of-time is a central concept in the theoretical framework, it varies significantly across individuals and situations. A business traveler may place a high value on saving time, while a tourist may be more willing to accept delays. Similarly, the length and purpose of a trip can influence how much a traveler is willing to pay to avoid congestion. In practice, it is neither desirable nor politically feasible to charge different tolls to different individuals based on personal characteristics. As a result, most systems apply uniform tolls that approximate average external costs, even though this introduces inefficiencies compared to the ideal pricing scheme.

These practical limitations mean that real-world implementations of congestion pricing often involve compromises. Instead of perfectly reflecting marginal social costs, tolls are typically set using simplified models and applied only on selected links or during peak periods. Despite these imperfections, congestion pricing remains a powerful tool for managing demand and improving the efficiency of transportation systems. The challenge lies in designing systems that are both theoretically grounded and practically implementable.

9.3 Summary

This chapter has explored the concept of congestion pricing from both theoretical and practical perspectives. At its core, the theoretical framework is based on a simple yet powerful idea: travelers should be charged for the congestion they cause. Each additional vehicle on a road contributes to longer travel times for others, and this external cost is not accounted for in individual route choices under normal conditions. By charging a toll equal to the marginal congestion cost, it is possible to internalize this externality, thereby aligning private incentives with the social optimum.

Mathematically, we have seen that the system-optimal traffic assignment, which minimizes the total travel time across the network, can be reproduced through a user equilibrium if travelers respond not to their private travel times, but to adjusted cost functions that include the marginal social cost. This result provides a theoretical justification for congestion pricing: if users experience these corrected costs – either through information or monetary charges — their selfish behavior leads to a collectively optimal outcome.

In practice, however, the implementation of congestion pricing faces a number of challenges. The chapter examined the case of Stockholm, where a congestion pricing scheme was successfully introduced after a trial period and a public referendum. This example illustrated that, while technically feasible, such policies are politically sensitive and can provoke significant public debate. Issues such as fairness, transparency, and the use of revenues must be carefully managed to build and maintain public support.

Moreover, the practical realization of the theoretical ideal is constrained by technological and behavioral complexities. It is difficult to equip every link in a road network with tolling infrastructure, and even more difficult to communicate dynamically changing tolls to travelers in a timely and comprehensible way. Additionally, the value of time varies across individuals and contexts, but charging personalized tolls is both controversial and ethically problematic.

Despite these difficulties, congestion pricing remains one of the most effective tools for managing demand in urban transportation systems. It offers a mechanism to reduce congestion, encourage modal shifts, and improve overall system efficiency. The key lies in finding a balance between economic theory and political and operational feasibility, designing systems that are grounded in rigorous analysis but flexible enough to work in the real world.

Chapter 10

Freight transportation: a short introduction

Freight transportation plays a critical role in the functioning of modern economies. It ensures that raw materials reach factories, that finished goods are delivered to stores, and that essential supplies are distributed across regions and countries. Freight transport and passenger transport differ in fundamental ways that justify distinct approaches to their study and management.

One of the most striking differences between transporting people and goods lies in the nature of the transported entities themselves. While individuals make travel decisions based on personal preferences, constraints, and behavioral tendencies, goods have no behavior of their own. A container of electronics or a pallet of food does not choose how, when, or by whom it is transported. Instead, all relevant decisions are made by shippers, carriers, logistics providers, and supply chain managers.

This leads to another key distinction: decision-making in freight transportation is typically centralized or at least highly coordinated. Unlike personal travel, where decisions are distributed across millions of individuals, freight movements are often the result of deliberate, optimized planning. A single logistics manager may determine the routing of thousands of tons of goods, leveraging information systems, contractual relationships, and operational constraints to do so efficiently.

Furthermore, the metrics that govern freight transportation are different from those that apply to passenger transport. While travelers value time, convenience, comfort, and even aesthetics, the movement of goods is primarily driven by economic considerations. Cost is often the dominant criterion in choosing between modes, routes, or carriers. Other factors such as reliability, speed, and risk may play a role, but they are usually evaluated through their

impact on cost or service level commitments.

Comfort, convenience, and user experience — central to the design and analysis of passenger transportation systems — are largely irrelevant in freight. What matters instead is how to move goods from origin to destination in a way that meets business objectives, complies with regulations, and minimizes disruption. This focus on efficiency and coordination fundamentally shapes the tools, models, and policies used in freight transport planning.

Firms involved in freight transportation make a wide range of decisions at different temporal and organizational scales. These decisions are illustrated in Figure 10.1, which extends the decision-making framework introduced in Figure 5.1 typically used for passenger transport to include the freight domain.

At the long-term level, firms engage in strategic planning activities that shape the entire structure of their logistics and supply chain operations. One of the most critical decisions at this level concerns the *design of the supply chain*, including the number, type, and location of facilities such as warehouses, distribution centers, and production sites. These choices determine the spatial distribution of logistics infrastructure and heavily influence operational costs and service levels for years to come.

In the midterm, firms must translate their strategic infrastructure into operational capabilities. This includes decisions about *fleet size* — how many vehicles to operate and of what types — as well as the *transport modes* to be used, such as road, rail, air, or maritime freight. These choices are shaped by expected demand, regulatory constraints, contractual agreements, and cost considerations. Unlike long-term decisions, which are relatively fixed, midterm decisions can be revised periodically to adapt to evolving conditions.

Finally, in the short term, firms handle the day-to-day execution of freight movements. This involves solving *vehicle routing problems*, scheduling pickups and deliveries, and responding to operational disruptions. Efficient *routing and scheduling* are essential to minimize costs, meet delivery time windows, and ensure customer satisfaction. These decisions are increasingly supported by real-time information systems that allow for dynamic adjustments in response to traffic, weather, or unexpected events.

Overall, the freight sector operates through a hierarchy of interdependent decisions that must be carefully coordinated. Each level builds on the previous one, and improvements at one level can unlock efficiencies at others.

The concepts of logistics and supply chain management are central to the organization and efficiency of freight transportation systems. Although often used interchangeably in casual conversation, these terms refer to distinct but related aspects of the movement and coordination of goods.

Logistics refers to the management of the flow of goods, information, and

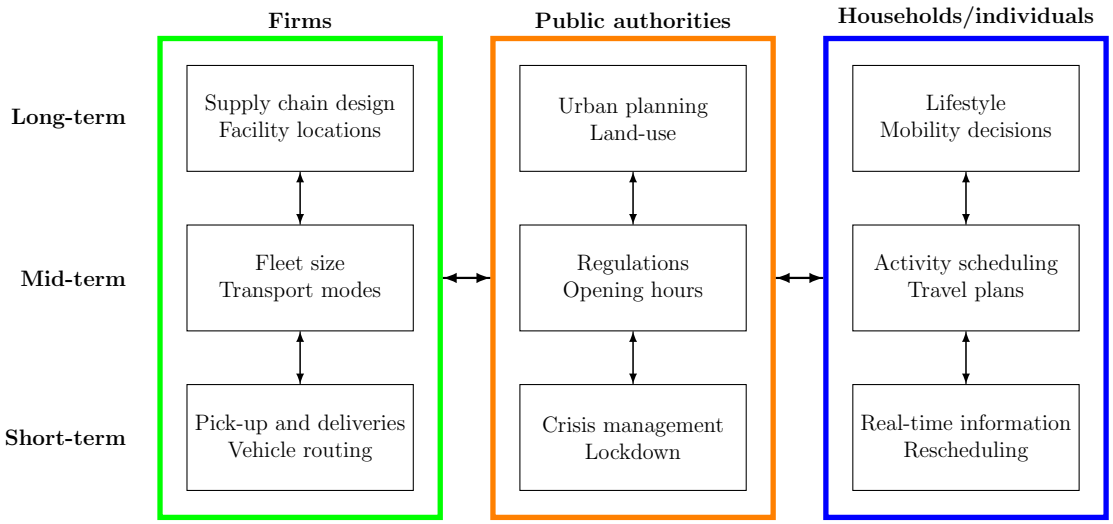


Figure 10.1: Choices and decisions made by public authorities, households, and firms. This figure extends the decision framework introduced earlier to include freight-related decisions made by firms at different time scales.

resources between the point of origin and the point of consumption. Its focus is on planning, implementing, and controlling these flows efficiently and effectively to meet customer requirements. This includes tasks such as inventory management, warehousing, order fulfillment, transportation, and distribution. In essence, logistics is concerned with ensuring that the right products are delivered in the right quantity, to the right place, at the right time. The underlying philosophy emphasizes precision, reliability, and responsiveness to demand.

From an organizational perspective, logistics typically operates within a single entity. For example, a retailer may manage its own inventory and coordinate the delivery of goods from its warehouses to its stores or directly to customers. In this context, logistics is a function internal to the firm, focused on optimizing operations within a well-defined organizational boundary.

In contrast, the term *supply chain* encompasses a broader and more complex network of interactions among multiple actors. A supply chain includes not just the internal logistics of a single firm, but also the coordination between suppliers, manufacturers, distributors, retailers, and customers. It involves the movement of goods, the exchange of information, and the management of financial flows across organizational boundaries. Effective supply chain management requires collaboration and synchronization among all stakeholders to ensure that the final product reaches the end user in a timely and cost-effective manner.

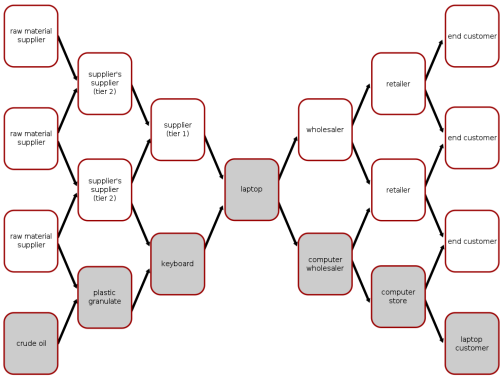


Figure 10.2: Supply chain

While logistics is a critical component of the supply chain, the supply chain itself extends beyond logistics to include strategic activities such as sourcing, procurement, production planning, and customer relationship management. The distinction can be summarized by noting that logistics is often the responsibility of one actor, whereas the supply chain represents a system of interdependent organizations working together to create and deliver value.

I this course, we provide an introduction to the fundamental decision problems encountered in the field of freight transportation and logistics. To offer a structured perspective, we examine these problems across three key time horizons: long-term, medium-term, and short-term. Each of these time scales corresponds to distinct types of decisions, methods of analysis, and operational objectives.

At the *long-term* level, we focus on the *facility location problem*. This is a strategic decision that involves determining where to place warehouses, distribution centers, or other logistical facilities. These choices are typically infrequent but highly consequential, as they have long-lasting effects on the structure and performance of supply chains.

In the *medium-term*, we turn to *inventory management*. This area deals with determining how much stock to hold, when to reorder, and how to balance the trade-off between holding costs and service levels. Inventory decisions must account for variability in demand, lead times, and storage capacity.

At the *short-term* level, we address the *vehicle routing problem* (VRP), a classic and widely studied problem in operations research. The VRP concerns the optimal assignment and sequencing of deliveries or pick-ups using a fleet of vehicles. It is central to day-to-day logistics operations such as parcel delivery, food distribution, or waste collection.

By analyzing these three problems — facility location, inventory management, and vehicle routing — we cover a representative set of decisions that define the performance and efficiency of freight systems across time scales.

10.1 Facility location

One of the most fundamental strategic problems in logistics is the *facility location problem*. This problem arises whenever an organization must decide where to locate its distribution centers, warehouses, or service depots in order to efficiently serve a set of customers. These long-term decisions have significant implications for cost, service quality, and operational flexibility, and they are difficult to reverse once implemented.

To formulate the problem, we begin with a set of customers, denoted by $\mathcal{C}$. Each customer $\mathtt{j} \in \mathcal{C}$ is associated with a demand quantity $\mathtt{d}_\mathtt{j}$, representing how much they require over a certain planning horizon. The goal is to serve these demands from a set of potential depot locations, represented by $\mathcal{D}$. These are candidate sites where facilities could be opened, but not all of them will necessarily be used.

Each potential depot $\mathtt{i} \in \mathcal{D}$ is associated with a *setup cost* $\mathtt{c}_\mathtt{i}$, which is the fixed cost incurred if the facility is opened. In addition, each depot has a maximum *capacity* $\ell_\mathtt{i}$, indicating the total demand it can serve. Finally, for each pair $(\mathtt{i}, \mathtt{j}) \in \mathcal{D} \times \mathcal{C}$, there is a *trip duration* or transportation cost $\mathtt{t}_{\mathtt{ij}}$ that quantifies the effort required to serve customer $\mathtt{j}$ from depot $\mathtt{i}$. This could represent distance, time, or monetary cost.

The facility location problem involves two interconnected decisions. First, we must determine which subset of the potential depots should be opened. This decision balances the fixed setup costs against the benefits of having facilities close to customers. Second, for each customer, we must decide from which depot they will be served. This assignment must respect capacity limits and aims to minimize transportation costs or durations.

The result is a combinatorial optimization problem in which the goal is typically to minimize the total cost, combining setup costs and service costs, while satisfying customer demands and facility constraints. The problem can be formulated mathematically and solved using exact algorithms for small instances or heuristics and approximation methods for larger, real-world applications.

To better understand the facility location problem, let us consider a concrete example inspired by a real-world urban context. Suppose a logistics company is responsible for delivering goods to a set of 20 retail shops located throughout the city of Lausanne. These shops represent the customers to

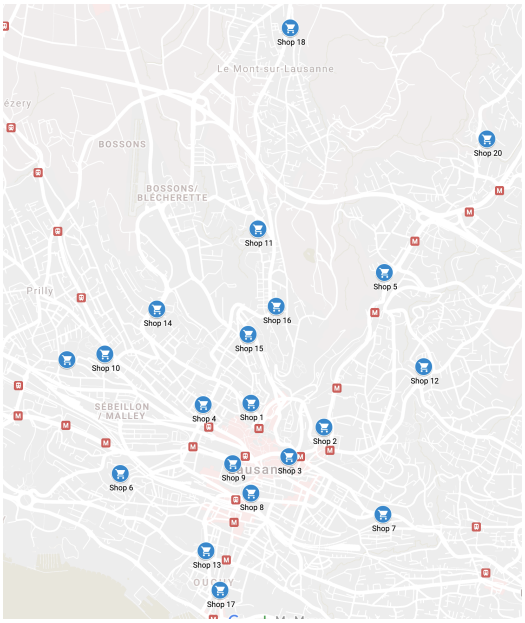


Figure 10.3: Clients: 20 shops in Lausanne

be served, and their geographic distribution is shown in Figure 10.3. Each shop requires regular deliveries, and the company must determine how best to structure its distribution network to meet this demand efficiently.

Rather than operating from a single central warehouse, the company has identified five possible sites within or near the city where depots could be established. These potential depot locations are illustrated in Figure 10.4. Each site offers a different trade-off in terms of cost, accessibility, and capacity, and the company must choose which of these to activate in its final network design.

Figure 10.5 shows an example of a feasible and reasonably efficient solution to this problem. In this particular configuration, the company has chosen to open 4 out of the 6 available depot sites. Each open depot is assigned exactly 5 clients to serve.

This example highlights the two central decisions in the facility location problem: selecting which facilities to open and determining how to assign customers to those facilities.

We now present a formal mathematical model for the facility location problem introduced earlier. This model captures the essential trade-offs between fixed infrastructure costs and operational delivery costs, while respecting service and capacity constraints. It allows us to determine both which depots to open and how to assign customer demand to the selected facilities

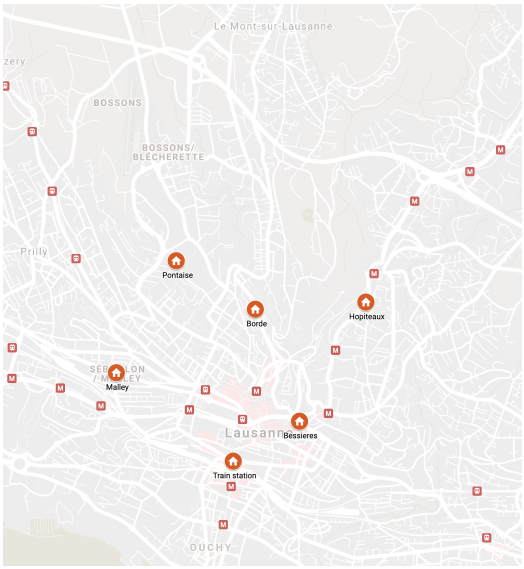


Figure 10.4: Depots: 6 potential locations

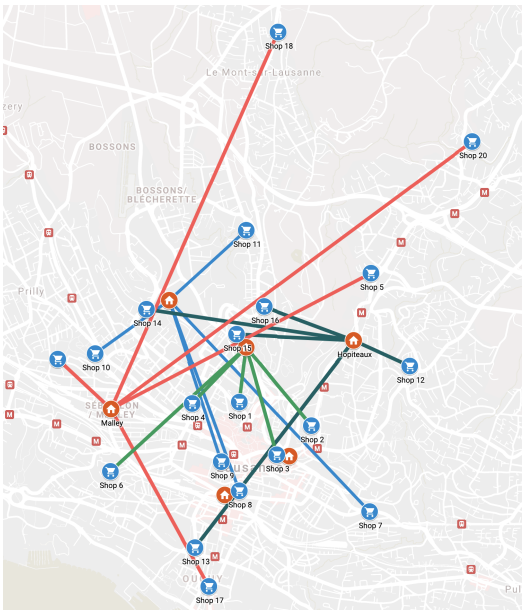


Figure 10.5: Example of solution: 4 depots are open, each serving 5 clients

in a cost-effective manner.

Let $\mathcal{C}$ denote the set of customers to be served, and $\mathcal{D}$ the set of potential depot locations. Each customer $j \in \mathcal{C}$ has a known demand d_j, and each depot $i \in \mathcal{D}$ is associated with a fixed setup cost c_i and a capacity limit ℓ_i. The travel time or transportation cost between depot i and customer j is denoted by t_{ij}, and γ represents the cost per unit of demand per unit of travel time.

The model uses two sets of decision variables. The binary variable $x_i \in \{0, 1\}$ indicates whether depot i is opened ($x_i = 1$) or not ($x_i = 0$). The continuous variable $y_{ij} \in \mathbb{R}_+$ denotes the proportion of customer j's demand that is served by depot i. The quantity $d_j y_{ij}$ thus represents the actual volume of goods transported from depot i to customer j.

The objective of the model is to minimize the total cost, which is composed of two parts: the fixed costs of opening depots, and the variable costs of transporting goods from depots to customers. This leads to the following objective function:

$$\min_{x,y} \sum_{i \in \mathcal{D}} c_i x_i + \gamma \sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{C}} t_{ij} d_j y_{ij}.$$

The first term represents the total fixed cost incurred for opening depots. The second term accounts for the transportation cost, where each unit of demand incurs a cost proportional to the travel time between the assigned depot and the customer.

The model is subject to several constraints. First, each customer's demand must be fully satisfied, which means the total proportion of their demand served across all depots must equal one:

$$\sum_{i \in \mathcal{D}} y_{ij} = 1, \quad \forall j \in \mathcal{C}.$$

Second, no depot may exceed its capacity. The total demand served by depot i cannot surpass ℓ_i, and this constraint only applies if the depot is open:

$$\sum_{j \in \mathcal{C}} d_j y_{ij} \leqslant \ell_i x_i, \quad \forall i \in \mathcal{D}.$$

Finally, we enforce the binary nature of the facility-opening decision and the non-negativity of the assignment variables:

$$\begin{aligned} x_i &\in \{0, 1\}, && \forall i \in \mathcal{D}, \\ y_{ij} &\geqslant 0, && \forall i \in \mathcal{D}, \forall j \in \mathcal{C}. \end{aligned}$$

This formulation is a classic example of a mixed-integer linear problem (MILP), combining binary decisions with continuous variables.

10.1.1 Numerical example

To illustrate the facility location model with a concrete numerical example, we consider the scenario involving 20 shops located throughout the city of Lausanne. These shops must be supplied from a set of candidate depot locations. Table 10.1 reports the travel times, in seconds, between each shop and each of the six potential depot sites: Bessières, Borde, Hôpitaux, Malley, Pontaise, and the train station. These values were obtained from OpenStreetMap and reflect realistic travel durations under typical traffic conditions.

This data serves as input for the transportation component of the facility location model. Specifically, the travel times t_{ij} between depot i and customer j determine the variable component of the cost function.

	Bessières	Borde	Hôpitaux	Malley	Pontaise	Train station
Shop 1	141.7	125.8	262.3	177.6	121.4	220.1
Shop 2	92.5	197.2	115.2	319.9	320.4	248.7
Shop 3	139.2	177.7	257.9	207.8	300.9	192.4
Shop 4	247.9	227.8	385.2	149.9	181.8	192.4
Shop 5	237.4	342.1	88.0	464.8	366.3	393.6
Shop 6	257.6	321.2	394.9	146.0	275.2	157.1
Shop 7	98.3	242.5	249.8	285.7	350.2	147.3
Shop 8	151.8	289.5	297.0	198.0	273.8	17.3
Shop 9	113.3	243.1	250.6	153.8	229.6	116.7
Shop 10	357.7	361.2	495.0	170.9	239.4	321.7
Shop 11	326.1	184.7	416.1	366.7	215.3	409.2
Shop 12	265.3	370.0	277.8	492.7	493.2	421.5
Shop 13	234.2	369.4	379.4	247.6	323.4	102.6
Shop 14	280.7	216.3	401.3	172.3	79.1	282.2
Shop 15	214.9	186.6	335.5	250.8	134.2	293.3
Shop 16	280.4	139.0	370.4	321.0	203.6	363.5
Shop 17	224.9	362.6	370.1	302.4	378.2	138.5
Shop 18	462.1	320.7	438.2	502.7	333.4	545.2
Shop 19	383.7	412.4	521.0	140.2	325.9	324.4
Shop 20	452.2	556.9	302.8	662.2	477.3	608.4

Table 10.1: Travel time in seconds between 20 shops in Lausanne and 6 potential depot locations.

Source: OpenStreetMap.org

Scenario 1

To explore the behavior of the facility location model in a simplified setting, we begin with a first scenario where all depot locations are made equally and fully available. In this configuration, all setup costs are set to zero, which means the model is free to open any number of depots without incurring a fixed penalty. This allows us to isolate and study the impact of transportation costs alone on the optimal facility and assignment decisions.

The setup costs and capacities for each of the six candidate depot locations are summarized in Table 10.2. All depots are given identical capacities of 50 units, which is more than sufficient to serve the total demand of 20 units. As a result, capacity constraints are not binding in this scenario.

Depot	**Setup cost**	**Capacity**
Train station	0	50
Pontaise	0	50
Hôpitaux	0	50
Malley	0	50
Bessières	0	50
Borde	0	50

Table 10.2: Scenario 1: Setup costs and capacities for candidate depots.

On the demand side, we assume that each of the 20 shops requires exactly 1 unit of goods. This uniform demand simplifies the allocation logic and ensures that no particular customer has disproportionate influence on the solution. The demand values are reported in Table 10.3.

Shop 1: 1	**Shop 2**: 1	**Shop 3**: 1	**Shop 4**: 1
Shop 5: 1	**Shop 6**: 1	**Shop 7**: 1	**Shop 8**: 1
Shop 9: 1	**Shop 10**: 1	**Shop 11**: 1	**Shop 12**: 1
Shop 13: 1	**Shop 14**: 1	**Shop 15**: 1	**Shop 16**: 1
Shop 17: 1	**Shop 18**: 1	**Shop 19**: 1	**Shop 20**: 1

Table 10.3: Scenario 1: Demand values for the 20 shops in Lausanne.

Lastly, the conversion parameter γ is set to 0.01, which translates travel time in seconds into a monetary or cost-equivalent unit for use in the objective function.

The optimal solution obtained for Scenario 1 is illustrated in Figure 10.6. In this scenario, there are no setup costs associated with opening depots, and each depot has sufficient capacity to serve all or part of the total demand.

As a result, the optimization model minimizes only the transportation cost, which is proportional to the distance (or travel time) between depots and customers.

Because there is no penalty for opening facilities, the optimal solution involves activating all six available depot locations. This allows the model to assign each customer to the closest depot, thereby minimizing delivery distances and overall transportation cost. Each customer is assigned to exactly one depot, and depot assignments reflect spatial proximity in a nearly intuitive and geographically balanced way.

The total cost for this solution is 29.44 cost units, which represents the sum of the travel cost contributions across all customer-depot pairs. This scenario serves as a benchmark for comparison with later cases where additional constraints — such as setup costs or limited capacities — will make the optimization more complex and less spatially symmetric.

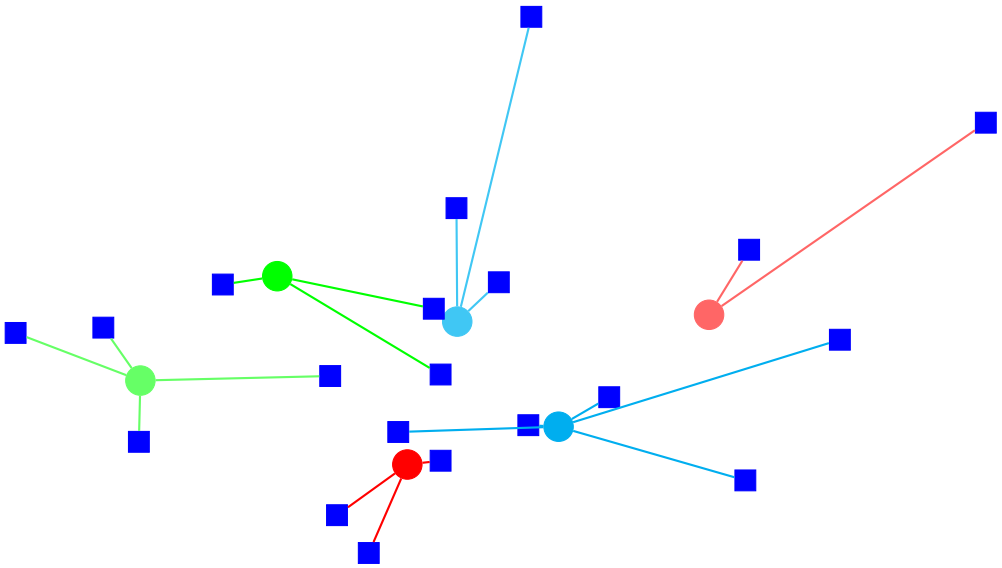


Figure 10.6: Optimal facility and customer assignment for Scenario 1. All depots are open, and each customer is served by the closest depot.

Scenario 2

In the second scenario, we introduce fixed costs for opening depots in order to more realistically reflect infrastructure or operational investments associated with establishing a logistics site. Unlike in Scenario 1, where all depots could be opened freely, we now assume that opening any of the candidate depot locations incurs a cost. This modification adds a new layer of trade-

offs to the problem, forcing the model to balance fixed facility costs against transportation costs when choosing the optimal configuration.

The setup costs and capacities for each depot are summarized in Table 10.4. All depots have the same capacity of 50 units, sufficient to serve the overall demand, and a uniform setup cost of 5 units. These fixed costs encourage the model to minimize the number of open depots, possibly leading to longer delivery distances but lower overall infrastructure costs.

Depot	**Setup cost**	**Capacity**
Train station	5	50
Pontaise	5	50
Hôpitaux	5	50
Malley	5	50
Bessières	5	50
Borde	5	50

Table 10.4: Scenario 2: Setup costs and capacities for candidate depots.

As in Scenario 1, the demand for each of the 20 shops is set to 1 unit. This uniform demand ensures comparability between scenarios and maintains a consistent workload across customers. The demand data is displayed compactly in Table 10.5.

Shop 1: 1	**Shop 2**: 1	**Shop 3**: 1	**Shop 4**: 1
Shop 5: 1	**Shop 6**: 1	**Shop 7**: 1	**Shop 8**: 1
Shop 9: 1	**Shop 10**: 1	**Shop 11**: 1	**Shop 12**: 1
Shop 13: 1	**Shop 14**: 1	**Shop 15**: 1	**Shop 16**: 1
Shop 17: 1	**Shop 18**: 1	**Shop 19**: 1	**Shop 20**: 1

Table 10.5: Scenario 2: Demand values for the 20 shops in Lausanne.

The conversion parameter remains $\gamma = 0.01$, meaning that travel times are weighted by this factor to obtain the total transportation cost component of the objective function.

In Scenario 2, the model is no longer free to open all depots; it must carefully weigh the trade-off between minimizing transportation costs and reducing the fixed costs associated with operating multiple sites.

Figure 10.7 illustrates the optimal solution for this scenario. The total cost of this solution is 49.88 units, which includes both the transportation and setup costs. Compared to Scenario 1, the optimizer chooses to open only three depots: the Train Station, Pontaise, and Hôpitaux. As a result, these depots are assigned to a larger number of shops each, including some that

are not geographically closest. This consolidation strategy allows the model to limit fixed costs at the expense of slightly higher transport distances for certain customers.

The effect of introducing setup costs is clearly visible: some previously open depots (e.g., Malley, Bessières, and Borde) remain closed in this solution, while the remaining ones serve more dispersed areas. This outcome highlights the essential role of setup costs in encouraging economies of scale in logistics planning and illustrates how such costs influence not just which depots to open, but also how the delivery workload is distributed across the network.

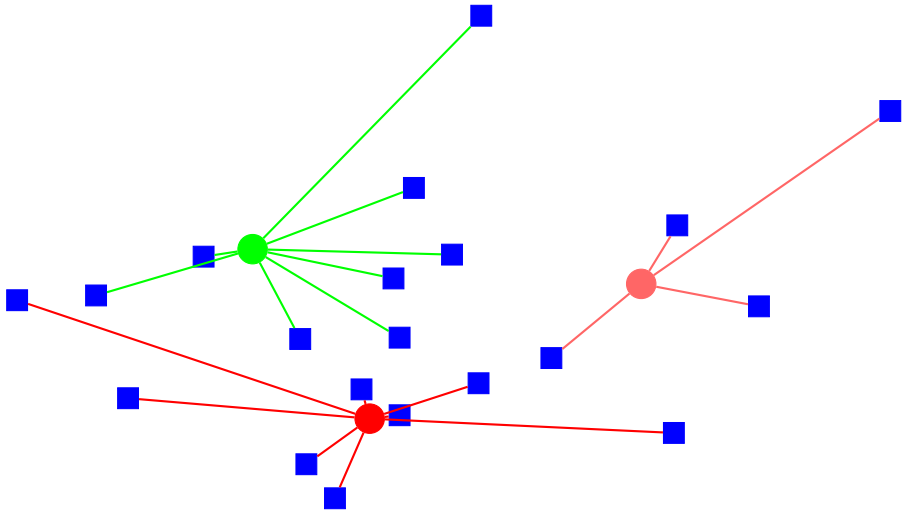


Figure 10.7: Optimal facility and customer assignment for Scenario 2. Three depots are open, balancing setup and transportation costs.

Scenario 3

In the third scenario, we build upon the previous configuration by introducing not only fixed setup costs but also strict capacity limitations at each depot. This scenario models a situation where depot sites are relatively small or constrained in their ability to process and dispatch goods, requiring a more distributed network to satisfy demand. The trade-off between opening costs and limited service capacity makes this scenario particularly illustrative of real-world logistical challenges.

Table 10.6 summarizes the setup costs and capacities of each candidate depot. As in Scenario 2, all depots have a fixed opening cost of 5 units. However, unlike before, each depot is now limited to a maximum capacity of only 5 units. Given that the total demand remains 20 units, this implies that

at least four depots must be opened to serve all shops, even in the best-case spatial configuration.

Depot	Setup cost	Capacity
Train station	5	5
Pontaise	5	5
Hôpitaux	5	5
Malley	5	5
Bessières	5	5
Borde	5	5

Table 10.6: Scenario 3: Setup costs and limited capacities for candidate depots.

The demand side of the problem remains unchanged from the earlier scenarios. Each of the 20 shops in Lausanne requests 1 unit of goods, for a total demand of 20 units. The demand distribution is shown in compact form in Table 10.7.

Shop 1: 1	**Shop 2**: 1	**Shop 3**: 1	**Shop 4**: 1
Shop 5: 1	**Shop 6**: 1	**Shop 7**: 1	**Shop 8**: 1
Shop 9: 1	**Shop 10**: 1	**Shop 11**: 1	**Shop 12**: 1
Shop 13: 1	**Shop 14**: 1	**Shop 15**: 1	**Shop 16**: 1
Shop 17: 1	**Shop 18**: 1	**Shop 19**: 1	**Shop 20**: 1

Table 10.7: Scenario 3: Demand values for the 20 shops in Lausanne.

The conversion parameter remains set to $\gamma = 0.01$, translating travel times into cost units. In this scenario, the model must now balance the cost of opening more depots against the necessity to do so due to capacity constraints.

The optimal solution, presented in Figure 10.8, demonstrates how the model responds to this new challenge. The total cost of the solution is 53 units. To satisfy the demand of 20 shops with depots each limited to five customers, the model must open at least four depots. In this instance, it activates four depots: the Train Station, Hôpitaux, Malley, and Borde. Each of them is assigned exactly five customers, making full use of their allowed capacity.

This scenario highlights how the inclusion of capacity constraints can drive the solution away from the purely cost-optimal configuration found in earlier scenarios. Some customers are no longer assigned to their closest depot if it is already at capacity, which may increase their transportation

cost. Nevertheless, the solution remains feasible and efficient, demonstrating the robustness of the optimization model.

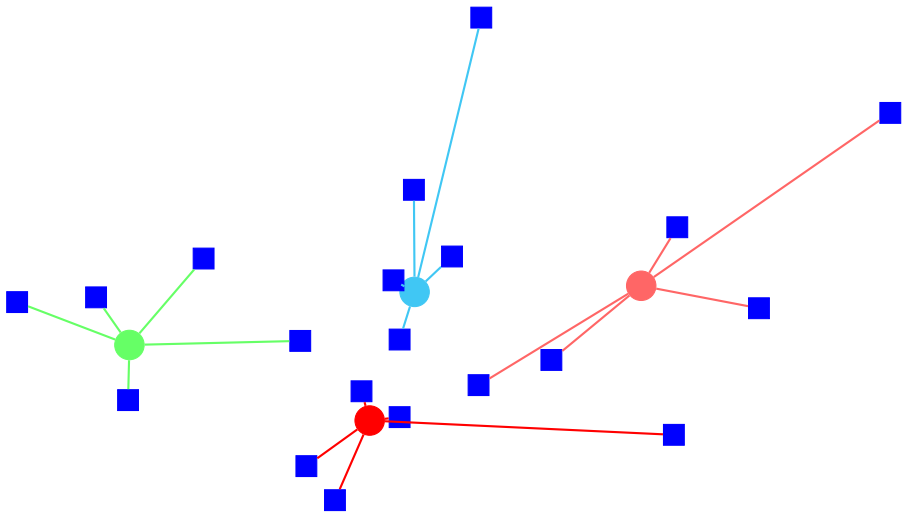


Figure 10.8: Optimal facility and customer assignment for Scenario 3. Capacity constraints require the use of four depots, each serving five customers.

Scenario 4

In the fourth scenario, we consider a situation in which one customer has a demand that exceeds the capacity of any individual depot. This introduces a new modeling challenge and showcases the flexibility of the facility location model to allow fractional assignments, where a customer is served by multiple depots.

As in previous scenarios, we assume non-zero setup costs and limited capacities for all depot locations. Table 10.8 presents the setup cost and capacity values for each candidate depot. Each depot has a capacity of 5 units and a fixed setup cost of 5 units. The overall system demand, however, now exceeds 20 units.

The demand values for the 20 shops in Lausanne are listed in Table 10.9. In contrast to the previous scenarios, Shop 1 has a demand of 6 units, while all other shops maintain a demand of 1 unit each. The total system demand is therefore 25 units, requiring at least five depots to be opened. Since the capacity of each depot is only 5 units, no single facility can fully serve Shop 1, making it necessary for this shop to be served by multiple depots.

The conversion parameter used in the objective function remains $\gamma = 0.01$, converting travel times into cost units. This scenario demonstrates how the model accommodates heterogeneous demand levels and capacity

Depot	Setup cost	Capacity
Train station	5	5
Pontaise	5	5
Hôpitaux	5	5
Malley	5	5
Bessières	5	5
Borde	5	5

Table 10.8: Scenario 4: Setup costs and capacities for candidate depots.

Shop 1: 6	**Shop 2**: 1	**Shop 3**: 1	**Shop 4**: 1
Shop 5: 1	**Shop 6**: 1	**Shop 7**: 1	**Shop 8**: 1
Shop 9: 1	**Shop 10**: 1	**Shop 11**: 1	**Shop 12**: 1
Shop 13: 1	**Shop 14**: 1	**Shop 15**: 1	**Shop 16**: 1
Shop 17: 1	**Shop 18**: 1	**Shop 19**: 1	**Shop 20**: 1

Table 10.9: Scenario 4: Demand values for the 20 shops in Lausanne.

constraints, making use of fractional assignments when required to ensure feasibility and cost efficiency.

Figure 10.9 presents the optimal solution to this scenario, with a total cost of 63.22 units. The solution shows that Shop 1 is served by three depots: Pontaise (with 1/2 of its demand), Malley (1/6), and Borde (1/3), as indicated by the annotated arrows in the diagram. All six depots are active in the solution, reflecting the high service requirements and limited capacity of each facility.

This outcome highlights the model's flexibility to allocate fractional flows when total demand exceeds individual depot capacity, ensuring that every shop is fully served while keeping the overall cost as low as possible.

10.1.2 Summary

The facility location problem is a foundational topic in logistics and transportation systems. It addresses the challenge of determining which depots should be opened and how customers should be assigned to them. This decision has significant implications for both operational efficiency and cost-effectiveness, as it directly impacts transportation costs, service quality, and the utilization of resources such as staff and infrastructure.

At the heart of the problem lies an optimization model that balances two types of costs: fixed setup costs associated with opening depots, and variable transportation costs incurred when serving customers from those

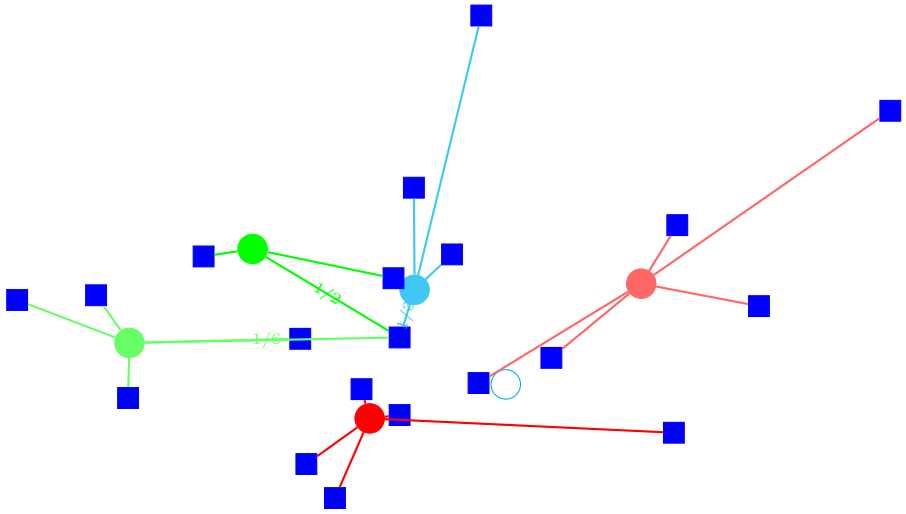


Figure 10.9: Optimal assignment in Scenario 4. Shop 1's demand is distributed across three depots due to capacity limitations.

depots. The decision-making process must also respect constraints such as depot capacity and the requirement that each customer be fully served.

Despite its relatively simple formulation, the facility location problem is inherently complex. It belongs to the class of combinatorial optimization problems, meaning that the time required to solve it grows rapidly with the size of the input data. As a result, practical applications often rely on advanced optimization solvers or heuristic methods to find high-quality solutions within a reasonable amount of time.

The basic version of the problem can be extended in numerous ways to better reflect real-world challenges. For example, demand may be uncertain or vary over time, which introduces stochastic elements into the model. Depots might have different levels of reliability, opening hours, or environmental impacts. Customers may need to be served within specific time windows or by vehicles with limited capacities. These variations give rise to a rich family of models, each tailored to specific applications in retail, healthcare, disaster relief, and beyond.

10.2 Inventory management

Inventory management is a crucial component of logistics and supply chain operations. It addresses a central issue: goods are not always consumed immediately upon production, nor are they delivered exactly when needed. Instead, items are typically stored — either temporarily or for extended

periods — on both the production and consumption sides of the supply chain. This necessary buffering introduces various types of costs that must be carefully managed.

There are two primary categories of inventory-related costs. The first involves *storage costs*, which include expenses associated with warehousing, such as physical space, equipment, staffing, and insurance. The second concerns *waiting costs*, which account for the opportunity cost of capital tied up in inventory, the risk of product obsolescence, and potential degradation or spoilage of goods over time. These costs can be significant, and minimizing them without jeopardizing service quality is a fundamental goal of inventory management.

10.2.1 Fixed consumption

To explore how inventory should be managed, we begin with a simplified context. Suppose that items are produced at a single origin point and consumed at a different destination. The demand at the consumption site is assumed to be constant, denoted by the rate f, over a fixed planning horizon of length $\mathsf{t_H}$. The question arises: how frequently should shipments be made? Should goods be shipped individually, in small batches, or in larger quantities at less frequent intervals?

Let s represent the quantity of items shipped in each batch. The *headway*, or the time between two consecutive shipments, is then given by $\mathsf{h} = \frac{\mathsf{s}}{\mathsf{f}}$, as the consumption rate remains constant. This relationship highlights the trade-off between shipment size and shipment frequency: smaller, more frequent shipments reduce inventory levels but increase transportation costs, while larger, less frequent shipments reduce transportation frequency but increase storage requirements.

Figure 10.10 provides a visual representation of the dynamics of inventory management and shipment scheduling over time. It shows how items are produced, shipped, and consumed in a stylized context with constant demand and regular shipments.

The horizontal axis represents time, while the vertical axis indicates the cumulative number of items involved in the process. Two main diagonal lines dominate the figure: one for production and one for consumption.

The *production line*, shown in red and starting at the origin, has a slope corresponding to the constant production rate. This line represents the cumulative number of items produced at the origin over time. Similarly, a second red line, shifted horizontally, represents the cumulative number of items consumed at the destination, also with a constant slope matching the consumption rate.

The items are not shipped individually as they are produced, but rather in batches. These shipments are illustrated by a stepwise curve in purple. When items are produced, they are first stored temporarily. This is shown by the horizontal segments of the step curve, during which the number of shipped items remains constant while time progresses—indicating inventory buildup. Then, all stored items are shipped at once, represented by the vertical jumps in the step function. These vertical lines correspond to instantaneous transfers of items from inventory to the shipping process.

At the destination, a similar step function describes the reception of items. After a time delay equal to the travel time $\mathtt{t}$, the items arrive at their destination, again in batches. Here too, the horizontal portions of the curve indicate waiting inventory at the receiving end before the items are gradually consumed.

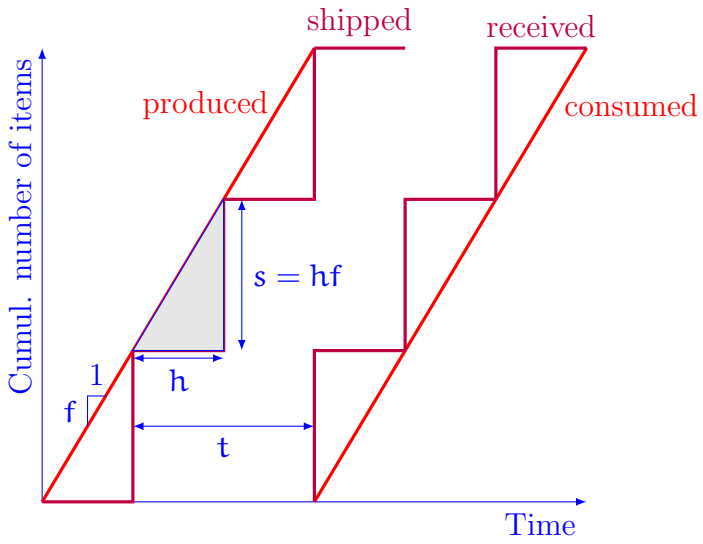


Figure 10.10: Production, shipment, and consumption over time in an inventory management system

In inventory management, the decision variable is the *shipment size*, denoted $\mathtt{s}$, which corresponds to the number of items shipped at once. Given a constant demand rate $\mathtt{f}$, and a headway $\mathtt{h}$ (i.e., the time between two consecutive shipments), the shipment size satisfies the relationship $\mathtt{s} = \mathtt{hf}$. This expresses that within each interval $\mathtt{h}$, the number of items consumed is precisely the amount shipped.

From this, we can derive the *number of shipments* required to cover a planning horizon of duration $\mathtt{t}_\mathtt{H}$. This number is given by $\mathtt{t}_\mathtt{H}/\mathtt{h}$, or equivalently $\mathtt{ft}_\mathtt{H}/\mathtt{s}$, since both expressions represent the total number of demand

units divided by the quantity per shipment.

Each shipment incurs certain costs. Let c_f be the fixed cost of transportation (regardless of the number of items), and c_v be the variable cost per item transported. Then the cost of one shipment is $c_f + c_v s$. Over the entire planning horizon, the *total shipment costs* become:

$$\frac{t_H}{h} c_f + \frac{t_H}{h} c_v s = \frac{t_H c_f}{h} + t_H c_v f.$$

Note that only the fixed component depends on the headway h; the total variable component remains constant, as it depends on the total number of items $f t_H$, regardless of how frequently they are shipped.

Next, consider *storage costs*. When items are produced but not yet shipped, they are temporarily held in inventory. Since each shipment involves a quantity $s = hf$, and items are stored on average for a time proportional to h, the storage cost at the origin is $c_r s = c_r hf$, where c_r is the storage cost per item.

Additionally, each item experiences a delay between production and consumption, consisting of the inventory time h and the travel time t. The associated *waiting costs* are given by $c_w f(h + t)$, where c_w denotes the cost per item per time unit of waiting. This includes the opportunity cost of capital, depreciation, or loss in value due to obsolescence.

Combining all these elements, the *total cost* over the planning horizon is expressed as:

$$c_r hf + c_w f(h + t) + \frac{t_H c_f}{h} + t_H c_v f.$$

This total includes storage costs, waiting costs, fixed transportation costs, and variable transportation costs.

To find the *optimal headway* that minimizes total costs, we differentiate the total cost function with respect to h. The derivative is:

$$\frac{d}{dh} \text{Total Cost} = c_r f + c_w f - \frac{t_H c_f}{h^2}.$$

Setting the derivative equal to zero and solving for h yields the optimal headway:

$$h^* = \sqrt{\frac{c_f t_H}{f(c_r + c_w)}}.$$

This expression reveals the underlying economics of the inventory decision. When the fixed cost c_f is high, as is the case with large container ships or bulk transport, the optimal strategy is to ship in large batches infrequently — leading to a larger headway. Conversely, when holding costs

$c_r + c_w$ are dominant, as in the case of expensive or perishable goods, the optimal solution involves more frequent shipments with smaller quantities, thereby reducing headway.

This trade-off lies at the heart of many real-world logistics decisions, from maritime shipping to just-in-time production systems, and illustrates the importance of quantitative models in guiding operational strategy.

10.2.2 Variable consumption

The simple inventory model described earlier assumes that production and consumption occur at identical rates over time, resulting in perfect synchronization. However, this assumption is often unrealistic. In real-world systems, even if the average production rate equals the average consumption rate, fluctuations inevitably occur. Sometimes, consumption temporarily outpaces production, and sometimes the reverse happens. These mismatches generate new challenges that must be accounted for in effective inventory planning.

Figure 10.11 illustrates this situation. The horizontal axis represents time, and the vertical axis represents the cumulative number of items produced or consumed. The solid red line on the left shows cumulative production, which increases steadily over time, reflecting a constant production rate. In contrast, the solid red line represents the cumulative demand, which follows a piecewise linear trajectory with variable slope. These variations capture the fact that consumers do not always consume at a uniform rate. The dotted red line shows the consumption at constant rate assumed in the previous model. Note that the two line meet at the end, illustrating that the total consumption is the same as before.

The purple staircase line again represents the shipment process, with inventory accumulating over time at the production location, and being dispatched in batches. The corresponding staircase curve at the receiving end shows when items arrive and are available for consumption.

However, unlike the earlier scenario where production and consumption matched exactly at all times, the discrepancy here introduces periods during which demand exceeds available inventory. These periods are visually marked in Figure 10.11 by shaded grey triangles. These areas represent the cumulative volume of unsatisfied demand — that is, items that were requested but not available at the time of need.

This situation highlights a critical operational issue: stockouts. In practice, such stockouts might result in lost sales, production delays, unmet service level agreements, or customer dissatisfaction. The model must therefore be extended to incorporate mechanisms for handling variability in demand,

such as safety stocks, responsive shipment scheduling, or dynamic inventory buffers.

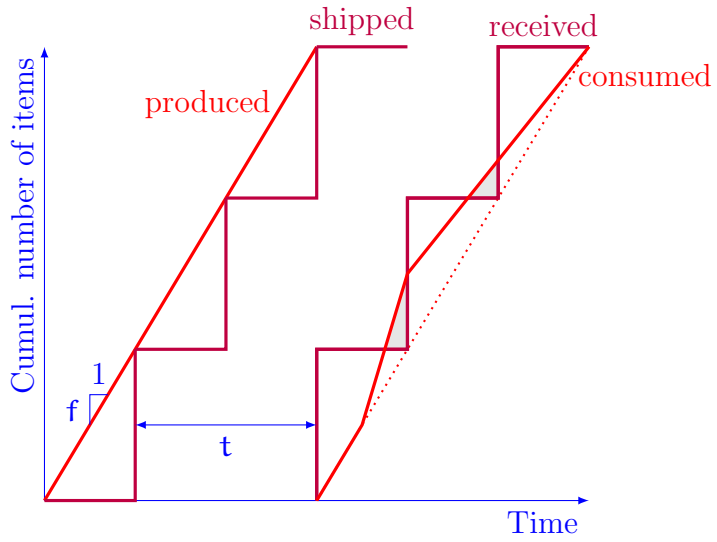


Figure 10.11: Mismatch between production and consumption: grey areas indicate unsatisfied demand

Decisions must adapt to changing demand in order to avoid costly disruptions such as stockouts. To address this, we introduce a more flexible and realistic modeling framework: the *multi-period inventory model*.

The key idea of the multi-period model is to divide the overall time horizon, denoted $[0, t_H]$, into K consecutive periods, indexed by $t = 1, 2, \ldots, K$. Within each period, we assume that the relevant quantities — demand, costs, and decisions — remain constant. However, they may change from one period to the next, thereby capturing temporal variations in the system.

Each period t has a known duration δ_t, and is characterized by a known demand rate f_t, expressed in items per unit of time. This setup allows for modeling seasonal trends, promotional surges, or simply unpredictable fluctuations in consumption.

The system also tracks the inventory level over time. We denote by i_t the inventory at the end of period t, and assume that the initial inventory i_0 is known. The main decision variable is the quantity s_t to order (or ship) at the beginning of period t. The inventory evolution is then described by the balance equation:

$$i_t = i_{t-1} - f_t \delta_t + s_t.$$

This equation expresses that the inventory at the end of the current period is equal to the inventory carried from the previous period, minus the amount

consumed during the period (which is $f_t \delta_t$), plus the quantity newly received or produced.

The cost structure is also extended to reflect the period-specific nature of operations. Each period may incur a *fixed ordering cost* c_f^t whenever an order is placed, a *variable ordering cost* c_v^t proportional to the number of items ordered, and a *waiting or holding cost* c_w^t that accounts for the cost of storing or financing inventory.

This framework allows for modeling real-world scenarios where the timing and magnitude of orders are critical decisions. It can be further refined to include constraints on order capacity, delivery delays, perishability, or stochastic variations in demand. The goal, however, remains consistent: to determine the optimal sequence of shipment decisions $s_1, s_2, \ldots, s_K$ that minimizes the total cost over the planning horizon, while avoiding periods of unsatisfied demand.

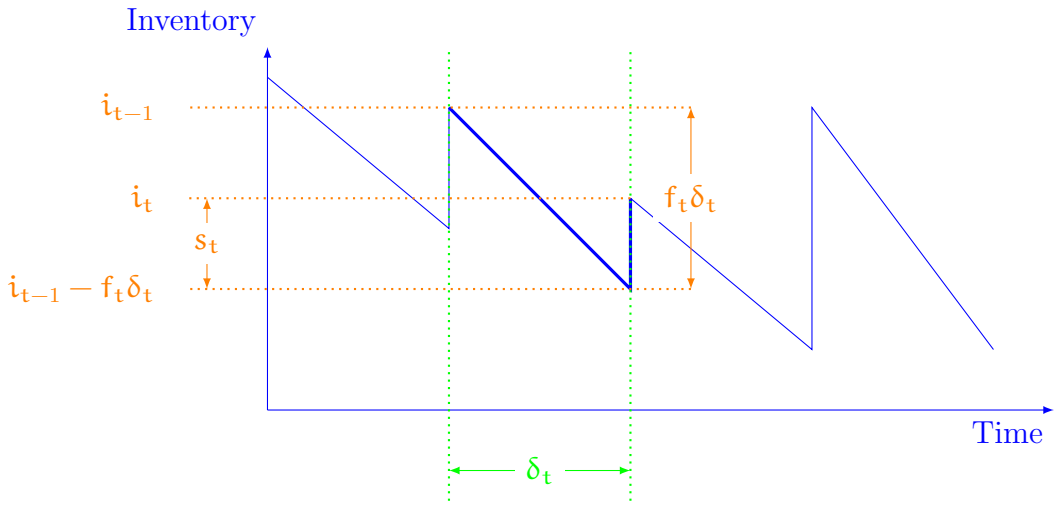


Figure 10.12: Inventory evolution in a multi-period model with ordering decision

Figure 10.12 illustrates how inventory evolves in a single period within a multi-period inventory management model. The horizontal axis represents time, and the vertical axis indicates the level of inventory.

At the beginning of period t, the system holds an inventory level of i_{t-1} (orange horizontal line). Over the duration of the period, represented by δ_t, consumption occurs at a constant rate f_t, which results in a downward slope of the inventory trajectory. The inventory at the end of the period, before any restocking, would be $i_{t-1} - f_t \delta_t$.

To avoid running out of stock, an order is placed during the period. The incoming shipment has a size s_t and is added to the remaining inventory,

bringing the final level up to i_t. The diagram clearly separates the consumption over time and the replenishment point. The thick segments of the inventory curve emphasize the period of interest.

During the period, the inventory is gradually decreasing due to consumption. The average inventory level can be approximated by the area under the inventory curve divided by the period duration δ_t. This average is:

$$\frac{1}{\delta_t}\left(i_{t-1}\delta_t - \frac{1}{2}f_t\delta_t^2\right) = i_{t-1} - \frac{1}{2}f_t\delta_t.$$

This reflects the typical number of items that must be stored during the period.

The waiting costs are the costs associated with keeping items in storage before they are consumed. They are proportional to the average number of stored items and are computed as:

$$c_w^t\left(i_{t-1} - \frac{1}{2}f_t\delta_t\right),$$

where c_w^t is the per-unit holding cost for period t.

The transportation costs consist of two components: a fixed cost c_f^t incurred for placing an order, regardless of its size, and a variable cost c_v^t that scales with the number of items shipped. The total transportation cost in period t is thus:

$$c_f^t + c_v^t s_t.$$

We now combine all the relevant quantities and constraints into a single optimization framework. The objective is to determine the optimal shipment sizes s_t for each period $t = 1, \ldots, K$ so as to minimize total costs while ensuring that demand is met and inventory remains non-negative throughout the planning horizon.

The objective function of the multi-period inventory management problem contains four distinct terms, each representing a component of the total cost incurred during the planning horizon.

$$\min_{s,i} \sum_{t=1}^{K}\left(c_w^t i_{t-1} - \frac{1}{2}c_w^t f_t\delta_t + c_f^t + c_v^t s_t\right)$$

The first term, $c_w^t i_{t-1}$, corresponds to the *waiting cost* for the items stored during period t. It is proportional to the inventory level i_{t-1} at the beginning of the period and reflects costs such as capital immobilization, depreciation, and insurance. Since inventory generates costs even when unused,

this term penalizes large inventories and encourages efficient, just-in-time ordering strategies.

The second term, $\frac{1}{2}c_w^t f_t \delta_t$, accounts for the waiting cost of the items consumed during period t, assuming a linear depletion of stock. On average, these items wait for half of the period before being used. While this cost is real, it depends only on the demand f_t and the duration δ_t, both of which are fixed parameters. As such, it does not depend on the decision variables s_t or i_t, and its inclusion does not influence the optimal solution. Therefore, we can safely omit this term from the objective function.

The third term, c_f^t, represents the fixed cost of placing an order in period t. However, in the current model, we assume that shipments are made in every period, regardless of size. Since the number of orders is fixed, the total fixed cost is constant and independent of the decision variables. Like the previous term, it can be excluded from the optimization without affecting the outcome.

The fourth and final term, $c_v^t s_t$, is the *variable transportation cost*, proportional to the quantity shipped during period t. It captures the cost per unit of items transported and contributes directly to the optimization. This term, along with the waiting cost on inventory $c_w^t i_{t-1}$, plays a central role in determining the trade-off between frequent small shipments and infrequent large ones.

After removing the two constant terms, the simplified and relevant objective function becomes:

$$\min_{s,i} \sum_{t=1}^{K} \left(c_w^t i_{t-1} + c_v^t s_t \right)$$

where i_{t-1} is the inventory level at the beginning of period t, and s_t is the shipment size.

This objective function is subject to two types of constraints. The first is the *inventory dynamics* constraint, which tracks how the inventory evolves from one period to the next. For each period $t = 1, \ldots, K$, we require:

$$i_t = i_{t-1} - f_t \delta_t + s_t.$$

This equation ensures that the inventory at the end of the period equals the starting inventory minus the demand over the period, plus any new shipments.

The second type of constraint ensures that demand is satisfied without delay. We require that the available inventory before consumption in each period is sufficient:

$$i_{t-1} - f_t \delta_t \geqslant 0, \quad \text{for all } t = 1, \ldots, K.$$

This condition prevents stockouts by enforcing that enough inventory is present to cover the full demand in the period.

Putting all these components together, we obtain a linear optimization problem in the decision variables s_t and i_t, for $t = 1, \ldots, K$. This formulation is computationally tractable and can be solved efficiently using standard linear optimization techniques such as the simplex algorithm.

Scenario 1

We now illustrate the multi-period inventory management model using a simple numerical example.

The scenario is defined over $K = 5$ consecutive time periods, each of duration 1 unit. The initial inventory at the beginning of the first period is set to 14 units. The demand across the five periods is given by the vector $[10, 5, 20, 1, 30]$, while both the unit waiting costs and the variable transportation costs are uniform, set to 1 for all periods. The objective is to determine the optimal shipping and inventory policy that minimizes the total cost over the planning horizon.

Figure 10.13 shows the optimal solution. Each line segment illustrates the inventory level at the start and end of each period. The orange numbers indicate the quantities shipped at the beginning of the respective periods.

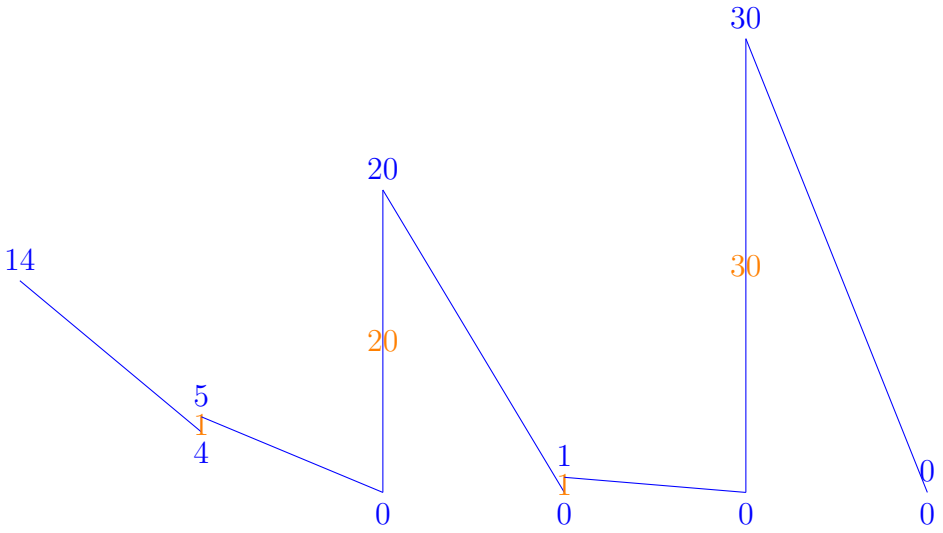


Figure 10.13: Optimal inventory and shipment plan for Scenario 1

This solution reflects a *just-in-time* inventory management strategy. The transportation costs are sufficiently low that the model prefers to match

inventory to demand as closely as possible in each period. This minimizes the amount of inventory held and therefore reduces waiting costs.

We observe that, at the beginning of each period, the inventory level exactly equals the demand for that period. This is made possible by perfect anticipation of future demand and the absence of any constraint preventing small, frequent shipments. For instance, at the beginning of period 3, a shipment of 20 units is made to precisely cover the high demand of that period.

This behavior illustrates the model's preference to avoid unnecessary storage by relying on flexible and timely shipments. In practical terms, this policy is viable when transportation is inexpensive and logistics operations can accommodate frequent deliveries.

Scenario 2

We now consider a variation of the previous scenario to highlight how the inventory management model adapts when transportation costs vary across time. The setup remains the same: the planning horizon is divided into $\mathsf{K} = 5$ equal periods of unit duration, with an initial inventory of 14 units and demand given by the vector $[10, 5, 20, 1, 30]$. Waiting costs remain constant across all periods, set to 1. However, the transportation costs vary: while they are equal to 1 in all periods except the fourth, the cost in the fourth period is significantly higher, set to 100. This models a situation where, for instance, deliveries are much more expensive during a holiday, a weekend, or a special event.

Figure 10.14 shows the optimal solution. As before, the segments represent the evolution of inventory across time, and the orange labels indicate the shipment quantities received at the start of each period.

The solution demonstrates a strategic response to the spike in transportation costs in the fourth period. To avoid placing an order during that expensive window, the model anticipates future demand by increasing the shipment quantity in period 3. Specifically, the 31 units shipped at the start of period 3 are enough to cover the demand for both period 4 (1 unit) and period 5 (30 units).

This anticipatory behavior leads to increased inventory holding, but the trade-off is favorable because holding costs are much lower than the high transportation cost of period 4. As a result, the solution minimizes the total cost by shifting shipments away from the costly period.

This example highlights the value of *perfect anticipation* and the ability of the model to balance between inventory and transportation costs. It also underscores the importance of temporal variability in logistics planning,

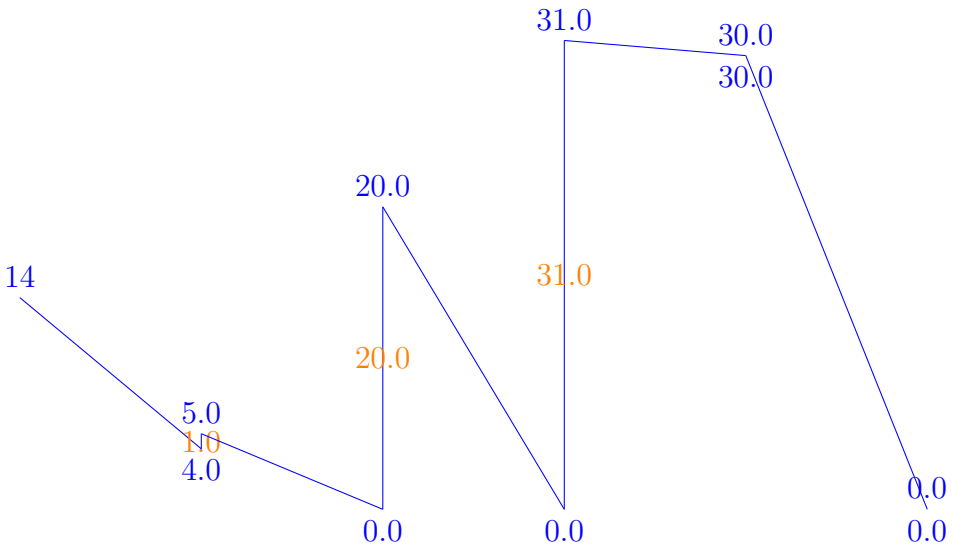


Figure 10.14: Optimal inventory and shipment plan for Scenario 2

where cost-efficient solutions often require adapting order schedules to external fluctuations.

Scenario 3

We now examine a third scenario that demonstrates how the inventory management model adapts when transportation costs are low in one period and remain high for all subsequent periods. This type of situation might arise due to seasonal pricing, limited access to transportation infrastructure, or scheduled disruptions that increase delivery costs over several intervals.

In this scenario, the planning horizon again consists of $\mathsf{K} = 5$ periods of unit duration, with an initial inventory of 14 units. Demand over the five periods is given by the vector $[10, 5, 20, 1, 30]$, and both the waiting costs and demand pattern remain unchanged from the previous examples. However, the transportation costs are now highly asymmetric: they are low (equal to 1) only during the first period, and then increase drastically to 100 for all remaining periods.

Figure 10.15 presents the optimal solution under these conditions. Each segment traces the evolution of the inventory level across time, and the orange annotation marks the quantity received at the start of the first period.

In this configuration, the entire demand over the five periods (totaling 66 units) must be met using only one economical shipment. The model exploits the low cost in period 1 to receive a large shipment of 52 units, on top of the 14 units initially in inventory.

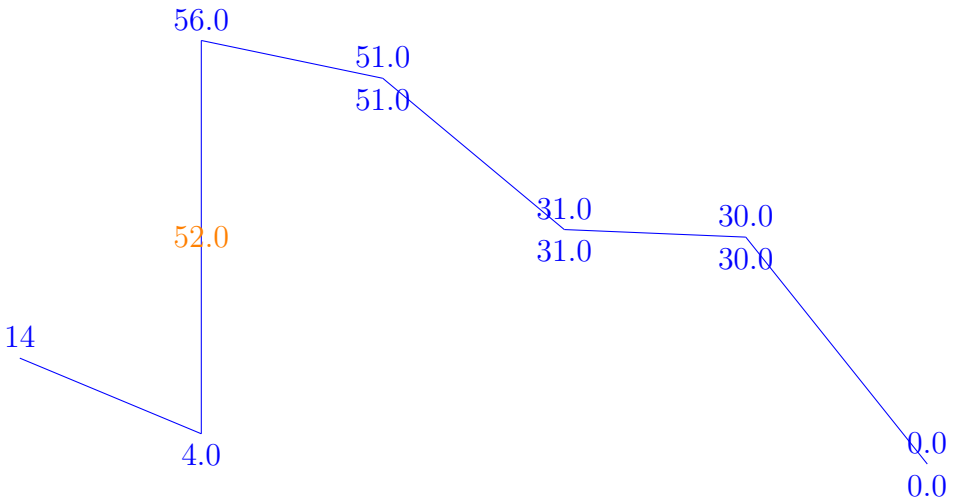


Figure 10.15: Optimal inventory and shipment plan for Scenario 3

From that point on, no further shipments are made, because transportation becomes prohibitively expensive. The entire consumption sequence is covered by drawing down the inventory accumulated at the beginning. This comes at the expense of increased holding costs — inventory is kept for longer, especially for the later demand. However, this trade-off is cost-effective because holding costs remain relatively low compared to the steep transportation costs.

This scenario highlights a key insight: when future delivery costs are predictable and prohibitively high, it is optimal to concentrate shipments in a cheaper period, even if it implies higher storage and waiting costs. The result is an extreme version of a *just-in-advance* strategy, made possible by the model's assumption of *perfect anticipation*.

Scenario 4

In this final scenario, we examine how the inventory management model reacts to a case in which both transportation and waiting costs vary across time. Specifically, transportation is inexpensive in the first period but becomes expensive for the remainder of the planning horizon, while waiting costs remain low except during the fourth period, when they spike significantly.

The problem consists of $\mathsf{K} = 5$ time intervals of duration 1, with initial inventory set to 14 units. The demand sequence is again given by $[10, 5, 20, 1, 30]$. Transportation costs are low only in the first period, taking the value 1, and then jump to 100 in the following periods. Meanwhile,

waiting costs are constant at 1 in every period, except for the fourth period where they sharply increase to 100. These parameters aim to simulate a situation where not only transportation becomes constrained after a certain time, but storing inventory becomes particularly undesirable during one specific interval.

The optimal solution is depicted in Figure 10.16. Inventory evolution and shipment quantities are marked, with the orange labels indicating the size of the orders.

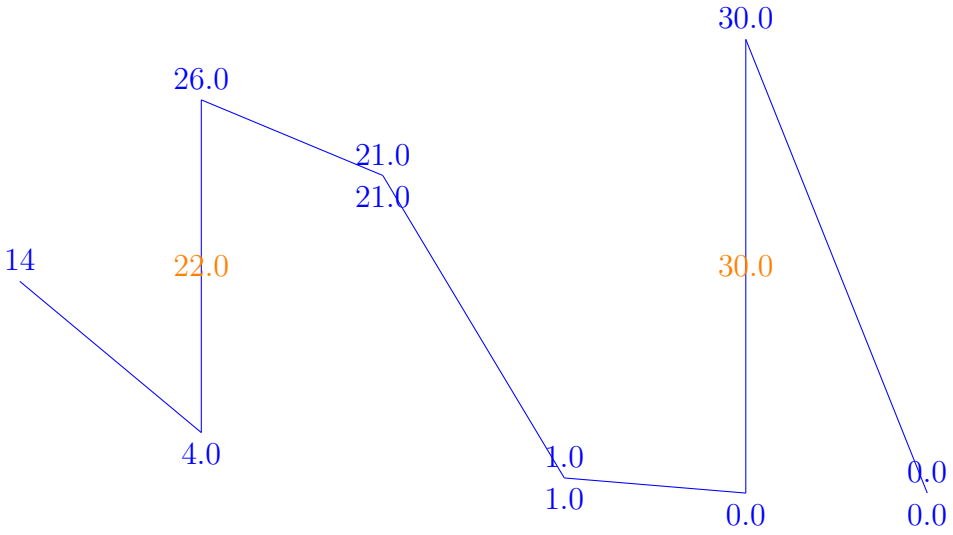


Figure 10.16: Optimal inventory and shipment plan for Scenario 4

The solution demonstrates several key behaviors of the model. In the first period, a shipment of 22 units is received, covering the demand for periods 2, 3, and partially for 1. This anticipatory behavior avoids incurring transportation costs later, when they become prohibitively high. Importantly, the model ensures that inventory is fully depleted by the end of period 3 to avoid the expensive waiting costs in period 4.

Then, rather than maintaining any stock into period 4, a large shipment of 30 units is ordered at the start of period 5 to meet the final demand. This is done despite the high transportation cost in period 5, because the alternative—storing inventory during the expensive period 4—would have resulted in even greater waiting costs. By keeping inventory low when storage is costly, the model minimizes total cost.

This scenario illustrates the powerful role of anticipation in inventory management. The optimizer balances competing cost components over time, and the resulting policy avoids both high transportation and high storage

expenses. It is another example of how the model takes advantage of foresight to make globally optimal decisions.

10.2.3 Summary

The inventory management section has shown that the movement of goods from producers to consumers involves two critical operations: transportation and storage. These operations are not merely logistical necessities but key cost drivers within a supply chain. Whenever items are not instantly consumed after being produced, they must be held in inventory. Similarly, they must be physically transported to their destination. Both of these actions incur costs — storage costs on one side, and transportation costs on the other.

Storage costs include the expenses associated with physical space, handling, insurance, depreciation of goods, and the opportunity cost of capital. As such, reducing the amount of time items remain in storage tends to reduce overall storage costs. One natural implication of this is the preference for shorter headways — that is, more frequent shipments of smaller quantities. This minimizes the amount of inventory held at any given time.

In contrast, transportation costs often favor the opposite strategy. Shipping in larger quantities over longer intervals can reduce costs per item, especially when fixed costs are involved in every shipment, such as preparing a truck, booking freight space, or administrative overhead. Thus, minimizing transportation costs pushes toward longer headways and larger shipment sizes.

This sets up a fundamental trade-off: minimizing storage costs suggests more frequent deliveries, while minimizing transportation costs encourages more infrequent shipments. The balance between these two opposing forces is captured through a formal optimization problem, where decision variables such as headway and shipment size are chosen to minimize the total cost over a given planning horizon.

In our first model, we assumed that demand was constant and perfectly known, which allowed us to derive closed-form expressions for the optimal shipment strategy. However, in practice, demand is rarely constant. Even if the average demand is stable, short-term variations and seasonality create discrepancies between production and consumption rates at any given moment. This necessitates more flexible models.

To address this, we introduced a multi-period inventory management model. This model divides the planning horizon into discrete periods and allows production, consumption, and cost parameters to vary from one period to the next. Importantly, this formulation remains a linear optimization

problem and can be solved efficiently using standard mathematical optimization tools. It also introduces time-varying decision variables and can incorporate changes in transportation and storage costs.

Nonetheless, another significant issue in inventory management is uncertainty — particularly uncertainty in demand forecasts. Real-world demand is not only variable but often difficult ot predict. To manage inventory effectively under such uncertainty, it becomes necessary to move beyond deterministic models and adopt stochastic versions. These incorporate probability distributions for demand and may involve concepts like safety stock, service levels, and risk management.

In conclusion, inventory management is a delicate balancing act between opposing cost structures, shaped by temporal dynamics and uncertainty. Mathematical models provide valuable tools to support decision-making, offering structured ways to navigate this trade-off under different assumptions about demand, cost, and system flexibility.

10.3 Vehicle routing problem

The vehicle routing problem arises in the final stage of the distribution process, where goods must be delivered to a set of customers using a fleet of vehicles. The central question is how to organize these deliveries in an efficient way. More precisely, we must determine which vehicle serves which customer, and in what order the customers should be visited. This problem is critical in logistics and transportation systems and has a direct impact on operational costs, service levels, and environmental outcomes.

The setting is as follows. We consider a set of customers, denoted by $\mathcal{C}$, who each require a certain quantity of goods. The vehicles that perform the deliveries are based at a central depot. To include both the customers and the depot in a unified way, we define the set of locations as $\mathcal{C}^+ = \mathcal{C} \cup \{\text{depot}\}$, and we assign the depot the index 0. Each customer $j \in \mathcal{C}$ has a known demand d_j. A total of q vehicles are available, each with the same capacity ℓ.

The travel time between any two distinct locations i and j is denoted by t_{ij}. We assume the travel times are symmetric, meaning that $t_{ij} = t_{ji}$ for all $i, j \in \mathcal{C}^+$. This assumption simplifies the modeling and corresponds to many realistic scenarios where the distance or time between two locations is the same in both directions.

To formulate this as a mathematical optimization problem, we define the decision variables x_{ij} for all $i, j \in \mathcal{C}^+$. The variable x_{ij} takes value 1 if location j is visited immediately after location i by one of the vehicles, and 0

otherwise. These binary variables encode both the assignment of customers to routes and the sequencing of visits within each route.

The objective is to minimize the total duration of all trips, which is expressed as the sum of travel times over all arcs used in the solution. This yields the objective function

$$\min \sum_{i,j \in \mathcal{C}^+} t_{ij} x_{ij}.$$

Several constraints are required to ensure that the solution is feasible. First, each customer must be visited exactly once. This is enforced by two types of constraints: one that ensures each customer has exactly one successor in a route,

$$\sum_{j \in \mathcal{C}^+} x_{ij} = 1, \quad \forall i \in \mathcal{C},$$

and one that ensures each customer has exactly one predecessor,

$$\sum_{i \in \mathcal{C}^+} x_{ij} = 1, \quad \forall j \in \mathcal{C}.$$

Additionally, we need to specify how many vehicles are used. Since each vehicle starts its route at the depot, the number of arcs leaving the depot must equal the number of vehicles. This leads to the constraint

$$\sum_{j \in \mathcal{C}} x_{0j} = q.$$

These constraints define the basic structure of the vehicle routing problem. However, they are not sufficient. First, the capacity of the vehicles are ignored.

To better understand theotheer missing constraints, we consider a numerical example based on realistic geographic and demand data. The depot is located at the train station, and there are 20 customers to serve, each with an identical demand of 1 unit. A fleet of three vehicles is available to perform the deliveries, each with a capacity of 20 units. This means that each vehicle, in theory, could serve all customers on its own, but we distribute the deliveries across multiple vehicles to minimize travel time. The travel durations between all locations — customers and depot — are derived from OpenStreetMap, reflecting actual distances in a city context.

The spatial distribution of the depot and the customers is shown in Figure 10.17. The depot, drawn as an orange circle, is centrally positioned, while the customers are represented by small blue squares. The layout illustrates a

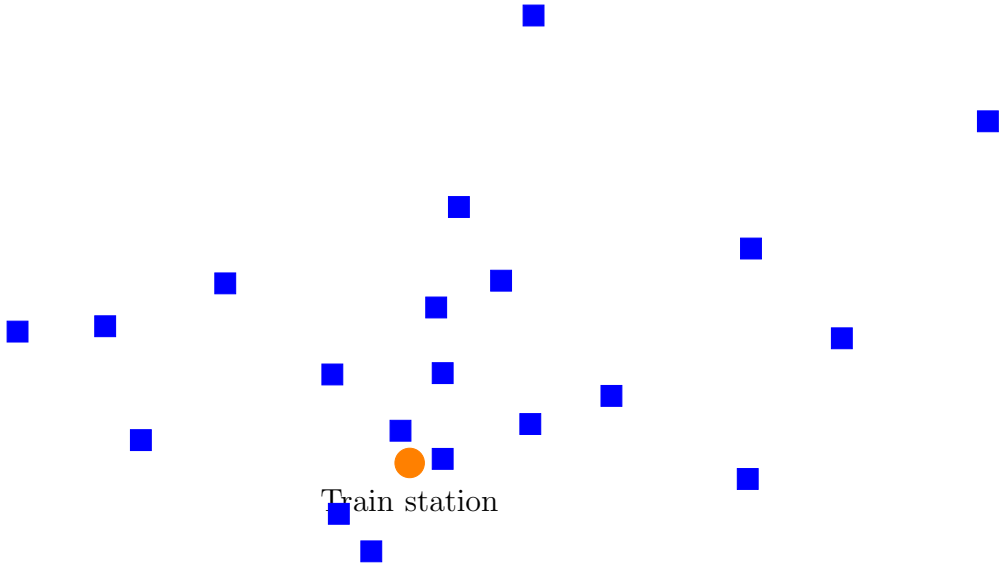


Figure 10.17: Depot and customer locations in the routing problem

diverse spatial spread of customers, requiring a thoughtful routing strategy to ensure efficiency.

Figure 10.18 illustrates the optimal solution of the above model, where the arrows indicate the direction in which customers are visited. This solution is actually invalid. Indeed, it contains *subtours*, which are closed loops among a subset of customers that are disconnected from the depot. These subtours imply that certain groups of customers are visited in a cycle without any connection to the depot, violating the requirement that every delivery tour starts and ends at the depot.

To ensure the correctness and feasibility of vehicle routing solutions, it is necessary to introduce additional constraints into the model. In particular, we must address two crucial issues: enforcing vehicle capacity limits and eliminating subtours. These enhancements are inspired by methods used for the well-known traveling salesman problem (TSP), which can be viewed as a special case of the vehicle routing problem (VRP) with only one vehicle. In both problems, the potential for invalid cycles that do not involve the depot must be eliminated, and this is typically done through the introduction of auxiliary variables and additional constraints.

To achieve this, we define new decision variables $\mathfrak{u}_\mathfrak{i}$ for each customer $\mathfrak{i} \in \mathcal{C}$. The variable $\mathfrak{u}_\mathfrak{i}$ represents the load of the vehicle upon arrival at customer $\mathfrak{i}$ — that is, the cumulative quantity of items already on board when the vehicle visits this location. These variables allow us to keep track of the vehicle's content and ensure that it respects both the capacity of the

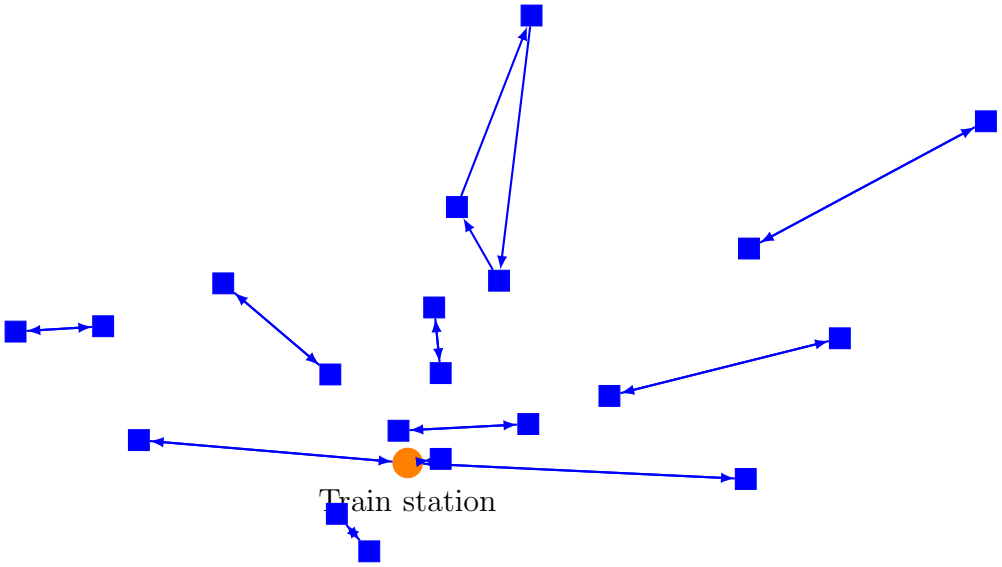


Figure 10.18: Illustration of an invalid solution with subtours and no capacity enforcement

vehicle and the demand of each customer.

The first constraint ensures that the vehicle does not exceed its capacity when reaching any customer. This is expressed as

$$u_j \leqslant \ell, \quad \forall j \in \mathcal{C}.$$

This upper bound guarantees that no customer is visited with a load that would violate the vehicle's capacity limit.

Next, we require that when a vehicle reaches a customer, it has at least as much load as is needed to satisfy the customer's demand:

$$u_j \geqslant d_j, \quad \forall j \in \mathcal{C}.$$

This constraint ensures that the vehicle always arrives prepared to fulfill the customer's request.

To properly define the role of the variables u_i in the context of the route, we introduce a constraint that links the load at customer i to the load at customer j, if the vehicle travels from i to j:

$$u_i - u_j + \ell x_{ij} \leqslant \ell - d_j, \quad \forall i, j \in \mathcal{C}.$$

This constraint serves a dual purpose. If $x_{ij} = 1$, meaning that the vehicle goes directly from i to j, then the constraint becomes

$$u_i - u_j + \ell \leqslant \ell - d_j,$$

which simplifies to

$$u_j \geqslant u_i + d_j.$$

This implies that the load upon arrival at customer j must be at least the load at i plus the demand of j, effectively enforcing a strictly increasing load along the route. Such a property prevents the creation of subtours—cycles among customers that are disconnected from the depot — because in a subtour, the vehicle would return to the same location without having visited the depot, which would contradict the strictly increasing nature of u_j.

If, on the other hand, $x_{ij} = 0$, then the constraint becomes

$$u_i - u_j \leqslant \ell - d_j,$$

which holds trivially due to the previously stated capacity and demand constraints. Specifically, since $u_j \geqslant d_j$ and $u_i \leqslant \ell$, we have

$$d_j - u_j \leqslant \ell - u_i,$$

which confirms that this inequality does not restrict the solution when the arc from i to j is not used.

The final form of the model includes all these constraints, along with the original routing constraints, to yield a valid and complete formulation of the vehicle routing problem:

$$\min_{x,u} \sum_{i,j \in \mathcal{C}^+} t_{ij} x_{ij},$$

subject to

$$\begin{aligned}
\sum_{j \in \mathcal{C}^+} x_{ij} &= 1, && \forall i \in \mathcal{C}, \\
\sum_{i \in \mathcal{C}^+} x_{ij} &= 1, && \forall j \in \mathcal{C}, \\
\sum_{j \in \mathcal{C}} x_{0j} &= q, && \\
u_j &\leqslant \ell, && \forall j \in \mathcal{C}, \\
u_j &\geqslant d_j, && \forall j \in \mathcal{C}, \\
u_i - u_j + \ell x_{ij} &\leqslant \ell - d_j, && \forall i, j \in \mathcal{C}, \\
x_{ij} &\in \{0, 1\}, && \forall i, j \in \mathcal{C}^+.
\end{aligned}$$

This formulation effectively addresses the practical aspects of vehicle routing, such as load management and route connectivity, enabling the computation of feasible and efficient delivery plans.

10.3.1 Scenario 1

The result of the vehicle routing problem applied to our numerical example is shown in Figure 10.19. The instance includes a depot located at the train station and 20 customers dispersed in a city. Each customer has a unit demand, and the depot has access to a fleet of three vehicles, each with a capacity of 20 items. Distances between locations are derived from real travel times using OpenStreetMap data.

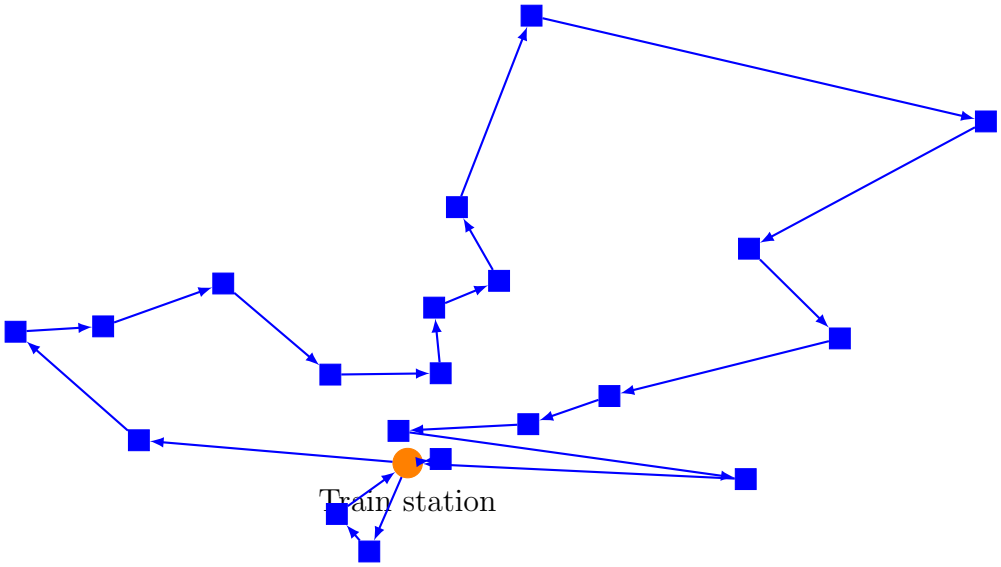


Figure 10.19: Solution of the vehicle routing problem in Scenario 1.

From the figure, we observe that all customers are served, and all routes start and end at the depot. However, the solution is not balanced across the three available vehicles. One vehicle serves 17 customers, which is close to the capacity limit. Another vehicle is assigned only two customers, and the third one is responsible for a single delivery.

10.3.2 Scenario 2

In this scenario, we revisit the same vehicle routing problem as before but introduce a more restrictive fleet capacity. Specifically, each of the three available vehicles can now serve at most nine customers, instead of the previous capacity of twenty. The depot remains located at the train station, and the demand at each customer location is still equal to one unit. Figure 10.20 displays the resulting delivery plan obtained by solving the optimization model with this tighter constraint.

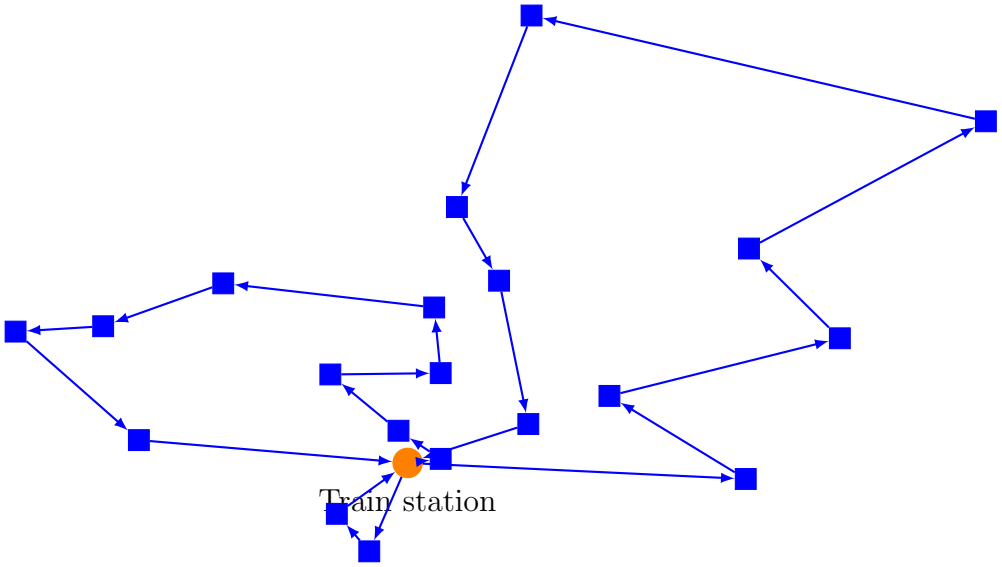


Figure 10.20: Vehicle routing solution for Scenario 2 with capacity limited to nine customers per vehicle.

The solution reflects a better distribution of the workload across the fleet compared to Scenario 1. Two of the vehicles are each assigned to serve exactly nine customers, fully utilizing their capacity. The third vehicle serves the remaining two customers, ensuring all twenty demands are met.

10.3.3 Scenario 3

In this scenario, we consider a more constrained variant of the vehicle routing problem. The number of vehicles available is increased to four, but each vehicle has a more limited capacity: it can now serve only eight customers. The total number of customers remains twenty, and as before, each customer requires the delivery of one unit. The depot is still located at the train station, and all deliveries must start and end there. The objective is again to assign customers to vehicles and sequence their visits so as to minimize the overall travel time, subject to capacity constraints.

The solution to this scenario is depicted in Figure 10.21. Each route, represented by a path of directed arrows, corresponds to the trip made by one vehicle. The figure shows how the twenty customers are partitioned across the four available vehicles in a way that satisfies all problem constraints.

As expected from the tighter capacity limit, the solution splits the deliveries more evenly among the fleet. The first two vehicles each serve eight customers, which is the maximum their capacity allows. The third vehicle

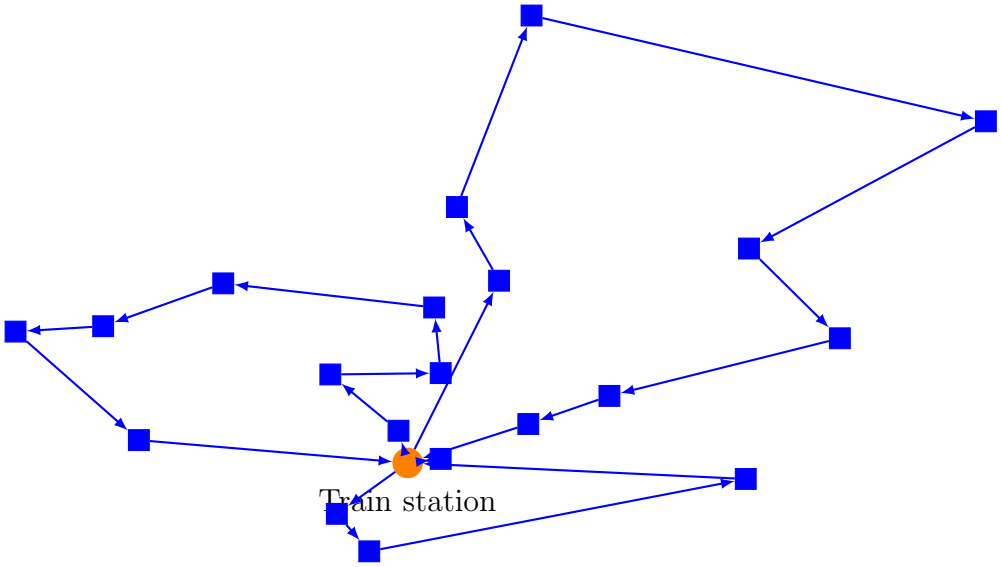


Figure 10.21: Vehicle routing solution for Scenario 3 with four vehicles, each having a capacity of eight customers.

serves three customers, while the fourth handles only one. This distribution ensures feasibility but also reflects the trade-off involved: increasing the number of vehicles can help accommodate tighter constraints, but it may lead to some underutilization, as seen in the case of the last vehicle.

This example illustrates how the model dynamically adapts to both capacity and fleet size. By adjusting the number of vehicles and their load capacities, the optimization algorithm constructs routes that are feasible, efficient, and responsive to logistical constraints. It also shows that, in practical applications, a balance must often be struck between minimizing travel time and using resources efficiently, especially when some vehicles are not filled to capacity.

10.3.4 Scenario 4

In this fourth scenario, we explore a variant of the vehicle routing problem where the number of vehicles is fixed to four, and the capacity of each vehicle is limited to ten units. While the overall number of customers remains twenty, we introduce a new difficulty: one of the customers — specifically customer 18 — requires a significantly larger delivery, with a demand of eight units. All other customers continue to require one unit each. This asymmetry in demand introduces a substantial increase in complexity.

Figure 10.22 displays the resulting optimal solution. As in previous cases,

all delivery tours start and end at the depot located at the train station. The figure highlights the paths followed by each vehicle to visit its assigned customers while respecting both the demand constraints and the vehicle capacity constraints. Notably, customer 18 is distinguished in the figure: their node is represented in orange, reflecting their high demand relative to the others.

Figure 10.22: Vehicle routing solution for Scenario 4 with four vehicles of capacity ten and a high-demand customer.

The solution reflects the impact of the large demand associated with customer 18. Because a single delivery of eight units nearly fills an entire vehicle, this customer must be carefully assigned to avoid violating the capacity constraint. In this instance, customer 18 is served by vehicle 3, which handles only three customers in total.

Meanwhile, the other three vehicles are more densely utilized: vehicle 1 serves nine customers, nearly reaching its full capacity, while vehicle 2 and vehicle 4 serve seven and one customer respectively. This asymmetric distribution of workload results from the need to maintain feasibility under both capacity and demand constraints.

A striking feature of this scenario is the computational complexity required to identify the optimal solution. Solving this instance took more than ten hours on EPFL's Scitas High Performance Computing infrastructure. This emphasizes the combinatorial nature of the problem. It also underlines the value of exact optimization approaches in logistics planning — especially in contexts where respecting delivery constraints is critical.

10.3.5 Summary

The vehicle routing problem is a central question in transportation and logistics. It arises when a set of items must be delivered from a central location — referred to as the depot — to a group of geographically dispersed customers. To perform these deliveries, a fleet of vehicles is available at the depot. The key challenge is to determine both the assignment of customers to vehicles and the sequence in which each vehicle should visit its assigned customers. The goal is typically to minimize the total cost or total distance traveled while respecting various operational constraints.

The core of the problem consists in ensuring that each customer is visited exactly once, that the total demand assigned to each vehicle does not exceed its capacity, and that all routes start and end at the depot. The resulting optimization problem is highly combinatorial: the number of possible assignments and routes grows extremely rapidly with the number of customers.

In practice, the basic vehicle routing problem admits many variants that reflect the diversity of real-world logistics operations. For instance, the fleet may be heterogeneous, with vehicles differing in capacity, cost, or speed. Some applications involve not just deliveries, but also pick-ups, leading to pick-up and delivery problems. Other settings impose time windows: customers must be visited within specific time intervals. In some cases, deliveries can be split across multiple vehicles if necessary. More complex networks may also involve multiple depots. These and many other variants make the vehicle routing problem a highly flexible modeling tool for real-life transportation systems.

The mathematical structure of the vehicle routing problem presents significant challenges. The model includes constraints that prevent subtours — closed loops that do not include the depot — and that enforce vehicle capacity limits. Although the model presented here is one possible formulation, other formulations — such as flow-based or path-based models — may be more efficient for solving large-scale instances.

Due to this complexity, exact methods can become computationally infeasible for large problems or in settings where solutions must be computed frequently in response to new data. In such cases, heuristic methods — such as local search, metaheuristics, or machine learning-based approaches — are often employed. These methods do not guarantee optimality but can provide high-quality solutions within acceptable time limits. The frequent need to recompute solutions in operational contexts further motivates the use of fast, approximate solution techniques.

10.4 Summary

Freight transportation differs markedly from passenger transportation in that it involves significantly less behavioral modeling and relies much more heavily on optimization techniques. In the movement of goods, decisions are typically driven by cost efficiency, logistical constraints, and operational feasibility, rather than by individual preferences or choices. As a result, freight transportation problems are often framed as well-defined optimization problems with clear objectives and constraints.

Throughout this chapter, we have explored three illustrative examples that reflect the different time horizons involved in freight logistics. At the strategic, long-term level, the facility location problem focuses on deciding where to place depots or distribution centers in order to minimize delivery costs while satisfying customer demand. This decision has long-lasting implications and requires careful planning based on current and projected demand

patterns.

At the medium-term level, inventory management addresses the trade-off between storage and transportation costs. It aims to determine how much to order and when, so as to meet demand efficiently while minimizing holding costs and avoiding stockouts. We have seen both simplified models with constant demand and more realistic multi-period models that accommodate demand variability over time.

Finally, at the operational, short-term level, the vehicle routing problem deals with how to deliver goods to customers using a fleet of vehicles. This problem entails assigning customers to routes and determining the order of visits, all while respecting vehicle capacities and minimizing total travel time or cost.

These three problems—facility location, inventory management, and vehicle routing—provide a structured view of the various planning layers in freight transportation. Together, they highlight how mathematical modeling and optimization can be leveraged to design efficient, reliable, and cost-effective logistics systems.

Chapter 11

Cost benefit analysis

Cost-benefit analysis (CBA) is a fundamental tool used in the evaluation of transportation systems and projects. It provides a structured approach to compare the total expected costs of a project with its anticipated benefits, expressed in monetary terms. The goal is to determine whether the benefits outweigh the costs, and by how much, thereby supporting informed decision-making regarding the allocation of resources.

11.1 A simple example

Let us begin with a simplified example to illustrate the basic principles of cost-benefit analysis in the context of a transportation project. Suppose an airline company is considering the launch of a new route between Geneva (GVA) and London Heathrow (LHR). To do so, it would need to acquire and operate an additional aircraft. The central question is whether this investment is economically justified: in other words, do the expected benefits from operating the new line outweigh the associated costs?

The project involves the purchase of a Boeing 737-300 aircraft, a commonly used model for short-haul flights. The financial evaluation distinguishes between two main types of costs: fixed costs and variable costs. Fixed costs refer to expenditures that are independent of the actual usage of the aircraft, such as the loan repayment for acquiring the plane. In this example, the fixed costs are estimated at $403,765 per year, based on a financing period of 120 months.

Variable costs, on the other hand, depend on the operation of the aircraft. These include maintenance expenses, fuel consumption, and other costs that scale with flight activity. For the proposed route, the annual variable costs are projected to reach $2,875,072. Adding the fixed and variable components,

the total cost of operating the aircraft for one year amounts to $3,278,837.

To complete our analysis, we must now examine the operational details of the proposed airline service, the associated crew costs, and the expected revenues. The planned service consists of two daily round-trip flights between Geneva (GVA) and London Heathrow (LHR), each with a duration of approximately 1 hour and 35 minutes. Flights are scheduled five days a week, leading to a total of 520 flights per year. This represents 823.3 flight hours annually.

A major component of operating costs in aviation is crew remuneration. For this example, we assume an average cost of $2,000 per flight hour, covering both cockpit and cabin crew. Based on the annual number of flight hours, the total crew costs amount to $1,646,666 per year. These costs are variable in nature and scale directly with the number of flights operated.

We then turn to the revenue side of the analysis. With an average of 120 passengers per flight and a mean ticket price of $100, each flight is expected to generate $12,000 in revenue. Multiplying by the annual number of flights (520), the total annual revenue reaches $6,240,000. This simple revenue model assumes consistent load factors and pricing throughout the year.

Bringing all these components together, we can calculate the annual costs and benefits. The total cost of operating the aircraft, including both fixed and variable costs, is $3,278,837. Adding the crew costs yields total annual expenditures of $4,925,503. With revenues of $6,240,000, the operation produces an annual surplus, or benefit, of $1,314,497.

Beyond simply calculating the annual surplus, cost-benefit analysis can also yield valuable operational insights through the computation of break-even indicators. These indicators help decision-makers understand the conditions under which the project neither generates a profit nor incurs a loss. In other words, the break-even point corresponds to the level of revenue that exactly offsets total costs.

In our airline example, the total annual cost of operating the new service is $4,926,000. To determine the break-even revenue per flight, we divide this amount by the total number of annual flights, which is 520. This results in a break-even revenue of approximately $9,473 per flight. If each flight can generate this amount in ticket sales, the airline covers all operational costs but does not earn a profit.

From this, we can derive alternative scenarios that achieve this break-even point. For example, if every flight carries 120 passengers — the assumed aircraft capacity – the required average ticket price to break even is approximately $79. Alternatively, if the average ticket price remains at $100, the break-even load factor drops to 94 passengers per flight. These indicators provide concrete targets for pricing and occupancy that can guide managerial

decisions.

Such break-even analysis is particularly useful in the early stages of project planning. It supports the evaluation of pricing strategies, demand forecasts, and operational viability. Moreover, it enables the identification of risk thresholds: if actual performance falls below the break-even levels, the project may need to be restructured or reconsidered. In this sense, break-even indicators complement surplus calculations by offering a more nuanced understanding of what is required for a transportation service to be financially sustainable.

This example illustrates how cost-benefit analysis can be used to evaluate a transportation project in financial terms. While simplified, the calculation reflects key principles: the separation of fixed and variable costs, the role of operational scheduling, and the estimation of revenues based on demand assumptions. The result — a net positive annual benefit — suggests that, under these assumptions, the proposed route is economically viable. Further analysis might refine these estimates, incorporate risk and uncertainty, or include external benefits such as improved connectivity or reduced environmental impacts. Nonetheless, this scenario provides a concrete foundation for understanding the rationale behind investment decisions in the transport sector.

Cost-benefit analysis is a central component of business decision-making, particularly in capital-intensive industries like aviation. Whether assessing the launch of a new service or the acquisition of a major asset, businesses rely on structured financial evaluations to gauge the viability of their options. In practice, this type of analysis is often implemented using simple spreadsheet tools, where assumptions, parameters, and equations can be clearly laid out and easily adjusted. Such tools facilitate transparency and reproducibility, while enabling sensitivity testing to explore how results vary under different conditions.

Despite its apparent simplicity, this approach is not without challenges. One of the main difficulties lies in defining credible scenarios. For example, future demand, ticket pricing, or operating conditions may be subject to considerable uncertainty. Another difficulty concerns the accurate estimation of costs, especially for new or infrequent operations where historical data may be scarce or not directly transferable. Finally, it is often hard to ensure that all relevant factors have been taken into account. Costs and benefits may span multiple dimensions—financial, operational, environmental, and social—and capturing them comprehensively requires careful judgment and interdisciplinary input.

11.2 A more complex example

To illustrate the broader scope of cost-benefit analysis in public infrastructure, we turn to a more complex and large-scale example: the Gotthard Base Tunnel in Switzerland. This railway tunnel, which officially opened in 2016, is one of the most ambitious transportation projects in recent European history. Extending over 57 kilometers between the cantons of Uri and Ticino, the tunnel forms a vital part of the north-south trans-Alpine rail corridor, significantly improving connectivity between Zurich and Milan. With two separate tubes to accommodate bidirectional traffic, it represents a major technological and engineering achievement.

One of the tunnel's most tangible impacts is the reduction in travel time. The new infrastructure shortens the journey between Zurich and Milan by approximately 30 minutes, a change that affects not only passenger trains but also freight services. In doing so, it enhances the overall efficiency and attractiveness of rail transport along this corridor, potentially shifting traffic away from road to rail. The total investment required to construct the tunnel amounted to approximately 12.2 billion Swiss francs, funded largely by public sources.

One might be tempted to evaluate the financial viability of the Gotthard Base Tunnel using a similar approach as for a commercial airline route, by calculating the potential revenues from a hypothetical toll system. For example, if each train — whether freight or passenger — were charged a fee of 1,000 Swiss francs for using the tunnel, and if 25,000 freight trains and 20,000 passenger trains used the tunnel annually, the resulting revenue would amount to 45 million francs per year. Based on the tunnel's construction cost of 12.2 billion francs, it would take approximately 271 years to recover the investment through such toll revenues alone.

However, this approach is clearly inadequate for assessing the viability of a public infrastructure project of this magnitude. First, the revenue estimate assumes that demand remains unchanged, even after the introduction of a toll. In reality, economic theory and empirical evidence suggest that introducing a user charge would reduce the number of trains using the tunnel, as some services may be rerouted or canceled in response to higher costs. As a result, the projected revenue is likely to be overstated, and the actual break-even period would be even longer than the simple calculation suggests. But more importantly, this kind of narrow financial evaluation fails to capture the broader economic and societal rationale for building the tunnel.

The limitations of this revenue-based assessment were already acknowledged in 1974 by the Swiss Federal Railways (SBB). At the time, analysts pointed out that while it is possible to define a *utilization threshold* — a

point at which operational revenues equal or exceed additional costs — such a threshold says little about the project's true economic value (Diemant, 1974).

As the SBB report noted, it is fundamentally impossible to demonstrate the economic profitability of the tunnel in the traditional, commercial sense. Instead, evaluating the Gotthard Base Tunnel requires an *overall societal perspective*, which considers a wide range of benefits beyond direct revenues. These may include improved connectivity across regions and countries, reduced environmental impacts due to shifts from road to rail, increased safety, and the long-term resilience of freight and passenger transport networks. Such benefits, while harder to quantify, are essential for understanding why societies undertake large infrastructure investments that may not be justifiable on financial grounds alone.

This example underscores the distinction between financial and economic appraisal. While cost-benefit analysis remains a central tool, its application in the context of public infrastructure must extend beyond balance-sheet considerations. Public investment decisions are not solely about maximizing profit; rather, they must reflect broader societal goals such as equity, environmental sustainability, and national or regional cohesion. Unlike private businesses, which focus on financial returns to shareholders, governments are responsible for serving the collective interest of their citizens.

Evaluating large-scale infrastructure projects therefore requires a careful balance between rigorous technical analysis and political judgment. Cost-benefit analysis provides a structured framework for identifying and quantifying the expected impacts of a project, but it does not dictate the final decision. Political prerogatives — such as promoting regional development, reducing emissions, or enhancing national resilience — may justify projects that would not be considered profitable in a narrow financial sense.

The primary objective of such evaluation is not to produce a definitive yes-or-no answer, but to inform the decision-making process. A well-conducted analysis helps policymakers understand the trade-offs involved and the distribution of costs and benefits across different segments of society. In doing so, it supports the design of decisions that aim to achieve the greatest public good, in line with societal values and long-term strategic priorities. Ultimately, cost-benefit analysis is a tool for transparency and accountability in public investment, not a substitute for democratic governance.

11.3 Methodology

Cost-benefit analysis is a structured methodological framework used to guide decision-making in the planning and evaluation of transportation projects.

Its primary purpose is to provide a rational basis for determining whether a project should proceed, and if so, which among several alternatives offers the greatest overall value. The analysis aims to assess all relevant costs and benefits associated with each option, translating them into a common metric to support transparent and informed choices.

Two key objectives typically motivate a cost-benefit analysis. First, it can be used to support a go/no-go decision — that is, to evaluate whether a proposed project is worthwhile from an economic standpoint. Second, it provides a basis for comparing different variants of a project. For example, planners might assess alternative routes, technologies, or service levels, with the goal of selecting the option that maximizes societal benefit relative to cost.

A critical first step in any cost-benefit analysis is the collection of appropriate data. This begins with the identification of stakeholders — those who will be affected by the project, either directly or indirectly. Understanding who the relevant actors are helps ensure that the analysis captures a comprehensive range of perspectives. Next, it is important to determine what aspects of the project matter most. This includes impacts such as travel time savings, safety improvements, environmental effects, and broader economic or social consequences. Once these aspects are identified, they must be translated into measurable indicators. The choice of indicators depends on the nature of the project and the objectives of the analysis, and may include metrics such as vehicle operating costs, accident rates, emissions, or property values.

Once data have been collected and relevant indicators identified, the analysis proceeds by combining the indicators into aggregate measures of cost and benefit. This involves forecasting future trends, applying appropriate discount rates to account for the time value of money, and estimating the net present value of each alternative. The resulting figures allow for a systematic comparison of options, highlighting the trade-offs involved and helping decision-makers identify the most advantageous course of action.

Throughout the process, transparency and consistency are essential. Assumptions must be clearly stated, methods rigorously applied, and uncertainties openly acknowledged. By doing so, cost-benefit analysis can serve not only as a technical tool, but also as a means to foster accountability and trust in public investment decisions.

11.3.1 Stakeholders

A central aspect of cost-benefit analysis is the recognition that transportation projects affect a wide range of stakeholders. Identifying and understanding

these stakeholders is important for conducting a comprehensive and balanced evaluation. The impacts of a project are rarely confined to a single group, and the perceived costs and benefits can vary significantly depending on one's role, perspective, and interests.

One key stakeholder group consists of travelers — individuals who use the transport system for commuting, leisure, business, or other purposes. For them, the primary benefits of a project may include reduced travel times, improved comfort, enhanced reliability, or increased safety. At the same time, changes to routes, services, or pricing structures may impose new costs or inconveniences.

Transport operators, both public and private, form another important group. These stakeholders are directly affected by changes in demand, operating costs, infrastructure access, and regulatory conditions. A new rail line or tunnel, for instance, might offer opportunities for increased revenue and efficiency, but could also entail adjustments to service patterns or investments in new rolling stock.

Public authorities — such as transport ministries, regional governments, or municipal planning agencies — are responsible for planning, funding, and regulating transportation infrastructure. Their perspective extends beyond the immediate users to include broader social, economic, and environmental objectives. For these actors, cost-benefit analysis is a tool to guide resource allocation and ensure accountability in public spending.

Finally, the effects of transportation projects often extend to society at large. Environmental impacts, land use changes, noise pollution, and greenhouse gas emissions may affect communities that are not directly involved in the transport system. These indirect effects must also be considered in the analysis, as they represent real costs and benefits distributed across the population.

It is important to acknowledge that a cost to one stakeholder may represent a benefit to another. For example, a toll on a new roadway may be a burden to individual drivers but a source of funding for the public authority. Similarly, a shift from road to rail freight might reduce emissions and improve safety for the general public, even if it results in higher logistics costs for some firms. Recognizing these trade-offs is essential to avoid biased or incomplete assessments. A robust cost-benefit analysis seeks to account for all stakeholder perspectives, thereby supporting decisions that reflect the collective interest.

11.3.2 Indicators

The foundation of any cost-benefit analysis lies in the identification and quantification of relevant indicators. These indicators serve as measurable representations of the various costs and benefits associated with a transportation project.

Cost indicators can be divided into two main categories: long-term and short-term. Long-term costs typically involve investments with a duration of one year or more. These include expenditures related to the design and engineering of infrastructure, the construction of physical assets such as tunnels, bridges, or stations, and the acquisition of vehicles or rolling stock. Because these investments are incurred over extended periods and are often subject to changes in price levels, it is important to adjust long-term cost estimates for inflation, using appropriate financial and economic discounting methods.

Short- and medium-term costs, on the other hand, are typically recurring and operational in nature. These include the day-to-day costs of operating services, maintaining infrastructure and vehicles, and managing personnel and logistics. Although they are smaller in scale compared to capital expenditures, operational costs accumulate over the project's lifetime and can significantly affect its overall economic viability.

In addition to classifying costs by time horizon, indicators can also be categorized as monetary or non-monetary. Monetary indicators are relatively straightforward to quantify and include fares paid by users, tolls collected on infrastructure, and taxes levied to support the operation or construction of transport services. These indicators reflect direct financial flows and are often easily derived from market data or existing budgetary frameworks.

Non-monetary indicators, however, capture impacts that are not directly priced in markets but are nonetheless critical to evaluating the societal value of a project. These include travel time savings for passengers, which reflect the efficiency of the system; reductions in traffic accidents, which contribute to public safety; and environmental impacts such as noise, air pollution, and CO_2 emissions. Other indicators may reflect changes in land use, urban development patterns, or spatial equity — for instance, how a new project improves access to jobs or services in underserved areas.

Because non-monetary impacts are not inherently expressed in monetary terms, cost-benefit analysis often requires converting them using standardized methods or willingness-to-pay estimates. While this introduces uncertainty and complexity, it allows for a more holistic comparison of project alternatives. Ultimately, a well-designed set of indicators — covering both costs and benefits, monetary and non-monetary, short- and long-term — ensures that the analysis reflects the multifaceted nature of transportation

investments and their implications for society.

11.3.3 Illustration

Let us consider a hypothetical transportation project of high technological ambition: the construction of a Hyperloop system between Geneva and Zürich. The concept of the Hyperloop involves a vacuum-sealed tube through which pressurized pods travel at extremely high speeds using magnetic propulsion. In this scenario, the proposed system would cover the 280 kilometers between the two cities in just 30 minutes, dramatically reducing current travel times and redefining intercity mobility.

Such an ambitious infrastructure project would involve a wide array of stakeholders, each of whom would be affected in different ways. Travelers stand to benefit from a substantial reduction in travel time, potentially improving productivity, convenience, and access to employment opportunities. However, these benefits may be offset by higher fares if the service is priced as a premium product. For transport operators, the project represents both a major capital investment and a potential source of future revenue through fare collection and network expansion. Public authorities may contribute to the financing of the infrastructure and play a role in regulatory oversight, while also benefiting from tax revenues or strategic gains linked to regional development.

Beyond the directly involved parties, the project would have significant implications for society at large. These include environmental impacts, such as potential reductions in air pollution and CO_2 emissions if the Hyperloop replaces car or air travel. On the other hand, construction and land acquisition could cause ecological disruptions or raise questions of land use fairness. Spatial impacts also merit attention, as the improved connectivity between the cities may influence housing markets, labor mobility, and regional equity.

Table 11.1 summarizes the key indicators associated with the Geneva–Zürich Hyperloop project and how they map across different stakeholder groups. The entries represent either costs or benefits that each group may incur or enjoy. Note that some effects are ambiguous or may involve trade-offs, such as the societal consequences of faster travel and infrastructure expansion. While simplified, the table provides a useful starting point for organizing the evaluation process.

11.3.4 Issues

While cost-benefit analysis offers a powerful framework for evaluating transportation projects, its effective application raises several important issues.

	Travelers	Operators	Authorities	Society
Capital investment		Cost	Cost	
Operations and maintenance		Cost		
Fare or toll	Cost	Benefit		
Taxes		Cost	Benefit	
Travel time savings	Benefit			Cost/Benefit
Pollution				Cost/Benefit
Land use				Cost
Spatial impacts				Cost/Benefit

Table 11.1: Stakeholders and indicators for a Hyperloop project between Geneva and Zürich

These challenges must be understood and addressed to ensure that the analysis provides meaningful guidance for decision-making.

A first point of concern relates to the distinction between private and public projects. In the case of a commercial enterprise, such as the airline example described in Section 11.1, the analysis is typically conducted from the perspective of the operator. The primary focus is on the financial viability of the investment: whether revenues will exceed costs, and what return on investment can be expected. However, public infrastructure projects serve a broader purpose. Their evaluation must consider a wider range of stakeholders, including users, operators, public authorities, and society at large. This means that all dimensions of impact — not just financial flows to the operator — must be taken into account. Public investment decisions must be guided not only by profitability, but also by considerations of equity, environmental sustainability, and strategic importance.

Another recurrent issue in cost-benefit analysis is the accurate estimation of monetary costs, particularly for large infrastructure projects. There is extensive empirical evidence showing that such projects often suffer from significant cost overruns. Initial estimates tend to be overly optimistic, either because of insufficient data, methodological weaknesses, or strategic underestimation. Underestimating costs can lead to poor investment decisions, misallocation of public resources, and erosion of public trust. For this reason, analysts must apply rigorous and conservative approaches to cost

forecasting, incorporating risk margins and learning from comparable past projects.

In addition to monetary costs, transportation projects generate a wide range of non-monetary costs and benefits. These include improvements in travel time, safety, air quality, and land use, among others. Unlike monetary indicators, these effects are not naturally expressed in a common unit, making them difficult to compare or aggregate. A key challenge is to find appropriate ways to quantify and, when necessary, monetize these impacts. This often involves the use of shadow prices, willingness-to-pay estimates, or scoring systems. Even when monetary conversion is not possible or appropriate, analysts must find transparent methods to account for these effects in the final decision.

The combination of monetary and non-monetary indicators also raises normative questions. Should all effects be reduced to a single monetary value? Or should some impacts — such as environmental preservation or social equity — be treated as constraints or objectives in their own right? These questions do not have simple answers and often require the integration of cost-benefit analysis with broader decision-making frameworks. Ultimately, the purpose of evaluation is not to produce a definitive verdict, but to provide decision-makers with the best available information to make choices that align with the public interest.

The rest of the chapter is dedicated to the discussion of those issues.

11.4 Estimation of monetary costs

One of the most persistent and well-documented issues in the evaluation of large infrastructure projects is the systematic underestimation of monetary costs. This phenomenon has been observed across a wide range of sectors and countries, and it undermines the credibility of cost-benefit analyses when used as a decision-support tool. While uncertainties are inevitable in forecasting long-term investments, the scale and consistency of cost overruns suggest deeper structural problems in how project costs are estimated and communicated.

A striking illustration of this issue can be found in the organization of the Olympic Games (Andreff, 2012), which often involve substantial public investment in transportation, sports venues, accommodation, and security infrastructure. Table 11.2 presents a comparison between the announced budgets and actual costs for several recent Olympic Games. In London 2012, for example, the initial budget was 3.4 billion pounds, but the final cost reached 11.6 billion pounds — an increase by a factor of 3.4. The discrepancy

was even more dramatic for Beijing 2008, where the estimated cost was just under 2 billion dollars, while the real cost exceeded 43 billion, resulting in an overrun of more than twentyfold. Although not all cases are as extreme, even the relatively moderate overrun in Athens 2004 (30 percent) and Sydney 2000 (90 percent) illustrates how frequent and significant these deviations can be.

Olympic Games	Budget	Real Costs	Cost Multiplier
London 2012	3.4 billion £	11.6 billion £	×3.4
Beijing 2008	1.9 billion $	43–45 billion $	×23.7
Athens 2004	4.6 billion €	6 billion €	×1.3
Sydney 2000	3.4 billion $	6.6 billion $	×1.9

Table 11.2: Budgeted vs. actual costs of recent Olympic Games

This pattern is not confined to sports events. In transportation and infrastructure more broadly, empirical research (e. g. Flyvbjerg et al., 2003) shows that 90 percent of very large infrastructure projects experience cost overruns. Rail projects tend to exceed their budgets by an average of 45 percent, tunnels and bridges by 34 percent, and road projects by 20 percent. Across the transportation sector, the average overrun is approximately 28 percent. Alarmingly, this tendency has not shown any significant improvement over time. Despite decades of experience and advances in planning methodologies, no systematic learning seems to be taking place.

The consequences of poor cost estimation are far-reaching. Underestimated costs can lead to flawed investment decisions, budgetary shortfalls, and reduced public trust. When costs balloon, projects may need to be scaled back, delayed, or require additional funding, sometimes at the expense of other public priorities. This undermines the reliability of cost-benefit analysis, especially when its results are used to justify large public expenditures.

To mitigate these risks, analysts and decision-makers must approach cost estimation with a critical mindset. Conservative assumptions, benchmarking against comparable projects, and formal risk assessment procedures should become standard practice. Furthermore, transparency in how cost figures are derived—and accountability for their accuracy—are essential to fostering responsible project planning and evaluation.

11.5 Estimation of non-monetary costs: transforming everything into monetary units

When evaluating the benefits and costs of transportation projects, one often faces the challenge that many crucial impacts are measured in different units. For example, improvements in travel time, safety, and environmental quality cannot be directly added together because they do not share a common metric. There are two principal approaches to overcome this difficulty. The first approach is to transform all indicators into a single common metric, usually monetary units. This monetization process involves estimating the economic value of non-monetary effects — such as assigning a value to each minute of travel time saved or each kilogram of CO_2 reduced — so that these effects can be aggregated with conventional financial costs and benefits. This method allows decision-makers to compare diverse impacts on a unified scale and is particularly useful when making cost-benefit comparisons between different projects. However, it also relies on assumptions about the value of non-market effects, which may introduce uncertainty.

The second approach is multi-criteria analysis, where each indicator is retained in its original unit. While this section focuses on the monetization strategy, it is important to recognize that multi-criteria analysis offers an alternative that might be more appropriate when monetizing certain non-monetary outcomes is problematic or controversial. We discuss this multi-criteria method in greater detail in Section 11.6.

There are several approaches available to transform non-monetary indicators into monetary units, each grounded in a different conceptual foundation. The choice of method depends on the nature of the impact being considered, the availability of data, and the objectives of the analysis (see, for instance, Duong, 2009). The four principal methods are the *behavioral approach*, the *cost for society*, the use of a *shadow price*, and the reference to an existing *market price*. These approaches offer distinct perspectives on valuation and are often applied in complementary ways within the framework of cost-benefit analysis.

In the following sections, we explore each of these methods in more detail.

11.5.1 Behavioral approach

The behavioral approach to monetizing non-monetary indicators relies on observing or inferring individuals' preferences and behaviors in order to estimate the value they assign to specific changes in their environment. The fundamental idea is that people's willingness to pay for certain improvements

— or to avoid certain harms — can reveal the implicit monetary value they place on non-monetary aspects such as time, safety, or environmental quality. Within this approach, three main techniques are commonly used: consumer surplus analysis, contingent valuation, and risk mitigation valuation.

The first technique is based on the concept of *consumer surplus* (see Section 2.3), which represents the difference between what an individual is willing to pay for a good or service and what they actually pay. In the context of transportation, this can be used to estimate the benefits of a new project by comparing the generalized cost of travel (including time and money) before and after the project. One widely used method for estimating consumer surplus is the *rule of half*, which assumes that for users switching from one mode or route to another due to a policy change, the average benefit is half the difference in generalized costs, if the supply and demand curves are linear (see Figure 2.12). This approach is particularly useful for approximating travel time savings and pricing effects in large-scale transportation models.

The second technique is known as *contingent valuation*. This method uses surveys to directly ask individuals about their willingness to pay for a hypothetical improvement or to accept compensation for a hypothetical loss. For example, the value of time (see Section 3.2) measures how much individuals would be willing to pay to reduce their daily commute by ten minutes. Similarly, a "value of reliability" is the willingness to pay to improve the reliability of a public transport service. Contingent valuation is particularly useful for valuing attributes that are not traded in any market and where no observable behavior exists. However, it is subject to several limitations, including hypothetical bias and the potential discrepancy between willingness to pay and willingness to accept, the latter often being significantly higher due to loss aversion or emotional attachment.

The third technique within the behavioral approach involves valuing *risk mitigation*. This refers to the estimation of how much individuals are willing to pay to reduce risks to their health or safety. For example, the *value of statistical life* is a key parameter in transportation safety analysis and is typically derived from studies of labor markets, where wage differentials are used to estimate how much extra income workers demand to accept higher job-related risks. Similarly, the value of reducing the probability of injuries can be inferred from consumer behavior, such as the purchase of safety equipment or insurance. These values are then used to assign a monetary equivalent to reductions in accident rates or improvements in safety resulting from a project.

Let's illustrate the valuation of life and health improvements, in the context of risk mitigation. This involves estimating how much individuals are willing to pay to reduce the probability of death or injury. The resulting

measure, often referred to as the *Value of a Statistical Life* (VSL), plays a critical role in evaluating policies or projects that affect safety, such as investments in road infrastructure, public transport safety enhancements, or environmental regulation.

In Switzerland, the official VSL is derived from studies conducted by the OECD using stated preference methods (OECD - Organisation for Economic Co-operation and Development, 2012, ARE - Office fédéral du développement territorial, 2022). These methods involve survey-based techniques in which individuals are asked how much they would be willing to pay for small reductions in their mortality risk. The aggregated results are then scaled up to reflect the value of avoiding one statistical death in a population. According to official estimates, the VSL in Switzerland was 6.4 million Swiss francs in 2010 and increased to 6.9 million by 2021. These values are used in official economic appraisals to quantify the benefits of interventions that improve public safety.

In the United States, historical estimates of the value of life also reflect a behavioral approach, but often rely on revealed preferences — particularly wage-risk tradeoffs in the labor market. In this context, economists examine how much extra income workers demand to accept more dangerous jobs, thereby inferring how much people value marginal changes in mortality risk. Table 11.3 shows estimates of the VSL in the U.S. from 1940 to 1980, in thousands of 1990 U.S. dollars. These figures, ranging from approximately $700,000 in 1940 to over $5 million in 1980, reflect both increased income levels and a growing societal concern for safety over time.

Table 11.3: Estimates of the Value of a Statistical Life in the United States

Year	Lower bound ($K)	Upper bound ($K)
1940	713	996
1950	1,122	1,755
1960	1,085	2,132
1970	2,792	4,937
1980	4,144	5,347

In thousands of 1990 U.S. dollars. Source: Costa and Kahn, 2004

While the VSL captures the value of avoiding a statistical death, it is also useful to consider the value of extending life through medical or policy interventions. For example, a study reported by Kingsbury (2008) found that if Medicare in the United States spent an additional $129,000 to treat a specific group of patients, it would result, on average, in one additional quality-adjusted life year (QALY) per person. In this case, the QALY serves

as a unit that adjusts life expectancy gains by the quality of health experienced during those years, making it a valuable tool in health economics.

These examples underscore how the concept of willingness to pay can be applied not only to convenience and comfort, but also to life-and-death considerations. By grounding valuations in actual or stated behavior, the behavioral approach provides a framework for incorporating safety and health improvements into economic evaluations in a systematic, albeit ethically sensitive, manner.

Together, these techniques provide a rigorous framework for assigning monetary values to otherwise intangible benefits. While each has its own methodological challenges, the behavioral approach remains one of the most widely accepted and empirically grounded methods for integrating non-monetary impacts into cost-benefit analysis.

11.5.2 Shadow price

An alternative to the behavioral approach for monetizing non-monetary indicators is the use of *shadow prices*. A shadow price is an artificial or imputed value assigned to a good or externality that does not have a clearly observable market price. It is typically established through expert judgment, policy negotiation, or regulatory consensus, with the aim of internalizing costs or benefits that are otherwise external to market transactions.

One widely used example of a shadow price is the valuation of carbon dioxide (CO_2) emissions. Because CO_2 contributes to climate change but is not traded like a conventional commodity in most markets, governments often assign a notional[1] cost per ton of CO_2 to reflect its environmental impact. These values vary by country and over time, depending on policy priorities, climate targets, and methodological choices. In Switzerland, for instance, the official shadow price for CO_2 rose from CHF 96 per ton in 2021 to CHF 120 in 2022, reflecting a growing commitment to climate policy. In contrast, the European Union used a lower value of CHF 46 per ton in 2021. Finland was the first country to adopt a carbon pricing mechanism in 1990, pioneering the use of shadow prices in climate-related cost assessments.

While shadow prices can be a valuable tool for bringing non-market effects into the scope of economic evaluation, their use is not without limitations. A fundamental concern is the degree of *subjectivity* involved in determining the appropriate value. Unlike market prices, which emerge from the interaction of supply and demand, shadow prices are often derived through negotiation or expert modeling. This makes them highly sensitive to the assumptions,

[1] estimated or policy-based.

values, and institutional contexts that underpin their construction.

Moreover, shadow prices can be vulnerable to *conflicts of interest* and political influence. Stakeholders with divergent agendas — such as industry groups, environmental organizations, and government ministries — may advocate for different valuations based on how they expect the resulting policy implications to affect them. In some cases, powerful lobbies may exert pressure to keep prices artificially low in order to minimize compliance costs or to protect vested interests. Conversely, advocacy groups may push for higher valuations to highlight long-term environmental risks or social costs.

As a result, the legitimacy and credibility of shadow prices depend critically on the transparency and rigor of the processes through which they are defined. Ideally, shadow prices should be grounded in robust scientific evidence and updated regularly to reflect changing knowledge and conditions. At the same time, they must strike a balance between economic theory and political feasibility, acknowledging that some level of compromise is often unavoidable in public policy.

In summary, shadow prices are a practical way to integrate otherwise invisible costs into cost-benefit analysis. However, their effectiveness depends on how well they are designed and governed. When used thoughtfully and transparently, they can support more comprehensive and equitable decision-making. But when used without scrutiny, they risk introducing bias or masking underlying trade-offs.

11.5.3 Market price

Another approach to monetizing non-monetary indicators relies on the use of market prices. When a direct market does not exist for a particular externality — such as pollution or noise — governments can intervene by creating an *artificial market*. The objective is to assign a monetary value to external costs by making them tradable commodities. This mechanism transforms previously unpriced negative externalities into goods that are subject to supply and demand, thereby revealing their economic value through the price of tradeable permits.

The creation of such markets typically follows a regulatory framework established by public authorities. First, the government identifies a harmful activity that produces a negative externality, such as air pollution or carbon emissions. It then sets a cap on the total allowable level of that externality and issues a limited number of permits that grant the right to produce a specified amount of it. These permits can be allocated for free or auctioned and are tradable in the open market. This system is commonly referred to as *cap-and-trade* (Flachsland et al., 2011). By introducing scarcity, the

market assigns a price to the externality, and economic agents are incentivized to reduce their emissions if the cost of doing so is lower than purchasing additional permits.

A number of real-world applications illustrate the effectiveness of this approach. One of the earliest examples is the lead phase-down program in the United States (1979–1996), which targeted the removal of lead from gasoline (Newell and Rogers, 2003). Refineries were issued tradeable permits limiting lead content, allowing flexibility in compliance while gradually reducing overall emissions. This market-based approach significantly accelerated the reduction of lead in gasoline, achieving in 1981 what would have otherwise taken until 1987 without such measures.

Another notable case is the Ecopoint system in Austria (1995–2006), aimed at limiting noise and air pollution from heavy goods vehicles crossing the country. Trucking companies were allocated a fixed number of Ecopoints, which they could trade among each other depending on their routes and vehicle types. The system effectively encouraged cleaner and quieter transportation technologies (Caveri, 2003).

A more recent and ongoing example is California's Low Emission Vehicle (LEV) and Zero Emission Vehicle (ZEV) programs, initiated in the 1990s. These programs require automobile manufacturers to sell a certain proportion of low or zero-emission vehicles, such as electric cars. Manufacturers who exceed the requirement can sell their excess credits to others who fall short, creating a dynamic market that assigns a price to clean vehicle technology and accelerates innovation and adoption by rewarding early movers (California Air Resources Board, 2024, McConnell and Leard, 2021).

11.5.4 Summary

Transforming non-monetary indicators into monetary values is a common and useful strategy in cost-benefit analysis, but it is not without limitations. As discussed, there is no single or universally accepted way to carry out this transformation. The value assigned to time, health, or environmental quality may vary depending on the context, methodology, cultural preferences, and institutional frameworks. Whether through behavioral inference, shadow pricing, or market-based instruments, each approach involves certain assumptions and degrees of simplification.

One major source of complexity is the unavoidable role of *subjectivity*. Estimating how much people value safety, clean air, or reduced noise often requires interpreting preferences through models, surveys, or policy negotiations. Even when supported by empirical data, the outcome is still sensitive to methodological choices. Different analysts or institutions may arrive at

different valuations for the same indicator, which can significantly affect the final evaluation of a project.

For these reasons, analysts and decision-makers sometimes seek alternative approaches that preserve the integrity of non-monetary indicators without converting them into a common monetary unit. One such alternative is the *multi-criteria approach*, which allows for the evaluation of multiple, diverse impacts side by side. Rather than aggregating everything into a single cost-benefit figure, this method recognizes that different objectives — such as environmental protection, social equity, or economic efficiency — may be best addressed using separate criteria.

In the following section, we introduce the principles and tools of multi-criteria analysis, explore its potential advantages in dealing with complex transport projects, and discuss how it can complement or enhance traditional cost-benefit methods.

11.6 Estimation of non-monetary costs: multicriteria analysis

Multi-criteria analysis (MCA) offers an alternative to monetary valuation when dealing with complex decisions that involve multiple, diverse impacts. Instead of converting all effects into a single monetary unit, MCA evaluates projects across several distinct indicators, each representing a specific objective or concern. This approach acknowledges that in many cases, costs, environmental impacts, and social benefits are inherently incommensurable and should not be forced into a common scale.

In MCA, a project $\mathtt{i}$ is evaluated using a set of indicators $\mathtt{q}^{\mathtt{i}}_1, \mathtt{q}^{\mathtt{i}}_2, \ldots, \mathtt{q}^{\mathtt{i}}_{\mathtt{K}}$, where each $\mathtt{q}^{\mathtt{i}}_{\mathtt{k}}$ measures performance with respect to criterion $\mathtt{k}$. These indicators may include financial costs, travel time, environmental effects (such as CO_2 emissions), user benefits (like consumer surplus), or other relevant impacts. The guiding principle is that lower values indicate better performance — hence, we seek to minimize each indicator. If an indicator reflects a quantity that is better when higher (such as consumer surplus), it can simply be multiplied by -1 so that the objective of minimization is preserved across all indicators.

Importantly, in this framework, the indicators are treated as incomparable: no attempt is made to combine them into a single score. Instead, each criterion retains its original unit and meaning, preserving the richness of the decision-making context. This allows analysts and stakeholders to consider trade-offs transparently. For example, suppose project $\mathtt{i}$ has lower costs and

lower CO_2 emissions than project j, but offers slightly longer travel times. Whether project i is preferable to project j depends not on a pre-defined formula, but on how decision-makers value the trade-off between environmental benefit and user convenience.

To make this concrete, consider the following example: q_1^i and q_1^j represent the financial cost of two projects; q_2^i and q_2^j reflect the average travel time for users; q_3^i and q_3^j quantify the expected CO_2 emissions; and $-q_4^i$ and $-q_4^j$ correspond to the consumer surplus, which is better when higher. In this setting, a project may dominate another if it performs better on all indicators, but more often, trade-offs must be made—such as accepting higher costs in exchange for lower emissions.

A central concept in multi-criteria analysis is that of *dominance*, which provides a formal basis for comparing alternatives when multiple indicators are involved.

Formally, consider two projects i and j, each evaluated according to a set of K indicators. We say that project i *dominates* project j, denoted $i \prec j$, if two conditions are satisfied. First, project i must be no worse than project j in all indicators:

$$\forall k \in \{1, \dots, K\}, \quad q_k^i \leqslant q_k^j.$$

Second, project i must be strictly better than project j in at least one indicator:

$$\exists k \in \{1, \dots, K\}, \quad q_k^i < q_k^j.$$

This means that project i either matches or outperforms project j on every front, and strictly improves upon it in at least one dimension. In such cases, project j can be discarded from further consideration, since there exists another option that is objectively better.

The dominance relation has several useful properties that shape how we interpret it in analysis. First, it is *not reflexive*: a project does not dominate itself, i.e., $i \not\prec i$. Second, it is *not symmetric*: if $i \prec j$, it does not follow that $j \prec i$; in fact, the opposite must be true — if $i \prec j$, then $j \not\prec i$. This expresses the asymmetry inherent in a "better than" relationship.

Dominance is also *transitive*: if project i dominates project j, and project j dominates project ℓ, then it must be the case that project i dominates project ℓ; formally, $i \prec j$ and $j \prec \ell \Rightarrow i \prec \ell$. This transitivity allows for consistent pruning of inferior alternatives from the set of feasible projects.

However, the dominance relation is *not complete*. There may exist projects i and j such that neither dominates the other — that is, $i \not\prec j$ and $j \not\prec i$. This situation arises frequently in real-world decision-making, where trade-offs exist between indicators. For example, one project may be cheaper but

more polluting, while another is cleaner but more expensive. In such cases, further analysis is needed to guide the selection process.

The concept of dominance can be illustrated with a simple example involving four projects, each evaluated using two indicators: cost and travel time. These indicators are both to be minimized, meaning that a project is preferred if it offers lower cost and/or shorter travel time. The positions of the projects in the cost-time plane are shown in Figure 11.1, where each dot corresponds to a project alternative, and their relative coordinates reflect their performance in terms of the two criteria.

We begin by examining project $\mathtt{i}_3$, located at the bottom-left corner of the shaded grid. This project has both lower cost and lower travel time compared to project $\mathtt{i}_2$, which lies at the top-right. Since $\mathtt{i}_3$ is no worse in either criterion and strictly better in both, we say that $\mathtt{i}_3$ *dominates* $\mathtt{i}_2$, written $\mathtt{i}_3 \prec \mathtt{i}_2$. Similarly, $\mathtt{i}_3$ also dominates $\mathtt{i}_1$, as it achieves the same travel time with a lower cost.

However, when comparing project $\mathtt{i}_1$ and project $\mathtt{i}_4$, we find that neither dominates the other. Project $\mathtt{i}_1$ has a lower cost but a longer travel time than $\mathtt{i}_4$, while $\mathtt{i}_4$ offers a time advantage at a higher cost. In this case, the dominance condition is not satisfied in either direction: $\mathtt{i}_1 \not\prec \mathtt{i}_4$ and $\mathtt{i}_4 \not\prec \mathtt{i}_1$. This situation exemplifies the presence of a trade-off between criteria — no project is strictly better across the board, and further analysis or stakeholder preferences are needed to decide between them.

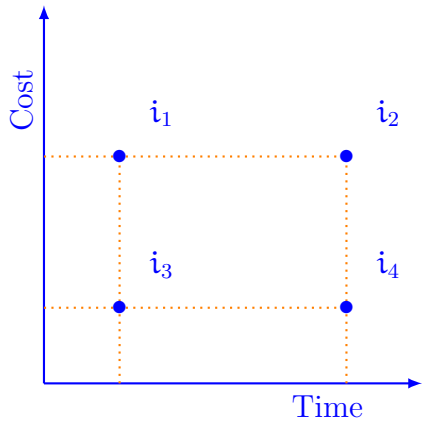


Figure 11.1: Example of dominance relationships between four projects evaluated by cost and travel time.

The concept of dominance leads to the concept of *Pareto optimality*, which serves as a criterion for identifying efficient solutions when multiple, conflicting objectives must be considered. Unlike in single-objective optimization —

where the goal is to find a unique best solution — multi-criteria problems typically give rise to a set of alternatives that are all, in a certain sense, optimal. Pareto optimality provides the formal foundation for this idea.

Let $\mathcal{C}$ denote the set of all feasible projects. A project $\mathfrak{i}^* \in \mathcal{C}$ is said to be *Pareto optimal* if it is not dominated by any other project in the set. Formally, this means:

$$\nexists \mathfrak{j} \in \mathcal{C} \text{ such that } \mathfrak{j} \prec \mathfrak{i}^*.$$

In other words, there is no project $\mathfrak{j}$ in the set that is at least as good as $\mathfrak{i}^*$ in all indicators and strictly better in at least one.

The intuition behind Pareto optimality is straightforward yet powerful: a project is Pareto optimal if no improvement can be made in any one criterion without incurring a degradation in at least one other. For instance, if a project offers the lowest cost but a longer travel time, it might still be considered optimal if reducing the travel time would necessarily raise the cost. Such solutions reflect the inherent trade-offs in real-world decisions.

The set of all Pareto optimal solutions forms what is known as the *Pareto frontier* or *efficient frontier*. This frontier delineates the boundary of achievable performance: any movement beyond it in one dimension must come at the expense of performance in another.

In practice, the application of Pareto optimality does not end with the identification of efficient solutions. Instead, it marks the beginning of a more nuanced decision-making process. The first step is to compute the *Pareto optimal set*:

$$\mathsf{P}^* = \{\mathfrak{i}^* \in \mathcal{C} \mid \nexists \mathfrak{j} \in \mathcal{C} \text{ such that } \mathfrak{j} \prec \mathfrak{i}^*\},$$

that is, the set of all projects that are not dominated by any other in the feasible set $\mathcal{C}$. These projects form the Pareto frontier and represent the best trade-offs available given the considered indicators.

Once the Pareto optimal set is identified, the next step involves selecting a final project from among these efficient alternatives. This selection process is inherently subjective, as it requires decision-makers to articulate and apply preferences among competing objectives. Political priorities, considerations of social equity, environmental values, and stakeholder input all play critical roles at this stage. Since no project in the Pareto set is objectively superior to the others across all criteria, the final choice reflects a deliberate prioritization of certain impacts over others.

The procedure typically involves focusing attention exclusively on the Pareto optimal set, thereby reducing the complexity of the decision space while retaining all meaningful alternatives. Within this reduced set, decision-makers can engage in transparent deliberation about the relative importance of each indicator. If needed, additional indicators can be introduced to better

capture specific concerns or to refine the evaluation. When such new criteria are added, the Pareto set must be updated accordingly, as the introduction of new dimensions may alter the dominance relationships among projects.

Pareto optimality thus reframes decision-making by shifting the emphasis away from identifying a single "best" project and toward managing the trade-offs among equally efficient options. This approach promotes transparency, as it makes explicit which compromises are necessary and which preferences guide the final decision. It also supports accountability, by clearly distinguishing the technical analysis that defines the Pareto frontier from the political and social judgments that inform the final selection.

11.6.1 Example

To illustrate the practical use of multi-criteria analysis, consider the case of railway timetable rescheduling following a major disruption in operations. Such disruptions — caused by events like technical failures, accidents, or severe weather — require the rapid implementation of a temporary or *disposition timetable*. The goal is to restore operations in a way that minimizes negative impacts while respecting operational constraints.

In this context, the disposition timetable may involve various corrective actions: trains may be fully canceled, partially canceled (i.e., ending service before their final destination), delayed, or rerouted. In certain cases, emergency replacement trains might be dispatched to maintain service continuity. Each of these actions carries different implications for passengers, operators, and the broader transportation system.

The rescheduling problem is inherently multi-objective. First, there is the need to *minimize passenger inconvenience*, typically measured in terms of lost time, missed connections, or increased uncertainty. Second, there is a desire to *minimize costs*, which may include additional crew hours, energy consumption, or penalties for late arrivals. Third, operators aim to *minimize deviations from the original timetable*, preserving the planned structure of operations to the greatest extent possible. Maintaining this structure facilitates both passenger expectations and operational feasibility in subsequent hours or days.

These objectives are not necessarily aligned. For example, minimizing cost might lead to the cancellation of less critical trains, which may in turn increase inconvenience for affected passengers. Conversely, minimizing inconvenience could require extensive rerouting or the addition of emergency services, thereby raising operational costs. Similarly, any intervention that improves passenger experience or cost efficiency might require significant departures from the original timetable, complicating downstream operations.

We refer the interested reader to Binder et al. (2017) for a detailed analysis of this complex optimization problem in the context of disrupted railway operations. The approach adopted in that study illustrates how multi-criteria decision-making techniques can be applied to manage conflicting goals under operational stress.

Figure 11.2 provides two examples of Pareto frontiers generated for two different disruption scenarios. Each frontier represents the set of disposition timetables that are Pareto optimal — meaning that no other feasible solution improves one objective without worsening at least one of the others. In these examples, three distinct objectives are considered: minimizing operational cost, minimizing passenger inconvenience (often measured in total delay or lost time), and minimizing deviations from the original planned timetable.

To effectively visualize this three-dimensional trade-off, the figure uses a two-dimensional coordinate system for two of the objectives — cost and inconvenience — while the third objective, deviation from the original timetable, is represented using different curves on the same plot. Each curve corresponds to a fixed level of deviation, allowing one to explore the trade-off space within that constraint.

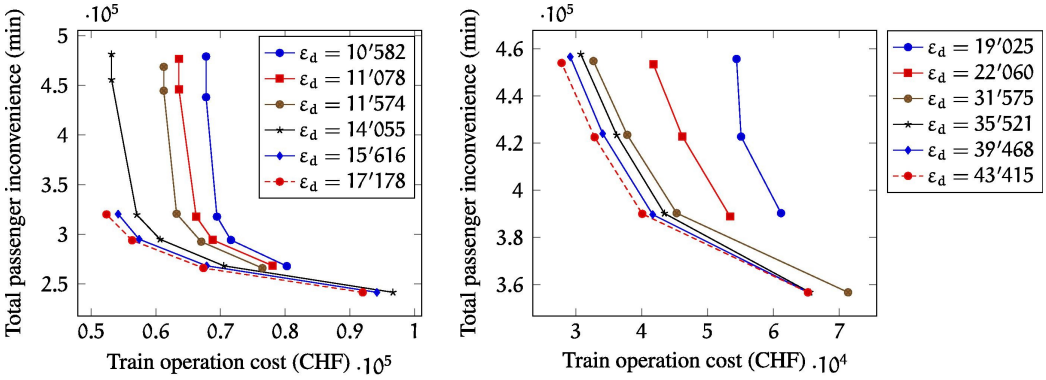
This representation helps reveal the structure of the decision space. For a given tolerance in timetable deviation, one can identify the disposition plans that offer the best compromise between cost and inconvenience. As the permitted deviation increases (i.e., moving from one curve to another), the potential for reducing cost or improving passenger outcomes also increases. However, this often comes at the price of greater disruption to the overall timetable structure, which may introduce downstream operational complications or reduced predictability for travelers.

Such visualizations are powerful tools for both analysts and decision-makers. They not only expose the efficient frontier of available solutions but also make the consequences of preference shifts or policy constraints immediately visible. This aids in making informed, transparent, and accountable decisions under complex trade-off conditions.

11.7 Conclusion

Cost-benefit analysis is a fundamental tool for evaluating public and private investment projects, particularly in the field of transportation and infrastructure. Its primary objective is to support key decisions — such as whether to proceed with a proposed project or to select among multiple competing alternatives — by systematically comparing anticipated benefits and costs.

The starting point of any cost-benefit analysis is the identification of rele-



(a) Pareto frontier for disruption Gouda-Utrecht (Gou-Utr).

(b) Pareto frontier for disruption Den Haag-Leiden- Schiphol (DeH-Lei-Sch).

Figure 11.2: Examples of Pareto frontiers for two disruption scenarios. Cost and inconvenience are shown on the axes, while deviation from the original timetable is represented by different curves.

vant *stakeholders*, including travelers, transport operators, public authorities, and society at large. Each of these groups may experience different types of impacts, and the analysis must be structured to reflect this diversity. Once stakeholders are defined, appropriate *indicators* must be selected to capture the effects of the project on each group. These may include direct monetary impacts (such as fares, taxes, or operating costs) as well as non-monetary effects (such as time savings, emissions, noise, or land use changes).

The analytical phase of cost-benefit analysis involves the systematic evaluation of these indicators. In some cases, all effects can be converted into a common unit — typically monetary — using methods such as willingness-to-pay, shadow pricing, or market-based instruments. This enables the aggregation of costs and benefits into a single net value, which simplifies comparisons across projects. However, this transformation is not always straightforward and may involve subjective assumptions or politically negotiated values.

When the conversion of indicators into monetary terms is either infeasible or undesirable, a *multi-criteria analysis* may be used instead. This approach preserves the distinct nature of each indicator and evaluates projects based on their performance across multiple dimensions. The analysis then identifies Pareto optimal solutions — those that cannot be improved in any criterion without sacrificing another — and supports selection among them based on explicit trade-offs and preferences.

Despite its structured methodology, cost-benefit analysis is not without challenges. A common issue is the systematic *underestimation of costs*, particularly in large infrastructure projects. This may result from optimism bias,

incomplete data, or strategic misrepresentation. Furthermore, even the most rigorous analyses must contend with the *role of subjectivity* in interpreting data, valuing impacts, and setting priorities. Recognizing these limitations is essential to ensure that the conclusions of cost-benefit analysis are interpreted appropriately and used to inform decisions in a transparent and accountable way.

Chapter 12

Conclusion

Transportation systems are inherently complex. They consist of interdependent components operating across multiple modes, time scales, and spatial layers. This complexity is amplified by the dynamic interactions between infrastructure, vehicles, and users, as well as by the influence of external factors such as geography, policy, and technology.

Beyond their physical and technical structure, transportation systems involve human and economic dimensions. Travel behavior, preferences, and choices play a central role in determining demand and shaping system performance. Economic principles such as elasticity, surplus, and equilibrium help explain how individuals and systems respond to changes in cost, service levels, or infrastructure.

Designing, maintaining, and operating transportation systems requires the capacity to evaluate trade-offs between often conflicting objectives. Efficiency, equity, reliability, environmental impact, and cost-effectiveness must be balanced using a variety of indicators. Quantitative models, data analysis, and scenario evaluation are essential tools in this process.

The role of the engineer in this context is to provide objective, evidence-based analysis to inform and support decision-making. While the ultimate choices are made in a policy context, engineering input ensures that these decisions are grounded in a rigorous understanding of system behavior, constraints, and possibilities.

This course has provided a brief introduction to some of the modeling tools used to address the complexity of transportation systems. The analysis of such systems is a broad field, rich with challenges that draw on diverse areas of knowledge, including engineering, optimization, data science, economics, and computer science. It is inherently polytechnical, requiring an integrated perspective to understand and shape mobility in a rapidly evolving world.

Bibliography

Akçelik, R. (1991). Travel time functions for transport planning purposes: Davidson's function, its time dependent form and an alternative travel time function, *Australian Road Research* **21**(3): 49–59.

Andreff, W. (2012). Pourquoi le coût des jeux olympiques est-il toujours sous-estimé ? la "malédiction du vainqueur de l'enchère" (winners's curse)., *Papeles de Europa* **25**.

ARE - Office fédéral du développement territorial (2022). Coûts et bénéfices externes des transports en suisse. Transports par la route et le rail, par avion et par bateau 2019.

Baker, L. (2009). Removing roads and traffic lights speeds urban travel., *Scientific American* .

Binder, S., Maknoon, Y. and Bierlaire, M. (2017). The multi-objective railway timetable rescheduling problem, *Transportation Research Part C: Emerging Technologies* **78**: 78–94. DOI: 10.1016/j.trc.2017.02.001.

Börjesson, M., Eliasson, J., Hugosson, M. B. and Brundell-Freij, K. (2012). The stockholm congestion charges—5 years on. effects, acceptability and lessons learnt, *Transport Policy* pp. 1–12.

Bureau of Public Roads (1964). *Traffic Assignment Manual*, U.S. Department of Commerce.

California Air Resources Board (2024). Zero emission vehicle (zev) program, `https://ww2.arb.ca.gov/our-work/programs/zero-emission-vehicle-program/about`. Accessed April 2024.

Caveri, L. (2003). Report on the proposal for a regulation of the european parliament and of the council on common rules for access to the market in the carriage of goods by road, *Session document A5-0019/2003*, European Parliament. Accessed April 2024.

Costa, D. L. and Kahn, M. E. (2004). Changes in the value of life, 1940–1980., *Journal of Risk and Uncertainty* **29**(2): 159–180.

Danalet, A., Justen, A. and Mathys, N. A. (2021). Trip generation in Switzerland, *Proceedings of the 21th Swiss Transportation Research Conference.* **URL:** *www.strc.ch*

Davidson, K. B. (1966). A flow travel time relationship for use in transportation planning, *Australian Road Research* **2**(1): 3–19.

Diemant, H. (1974). Informationstagung über die gotthardbasislinie vom 5. märz 1974., *Finanzabteilung*, SBB.

Duong, M. H. (2009). What is the price of carbon? five definitions, *SAPI EN. S. Surveys and Perspectives Integrating Environment and Society* (2.1).

Eliasson, J. (2012). How to solve traffic jams?, TED Talk.

Emberger, G., König, I., Krpata, R., Rollinger, W., Grundner, M. and Linien, W. (2013). Public transport for (disabled) people–the Vienna experience, *13th World Conference on Transport Research (WCTR), Rio de Janeiro*, Vol. 20.

Flachsland, C., Brunner, S., Edenhofer, O. and Creutzig, F. (2011). Climate policies for road transport revisited (ii): Closing the policy gap with cap-and-trade, *Energy Policy* **39**(4): 2100–2110. DOI: https://doi.org/10.1016/j.enpol.2011.01.053.

Flyvbjerg, B., Skamris Holm, M. K. and Buhl, S. L. (2003). How common and how large are cost overruns in transport infrastructure projects?, *Transport reviews* **23**(1): 71–88.

Garrow, L., Jones, S. and Parker, R. (2006). How much airline customers are willing to pay: an analysis of price sensitivity in online distribution channels, *Journal of Revenue and Pricing Management* **5**(4): 271–290.

Hidalgo, D., Pereira, L., Estupiñán, N. and Jiménez, P. L. (2013). Transmilenio BRT system in Bogota, high performance and positive impact – main results of an ex-post evaluation, *Research in Transportation Economics* **39**(1): 133–138. DOI: https://doi.org/10.1016/j.retrec.2012.06.005.

Khisty, C. J. and Lall, B. K. (2003). *Transportation Engineering. An introduction.*, 3rd edn, Prentice Hall.

Kingsbury, K. (2008). The value of a human life: $129,000, *Time Magazine* **20**.

Knödel, W. (1969). *Graphentheoretische Methoden und ihre Anwendungen*, Springer, Berlin.

Kolata, G. (1990). What if they closed 42d street and nobody noticed?, *The New York Times* p. 38.

McConnell, V. and Leard, B. (2021). Pushing new technology into the market: California's zero emissions vehicle mandate, *Review of Environmental Economics and Policy* **15**(1): 169–179. DOI: 10.1086/713055.

Messerli, F. (2012). Chocolate consumption, cognitive function, and Nobel laureates, *The New England Journal of Medicine* **367**: 1562–1564.

Nash, J. F. (1950a). Equilibrium points in n-person games, *Proceedings of the National Academy of Sciences* **36**(1): 48–49. DOI: 10.1073/pnas.36.1.48.

Nash, J. F. (1950b). *Non-cooperative games*, PhD thesis, Princeton University. PhD thesis.
URL: *https://www.princeton.edu/ hasselbo/papers/Nash.pdf*

Newell, R. G. and Rogers, K. (2003). The market-based lead phasedown, *Technical Report DP 03-37*, Resources for the Future.
URL: *https://media.rff.org/documents/RFF-DP-03-37.pdf*

OECD - Organisation for Economic Co-operation and Development (2012). *Mortality risk valuation in environment, health and transport policies*, OECD Publishing.

Papadimitriou, C. (2001). Algorithms, games, and the internet, *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*, STOC '01, Association for Computing Machinery, New York, NY, USA, pp. 749–753. DOI: 10.1145/380752.380883.
URL: *https://doi.org/10.1145/380752.380883*

Setyawan, A., Nainggolan, J. and Budiarto, A. (2015). Predicting the remaining service life of road using pavement condition index, *Procedia Engineering* **125**: 417–423. DOI: https://doi.org/10.1016/j.proeng.2015.11.108.