FUNDAMENTALS OF TRANSPORTATION



David Levinson et al.

Wikibook



Fundamentals of Transportation

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden borne from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by NICE CXone Expert and were supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org or visit our main website at https://LibreTexts.org.

The LibreTexts name, logo, and book covers are trademarked by LibreTexts, Inc. (a 501(c)(3) not-for-profit organization) and protected against unauthorized use.

This text was compiled on 09/22/2026

TABLE OF CONTENTS

About the Authors

1: Introduction and Planning

1.1: Fundamentals of Transportation/Introduction

1.2: Fundamentals of Transportation and Decision Making

1.3: Evaluation

2: Planning Models

2.1: Modeling

2.2: Data

2.3: Networks

2.4: Modes

3: Modeling Methods

3.1: Agent-based Modeling

3.2: Choice Modeling

3.3: Land Use Forecasting

3.4: Trip Generation

3.5: Destination Choice

3.6: 3-6 Route Choice

4: Transit

4.1: Transit Demand

4.2: Transit Operations and Capacity

4.3: Network Design and Frequency

4.4: Timetabling and Scheduling

4.5: Traveler Information for Transit

5: Traffic

5.1: Queueing

5.2: Traffic Flow

5.3: Queueing and Traffic Flow

6: Traffice Control

6.1: Shockwaves

6.2: Traffic Signals

6.3: Traffic Control Devices

6.4: Metering and Analogs

7: Geometric Design

7.1: Sight Distance

7.2: Grade

7.3: Earthwork
7.4: Horizontal Curves
7.5: Vertical Curves

Index

Glossary

This page titled Fundamentals of Transportation is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. via source content that was edited to the style and standards of the LibreTexts platform.

About the Authors

Authors of this book include David Levinson, Henry Liu of the University of Michigan, William Garrison,Mark Hickman, Adam Danczyk, Michael Corbett, Brendan Nee. Karen Dixon at Oregon State University contributed Flash animations developed by herself and her students linked to in this book.

This page titled About the Authors is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

1: INTRODUCTION AND PLANNING

1.1: Fundamentals of Transportation/Introduction
1.2: Fundamentals of Transportation and Decision Making
1.3: Evaluation

This page titled 1: Introduction and Planning is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

1.1: Fundamentals of Transportation/Introduction

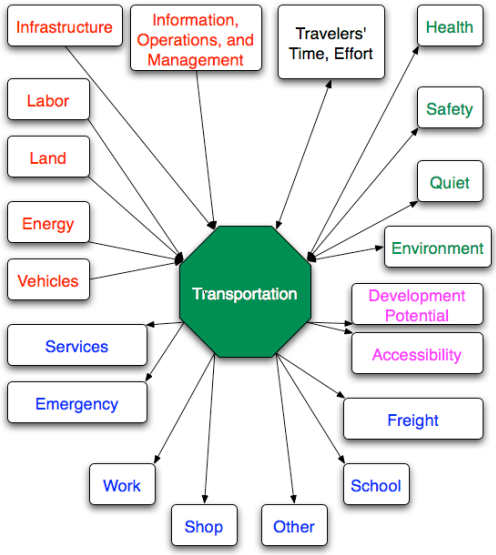
Transport is the movement of humans,animals and goods from one location to another.Transportation moves people and goods from one place to another using a variety of vehicles across different infrastructure systems. It does this using not only technology (namely vehicles, energy, and infrastructure), but also people's time and effort; producing not only the desired outputs of passenger trips and freight shipments, but also adverse outcomes such as air pollution, noise, congestion, crashes, injuries, and fatalities.If agriculture and industries are supposed to be the body of country,transport may be said to be the nerves and veins of the economy.

Figure 1 illustrates the inputs, outputs, and outcomes of transportation. In the upper left are traditional inputs (infrastructure including pavements, bridges, etc.), labor required to produce transportation, land consumed by infrastructure, energy inputs, and vehicles). Infrastructure is the traditional preserve of civil engineering, while vehicles are anchored in mechanical engineering. Energy, to the extent it is powering existing vehicles is a mechanical engineering question, but the design of systems to reduce or minimize energy consumption require thinking beyond traditional disciplinary boundaries.

On the top of the figure are Information, Operations, and Management, and Travelers' Time and Effort. Transportation systems serve people, and are created by people, both the system owners and operators, who run, manage, and maintain the system and travelers who use it. Travelers' time depends both on freeflow time, which is a product of the infrastructure design and on delay due to congestion, which is an interaction of system capacity and its use. On the upper right side of the figure are the adverse outcomes of transportation, in particular its negative externalities:

- by polluting, systems consume health and increase morbidity and mortality;
- by being dangerous, they consume safety and produce injuries and fatalities;
- by being loud they consume quiet and produce noise (decreasing quality of life and property values); and
- by emitting carbon and other pollutants, they harm the environment.

All of these factors are increasingly being recognized as costs of transportation, but the most notable are the environmental effects, particularly with concerns about global climate change. The bottom of the figure shows the outputs of transportation. Transportation is central to economic activity and to people's lives, it enables them to engage in work, attend school, shop for food and other goods, and participate in all of the activities that comprise human existence. More transportation, by increasing accessibility to more destinations, enables people to better meet their personal objectives, but entails higher costs both individually and socially. While the "transportation problem" is often posed in terms of congestion, that delay is but one cost of a system that has many costs and even more benefits. Further, by changing accessibility, transportation gives shape to the development of land.



Modalism and Intermodalism

Transportation is often divided into infrastructure modes: e.g. highway, rail, water, pipeline and air. These can be further divided. Highways include different vehicle types: cars, buses, trucks, motorcycles, bicycles, and pedestrians. Transportation can be further separated into freight and passenger, and urban and inter-city. Passenger transportation is divided in public (or mass) transit (bus, rail, commercial air) and private transportation (car, taxi, general aviation).

These modes of course intersect and interconnect. At-grade crossings of railroads and highways, inter-modal transfer facilities (ports, airports, terminals, stations).

Different combinations of modes are often used on the same trip. I may walk to my car, drive to a parking lot, walk to a shuttle bus, ride the shuttle bus to a stop near my building, and walk into the building where I take an elevator.

Transportation is usually considered to be between buildings (or from one address to another), although many of the same concepts apply within buildings. The operations of an elevator and bus have a lot in common, as do a forklift in a warehouse and a crane at a port.

Motivation

Transportation engineering is usually taken by undergraduate Civil Engineering students. Not all aim to become transportation professionals, though some do. Loosely, students in this course may consider themselves in one of two categories: Students who intend to specialize in transportation (or are considering it), and students who don't. The remainder of civil engineering often divides into two groups: "Wet" and "Dry". Wets include those studying water resources, hydrology, and environmental engineering, Drys are those involved in structures and geotechnical engineering.

Transportation students

Transportation students have an obvious motivation in the course above and beyond the fact that it is required for graduation. Transportation Engineering is a pre-requisite to further study of Highway Design, Traffic Engineering, Transportation Policy and Planning, and Transportation Materials. It is our hope, that by the end of the semester, many of you will consider yourselves Transportation Students. However not all will.

"Wet Students"

I am studying Environmental Engineering or Water Resources, why should I care about Transportation Engineering?

Transportation systems have major environmental impacts (air, land, water), both in their construction and utilization. By understanding how transportation systems are designed and operate, those impacts can be measured, managed, and mitigated.

"Dry Students"

I am studying Structures or Geomechanics, why should I care about Transportation Engineering?

Transportation systems are huge structures of themselves, with very specialized needs and constraints. Only by understanding the systems can the structures (bridges, footings, pavements) be properly designed. Vehicle traffic is the dynamic structural load on these structures.

Citizens and Taxpayers

Everyone participates in society and uses transportation systems. Almost everyone complains about transportation systems. In developed countries you seldom hear similar levels of complaints about water quality or bridges falling down. Why do transportation systems engender such complaints, why do they fail on a daily basis? Are transportation engineers just incompetent? Or is something more fundamental going on? By understanding the systems as citizens, you can work toward their improvement. Or at least you can entertain your friends at parties.

Goal

It is often said that the goal of Transportation Engineering is "The Safe and Efficient Movement of People and Goods."

But that goal (safe and efficient movement of people and goods) doesn't answer:

Who, What, When, Where, How, Why?

Overview

This libretext is broken into 4 major units

- Transportation Planning: Forecasting, determining needs and standards.
- Transit: Transit demand, service planning, and operations.
- Traffic Engineering (Operations): Queueing, Traffic Flow Highway Capacity and Level of Service (LOS)
- Highway Engineering (Design): Vehicle Performance/Human Factors, Geometric Design

Thought Questions

- What constraints keeps us from achieving the goal of transportation systems?
- What is the "Transportation Problem"?

Answer available at the end of this section.

Sample Problem

- Identify a transportation problem (local, regional, national, or global) and consider solutions. Research the efficacy of various solutions. Write a one-page memo documenting the problem and solutions, documenting your references.

Answer available at the end of this section.

Abbreviations/Key Terms

- LOS - Level of Service
- ITE - Institute of Transportation Engineers
- TRB - Transportation Research Board
- TLA - Three letter abbreviation
- Hierarchy of Roads
- Functional Classification
- Modes
- Vehicles
- Freight, Passenger
- Urban, Intercity
- Public, Private

Answers

Thought Questions

Sample Problem

This page titled 1.1: Fundamentals of Transportation/Introduction is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

1.2: Fundamentals of Transportation and Decision Making

Decision Making is the process by which one alternative is selected over another. Decision making generally occurs in the planning phases of transportation projects, but last minute decision making has been shown to occur, sometimes successfully. Several procedures for making decisions have been outlined in effort to minimize inefficiencies or redundancies. These are idealized (or normative) processes, and describe how decisions might be made in an ideal world, and how they are described in official documents. Real-world processes are not as orderly.

Applied systems analysis is the use of rigorous methods to assist in determining optimal plans, designs and solutions to large scale problems through the application of analytical methods. Applied systems analysis focuses upon the use of methods, concepts and relationships between problems and the range of techniques available. Any problem can have multiple solutions. The optimal solution will depend upon technical feasibility (engineering) and costs and valuation (economics). Applied systems analysis is an attempt to move away from the engineering practice of design detail and to integrate feasible engineering solutions with desirable economic solutions. The systems designer faces the same problem as the economist, "efficient resource allocation" for a given objective function.

Systems analysis emerged during World War II, especially with the deployment of radar in a coordinated way. It spread to other fields such as fighter tactics, mission planning and weapons evaluation. Ultimately the use of mathematical techniques in such problems came to be known as operations research, while other statistical and econometric techniques are being applied. Optimization applies to cases where data is under-determined (fewer observations than dependent variables) and statistics where data is over-determined (more observations than dependent variables). After World War II, techniques spread to universities. Systems analysis saw further mathematical development and application to a broad variety of problems.

It has been said of Systems Analysis, that it is:

- "A coordinated set of procedures which addresses the fundamental issues of design and management: that of specifying how men, money and materials should be combined to achieve a higher purpose" - De Neufville
- "... primarily a methodology, a philosophical approach to solving problems for and for planning innovative advances" - Baker
- "Professionals who endeavor to analyze systematically the choices available to public and private agencies in making changes in the transportation system and services in a particular region" - Manheim
- "Systems analysis is not easy to write about: brief, one sentence definitions frequently are trivial" - Thomas

The most prominent decision-making process to emerge from systems analysis is **rational planning**, which will be discussed next, followed by some critiques and alternatives.

TE-Systems-RationalPlanningAndDecisionMaking.png

How does one (rationally) decide what to do?

The figure identifies three layers of abstraction. The first layer (top row) describes the high level process, which we can summarize in six steps. A second layer details many of the components of the first layer. A third layer, identified by the blue box, "abstract into model or framework" depends on the problem at hand.

Overview data

The first step is observational, review and gather data about the system under consideration. An understanding of the world around is required, including specifying the system.

The problem (defined in the next step) lies within a larger system, that comprises

1. Objectives - measure the effectiveness or performance
2. Environment - things which affect the system but are not affected by it
3. Resources - factor inputs to do the work
4. Components - set of activities or tasks of the system
5. Management - sets goals, allocates resources and exercises control over components
6. Model of how variables in 1-5 relate to each other

the detailed objectives are identified in the following step, and the detailed model for analysis of the problem is specified in the step after that.

Note

For instance in the case of intercity transportation in California, data about existing demand conditions, existing supply conditions, future demand expectations, and proposed changes to supply would be important inputs. Changes in technology and environmental conditions are important considerations for long-term projects. We would also want to know the certainty of the forecasts, not just a central tendency, and the potential for alternative scenarios which may be widely different.

Define the problem

The second step is to *define the problem* more narrowly, in a sense to *identify needs*.

Note

Rather than an amorphous issue (intercity transportation), we might be interested in a more detailed question, e.g. how to serve existing and future demands between two cities (say metropolitan Los Angeles and San Francisco). The problem might be that demand is expected to grow and outstrip supply.

Formulate goal

The third step is to formulate a goal. For major transportation projects, or projects with intense community interest, this may involve the public. For instance

Note

To serve future passenger demand between Los Angeles and San Francisco, quickly, safely, cleanly, and inexpensively.

The goal will need to be testable, the process below "formulate goal" in the flowchart suggests this process in more detail.

The first aspect is to operationalize the goal. We need to measure the adverbs in the goal (e.g. how do we measure "quickly", "safely", "cleanly", or "inexpensively"). Some are straight-forward. "Quickly" is a measure of travel time or speed. But it needs to account for both the access and egress time, the waiting time, and the travel time, and these may not be weighted the same.

The second step is identifying the decision criteria. Each adverb may have a certain value, but it might be that an alternative has not merely have the most points in one area, but establish at least minimum satisfactory points in all areas. So a very fast mode must meet a specific safety test, and going faster does not necessarily mean it can also be more dangerous (despite what a rational economist might think about trade-offs).

The third is to weight those criteria. E.g. how important is speed vs. safety? This is in many ways a value question, though economics can try to value each of these aspects in monetary form, enabling Evaluation. For instance, many Negative externalities have been monetized, giving a value of time in delay, a value of pollution damages, and a value of life.

Generate alternatives

Examining, evaluating, and recommending alternatives is often the job of professionals, engineers, planners, and economists. Final selection is generally the job of elected or appointed officials for important projects.

There are several sub-problems here, the first is to *generate alternatives*. This may require significant creativity. Within major alternatives, there may be many sub-alternatives, e.g. the main alternative may be mode of travel, the sub-alternatives may be different alignments. For network problems there may be many combinations of alternative alignments. If the analyst is lucky,

these are *separable problems*, that is, the choice of one sub-alignment is independent of the choice of alternative sub-alignments.

1. Algorithms-systematic search over available alternatives
    1. Analytical
    2. Exact numerical
    3. Heuristic numerical
2. Generate alternatives selectively, evaluate subjectively
    1. Fatal flaw analysis
    2. Simple rating schemes
    3. **Delphi exercises**
3. Generate alternatives judgmentally, evaluate scientifically using system model

A critical issue is how many alternatives to consider. In principle, an infinite number of more or less similar alternatives may be generated, not all are practical, and some may be minor variations. In practice a stopping rule to consider a reasonable number of alternatives is used. Major exemplars of the alternatives may be used, with fine-tuning awaiting a later step after the first set of alternatives is analyzed. The process may be iterative, winnowing down alternatives and detailing alternatives as more information is gained throughout the analysis.

Note

Several major alternatives may be suggested, expand highways, expand air travel, or construct new high-speed rail line, along with a no-build alternative.

Abstract into model or framework

"All Models are Wrong, Some Models are Less Wrong than Others"—Anonymous

"All Models are Wrong, Some Models are Useful"—George Box

The term **Model** refers here to a *mathematical representation of a system*, while a **Framework** is a *qualitative organizing principle for analyzing a system*. The terms are sometimes used interchangeably.

Framework Example: Porter's Diamond of Advantage

To illustrate the idea of a framework, consider Porter's *Diamond of Advantage*

Michael Porter proposes four key determinants of competitiveness, which he calls the "Diamond of Advantage," based on cases from around the world:

1. factor conditions, such as a specialized labor pool, specialized infrastructure and sometimes selective disadvantages that drive innovation;
2. home demand, or local customers who push companies to innovate, especially if their tastes or needs anticipate global demand;
3. related and supporting industries, specifically internationally competitive local supplier industries, creating a high quality, supportive business infrastructure, and spurring innovation and spin-off industries; and
4. industry strategy/rivalry, involving both intense local rivalry among area industries that is more motivating than foreign competition and as well as a local "culture" which influences individual industries' attitudes toward innovation and competition.

Model Example: The Four-Step Urban Transportation Planning System

Within the rational planning framework, transportation forecasts have traditionally followed the sequential four-step model or urban transportation planning (UTP) procedure, first implemented on mainframe computers in the 1950s at the Detroit Area Transportation Study and Chicago Area Transportation Study (CATS).

Land use forecasting sets the stage for the process. Typically, forecasts are made for the region as a whole, e.g., of population growth. Such forecasts provide control totals for the local land use analysis. Typically, the region is divided into zones and by trend or regression analysis, the population and employment are determined for each.

The four steps of the classical urban transportation planning system model are:

- **Trip generation** determines the frequency of origins or destinations of trips in each zone by trip purpose, as a function of land uses and household demographics, and other socio-economic factors.
- **Destination choice** matches origins with destinations, often using a gravity model function, equivalent to an entropy maximizing model. Older models include the fratar model.
- **Mode choice** computes the proportion of trips between each origin and destination that use a particular transportation mode. This model is often of the logit form, developed by Nobel Prize winner Daniel McFadden.
- **Route choice** allocates trips between an origin and destination by a particular mode to a route. Often (for highway route assignment) Wardrop's principle of user equilibrium is applied, wherein each traveler chooses the shortest (travel time) path, subject to every other driver doing the same. The difficulty is that travel times are a function of demand, while demand is a function of travel time.

Ascertain performance

This is either an output of the analytical model, or the result of subjective judgment.

Sherden identifies a number of major techniques for technological forecasting that can be used to ascertain expected performance of particular technologies, but that can be used within a technology to ascertain the performance of individual projects. These are listed in the following box:

> "Major techniques for technological forecasting [3]
>
> - Delphi method: a brain-storming session with a panel of experts.
> - Nominal group process: a variant of the Delphi method with a group leader.
> - Case study method: an analysis of analogous developments in other technologies.
> - Trend analysis: the use of statistical analysis to extend past trends into the future.
> - S-curve: a form of trend analysis using an s-shaped curve to extend past trends into the future.
> - Correlation analysis: the projection of development of a new technology past developments in similar technologies.
> - Lead-user analysis: the analysis of leading-edge users of a new technology predict how the technology will develop.
> - Analytic hierarchy process: the projection of a new technology by analyzing a hierarchy of forces influencing its development.
> - Systems dynamics: the use of a detailed model to assess the dynamic relationships among the major forces influencing the development of the technology.
> - Cross-impact analysis: the analysis of potentially interrelated future events that may affect the future development of a technology.
> - Relevance trees: the breakdown of goals for a technology into more detailed goals and then assigning probabilities that the technology will achieve these detail goals.
> - Scenario writing: the development of alternative future views on how the new technology could be used."

Rate alternatives

The performance of each of the alternatives is compared across decision criteria, and weighted depending on the importance of those criteria. The alternative with the highest ranking would be identified, and this information would be brought forward to

decision-makers.

Compute optimal decision

The analyst is generally not the decision maker. The actual influence of the results of the analysis in actual decisions will depend on:

1. Determinacy of evaluation
2. Confidence in the results on the part of the decision maker
3. Consistency of rating among alternatives

Implement alternatives

A decision is made. A project is constructed or a program implemented.

Evaluate outcome

Evaluating outcomes of a project includes comparing outcome against goals, but also against predictions, so that forecasting procedures can be improved. Analysis and implementation experience lead to revisions in systems definition, and may affect the values that underlay that definition. The output from this "last" step in is used as input to earlier steps in subsequent analyses. See e.g. Parthasarathi, Pavithra and David Levinson (2010) Post-Construction Evaluation of Traffic Forecast Accuracy. *Transport Policy*

Relationship to other models

We need a tool to "Identify Needs" and "Evaluate Options". This may be the transportation forecasting model.

Problem PRT: Skyweb Express

The Metropolitan Council of Governments (the region's main transportation planning agency) is examining whether the Twin Cities should build a new Personal Rapid Transit system in downtown Minneapolis, and they have asked you to recommend how it should be analyzed

1. What kind of model should be used. Why?

2. What data should be collected.

Form groups of 3 and take 15 minutes and think about what kinds of models you want to run and what data you want to collect, what questions you would ask, and how it should be collected. Each group should have a note-taker, but all members of the group should be able to present findings to the class.

Thought Questions

- Is the "rational planning" process rational?
- Compare and contrast the rational planning process with the scientific method?

Some Issues with Rational Planning

Nevertheless, some issues remain with the rational planning model:

Problems of incomplete information

- Limited Computational Capacity
- Limited Solution Generating Capacity
- Limited input data
- Cost of Analysis

Problems of incompatible desires

- Conflicting Goals
- Conflicting Evaluation Criteria
- Reliance on Experts (What about the People?)

Alternative Planning Decision Making Paradigms: Are They Irrational?

No one really believes the rational planning process is a good description of most decision making, as it is highly idealized. Alternatives normative and positive paradigms to the rational planning process include:

Several strategies normatively address the problems associated with incomplete information:

- Satisficing
- Decomposition hierarchically into Strategy/Tactics/Operations.

Other strategies describe how organizations and political systems work:

- Organizational Process
- Political Bargaining

Some do both:

- Incrementalism

The paper Montes de Oca, Norah and David Levinson (2006) Network Expansion Decision-making in the Twin Cities. *Journal of the Transportation Research Board: Transportation Research Record* #1981 pp 1-11 describes the actual decision making process for road construction in the Twin Cities.

This page titled 1.2: Fundamentals of Transportation and Decision Making is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

1.3: Evaluation

A **benefit-cost analysis** (BCA)[1] is often required in determining whether a project should be approved and is useful for comparing similar projects. It determines the stream of quantifiable economic benefits and costs that are associated with a project or policy. If the benefits exceed the costs, the project is worth doing; if the benefits fall short of the costs, the project is not. Benefit-cost analysis is appropriate where the technology is known and well understood or a minor change from existing technologies is being performed. BCA is not appropriate when the technology is new and untried because the effects of the technology cannot be easily measured or predicted. However, just because something is new in one place does not necessarily make it new, so benefit-cost analysis would be appropriate, e.g., for a light-rail or commuter rail line in a city without rail, or for any road project, but would not be appropriate (at the time of this writing) for something truly radical like teleportation.

The identification of the costs, and more particularly the benefits, is the chief component of the "art" of Benefit-Cost Analysis. This component of the analysis is different for every project. Furthermore, care should be taken to avoid double counting; especially counting cost savings in both the cost and the benefit columns. However, a number of benefits and costs should be included at a minimum. In transportation these costs should be separated for users, transportation agencies, and the public at large. Consumer benefits are measured by consumer's surplus. It is important to recognize that the demand curve is downward sloping, so there a project may produce benefits both to existing users in terms of a reduction in cost and to new users by making travel worthwhile where previously it was too expensive.

Agency benefits come from profits. But since most agencies are non-profit, they receive no direct profits. Agency construction, operating, maintenance, or demolition costs may be reduced (or increased) by a new project; these cost savings (or increases) can either be considered in the cost column, or the benefit column, but not both.

Society is impacted by transportation project by an increase or reduction of negative and positive externalities. Negative externalities, or social costs, include air and noise pollution and accidents. Accidents can be considered either a social cost or a private cost, or divided into two parts, but cannot be considered in total in both columns.

If there are network externalities (i.e. the benefits to consumers are themselves a function of the level of demand), then consumers' surplus for each different demand level should be computed. Of course this is easier said than done. In practice, positive network externalities are ignored in Benefit Cost Analysis.

Background

Early Beginnings

When Benjamin Franklin was confronted with difficult decisions, he often recorded the pros and cons on two separate columns and attempted to assign weights to them. While not mathematically precise, this "moral or prudential algebra", as he put it, allowed for careful consideration of each "cost" and "benefit" as well as the determination of a course of action that provided the greatest benefit. While Franklin was certainly a proponent of this technique, he was certainly not the first. Western European governments, in particular, had been employing similar methods for the construction of waterway and shipyard improvements.

Ekelund and Hebert (1999) credit the French as pioneers in the development of benefit-cost analyses for government projects. The first formal benefit-cost analysis in France occurred in 1708. Abbe de Saint-Pierre attempted to measure and compare the incremental benefit of road improvements (utility gained through reduced transport costs and increased trade), with the additional construction and maintenance costs. Over the next century, French economists and engineers applied their analysis efforts to canals (Ekelund and Hebert, 1999). During this time, The École Polytechnique had established itself as France's premier educational institution, and in 1837 sought to create a new course in "social arithmetic": "...the execution of public works will in many cases tend to be handled by a system of concessions and private enterprise. Therefore our engineers must henceforth be able to evaluate the utility or inconvenience, whether local or general, or each enterprise; consequently they must have true and precise knowledge of the elements of such investments." (Ekelund and Hebert, 1999, p. 47). The school also wanted to ensure their students were aware of the effects of currencies, loans, insurance, amortization and how they affected the probable benefits and costs to enterprises.

In the 1840s French engineer and economist Jules Dupuit (1844, tr. 1952) published an article "On Measurement of the Utility of

Public Works", where he posited that benefits to society from public projects were not the revenues taken in by the government (Aruna, 1980). Rather the benefits were the difference between the public's willingness to pay and the actual payments the public made (which he theorized would be smaller). This "relative utility" concept was what Alfred Marshall would later rename with the more familiar term, "consumer surplus" (Ekelund and Hebert, 1999).

Vilfredo Pareto (1906) developed what became known as Pareto improvement and Pareto efficiency (optimal) criteria. Simply put, a policy is a Pareto improvement if it provides a benefit to at least one person without making anyone else worse off (Boardman, 1996). A policy is Pareto efficient (optimal) if no one else can be made better off without making someone else worse off. British economists Kaldor and Hicks (Hicks, 1941; Kaldor, 1939) expanded on this idea, stating that a project should proceed if the losers could be compensated in some way. It is important to note that the Kaldor-Hicks criteria states it is sufficient if the winners could potentially compensate the project losers. It does not require that they be compensated.

Benefit-cost Analysis in the United States

Much of the early development of benefit-cost analysis in the United States is rooted in water related infrastructure projects. The US Flood Control Act of 1936 was the first instance of a systematic effort to incorporate benefit-cost analysis to public decision-making. The act stated that the federal government should engage in flood control activities if "the benefits to whomsoever they may accrue [be] in excess of the estimated costs," but did not provide guidance on how to define benefits and costs (Aruna, 1980, Persky, 2001). Early Tennessee Valley Authority (TVA) projects also employed basic forms of benefit-cost analysis (US Army Corp of Engineers, 1999). Due to the lack of clarity in measuring benefits and costs, many of the various public agencies developed a wide variety of criteria. Not long after, attempts were made to set uniform standards.

The U.S. Army Corp of Engineers "Green Book" was created in 1950 to align practice with theory. Government economists used the Kaldor-Hicks criteria as their theoretical foundation for the restructuring of economic analysis. This report was amended and expanded in 1958 under the title of "The Proposed Practices for Economic Analysis of River Basin Projects" (Persky, 2001).

The Bureau of the Budget adopted similar criteria with 1952's Circular A-47 - "Reports and Budget Estimates Relating to Federal Programs and Projects for Conservation, Development, or Use of Water and Related Land Resources".

Modern Benefit-cost Analysis

During the 1960s and 1970s the more modern forms of benefit-cost analysis were developed. Most analyses required evaluation of:

1. The present value of the benefits and costs of the proposed project at the time they occurred
2. The present value of the benefits and costs of alternatives occurring at various points in time (opportunity costs)
3. Determination of risky outcomes (sensitivity analysis)
4. The value of benefits and costs to people with different incomes (distribution effects/equity issues) (Layard and Glaister, 1994)

The Planning Programming Budgeting System (PPBS) - 1965

The Planning Programming Budgeting System (PPBS) developed by the Johnson administration in 1965 was created as a means of identifying and sorting priorities. This grew out of a system Robert McNamara created for the Department of Defense a few years earlier (Gramlich, 1981). The PPBS featured five main elements:

1. A careful specification of basic program objectives in each major area of governmental activity.
2. An attempt to analyze the outputs of each governmental program.
3. An attempt to measure the costs of the program, not for one year but over the next several years ("several" was not explicitly defined).
4. An attempt to compare alternative activities.
5. An attempt to establish common analytic techniques throughout the government.

Office of Management and Budget (OMB) – 1977

Throughout the next few decades, the federal government continued to demand improved benefit-cost analysis with the aim

of encouraging transparency and accountability. Approximately 12 years after the adoption of the PPBS system, the Bureau of the Budget was renamed the Office of Management and Budget (OMB). The OMB formally adopted a system that attempts to incorporate benefit-cost logic into budgetary decisions. This came from the Zero-Based Budgeting system set up by Jimmy Carter when he was governor of Georgia (Gramlich, 1981).

Recent Developments[edit]

Executive Order 12292, issued by President Reagan in 1981, required a regulatory impact analysis (RIA) for every major governmental regulatory initiative over $100 million. The RIA is basically a benefit-cost analysis that identifies how various groups are affected by the policy and attempts to address issues of equity (Boardman, 1996).

According to Robert Dorfman, (Dorfman, 1997) most modern-day benefit-cost analyses suffer from several deficiencies. The first is their attempt "to measure the social value of all the consequences of a governmental policy or undertaking by a sum of dollars and cents". Specifically, Dorfman mentions the inherent difficultly in assigning monetary values to human life, the worth of endangered species, clean air, and noise pollution. The second shortcoming is that many benefit-cost analyses exclude information most useful to decision makers: the distribution of benefits and costs among various segments of the population. Government officials need this sort of information and are often forced to rely on other sources that provide it, namely, self-seeking interest groups. Finally, benefit-cost reports are often written as though the estimates are precise, and the readers are not informed of the range and/or likelihood of error present.

The Clinton Administration sought proposals to address this problem in revising Federal benefit-cost analyses. The proposal required numerical estimates of benefits and costs to be made in the most appropriate unit of measurement, and "specify the ranges of predictions and shall explain the margins of error involved in the quantification methods and in the estimates used" (Dorfman, 1997). Executive Order 12898 formally established the concept of Environmental Justice with regards to the development of new laws and policies, stating they must consider the "fair treatment for people of all races, cultures, and incomes." The order requires each federal agency to identify and address "disproportionately high and adverse human health or environmental effects of its programs, policies and activities on minority and low-income populations."

Probabilistic Benefit-Cost Analysis

In recent years there has been a push for the integration of sensitivity analyses of possible outcomes of public investment projects with open discussions of the merits of assumptions used. This "risk analysis" process has been suggested by Flyvbjerg (2003) in the spirit of encouraging more transparency and public involvement in decision-making.

The Treasury Board of Canada's Benefit-Cost Analysis Guide recognizes that implementation of a project has a probable range of benefits and costs. It posits that the "effective sensitivity" of an outcome to a particular variable is determined by four factors:

- the responsiveness of the Net Present Value (NPV) to changes in the variable;
- the magnitude of the variable's range of plausible values;
- the volatility of the value of the variable (that is, the probability that the value of the variable will move within that range of plausible values); and
- the degree to which the range or volatility of the values of the variable can be controlled.

It is helpful to think of the range of probable outcomes in a graphical sense, as depicted in Figure 1 (probability versus NPV).

Once these probability curves are generated, a comparison of different alternatives can also be performed by plotting each one on the same set of ordinates. Consider for example, a comparison between alternative A and B (Figure 2).

In Figure 2, the probability that any specified positive outcome will be exceeded is always higher for project B than it is for project A. The decision maker should, therefore, always prefer project B over project A. In other cases, an alternative may have a much broader or narrower range of NPVs compared to other alternatives (Figure 3).

Some decision-makers might be attracted by the possibility of a higher return (despite the possibility of greater loss) and therefore might choose project B. Risk-averse decision-makers will be attracted by the possibility of lower loss and will therefore be inclined to choose project A.

Figure 1: Probability-density distribution of net present values approximated by a normal curve. Source: Treasury Board of Canada, Benefit-Cost Analysis Guide, 1998

Figure 2: Probability distribution curves for the NPVs of projects A and B. Source: Treasury Board of Canada, Benefit-Cost Analysis Guide, 1998.

Figure 3: Probability distribution curves for the NPVs of projects A and B, where Project A has a narrower range of possible NPVs. Source: Treasury Board of Canada, Benefit-Cost Analysis Guide, 1998

Discount rate

Both the costs and benefits flowing from an investment are spread over time. While some costs are one-time and borne up front, other benefits or operating costs may be paid at some point in the future, and still others received as a stream of payments collected over a long period of time. Because of inflation, risk, and uncertainty, a dollar received now is worth more than a dollar received at some time in the future. Similarly, a dollar spent today is more onerous than a dollar spent tomorrow. This reflects the concept of time preference that we observe when people pay bills later rather than sooner. The existence of real interest rates reflects this time preference. The appropriate discount rate depends on what other opportunities are available for the capital. If simply putting the money in a government insured bank account earned 10% per year, then at a minimum, no investment earning less than 10% would be worthwhile. In general, projects are undertaken with those with the highest rate of return first, and then so on until the cost of raising capital exceeds the benefit from using that capital. Applying this efficiency argument, no project should be undertaken on cost-benefit grounds if another feasible project is sitting there with a higher rate of return.

Three alternative bases for the setting the government's test discount rate have been proposed:

1. The social rate of time preference recognizes that a dollar's consumption today will be more valued than a dollar's consumption at some future time for, in the latter case, the dollar will be subtracted from a higher income level. The amount of this difference per dollar over a year gives the annual rate. By this method, a project should not be undertaken unless its rate of return exceeds the social rate of time preference.
2. The opportunity cost of capital basis uses the rate of return of private sector investment, a government project should not be undertaken if it earns less than a private sector investment. This is generally higher than social time preference.
3. The cost of funds basis uses the cost of government borrowing, which for various reasons related to government insurance and its ability to print money to back bonds, may not equal exactly the opportunity cost of capital.

Typical estimates of social time preference rates are around 2 to 4 percent while estimates of the social opportunity costs are around 7 to 10 percent.

Generally, for Benefit-Cost studies an acceptable rate of return (the government's test rate) will already have been established. An alternative is to compute the analysis over a range of interest rates, to see to what extent the analysis is sensitive to variations in this factor. In the absence of knowing what this rate is, we can compute the rate of return (internal rate of return) for which the project breaks even, where the net present value is zero. Projects with high internal rates of return are preferred to those with low rates.

Determine a present value

The basic math underlying the idea of determining a present value is explained using a simple compound interest rate problem as the starting point. Suppose the sum of $100 is invested at 7 percent for 2 years. At the end of the first year the initial $100 will have earned $7 interest and the augmented sum ($107) will earn a further 7 percent (or $7.49) in the second year. Thus at the end of 2 years the $100 invested now will be worth $114.49.

The discounting problem is simply the converse of this compound interest problem. Thus, $114.49 receivable in 2 years time, and discounted by 7 per cent, has a present value of $100.

Present values can be calculated by the following equation:

(1)

$$P = \frac{F}{(1+i)^n} \tag{1.3.1}$$

where:

- F = future money sum

- P = present value
- i = discount rate per time period (i.e. years) in decimal form (e.g. 0.07)
- n = number of time periods before the sum is received (or cost paid, e.g. 2 years)

Illustrating our example with equations we have:

$$P = \frac{F}{(1+i)^n} = \frac{114.49}{(1+0.07)^2} = 100.00 \tag{1.3.2}$$

The present value, in year 0, of a stream of equal annual payments of A starting year 1, is given by the reciprocal of the equivalent annual cost. That is, by:

(2)

$$P = A\left[\frac{(1+n)^n - 1}{i(1+i)^n}\right] \tag{1.3.3}$$

where:

- A = Annual Payment

For example: 12 annual payments of $500, starting in year 1, have a present value at the middle of year 0 when discounted at 7% of: $3971

$$P = A\left[\frac{(1+n)^n - 1}{i(1+i)^n}\right] = 500\left[\frac{(1+0.07)^{12} - 1}{0.07(1+0.07)^{12}}\right] = 3971 \tag{1.3.4}$$

The present value, in year 0, of m annual payments of A, starting in year n + 1, can be calculated by combining discount factors for a payment in year n and the factor for the present value of m annual payments. For example: 12 annual mid-year payments of $250 in years 5 to 16 have a present value in year 4 of $1986 when discounted at 7%. Therefore in year 0, 4 years earlier, they have a present value of $1515.

$$P_{Y=4} = A\left[\frac{(1+n)^n - 1}{i(1+i)^n}\right] = 250\left[\frac{(1+0.07)^{12} - 1}{0.07(1+0.07)^{12}}\right] = 1986 \tag{1.3.5}$$

$$P_{Y=0} = \frac{F}{(1+i)^n} = \frac{P_{Y=4}}{(1+i)^n} = \frac{1986}{(1+0.07)^4} = 1515 \tag{1.3.6}$$

Evaluation criterion

Three equivalent conditions can tell us if a project is worthwhile

1. The discounted present value of the benefits exceeds the discounted present value of the costs
2. The present value of the net benefit must be positive.
3. The ratio of the present value of the benefits to the present value of the costs must be greater than one.

However, that is not the entire story. More than one project may have a positive net benefit. From the set of mutually exclusive projects, the one selected should have the highest net present value. We might note that if there are insufficient funds to carry out all mutually exclusive projects with a positive net present value, then the discount used in computing present values does not reflect the true cost of capital. Rather it is too low.

There are problems with using the internal rate of return or the benefit/cost ratio methods for project selection, though they provide useful information. The ratio of benefits to costs depends on how particular items (for instance, cost savings) are ascribed to either the benefit or cost column. While this does not affect net present value, it will change the ratio of benefits to costs (though it cannot move a project from a ratio of greater than one to less than one).

Examples

Example 1: Benefit Cost Application

This problem, adapted from Watkins (1996), illustrates how a Benefit Cost Analysis might be applied to a project such as a highway widening. The improvement of the highway saves travel time and increases safety (by bringing the road to modern standards). But there will almost certainly be more total traffic than was carried by the old highway. This example excludes external costs and benefits, though their addition is a straightforward extension. The data for the "No Expansion" can be collected from off-the-shelf sources. However the "Expansion" column's data requires the use of forecasting and modeling. Assume there are 250 weekdays (excluding holidays) each year and four rush hours per weekday.

Table 1: Data

		No Expansion	Expansion
Peak			
Passenger Trips	(per hour)	18,000	24,000
Trip Time	(minutes)	50	30
Off-peak			
Passenger Trips	(per hour)	9,000	10,000
Trip Time	(minutes)	35	25
Traffic Fatalaties	(per year)	2	1

Note: the operating cost for a vehicle is unaffected by the project, and is $4.

Table 2: Model Parameters

Peak Value of Time	($/minute)	$0.15
Off-Peak Value of Time	($/minute)	$0.10
Value of Life	($/life)	$3,000,000

What is the benefit-cost relationship?

Solution

Figure 1: Change in Consumers' Surplus

A 50 minute trip at $0.15/minute is $7.50, while a 30 minute trip is only $4.50. So for existing users, the expansion saves $3.00/trip. Similarly in the off-peak, the cost of the trip drops from $3.50 to $2.50, saving $1.00/trip.

Consumers' surplus increases both for the trips which would have been taken without the project and for the trips which are stimulated by the project (so-called "induced demand"), as illustrated above in Figure 1. Our analysis is divided into Old and New Trips, the benefits are given in Table 3.

Table 3: Hourly Benefits

TYPE	Old trips	New Trips	Total
Peak	$54,000	$9000	$63,000
Off-Peak	$9,000	$500	$9,500

Note: Old Trips: For trips which would have been taken anyway the benefit of the project equals the value of the time saved multiplied by the number of trips. New Trips: The project lowers the cost of a trip and public responds by increasing the number of trips taken. The benefit to new trips is equal to one half of the value of the time saved multiplied by the increase in the number of trips. There are 1000 peak hours per year. With 8760 hours per year, we get 7760 offpeak hours per year. These numbers permit the calculation of annual benefits (shown in Table 4).

Table 4: Annual Travel Time Benefits

TYPE	Old trips	New Trips	Total
Peak	$54,000,000	$9,000,000	$63,000,000
Off-Peak	$69,840,000	$3,880,000	$73,720,000
Total	$123,840,000	$12,880,000	$136,720,000

The safety benefits of the project are the product of the number of lives saved multiplied by the value of life. Typical values of life are on the order of $3,000,000 in US transportation analyses. We need to value life to determine how to trade off between safety investments and other investments. While your life is invaluable to you (that is, I could not pay you enough to allow me to kill you), you don't act that way when considering chance of death rather than certainty. You take risks that have small probabilities of very bad consequences. You do not invest all of your resources in reducing risk, and neither does society. If the project is expected to save one life per year, it has a safety benefit of $3,000,000. In a more complete analysis, we would need to include safety benefits from non-fatal accidents.

The annual benefits of the project are given in Table 5. We assume that this level of benefits continues at a constant rate over the life of the project.

Table 5: Total Annual Benefits

Type of Benefit	Value of Benefits Per Year
Time Saving	$136,720,000
Reduced Risk	$3,000,000
Total	$139,720,000

Costs

Highway costs consist of right-of-way, construction, and maintenance. Right-of-way includes the cost of the land and buildings that must be acquired prior to construction. It does not consider the opportunity cost of the right-of-way serving a different purpose. Let the cost of right-of-way be $100 million, which must be paid before construction starts. In principle, part of the right-of-way cost can be recouped if the highway is not rebuilt in place (for instance, a new parallel route is constructed and the old highway can be sold for development). Assume that all of the right-of-way cost is recoverable at the end of the thirty-year lifetime of the project. The $1 billion construction cost is spread uniformly over the first four-years. Maintenance costs $2 million per year once the highway is completed.

The schedule of benefits and costs for the project is given in Table 6.

Table 6: Schedule of Benefits And Costs ($ millions)

Time (year)	Benefits	Right-of-way costs	Construction Costs	Maintenance costs
0	0	100	0	0
1-4	0	0	250	0
5-29	139,72	0	0	2
30	139.72	-100	0	2

Conversion to Present Value

The benefits and costs are in constant value dollars. Assume the real interest rate (excluding inflation) is 2%. The following equations provide the present value of the streams of benefits and costs.

To compute the Present Value of Benefits in Year 5, we apply equation (2) from above.

$$P = A\left[\frac{(1+n)^n - 1}{i(1+i)^n}\right] = 139.72\left[\frac{(1+0.02)^{26} - 1}{0.02(1+0.02)^{26}}\right] = 2811.31 \tag{1.3.7}$$

To convert that Year 5 value to a Year 1 value, we apply equation (1)

$$P = \frac{F}{(1+i)^n} = \frac{2811.31}{(1+0.02)^4} = 2597.21 \tag{1.3.8}$$

The present value of right-of-way costs is computed as today's right of way cost ($100 M) minus the present value of the recovery of those costs in Year 30, computed with equation (1):

$$P = \frac{F}{(1+i)^n} = \frac{100}{(1+0.02)^{30}} = 55.21 \tag{1.3.9}$$

$$100 - 55.21 = 44.79 \tag{1.3.10}$$

The present value of the construction costs is computed as the stream of $250M outlays over four years is computed with equation (2):

$$P = A\left[\frac{(1+n)^n - 1}{i(1+i)^n}\right] = 250\left[\frac{(1+0.02)^4 - 1}{0.02(1+0.02)^4}\right] = 951.93 \tag{1.3.11}$$

Maintenance Costs are similar to benefits, in that they fall in the same time periods. They are computed the same way, as follows: To compute the Present Value of $2M in Maintenance Costs in Year 5, we apply equation (2) from above.

$$P = A\left[\frac{(1+n)^n - 1}{i(1+i)^n}\right] = 2\left[\frac{(1+0.02)^{26} - 1}{0.02(1+0.02)^{26}}\right] = 40.24 \tag{1.3.12}$$

To convert that Year 5 value to a Year 1 value, we apply equation (1)

$$P = \frac{F}{(1+i)^n} = \frac{40.24}{(1+0.02)^4} = 37.18 \tag{1.3.13}$$

As Table 7 shows, the benefit/cost ratio of 2.5 and the positive net present value of $1563.31 million indicate that the project is worthwhile under these assumptions (value of time, value of life, discount rate, life of the road). Under a different set of assumptions, (e.g. a higher discount rate), the outcome may differ.

Table 7: Present Value of Benefits and Costs ($ millions)

	Present Value
Benefits	2,597.21
Costs	
Right-of-Way	44.79
Construction	951.93
Maintenance	37.18
Costs SubTotal	1,033.90
Net Benefit(B-C)	1,563.31
Benefit/Cost Ratio	2.5

Thought Questions

Decision Criteria

Which is a more appropriate decision criteria: Benefit/Cost or Benefit - Cost? Why?

Is it only money that matters?

Problem

Is money the only thing that matters in Benefit-Cost Analysis? Is "converted" money the only thing that matters? For example, the value of human life in dollars?

Solution

Absolutely not. A lot of benefits and costs can be converted to monetary value, but not all. For example, you can put a price on human safety, but how can you put a price on, say, aesthetics—something that everyone agrees is beneficial. What else can you think of?

Answer available at the end of this section.

Can small units of time be given the same value of time as larger units of time?

In other words, do 60 improvements each saving a traveler 1 minute equal 1 improvement saving a traveler 60 minutes? Similarly, does 1 improvement saving a 1000 travelers 1 minute equal the value of time of a single traveler of 1000 minutes. These are different problems, one is intra-traveler and one is inter-traveler, but related.

Several issues arise.

A. Is value of time linear or non-linear? To this we must conclude the value of time is surely non-linear. I am much more agitated waiting 3 minutes at a red light than 2, and I begin to suspect the light is broken. Studies of ramp meters show a similar phenomena.[2]

B. How do we apply this in a benefit-cost analysis? If we break one project into 60 smaller projects, each with a smaller value of travel time saved, and then we added the gains, we would get a different result than the what obtains with a single large project. For analytical convenience, we would like our analyses to be additive, not sub-additive, otherwise arbitrarily dividing the project changes the result. In particular many smaller projects will produce an undercount that is quite significant, and result in a much lower benefit than if the projects were bundled.

As a practical matter, every Benefit/Cost Analysis assumes a single value of time, rather than assuming non-linear value of time. This also helps avoiding biasing public investments towards areas with people who have a high value of time (the rich)

On the other hand, mode choice analyses do however weight different components of travel time differently, especially transit time (i.e. in-vehicle time is less onerous than waiting time). The implicit value of time for travelers does depend on the type of time (though generally not the amount of time). Using the log-sum of the mode choice model as a measure of benefit would implicitly account for this.

Are sunk costs sunk, is salvage value salvageable? A paradox in engineering economics analysis.

Salvage value is defined as "The estimated value of an asset at the end of its useful life." Sunk cost is defined as "Cost already incurred which cannot be recovered regardless of future events."

It is often said in economics that "sunk costs are sunk", meaning they should not be considered a cost in economic analysis, because the money has already been spent.

Now consider two cases

In **case 1**, we have a road project that costs $10.00 today, and at the end of 10 years has some economic value remaining, let's say a salvage value of $5.00, which when discounted back to the present is $1.93 (at 10% interest). This value is the residual value of the road. Thus, the total present cost of the project $10.00 - $1.93 = $8.07. Clearly the road cannot be moved. However, its presence makes it easier to build future roads ... the land has been acquired and graded, some useful material for aggregate is on-site perhaps, and can be thought of as the amount that it reduces the cost of future generations to build the road. Alternatively, the land could be sold for development if the road is no longer needed, or turned into a park.

Assume the present value of the benefit of the road is $10.00. The benefit/cost ratio is $10.00 over $8.07 or 1.23. If we treat the salvage value as a benefit rather than cost, the benefit is $10.00 + $1.93 = $11.93 and the cost is $10, and the B/C is 1.193.

In 10 years time, the community decides to replace the old worn out road with a new road. This is a new project. The salvage value from the previous project is now the sunk cost of the current project (after all the road is there and could not be moved, and so does not cost the current project anything to exploit). So the cost of the project in 10 years time would be $10.00 - $5.00 = $5.00. Discounting that to the present is $1.93.

The benefit in 10 years time is also $10.00, but the cost in 10 years time was $5.00, and the benefit/cost ratio they perceive is $10.00/$5.00 = 2.00

Aggregating the two projects

- the benefits are $10 + $3.86 = $13.86
- the costs are $8.07 + $1.93 = $10.00
- the collective benefit/cost ratio is 1.386
- the NPV is benefits - costs = $3.86

One might argue the salvage value is a benefit, rather than a cost reduction. In that case

- the benefits are $10.00 + $1.93 + $3.86 = $15.79
- the costs are $10.00 + $1.93 = $11.93
- the collective benefit/cost ratio is 1.32
- the NPV remains $3.86

Case 2 is an identical road, but now the community has a 20 year time horizon to start. The initial cost is $10, and the cost in 10 years time is $5.00 (discounted to $1.93). The benefits are $10 now and $10 in 10 years time (discounted to $3.86). There is no salvage value at the end of the first period, nor sunk costs at the beginning of the second period. What is the benefit cost ratio?

- the costs are $11.93
- the benefits are still $13.86
- the benefit/cost ratio is 1.16
- the NPV is $1.93.

If you are the community, which will you invest in? Case 1 has an initial B/C of 1.23 (or 1.193), Case 2 has a B/C of 1.16. But the real benefits and real costs of the roads are identical.

The salvage value in this example is, like so much in economics (think Pareto optimality), an accounting fiction. In this case no transaction takes place to realize that salvage value. On the other hand, excluding the salvage value over-estimates the net cost of the project, as it ignores potential future uses of the project.

Time horizons on projects must be comparable to correctly assess relative B/C ratio, yet not all projects do have the same benefit/cost ratio.

Software Tools for Impact Analysis

The majority of economic impact studies for highway capacity projects are undertaken using conventional methods. These methods tend to focus on the direct user impacts of individual projects in terms of travel costs and outcomes, and compare sums of quantifiable, discounted benefits and costs. Inputs to benefit-cost analyses can typically be obtained from readily available data sources or model outputs (such as construction and maintenance costs, and before and after estimates of travel demand, by vehicle class, along with associated travel times). Valuation of changes in external, somewhat intangible costs of travel (e.g., air pollution and crash injury) can usually be accommodated by using *shadow price* estimates, such as obtained from FHWA-suggested values, based on recent empirical studies.

The primary benefits included in such studies are those related to reductions in user cost, such as travel time savings and vehicle operating costs (e.g. fuel costs, vehicle depreciation, etc.). Additional benefits may stem from reductions in crash rates, vehicle emissions, noise, and other costs associated with vehicle travel. Project costs are typically confined to expenditures on capital investment, along with ongoing operations and maintenance costs.

A number of economic analysis tools have been developed under the auspices of the United States Federal Highway Administration (FHWA) permitting different forms of benefit-cost analysis for different types of projects, at different levels of evaluation. Several of these tools are prevalent in past impact analyses, and are described here. However, none identifies the effects of infrastructure on the economy and development.

MicroBENCOST

MicroBENCOST is a sketch planning tool for estimating basic benefits and costs of a range of highway improvement projects, including capacity addition projects. In each type of project, attention is focused on corridor traffic conditions and their resulting impact on motorist costs with and without a proposed improvement. This type of approach may be appropriate for situations where projects have relatively isolated impacts and do not require regional modeling.

SPASM

The Sketch Planning Analysis Spreadsheet Model (SPASM) is a benefit-cost tool designed for screening level analysis. It outputs estimates of project costs, cost-effectiveness, benefits, and energy and air quality impacts. SPASM is designed to allow for comparison among multiple modes and non-modal alternatives, such as travel demand management scenarios. The model is comprised of three modules (worksheets) relating to public agency costs, characteristics of facilities and trips, and a travel demand component. Induced traffic is dealt with through the use of elasticity-based methods, where an elasticity of vehicle-miles of travel (VMT) with respect to travel time is defined and applied. Vehicle emissions are estimated based on calcuations of VMT, trip length and speeds, and assumed shares of travel occuring in cold start, hot start, and hot stabilized conditions. Analysis is confined to a corridor level, with all trips having the same origin, destination and length. This feature is appropriate for analysis of linear transportation corridors, but also greatly limits the ability to deal with traffic drawn to or diverted from outside the corridor. DeCorla-Souza et al. (1996) describe the model and its application to a freeway corridor in Salt Lake City, Utah.

STEAM

The Surface Transportation Efficiency Analysis Model (STEAM) is a planning-level extension of the SPASM model, designed for a fuller evaluation of cross-modal and demand management policies. STEAM was designed to overcome the most important limitations of its predecessor, namely the assumption of average trip lengths within a single corridor and the inability to analyze systemwide effects. The enhanced modeling capabilities of STEAM feature greater compatibility with existing four-step travel demand models, including a trip table module that is used to calculate user benefits and emissions estimates based on changes in network conditions and travel behavior. Also, the package features a risk analysis component to its evaluation summary module, which calculates the likelihood of various outcomes such as benefit-cost ratios. An overview of STEAM and a hypothetical application are given by DeCorla-Souza et al. (1998).

SMITE

The Spreadsheet Model for Induced Travel Estimation (SMITE) is a sketch planning application that was designed for inclusion with STEAM in order to account for the effects of induced travel in traffic forecasting. SMITE's design as a simple spreadsheet application allows it to be used in cases where a conventional, four-step travel demand model is unavailable or cannot account for

induced travel effects in its structure. SMITE applies elasticity measures that describe the response in demand (VMT) to changes in travel time and the response in supply (travel time) to changes in demand levels.

SCRITS

As a practical matter, highway corridor improvements involving intelligent transportation systems (ITS) applications to smooth traffic flow can be considered capacity enhancements, at least in the short term. The FHWA's SCRITS (SCReening for ITS) is a sketch planning tool that offers rough estimates of ITS benefits, for screening-level analysis. SCRITS utilizes aggregate relationships between average weekday traffic levels and capacity to estimate travel speed impacts and vehicle-hours of travel (VHT). Like many other FHWA sketch planning tools, it is organized in spreadsheet format and can be used in situations where more sophisticated modeling systems are unavailable or insufficient.

HERS

In addition to helping states plan and manage their highway systems, the FHWA's Highway Economic Requirements System for states (HERS-ST) offers a model for economic impacts evaluation. In one case, Luskin (2005) use HERS-ST to conclude that Texas is under-invested in highways – particularly urban systems and lower-order functional classes – by 50 percent. Combining economic priniciples with engineering criteria, HERS evaluates competing projects via benefit-cost ratios. Recognizing user benefits, emissions levels, and construction and maintenance costs, HERS operates within a GIS environment and will be evaluated under this project, for discussion in project deliverables. Well established software like HERS offer states and regions an oportunity to readily pursue standardized economic impact evaluations on all projects, a key advantage for many users, as well as the greater community.

Summary of Software Tools

Many analytical tools, like those described above, are favored due to their relative ease of use and employment of readily available or easily acquired data. However, several characteristics limit their effectiveness in evaluating the effects of new highway capacity. First, they are almost always insufficient to describe the full range of impacts of new highway capacity. Such methods deliberately reduce economic analysis to the most important components, resorting to several simplifying assumptions. If a project adds capacity to a particularly important link in the transportation network, its effects on travel patterns may be felt outside the immediate area. Also, the effects of induced travel, in terms of either route switching or longer trips, may not be accounted for in travel models based on a static, equilibrum assignment of traffic. In the longer term, added highway capacity may lead to the spatial reorganization of activities as a result of changes in regional accessibility. These types of changes cannot typically be accounted for in analysis methods.

Second, there is the general criticism of methods based on benefit-cost analysis that they cannot account for all possible impacts of a project. Benefit-cost methods deliberately reduce economic analysis to the most important components and often must make simplifying assumptions. The project-based methods described here generally do not describe the economic effects of a project on different user or non-user groups. Winners and losers from a new capacity project cannot be effectively identified and differentiated.

Third, a significant amount of uncertainty and risk is involved in the employment of project-based methods. Methods that use benefit-cost techniques to calculate B/C ratios, rates of return, and/or net present values are often sensitive to certain assumptions and inputs. With transportation infrastructure projects, the choice of discount rate is often critical, due to the long life of projects and large, up-front costs. Also, the presumed value of travel time savings is often pivotal, since it typically reflects the majority of project benefits. Valuations of travel time savings vary dramatically across the traveler population, as a function of trip purpose, traveler wage, household income, and time of day. It is useful to test several plausible values.

Assessment procedures in the UK and other parts of Europe have moved towards a multi-criteria approach, where economic development is only one of several appraisal criteria. Environmental, equity, safety, and the overall integration with other policy sectors are examined in a transparent framework for decision makers. In the UK, the Guidance on the Methodologies for Multi-Modal Studies (2000) provides such a framework. These procedures require a clear definition of project goals and objectives, so that actual effects can be tied to project objectives, as part of the assessment procedure. This is critical for understanding induced travel effects. Noland (2007) has argued that this implies that comprehensive economic assessment, including estimation of land valuation effects, is the only way to fully assess the potential beneficial impacts of projects.

Sample Problems

Problem 1:

A new transportation project is proposed to the city. This project is a form of "guide wire", where cars can hook to these moving, below-ground wires and be transported for free around town. This project is proven to reduce gas consumption by $5 million in its completion year. The city's preferred contractor says that it will take 10 years to build the thing and cost $500,000 a year, which is "perfect" because costs would add up to the benefit. Knowing that inflation is 3 percent, as an expert evaluator, is this a wise decision?

Answer available at the end of this section.

Problem 2:

A new Southstar rail line is proposed. This project is expected to reduce travel time for 2,000 commuters by 30 minutes per day, in its completion year. The line only operates on weekdays (Monday-Friday). It will take 2 years to build and cost $320,000,000 total (Net Present Value). If the interest rate is 3 percent, above what value of time must average value of time for SouthStar Passengers be in order for the benefit/cost ratio to exceed 1.

- Assume a 30 year lifespan. The interest rate is annual*

Answer available at the end of this section.

Key Terms

- Benefit-Cost Analysis
- Profits
- Costs
- Discount Rate
- Present Value
- Future Value

External Exercises

Use the SAND software at the STREET website to learn how to evaluate network performance given a changing network scenario.

Answers

Is it only money that matters?

Problem 1:

Convert everything to Present Value and see just how great a deal this is.

For the present value of the $5,000,000 benefit (gas reduction):

$$P = \frac{F}{(1+i)^n} = \frac{\$5,000,000}{(1+0.03)^{10}} = \$3,720,469.57 \tag{1.3.14}$$

The benefit in present-day value is $3,720,469.57.

For the money (cost) being sent to the contractor, a payment of $470,000 per year, the present value would be:

$$P = A\left[\frac{(1+n)^n - 1}{i(1+i)^n}\right] = \$470,000\left[\frac{(1+0.03)^{10} - 1}{0.03(1+0.03)^{10}}\right] = \$4,009,195.33 \tag{1.3.15}$$

The cost in present-day value is $4,009,195.33. Therefore, this detail, while shiny in appearance at first, is NOT a wise decision, since costs exceed benefits.

Problem 2:

Benefits considered are only Travel Time Savings.

Travel Time Savings = 2,000 Commuters x 0.5 Hours/day (30 minutes/day)

Travel Time Savings = 1000 Commuters-Hours/day x 5 days/week 52 weeks/year

Travel Time Savings = 260,000 Commuters-Hours/year

Travel Time Savings start after two years (given in the problem statement). Other assumptions are: Present Year is 0, no growth (constant commuters ev- ery year, and thus constant savings), and constant Value of Travel Time (VOT). Therefore, Benefits must be discounted to Present value for each year and added for a total during the lifespan of the project considered (30 years).

Adding up all Present Value of Travel Time

Present Value of Travel Time =

$$(VOT)(260,000)(\frac{1}{[1+0.03]^2}+\frac{1}{[1+0.03]^3}+...+\frac{1}{[1+0.03]^{29}}+\frac{1}{[1+0.03]^{30}}) \tag{1.3.16}$$

You can sum it up in an excel spreadsheet or recognize that this is a geometric series.

The sum is 18.63 inside the parentheses.

Total Present Value of Travel Time = (4,843,687.57)(VOT)

Costs are given by Total Present Value of $320,000,000.

Benefits/Costs = 1

Thus

$$\frac{(4,843,687.57)(VOT)}{320,000,000}=1 \tag{1.3.17}$$

VOT = USD$66/Hr.

The VOT must be at least 66 US dollars.

This page titled 1.3: Evaluation is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

2: PLANNING MODELS

2.1: Modeling
2.2: Data
2.3: Networks
2.4: Modes

This page titled 2: Planning Models is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

2.1: Modeling

All forecasts are wrong; some forecasts are more wrong than others. - anonymous

Modeling is a means for representing reality in an abstracted way. Your mental models are your **world view**: your outlook on life, and the world. The world view is your *internal model of how the world works*; it is employed every time you make a prediction: what do you expect, what is a surprise. The expression “Where you stand depends on where you sit” epitomizes this idea. Your world view is shaped by your experience and your position.

When modeling, the issue of **Point of View** should be considered. It must be clear who (and what) the results are for. If you are modeling for personal pleasure, it will naturally reflect your own worldview, but if you are working for an employer or client, their point of view must also be considered, if the inputs or results deviate significantly from their worldview, they may adapt their worldview, but more likely will dismiss the model.

Modeling can be conducted both for subjective **advocacy** and for objective **analysis**. The same methods may be employed in either, and the ethical modeler will produce the same results in either case, but they may be used differently.

Why Model?

There are a variety of reasons to model. Modeling helps

- gain insight into complex situations by understanding simpler situations resembling them
- optimize the use of resources in building or maintaining systems
- operate system, particularly by testing alternative operational scenarios
- educate and provide experience for model-builders
- provide a platform for testing contending ideas and use in negotiations.

Particular applications in transportation include:

- Forecasting traffic
- Testing scenarios (alternative land uses, networks, policies)
- Planning projects/corridor studies
- Regulating land use: Growth management/public facility adequacy
- Managing complexity, when eyeballs are insufficient, (different people have different intuitions)
- Understanding travel behavior
- Influencing decisions
- Estimating traffic in the absence of data

Developing Models

As an engineer, economist, or planner you may be given a model to use. But that model was not spontaneously generated, it came from other engineers, economists, or planners who undertook a series of steps to translate raw data into a tool that could be used to generate useful information.

The first step, specification, tells us what depends on what (the number of trips leaving a zone may depend on the number of households). The estimation step tells us how strong these relationships mathematically are, (each household generates two trips in the peak hour). Implementation takes those relationships and puts them into a computer program. Calibration looks at the output of the computer program and compares it to other data, if they don't match exactly, adjustments to the models may be made. Validation compares the results to data at another point in time. Finally the model is applied to look at a project (e.g. how much traffic will use Granary Road after Washington Avenue is closed to traffic).

- Specification:
  - $$y = f(X) \tag{2.1.1}$$
- Estimation:

- $$y = mX + b; m = 1, b = 2 \tag{2.1.2}$$
- Implementation
  - If Z > W, Then
$$y = mX + b \tag{2.1.3}$$
- Calibration
  - $$y_{predicted} + k = y_{observed} \tag{2.1.4}$$
- Validation
  - $$y_{predicted1990} + k = y_{observed1990} \tag{2.1.5}$$
- Application

Specification

When building a model system, numerous decisions must be made. These are discussed below:

Types of Models

There are numerous types of models, a short list is below. Each has different applicability, multiple methods may be used in pursuit of the same question, sometimes they are complementary, and sometimes competitive techniques.

- Network analysis
- Linear Programming
- Nonlinear Programming
- Simulation
- Deterministic queuing
- Probabilistic queuing
- Regression
- Neural Nets
- Genetic Algorithm
- Cost/ Benefit Analysis
- Life-cycle costing
- System Dynamics
- Control Theory
- Difference Equations
- Differential Equations
- Probabilistic Risk Assessment
- Supply/Demand Equilibrium
- Game Theory
- Statistical Decision Theory
- Markov Models
- Cellular Automata
- Etc.

Model Trade-offs

Building a model requires trading-off time and resource constraints. One could always be more detailed, more accurate, or more comprehensive if resources were not constrained. However, the following must also be considered.

- Money,
- Data,
- Computation,
- Labor,
- Ease of Use,
- Convincing (e.g. Graphic Displays),
- Extendable,

- Evidence of Model Benefits,
- Measuring Model Success

Organization of Model System

- Hierarchy of Models
- Centralized vs. Decentralized (Optimization (Global) vs. Agent, Local Optimization)

Time

- Time Frame
- Static vs. Dynamic
- Real Time vs. Offline
- Short Term vs. Long Term (Partial vs. General Equilibrium)
- Proactive vs. Reactive (Predictive vs. Responsive)

Space

- Scale/Detail
- Spatial Extent
- Boundaries (Boundary Effects)
- Macroscopic vs. Microscopic (Zones, Flows vs. Individuals, Vehicles)

Process

- Stochastic vs. Deterministic
- Linear vs. Nonlinear
- Continuous vs. Discrete
- Numerical Simulation vs. Closed Form Solution
- Equilibrium vs. Disequilibrium

Type

- Behavioral vs. Aggregate Model
- Physical vs. Mathematical Models

Solution Techniques

When solving the model, the system as a whole must be understood. Several questions arise:

- Does the solution exist?
- Is the solution unique?
- Is the solution feasible?

Solution techniques often trade-off accuracy vs. speed. Some solution techniques may only guarantee a local optima, while others (such as brute force techniques) can guarantee a global optimum, but may be much slower.

“Four-Step” Urban Transportation Planning Models

We want to answer a number of related questions (who, what, where, when, why, and how):

- Who is traveling or what is being shipped?
- Where are the origin and destination of those trips, shipments?
- When do those trips begin and end (how long do they take, how far away are the trip ends)?

- Why are the trips being made, what is their purpose?
- How are the trips getting there, what routes or modes are they taking?

If we know the answers to those questions, we want to know what are the answers to those questions a function of?

- Cost: Money, Time spent on the trip,
- Cost: Money and Time of alternatives.
- Benefit (utility) of trip (e.g. the activity at the destination)
- Benefit of alternatives

The reason for this is to understand what will happen under various circumstances:

- How much “induced demand” will be generated if a roadway is expanded?
- How many passengers will be lost if bus services are cut back?
- How many people will be “priced off” if tolls are implemented?
- How much traffic will a new development generate?
- How much demand will I lose if I raise shipping costs to my customers?

In short, for urban passenger travel, we are trying to predict the number of trips by:

- Origin Activity,
- Destination Activity,
- Origin Zone,
- Destination Zone,
- Mode,
- Time of Day, and
- Route.

This is clearly a multidimensional problem.

In practice, the mechanism to do this is to employ a "four-step" urban transportation planning model, each step will be detailed in subsequent modules. These steps are executed in turn, though there may be feedback between steps:

- Trip Generation - How many trips T_i or T_j are entering or leaving zone i or j
- Trip Distribution or Destination Choice - How many trips T_{ij} are going from zone i to zone j
- Mode Choice - How many trips T_{ijm} from i to j are using mode m
- Route Choice - Which links are trips T_{ijmr} from i to j by mode m using route r

Thought Questions

- Is past behavior reflective of future behavior?
- Can the future be predicted?
- Is the future independent of decisions, or are prophecies self-fulfilling?
- How do we know if forecasts were successful?
- Against what standard are they to be judged?
- What values are embedded in the planning process?
- What happens when values change?

Additional Problems

Homework

1. Explain the four-step transportation planning model?

2. What are the outputs of the planning model and what they are used for?

3. What are the strength and weaknesses of this type of the forecasting procedure?

4. List 4 applications of transportation planning models.

5. List 4 typical data sources used in estimating, calibrating, and validating transportation planning models.

6. What is the rational planning model, and how does the four-step transportation planning model relate to it?

7. Identify 5 ways in which the conventional 4-step transportation planning model is imperfect.

Additional Problems

1. List five inputs to transportation planning models – are they measured or forecast?
2. Why is modeling important, (give reasons, purposes of modeling)?
3. What would be a possible complication or problem that would arise if in the planning of a project no forecasting was used?
4. What are the steps in the rational planning model?
5. Why is the rational planning model not always used (examples when it is not used)?
6. What are some of the factors preventing accurate forecasting?
7. List the steps in the four-step urban transportation planning model?
8. Explain the advantages and disadvantages of public vs. privately owned transportation?
9. What types of matrices are there in the transportation model, how do they differ?
10. How does the effect of population increase effect transportation planning?
11. What are the step in the generic modeling process?
12. Are all transportation modes publicly owned? Give examples of public and private.
13. When would it be reasonable to use the rational planning model?
14. What significance does the equilibrium point in transportation economics have in design?
15. What makes forecasting with one method better than another, and how are differences resolved?
16. How do measured inputs affect the forecasting inputs?
17. What is the idealized model to plan for a transportation project?
18. On what type of economic model does transportation depend?
19. Who, other than planners and engineers, should be involved in defining the needs and objectives for future projects?
20. Can we steer development and growth trends through transportation planning decision?
21. Give three examples of intermodalism?
22. Are actual transportation routes planned using the Rational Planning model?
23. List at least 10 things that must be considered before undertaking a major transit project?
24. How is the rational planning model used?
25. Why do we combine vector matrices to a full matrix?
26. Is household income measured or forecast?
27. What statistical methods play a role in transportation engineering/planning?
28. Who measures transit use or travel times?
29. What is a zone centroid?
30. What is forecasting?
31. How are various transportation needs provided?
32. What is considered a carrier?
33. How does vehicle occupancy influence our design process?
34. Is rational planning process more likely to be used on major projects or small ones?
35. What are the axes labels on the supply / demand curve?
36. Explain why all forecasts are wrong?
37. How does supply and demand work? How do they relate?
38. How does intermodealism affect forecasting?
39. What is the difference between modalism and intermodalism?

Key Terms

- Rational Planning
- Transportation planning model
- Matrix, Full Matrix, Vector Matrix, Scalar Matrix
- Trip table
- Travel time matrix
- Origin, Destination
- Purpose
- Network
- Zone (Traffic Analysis Zone or Transportation Analysis Zone, or TAZ)
- External station or external zone
- Centroid
- Node
- Link
- Turn
- Route
- Path
- Mode

Video

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

This page titled 2.1: Modeling is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

2.2: Data

There are a variety of types of transportation data used in analysis. Some are listed below:

- Infrastructure Status
- Traffic Counts
- Travel Behavior Inventory
- Land Use Inventory
- Truck/Freight Demand
- External/Internal Demand (by Vehicle Type)
- Special Generators

Revealed Preference

Household travel surveys which ask people what they actually did are a type of *Revealed Preference* survey data that have been obtained by direct observation of the choice that individuals make with respect to travel behavior. *Travel Cost Analysis* uses the prices of market goods to evaluate the value of goods that are not traded in the market.

Hedonic Pricing uses the market price of a traded good and measures of its component attributes to calculate value. There are other methods to attain Revealed Preference information, but surveys are the most common in travel behavior.

Travel Behavior Inventory

While the Cleveland Regional Area Traffic Study in 1927 was the first metropolitan planning attempt sponsored by the US federal government, the lack of comprehensive survey methods and standards at that time precluded the systematic collection of information such as travel time, origin and destination, and traffic counts. The first US travel surveys appeared in urban areas after the Federal-Aid Highway Act of 1944 permitted the spending of federal funds on urban highways.[1] A new home- interview origin-destination survey method was developed in which households were asked about the number of trips, purpose, mode choice, origin and destination of the trips conducted on a daily basis. In 1944, the US Bureau of Public Roads printed the Manual of Procedures for Home Interview Traffic Studies.[2] This new procedure was first implemented in several small to mid-size areas. Highway engineers and urban planners made use of the new data collected after the 1944 Highway Act extended federally sponsored planning to travel surveys as well as traffic counts, highway capacity studies, pavement condition studies and cost-benefit analysis.

Attributes of a household travel survey, or travel behavior inventory include:

- Travel Diary of ~ 1% sample of population (all trips made on one day) every 10 years
- Socioeconomic/demographic data of survey respondents
- Collection methodology:
  - Phone,
  - Mail,
  - In-Person at Home,
  - In-Person at Work,
  - Roadside

Many such surveys are available online at: Metropolitan Travel Survey Archive

Thought Question

What are the advantages and disadvantages of Revealed Preference surveys?

Stated Preference

In contrast with revealed preference, *Stated preference* is a group of techniques used to calculate the utility functions of transport options based on the response of an individual decision-maker to certain given options. The options generally are based on descriptions of the transport scenario or are constructed by the researcher

- *Contingent valuation* is based on the assumption that the best way to find out the value that an individual places on something is known by asking.
- *Compensating variation* is the compensating payment that leaves an individual as well off as before the economic change.
- *Equivalent variation* for a benefit is the minimum amount of money one would have to be compensated to leave the person as well as they would be after the change.
- *Conjoint analysis* refers to the application of the design of experiments to obtain the preferences of the individual, breaking the task into a list of choices or ratings that enable us to compute the relative importance of each of the attributes studied

Thought Question

What are the advantages and disadvantages of Stated Preference surveys?

Pavement conditions

adapted from Xie, Feng and David Levinson (2008) The Use of Road Infrastructure Data for Urban Transportation Planning: Issues and Opportunities. Published in Infrastructure Reporting and Asset Management Edited by Adjo Amekudzi and Sue McNeil. pp- 93-98. Publisher: American Society of Civil Engineers, Reston, Virginia.

Road infrastructure represents the supply side of an urban transportation system. Pavement condition is a critical indicator to the quality of road infrastructure in terms of providing a smooth and reliable driving environment on roads. A series of indices have been developed to evaluate the pavement conditions of road segments in their respective jurisdictions: Pavement Condition Index (PCI) is scored as a perfect roadway (100 points) minus point deductions for “distresses” that are observed; Present Serviceability Rating (PSR) is measured as vertical movement per unit horizontal movement (e.g. millimeters of vertical displacement per meter of horizontal displacement) as one drives along the road; (SR) Surface Rating is calculated by reviewing images of the roadway based on the frequency and severity of defects; Pavement Quality Index (PQI) is calculated using the PSR and SR to evaluate the general condition of the road. A high PQI (up to 4.5) means a road will most likely not need maintenance soon, whereas a low PQI means it can be selected for maintenance.

These indices of pavement quality are basic measures for road maintenance and preservation, for which each county develops its own performance standards to evaluate pavement conditions and make decisions on maintenance and preservation projects. Typically, pavement preservation projects are prioritized based on PCI of road segments: the lower the PCI, the higher the likelihood of selection. Taking Washington County, Minnesota as an example, the county has determined that a reasonable standard to maintain is an average PCI of 72. Thus any road segment with its PCI below 72 has a chance to be selected for preservation. Dakota County, Minnesota on the other hand, scores its preservation projects according to the measure of PQI: a road segment will be allocated 17 points (out of a possible 100) if its PQI falls lower than 3.1.

The pavement data structure is incompatible with the link-node structure of the planning road network used by the Metropolitan Council and other planning agencies. Typically, the measures of PCI, PSR, and PQI are stored in records indexed by “highway segment numbers” along each highway route. Highway sections with the same highway segment number are differentiated by their starting and ending stations. There is no exact match between highway segments in the actual road network and links in the planning network, as stationing is a position along the curved centerline of a highway while a planning network is a simplified structure consisting of only straight lines intersecting at points. Historic pavement data is generally unavailable in electronic format, although the information on pavement history such as pavement life and the duration since last repaving are important to estimate the cost of a preservation project, also affecting the decision whether a specific project will get selected and how much funding will be allocated.

Traffic flow

Traffic conditions reflect the travel demand loaded on a given road infrastructure. Traffic flows on roads, together with road

capacity, can be used to calculate the volume/capacity (V/C) ratio, which is an approximate indicator for the level of service of road infrastructure, and is commonly adopted by the jurisdictions in their respective decision making processes. The traffic flows on the planning road network are predicted by the transportation planning model, but the results have to be calibrated with actual traffic data.

Loop detectors are the primary technology currently employed in many US metropolitan areas to collect actual traffic data. E.g. In the Twin Cities of Minneapolis-St. Paul, about one thousand detector stations have been buried on major highways, through which Mn/DOT's Traffic Management Center collects, stores, and makes public traffic data every 30 seconds, including measured volume (flow) and occupancy, and calculated speed data for each detector station. Although the estimates of Annual Average Daily Traffic (AADT) for the planning road network are readily available, loop detectors provide more accurate measures of traffic volume, since they are collecting real-time data on a continuous basis. It also allows for calibrating models to hourly rather than daily conditions.

Due to the limited ability to convert raw data collected by loop detectors, however, most forecasting models rely on AADT data. The raw data are stored in a 30- second interval in binary codes. For planning uses they have to be converted and aggregated into desired measures, such as peak hour average, average for a particular month or a particular season, etc, in a systematic way.

Another issue in integrating loop detector data into a planning road network is to match the detector stations with the links in planning networks. Similar to the problem encountered in translating pavement data, detectors are located along major highways and mapped on the actual geometry of the network, while the planning road network is a simplified structure with only straight lines.

Sampling Issues (Statistics)

- Sample Size,
- Population of Interest
- Sampling Method,
- Error:
  - Measurement,
  - Sampling,
  - Computational,
  - Specification,
  - Transfer,
  - Aggregation
- Bias,
- Oversampling
- Extent of Collection
  - Spatial
  - Temporal
- Span of Data
  - Cross-section,
  - Time Series, and
  - Panel

Metadata

Adapted from Levinson, D. and Zofka, Ewa. (2006) "The Metropolitan Travel Survey Archive: A Case Study in Archiving" in Travel Survey Methods: Quality and Future Directions, Proceedings of 5th International Conference on Travel Survey Methods *(Peter Stopher and Cheryl Stecher, editors)*

Metadata allows data to function together. Simply put, metadata is information about information – labeling, cataloging and descriptive information structured to permit data to be processed. Ryssevik and Musgrave (1999) argue that high quality metadata standards are essential as metadata is the launch pad for any resource discovery, maps complex data, bridges the gap between data producers and consumers, and links data with its resultant reports and scientific studies produced about it. To meet the increasing needs for the proper data formats and encoding standards, the World Wide Web Consortium (W3C) has developed the generic Resource Description Framework (RDF) (W3C 2002). RDF treats metadata more generally, providing a standard way to use Extended Markup Language (XML) to "represent metadata in the form of statements about properties and relationships of

items" (W3C 2002). Resources can be almost any type of file, including of course, travel surveys. RDF delivers detailed and unified data description vocabulary.

Applying these tools specifically to databases, the Data Documentation Initiative (DDI) for Document Type Definitions (DTD) applies metadata standards used for documenting datasets. DDI was first developed by European and North American data archives, libraries and official statistics agencies. "The Data Documentation Initiative (DDI) is an effort to establish an international XML-based standard for the content, presentation, transport, and preservation of documentation for datasets in the social and behavioral sciences" (Data Documentation Initiative 2004). As this international standardization effort gathers momentum it is expected more and more datasets to be documented using DDI as the primary metadata format. With DDI, searching data archives on the Internet no longer depends on an archivist's skill at capturing the information that is important to researchers. The standard of data description provides sufficient detail sorted in a user-friendly manner.

This page titled 2.2: Data is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

2.3: Networks

Transportation systems have specific structure. Roads have length, width, and depth. The characteristics of roads depends on their purpose.

Roads

A road is a path connecting two points. The English word 'road' comes from the same root as the word 'ride' –the Middle English 'rood' and Old English 'rad' –meaning the act of riding. Thus a road refers foremost to the right of way between an origin and destination. In an urban context, the word street is often used rather than road, which dates to the Latin word 'strata', meaning pavement (the additional layer or stratum that might be on top of a path).

Modern roads are generally paved, and unpaved routes are considered trails in some countries. The pavement of roads began early in history. Approximately 2600 BCE, the Egyptians constructed a paved road out of sandstone and limestone slabs to assist with the movement of stones on rollers between the quarry and the site of construction of the pyramids. The Romans and others used brick or stone pavers to provide a more level, and smoother surface, especially in urban areas, which allows faster travel, especially of wheeled vehicles. The innovations of Thomas Telford and John McAdam reinvented roads in the early nineteenth century, by using less expensive smaller and broken stones, or aggregate, to maintain a smooth ride and allow for drainage. Later in the nineteenth century, application of tar (asphalt) further smoothed the ride. In 1824, asphalt blocks were used on the Champs-Élysées in Paris. In 1872, the first asphalt street (Fifth Avenue) was paved in New York (due to Edward de Smedt), but it wasn't until bicycles became popular in the late nineteenth century that the "Good Roads Movement" took off. Bicycle travel, more so than travel by other vehicles at the time, was sensitive to rough roads. Demands for higher quality roads really took off with the widespread adoption of the automobile in the United States in the early twentieth century.

The first good roads in the twentieth century were constructed of Portland cement concrete (PCC). The material is stiffer than asphalt (or asphalt concrete) and provides a smoother ride. Concrete lasts slightly longer than asphalt between major repairs, and can carry a heavier load, but is more expensive to build and repair. While urban streets had been paved with concrete in the US as early as 1889, the first rural concrete road was in Wayne County, Michigan, near to Detroit in 1909, and the first concrete highway in 1913 in Pine Bluff, Arkansas. By the next year over 2300 miles of concrete pavement had been laid nationally. However over the remainder of the twentieth century, the vast majority of roadways were paved with asphalt. In general only the most important roads, carrying the heaviest loads, would be built with concrete.

Roads are generally classified into a hierarchy. At the top of the hierarchy are freeways, which serve entirely a function of moving vehicles between other roads. Freeways are grade-separated and limited access, have high speeds and carry heavy flows. Below freeways are arterials. These may not be grade-separated, and while access is still generally limited, it is not limited to the same extent as freeways, particularly on older roads. These serve both a movement and an access function. Next are collector/distributor roads. These serve more of an access function, allowing vehicles to access the network from origins and destinations, as well as connecting with smaller, local roads, that have only an access function, and are not intended for the movement of vehicles with neither a local origin nor destination. Local roads are designed to be low speed and carry relatively little traffic.

The class of the road determines which level of government administers it. The highest roads will generally be owned, operated, or at least regulated (if privately owned) by the higher level of government involved in road operations; in the United States, these roads are operated by the individual states. As one moves down the hierarchy of roads, the level of government is generally more and more local (counties may control collector/distributor roads, towns may control local streets). In some countries freeways and other roads near the top of the hierarchy are privately owned and regulated as utilities, these are generally operated as toll roads. Even publicly owned freeways are operated as toll roads under a toll authority in other countries, and some US states. Local roads are often owned by adjoining property owners and neighborhood associations.

The design of roads is specified in a number of design manual, including the AASHTO Policy on the Geometric Design of Streets and Highways (or Green Book). Relevant concerns include the alignment of the road, its horizontal and vertical curvature, its super-elevation or banking around curves, its thickness and pavement material, its cross-slope, and its width.

Freeways

A motorway or freeway (sometimes called an expressway or thruway) is a multi-lane divided road that is designed to be high-speed free flowing, access-controlled, built to high standards, with no traffic lights on the mainline. Some motorways or freeways are financed with tolls, and so may have tollbooths, either across the entrance ramp or across the mainline. However in the United States and Great Britain, most are financed with gas or other tax revenue.

Though of course there were major road networks during the Roman Empire and before, the history of motorways and freeways dates at least as early as 1907, when the first limited access automobile highway, the Bronx River Parkway began construction in Westchester County, New York (opening in 1908). In this same period, William Vanderbilt constructed the Long Island Parkway as a toll road in Queens County, New York. The Long Island Parkway was built for racing and speeds of 60 miles per hour (96 km/hr) were accommodated. Users however had to pay a then expensive $2.00 toll (later reduced) to recover the construction costs of $2 million. These parkways were paved when most roads were not. In 1919 General John Pershing assigned Dwight Eisenhower to discover how quickly troops could be moved from Fort Meade between Baltimore and Washington to the Presidio in San Francisco by road. The answer was 62 days, for an average speed of 3.5 miles per hour (5.6 km/hr). While using segments of the Lincoln Highway, most of that road was still unpaved. In response, in 1922 Pershing drafted a plan for an 8,000 mile (13,000 km) interstate system which was ignored at the time.

The US Highway System was a set of paved and consistently numbered highways sponsored by the states, with limited federal support. First built in 1924, they succeeded some previous major highways such as the Dixie Highway, Lincoln Highway and Jefferson Highway that were multi-state and were constructed with the aid of private support. These roads however were not in general access-controlled, and soon became congested as development along the side of the road degraded highway speeds.

In parallel with the US Highway system, limited access parkways were developed in the 1920s and 1930s in several US cities. Robert Moses built a number of these parkways in and around New York City. A number of these parkways were grade separated, though they were intentionally designed with low bridges to discourage trucks and buses from using them. German Chancellor Adolf Hitler appointed a German engineer Fritz Todt Inspector General for German Roads. He managed the construction of the German Autobahns, the first limited access high-speed road network in the world. In 1935, the first section from Frankfurt am Main to Darmstadt opened, the total system today has a length of 11,400 km. The Federal-Aid Highway Act of 1938 called on the Bureau of Public Roads to study the feasibility of a toll-financed superhighway system (three east-west and three north-south routes). Their report Toll Roads and Free Roads declared such a system would not be self-supporting, advocating instead a 43,500 km (27,000 mile) free system of interregional highways, the effect of this report was to set back the interstate program nearly twenty years in the US.

The German autobahn system proved its utility during World War II, as the German army could shift relatively quickly back and forth between two fronts. Its value in military operations was not lost on the American Generals, including Dwight Eisenhower.

On October 1, 1940, a new toll highway using the old, unutilized South Pennsylvania Railroad right-of-way and tunnels opened. It was the first of a new generation of limited access highways, generally called superhighways or freeways that transformed the American landscape. This was considered the first freeway in the US, as it, unlike the earlier parkways, was a multi-lane route as well as being limited access. The Arroyo Seco Parkway, now the Pasadena Freeway, opened December 30, 1940. Unlike the Pennsylvania Turnpike, the Arroyo Seco parkway had no toll barriers.

A new National Interregional Highway Committee was appointed in 1941, and reported in 1944 in favor of a 33,900 mile system. The system was designated in the Federal Aid Highway Act of 1933, and the routes began to be selected by 1947, yet no funding was provided at the time. The 1952 highway act only authorized a token amount for construction, increased to $175 million annually in 1956 and 1957.

The US Interstate Highway System was established in 1956 following a decade and half of discussion. Much of the network had been proposed in the 1940s, but it took time to authorize funding. In the end, a system supported by gas taxes (rather than tolls), paid for 90% by the federal government with a 10% local contribution, on a pay-as-you-go" system, was established. The Federal Aid Highway Act of 1956 had authorized the expenditure of $27.5 billion over 13 years for the construction of a 41,000 mile interstate highway system. As early as 1958 the cost estimate for completing the system came in at $39.9 billion and the end date slipped into the 1980s. By 1991, the final cost estimate was $128.9 billion. While the freeways were seen as positives in most parts of the US, in urban areas opposition grew quickly into a series of freeway revolts. As soon as 1959, (three years after the Interstate act), the San Francisco Board of Supervisors removed seven of ten freeways from the city's master plan, leaving the Golden Gate bridge unconnected to the freeway system. In New York, Jane Jacobs led a successful freeway revolt against the Lower Manhattan Expressway sponsored by business interests and Robert Moses among others. In Baltimore, I-70, I-83, and

I-95 all remain unconnected thanks to highway revolts led by now Senator Barbara Mikulski. In Washington, I-95 was rerouted onto the Capital Beltway. The pattern repeated itself elsewhere, and many urban freeways were removed from Master Plans.

In 1936, the Trunk Roads Act ensured that Great Britain's Minister of Transport controlled about 30 major roads, of 7,100 km (4,500 miles) in length. The first Motorway in Britain, the Preston by-pass, now part of the M-6, opened in 1958. In 1959, the first stretch of the M1 opened. Today there are about 10,500 km (6300 miles) of trunk roads and motorways in England.

Australia has 790 km of motorways, though a much larger network of roads. However the motorway network is not truly national in scope (in contrast with Germany, the United States, Britain, and France), rather it is a series of local networks in and around metropolitan areas, with many intercity connection being on undivided and non-grade separated highways. Outside the Anglo-Saxon world, tolls were more widely used. In Japan, when the Meishin Expressway opened in 1963, the roads in Japan were in far worse shape than Europe or North American prior to this. Today there are over 6,100 km of expressways (3,800 miles), many of which are private toll roads. France has about 10,300 km of expressways (6,200 miles) of motorways, many of which are toll roads. The French motorway system developed through a series of franchise agreements with private operators, many of which were later nationalized. Beginning in the late 1980s with the wind-down of the US interstate system (regarded as complete in 1990), as well as intercity motorway programs in other countries, new sources of financing needed to be developed. New (generally suburban) toll roads were developed in several metropolitan areas.

An exception to the dearth of urban freeways is the case of the Big Dig in Boston, which relocates the Central Artery from an elevated highway to a subterranean one, largely on the same right-of-way, while keeping the elevated highway operating. This project is estimated to be completed for some $14 billion; which is half the estimate of the original complete US Interstate Highway System.

As mature systems in the developed countries, improvements in today's freeways are not so much widening segments or constructing new facilities, but better managing the road space that exists. That improved management, takes a variety of forms. For instance, Japan has advanced its highways with application of Intelligent Transportation Systems, in particular traveler information systems, both in and out of vehicles, as well as traffic control systems. The US and Great Britain also have traffic management centers in most major cities that assess traffic conditions on motorways, deploy emergency vehicles, and control systems like ramp meters and variable message signs. These systems are beneficial, but cannot be seen as revolutionizing freeway travel. Speculation about future automated highway systems has taken place almost as long as highways have been around. The Futurama exhibit at the New York 1939 World's Fair posited a system for 1960. Yet this technology has been twenty years away for over sixty years, and difficulties remain.

Layers of Networks

	Data Unit	Layer	Function
Host Layers	Data	7. Application	Network process to application
Host Layers	Data	6. Presentation	Data representation,encryption and decryption
Host Layers	Data	5. Session	Interhost communication
Host Layers	Segments	4. Transport	End-to-end connections and reliability,Flow control
Media Layers	Packet	3. Network	Path determination and logical addressing
Media Layers	Frame	2. Data Link	Physical addressing
Media Layers	Bit	1. Physical	Media, signal and binary transmission

All networks come in layers. The OSI Reference Model for the Internet is well-defined. Roads too are part of a layer of subsystems of which the pavement surface is only one part. We can think of a hierarchy of systems.

- Places
- Trip Ends
- End to End Trip
- Driver/Passenger
- Service (Vehicle & Schedule)
- Signs and Signals
- Markings

- Pavement Surface
- Structure (Earth & Pavement and Bridges)
- Alignment (Vertical and Horizontal)
- Right-Of-Way
- Space

At the base is *space*. On space, a specific *right-of-way* is designated, which is property where the road goes. Originally right-of-way simply meant legal permission for travelers to cross someone's property. Prior to the construction of roads, this might simply be a well-worn dirt path.

On top of the right-of-way is the *alignment*, the specific path a transportation facility takes within the right-of-way. The path has both vertical and horizontal elements, as the road rises or falls with the topography and turns as needed.

Structures are built on the alignment. These include the roadbed as well as bridges or tunnels that carry the road.

Pavement surface is the gravel or asphalt or concrete surface that vehicles actually ride upon and is the top layer of the structure. That surface may have *markings* to help guide drivers to stay to the right (or left), delineate lanes, regulate which vehicles can use which lanes (bicycles-only, high occupancy vehicles, buses, trucks) and provide additional information. In addition to marking, signs and signals to the side or above the road provide additional regulatory and navigation information.

Services use roads. Buses may provide scheduled services between points with stops along the way. Coaches provide scheduled point-to-point without stops. Taxis handle irregular passenger trips.

Drivers and *passengers* use services or drive their own vehicle (producing their own transportation services) to create an *end-to-end trip*, between an origin and destination. Each origin and destination comprises a *trip end* and those trip ends are only important because of the *places* at the ends and the activity that can be engaged in. As transportation is a derived demand, if not for those activities, essentially no passenger travel would be undertaken.

With modern information technologies, we may need to consider additional systems, such as Global Positioning Systems (GPS), differential GPS, beacons, transponders, and so on that may aide the steering or navigation processes. Cameras, in-pavement detectors, cell phones, and other systems monitor the use of the road and may be important in providing feedback for real-time control of signals or vehicles.

Each layer has rules of behavior:

- some rules are physical and never violated, others are physical but probabilistic
- some are legal rules or social norms which are occasionally violated

Hierarchy of Roads

Even within each layer of the system of systems described above, there is differentiation.

Transportation facilities have two distinct functions: through movement and land access. This differentiation:

- permits the aggregation of traffic to achieve economies of scale in construction and operation (high speeds);
- reduces the number of conflicts;
- helps maintain the desired quiet character of residential neighborhoods by keeping through traffic away from homes;
- contains less redundancy, and so may be less costly to build.

Hierarchy of Roads

Functional Classification	Types of Connections	Relation to Abutting Property	Minnesota Examples
Limited Access (highway)	Through traffic movement between cities and across cities	Limited or controlled access highways with ramps and/or curb cut controls.	I-94, Mn280

Linking (arterial:principal and minor)	Traffic movement between limited access and local streets.	Direct access to abutting property.	University Avenue, Washington Avenue
Local (collector and distributor roads)	Traffic movement in and between residential areas	Direct access to abutting property.	Pillsbury Drive, 17th Avenue

Network

- Zone Centroid - special node whose number identifies a zone, located by an "x" "y" coordinate representing longitude and latitude (sometimes "x" and "y" are identified using planar coordinate systems).
- Node (vertices) - intersection of links , located by *x* and *y* coordinates
- Links (arcs) - short road segments indexed by from and to nodes (including centroid connnectors), attributes include lanes, capacity per lane, allowable modes
- Turns - indexed by at, from, and to nodes
- Routes, (paths) - indexed by a series of nodes from origin to destination. (e.g. a bus route)
- Modes - car, bus, HOV, truck, bike, walk etc.

Matrices

Scalar

A scalar is a single value that applies model-wide; e.g. the price of gas or total trips.

	Total Trips
Variable	T

Vectors

Vectors are values that apply to particular zones in the model system, such as trips produced or trips attracted or number of households. They are arrayed separately when treating a zone as an origin or as a destination so that they can be combined into full matrices.

- vector (origin) - a column of numbers indexed by traffic zones, describing attributes at the origin of the trip (e.g. the number of households in a zone)

	Trips Produced at Origin Zone
Origin Zone 1	T_{i1}
Origin Zone 2	T_{i2}
Origin Zone 3	T_{i3}

- vector (destination) - a row of numbers indexed by traffic zones, describing attributes at the destination

	Destination Zone 1

Trips Attracted to Destination Zone	T_{j1}

Full Matrices

A full or interaction matrix is a table of numbers, describing attributes of the origin-destination pair

	Destination Zone 1
Origin Zone 1	T_{11}
Origin Zone 2	T_{21}
Origin Zone 3	T_{31}

Thought Questions

- Identify the rules associated with each layer?
- Why aren't all roads the same?
- How might we abstract the real transportation system when representing it in a model for analysis?
- Why is abstraction useful?

Abbreviations

- SOV - single occupant vehicle
- HOV - high occupancy vehicle (2+, 3+, etc.)
- TAZ - transportation analysis zone or traffic analysis zone

Variables

- msXX - scalar matrix
- moXX - origin vector matrix
- mdXX - destination vector matrix
- mfXX - full vector matrix
- T - Total Trips
- T_i - Trips Produced from Origin Zone i
- T_j - Trips Attracted to Destination Zone j
- T_{ij} - Trips Going Between Origin Zone i and Destination Zone j

Key Terms

- Zone Centroid
- Node

- Links
- Turns
- Routes
- Modes
- Matrices
- Right-of-way
- Alignment
- Structures
- Pavement Surface
- Markings
- Signs and Signals
- Services
- Driver
- Passenger
- End to End Trip
- Trip Ends
- Places

External Exercises

Use the ADAM software at the STREET website and examine the network structure. Familiarize yourself with the software, and edit the network, adding at least two nodes and four one-way links (two two-way links), and deleting nodes and links. What are the consequences of such network adjustments? Are some adjustments better than others?

Video

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

This page titled 2.3: Networks is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

2.4: Modes

Wikipedia's article gives a list of a variety of modes largely defined by their technology (are they powered by animals or engines, are the engines on the vehicle or is the vehicle powered by a cable, do the vehicles travel on land, sea, or water, etc. While this is a reasonably comprehensive list and a reasonable organization of the subject for Wikipedia, it does not really get at the transportation aspect of modes, focusing instead on their mechanical aspects.

What characteristics describe and differentiate modes? Every mode must differ from every other mode on at least one dimension (otherwise they would be the same mode). This is analogous to the idea of speciation in biology. The Figure is a first cut at this for surface passenger transportation.

This distinguishes primarily on the non-mechanical (non-propulsion) characteristics of the service. Of course not every possible dimension is identified, and a few of the circles contain multiple modes which are otherwise obviously distinct (e.g. gondolas and subways are much the same from a transportation service perspective but for one is underground and uses a train and the other is suspended by a cable which moves it). This taxonomy differentiates things that were qualitatively different rather than quantitatively different.

So the first cut is about time, is a reservation required or not (i.e. does it need some advance planning).

The second cut is about time as well, is the service scheduled or dynamic.

The third cut is about space, are the routes fixed or dynamic.

If the route is fixed, are stops fixed (i.e. does the vehicle stop at every stop, or only when called, like a bus).

Otherwise if the routes are dynamic, things get a bit more ad-hoc, as the key question changes.

Some traditional distinctions (access mode vs. primary mode, such as walk to transit vs. drive to transit) are not distinguished here, rather that would be thought of as at least two trips, one where you walk or drive to some place (with the purpose of changing modes), and second where you take some form of transit.

This page titled 2.4: Modes is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

3: MODELING METHODS

3.1: Agent-based Modeling
3.2: Choice Modeling
3.3: Land Use Forecasting
3.4: Trip Generation
3.5: Destination Choice
3.6: 3-6 Route Choice

This page titled 3: Modeling Methods is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

3.1: Agent-based Modeling

Agent-based modeling

Transportation engineers and planners rely on transportation forecasting models to address a wide range of increasingly complicated issues, from congestion and air quality, to social equity concerns. Two major strands of travel demand models have emerged over the past several decades, trip-based and activity-based approaches.

The traditional four-step travel demand model, often referred to as the trip-based approach, takes individual trips as the elementary subjects and considers aggregate travel choices in four steps: trip generation, trip distribution, modal split, and route assignment. This sequential travel demand modeling paradigm, which originated in the 1950s when limited data, computational power, and algorithms were available, ignores the diversity across individuals and does not have solid foundation in travel behavior theory. Discrete choice analysis, describes travel demand as a multi-dimensional hierarchical choice process, including residential and business location choice, trip origin, trip destination, travel modes and etc. Although discrete choice models could improve travel demand prediction by classifying travelers according to certain attributes such as age, gender and household incomes, it still in the end focuses on aggregate travel behaviors and ignores individual decision-making processes. Another flaw of four-step model lies in the fact that this sequential modeling process ignores the interaction between steps and could not predicts certain phenomenon such as induced travel or demand, which can be thought of as a feedback from traffic assignment to trip generation, distribution, and mode split. Although introducing feedback and iteratively applying four-step approach could mitigate this problem, researchers believe that a coherent framework should be introduced to address four steps simultaneously.

To overcome these inadequacies of conventional four-step modeling, activity-based models have been applied in travel demand analysis since the 1970s. Activity-based models predict activities and related travel choices by considering time and space constraints as well as individual characteristics. Individuals will follow a sequence of activities and make corresponding trips connecting those activities to maximize their utilities. Macroscopic travel patterns are predicted through aggregation of individual travel choice.

Although activity-based models have the potential to bridge the gap between individual decision-making processes and macroscopic travel demand, these models require solving many optimization problems simultaneously, which is computationally difficult and behaviorally unrealistic. Therefore, some models employs external aggregate methods such as User Equilibrium (Deterministic (DUE) or Stochastic (SUE)) to address route choice, which compromises their claims as microscopic decision-making models.

The agent-based travel demand model has emerged as a new generation of transportation forecasting tools and provides an alternative to address the topic of travel demand modeling. This modeling approach is flexible and capable to model individual decision making process. There have been many applications of agent-based model in transportation (Transportation Research Part C (2002) dedicated a special issue to this topic). This modeling strategy, however, has not yet been widely adopted in travel demand modeling practice.

To build a pedagogically appropriate model, this chapter introduces an Agent-based Demand and Assignment Model (ADAM), extending Zhang and Levinson (2004), which addresses the destination choice and route choice problems with consideration of congestion. Students have the opportunity to work with the ADAM model for several exercises.

Introduction to agent-based models for transport

While agent-based models are not commonly used in travel demand forecasting as such, many activity-based models are agent-based models of a sort, at least in part, though the behaviors of the agents are typically very complex. Historically, agent-based models come from different fields such as genetics, artificial intelligence, cognitive science, social science. The advantage of using them in transportation begins first with the intuition they provide. It makes more sense to people to think of individual travelers behaving rather than flows. This is in part because it is also more realistic, in that it can be formulated to capture the process by which travelers make decisions, and because it is tracking individuals, can be internally consistent (so that a given traveler has a particular set of constraints (like income, obligations, and time available)

There are several elements in an agent-based model:

- *Agents* are like people who have characteristics, goals and behavioral rules. The actions of agents depend on the environment they inhabit.
- The *environment* provides a space where agents live. The environment is shaped by the actions of agents.
- *Interaction rules* describe how agents and the environment interact

An agent-based model evolves by itself once those micro-level elements are specified. Macro-level properties emerge from this evolutionary process.

An exploratory agent-based model is presented below. The advantage of this model is its simplicity. Clearly, it will lose some predictive detail, but hopefully gives you a flavor of the kinds of modeling approaches and things that can be modeled with agent-based models in the realm of travel demand.

Agent-based Demand and Assignment Model (ADAM)

The agent-based modeling approach assumes that aggregate urban travel demand patterns emerge from multi-dimensional choice process of individuals. All agents have individual characteristics, goals, and rules of travel behaviors. Agents exchange information with the environment on their travel experiences and adjust their travel choices according to available information. In ADAM travelers are active agents and nodes are fixed point agents, while links comprise the environment.

ADAM can be thought of as modeling the AM commute. As shown in Figure 1, ADAM examines the status of each traveler after updating turning matrices at nodes. If a traveler has not found a satisfactory job (status = 1), that traveler will continue the random process of job searching following the rules presented later in this paper. The process will repeat until either all travelers have found jobs (chosen a destination) or some maximum number of iterations are reached. The key components of the agent-based model are introduced in turn below.

Agents

Travelers aim to find a job on the network and a route leading from their origin to this destination with the lowest cost. In the searching process, each traveler visits a node and decides to either accept or reject a job available at that node according to rules discussed later in this paper. If they reject a job at that node, they proceed to another node. Travelers learn current link travel times in the neighborhood of the node when they visit a node through this link and they will only proceed through one link at each step. By accumulating link travel time information during the trip, travelers could derive travel cost between any two nodes they visited.

Nodes are geographic locations where links intersect in the real world. In this model, they also represent the abstract centroids of traffic zone where travelers originate from and are destined to. Furthermore, nodes are carriers of pooled, collective knowledge, including both shortest path information and attractiveness of adjacent nodes. Travelers would exchange knowledge with nodes once they arrived at a new node. The knowledge and exchanging behavior is an abstraction of information spread in a community and communications among travelers in the real world. Links represents roads in the real world and have attributes such as length, free flow travel time, and capacity. Links also provide information about traffic flow and travel time to travelers passing by, which abstracts travelers' observation of traffic condition in the real world. Links impose geographic constraints to travelers since they are only able to visit adjacent nodes directly connected by a link with the node they are currently visiting.

Rules

Rules are the most important attributes of an agent-based model, which drives the evolution of the model given initial condition. There are two fundamental rules in ADAM: turning rules for finding a destination and information exchange rules for improving paths.

Destination selection rules: Network Origin-Destination Exploration

The first element of ADAM is for each traveler, the discovery of a destination. The model which does this Network Origin-Destination Exploration (NODE) is described below.

Nodes provide turning guidance matrices to travelers, which decide the probability for each traveler to accept a job or proceed to

the next node and which direction to go in the later case. Each node (i) has a set of supply nodes $S(s_1 \dots s_S)$ and a set of demand nodes $D(d_1, \dots, d_D)$. Therefore, a SD matrix will be provided and each term, $p_{s,d}$ (for simplicity, i is omitted), represents the probability to move from supply node s to demand node d.

(1)
$$P = \begin{bmatrix} p_{s_1,d_1} & \cdots & p_{s_1,d_D} \\ \cdots & \cdots & \cdots \\ p_{s_S,d_1} & \cdots & p_{s_S,d_D} \end{bmatrix}$$

The probability is determined by many factors, including travelers' characteristics (Ω_t), the opportunity (or attractiveness) at the current node (b_i), the opportunity at demand nodes (b_d) and the ease of reaching those opportunities (A).

(2)
$$P = f(\Omega_t, b_i, b_d, A) \tag{3.1.1}$$

Different definitions of turning probability reflect assumptions of different underlying decision-making processes of travelers regarding where to work and may lead to very different travel demand patterns on the network. Zhang and Levinson (2004) assumed that this probability is proportional to jobs available at each node and ignored the ease of reaching them (travel cost). Another disadvantage of this assumption is that if a node does not have available jobs, travelers will never search in this direction even though more jobs may be available via this node.

Extending Zhang and Levinson (2004), a Logit-form probability is used, where c_d represents travel cost to a destination while c_i is the corresponding intrazonal travel cost. The parameter θ indicates the importance of travel cost when travelers evaluate possible destinations, while β is related to people's relative willingness to travel. A larger β implies that travelers are more likely to accept jobs at current node, thus have a shorter travel length.

(3)
$$p_{s,d} = 0 if s = d \tag{3.1.2}$$

(4) \[p_{s,d}=\dfrac{b_de^{-\theta c_d}{\beta b_ie^\theta c_d+\displaystyle \sum_{d\epsilon,d\cancel=i}b_de^{-\thetac_d}}

(5)
$$p_{s,d} = \frac{\beta b_d e^{-\theta c_d}}{\beta b_i e^{-\theta c_d} + \sum_{d \in D, d \neq i} b_d e^{-\theta c_d}} \text{ if } s \neq d \text{ and } d \neq i$$

The variable b_d reflects the opportunity or attractiveness of a node and can be further generalized beyond the number of jobs. We could define it as the summation of jobs on all nodes adjacent to node d, which abstracts the regional accessibility discussed in many previous studies (Handy, 1993). This definition could mitigate the aforementioned problem of search direction. Using accessibility to the whole network is another possibility. However, this may lead to essentially random search since accessibility to the whole network of nearby nodes may be quite similar. In this study we adopt regional accessibility as the indicator of attractiveness to next node, while willingness to accept a job (to stay) is proportional to jobs available at current node.

Path learning rule: Agent-based route choice

The other important rule in ADAM is the path learning rule. Travelers will learn travel cost of links on their travel route, while nodes keep information about the shortest path from itself to all other nodes which have been visited by travelers to that node. Once a traveler arrives at a new node, that traveler compares their knowledge about travel cost from the current node to each node on the traveler's travel route. Both of them will keep the shorter "shortest path" after knowledge exchange. Although nodes originally have very limited knowledge about the routes in the remaining network, information spreads rapidly on the network. With the congested link travel time, which can be simply defined as any available travel time-flow relationship, each traveler's choice will change the link travel time on the network and thus affect destination and route choice of other travelers. Travelers' route adjustments will trigger more significant change on the network thus other travelers' behavior. This mechanism reflects the complexity of the real

world.

The initial route choice can be either given or generated by a random-walk route searching process at iteration 0. In the random walk scenario, travelers set off from their origins and travel in a randomly chosen direction, updating directions after arriving at each node. However, directed cycles and U-turns are prevented. Once travelers arrive at the destination, their travel routes become the initial travel route and will be updated in subsequent iterations. The randomness of searching direction and the large number of travelers will ensure the diversity of initial route choices, which comprises the knowledge based on subsequent iterations.

On subsequent iterations, each traveler follows a fixed route chosen at the end of the previous iteration. Once arriving at a destination centroid, travelers will enrich the information set with their individual knowledge while benefiting from the pooled knowledge at the same time by exchanging both shortest path and toll information with centroids. Those travelers will also bring that updated information back to their origin and repeat the exchange process. The information exchange mechanism is illustrated by Figure 1.

Agent-based route choice with Learning and Exchange of Information (Illustration of specific model in wikibook Fundamentals of Transportation

As illustrated in Figure 1, suppose that the traveler originating at node 1 is traveling to node 5, initially via node 4. His initial shortest path knowledge is 1-3-4-5. Suppose the shortest path information stored at node 5 is 4-5, 3-5, 2-3-5 and 1-2-3-5, respectively from nodes 4, 3, 2 and 1. The comparison starts from the node closest to the current node along the path chain in traveler's memory and repeats for each node on this chain until reaching the origin. After comparing the path from node 3 to 5, the traveler's path information is updated to 1-3-5 since the shortest path for this path segment proposed by the node is shorter than that held by the traveler. Notice that this improvement has also changed the shortest path from node 1 to 5 in the traveler's memory. Consequently, the node will adopt the path from node 1 proposed by the traveler since 1-3-5 is better than 1-2-3-5. The updated path from node 1 to 5 then becomes part of the traveler's shortest path information. This information exchange mechanism will naturally mutate the path chain and generate the most efficient route, sometimes better than all known existing routes. Since nodes store K alternative paths, nodes will insert the path proposed by the visitor in their information pool as long as this path is better than the longest path stored. This information will also be shared with those travelers visiting node 5 at subsequent steps.

After stopping at the destination node, travelers compare their travel route determined at the end of previous iterations and shortest path learned during the currently iteration. The path length is evaluated in dollar value by each traveler, considering their individual value of time and the toll charged by each link segment. Since travelers have different values of time, the cost of K alternatives should be reevaluated and sorted for each traveler. If the path suggested by the destination node is better than their current route, the travelers have a probability to switch to the better route that iteration. In general,

$$P = f(\sigma, \Delta, T) \tag{3.1.3}$$

To apply this model, we choose a specific form:

{\displaystyle \left\{{\begin{array}{l}{\begin{array}{*{20}c}{P=\sigma (1-e^{-\gamma \Delta })}&{{\begin{array}{*{20}c}{\rm {if}}&{\rm {\Delta
T}}\\\end{array}}{\rm {}}}\\\end{array}}{\rm {}}\\{\begin{array}{*{20}c}{P=0}&

ParseError: invalid DekiScript (click for details)

```
Callstack:
    at (Bookshelves/Civil_Engineering/Fundamentals_of_Transportation/03:_Modeling_Methods/3.01
```

\\end{array}}{\rm {}}\\end{array}}\right.}" src="https://wikimedia.org/api/rest_v1/me...433bb7fd723077">

Where:

- Δ represents the potential benefit by switching routes, which is defined as the time or money saving by choosing route proposed by the destination node instead of sticking to the current route.
- T is the threshold of benefit perception, which reflects both the incapability to perceive small benefit and the inertia for people to change route.
- σ denotes the probability of perceiving an existing better route in a given day, and captures the differentiation in the effectiveness of social networks defines the shape of the probability curve.

ARC simulates the day-to-day route choice behavior of travelers and this probability curve must account for two factors:

1. the probability a traveler perceives this better path once its information is available and
2. the probability a traveler takes this path once it is learned. It should be noted that information spreading takes time and not everyone learns immediately.

Travelers with more effective social networks are more likely to be exposed to such information and thus have a higher probability of learning the better path. Once a new road opens, it takes weeks or even months before the flow reaches a stable level. Even when people learn a better alternative, route change involves a certain switching cost preventing travelers from changing routes immediately. Or travelers may just resist changing because of inertia. Considering these factors, this curve should increase as benefits increase and reach some upper limit predicted by the willingness to learn. Estimation of this curve through survey or other psychological studies will enhance the empirical foundation of the model.

Figure 2 illustrates the flow chart of ARC. After travelers choose their routes according to the aforementioned probability, link flow and link travel time will be updated. Consequently, the cost of all possible paths stored both at nodes and travelers will be updated without changing the choice set. Then travelers will follow their new route and repeat the described process until an equilibrium pattern is reached (equilibrium is defined here as link flow variance smaller than a pre-determined threshold ϵ, we arbitrarily choose $\epsilon = 5$. Once this equilibrium is reached, no traveler has the incentive to change their travel route according to their behavioral rules and available information. Thus a link flow pattern would be reached and could be provided to other model components under a more comprehensive framework.

Iterations

Traditional travel demand models disentangled this complexity by formulating an optimization problem, using either Deterministic or Stochastic User Equilibrium. However, algorithms employed to solve such optimization problems are computationally cumbersome and behaviorally unrealistic. Instead, ADAM introduces a heuristic learning process to address this challenge. Under this framework, travelers will reenter the network and choose their destination and route again according to the link travel time resulting from their previous choices. Updated shortest path information will be learned and spread by travelers. This process mimics people's job change and route change behavior. Given the initial condition, ADAM evolves with previously defined rules and a pattern may be achieved according to certain convergence rules, from which macroscopic information such as trip distribution and traffic assignment can be extracted by summing up individual choices.

This page titled 3.1: Agent-based Modeling is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

3.2: Choice Modeling

Mode choice analysis is the third step in the conventional four-step transportation forecasting model, following Trip Generation and Destination Choice but before Route Choice. While trip distribution's zonal interchange analysis yields a set of origin destination (OD) tables which tells where the trips will be made, mode choice analysis allows the modeler to determine what mode of transport will be used.

Disaggregate Travel Demand models

Travel demand theory was introduced in the appendix on traffic generation. The core of the field is the set of models developed following work by Stan Warner in 1962 (Strategic Choice of Mode in Urban Travel: A Study of Binary Choice). Using data from the CATS, Warner investigated classification techniques using models from biology and psychology. Building from Warner and other early investigators, disaggregate demand models emerged. Analysis is disaggregate in that individuals are the basic units of observation, yet aggregate because models yield a single set of parameters describing the choice behavior of the population. Behavior enters because the theory made use of consumer behavior concepts from economics and parts of choice behavior concepts from psychology. Researchers at the University of California, Berkeley (especially Daniel McFadden, who won a Nobel Prize in Economics for his efforts) and the Massachusetts Institute of Technology (Moshe Ben-Akiva) (and in MIT associated consulting firms, especially Cambridge Systematics) developed what has become known as choice models, direct demand models (DDM), Random Utility Models (RUM) or, in its most used form, the multinomial logit model (MNL).

Choice models have attracted a lot of attention and work; the Proceedings of the International Association for Travel Behavior Research chronicles the evolution of the models. The models are treated in modern transportation planning and transportation engineering textbooks.

One reason for rapid model development was a felt need. Systems were being proposed (especially transit systems) where no empirical experience of the type used in diversion curves was available. Choice models permit comparison of more than two alternatives and the importance of attributes of alternatives. There was the general desire for an analysis technique that depended less on aggregate analysis and with a greater behavioral content. And, there was attraction too, because choice models have logical and behavioral roots extended back to the 1920s as well as roots in Kelvin Lancaster's consumer behavior theory, in utility theory, and in modern statistical methods.

The Logit Model

The Logit Model, widely used for transportation forecasting in various forms, was first theorized by Daniel McFadden. The Logit model says, the probability that a certain mode choice will be taken is proportional to e raised to the utility over the sum of e raised to the utility.

$$P_m = \frac{e^{u_{ijm}}}{\sum e^{u_{ijm}}} \tag{3.2.1}$$

For any Logit Model the sum of the probability of all modes will equal 1.

$$1 = \sum P_m \tag{3.2.2}$$

The Logit Model also says that if a new mode of transportation is added to a system (or taken away) then the original modes will lose (or gain) an amount of travels proportional to their share originally.

Steps for the Logit Model:

- Compute the Utility for each OD pair and mode
- Compute Exponentiated utilities for each OD pair and mode
- Sum Exponentiated utilities for each OD pair
- Compute Probability for each mode by OD pair
- Multiply Probability for OD pair by number of trips for each OD pair

Psychological roots

Early psychology work involved the typical experiment: Here are two objects with weights, w_1 and w_2, which is heavier? The finding from such an experiment would be that the greater the difference in weight, the greater the probability of choosing correctly. Graphs similar to the one on the right result.

Distribution of perceived weights

Louis Leon Thurstone proposed (in the 1920s) that perceived weight,

$$w = v + e \tag{3.2.3}$$

where v is the true weight and e is random with $E(e) = 0$

The assumption that e is normally and identically distributed (NID) yields the binary probit model.

Econometric formulation

Economists deal with utility rather than physical weights, and say that

observed utility = mean utility + random term.

Utility in this context refers to the total satisfaction (or happiness) received from making a particular choice or consuming a good or service.

The characteristics of the object, x, must be considered, so we have

$$u(x) = v(x) + e(x) \tag{3.2.4}$$

If we follow Thurston's assumption, we again have a probit model.

An alternative is to assume that the error terms are independently and identically distributed with a Weibull, Gumbel Type I, or double exponential distribution (They are much the same, and differ slightly in their tails (thicker) from the normal distribution). This yields the multinomial logit model (MNL). Daniel McFadden argued that the Weibull had desirable properties compared to other distributions that might be used. Among other things, the error terms are normally and identically distributed. The logit model is simply a log ratio of the probability of choosing a mode to the probability of not choosing a mode.

$$log\left(\frac{P_i}{1-P_i}\right) = v(x_i) \tag{3.2.5}$$

Observe the mathematical similarity between the logit model and the S-curves we estimated earlier, although here share increases with utility rather than time. With a choice model we are explaining the share of travelers using a mode (or the probability that an individual traveler uses a mode multiplied by the number of travelers).

The comparison with S-curves is suggestive that modes (or technologies) get adopted as their utility increases, which happens over time for several reasons. First, because the utility itself is a function of network effects, the more users, the more valuable the service, higher the utility associated with joining the network. Second, because utility increases as user costs drop, which happens when fixed costs can be spread over more users (another network effect). Third, technological advances, which occur over time and as the number of users increases, drive down relative cost.

An illustration of a utility expression is given:

$$log\left(\frac{P_A}{1-P_A}\right) = \beta_0 + \beta_1(c_A - c_T) + \beta_2(t_A - t_T) + \beta_3 I + \beta_4 N = v_A \tag{3.2.6}$$

where

- P_i = Probability of choosing mode i.
- P_A = Probability of taking auto

- c_A,c_T = cost of auto, transit
- t_A,t_T = travel time of auto, transit
- I = income
- N = Number of travelers

With algebra, the model can be translated to its most widely used form:

$$\frac{P_A}{1-P_A} = e^{v_A} \tag{3.2.7}$$

$$P_A = e^{v_A} - P_A e^{v_A} \tag{3.2.8}$$

$$P_A(1+e^{v_A}) = e^{v_A} \tag{3.2.9}$$

$$P_A = \frac{e^{v_A}}{1+e^{v_A}} \tag{3.2.10}$$

It is fair to make two conflicting statements about the estimation and use of this model:

1. It's a "house of cards", and
2. Used by a technically competent and thoughtful analyst, it's useful.

The "house of cards" problem largely arises from the utility theory basis of the model specification. Broadly, utility theory assumes that (1) users and suppliers have perfect information about the market; (2) they have deterministic functions (faced with the same options, they will always make the same choices); and (3) switching between alternatives is costless. These assumptions don't fit very well with what is known about behavior. Furthermore, the aggregation of utility across the population is impossible since there is no universal utility scale.

Suppose an option has a net utility u_{jk} (option k, person j). We can imagine that having a systematic part v_{jk} that is a function of the characteristics of an object and person j, plus a random part e_{jk}, which represents tastes, observational errors, and a bunch of other things (it gets murky here). (An object such as a vehicle does not have utility, it is characteristics of a vehicle that have utility.) The introduction of e lets us do some aggregation. As noted above, we think of observable utility as being a function:

$$v_A = \beta_0 + \beta_1(c_A - c_T) + \beta_2(t_A - t_T) + \beta_3 I + \beta_4 N \tag{3.2.11}$$

where each variable represents a characteristic of the auto trip. The value β_0 is termed an alternative specific constant. Most modelers say it represents characteristics left out of the equation (e.g., the political correctness of a mode, if I take transit I feel morally righteous, so β_0 may be negative for the automobile), but it includes whatever is needed to make error terms NID.

Econometric Estimation

Turning now to some technical matters, how do we estimate $v(x)$? Utility ($v(x)$) isn't observable. All we can observe are choices (say, measured as 0 or 1), and we want to talk about probabilities of choices that range from 0 to 1. (If we do a regression on 0s and 1s we might measure for j a probability of 1.4 or -0.2 of taking an auto.) Further, the distribution of the error terms wouldn't have appropriate statistical characteristics.

Figure: Likelihood Function for the Sample {1,1,1,0,1}.

The MNL approach is to make a maximum likelihood estimate of this functional form. The likelihood function is:

$$L^* = \prod_{n=1}^{N} f(y_n|x_n,\theta) \tag{3.2.12}$$

we solve for the estimated parameters

$$\hat{\theta} \tag{3.2.13}$$

that max L*. This happens when:

$$\frac{\partial L}{\partial \hat{\theta}_N} = 0 \tag{3.2.14}$$

The log-likelihood is easier to work with, as the products turn to sums:

$$lnL^* = \prod_{n=1}^{N} lnf(y_n|x_n, \theta) \tag{3.2.15}$$

Consider an example adopted from John Bitzan's Transportation Economics Notes. Let *X* be a binary variable that is *gamma* and 0 with probability (1- *gamma*). Then f(0) = (1- *gamma*) and f(1) = *gamma*. Suppose that we have 5 observations of X, giving the sample {1,1,1,0,1}. To find the maximum likelihood estimator of *gamma* examine various values of *gamma*, and for these values determine the probability of drawing the sample {1,1,1,0,1} If *gamma* takes the value 0, the probability of drawing our sample is 0. If *gamma* is 0.1, then the probability of getting our sample is: f(1,1,1,0,1) = f(1)f(1)f(1)f(0)f(1) = 0.1*0.1*0.1*0.9*0.1=0.00009. We can compute the probability of obtaining our sample over a range of *gamma* – this is our likelihood function. The likelihood function for n independent observations in a logit model is

$$L^* = \prod_{n=1}^{N} P_i^{Y_i}(1 - P_i)^{1-Y_i} \tag{3.2.16}$$

where: Y_i = 1 or 0 (choosing e.g. auto or not-auto) and Pi = the probability of observing Yi=1

The log likelihood is thus:

$$\ell = lnL^* = \prod_{i=1}^{n}[Y_i lnP_i + (1 - Y_i)ln(1 - P_i)] \tag{3.2.17}$$

In the binomial (two alternative) logit model,

\[P_auto=\dfrac{e^{v(x_{auto})}{1+e^{v(x_{auto})}\], so

$$\ell = lnL^* = \prod_{i=1}^{n}[Y_i v(x_{auto}) - ln(1 + e^{v(x_{auto})}] \tag{3.2.18}$$

The log-likelihood function is maximized setting the partial derivatives to zero:

$$\frac{\partial L}{\partial \beta} = \prod_{i=1}^{n}(Y_i - \hat{P}_i) = 0 \tag{3.2.19}$$

The above gives the essence of modern MNL choice modeling.

Independence of Irrelevant Alternatives (IIA)

Independence of Irrelevant Alternatives is a property of Logit, but not all Discrete Choice models. In brief, the implication of IIA is that if you add a mode, it will draw from present modes in proportion to their existing shares. (And similarly, if you remove a mode, its users will switch to other modes in proportion to their previous share). To see why this property may cause problems, consider the following example: Imagine we have seven modes in our logit mode choice model (drive alone, carpool 2 passenger, carpool 3+ passenger, walk to transit, auto driver to transit (park and ride), auto passenger to transit (kiss and ride), and walk or bike). If we eliminated Kiss and Ride, a disproportionate number may use Park and Ride or carpool.

Consider another example. Imagine there is a mode choice between driving and taking a red bus, and currently each has 50% share. If we introduce another mode, let's call it a blue bus with identical attributes to the red bus, the logit mode choice model would give each mode 33.3% of the market, or in other words, buses will collectively have 66.7% market share. Logically, if the mode is truly identical, it would not attract any additional passengers (though one can imagine scenarios where adding capacity would increase bus mode share, particularly if the bus was capacity constrained.

There are several strategies that help with the IIA problem. Nesting of choices allows us to reduce this problem. However, there is an issue of the proper Nesting structure. Other alternatives include more complex models (e.g. Mixed Logit) which are more difficult to estimate.

Returning to roots

The discussion above is based on the economist's utility formulation. At the time MNL modeling was developed there was some attention to psychologist's choice work (e.g., Luce's choice axioms discussed in his Individual Choice Behavior, 1959). It has an analytic side in computational process modeling. Emphasis is on how people think when they make choices or solve problems (see Newell and Simon 1972). Put another way, in contrast to utility theory, it stresses not the choice but the way the choice was made. It provides a conceptual framework for travel choices and agendas of activities involving considerations of long and short term memory, effectors, and other aspects of thought and decision processes. It takes the form of rules dealing with the way information is searched and acted on. Although there is a lot of attention to behavioral analysis in transportation work, the best of modern psychological ideas are only beginning to enter the field. (e.g. Golledge, Kwan and Garling 1984; Garling, Kwan, and Golledge 1994).

Modal Split

This page describes historical, but no longer standard, practice in Mode Choice models.

The early transportation planning model developed by the Chicago Area Transportation Study (CATS) focused on transit, it wanted to know how much travel would continue by transit. The CATS divided transit trips into two classes: trips to the CBD (mainly by subway/elevated transit, express buses, and commuter trains) and other (mainly on the local bus system). For the latter, increases in auto ownership and use were trade off against bus use; trend data were used. CBD travel was analyzed using historic mode choice data together with projections of CBD land uses. Somewhat similar techniques were used in many studies. Two decades after CATS, for example, the London study followed essentially the same procedure, but first dividing trips into those made in inner part of the city and those in the outer part. This procedure was followed because it was thought that income (resulting in the purchase and use of automobiles) drove mode choice.

Diversion Curve techniques

The CATS had diversion curve techniques available and used them for some tasks. At first, the CATS studied the diversion of auto traffic from streets and arterial to proposed expressways. Diversion curves were also used as bypasses were built around cities to establish what percentage of the traffic would use the bypass. The mode choice version of diversion curve analysis proceeds this way: one forms a ratio, say:

$$\frac{c_{transit}}{c_{auto}} = R \tag{3.2.20}$$

where:

c_m = travel time by mode m and

R is empirical data in the form:

Figure: Mode choice diversion curve

Given the R that we have calculated, the graph tells us the percent of users in the market that will choose transit. A variation on the technique is to use costs rather than time in the diversion ratio. The decision to use a time or cost ratio turns on the problem at hand. Transit agencies developed diversion curves for different kinds of situations, so variables like income and population density entered implicitly.

Diversion curves are based on empirical observations, and their improvement has resulted from better (more and more pointed) data. Curves are available for many markets. It is not difficult to obtain data and array results. Expansion of transit has motivated data development by operators and planners. Yacov Zahavi's UMOT studies contain many examples of diversion curves.

In a sense, diversion curve analysis is expert system analysis. Planners could "eyeball" neighborhoods and estimate transit ridership by routes and time of day. Instead, diversion is observed empirically and charts can be drawn.

Examples

Example 1: Mode Choice Model

You are given this mode choice model

$$U_{ijm} = -0.412(C_c/w) - 0.0201 * C_{ivt} - 0.0531 * C_{ovt} - 0.89 * D_1 - 1.783D_3 - 2.15D_4 \quad (3.2.21)$$

Where:

- C_c/w = cost of mode (cents) / wage rate (in cents per minute)
- C_{ivt} = travel time in-vehicle (min)
- C_{ovt} = travel time out-of-vehicle (min)
- D = mode specific dummies: (dummies take the value of 1 or 0)
  - D_1 = driving,
  - D_2 = transit with walk access, [base mode]
  - D_3 = transit with auto access,
  - D_4 = carpool

With these inputs:

	Driving	**Walk Connect Transit**	**Auto Connect Transit**	**Carpool**
t = travel time in-vehicle (min)	10	30	15	12
t0 = travel time out-of-vehicle (min)	0	15	10	3
D_1 = driving,	1	0	0	0
D_2 = transit with walk access, [base mode]	0	1	0	0
D_3 = transit with auto access,	0	0	1	0
D_4 = carpool	0	0	0	1
COST	25	100	100	150
WAGE	60	60	60	60

What are the resultant mode shares?

Solution

Outputs	**1: Driving**	**2: Walk Connect Transit**	**3: Auto Connect Transit**	**4: Carpool**	**Sum**
Utilities	-1.26	-2.09	-3.30	-3.58	

EXP(V)	0.28	0.12	0.04	0.03	
P(V)	59.96%	26.31%	7.82%	5.90%	100%

Interpretation

Value of Time:

\[.0411/2.24=$0.0183/*min*=$1.10/hour\]

(in 1967 $, when the wage rate was about $2.85/hour)

implication, if you can improve the travel time (by more buses, less bottlenecks, e.g.) for less than $1.10/hour/person, then it is socially worthwhile.

Example 2: Mode Choice Model Interpretation

What mode would a perfectly rational, perfectly informed traveler choose in a deterministic world given these facts:

Case 1

'	**Bus**	**Car**	**Parameter**
Tw	10 min	5 min	-0.147
Tt	40 min	20 min	-0.0411
C	$2	$1	-2.24

Car always wins (independent of parameters as long as all are < 0)

Case 2

'	**Bus**	**Car**	**Parameter**
Tw	5 min	5 min	-0.147
Tt	40 min	20 min	-0.0411
C	$2	$4	-2.24
Results	-6.86	-10.51	

Sample Problem

Problem 1

You are given the following mode choice model.

$$U_{ijm} = -1C_{ijm} + 5D_T \tag{3.2.22}$$

Where:

- C_{ijm} = travel cost between i and j by mode m
- D_T = dummy variable (alternative specific constant) for transit

and Travel Times

Auto Travel Times

Origin/Destination	Dakotopolis	New Fargo
Dakotopolis	5	7
New Fargo	7	5

Transit Travel Times

Origin/Destination	Dakotopolis	New Fargo
Dakotopolis	10	15
New Fargo	15	8

A. Using a logit model, determine the probability of a traveler driving.

B. Using the results from the previous problem (#2), how many car trips will there be?

Answer available at the end of this section.

Problem 2

Prior to the collapse, there were two modes serving the Marcytown-Rivertown corridor: driving alone (d) and carpool (c), which takes advantage of an uncongested carpool lane. The utilities of the modes are as given below.

$U_d = -t_d$

$U_c = -12 - t_c$

where t is the travel time

Assuming a multinomial logit model, and

A) That the congested time by driving alone was 15 minutes and time by carpool was 5 minutes. What was the modeshare prior to the collapse?

B) How would you interpret the constant of -12 in the expression for Uc?

C) After the collapse, because of a shift in travelers from other bridges, the travel time by both modes increased by 12 minutes. What is the post-collapse modeshare?

Answer available at the end of this section.

Additional Questions

Homework

1. Identify five independent variables that you believe affect mode choice. Pose hypotheses about how each variable affects share of travel by each mode.

2. Explain Independence from Irrelevant Alternatives

3. You are given the following mode choice model.

$$U_{ij} = -2C_{ijm} + 5D_T - 5D_W \tag{3.2.23}$$

Where: C_{ijm} = trabel time between i and j by mode m

D_T = dummy variable (alternative specific constant) for transit

D_W = dummy variable (alternative specific constant) for walking

Here auto is assumed the baseline mode.

The auto travel time between zones (in minutes) is given by the following matrix:

Origin \ Destination	**To Dakotopolis**	**To New Fargo**
From Dakotopolis	5	9
From New Fargo	9	4

The transit travel time between zones (in minutes) is given by the following matrix:

Origin \ Destination	**To Dakotopolis**	**To New Fargo**
From Dakotopolis	10	15
From New Fargo	15	8

The walk travel time between zones (in minutes) is given by the following matrix:

Origin \ Destination	**To Dakotopolis**	**To New Fargo**
From Dakotopolis	10	30
From New Fargo	30	10

Using a logit model, what is the probability of a traveler taking transit?

Additional Questions

1. What factors determine mode choice, utility U_{mij}
2. What role does socio-economics play in mode choice?
3. What is an aggregate model, disaggregate model?
4. How is probability (P_m) calculated?
5. What happens to probability when utility increases or decreases?
6. Why is it hard for transportation engineers to determine how many people will be taking one mode?
7. Who developed the logit model for transportation mode choice?
8. What is “kiss and ride”
9. What happens when a new mode is introduced?
10. Explain the red bus/blue bus paradox, and Independence of Irrelevant Alternatives feature of multinomial logit models.
11. Compare the logit model with the gravity model.
12. What is an alternative specific constant?

Variables

- U_{ijm}- Utility of traveling from i to j by mode m
- D_n = mode specific dummies: (dummies take the value of 1 or 0)
- P_m = Probability of mode m
- C_c/w = cost of mode (cents) / wage rate (in cents per minute)
- C_{ivt} = travel time in-vehicle (min)

- C_{ovt} = travel time out-of-vehicle (min)
- D_n = mode specific dummies: (dummies take the value of 1 or 0)

Abbreviations

- WCT - walk connected transit
- ADT - auto connect transit (drive alone/park and ride)
- APT - auto connect transit (auto passenger/kiss and ride)
- AU1 - auto driver (no passenger)
- AU2 - auto 2 occupants
- AU3+ - auto 3+ occupants
- WK/BK - walk/bike
- IIA - Independence of Irrelevant Alternatives

Key Terms

- Mode choice
- Logit
- Probability
- Independence of Irrelevant Alternatives (IIA)
- Dummy Variable (takes value of 1 or 0)

Videos

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Answers

Problem 1

A. Using a logit model, determine the probability of a traveler driving.

Solution Steps

1. Compute Utility for Each Mode for Each Cell
2. Compute Exponentiated Utilities for Each Cell
3. Sum Exponentiated Utilities
4. Compute Probability for Each Mode for Each Cell
5. Multiply Probability in Each Cell by Number of Trips in Each Cell

Auto Utility: U_{auto}

Origin\Destination	Dakotopolis	New Fargo
Dakotopolis	-5	-7
New Fargo	-7	-5

Transit Utility: $U_{transit}$

Origin\Destination	Dakotopolis	New Fargo
Dakotopolis	-5	-10
New Fargo	-10	-3

$e^{U_{auto}}$

Origin\Destination	Dakotopolis	New Fargo
Dakotopolis	0.0067	0.0009
New Fargo	0.0009	0.0067

$e^{U_{transit}}$

Origin\Destination	Dakotopolis	New Fargo
Dakotopolis	0.0067	0.0000454
New Fargo	0.0000454	0.0565

Sum: $e^{U_{transit}} + e^{U_{auto}}$

Origin\Destination	Dakotopolis	New Fargo
Dakotopolis	0.0134	0.0009454
New Fargo	0.0009454	0.0498

P(Auto) = $e^{U_{auto}}/(e^{U_{auto}} + e^{U_{transit}})$

Origin\Destination	Dakotopolis	New Fargo
Dakotopolis	0.5	0.953
New Fargo	0.953	0.12

P(Transit) = $e^{U_{transit}}/(e^{U_{auto}} + e^{U_{transit}})$

Origin\Destination	Dakotopolis	New Fargo
Dakotopolis	0.5	0.047
New Fargo	0.047	0.88

Part B

B. Using the results from the previous problem (#2), how many car trips will there be?

Recall

Total Trips

Origin\Destination	Dakotopolis	New Fargo
Dakotopolis	9395	5606
New Fargo	6385	15665

Total Trips by Auto = $T_{ij} * P(Auto)$

Origin\Destination	Dakotopolis	New Fargo
Dakotopolis	4697	5339
New Fargo	6511	1867

Problem 2

A) That the congested time by driving alone was 15 minutes and time by carpool was 5 minutes. What was the modeshare prior to the collapse?

Ud = -15

Uc = -17

Pd = 0.88

Pc = 0.12

B) How would you interpret the constant of -12 in the expression for Uc?

The constant -12 is an alternative specific constant. In this example, even if the travel time for the drive and carpool modes were the same, the utility of the carpool mode is lesser than the drive alone due to the negative constant -12. This indicates that there is a lesser probability of an individual choosing the carpool mode due to its lower utility.

C) After the collapse, because of a shift in travelers from other bridges, the travel time by both modes increased by 12 minutes. What is the post-collapse modeshare?

In this question since the travel time for both modes increase by the same 12 minutes, the post-collapse mode share will be the same as before

D) The transit agency decides to run a bus to help out the commuters after the collapse. Again assuming the multinomial logit model holds, without knowing how many travelers take the bus, what proportion of travelers on the bus previously took the car? Why? Comment on this result. Does it seem plausible?

The logit model will indicate that 88% of the bus riders previously drove alone due to the underlying Independence of Irrelevant Alternatives (IIA) property. The brief implication of IIA is that when you add a new mode, it will draw from the existing modes in proportion to their existing shares. This doesnt seem plausible since the bus is more likely to draw from the carpool mode than the drive alone mode.

This page titled 3.2: Choice Modeling is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

3.3: Land Use Forecasting

Land use forecasting undertakes to project the distribution and intensity of trip generating activities in the urban area. In practice, land use models are demand driven, using as inputs the aggregate information on growth produced by an aggregate economic forecasting activity. Land use estimates are inputs to the transportation planning process.

The discussion of land use forecasting to follow begins with a review of the Chicago Area Transportation Study (CATS) effort. CATS researchers did interesting work, but did not produce a transferable forecasting model, and researchers elsewhere worked to develop models. After reviewing the CATS work, the discussion will turn to the first model to be widely known and emulated: the Lowry model developed by Ira S. Lowry when he was working for the Pittsburgh Regional Economic Study. Second and third generation Lowry models are now available and widely used, as well as interesting features incorporated in models that are not widely used.

Today, the transportation planning activities attached to metropolitan planning organizations are the loci for the care and feeding of regional land use models. In the US, interest in and use of models is spotty, for most agencies are concerned with short run planning and day-to-day decisions. Interest is higher in Europe and elsewhere.

Even though there isn't much use of full blown land use modeling in the US today, we need to understand the subject: the concepts and analytic tools pretty much define how land use-transportation matters are thought about and handled; there is a good bit of interest in the research community where there have been important developments; and when the next upturn in infrastructure development comes along the present models will form the starting place for work.

Land Use Analysis at the Chicago Area Transportation Study

In brief, the CATS analysis of the 1950s was "by mind and hand" distribute growth. The product was maps developed with a rule-based process. The rules by which land use was allocated were based on state-of-the art knowledge and concepts, and it hard to fault CATS on those grounds. The CATS took advantage of Colin Clark's extensive work on the distribution of population densities around city centers. Theories of city form were available, sector and concentric circle concepts, in particular. Urban ecology notions were important at the University of Chicago and University of Michigan. Sociologists and demographers at the University of Chicago had begun its series of neighborhood surveys with an ecological flavor. Douglas Carroll, the CATS director, had studied with Amos Hawley, an urban ecologist at Michigan.

Stylized Urban Density Gradient

Colin Clark studied the population densities of many cities, and he found traces similar to those in the figure. Historic data show how the density line has changed over the years. To project the future, one uses changes in the parameters as a function of time to project the shape of density in the future, say in 20 years. The city spreads glacier-like. The area under the curve is given by population forecasts.

The CATS did extensive land use and activity surveys, taking advantage of the City work done by the Chicago Planning Commission. Hock's work forecasting activities said what the land uses-activities were that would be accommodated under the density curve. Existing land use data were arrayed in cross section. Land uses were allocated in a manner consistent with the existing pattern.

The study area was divided into transportation analysis zones: small zones where there was a lot of activity, larger zones elsewhere. The original CATS scheme reflected its Illinois State connections. Zones extended well away from the city. The zones were defined to take advantage of Census data at the block and minor civil division levels. They also strove for homogeneous land use and urban ecology attributes.

The first land use forecasts at CATS arrayed developments using "by hand" techniques, as stated. We do not fault the "by hand" technique – the then state of computers and data systems forced it. It was a rule based land use allocation. Growth was the forcing function, as were inputs from the economic study. Growth said that the population density envelope would have to shift. The land uses implied by the mix of activities were allocated from "Where is the land available?" and "What's the use now?" Considerations. Certain types of activities allocate easily: steel mills, warehouses, etc.

Conceptually, the allocation rules seem important. There is lot of spatial autocorrelation in urban land uses; it's driven by historical path dependence: this sort of thing got started here and seeds more of the same. This autocorrelation was lost somewhat in the step from "by hand" to analytic models.

The CATS procedure was not viewed with favor by the emerging Urban Transportation Planning professional peer group, and in the late 1950s there was interest in the development of analytic forecasting procedures. At about the same time, similar interests emerged to meet urban redevelopment and sewer planning needs, and interest in analytic urban analysis emerged in political science, economics, and geography.

Lowry Model

Flowchart of Lowry Model

Hard on the heels of the CATS work, several agencies and investigators began to explore analytic forecasting techniques, and between 1956 and the early 1960s a number of modeling techniques evolved. Irwin (1965) provides a review of the status of emerging models. One of the models, the Lowry model, was widely adopted.

Supported at first by local organizations and later by a Ford Foundation grant to the RAND Corporation, Ira S. Lowry undertook a three-year study in the Pittsburgh metropolitan area. (Work at RAND will be discussed later.) The environment was data rich, and there were good professional relationships available in the emerging emphasis on location and regional economies in the Economics Department at the University of Pittsburgh under the leadership of Edgar M. Hoover. The structure of the Lowry model is shown on the flow chart.

The flow chart gives the logic of the Lowry model. It is demand driven. First, the model responds to an increase in basic employment. It then responds to the consequent impacts on service activities. As Lowry treated his model and as the flow chart indicates, the model is solved by iteration. But the structure of the model is such that iteration is not necessary.

Although the language giving justification for the model specification is an economic language and Lowry is an economist, the model is not an economic model. Prices, markets, and the like do not enter.

A review of Lowry's publication will suggest reasons why his approach has been widely adopted. The publication was the first full elaboration of a model, data analysis and handling problems, and computations. Lowry's writing is excellent. He is candid and discusses his reasoning in a clear fashion. One can imagine an analyst elsewhere reading Lowry and thinking, "Yes, I can do that."

The diffusion of innovations of the model is interesting. Lowry was not involved in consulting, and his word of mouth contacts with transportation professionals were quite limited. His interest was and is in housing economics. Lowry did little or no "selling." We learn that people will pay attention to good writing and an idea whose time has come.

The model makes extensive use of gravity or interaction decaying with distance functions. Use of "gravity model" ideas was common at the time Lowry developed his model; indeed, the idea of the gravity model was at least 100 years old at the time. It was under much refinement at the time of Lowry's work; persons such as Alan Voorhees, Mort Schneider, John Hamburg, Roger Creighon, and Walter Hansen made important contributions. (See Carrothers 1956).

The Lowry Model provided a point of departure for work in a number of places. Goldner (1971) traces its impact and modifications made. Steven Putnam at the University of Pennsylvania used it to develop PLUM (Projective Land Use Model) and I(incremental)PLUM. We estimate that Lowry derivatives are used in most MPO studies, but most of today's workers do not recognize the Lowry heritage, the derivatives are one or two steps away from the mother logic.

Penn-Jersey Model

Flowchart of Penn-Jersey land use forecasting model

The P-J (Penn-Jersey, greater Philadelphia area) analysis had little impact on planning practice. It will now be discussed, even so, because it illustrates what planners might have done, given available knowledge building blocks. It is an introduction to some of the work by researchers who are not practicing planners.

The P-J study scoped widely for concepts and techniques. It scoped well beyond the CATS and Lowry efforts, especially taking advantage of things that had come along in the late 1950s. It was well funded and viewed by the State and the Bureau of Public Roads as a research and a practical planning effort. Its Director's background was in public administration, and leading personnel

were associated with the urban planning department at the University of Pennsylvania. The P-J study was planning and policy oriented.

The P-J study drew on several factors "in the air". First, there was a lot of excitement about economic activity analysis and the applied math that it used, at first, linear programming. T. J. Koopmans, the developer of activity analysis, had worked in transportation. There was pull for transportation (and communications) applications, and the tools and interested professionals were available.

There was work on flows on networks, through nodes, and activity location. Orden (1956) had suggested the use of conservation equations when networks involved intermediate modes; flows from raw material sources through manufacturing plants to market were treated by Beckmann and Jacob Marschak (1955) and Goldman (1958) had treated commodity flows and the management of empty vehicles.

Maximal flow and synthesis problems were also treated (Boldreff 1955, Gomory and Hu 1962, Ford and Fulkerson 1956, Kalaba and Juncosa 1956, Pollack 1964). Balinski (1960) considered the problem of fixed cost. Finally, Cooper (1963) considered the problem of optimal location of nodes. The problem of investment in link capacity was treated by Garrison and Marble (1958) and the issue of the relationship between the length of the planning time-unit and investment decisions was raised by Quandt (1960) and Pearman (1974).

A second set of building blocks was evolving in location economics, regional science, and geography. Edgar Dunn (1954) undertook an extension of the classic von Thünen analysis of the location of rural land uses. Also, there had been a good bit of work in Europe on the interrelations of economic activity and transportation, especially during the railroad deployment era, by German and Scandinavian economists. That work was synthesized and augmented in the 1930's by August Lösch, and his The Location of Economic Activities was translated into English during the late 1940s. Edgar Hoover's work with the same title was also published in the late 1940s. Dunn's analysis was mainly graphical; static equilibrium was claimed by counting equations and unknowns. There was no empirical work (unlike Garrison 1958). For its time, Dunn's was a rather elegant work.

William Alonso's (1964) work soon followed. It was modeled closely on Dunn's and also was a University of Pennsylvania product. Although Alonso's book was not published until 1964, its content was fairly widely known earlier, having been the subject of papers at professional meetings and Committee on Urban Economics (CUE) seminars. Alonso's work became much more widely known than Dunn's, perhaps because it focused on "new" urban problems. It introduced the notion of bid rent and treated the question of the amount of land consumed as a function of land rent.

Wingo (1961) was also available. It was different in style and thrust from Alonso and Dunn's books and touched more on policy and planning issues. Dunn's important, but little noted, book undertook analysis of location rent, the rent referred to by Marshall as situation rent. Its key equation was:

$$R - Y(P - c) - Ytd \tag{3.3.1}$$

where: R = rent per unit of land, P = market price per unit of product, c = cost of production per unit of product, d = distance to market, and t = unit transportation cost.

In addition, there were also demand and supply schedules.

This formulation by Dunn is very useful, for it indicates how land rent ties to transportation cost. Alonso's urban analysis starting point was similar to Dunn's, though he gave more attention to market clearing by actors bidding for space.

The question of exactly how rents tied to transportation was sharpened by those who took advantage of the duality properties of linear programming. First, there was a spatial price equilibrium perspective, as in Henderson (1957, 1958) Next, Stevens (1961) merged rent and transportation concepts in a simple, interesting paper. In addition, Stevens showed some optimality characteristics and discussed decentralized decision-making. This simple paper is worth studying for its own sake and because the model in the P-J study took the analysis into the urban area, a considerable step.

Stevens 1961 paper used the linear programming version of the transportation, assignment, translocation of masses problem of Koopmans, Hitchcock, and Kantorovich. His analysis provided an explicit link between transportation and location rent. It was quite transparent, and it can be extended simply. In response to the initiation of the P-J study, Herbert and Stevens (1960) developed the core model of the P-J Study. Note that this paper was published before the 1961 paper. Even so, the 1961 paper came first in Stevens' thinking.

The Herbert-Stevens model was housing centered, and the overall study had the view that the purpose of transportation

investments and related policy choices was to make Philadelphia a good place to live. Similar to the 1961 Stevens paper, the model assumed that individual choices would lead to overall optimization.

The P-J region was divided into *u* small areas recognizing *n* household groups and *m* residential bundles. Each residential bundle was defined on the house of apartment, the amenity level in the neighborhood (parks, schools, etc.), and the trip set associated with the site. There is an objective function:

$$maxZ = \sum_{k=1}^{u}\sum_{i=1}^{n}\sum_{h=1}^{m} x_{ih}^{k}(b_{ih} - c_{ih}^{k}) \tag{3.3.2}$$

$$x_{ih}^{k} \geq 0 \tag{3.3.3}$$

wherein x_{ihk} is the number of households in group *i* selecting residential bundle *h* in area *k*. The items in brackets are b_{ih} (the budget allocated by *i* to bundle *h*) and c_{ihk}, the purchase cost of *h* in area *k*. In short, the sum of the differences between what households are willing to pay and what they have to pay is maximized; a surplus is maximized. The equation says nothing about who gets the surplus: it is divided between households and those who supply housing in some unknown way. There is a constraint equation for each area limiting the land use for housing to the land supply available.

$$\sum_{i=1}^{n}\sum_{h=1}^{m} s_{ih}x_{ih}^{k} \leq L^{k} \tag{3.3.4}$$

where: s_{ih} = land used for bundle *h* L_k = land supply in area *k*

And there is a constraint equation for each household group assuring that all folks can find housing.

$$\sum_{k=1}^{u}\sum_{h=1}^{m} x_{ih}^{k} = N_i \tag{3.3.5}$$

where: N_i = number of households in group *i*

A policy variable is explicit, the land available in areas. Land can be made available by changing zoning and land redevelopment. Another policy variable is explicit when we write the dual of the maximization problem, namely:

$$minZ' = \sum_{k=1}^{u} r^{k}L^{k} + \sum_{i=1}^{n} v_i(-N_i) \tag{3.3.6}$$

Subject to:

$$s_{ih}r^{k} - v_i \geq b_{ih} - c_{ih}^{k} \tag{3.3.7}$$

$$r^{k} \geq 0 \tag{3.3.8}$$

The variables are r^k (rent in area *k*) and v_i an unrestricted subsidy variable specific to each household group. Common sense says that a policy will be better for some than others, and that is reasoning behind the subsidy variable. The subsidy variable is also a policy variable because society may choose to subsidize housing budgets for some groups. The constraint equations may force such policy actions.

It is apparent that the Herbert-Stevens scheme is a very interesting one. Its also apparent that it is housing centered, and the tie to transportation planning is weak. That question is answered when we examine the overall scheme for study, the flow chart of a single iteration of the model. How the scheme works requires little study. The chart doesn't say much about transportation. Changes in the transportation system are displayed on the chart as if they are a policy matter.

The word "simulate" appears in boxes five, eight, and nine. The P-J modelers would say, "We are making choices about transportation improvements by examining the ways improvements work their way through urban development. The measure of merit is the economic surplus created in housing."

Academics paid attention to the P-J study. The Committee on Urban Economics was active at the time. The committee was funded by the Ford Foundation to assist in the development of the nascent urban economics field. It often met in Philadelphia for review of the P-J work. Stevens and Herbert were less involved as the study went along. Harris gave intellectual leadership, and

he published a fair amount about the study (1961, 1962). However, the P-J influence on planning practice was nil. The study didn't put transportation up front. There were unsolvable data problems. Much was promised but never delivered. The Lowry model was already available.

Kain Model

Figure - Causal arrow diagram illustrating Kain's econometric model for transportation demand

About 1960, the Ford Foundation made a grant to the RAND Corporation to support work on urban transportation problems (Lowry's work was supported in part by that grant). The work was housed in the logistics division of RAND, where the economists at RAND were housed. The head of that division was then Charles Zwick, who had worked on transportation topics previously.

The RAND work ranged from new technology and the cost of tunneling to urban planning models and analyses with policy implications. Some of the researchers at RAND were regular employees. Most, however, were imported for short periods of time. The work was published in several formats: first in the RAND P series and RM series and then in professional publications or in book form. Often, a single piece of work is available in differing forms at different places in the literature.

In spite of the diversity of topics and styles of work, one theme runs through the RAND work – the search for economic policy guides. We see that theme in Kain (1962), which is discussed by de Neufville and Stafford, and the figure is adapted from their book.

Kain's model dealt with direct and indirect affects. Suppose income increases. The increase has a direct effect on travel time and indirect affects through the use of land, auto ownership, and choice of mode. Work supported at RAND also resulted in Meyer, Kain and Wohl (1964). These parts of the work at RAND had considerable influence on subsequent analysis (but not so much on practice as on policy). John Meyer became President of the National Bureau of Economic Research and worked to refocus its lines of work. Urban analysis Kain-style formed the core of a several-year effort and yielded book length publications (see, e.g., G. Ingram, et al., The NBER Urban Simulation Model, Columbia Univ. Press, 1972). After serving in the Air Force, Kain moved to Harvard, first to redirect the Urban Planning Department. After a time, he relocated at the Kennedy School, and he, along with José A. Gómez-Ibáñez, John Meyer, and C. Ingram, lead much work in an economic-policy analysis style. Martin Wohl moved on from RAND, eventually, to Carnegie-Mellon University, where he continued his style of work (e.g. Wohl 1984).

Policy Oriented Gaming

The notion that the impact of policy on urban development might be simulated was the theme for a conference at Cornell in the early 1960s; collegiums were formed, several streams of work emerged. Several persons developed rather simple (from today's view) simulation games. Land use development was the outcome of gravitational type forces and the issue faced was that of conflicts between developers and planners when planners intervened in growth. CLUG and METROPOLIS are two rather well known products from this stream of work (they were the SimCity of their day); there must be twenty or thirty other similar planner vs. developer in the political context games. There seems to have been little serious attempt to analyze use of these games for policy formulation and decision-making, except for work at the firm Environmetrics.

Peter House, one of the Cornell Conference veterans, established Environmetrics early in the 1960s. It, too, started with relatively simple gaming ideas. Over about a ten-year period, the comprehensiveness of gaming devices was gradually improved and, unlike the other gaming approaches, transportation played a role in their formulation. Environmetrics' work moved into the Environmental Protection Agency and was continued for a time at the EPA Washington Environmental Studies Center.

A model known as River Basin was generalized to GEM (General Environmental Assessment Model) and then birthed SEAS (Strategic Environmental Assessment Model) and SOS (Son of SEAS). There was quite a bit of development as the models were generalized, too much to be discussed here.

The most interesting thing to be noted is change in the way the use of the models evolved. Use shifted from a "playing games" stance to an "evaluate the impact of federal policy" stance. The model (both equations and data) is viewed as a generalized city or cities. It responds to the question: What would be the impact of proposed policies on cities?

An example of generalized question answering is LaBelle and Moses (1983) La Belle and Moses implement the UTP process on typical cities to assess the impact of several policies. There is no mystery why this approach was used. House had moved from the EPA to the DOE, and the study was prepared for his office.

University of North Carolina

A group at Chapel Hill, mainly under the leadership of Stuart Chapin, began its work with simple analysis devices somewhat similar to those used in games. Results include Chapin (1965), Chapin and H. C. Hightower (1966) and Chapin and Weiss (1968). That group subsequently focused on (1) the ways in which individuals make tradeoffs in selecting residential property, (2) the roles of developers and developer decisions in the urban development process, and (3) information about choices obtained from survey research. Lansing and Muller (1964 and 1967) at the Survey Research Center worked in cooperation with the Chapel Hill Group in developing some of this latter information.

The first work was on simple, probabilistic growth models. It quickly moved from this style to game-like interviews to investigate preferences for housing. Persons interviewed would be given “money” and a set of housing attributes – sidewalks, garage, numbers of rooms, lot size, etc. How do they spend their money? This is an early version of the game The Sims. The work also began to examine developer behavior, as mentioned. (See: Kaiser 1972).

Reviews and Surveys

In addition to reviews at CUE meetings and sessions at professional meetings, there have been a number of organized efforts to review progress in land use modeling. An early effort was the May 1965 issue of the Journal of the American Institute of Planners edited by B. Harris. The next major effort was a Highway Research Board Conference in June, 1967 (HRB 1968) and this was most constructive. This reference contains a review paper by Lowry, comments by Chapin, Alonso, and others. Of special interest is Appendix A, which listed several ways that analysis devices had been adapted for use. Robinson (1972) gives the flavor of urban redevelopment oriented modeling. And there have been critical reviews (e.g. Brewer 1973, Lee 1974). Pack (1978) addresses agency practice; it reviews four models and a number of case studies of applications. (See also Zettel and Carll 1962 and Pack and Pack 1977). The discussion above has been limited to models that most affected practice (Lowry) and theory (P-J, etc.) there are a dozen more that are noted in reviews. Several of those deal with retail and industry location. There are several that were oriented to urban redevelopment projects where transportation was not at issue.

Discussion

Lowry-derived land use analysis tools reside in the MPOs. The MPOs also have a considerable data capability including census tapes and programs, land use information of varied quality, and survey experiences and survey-based data. Although large model work continues, fine detail analysis dominates agency and consultant work in the US. One reason is the requirement for environmental impact statements. Energy, noise, and air pollution have been of concern, and techniques special to the analysis of these topics have been developed. Recently, interest has increased in the uses of developer fees and/or other developer transportation related actions. Perceived shortages for funds for highways and transit are one motive for extracting resources or actions from developers. There’s also the long-standing ethic that those who occasion costs should pay. Finally, there is a small amount of theoretical or academic work. Small is the operative word. There are few researchers and the literature is limited.

The discussion to follow will first emphasize the latter, theory-oriented work. It will then turn to a renewed interest in planning models in the international arena. Modern behavioral, academic, or theory-based analysis of transportation and land use date from about 1965. By modern we mean analysis that derives aggregate results from micro behavior. First models were Herbert-Stevens in character. Similar to the P-J model, they:

- Treated land as the constraining resource and land use choices given land rent variations as the critical behavior.
- Imagined roles for policy makers.
- Emphasized residential land uses and ignored interdependencies in land uses.
- Used closed system, comparative statics ways of thinking.
- And gave no special attention to transportation.

There have been three major developments subsequently:

1. Consideration of transportation activities and labor and capital inputs in addition to land inputs,
2. Efforts to use dynamic, open system ways of thinking, and
3. Inquiry into how micro choice behavior yields macro results.

The Herbert-Stevens model was not a behavioral model in the sense that it did not try to map from micro to macro behavior. It did assume rational, maximizing behavior by locators. But that was attached to macro behavior and policy by assumed some

centralized authority that provided subsidies. Wheaton (1974) and Anderson (1982) modified the Herbert-Stevens approach in different, but fairly simple, ways to deal with the artificiality of the Herbert-Stevens formulation.

An alternative to the P-J, Herbert-Stevens tradition was seeded when Edwin S. Mills, who is known as the father of modern urban economics, took on the problem of scoping more widely. Beginning with Mills (1972), Mills has developed a line of work yielding more publications and follow on work by others, especially his students.

Using a Manhattan geometry, Mills incorporated a transportation component in his analysis. Homogeneous zones defined by the transportation system were analyzed as positioned x integer steps away from the central zone via the Manhattan geometry. Mills treated congestion by assigning integer measures to levels of service, and he considered the costs of increasing capacity. To organize flows, Mills assumed a single export facility in the central node. He allowed capital-land rent trade offs yielding the tallest buildings in the central zones.

Stating this in a rather long but not difficult to understand linear programming format, Mills' system minimizes land, capital, labor, and congestion costs, subject to a series of constraints on the quantities affecting the system. One set of these is the exogenously gives vector of export levels. Mills (1974a,b) permitted exports from non-central zones, and other modifications shifted the ways congestion is measured and allowed for more than one mode of transport.

With respect to activities, Mills introduced an input-output type coefficient for activities; aqrs, denotes land input q per unit of output r using production technique s. T.J. Kim (1979) has followed the Mills tradition through the addition of articulating sectors. The work briefly reviewed above adheres to a closed form, comparative statics manner of thinking. This note now will turn to dynamics.

The literature gives rather varied statements on what consideration of dynamics means. Most often, there is the comment that time is considered in an explicit fashion, and analysis becomes dynamic when results are run out over time. In that sense, the P-J model was a dynamic model. Sometimes, dynamics are operationalized by allowing things that were assumed static to change with time. Capital gets attention. Most of the models of the type discussed previously assume that capital is malleable, and one considers dynamics if capital is taken as durable yet subject to ageing – e.g., a building once built stays there but gets older and less effective. On the people side, intra-urban migration is considered. Sometimes too, there is an information context. Models assume perfect information and foresight. Let's relax that assumption.

Anas (1978) is an example of a paper that is "dynamic" because it considers durable capital and limited information about the future. Residents were mobile; some housing stock was durable (outlying), but central city housing stock was subject to obsolescence and abandonment.

Persons working in other traditions tend to emphasize feedbacks and stability (or the lack of stability) when they think "dynamics," and there is some literature reflecting those modes of thought. The best known is Forester (1968), which set off an enormous amount of critique and some follow on thoughtful extensions (e.g., Chen (ed), 1972)

Robert Crosby in the University Research Office of the US DOT was very much interested in the applications of dynamics to urban analysis, and when the DOT program was active some work was sponsored (Kahn (ed) 1981). The funding for that work ended, and we doubt if any new work was seeded.

The analyses discussed use land rent ideas. The direct relation between transportation and land rent is assumed, e.g., as per Stevens. There is some work that takes a less simple view of land rent. An interesting example is Thrall (1987). Thrall introduces a consumption theory of land rent that includes income effects; utility is broadly considered. Thrall manages both to simplify analytic treatment making the theory readily accessible and develop insights about policy and transportation.

Wachs summarizes Lee's (1973) "Requiem for Large Scale Models."

Note

Lee couched his argument, as many of you will recall, in terms of what he called the seven deadly sins of modeling. The seven sins were:

1) Hypercomprehensiveness: Meaning that the models tried to replicate too complex a system in a single shot, and were expected to serve too many different purposes at the same time.

> 2) Grossness: In a way, the converse of hypercomprehensiveness. Even though they tried to do too much and serve too many purposes, their results or outputs were too coarse and aggregate, too simplistic to be useful for complicated and sophisticated policy requirements.
>
> 3) Data Hungriness: Even to produce, gross outputs (a few variables), the models required us to input many variables for many geographic units, and from at least several time periods in order to produce approximate projections, and very often we could not afford the data collection efforts needed to run the models. In other instances, data simply didn't exist at the levels of specificity which would be appropriate to run them.
>
> 4) Wrongheadedness: Lee meant that the models suffered from substantial and largely unrecognized deviations between the behavior claimed for them and the variables and equations which actually determined their behavior. As an example, when regional averages were used to calibrate models, but forecasts were made for local areas, the models deviated from reality because of specification errors which were often not even recognized by their users.
>
> 5) Complicatedness: Even though when you looked at them through one set of lenses the models seemed terribly simplistic, when looked at through another set of lenses they were outrageously complex. Too simplistic in replicating urban economic and social processes, the models were too complex in their computational algorithms. Errors were multiplied because there were so many equations, spatial units, and time periods. Even the theoretical notion of the model or its representation of an urban process was grossly simplistic compared with reality. Often, the user didn't know how the errors were propagated through series of sequential operations; and sometimes we needed to use systematic adjustments or "correction factors" to make the models more realistic even though we did not completely comprehend the sources of all the errors and could not interpret the correction factors in real-world terms.
>
> 6) Mechanicalness: Lee meant that we routinely went through many steps in a modeling process without completely understanding why we did so, and without fully comprehending the consequences in terms of validity or error magnification. He stated, for example, that even rounding errors could be compounded beyond reasonable bounds by mechanical steps taken to calibrate and apply many models without the user's knowledge.
>
> 7) Expensiveness: The costs of the models, derived from their grossness, data hungriness, complicatedness, and so on, placed them beyond the financial means of many agencies, or depleted the resources of agencies so much that the very use of models precluded having the resources available to improve them or to fine tune them to make them appropriate to their applications.
>
> Lee argued in 1973 that the models should be improved in four ways:
>
> 1) Models should be made more transparent to users and policymakers.
>
> 2) Models should combine strong theoretical foundations, objectives information, and wisdom or good judgment. Without these elements, they remain exercises in empty-headed empiricism, abstract theorizing, or false consciousness of what is actually going on in our urban areas.
>
> 3) We should start with problems and match our methods to the needs of particular situations, gathering no more information and using no more modeling complexity than is really needed.
>
> 4) We should build the simplest models possible, since complex models do not work well, and certainly are unlikely to be understood by those who are asked to act on the basis of the model outputs.
>
> – *Martin Wachs, Keynote Address: Evolution and Objectives of the Travel Model Improvement Program, in Travel Model Improvement Program Conference Proceedings August 14–17, 1994, edited by Shunk, Gordon and Bass, Patricia*

John Landis, has responded about seven challenges facing large scale models:

1. Models - microbehavioral (actors and agents) ... Social Benefit/Social Action
2. Simulation - multiple movies/scenarios
3. Respond to constraints and investments
4. Nonlinearity - path dependence in non-artifactual way (structure and outcomes, network effects)
5. spatial vs. real autocorrelation, emergence - new dynamics, threshold network effects
6. preference utility diversity and change over time

7. Useful beyond calibration periods. Embed innovators and norming agents. Strategic and response function.

This page titled 3.3: Land Use Forecasting is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

3.4: Trip Generation

Trip Generation is the first step in the conventional four-step transportation forecasting process (followed by Destination Choice, Mode Choice, and Route Choice), widely used for forecasting travel demands. It predicts the number of trips originating in or destined for a particular traffic analysis zone.

Every trip has two ends, and we need to know where both of them are. The first part is determining how many trips originate in a zone and the second part is how many trips are destined for a zone. Because land use can be divided into two broad category (residential and non-residential) we have models that are household based and non-household based (e.g. a function of number of jobs or retail activity).

For the residential side of things, trip generation is thought of as a function of the social and economic attributes of households (households and housing units are very similar measures, but sometimes housing units have no households, and sometimes they contain multiple households, clearly housing units are easier to measure, and those are often used instead for models, it is important to be clear which assumption you are using).

At the level of the traffic analysis zone, the language is that of land uses "producing" or *attracting* trips, where by assumption trips are "produced" by households and "attracted" to non-households. Production and attractions differ from origins and destinations. Trips are produced by households even when they are returning home (that is, when the household is a destination). Again it is important to be clear what assumptions you are using.

Activities

People engage in activities, these activities are the "purpose" of the trip. Major activities are home, work, shop, school, eating out, socializing, recreating, and serving passengers (picking up and dropping off). There are numerous other activities that people engage on a less than daily or even weekly basis, such as going to the doctor, banking, etc. Often less frequent categories are dropped and lumped into the catchall "Other".

Every trip has two ends, an origin and a destination. Trips are categorized by *purposes*, the activity undertaken at a destination location.

Observed trip making from the Twin Cities (2000-2001) Travel Behavior Inventory by Gender

Trip Purpose	Males	Females	Total
Work	4008	3691	7691
Work related	1325	698	2023
Attending school	495	465	960
Other school activities	108	134	242
Childcare, daycare, after school care	111	115	226
Quickstop	45	51	96
Shopping	2972	4347	7319
Visit friends or relatives	856	1086	1942
Personal business	3174	3928	7102
Eat meal outside of home	1465	1754	3219
Entertainment, recreation, fitness	1394	1399	2793
Civic or religious	307	462	769
Pick up or drop off passengers	1612	2490	4102

With another person at their activities	64	48	112

Some observations:

- Men and women behave differently on average, splitting responsibilities within households, and engaging in different activities,
- Most trips are not work trips, though work trips are important because of their peaked nature (and because they tend to be longer in both distance and travel time),
- The vast majority of trips are not people going to (or from) work.

People engage in activities in sequence, and may chain their trips. In the Figure below, the trip-maker is traveling from home to work to shop to eating out and then returning home.

HomeWorkShopEat.png

Specifying Models

How do we predict how many trips will be generated by a zone? The number of trips originating from or destined to a purpose in a zone are described by trip rates (a cross-classification by age or demographics is often used) or equations. First, we need to identify what we think the relevant variables are.

Home-end

The total number of trips leaving or returning to homes in a zone may be described as a function of:

$$T_h = f(housing\ units, household\ size, age, income, accessibility, vehicle\ ownership) \tag{3.4.1}$$

Home-End Trips are sometimes functions of:

- Housing Units
- Household Size
- Age
- Income
- Accessibility
- Vehicle Ownership
- Other Home-Based Elements

Work-end

At the work-end of work trips, the number of trips generated might be a function as below:

$$T_w = f(jobs(area\ of\ space\ by\ type, occupancy\ rate \tag{3.4.2}$$

Work-End Trips are sometimes functions of:

- Jobs
- Area of Workspace
- Occupancy Rate
- Other Job-Related Elements

Shop-end

Similarly shopping trips depend on a number of factors:

$$T_s = f(number\ of\ retail\ workers, type\ of\ retail, area, location, competition) \tag{3.4.3}$$

Shop-End Trips are sometimes functions of:

- Number of Retail Workers
- Type of Retail Available
- Area of Retail Available
- Location
- Competition
- Other Retail-Related Elements

Input Data

A forecasting activity conducted by planners or economists, such as one based on the concept of economic base analysis, provides aggregate measures of population and activity growth. Land use forecasting distributes forecast changes in activities across traffic zones.

Estimating Models

Which is more accurate: the data or the average? The problem with averages (or aggregates) is that every individual's trip-making pattern is different.

Home-end

To estimate trip generation at the home end, a cross-classification model can be used. This is basically constructing a table where the rows and columns have different attributes, and each cell in the table shows a predicted number of trips, this is generally derived directly from data.

In the example cross-classification model: The dependent variable is trips per person. The independent variables are dwelling type (single or multiple family), household size (1, 2, 3, 4, or 5+ persons per household), and person age.

The figure below shows a typical example of how trips vary by age in both single-family and multi-family residence types.

The figure below shows a moving average.

Non-home-end

The trip generation rates for both "work" and "other" trip ends can be developed using Ordinary Least Squares (OLS) regression (a statistical technique for fitting curves to minimize the sum of squared errors (the difference between predicted and actual value) relating trips to employment by type and population characteristics.

The variables used in estimating trip rates for the work-end are Employment in Offices (E_{off}), Retail (E_{ret}), and Other (E_{oth})

A typical form of the equation can be expressed as:

$$T_{D,k} = a_1 E_{off,k} + a_2 E_{oth,k} + a_3 E_{ret,k} \tag{3.4.4}$$

Where:

- $T_{D,k}$ - Person trips attracted per worker in Zone k
- $E_{off,i}$ - office employment in the ith zone
- $E_{oth,i}$ - other employment in the ith zone
- $E_{ret,i}$- retail employment in the ith zone
- a_1, a_2, a_3 - model coefficients

Normalization

For each trip purpose (e.g. home to work trips), the number of trips originating at home must equal the number of trips destined for work. Two distinct models may give two results. There are several techniques for dealing with this problem. One can either assume one model is correct and adjust the other, or split the difference.

It is necessary to ensure that the total number of trip origins equals the total number of trip destinations, since each trip interchange by definition must have two trip ends.

The rates developed for the home end are assumed to be most accurate,

The basic equation for normalization:

$$T'_{D,j} = T_{D,j} \frac{\sum i = 1^{I} T_{O,i}}{\sum j = 1^{J} T_{TD,j}} \tag{3.4.5}$$

Sample Problems

Problem 1:

Planners have estimated the following models for the AM Peak Hour

$T_{O,i} = 1.5 * H_i$

$T_{D,j} = (1.5 * E_{off,j}) + (1 * E_{oth,j}) + (0.5 * E_{ret,j})$

Where:

$T_{O,i}$ = Person Trips Originating in Zone i

$T_{D,j}$ = Person Trips Destined for Zone j

H_i = Number of Households in Zone i

You are also given the following data

Data

Variable	Dakotopolis	New Fargo
H	10000	15000
E_{off}	8000	10000
E_{oth}	3000	5000
E_{ret}	2000	1500

A. What are the number of person trips originating in and destined for each city?

B. Normalize the number of person trips so that the number of person trip origins = the number of person trip destinations. Assume the model for person trip origins is more accurate.

Answer available at the end of this section.

Problem 2:

Modelers have estimated that the number of trips leaving Rivertown (T_O) is a function of the number of households (H) and the number of jobs (J), and the number of trips arriving in Marcytown (T_D) is also a function of the number of households and number of jobs.

$T_O = 1H + 0.1J; R^2 = 0.9$

$T_D = 0.1H + 1J; R^2 = 0.5$

Assuming all trips originate in Rivertown and are destined for Marcytown and:

Rivertown: 30000 H, 5000 J

Marcytown: 6000 H, 29000 J

Determine the number of trips originating in Rivertown and the number destined for Marcytown according to the model.

Which number of origins or destinations is more accurate? Why?

Answer available at the end of this section.

Problem 3

Modelers have estimated that in the AM peak hour, the number of trip origins (T_O) is a function of the number of households (H) and the number of jobs (J), and the number of trip destinations (T_D) is also a function of the number of households and number of jobs.

$T_O = 1.0H + 0.1J; R^2 = 0.9$

$T_D = 0.1H + 1J; R^2 = 0.5$

Suburbia: 30000 H, 5000 J

Urbia: 6000 H, 29000 J

1) Determine the number of trips originating in and destined for Suburbia and for Urbia according to the model.

2) Does this result make sense? Normalize the result to improve its accuracy and sensibility?

3) If the interzonal travel cost (from Suburbia to Urbia or Urbia to Suburbia) is 10 minutes, and the intrazonal travel cost (trips within Suburbia or within Urbia) is 5 minutes, use a (doubly-constrained) gravity model wherein the impedance is $f(t_{ij}) = t_{ij}^{-2}$ $f(t_{ij}) = t_{ij}^{-2}$, calculate the impedance, balance the matrix to match trip generation, and determine the trip table (within 2 percent accuracy)

Variables

- $T_{O,i}$ - Person trips originating in Zone i
- $T_{D,j}$ - Person Trips destined for Zone j
- $T_{O,i'}$ - Normalized Person trips originating in Zone i
- $T_{D,j'}$ - Normalized Person Trips destined for Zone j
- T_h - Person trips generated at home end (typically morning origins, afternoon destinations)

- T_w - Person trips generated at work end (typically afternoon origins, morning destinations)
- T_s - Person trips generated at shop end
- H_i - Number of Households in Zone i
- $E_{off,k}$ - office employment in Zone k
- $E_{ret,k}$ - retail employment in Zone k
- $E_{oth,k}$ - other employment in Zone k
- B_n - model coefficients

Abbreviations

- H2W - Home to work
- W2H - Work to home
- W2O - Work to other
- O2W - Other to work
- H2O - Home to other
- O2H - Other to home
- O2O - Other to other
- HBO - Home based other (includes H2O, O2H)
- HBW - Home based work (H2W, W2H)
- NHB - Non-home based (O2W, W2O, O2O)

External Exercises

Use the ADAM software at the STREET website and try Assignment #1 to learn how changes in analysis zone characteristics generate additional trips on the network.

Additional Problems

Homework

1. Keep a travel diary for the next week. Analyze your personal travel diary. Summarize the number of trips by mode, by purpose, by time of day, etc. To do so, for each trip record the following information
    1. the start and end time (to the nearest minute)
    2. start and end location of each trip,
    3. primary mode you took (drive alone, car driver with passenger, car passenger, bus, LRT, walk, bike, motorcycle, taxi, Zipcar, other). (use the codes provided)
    4. purpose (to work, return home, work related business, shopping, family/personal business, school, church, medical/dental, vacation, visit friends or relatives, other social recreational, other) (use the codes provided)
    5. if you traveled with anyone else, and if so whether they lived in your household or not.

Bonus: Email your professor at the end of everyday with a detailed log of your travel diary. (+5 points on the first exam)

Additional Problems

1. Are number of destinations always less than origins?
2. Pose 5 hypotheses about factors that affect work, non-work trips? How do these factors affect accuracy, and thus normalization?
3. What is the acceptable level of error?
4. Describe one variable used in trip generation and how it affects the model.
5. What is the basic equation for normalization?
6. Which of these models (home-end, work-end) are assumed to be more accurate? Why is it important to normalize trip generation models

7. What are the different trip purposes/types trip generation?
8. Why is it difficult to know who is traveling when?
9. What share of trips during peak afternoon peak periods are work to home (>50%, <50%?), why?
10. What does ORIO abbreviate?
11. What types of employees (ORIO) are more likely to travel from work to home in the evening peak
12. What does the trip rate tell us about various parts of the population?
13. What does the "T-statistic" value tell us about the trip rate estimation?
14. Why might afternoon work to home trips be more or less than morning home to work trips? Why might the percent of trips be different?
15. Define frequency.
16. Why do individuals > 65 years of age make fewer work to home trips?
17. Solve the following problem. You have the following trip generation model:

$$Trips = B_1 Off + B_2 Ind + B_3 Ret \tag{3.4.6}$$

And you are given the following coefficients derived from a regression model.

```
B_1 = 0.61
B_2 = 0.15
B_3 = 0.123
```

If there are 600 office employees, 300 industrial employees, and 200 retail employees, how many trips are going from work to home?

Videos

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Answers

Problem 1:

A. What are the number of person trips originating in and destined for each city?

Solution to Trip Generation Problem Part A

	Households H_i	Office Employees E_{off}	Other Employees E_{oth}	Retail Employees E_{ret}	Origins $T_{O,i} = 1.5 * H_i$	Destinations $T_{D,j} = (1.5 * E_{off,j}) + (1$

Dakotopolis	10000	8000	3000	2000	15000	16000
New Fargo	15000	10000	5000	1500	22500	20750
Total	25000	18000	8000	3000	37500	36750

B. Normalize the number of person trips so that the number of person trip origins = the number of person trip destinations. Assume the model for person trip origins is more accurate.

Use:

$$T'_{D,j} = T_{D,j} \frac{\sum i = 1^{I} T_{O,i}}{\sum j = 1^{J} T_{TD,j}} => T_{D,j} \frac{37500}{36750} = T_{D,j} * 1.0204 \tag{3.4.7}$$

Solution to Trip Generation Problem Part B

	Origins $T_{O,i}$	Destinations $T_{D,j}$	Adjustment Factor	Normalized Destinations $T'_{D,j}$	Rounded
Dakotopolis	15000	16000	1.0204	16326.53	16327
New Fargo	22500	20750	1.0204	21173.47	21173
Total	37500	36750	1.0204	37500	37500

Problem 2:

Determine the number of trips originating in Rivertown and the number destined for Marcytown according to the model.

T_Rivertown =T_O ; T_O= 1(30000) + 0.1(5000) = 30500 trips

T_(MarcyTown)=T_D ; T_D= 0.1(6000) + 1(29000) = 29600 trips

Which number of origins or destinations is more accurate? Why?

Origins(T_{Rivertown}) because of the goodness of fit measure of the Statistical model (R^2=0.9).

This page titled 3.4: Trip Generation is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

3.5: Destination Choice

Everything is related to everything else, but near things are more related than distant things. - Waldo Tobler's 'First Law of Geography'

Destination Choice (or **trip distribution** or **zonal interchange analysis**), is the second component (after Trip Generation, but before Mode Choice and Route Choice) in the traditional four-step transportation forecasting model. This step matches tripmakers' origins and destinations to develop a "trip table", a matrix that displays the number of trips going from each origin to each destination. Historically, trip distribution has been the least developed component of the transportation planning model.

Table: Illustrative Trip Table

Origin \ Destination	**1**	**2**	**3**	**Z**
1	T_{11}	T_{12}	T_{13}	T_{1Z}
2	T_{21}			
3	T_{31}			
Z	T_{Z1}			T_{ZZ}

Where: T_{ij} = Trips from origin *i* to destination *j*.

Work trip distribution is the way that travel demand models understand how people take jobs. There are trip distribution models for other (non-work) activities, which follow the same structure.

Fratar Models

The simplest trip distribution models (Fratar or Growth models) simply extrapolate a base year trip table to the future based on growth,

$$T_{ij,y+1} = g * T_{ij,y} \tag{3.5.1}$$

where:

- $T_{ij,y}$ - Trips from i to j in year y
- g - growth factor

Fratar Model takes no account of changing spatial accessibility due to increased supply or changes in travel patterns and congestion.

Gravity Model

The gravity model illustrates the macroscopic relationships between places (say homes and workplaces). It has long been posited that the interaction between two locations declines with increasing (distance, time, and cost) between them, but is positively associated with the amount of activity at each location (Isard, 1956). In analogy with physics, Reilly (1929) formulated Reilly's law of retail gravitation, and J. Q. Stewart (1948) formulated definitions of demographic gravitation, force, energy, and potential, now called accessibility (Hansen, 1959). The distance decay factor of $distance^{-1}$ has been updated to a more comprehensive function of generalized cost, which is not necessarily linear - a negative exponential tends to be the preferred form. In analogy with Newton's law of gravity, a gravity model is often used in transportation planning.

The gravity model has been corroborated many times as a basic underlying aggregate relationship (Scott 1988, Cervero 1989, Levinson and Kumar 1995). The rate of decline of the interaction (called alternatively, the impedance or friction factor, or the utility or propensity function) has to be empirically measured, and varies by context.

Limiting the usefulness of the gravity model is its aggregate nature. Though policy also operates at an aggregate level, more accurate analyses will retain the most detailed level of information as long as possible. While the gravity model is very successful in explaining the choice of a large number of individuals, the choice of any given individual varies greatly from the predicted value. As applied in an urban travel demand context, the disutilities are primarily time, distance, and cost, although discrete choice models with the application of more expansive utility expressions are sometimes used, as is stratification by income or auto ownership.

Mathematically, the gravity model often takes the form:

$$T_{ij} = r_i s_j T_{O,i} T_{D,j} F(C_{ij}) \tag{3.5.2}$$

$$\sum j T_{ij} = T_{O,i}, \sum i T_{ij} = T_{D,j} \tag{3.5.3}$$

$$r_i = (\sum j s_j T_{D,j} f(C_{ij}))^{-1} \tag{3.5.4}$$

$$s_j = (\sum i r_i T_{O,i} f(C_{ij}))^{-1} \tag{3.5.5}$$

where

- T_{ij} = Trips between origin i and destination j
- $T_{O,i}$ = Trips originating at i
- $T_{D,j}$ = Trips destined for j
- C_{ij} = travel cost between i and j
- r_i, s_j = balancing factors solved iteratively.
- f = impedance or distance decay factor

It is doubly constrained so that Trips from i to j equal number of origins and destinations.

Balancing a matrix

Balancing a matrix can be done using what is called the **Furness Method**, summarized and generalized below.

1. Assess Data, you have $T_{O,i}$,$T_{D,j}$, C_{ij}

2. Compute $f(C_{ij})$, e.g.

$$f(C_{ij}) = C_{ij}^{-2} \tag{3.5.6}$$

$$f(C_{ij}) = e^{\beta C_{ij}} \tag{3.5.7}$$

3. Iterate to Balance Matrix

(a) Multiply Trips from Zone $i(T_i)$ by Trips to Zone $j(T_j)$ by Impedance in Cell $ij(f(C_{ij})$ for all ij

(b) Sum Row Totals $T'_{O,i}$, Sum Column Totals $T'_{D,j}$

(c) Multiply Rows by $NO, i = T_{O,i}/T'_{O,i}$

(d) Sum Row Totals $T'_{O,i}$, Sum Column Totals $T'_{D,j}$

(e) Compare $T_{O,i}$ and $T'_{O,i}$, $T_{D,j}$ $T'_{D,j}$ if within tolerance stop, Otherwise go to (f)

(f) Multiply Columns by $N_{D,j} = T_{D,j}/T'_{D,j}$

(g) Sum Row Totals $T'_{O,i}$, Sum Column Totals $T'_{D,j}$

(h) Compare $T_{O,i}$ and $T'_{O,i}$, $T_{D,j}$ and $T'_{D,j}$ if within tolerance stop, Otherwise go to (b)

Issues

Feedback

One of the key drawbacks to the application of many early models was the inability to take account of congested travel time on the road network in determining the probability of making a trip between two locations. Although Wohl noted as early as 1963 research into the feedback mechanism or the “interdependencies among assigned or distributed volume, travel time (or travel ‘resistance’) and route or system capacity”, this work has yet to be widely adopted with rigorous tests of convergence or with a so-called “equilibrium” or “combined” solution (Boyce et al. 1994). Haney (1972) suggests internal assumptions about travel time used to develop demand should be consistent with the output travel times of the route assignment of that demand. While small methodological inconsistencies are necessarily a problem for estimating base year conditions, forecasting becomes even more tenuous without an understanding of the feedback between supply and demand. Initially heuristic methods were developed by Irwin and Von Cube (as quoted in Florian et al. (1975)) and others, and later formal mathematical programming techniques were established by Evans (1976).

Feedback and time budgets

A key point in analyzing feedback is the finding in earlier research by Levinson and Kumar (1994) that commuting times have remained stable over the past thirty years in the Washington Metropolitan Region, despite significant changes in household income, land use pattern, family structure, and labor force participation. Similar results have been found in the Twin Cities by Barnes and Davis (2000).

The stability of travel times and distribution curves over the past three decades gives a good basis for the application of aggregate trip distribution models for relatively long term forecasting. This is not to suggest that there exists a constant travel time budget.

In terms of time budgets:

- 1440 Minutes in a Day
- Time Spent Traveling: ~ 100 minutes + or -
- Time Spent Traveling Home to Work: 20 – 30 minutes + or -

Research has found that auto commuting times have remained largely stable over the past forty years, despite significant changes in transportation networks, congestion, household income, land use pattern, family structure, and labor force participation. The stability of travel times and distribution curves gives a good basis for the application of trip distribution models for relatively long term forecasting.

Examples

Example 1: Solving for impedance

You are given the travel times between zones, compute the impedance matrix $f(C_{ij})$, assuming $f(C_{ij}) = C_{ij}^{-2}$.

Travel Time OD Matrix (C_{ij})

Origin Zone	Destination Zone 1	Destination Zone 2
1	2	5
2	5	2

Compute impedances ($f(C_{ij})$)

Solution

Impedance Matrix ($f(C_{ij})$)

Origin Zone	Destination Zone 1	Destination Zone 2
1	$\frac{1}{2^2} = 0.25$	$\frac{1}{5^2} = 0.04$
2	$\frac{1}{5^2} = 0.04$	$\frac{1}{2^2} = 0.25$

Example 2:

You are given the travel times between zones, trips originating at each zone (zone1 =15, zone 2=15) trips destined for each zone (zone 1=10, zone 2 = 20) and asked to use the classic gravity model $f(C_{ij}) = C_{ij}^{-2}$

Travel Time OD Matrix (C_{ij})

Origin Zone	Destination Zone 1	Destination Zone 2
1	2	5
2	5	2

Solution

(a) Compute impedances ($f(C_{ij})$)

Impedance Matrix ($f(C_{ij})$)

Origin Zone	Destination Zone 1	Destination Zone 2
1	0.25	0.04
2	0.04	0.25

(b) Find the trip table

Balancing Iteration 0 (Set-up)

Origin Zone	Trips Originating	Destination Zone 1	Destination Zone 2
Trips Destined		10	20
1	15	0.25	0.04
2	15	0.04	0.25

Balancing Iteration 1 ($T_{ij,iteration1} = T_{O,i}T_{D,j}f(C_{ij})$)

Origin Zone	Trips Originating	Destination Zone 1	Destination Zone 2	Row Total $T'_{O,i}$	Normalizing Factor $N_{O,i} = T_{O,i}/T'_{O,i}$

Trips Destined		10	20		
1	15	37.50	12	49.50	0.303
2	15	6	75	81	0.185
Column Total		43.50	87		

Balancing Iteration 2 (\(T_{ij,iteration2}=T_{ij,iteration1}*N_{O,i,iteration1}\)**)**

Origin Zone	**Trips Originating**	**Destination Zone 1**	**Destination Zone 2**	**Row Total** $T'_{O,i}$	**Normalizing Factor** $N_{O,i} = T_{O,i}/T'_{O,i}$
Trips Destined		10	20		
1	15	11.36	3.64	15.00	1.00
2	15	1.11	13.89	15.00	1.00
Column Total		12.47	17.53		
Normalizing Factor $N_{D,j} = T_{D,j}/T'_{D,j}$		0.802	1.141		

Balancing Iteration 3 (\(T_{ij,iteration3}=T_{ij,iteration2}*N_{D,j,iteration2}\)**)**

Origin Zone	**Trips Originating**	**Destination Zone 1**	**Destination Zone 2**	**Row Total** $T'_{O,i}$	**Normalizing Factor** $N_{O,i} = T_{O,i}/T'_{O,i}$
Trips Destined		10	20		
1	15	9.11	4.15	13.26	1.13
2	15	0.89	15.85	16.74	0.90
Column Total		10.00	20.00		
Normalizing Factor = $N_{D,j} = T_{D,j}/T'_{D,j}$		1.00	1.00		

Balancing Iteration 4 (\(T_{ij,iteration4}=T_{ij,iteration3}*N_{O,i,iteration3}\)**)**

Origin Zone	**Trips Originating**	**Destination Zone 1**	**Destination Zone 2**	**Row Total** $T'_{O,i}$	**Normalizing Factor** $N_{O,i} = T_{O,i}/T'_{O,i}$
Trips Destined		10	20		
1	15	10.31	4.69	15.00	1.00
2	15	0.80	14.20	15.00	1.00
Column Total		11.10	18.90		
Normalizing Factor = $N_{D,j} = T_{D,j}/T'_{D,j}$		0.90	1.06		

Balancing Iteration 16 (\(T_{ij,iteration16}=T_{ij,iteration15}*N_{D,i,iteration5}\))

Origin Zone	Trips Originating	Destination Zone 1	Destination Zone 2	Row Total $T'_{O,i}$	Normalizing Factor $N_{O,i} = T_{O,i}/T'_{O,i}$
Trips Destined		10	20		
1	15	9.39	5.61	15.00	1.00
1	15	9.39	5.61	15.00	1.00
Column Total		10.01	19.99		
Normalizing Factor = $N_{D,j} = T_{D,j}/T'_{D,j}$		1.00	1.00		

So while the matrix is not strictly balanced, it is very close, well within a 1% threshold, after 16 iterations. The threshold refers to the proximity of the normalizing factor to 1.0.

Additional Questions

Homework

1. Identify five independent variables that you believe affect trip generation. Pose hypotheses about how each variable affects number of trips generated.

2. Identify three different types of trip distribution models. Which one includes the most information? Which one is most common?

3. You are given the following situation: The towns of Saint Cloud and Minneapolis, separated by 110 km as the crow flies, are to be connected by a railroad, a freeway, and a rural highway. Answer the following questions related to this problem

Trip Generation and Distribution

Your planners have estimated the following models for the AM Peak Hour

$$T_{O,i} = 500 + 0.5 * HH_i \tag{3.5.8}$$

$$T_{D,j} = 250 + 0.5 * OFFEMP_j + 0.355 * OTHEMP_j + 0.094 * RETEMP_j \tag{3.5.9}$$

Where: $T_{O,i}$ = Person Trips Originating in Zone *i*

$T_{D,j}$ =Person Trips Destined for Zone *j*

HH_i = Number of Households in Zone *i*

$OFFEMP_j$ =Office Employees in Zone *j*

$OTHEMP_j$ =Other Employees in Zone *j*

$RETEMP_j$Retail Employees in Zone *j*

Your are also given the following data

	Saint Cloud	**Minneapolis**

HH_i	47,604	476,040
$OFFEMP_j$	33,675	336,750
$OTHEMP_j$	14,500	145,000
$RETEMP_j$	22,000	220,000

The travel time between zones (in minutes) is given by the following matrix:

	To Minneapolis	**To Saint Cloud**
From Minneapolis	5	70
From Saint Cloud	70	5

(a) (10) What are the number of AM peak hour person trips originating in and destined for Saint Cloud and Minneapolis.

(b) (10) Assuming the origins are more accurate, normalize the number of destination trips for Saint Cloud and Minneapolis.

(c) (10) Assume a gravity model where the impedance $f(C_{ij}) = C_{ij}^{-2}$. Estimate the proportion of trips that go from Saint Cloud to Minneapolis. Solve your matrix within 5 percent of a balanced solution.

Additional Questions

1. What are the different kinds of models for trip distribution? How do they differ?
2. What factors do conventional Trip Distribution models neglect?
3. In a day, how many minutes are spent traveling for the average person, traveling to work? Why might this be stable, not stable?
4. Do travelers have a travel time tolerance or a travel time budget?
5. What affects travel impedance?
6. If impedance increases, will willingness to travel increase or decrease? Is a person more likely to travel to a closer place or a farther place?
7. Why does willingness to travel have a negative exponential form?
8. Briefly describe the gravity model? How might the gravity model be extended to depend on more than just size (opportunities) and distance (impedance)?
9. Why is balancing a matrix an iterative procedure?
10. What importance does impedance play in balancing iterations
11. Give two functions of impedance (f(Cij)) used in gravity models.
12. How do you calculate impedance between zones?
13. What is congested travel time?
14. Conventional trip distribution models are estimated for which mode?
15. How is trip distribution applied?
16. What can a trip distribution curve tell you about maximum and average trip times and willingness to take trips?
17. What factors are multiplied to give a resulting trip distribution curve

Variables

- $T_{O,i}$ - Trips leaving origin i
- $T_{D,j}$ - Trips arriving at destination {\displaystyle j}j
- $T'_{D,j}$ - Effective Trips arriving at destination j, computed as a result for calibration to the next iteration
- T_{ij} - Total number of trips between origin i and destination j

- r_i - Calibration parameter for Origins
- s_j - Calibration parameter for Destinations
- $f(C_{ij})$ - Cost function between origin i and destination j

Videos

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

This page titled 3.5: Destination Choice is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

3.6: 3-6 Route Choice

Route assignment, **route choice**, or **traffic assignment** concerns the selection of routes (alternative called paths) between origins and destinations in transportation networks. It is the fourth step in the conventional transportation forecasting model, following Trip Generation, Destination Choice, and Mode Choice. The zonal interchange analysis of trip distribution provides origin-destination trip tables. Mode choice analysis tells which travelers will use which mode. To determine facility needs and costs and benefits, we need to know the number of travelers on each route and link of the network (a route is simply a chain of links between an origin and destination). We need to undertake traffic (or trip) assignment. Suppose there is a network of highways and transit systems and a proposed addition. We first want to know the present pattern of travel times and flows and then what would happen if the addition were made.

Link Performance Function

The cost that a driver imposes on others is called the marginal cost. However, when making decisions, a driver only faces his own cost (the average cost) and ignores any costs imposed on others (the marginal cost).

- $$AverageCost = \frac{S_T}{Q} \tag{3.6.1}$$
- $$MarginalCost = \frac{\delta S_T}{\delta Q} \tag{3.6.2}$$

where S_T is the total cost, and Q is the flow.

BPR Link Performance Function

Suppose we are considering a highway network. For each link there is a function stating the relationship between resistance and volume of traffic. The Bureau of Public Roads (BPR) developed a link (arc) congestion (or volume-delay, or link performance) function, which we will term $S_a(Q_a)$

$$S_a(Q_a) = t_a(1 + 0.15\frac{(}{Q_a}c_a)^4) \tag{3.6.3}$$

t_a = *free-flow travel time on link a per unit of time*

Q_a = *flow (or volume) of traffic on link a per unit of time (somewhat more accurately: flow attempting to use link a)*

c_a = *capacity of link a per unit of time*

$S_a(Q_a)$ is the average travel time for a vehicle on link *a*

There are other congestion functions. The CATS has long used a function different from that used by the BPR, but there seems to be little difference between results when the CATS and BPR functions are compared.

Can Flow Exceed Capacity?

On a link, the capacity is thought of as “outflow.” Demand is inflow.

If inflow > outflow for a period of time, there is queueing (and delay).

For Example, for a 1 hour period, if 2100 cars arrive and 2000 depart, 100 are still there. The link performance function tries to represent that phenomenon in a simple way.

Wardrop's Principles of Equilibrium

User Equilibrium

Each user acts to minimize his/her own cost, subject to every other user doing the same. Travel times are equal on all used routes and lower than on any unused route.

System optimal

Each user acts to minimize the total travel time on the system.

Price of Anarchy

The reason we have congestion is that people are selfish. The cost of that selfishness (when people behave according to their own interest rather than society's) is the *price of anarchy*.

The ratio of system-wide travel time under User Equilibrium and System Optimal conditions.

Price of Anarchy = {\displaystyle UE/SO1}" aria-hidden="true" src="wikimedia.org/api/rest_v1/me...d15c6e66395909">

For a two-link network with linear link performance functions (latency functions), Price of Anarchy is < 4/3.

Is this too much? Should something be done, or is 33% waste acceptable? [The loss may be larger/smaller in other cases, under different assumptions, etc.]

Conservation of Flow

An important factor in road assignment is the conservation of flow. This means that the number of vehicles entering the intersection (link segment) equals the number of vehicles exiting the intersection for a given period of time (except for sources and sinks).

Similarly, the number of vehicles entering the back of the link equals the number exiting the front (over a long period of time).

Auto assignment

Long-standing techniques

The above examples are adequate for a problem of two links, however real networks are much more complicated. The problem of estimating how many users are on each route is long standing. Planners started looking hard at it as freeways and expressways (motorways) began to be developed. The freeway offered a superior level of service over the local street system and diverted traffic from the local system. At first, diversion was the technique. Ratios of travel time were used, tempered by considerations of costs, comfort, and level of service.

The Chicago Area Transportation Study (CATS) researchers developed diversion curves for freeways versus local streets. There was much work in California also, for California had early experiences with freeway planning. In addition to work of a diversion sort, the CATS attacked some technical problems that arise when one works with complex networks. One result was the Moore algorithm for finding shortest paths on networks.

The issue the diversion approach didn't handle was the feedback from the quantity of traffic on links and routes. If a lot of vehicles try to use a facility, the facility becomes congested and travel time increases. Absent some way to consider feedback, early planning studies (actually, most in the period 1960-1975) ignored feedback. They used the Moore algorithm to determine shortest paths and assigned all traffic to shortest paths. That's called all or nothing assignment because either all of the traffic from *i* to *j* moves along a route or it does not.

The all-or-nothing or shortest path assignment is not trivial from a technical-computational view. Each traffic zone is connected to *n - 1* zones, so there are numerous paths to be considered. In addition, we are ultimately interested in traffic on links. A link may be a part of several paths, and traffic along paths has to be summed link by link.

An argument can be made favoring the all-or-nothing approach. It goes this way: The planning study is to support investments so that a good level of service is available on all links. Using the travel times associated with the planned level of service, calculations indicate how traffic will flow once improvements are in place. Knowing the quantities of traffic on links, the capacity to be supplied to meet the desired level of service can be calculated.

Heuristic procedures

To take account of the affect of traffic loading on travel times and traffic equilibria, several heuristic calculation procedures were developed. One heuristic proceeds incrementally. The traffic to be assigned is divided into parts (usually 4). Assign the first part of the traffic. Compute new travel times and assign the next part of the traffic. The last step is repeated until all the traffic is assigned. The CATS used a variation on this; it assigned row by row in the O-D table.

The heuristic included in the FHWA collection of computer programs proceeds another way.

- Step 0: Start by loading all traffic using an all or nothing procedure.
- Step 1: Compute the resulting travel times and reassign traffic.
- Step 2: Now, begin to reassign using weights. Compute the weighted travel times in the previous two loadings and use those for the next assignment. The latest iteration gets a weight of 0.25 and the previous gets a weight of 0.75.
- Step 3. Continue.

These procedures seem to work “pretty well,” but they are not exact.

Frank-Wolfe algorithm

Dafermos (1968) applied the Frank-Wolfe algorithm (1956, Florian 1976), which can be used to deal with the traffic equilibrium problem.

Equilibrium Assignment

To assign traffic to paths and links we have to have rules, and there are the well-known Wardrop equilibrium (1952) conditions. The essence of these is that travelers will strive to find the shortest (least resistance) path from origin to destination, and network equilibrium occurs when no traveler can decrease travel effort by shifting to a new path. These are termed user optimal conditions, for no user will gain from changing travel paths once the system is in equilibrium.

The user optimum equilibrium can be found by solving the following nonlinear programming problem

$$min \sum_a \int_0^{v_a} S_a(Q_a)\, dx \tag{3.6.4}$$

subject to:

$$Q_a = \sum_i \sum_j \sum_r \alpha_{ij}^{ar} Q_{ij}^r \tag{3.6.5}$$

$$sum_r Q_{ij}^r = Q_{ij} \tag{3.6.6}$$

$$Q_a \geq 0, Q_{ij}^r \geq 0 \tag{3.6.7}$$

where Q_{ij}^r is the number of vehicles on path *r* from origin *i* to destination *j*. So constraint (2) says that all travel must take place: *i = 1 ... n; j = 1 ... n*

α_{ij}^{ar}= 1 if link *a* is on path *r* from *i* to *j* ; zero otherwise.

So constraint (1) sums traffic on each link. There is a constraint for each link on the network. Constraint (3) assures no negative traffic.

Transit assignment

There are also methods that have been developed to assign passengers to transit vehicles. In an effort to increase the accuracy of transit assignment estimates, a number of assumptions are generally made. Examples of these include the following:

- All transit trips are run on a set and predefined schedule that is known or readily available to the users.
- There is a fixed capacity associated with the transit service (car/trolley/bus capacity).

Examples

Example 1

Figure 1: Two Route Network

Solve for the flows on Links *a* and *b* in the Simple Network of two parallel links just shown if the link performance function on link *a*:

$S_a = 5 + 2 * Q_a$

and the function on link *b*:

$S_b = 10 + Q_b$

where total flow between the origin and destination is 1000 trips.

Solution

Time (Cost) is equal on all used routes so $S_a = S_b$

And we have Conservation of flow so, $Q_a + Q_b = Q_o = Q_d = 1000$

$5 + 2 * (1000 - Q_b) = 10 + Q_b$

$1995 = 3Q_b$

$Q_b = 665; Q_a = 335$

Example 2

An example from Eash, Janson, and Boyce (1979) will illustrate the solution to the nonlinear program problem. There are two links from node 1 to node 2, and there is a resistance function for each link (see Figure 1). Areas under the curves in Figure 2 correspond to the integration from 0 to *a* in equation 1, they sum to 220,674. Note that the function for link *b* is plotted in the reverse direction.

$S_a = 15(1 + 0.15(\frac{Q_a}{1000})^4)$

$S_b = 20(1 + 0.15(\frac{Q_a}{3000})^4)$

$Q_a + Q_b = 8000$

Show graphically the equilibrium result.

Solution

Figure 2- Graphical Solution to the Equilibrium Assignment Problem

Figure 3- Allocation of Vehicles not Satisfying the Equilibrium Condition

At equilibrium there are 2,152 vehicles on link *a* and 5,847 on link *b*. Travel time is the same on each route: about 63.

Figure 3 illustrates an allocation of vehicles that is not consistent with the equilibrium solution. The curves are unchanged, but with the new allocation of vehicles to routes the shaded area has to be included in the solution, so the Figure 3 solution is larger than the solution in Figure 2 by the area of the shaded area.

Example 3

Assume the traffic flow from Milwaukee to Chicago, is 15000 vehicles per hour. The flow is divided between two parallel facilities, a freeway and an arterial. Flow on the freeway is denoted Q_f, and flow on the two-lane arterial is denoted Q_a.

The travel time (in minutes) on the freeway (C_f) is given by:

$C_f = 10 + Q_f/1500$

$C_a = 15 + Q_a/1000$

Apply Wardrop's User Equilibrium Principle, and determine the flow and travel time on both routes.

Solution

The travel times are set equal to one another

$C_f = C_a$

$10 + Q_f/1500 = 15 + Q_a/1000$

The total traffic flow is equal to 15000

$Q_f + Q_a = 15000$

$Q_a = 15000 - Q_f$

$10 + Q_f/1500 = 15 + (15000 - Q_f)/1000$

Solve for Q_f

$Q_f = 60000/5 = 12000$

$Q_a = 15000 - Q_f = 3000$

$C_f = 18$

$C_a = 18$

Thought Questions

- How can we get drivers to consider their marginal cost?
- Alternatively: How can we get drivers to behave in a “System Optimal” way?

Sample Problems

Problem 1

Given a flow of six (6) units from origin “o” to destination “r”. Flow on each route ab is designated with Qab in the Time Function. Apply Wardrop's Network Equilibrium Principle (Users Equalize Travel Times on all used routes)

A. What is the flow and travel time on each link? (complete the table below) for Network A

Link Attributes

Link	Link Performance Function	Flow	Time
o-p	$C_{op} = 5 * Q_{op}$		
p-r	$C_{pr} = 25 + Q_{pr}$		
o-q	$C_{oq} = 20 + 2 * Q_{oq}$		
q-r	$C_{qr} = 5 * Q_{qr}$		

B. What is the system optimal assignment?

C. What is the Price of Anarchy?

Answer available at the end of this section.

These four links are really 2 links O-P-R and O-Q-R, because by conservation of flow Qop = Qpr and Qoq = Qqr.

Link Attributes

Link	Link Performance Function	Flow	Time
o-p-r	$C_{opr} = 25 + 6 * Q_{opr}$		
o-q-r	$C_{oqr} = 20 + 7 * Q_{oqr}$		

By Wardrop's Equilibrium Principle, the travel time (cost) on each used route must be equal. Therefore $C_{opr} = C_{oqr}$

OR
$25 + 6 * Q_{opr} = 20 + 7 * Q_{oqr}$

$5 + 6 * Q_{opr} = 7 * Q_{oqr}$

$Q_{oqr} = 5/7 + 6 * Q_{opr}/7$

By the conservation of flow principle

$Q_{oqr} + Q_{opr} = 6$

$Q_{opr} = 6 - Q_{oqr}$

By substitution

\Q_{oqr}=5/7+6/7(6-Q_{oqr})=41/7-6*Q_{oqr}/7\)

$13 * Q_{oqr} = 41$

$Q_{oqr} = 41/13 = 3.15$

$Q_{opr} = 2.84$

Check

$42.01 = 25 + 6(2.84)$

$42.05 = 20 + 7(3.15)$

Check (within rounding error)

Link Attributes

Link	Link Performance Function	Flow	Time
o-p-r	$C_{opr} = 25 + 6 * Q_{opr}$	2.84	42.01
o-q-r	$C_{oqr} = 20 + 7 * Q_{oqr}$	3.15	42.01

or expanding back to the original table:

Link Attributes

Link	Link Performance Function	Flow	Time
o-p	$C_{op} = 5 * Q_{op}$	2.84	14.2
p-r	$C_{pr} = 25 + Q_{pr}$	2.84	27.84
o-q	$C_{oq} = 20 + 2 * Q_{oq}$	3.15	26.3
q-r	$C_{qr} = 5 * Q_{qr}$	3.15	15.75

User Equilibrium: Total Delay = 42.01 * 6 = 252.06

Part B

What is the system optimal assignment?

Conservation of Flow:

$Q_{opr} + Q_{oqr} = 6$

$TotalDelay = Q_{opr}(25 + 6 * Q_{oqr}) + Q_{oqr}(20 + 7 * Q_{oqr})$

Double subscripts: use braces to clarify

Double subscripts: use braces to clarify

$25Q_{opr} + 6Q_{opr}^2 + 372 - 62Q_{opr} - 42Q_{opr} + 7Q_{opr}^2$

$13Q_{opr}^2 - 79Q_{opr} + 372$

Analytic Solution requires minimizing total delay

$\deltaC/\deltaQ = 26Q_{opr} - 79 = 0$

$Q_{opr} = 79/26 - 3.04$

$Q_{oqr} = 6 - Q_{opr} = 2.96$

And we can compute the SO travel times on each path

$C_{opr,SO} = 25 + 6 * 3.04 = 43.24$

$C_{opr,SO} = 20 + 7 * 2.96 = 40.72$

Note that unlike the UE solution, $C_{opr,SO} \backslash g C_{oqr,SO}$

Total Delay = 3.04(25+ 6*3.04) + 2.96(20+7*2.96) = 131.45+120.53= 251.98

Note: one could also use software such as a "Solver" algorithm to find this solution.

Part C

What is the Price of Anarchy?

User Equilibrium: Total Delay =252.06 System Optimal: Total Delay = 251.98

Price of Anarchy = 252.06/251.98 = 1.0003 < 4/3

Problem 2

The Marcytown - Rivertown corridor was served by 3 bridges, according to the attached map. The bridge over the River on the route directly connecting Marcytown and Citytown collapsed, leaving two alternatives, via Donkeytown and a direct. Assume the travel time functions Cij in minutes, Qij in vehicles/hour, on the five links routes are as given.

Marcytown - Rivertown Cmr = 5 + Qmr/1000

Marcytown - Citytown (prior to collapse) Cmc = 5 + Qmc/1000

Marcytown - Citytown (after collapse) Cmr = ∞

Citytown - Rivertown Ccr = 1 + Qcr/500

Marcytown - Donkeytown Cmd = 7 + Qmd/500

Donkeytown - Rivertown Cdr = 9 + Qdr/1000

Also assume there are 10000 vehicles per hour that want to make the trip. If travelers behave according to Wardrops user equilibrium principle.

Answer available at the end of this section.

Variables

- C_T - total cost
- C_k - travel cost on link k
- Q_k - flow (volume) on link k

Abbreviations

- VDF - Volume Delay Function
- LPF - Link Performance Function
- BPR - Bureau of Public Roads
- UE - User Equilbrium
- SO - System Optimal
- DTA - Dynamic Traffic Assignment
- DUE - Deterministic User Equilibrium

- SUE - Stochastic User Equilibrium
- AC - Average Cost
- MC - Marginal Cost

Key Terms

- Route assignment, route choice, auto assignment
- Volume-delay function, link performance function
- User equilibrium
- System optimal
- Conservation of flow
- Average cost
- Marginal cost

External Exercises

Use the ADAM software at the STREET website and try Assignment #3 to learn how changes in network characteristics impact route choice.

Additional Questions

Homework

1. If trip distribution depends on travel times, and travel times depend on the trip table (resulting from trip distribution) that is assigned to the road network, how do we solve this problem (conceptually)?

2. Do drivers behave in a system optimal or a user optimal way? How can you get them to move from one to the other.

3. Identify a mechanism that can ensure the system optimal outcome is achieved in route assignment, rather than the user equilibrium. Why would we want such an outcome? What are the drawbacks to the mechanism you identified?

4. Assume the flow from Dakotopolis to New Fargo, is 5300 vehicles per hour. The flow is divided between two parallel facilities, a freeway and an arterial. Flow on the freeway is denoted Q_f, and flow on the two-lane arterial is denoted Q_r. The travel time on the freeway C_f is given by:

$C_f = 5 + Q_f/1000$

The travel time on the arterial (Cr) is given by

$C_r = 7 + Q_r/500$

(a) Apply Wardrop's User Equilibrium Principle, and determine the flow and travel time on both routes from Dakotopolis to New Fargo.

(b) Solve for the System Optimal Solution and determine the flow and travel time on both routes.

5. Given a flow of 10,000 vehicles from origin to destination traveling on three parallel routes. Flow on each route A, B, or C is designated with Q_a, Q_b, Q_c in the Time Function Respectively. Apply Wardrop's Network Equilibrium Principle (Users Equalize Travel Times on all used routes), and determine the flow on each route.

$T_A = 500 + 20Q_A$

$T_B = 1000 + 10Q_B$

$T_C = 2000 + 30Q_C$

Additional Questions

1. How does average cost differ from marginal cost?
2. How do System Optimal and User Equilibrium travel time differ?
3. Why do we want people to behave in an SO way?
4. How can you get people to behave in an SO way?
5. Who was John Glen Wardrop?
6. What are Wardrop's Two Principles?
7. What does conservation of flow require in route assignment?
8. Can Variable Message Signs be used to encourage System Optimal behavior?
9. What is freeflow travel time?
10. If a problem has more than two routes, where does the extra equation come from?
11. How can you determine if a route is unused?
12. What is the difference between capacity and flow
13. Draw a typical volume-delay function for a deterministic, static user equilibrium assignment.
14. Can Q be negative?
15. What is route assignment?
16. Is it important that the output travel times from route choice be consistent with the input travel times for destination choice and mode choice? Why?

Videos

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Answers

Problem 1

What is the flow and travel time on each link? Complete the table below for Network A:

Link Attributes

Link	Link Performance Function	Flow	Time
o-p	$C_{op} = 5 * Q_{op}$		
p-r	$C_{pr} = 25 + Q_{pr}$		
o-q	$C_{oq} = 20 + 2 * Q_{oq}$		
q-r	$C_{qr} = 5 * Q_{qr}$		

Problem 2

A) Prior to the collapse, how many vehicles used each route?

Route A (Marcytown-Rivertown) = Ca = 5 + Qa/1000

Route B (Marcytown-Citytown-Rivertown) = Cb = 5 + Qb/1000 + 1 + Qb/500 = 6 + 3Qb/1000

Route C (Marcytown-Donkeytown-Rivertown)= Cc = 7 + Qc/500 + 9 + Qc/1000 = 16 + 3Qc/1000

At equilibrium the travel time on all three used routes will be the same: Ca = Cb = Cc

We also know that Qa + Qb + Qc = 10000

Solving the above set of equations will provide the following results:

Qa = 8467;Qb = 2267;Qc = −867

We know that flow cannot be negative. By looking at the travel time equations we can see a pattern.

Even with a flow of 0 vehicles the travel time on route C(16 minutes) is higher than A or B. This indicates that vehicles will choose route A or B and we can ignore Route C.

Solving the following equations:

Route A (Marcytown-Rivertown) = Ca = 5 + Qa /1000

Route B (Marcytown-Citytown-Rivertown) = Cb = 6 + 3Qb /1000

Ca = Cb

Qa + Qb = 10000

We can the following values:

Qa = 7750; Qb = 2250; Qc = 0

B) After the collapse, how many vehicles used each route?

We now have only two routes, route A and C since Route B is no longer possible. We could solve the following equations:

Route A (Marcytown-Rivertown) = Ca = 5 + Qa /1000

Route C (Marcytown- Donkeytown-Rivertown) = Cc = 16 + 3Qc /1000

Ca = Cc

Qa+ Qc= 10000

But we know from above table that Route C is going to be more expensive in terms of travel time even with zero vehicles using that route. We can therefore assume that Route A is the only option and allocate all the 10,000 vehicles to Route A.

If we actually solve the problem using the above set of equations, you will get the following results:

Qa = 10250; Qc = -250

which again indicates that route C is not an option since flow cannot be negative.

C) After the collapse, public officials want to reduce inefficiencies in the system, how many vehicles would have to be shifted between routes? What is the “price of anarchy” in this case?

User Equilibrium

TotalDelayUE =(15)(10,000)=150,000

System Optimal

TotalDelaySO =(Qa)(5+Qa/1000)+(Qc)(16+3Qc/1000)

Using Qa + Qc = 10,000

TotalDelaySO =(Qa2)/250−71Qa+460000

Minimize total delay ∂((Qa2)/250 − 71Qa + 460000)/∂Qa = 0

Qa/125−7 → Qa = 8875 Qc = 1125 Ca = 13,875 Cc = 19,375

TotalDelaySO =144938

Price of Anarchy = 150,000/144,938 = 1.035

This page titled 3.6: 3-6 Route Choice is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

4: TRANSIT

4.1: Transit Demand
4.2: Transit Operations and Capacity
4.3: Network Design and Frequency
4.4: Timetabling and Scheduling
4.5: Traveler Information for Transit

This page titled 4: Transit is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

4.1: Transit Demand

In this module, we present some of the ways in which **transit demand** is measured. In this context, the word *demand* is synonymous with *usage* or other measures of use such as *ridership*. In addition, we discuss common methods of estimating ridership, particularly when service or fare changes are planned. A few notes and caveats about forecasting ridership for new transit systems are also provided.

Measuring transit demand

As was described in the first module, transit use is generally measured through the metrics of unlinked passenger trips and passenger-miles. These are the most commonly reported measures for public transit use in the US, as all agencies that receive federal support are required to report them to the Federal Transit Administration at least annually.

Transit agencies generally collect this information using a common device called a ride check. A sample of the vehicle trips (trips from one route terminal to another) is selected, and a person (the "checker") is placed on board the vehicle for the sampled scheduled trip(s). That person records the location and the number of people who get on (board) and who get off (alight) the vehicle at that location. An example is shown here for a route traveling 11.1 miles from one terminal (stop 1) to another (stop 10).

Stop	Distance (mi)	Boardings	Alightings	Load
1	0	6	0	6
2	0.5	8	0	14
3	1	3	2	15
4	1.6	4	0	19
5	5.4	2	5	16
6	6.4	7	2	21
7	7	3	1	23
8	9.6	2	7	18
9	10.4	1	10	9
10	11.1	0	9	0

To help in visualization, the graph below plots the load by route segment, between stops. This is plotted using the "Distance" and the "Load" columns from the table above.

Transit load

Calculating these demand measures, the total number of unlinked trips is simply the sum of the number of people who board. In this example, the total boardings are 36, giving the total unlinked trips of 36.

The total number of passenger-miles is the combination of the number of persons on board (the "load") between stops, multiplied by the distance between stops. The load is calculated after each stop, adding the number who board and subtracting the number who alight. This number is multiplied by the distance between stops to obtain the total passenger miles. This is just the shaded area in the graph. In this example:

$$Passenger-miles = (6\ passengers)\cdot(0.5\ miles) + (14\ passengers)\cdot(0.5\ miles) + (15\ passengers)\cdot(0.6\ miles) + (19\ passengers)\cdot(3.8\ miles) + (16\ passengers)\cdot(1\ mile) + (21\ passengers)\cdot(0.6\ miles) + (23\ passengers)\cdot(2.6\ miles) + (18\ passengers)\cdot(0.8\ miles) + (9\ passengers)\cdot(0.7\ miles) = 200.3\ passenger-miles$$

Other travel information can also be obtained using various survey instruments, such as origin-destination or on-board surveys, which can capture transfer information, trip purpose, origin, destination, and demographic characteristics.

Simple transit demand estimation methods

One of the more common issues for transit planners is how to estimate demand changes that might occur if various service characteristics or fares are changed. There are a wide variety of tools that could be used to estimate changes in demand, but one must be sure to apply a tool that fits the desired purpose. For a short planning horizon and with only one variable change, elasticities work reasonably well. Because of their widespread use in transit planning, we discuss these extensively here. For a longer time horizon, with multiple variables at play, more sophisticated models may be needed; these are mentioned briefly at the end of this unit.

Elasticities

For transit demand, there are many cases where the analyst seeks to identify the major effects of a service change, or a fare change, over a relatively short time period (6 months, a year). In these cases, one might be interested in looking at a single change (only one variable is changing) over this time, with the assumption that there is little or no effect from all other factors (i.e., "all other things being equal").

An "elasticity" is well-suited for this case. An elasticity is a measure of sensitivity of demand relative to a particular variable *X*. Specifically, an elasticity $e_{D,X}$ is defined as the percent change in demand *D* for a +1% change in some variable *X*. This can be specified by a percent change in demand, divided by the percent change in *X*. Mathematically:

$$e_{D,X} = \frac{\Delta D/D}{\Delta X/X} = \frac{\Delta D}{\Delta X} \cdot \frac{X}{D} \tag{4.1.1}$$

$$As\ \backslash \mathrm{DeltaX} \to 0, e_{D,X} = \frac{\partial D}{\partial X} \cdot \frac{X}{D} \tag{4.1.2}$$

Conceptually, the elasticity indicates the sensitivity of demand to changes in *X*.

- If the elasticity is positive, then an increase in *X* results in an increase in demand.
- If the elasticity is negative, then an increase in *X* results in a decrease in demand.
- The larger the absolute value of the elasticity, the more sensitive the demand is to *X*.
- We say that demand is "elastic" with respect to *X* if the absolute value of the elasticity is greater than 1.0. This occurs when demand changes by more than 1% if *X* changes by 1%.
- We say that demand is "inelastic" with respect to *X* if the absolute value of the elasticity is less than 1.0. This occurs when demand changes by less than 1% if *X* changes by 1%.

For transit elasticities, demand is most commonly measured through unlinked trips. Common variables (the *X*s) that a transit agency might consider changing are: fares; frequencies (the number of vehicles per hour visiting a stop or station); and, different measures of travel time, such as total time spent traveling, or time spent on board, waiting time, time spent walking to or from the stop or station, or time spent making transfers.

One of the common empirical observations is that the values of elasticities tend to be transferable; that is, the values of the elasticities often apply to similar circumstances, but in different locations. This makes values of elasticities useful for understanding transit demand in many urban areas.

An extensive set of reports on transit elasticities and other transit service measures has been compiled by the Transit Cooperative Research Program (TCRP) within the Transportation Research Board. Specific elasticities can be obtained from these reports.[1] However, some general observations can be made:

- Transit fare elasticities are negative and inelastic (e.g., $e_{D,Fare}$ = –0.3). Demand drops with an increase in fares, but the demand does not drop as much, percentage-wise, as the fare increase. So, increasing fares will generally create more revenue for a transit agency.
- Elasticities for frequency are positive and inelastic (e.g., $e_{D,Frequency}$ = +0.5). Demand increases with an increase in service frequency, but not as much, percentage-wise, as the frequency increases.
- Elasticities for travel times are negative, and can be inelastic or elastic. Transit users tend to be less sensitive (inelastic) to on-board travel times or total travel times (e.g., $e_{D,TotalTime}$ = –0.6), but much more sensitive to access and egress time, waiting time, and transfer time (e.g., $e_{D,WaitTime}$ = –1.2).

Elasticities also may differ on how sensitive the target rider population is to the service or fare change. Travelers in the morning and evening peak periods tend to be less sensitive to fare and service changes than those who ride in mid-day and in the evening or overnight. Also, travelers making discretionary travel tend to be more sensitive to fare and service changes than those making required trips for work or school.

Uses of elasticities

The real value of elasticities is that they provide a clear way to estimate demand changes. That is, with a proposed change in *X*, and an elasticity $e_{D,X}$, one can estimate the change in demand *D*. Re-arranging the equation for the elasticity,

$$\frac{\delta D}{D} = e_{D,X} \cdot \frac{\delta X}{X} \tag{4.1.3}$$

The values of *D* and *X* in this equation represent the current conditions, and the value of Δ*X* is the change in the variable *X*. Then, the change in demand Δ*D* can be estimated.

For example, consider a transit agency that is considering raising fares from $1.25 to $1.50. The current number of unlinked trips are 110,000 per day. If we use a fare elasticity of –0.35, what is the expected change in demand?

Answer:

$$\frac{\Delta D}{D} = e_{D,X} \cdot \frac{\Delta X}{X}$$

$$\frac{\Delta D}{110,000\ unlinked\ trips} = -0.35 \cdot \frac{\$1.50 - \$1.25}{\$1.25}$$

$$\Delta D = -7700\ unlinked\ trips$$

So, the transit agency could expect to lose 7700 unlinked trips per day. But, it is easy to show that, with this fare elasticity, they would make more revenue from the riders after this change.

Sample Problem

A transit agency has an average annual demand of 21,650,000 unlinked trips, with a posted fare of $1.25. Transfers are free. If 25% of the unlinked trips are actually transfers, what are the expected ridership and revenue changes from an increase in the fare to $1.50? Assume a fare elasticity of -0.27, and that all other service conditions remain the same.

Solution

If 25% of the unlinked trips are actually transfers, then the real number of paid boardings is only 75% of the 21,650,000 unlinked trips, or 16,237,500 paid (linked) trips. At a fare of $1.25, this brings in (16,237,500 linked trips)*($1.25/linked trip) = $20,296,875.

With a fare elasticity of -0.27, the change in linked trips will be (-0.27)*(20% increase in fare) = -5.4%. So, the change in linked trips is (-5.4%)*(16,237,500), or -876,825 linked trips, yielding a net of 15,360,675 linked trips. With 25% transfers in the unlinked trips, and assuming this fraction does not change, this is equivalent to 20,480,900 unlinked trips.

These trips will bring in ($1.50/linked trip)*(15,360,675 linked trips) = $23,041,012.50. Hence, there is a net decrease in ridership of almost 1.2 million riders per year, but the total revenue will increase by more than $2.8 million.

Advanced demand forecasting methods

When introducing major service changes, or a new service, more sophisticated demand models are needed. This is because such changes induce much larger changes in travelers' behavior, such as possible changes in overall trip-making, origins, destinations, modes of travel, routes, and time of travel (among others). The interested reader is encouraged to consider other chapters related to long-term planning methods, to learn how these models are commonly developed and applied.

Perhaps most controversial in the area of transit demand estimation is the forecasting of future ridership (unlinked trips) on proposed new service. Since the 1970's, many transit agencies have proposed new heavy rail and light rail systems, requiring an assessment of the potential ridership for these systems. In these cases, many forecasts have been overly optimistic in their forecasts of transit ridership, while underestimating the actual costs.

From these forecasting experiences, one might make a few observations:

- There will be errors in any ridership forecast. Some forecasts have strongly over-predicted the actual ridership, while others have under-predicted ridership. Sources of error in these forecasts are many. More transparency about these possible errors, or some sensitivity analysis on assumptions in the models, can be useful in explaining uncertainties.
- In the past, there were incentives to over-predict ridership, to make the rail transit investments seem more cost-effective in attracting riders to transit. This is especially true when the federal share of the capital costs (which often range in the hundreds of millions to billions of dollars) can be up to 80% or more.
- There are a large number of factors that can be considered in developing new rail transit systems. Attracting new riders to public transit remains only one of many possible factors that could influence the political decisions on new transit systems. Specifically, political, economic, and social welfare considerations can often weigh heavily in the decision to develop new rail transit systems.

Related books

Vukan R. Vuchic (2005). Urban Transit: Operations, Planning, and Economics. Hoboken: John Wiley and Sons.

This page titled 4.1: Transit Demand is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

4.2: Transit Operations and Capacity

From the supply side, most public transit systems operate service that has some common characteristics. The most salient of these are the *vehicle cycle* and the notion of *capacity*. Later, we address some operating characteristics of specific modes, notably separating rail modes from fixed-route bus service.

Concepts in operations and capacity

The vehicle cycle

It is useful to refer to the operation of a vehicle (van, bus, or train) through the course of a day of transit service; this is commonly referred to as the vehicle "cycle" because it tends to repeat itself from one day to the next. This cycle is illustrated in the figure below.

Caption The Vehicle Cycle

- The cycle begins when a vehicle is started from a depot (a garage, yard, or other location). The depot serves as a common location where vehicles are stored or maintained.
- The vehicle is positioned from the depot to a location where it can begin service. This is commonly at the terminus of a route. Such a movement, from the depot to this location, is called a "pull-out".
- The vehicle travels from its starting location to another route terminus, stopping at stations or stops along the route to allow passengers to board and to alight. We will call such movement of the vehicle a "trip" (or, vehicle trip).
- While moving along the route, the vehicle incurs both running time and dwell time. Running time is the time spent traveling between stops or stations, and dwell time is the time spent stopped at locations to allow passengers to board and alight.
- While moving along the route, the vehicle is engaged in "revenue service", so that time and miles spent along the route while providing passenger service are called "revenue hours" and "revenue miles".
- When it reaches a route terminus, the vehicle is re-positioned for further service. If it returns along the same route or another route starting from the same terminus, there may be a short time for recovery before reentering revenue service, called a "layover".
- A vehicle may also be moved between termini to start service on a different route, resulting in what is called a "deadhead" trip, not in revenue service, between the two termini.
- The vehicle continues in revenue service on the fixed routes, repeating the process of stopping at stations or stops to allow passengers to board and alight.
- When the vehicle has reached its final terminus for its set of trips, it returns to the depot. Such movement from a route terminus to the depot is called a "pull-in".

This vehicle cycle is common to fixed-route service, particularly for bus and rail transit systems. In the case of demand-responsive service, there are no formal "termini" for a route; rather, the "termini" represent specific locations where persons are picked up or dropped off. At any point, if the vehicle becomes empty, then it may dead-head to the next pick-up location, or it may return to the depot.

Capacity

Capacity in transit operations is measured as the maximum number of passengers that can be carried past a single point on a fixed route, in a given period of time. The most common measure of capacity is in terms of passengers per hour.

In transit operations, we define the time between vehicles past a given point as the *headway* (or time headway), usually in minutes. The inverse of the headway is called the *frequency*, essentially capturing the number of vehicles per unit time past a certain point on a route, usually measured in vehicles per hour.

$$Frequency\ (veh/hr)\ f = \frac{60\ min/hr}{h} \tag{4.2.1}$$

$$Headway\ (in\ min)\ h = \frac{60\ min/hr}{f} \tag{4.2.2}$$

As an example, a route with a 10-minute headway has a frequency of 6 vehicles per hour.

With this in mind, the capacity of a transit route is given as the product of the frequency and the maximum number of persons per vehicle:

Capacity (passengers/hr) on the route = $C_{route} = f\cdotN_{car}\cdotC_{car}$

where: f = frequency on the route (veh/hr)

N_{car} = number or cars per train (=1 for bus)

C_{car} = maximum number of persons per car (or bus)

As examples, a bus route with a frequency of 6 buses/hr and a maximum of 55 persons on each bus has a capacity of 330 passengers per hour. A rail line with a frequency of 15 trains/hr (4 min headways) with 6 cars per train and a maximum of 150 persons per car has a capacity of 13,500 passengers per hour.

Capacity is a very important concept in comparing different transit modes.

Bus operations

Bus operations generally follow the traditional vehicle cycle presented above. Operation of each bus is under control of an operator (the driver). That operator is responsible for operating the bus safely along the route, as well as managing passenger boarding and alighting at stops along the route. They may also have responsibility for managing fare payment on board, and for ensuring passenger safety and security.

One characteristic that is fairly common in bus operations is the notion of “hail-stop” operations. In practice, this means that a bus needs to stop at a stop or station only if a passenger wishes to board or alight; otherwise, the bus may simply bypass the stop at running speed. In this concept, passengers wishing to board or to alight must signal these intentions to the operator, so that he/she knows whether it is necessary to stop at the next stop or station.

In some cases, the bus may stop at all stops; this is more common where the bus has only a limited number of stops and/or high volumes of passengers boarding and alighting at these stops.

Capacity on a bus route is the product of the route’s frequency and the number of passengers on each bus. To get higher capacity, one may increase the frequency of buses or increase the passenger-carrying capacity of each bus, or both.

- In the case of frequency, the primary limitation on very high-frequency service (e.g., higher than one bus per minute) is the limitation on the number of berths for buses to stop at each station. Buses must stop for dwell time to allow passenger boarding and alighting at a stop, and the number of buses that can access the stop during that time depends on how many berths there are. If the average dwell time at a stop is 20 seconds, and the bus stop has only one berth, then the highest frequency of buses serving that stop would be 3 buses in a minute. Of course, reducing dwell times (e.g., boarding and alighting at multiple doors, eliminating paying the fare while boarding) allows for higher capacities in such cases. Also, having multiple berths, especially on route segments where many routes overlap, can also allow for higher frequency service.
- In the case of bus capacity, agencies may move away from a standard, single-level, 40-foot bus to higher-capacity buses, such as double-decked buses or articulated buses. These bus types may have significantly higher passenger loads than a standard bus.

Finally, when buses operate in mixed traffic, it is well known that maintaining a schedule can be a challenge. Signal timing, traffic congestion, traffic incidents, and other factors can disrupt the expected running times; the accommodation of passengers with bicycles or wheelchairs may similarly result in longer dwell times at stops. As a result, poor schedule adherence is a common challenge to service reliability. In many cases, transit agencies will allow for some “slack” in the normal schedule, including extra time in the schedule at various points along a route, and in the layover time at terminals, to allow any buses running behind schedule to catch up again.

It is well-known that higher-frequency bus service is prone to “bunching”: a preceding bus may fall behind schedule, and the subsequent bus runs ahead of schedule, until they form a pair of buses operating as one. To combat poor schedule adherence and bus bunching, bus operations may also include elements of real-time control. Such measures can include: (1) holding buses at stops to allow more equal headways; (2) encouraging drivers to speed up or slow down between stops to maintain the schedule;

(3) in more extreme cases, allowing buses to skip stops or to “short-turn” (i.e., reverse direction on a route at a location other than the terminus) to improve schedule adherence or provide more capacity when the passenger demand surges on a certain segment of a route.

Rail operations

Rail vehicles (trains) also follow the vehicle cycle shown earlier. Usually, however, the trains stop at every station on the route. Also, in many cases, fare collection is performed in the station, allowing all doors of the train to be used for passenger boarding and alighting.

Rail systems can run in a grade-separated environment with its own guideway and no conflicting traffic, in a mixed traffic environment with other road vehicles, or in some combination of these two environments. As may be obvious, having a grade-separated environment allows trains to operate with few limitations, allowing higher speeds and much more reliable service (with good schedule adherence and minimal travel time variability). Once rail becomes subject to mixed traffic, speeds can drop and service often becomes less reliable.

Rail systems most commonly operate on electric power being transmitted to the vehicle through either what is called a “third rail”, located just off the ground near the existing rails, or through an overhead catenary.

The operation of rail systems is usually controlled by signal systems which indicate appropriate operating conditions to the train operator, either through signals on the wayside or through signals in the train cab. Traditionally, these signal systems are arranged according to fixed segments of track called “blocks”. These blocks are defined in such a way as to control for safe operation of the train system: no two trains can be located in a single block at the same time. This may be left to the control of the train operator, or an automated or semi-automated system can be employed to enforce these block restrictions.

In a “fixed-block” signal system, the signals have indications with various colors and patterns that prescribe the proper train operation. In the system illustrated below, a red signal is used to indicate stop, a yellow signal to indicate to reduce speed for a red signal ahead, and a green signal to indicate clear to proceed. The most common logic dictating signal the signal indication is based on the occupancy of the block(s) ahead of the signal. In the case below, red indicates the subsequent block is occupied; yellow that the subsequent signal is red; and green that the two subsequent blocks are not occupied. As trains move from block to block (from time t to t+1 to t+2), the indications in the signal system change accordingly.

Caption Illustration of Rail Signaling

The capacity of a rail line is based on the frequency of trains, the number of cars per train, and the maximum number of persons per car.

- The frequency is usually constrained at the upper end by the amount of time a train occupies a block. Since two trains cannot occupy the same block, the longest time to traverse a block becomes the minimum headway between trains; the highest frequency is then the inverse of the minimum headway. To increase frequency (and hence increase capacity), one must reduce the block length or determine other ways to control vehicle operation.
- Usually, the number of cars per train is limited by the geometry of the station platforms, or of other route geometric features (e.g., the distance between two cross-streets in mixed operation). When the train stops at a station, all cars allow passengers to board and alight, meaning that all cars must fit on the platform. So, adding cars to trains is only possible if all stations on the line can allow the cars to fit on the platform.
- Car capacity depends in part on the length of the car and the arrangement of seating and standing space within the car. Greater standing space means more capacity, although passengers will be less comfortable.

Additionally, in electric rail systems, the number of trains and cars per train can in practice be limited by the electric power capacity of the distribution system, because each additional car increases the electrical load.

Finally, some rail systems have fully automated or semi-automated operation, where the train speed may be controlled through an automated system. Signals are transmitted through the rails or through the power system to the train motors, indicating the appropriate speed for a given block.

Sample Problem

A transit agency has an existing 8.5-mile bus route (in one direction) where the average speed *between* stops is 16.9 miles per hour. The route has 22 stops along the route. If the average dwell time at each stop is 18 seconds, what is the average bus speed along the route as a whole?

And, if a 5-minute layover is required at the far end of the route, what is the total round-trip travel time? Assume the time for the inbound trip is identical in time to the outbound trip.

Answer available at the end of this section.

Glossary

- Bunching: The phenomenon in which two buses become closer to one another on a route, in some cases even joining one another along the route.
- Deadhead: A trip between route termini while the vehicle is not in revenue service; typically taken to reposition the vehicle to start service on a new route.
- Hail-stop operation: A vehicle need only stop at a station if there are passengers desiring to board or to alight. Otherwise, the vehicle can continue without stopping.
- Layover: A short period of time between the end of revenue service on one trip and the resumption of revenue service on the next (vehicle) trip, at a common terminus.
- Pull-in: The movement of a vehicle from a route terminus back to the depot at the end of revenue service.
- Pull-out: The movement of a vehicle from the depot to a route terminus to begin revenue service.
- Revenue service: The movement of a vehicle along a route where passengers may board and alight.
- Schedule slack: a short amount of time inserted into the schedule on a route to improve schedule adherence

Related books

Avishai Ceder (2007). Public Transit Planning and Operation: Theory, Modeling, and Practice. Oxford: Butterworth-Heinemann.

Transportation Research Board (2003). Transit Capacity and Quality of Service Manual, Transit Cooperative Research Program, Report 100, 2nd Edition.

Vukan R. Vuchic (2005). Urban Transit: Operations, Planning, and Economics. Hoboken: John Wiley and Sons.

Answers

Sample Problem

Total travel time = (8.5 mi) / (16.9 mi/hr) + (22 stops)*(18 sec/stop)*(1 hr / 3600 sec) = 0.613 hr = 37 minutes (rounded up to the nearest minute)
The average speed = (8.5 mi / 0.613 hr) = 13.9 mph.
The total round-trip time = 37 min + 5 min layover + 37 min = 79 min

This page titled 4.2: Transit Operations and Capacity is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

4.3: Network Design and Frequency

In this unit on service planning, the strategic decisions of **network and route design**, stop layout, and **frequency determination** are described. In the next unit, the tactical decisions associated with creating a service schedule (timetabling), creating a schedule for vehicles to operate the service (vehicle scheduling), and creating work shifts for operators (crew scheduling) are presented.

Overview

The process of planning and designing public transit service is called “service planning”. There are a number of activities commonly associated with service planning:

1. Network design
2. Route design and stop layout
3. Frequency determination
4. Timetabling
5. Vehicle scheduling
6. Crew scheduling

These are ordered roughly in terms of a general sequence in time, as well as of a dependence upon higher-level activities. The general design of a transit network is the highest level activity, undertaken only rarely or when major new systems (e.g., rail or express bus) are introduced. The network design then feeds an element of route design and stop layout, in which the more specific physical facilities for the routes and stops or stations are implemented. Once routes are in place, the frequency of service may be determined, and a timetable for vehicle trips (the service) along the route can be constructed. Once the timetable is created, schedules for vehicles on the route and throughout the network (i.e., the vehicle cycle for each vehicle in the fleet) can be created. Finally, work shifts for operators can be generated (a crew schedule), and those operators are assigned to the work.

The first three activities, in network and route design and frequency determination, tend to be more strategic in nature, and may only be considered infrequently by transit planners. These decisions tend to be driven in part by political and economic considerations, and as such may require careful and strategic thinking on the part of transit planners. This unit discusses these strategic issues.

The activities of timetabling and vehicle and crew scheduling are considered more tactical decisions, as these decisions are made as often as the transit agency may consider reviewing or changing schedules. In many cases, these tactical activities are assisted by software tools that can generate high quality solutions in a short period of time, often with direct interaction with the planner. These topics are discussed in the next unit.

Network and route design

Network design

There are many different ways of structuring a set of transit routes into a network to provide transit service to a given area. These might include any combination of:

- Radial services, focusing on collecting passengers from outlying areas and bringing them into a major trip generator (e.g., a downtown area, major employment center, or other significant destination);
- Cross-town or grid-like routes, focusing on connecting passengers across the area, perhaps between radial services or among several smaller trip generators; and,
- Direct connections, focusing on moving passengers between major trip generators (e.g., between downtown and a high-density residential area, or between a major employer and downtown).
- Circulators, focusing on collecting and distributing passengers in smaller sub-areas.

Usually, transit networks are made up of some combination of these types of routes, allowing a satisfactory level of access to most of the region, with a high degree of direct service to major trip generators. Sometimes, this may involve a combination of modes.

For example, in Boston, the rail systems are strongly radial, while the bus system provides cross-town and circulator services. The scale image below shows the strong radial nature of these lines, directed to downtown Boston.

Boston Rail Lines

Other systems, particularly bus systems, can include both cross-town and radial services, allowing more coverage and greater direct access to destinations. One example is the bus network of New Jersey Transit in and around Newark, shown in the figure below. One can clearly see both radial routes as well as a strong grid-like structure on the west side of downtown.

Newark Bus Lines

Regardless, in the US, the network structure is usually strongly influenced by political and economic realities, with strong citizen and political input as well as hard budget constraints. However, from a transportation perspective, the resulting network may be characterized in several important ways:

- *Geographic coverage*, often influenced by political considerations as well as objectives to provide mobility for lower-mobility populations;
- *Temporal coverage*, determining what time periods on weekdays and on weekends to offer service; and,
- *Connectivity* (direct vs. indirect service), with the desire to connect major trip generators with direct routes but perhaps allowing transfer trips to serve lower-demand areas.

These characteristics play a major role in estimating demand for the service. Specifically, the analyst may wish to experiment with a variety of network structures and routes, in order to estimate the level of demand that each network might support. Well-specified travel demand models should be able to account for geographic and temporal coverage, as well as trip connectivity, in forecasting transit usage.

As an important caveat, most transit agencies do not usually approach network design as if from scratch. Most agencies do have some existing route patterns, and associated infrastructure (stops, terminals, guideways, etc.) that may often constrain certain route structures or at least strongly favor maintaining some of the existing routes. As a result, most agencies consider new elements of the network (new routes and services) as a complement to existing services, with more modest adjustments to the existing route structures. However, where new modes of service may be introduced (e.g., rail systems in traditionally bus-oriented networks), more significant restructuring of service may be possible, perhaps extending service to new geographic areas, intensifying service in existing service areas, or supplementing the new mode by directing routes to connect to and from the new mode.

Route design

In conjunction with network design, a planner must also consider specific routes and their purpose. The layout of individual routes usually involves some trade-offs in design, most notably:

- *Stop density*. Stop density involves the trade-off of passenger access vs. route speed. Higher stop densities mean that passengers will not have to walk or travel far to get to a stop, allowing easier access to transit service. However, higher stop densities also mean that the vehicle may be stopping frequently, reducing the overall operating speed.
- *Route length and circuitousness*. Route length involves the trade-off of direct service vs. service reliability. Longer routes allow passengers to get to more potential destinations, as the route provides direct service to a larger geographic area. But, longer routes may lead to poor schedule adherence, as service may be more prone to travel time variability and/or service disruptions.
- *Trip generators*. The route can serve major trip generators or more minor trip generators, or some combination of these. Often, routes have termini that coincide with major trip generators, with higher passenger flows occurring between these major generators. However, there may also be a need to serve certain geographic areas with modest demand.

In many cases, some combination of routes can be built into the transit network that provides some balance in each of these areas. Common route designs include the following:

- Line haul: high frequency and/or high capacity service on major travel corridors
- Loops (one-way or two-way): coverage for lower-density areas or circulation in and among activity centers
- Short turn routes: to complement a full route, additional service provided on a shorter segment (the “short turn”) of the route
- Branching (or split) routes: branching toward the end of the route, with areas served by only some buses
- Feeder routes: connect to/from line haul from/to lower density areas

- Limited and express routes: only stop at major stops/stations to improve travel times, balance loads
- Zonal service: only serve some sections in a major corridor

Stop density and location

As noted above, stop density can affect both access and the overall speed of transit service. Given the time and distance necessary for vehicle acceleration and deceleration, a high density of stops will lead to much slower overall speeds, when the vehicles must stop at many or all of the stops. Yet, this provides much easier access for passengers, who must only travel a short distance to their nearest stop.

General guidelines suggest higher stop densities where land uses are likewise at higher density (downtown areas, major activity centers) and much lower where land uses are at much lower densities. But, this guidance may depend on a number of factors. If the probability of stopping is low, then increasing the stop density may come at only a modest cost, in terms of providing additional stop infrastructure. In addition, the lack of easy stop access may also serve as a deterrent to ridership, or may require other accommodations for access, such as large garages or lots for park-and-ride access.

A second area for design involves the suitable location for stops. As may be obvious, the best location for stops is as close to supportive land uses as possible. So, for major trip generators such as high-density work locations (office buildings, retail areas, etc.) and other locations, having a stop directly at or adjacent to the location is best. In many other instances, having stops near intersections may improve pedestrian access to land uses at the intersection or in nearby areas on any cross-street.

In bus operations, stops at intersections include “near-side” and “far-side” stops, referring to whether the stop occurs before or after the bus traverses through the intersection, respectively. Because of the potential conflicts with traffic flow at intersections, however, there are some who might advocate for locating stops away from intersections (“mid-block” locations). An excellent discussion of the advantages and disadvantages of near-side, far-side, and mid-block stops can be found in a recent TCRP Report. Many agencies have specific policies regarding near-side or far-side stops, but these are in general much preferred to mid-block stops.

Frequency determination

There are a number of ways that planners may determine a reasonable frequency of service on a route. The most common methods are:

- **Policy headways**. Many transit agencies will determine specific headways that meet policy goals, usually related to providing a minimum level of service along a route when demand on the route is low. This could include 30-minute or 60-minute headways, for example.
- **Minimum headways**. As discussed in the unit on transit operations, bus and rail systems have minimum headways, based on capacity limitations. In cases when demand on a route is very high, these minimum headways can be used to assign the maximum frequencies on the route.
- **Load management**. In many cases where policy headways or minimum headways do not apply, the typical way of determining frequencies is based on managing load at the *peak load point* along the route. The peak load point is that point along the route that experiences the largest number of passengers per hour. Let *X* be the highest allowable ratio of demand to supply (volume to capacity), with values between 0 and 1 (with 1 when volume = capacity). Then:

$$\frac{P}{f \cdot N_{car} \cdot C_{car}} \leq X \text{ or, } f \geq \frac{P}{X \cdot N_{car} \cdot C_{car}} \tag{4.3.1}$$

with $P =$ volume at peak load point (passengers/hr)

$X =$ maximum allowable volume-to-capacity ratio, $0 \leq X \leq 1$

$N_{car} =$ number of cars in a train (=1 for buses)

$C_{car} =$ number of passengers per car

> For a bus route operated with 55-passenger buses, a demand of 360 passengers per hour past the peak load point, and an allowable load factor of X = 0.90, the frequency must be:

$$f \geq \frac{360\ passengers/hr}{0.90 \cdot 1 \cdot (55\ passengers/bus)} = 7.27\ buses/hr$$

This would give the planner options for this route such as 7.5 buses/hr (8-min headways) or 8 buses/hr (7.5-min headways), or even higher frequencies.

These are the most common methods for determining frequencies.

From a more rigorous perspective, it can be shown that if one combines operating costs with passenger waiting costs, the "optimal" (cost-minimizing) frequency is given by:

$$f = \sqrt{\frac{VOT \cdot p}{2 \cdot C \cdot T}} \tag{4.3.2}$$

with: $VOT =$ value of passenger waiting time ($ per passenger-hour)

$p =$ total boardings per hour on the route, both directions (passengers per hour)

$C =$ operating cost per hour per vehicle ($ per vehicle-hour)

$T =$ total round trip travel time along the route (hours)\)

The important observation from this formula is that the frequency increases with the square root of the total passenger boardings. The square root balances the operating costs, which grow linearly with the frequency, with the passenger waiting time, which decreases inversely with the frequency.

With a value of waiting time of $15/hr, passenger boardings of 750 passengers per hour, operating costs of $80/vehicle-hour, and a round-trip travel time of 1.75 hours, we have:

$$f = \sqrt{\frac{(\$15/passenger-hr) \cdot (750\ passengers/hr)}{2 \cdot (\$80/veh-hr) \cdot (1.75\ hr/veh)}} = 6.34\ veh/hr$$

which could be rounded to 6 veh/hr (10 min headways) or to 6.67 veh/hr (9 min headways), assuming that the transit agency would prefer to have integer-valued headways.

Vicious/Virtuous Cycles

When the elasticity of transit is appropriately high, Mohring's Formula may result in what is known as a "vicious circle". When service is reduced in order to reduce costs (or any other reason), if the elasticity is positive, then demand will also decrease, meaning fewer riders. If we apply Mohring's formula for this number of riders, then we will again see demand drop with the frequency and thus fewer riders again. This cycle is known as a "vicious" circle. On the other hand, increasing frequency can increase demand and thus the number of riders. If we apply Mohring's Formula to this new number, the frequency will again increase. This is what is known as a "virtuous cycle".

Sample Problem 1:

A transit agency is considering changing the frequency on a popular route. Currently, the route serves 1175 passengers in the peak hour, operating at a frequency of 10 buses per hour.

Using Mohring's formula, what is the "optimal" frequency on this route? Use a cost per bus-hour of operation of $66, a value of time of $11 per hour, and a route round-trip time of 95 minutes.

If a comparable demand elasticity with respect to frequency is +0.37 for the peak hour, estimate the total number of passengers in the peak hour for this "optimal" frequency.

Answer available at the end of this section.

Sample Problem 2:

A transit agency is considering reducing service on a route in order to minimize costs. Currently, the route serves 1088 passengers in the peak hour, operating at a frequency of 8 buses per hour. The route currently serves 1113 passengers per peak hour, operating at 8 buses per hour. Using Mohring's formula, what is the "optimal" frequency on this route? The agency uses a cost per bus-hour of operation of $70, a value of time of $11 per hour, and a route round-trip time of 95 minutes. If the demand elasticity is 0.35 for the peak hour, estimate the total number of passengers in the peak hour for this optimal frequency.

Answer available at the end of this section.

Related books

Avishai Ceder (2007). Public Transit Planning and Operation: Theory, Modeling, and Practice. Oxford: Butterworth-Heinemann.

Vukan R. Vuchic (2005). Urban Transit: Operations, Planning, and Economics. Hoboken: John Wiley and Sons.

Answers

Sample Problem 1:

The optimal frequency is given as:

$f = \sqrt{\frac{VOT \cdot p}{2 \cdot C \cdot T}} = \sqrt{\frac{\$11 \cdot 1175\ passengers}{2 \cdot \$66 \cdot 95 min/60 min/hr)}} = 7.86\ buses/hr$

which we might round up to 8 buses/hr.

At 8 buses/hr, the net change in the frequency is -20%, yielding a percentage change in ridership of (+0.37)*(-20%) = -7.4%, or (-7.4%)*(1175 passengers) = 87 passengers. The new ridership would then be 1088 passengers.

Sample Problem 2:

The optimal frequency is given as:

$f_{opt} = \sqrt{\frac{VOT \cdot p}{2 \cdot C \cdot T}} = \sqrt{\frac{(11)(1113)}{2(70)(\frac{95}{60})}} = 7.43\ bus/hr$

At 7.43 buses/hr, the change in the frequency is -0.257, giving a change in ridership of (+0.35)*(-24.2%) = -9.0%, or (-9.0%)*(1113 passengers) = -100 passengers. The new ridership would then be 1013 passengers. Using this new ridership, we can iterate the equation again to find a new optimal

frequency: $f_{opt} = \sqrt{\frac{VOT \cdot p}{2 \cdot C \cdot T}} = \sqrt{\frac{(11)(1013)}{2(70)(\frac{95}{60})}} = 7.10\ bus/hr$

Which corresponds to a loss of ridership of 16.2 riders, which could again be iterated to find a new, lower frequency, further decreasing the ridership.

This page titled 4.3: Network Design and Frequency is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

4.4: Timetabling and Scheduling

In the previous unit on service planning, the strategic decisions of network and route design, stop layout, and frequency determination were described. In this unit, the tactical decisions associated with creating a service schedule (**timetabling**), creating a schedule for vehicles to operate the service (**vehicle scheduling**), and creating work shifts for operators (**crew scheduling**) are presented. A practical guidebook and learning tool has been published recently.

The motivation for good solutions to these tactical decisions is to minimize the net operating costs to the agency. Once the timetable is determined, the number of vehicles required to be in revenue service can also be identified. When the vehicle schedule is determined, the total mileage and hours for the vehicle fleet are defined. Finally, when the crew schedule is determined, the total cost of labor (operators) is defined. Since these factors are the primary determinants of operating costs, finding efficient solutions has a direct effect on the bottom line.

In many cases, these tactical activities are assisted by software tools that can generate high quality solutions in a short period of time, often with direct interaction with the planner. As a result, the interested student may wish to consult other sources to identify and to investigate the specific software tools that might be available, such as those described in a recent publication.

Timetabling

The general idea behind timetabling is to create a schedule for service. As inputs, one would consider the frequency of service for the given route (see previous unit) and the expected travel times between stops on the route. The latter could be determined either by historical experience or through estimates based on traffic conditions, vehicle acceleration and deceleration characteristics, expected dwell times, etc.

Let *h* be the selected headway for a route, perhaps for a specific time period of the day.

Let t_{ij} be the time between stop *i* and stop *j* along the route, where *i* and *j* are adjacent stops. The travel times between stops, t_{ij}, can vary by time of day, particularly as they may be affected by traffic conditions. They may also reflect any slack time built into the schedule between stops, to allow for possible variability in travel times.

Finally, let t_0 be the dispatch time (departure time) of the first vehicle from a terminal.

Then, the timetable can be created simply using the following structure, with *n* stops on the route and *k+1* vehicles to dispatch:

Stop 1 (Terminal)	**Stop 2**	**Stop 3**	**Stop 4**	**...**	**Stop n (Terminal)**
t_0	$t_0 + t_{12}$	\(t_0+t_{12}+t_{23}	$t_0 + t_{12} + t_{23} + t_{34}$	...	$t_0 + \sum_{i=1}^{n-1} t_{i,i+1}$
$t_0 + h$	$t_0 + h + t_{12}$	$t_0 + h + t_{12} + t_{23}$	$t_0 + h + t_{12} + t_{23} + t_{34}$	...	$t_0 + h + \sum_{i=1}^{n-1} t_{i,i+1}$
$t_0 + 2h$	$t_0 + 2h + t_{12}$	$t_0 + 2h + t_{12} + t_{23}$	$t_0 + 2h + t_{12} + t_{23} + t_{34}$	...	$t_0 + 2h + \sum_{i=1}^{n-1} t_{i,i+1}$
$t_0 + 3h$	$t_0 + 3h + t_{12}$	$t_0 + 3h + t_{12} + t_{23}$	$t_0 + 3h + t_{12} + t_{23} + t_{34}$	...	$t_0 + 3h + \sum_{i=1}^{n-1} t_{i,i+1}$
...	...	...	...	...	...
$t_0 + k \times h$	$t_0 + k \times h + t_{12}$	$t_0 + k \times h + t_{12} + t_{23}$	$t_0 + k \times h + t_{12} + t_{23} + t_{34}$	...	$t_0 + k \times h + \sum_{i=1}^{n-1} t_{i,i+1}$

The primary decision variable here is the initial dispatch time, t_0. Different operating conditions might lead to a number of possible choices for t_0:

- **"Clockface" values**. Passengers may remember the schedule more clearly if the dispatch times fall at easily recognized times on the clock. For example, with 15-minute headways, there may be value to passengers in dispatching a vehicle on the :00, :15, :30, and :45 of each hour.
- **Coordination for improved vehicle scheduling**. When a vehicle finishes its trip at a terminal, it will often be turned around to continue onto the next trip in the opposite direction on the route. In this case, there is a need for sufficient layover time at the terminal. If the vehicle finishes a trip at time *t*, then completes the layover after an additional time t_L, then the vehicle

may start its return trip after $t + t_L$. Choosing the dispatch time to occur at or slightly after $t + t_L$ allows for higher vehicle utilization.

> One way of visualizing such a system uses a so-called “string diagram,” shown in this figure. The blue lines indicate the trajectory of a vehicle from the terminal at stop 1 to the terminal at stop *n*, with short dwell times at each stop. Vehicles arriving at stop *n* then can return along the route in the opposite direction (the red lines), after a layover (indicated by the black arrows). These diagrams can be useful in visualizing vehicle movements and crosses along the route.

Caption String diagram of a single route timetable

- **Coordination of passenger transfers**. In some cases, it may be desirable to choose the dispatch time so that passengers may connect to other routes in the network without unduly long waiting times for the transfer route. To do this consistently across the timetable, the two (or more) routes that connect for the transfer must have the same headway h. If this is the case, then the initial dispatch time t0 can be chosen so that the vehicle’s arrival time at the transfer point matches that of the vehicle on the connecting route.
- **Reduction of vehicle requirements**. The timetable will dictate how many vehicles are in operation at any time of the day. In some cases, minor adjustments in the dispatch times, coupled with changes in layovers and/or dead-heads of vehicles between terminals can lead to a reduction in the number of vehicles needed for service.

Vehicle scheduling

Vehicle scheduling, also called “blocking”, involves assigning vehicles to cover the trips associated with the timetable. A vehicle “block” is the schedule of travel of a vehicle for a given day, including: (1) a pull-out from the depot, (2) a sequence of trips from the timetable, (3) any dead-head trips, and (4) a pull-in back to the depot (recall the vehicle cycle from the unit on vehicle operations).

Generally, once the timetable is created, the time and mileage that vehicles spend in revenue service (i.e., completing the trips in the timetable) is fixed. So, the usual goal in vehicle scheduling is to minimize the time and/or distance that vehicles spend outside of revenue service: e.g., pull-ins, pull-outs, dead-heads, and layovers. These all represent time or mileage that are “unproductive”, and hence should be minimized.

Constraints on this process include the following:

- Each trip in the timetable must be made by a vehicle.
- A vehicle cannot be assigned more than one trip at any point in time.
- If a vehicle must be re-positioned for a trip, the associated travel time and distance from its current position to the new position must be observed.

To solve for the vehicle schedule, one might consider a simple “first-in-first-out” rule. In this case, a vehicle stays on the same route throughout the whole period, and is always assigned to the next trip after a layover. The string diagram above gives just such an arrangement.

As a simple example, suppose we have a route that runs from terminal A to terminal B and then back to terminal A. Travel time from A to B and from B to A, including running and dwell time, is 30 minutes, and a minimum 5 minute layover is needed at each terminal. Headways are 15 minutes.

Below is a timetable for this situation, for trips between 6:00 am and after 9:00 am. The left-hand side of the timetable shows vehicle trips from A to B, while the right-hand side shows vehicle trips from B to A.

Caption Timetable and vehicle schedule

The colors correspond to different vehicles used on the route. The gray color corresponds to the first vehicle of the day, leaving A at 6:00 am and continuing with the trip from B at 6:40, the trip from A at 7:15, etc. A total of five vehicles are required to cover all the trips in this timetable.

In addition to the trips from the timetable, the vehicle block also includes a pull-out and pull-in, so that the final block for the first vehicle (gray) could look like the following.

Caption Vehicle block

For networks with longer policy headways (e.g., 30- or 60-minute headways), longer layovers at terminals may be necessary if vehicles serve the same route throughout the block. As a result, other options can be considered, particularly in terms of shifting vehicles from one route to another. The timetable may allow vehicles to shift from one route to another, in order to reduce layover time and/or to avoid pull-outs or pull-ins. Specific activities in the block can include:

- **Interlining**: the process of switching a vehicle from one route to another at a terminal, when the routes share that common terminal.
- **Deadheading**: the process of switching a vehicle from one route to another, also requiring a re-location of the vehicle (traveling empty) to another terminal.

These methods can be quite effective under different timetabling conditions.

Crew scheduling

Crew scheduling (also called “run-cutting” in the transit industry) is the task of determining work shifts (so-called “duties” or “runs”) for operators. Generally, the primary interest in crew scheduling is to minimize the total cost of labor that meets the service requirements.

A significant fraction, typically 60-70%, of the total operating costs at a transit agency involves the cost of operators, including wages, benefits, and other premiums. With this in mind, small reductions in the number of operators, or in the total work hours, can result in more substantive reductions in the total operating cost. For this reason, the task of scheduling crew to vehicles is one area where many large transit agencies can achieve some efficiencies and potential cost savings.

Crew scheduling is complicated because operators often cannot simply be assigned to a vehicle for the entire vehicle block. First, the shift would often be much longer than a typical 8-hour work period; and, second, the operator may not get sufficient break time during vehicle layovers (e.g., for lunch). Instead, the duties have to consider more practical concerns of the operators.

In this regard, transit agencies have rules that dictate the kind of work shifts the operators may perform. In most cases in the US, the types of work shifts are governed by collective bargaining agreements (union work rules) that specify work conditions for transit operators. Possible examples of work rules could include restrictions like the following:

- A duty should start and end at the same terminal
- Crew needs at least 2 breaks during the day: a normal (15-min) break and a (30-min) lunch break
- A break is required after no more than 3 hours of work
- Each crew must have at least 8 hours off before resuming duties on the next day
- Only 20% of duties can be longer than 9 hours
- Only 25% of duties can be split into intervals with an unpaid break (e.g. a duty that only covers the AM and PM peak periods)
- Only 30% of duties can be covered by part-time operators

The general approach to creating a crew schedule begins by cutting each vehicle block into “pieces of work.” Each piece of work is a subset of trips in the block, forming the elemental unit of work (driving) for the operator. Then, according to the constraints from the work rules, these pieces of work are assembled into feasible duties. The hope is to assemble a full set of duties such that all pieces of work are covered and that the total cost is minimal. The cost of a duty can depend on both the traditional hourly rate of pay for the operator for hours worked. If the operator has a *straight* shift (no unpaid break), they are paid a certain amount, usually at a given hourly rate. Other costs can include:

- A minimum *guarantee* of hours of pay, if the guarantee exceeds the number of hours worked (e.g., 8 hours of pay, even if the operator works only 7 hours);
- Premiums for *overtime* (e.g., time in the duty over 8 hours);
- Premiums for *spread* time. Spread is the total time between the start and end of a duty. If this exceeds a certain maximum (e.g., 9 hours), the operator is entitled to extra pay;
- Premiums for *swing*. Swing occurs when the duty starts and ends at different locations (terminals, depots);
- Premiums for *split* duties, where the duty has an unpaid break. This can occur when an operator works only the AM and PM peak periods, without working in the mid-day;

These rules on pay suggest that the crew schedule should contain as many straight duties as possible. Small pieces of work

that remain after generating these straight duties can be allocated to part-time operators (if they are available), to avoid other premiums, or covered using split duties with associated split and/or spread penalties.

A second problem in crew scheduling is *rostering*, in which duties are assembled into a group of duties (the "roster") for each operator, by week. For example, one roster could include the same 8-hour duty for 5 weekdays. However, many possible combinations of duties could be considered, especially if weekend or evening service is provided. Once the rosters are created, operators choose from among these duty rosters.

Glossary

- Block: the sequence of trips made by a vehicle in the course of one day of operations, including both revenue and non-revenue trips.
- Duty (or, Run): a work shift for an operator for one day.
- Guarantee: the minimum pay hours for an operator, regardless of the number of hours worked.
- Piece of work: an operator work assignment extracted from a vehicle block.
- Relief: the change of operators during a vehicle block.
- Roster (also, rostering): the set of duties for a single operator in a week.
- Split: a duty covering at least two intervals of time with an unpaid break.
- Spread: the time between when an operator reports for duty and when they end their duty.
- Straight: a duty covering a single interval of time.
- Swing: a duty in which the operator begins and ends at different locations.
- Tripper: a short work assignment (e.g., 2–4 hours); generally much shorter than a typical straight.

Related books

Avishai Ceder (2007). Public Transit Planning and Operation: Theory, Modeling, and Practice. Oxford: Butterworth-Heinemann.

Transportation Research Board (2003). Transit Capacity and Quality of Service Manual, Transit Cooperative Research Program, Report 100, 2nd Edition.

This page titled 4.4: Timetabling and Scheduling is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

4.5: Traveler Information for Transit

Transit riders constantly worry about their trip. While using public transportation, passengers are under the control of the transit agency. Delays, route changes, and confusing schedules are just a few of the concerns transit riders have. “After 3 minutes, a passenger waiting for a bus that is not coming expects more information and begins to consider alternatives to fulfill their journey.”(US Department of Transportation, 2004) After walking or driving to a transit station that does not have adequate information, passengers do not know if they have just missed the last bus or when the next bus will arrive. This can make the wait time unbearable. When a transit vehicle does arrive, travelers need to verify that it serves the route they wish to take, which can be another place for anxiety. Once on board, passengers must determine when to alight, if any transfers are necessary, and the best walking route to their final destination. All of these factors can make taking transit a stressful experience, especially for the uninitiated.

Transit information problems are magnified for individuals with disabilities. “What may be a surmountable difficulty for the able bodied user may become an impossibility for one who is handicapped.” (Arthur et al., 1992) People with physical or mental disabilities rely heavily on transit. Besides relieving anxiety for new users, transit information is vital for disabled people.

Many of the stresses and worries transit riders experience can be alleviated through information. While the actual transit service does not change – the route and schedule remain the same – the addition of information eases customers’ worries, increases their enjoyment and appreciation of transit, and may actually make the trip seem shorter. By increasing the usefulness and pleasantness of transit, riders will be likely to use transit more often. Increased information makes it easier and more enjoyable for new riders to negotiate transit and for experienced riders to travel on new routes. Information systems make transit user-friendlier and can eliminate some of the barriers to taking transit.

There are three locations where it would seem that transit information is desired: pretrip, at transit stops, and in-vehicle. The type of information necessary at each location is different. Table 1 summarizes the types of information systems, the purposes that they fulfill and their usefulness. They are discussed in turn

	Pretrip	**Transit Stop**	**In-Vehicle**
Purpose	Route planning	Anxiety relief	Current location
Purpose	General Operations	Route info	Transfer information
Purpose	Fare information	Real-time Arrival and Delay Information	
Static	Printed Schedules	Schedules	Route Map
Static		Route Maps	System Map
Static		Transfer info	Station/Stop Identifier
Static		Station/Stop Identifier	
Static		Location Maps	
Dynamic	Kiosks	Kiosks	Cellular/Wireless
Dynamic	Phone Systems	Countdown	Variable Message
Dynamic	Websites	Variable Message Board	Board
Dynamic		CCTV	Announcements
Dynamic		Announcements	Dynamic Map
Dynamic		Attendants	Driver
Most Useful	Websites	Variable Message Boards	Variable Message Boards
Most Useful		CCTV	Announcements
Most Useful		Kiosks	

Pretrip Information

Pretrip information can be obtained from printed schedules and maps, a transit agency website, or a telephone information service. Pretrip information focuses on trip planning. This is a critical step for new riders and transit riders who are not familiar with the route. If accurate information cannot be easily obtained, it is not likely that new users will venture onto transit, unless that is their only choice.

Static Pretrip Information

Printed schedules do not require the use of any equipment or technology, but do require that the user obtain one prior to the trip and understand basic map reading and scheduling. Obviously, printed materials do not contain any real-time information or route planning assistance. Unless a user has schedules for all relevant routes, it is possible that an optimal route exists which is unknown to the user. Another disadvantage of printed schedules is that they can easily become outdated when route and schedule changes occur. Often only a few stops are listed and the maps attached are very small. Printed schedules often do not provide what users want: A quick explanation about how to get from point A to point B (Transportation Research Board, 2003).

Automated phone systems that do not provide any real-time information could be considered static in that they are an aural presentation of a printed schedule. These systems can still be useful for users who lack a printed schedule, but they have the same limitations as printed materials.

Cincinnati Bus Diagram 2011

Dynamic Pretrip Information

Phones were once the dominant method of obtaining pretrip information from home. The most effective phone transit information systems provide personal help through live assistance. The advantages to phone information systems are that phones are ubiquitous. This allows phone information systems to be accessed from home as well as en-route using a cellular phone. Also, phone systems offer the potential to interact with a live person, which may be more comforting than a web or printed interface. Customized route information can be obtained and all questions answered with one conversation. This direct inquiry method appeals to many people. In a study done on the London ROUTES call center, four percent of the calls resulted in trips that would otherwise not have used transit (Le Jeune, 1996). In addition, real-time information can be delivered over the phone. The disadvantages to phone information systems are high cost, limited hours, inefficient information retrieval, and limited disability access.

Transit agency websites and apps (hereafter referred to as website, whether the browser is a conventional browser or a specialized app) offer the potential to reach a very large audience with updated information for relatively low cost. Websites offer many advantages to phone or printed information systems. The number of people with internet access and the popularity of the internet continue to grow. Websites are available 24 hours a day, offering flexibility to users who want information at any time. Websites (and apps) are more efficient at distributing information than phone systems. A computer interface is ideal for quickly retrieving, saving, and printing information on routes, policies, and schedules. Phone systems can be awkward, confusing, and time-consuming. Websites also appeal to new users who “just want to look” at transit. While phone systems require that a user have a specific question or route, websites have a self-serve approach that encourages browsing thus making it more likely that a visitor will stumble across a route or service that they didn’t know existed. Finally, websites can reach a broad range of people. Computer interfaces provide better accessibility than phone systems to disabled persons (Transportation Research Board, 2002). Visitors and tourists can look up transit information from home.

Websites can offer personalized trip planning. This is perhaps the best aspect of internet-based information systems. Rather than guessing which routes are necessary and then downloading and cross-referencing timetables, users can enter origin and destination pairs and receive routing instructions including transfers and walking information. Transit website trip planners can incorporate multiple modes and transit agencies, adding to their value.

Finally, websites provide the best platform for distributing real-time transit information. Coupled with dynamic AVL-type systems that track vehicles, website visitors can see where transit vehicles currently are and their estimated arrival time for any stop. The computer provides a visual map with much more information readily accessible. Real-time information available online can effectively reduce passenger waiting time, the most onerous part of the trip, because an individual can time their departure to arrive just prior to the transit vehicle.

One disadvantage to transit websites is that internet access is not available to everyone, and even as smart phone availability rises, data plans can remain expensive. Choice riders are more likely to have internet access, and are also likely to be less familiar with the transit system thus making a transit trip planner more useful to them.

Because different types of riders obtain information through different sources, it is important to maintain a variety of pretrip planning sources. Frequent riders generally do not need detailed information; they can plan a trip with just the schedule and route map. Younger, more educated and more affluent riders tend to use websites for pretrip information. Also, disabled people rely on websites for transit information (Transportation Research Board, 2002). Robinson et al. (1979) found that females, older persons, and less educated people tended to use a phone information system more often because they liked direct inquiry. Finally, inexperienced riders need to know everything from fares, operating policies, route information and bus stop location (Transportation Research Board, 2003). Given the advantages that websites have for distributing information effectively and efficiently, further development of online transit resources will provide users with excellent pretrip planning information.

Transit Stop Information Systems

Transit stops are where pedestrians interface with transit. Information is crucial at these locations because much of a transit rider's anxiety occurs at these stops while waiting for a vehicle to arrive and determining which vehicle to board. Information provided at transit stops focuses on route-specific information and reliving the anxiety of passengers. Information can be provided statically through posted maps, schedules, route numbers, and station identifiers or dynamically through countdowns, displays, dynamic maps, and verbal announcements.

Static Transit Stop Information

Static information is an essential part of any transit stop, regardless of size. It answers the questions of "where" and "how". It can also provide a guess as to "when" transit will occur. At the most basic level of static information, some type of identifier needs to be present to denote that a transit stop exists. While a simple "Bus Stop" sign fulfills this requirement, it does not provide any assistance to users nor is it conducive to encouraging new users. The Transit Cooperative Research Program recommends as a minimum that every transit stop should display the route numbers of all vehicles that use that stop, a list of locations served and general operating hours (Transportation Research Board, 1999). Ideally, a basic schedule showing each time that a vehicle will arrive, approximate transit times to all stops, and a map of each route that serves that stop would be included as well. This is the basic information required for new riders to negotiate transit from that stop. The only disadvantage of providing static information is that it needs to be updated whenever a schedule or route change occurs. This can be a large expense if each stop must have updated information. However, with good planning and standardized information for each stop along the route, this expense can be minimized.

With this static information, many questions that new riders have would remain unanswered. A system map with operating hours, fare descriptions, and other policies posted at each stop would allow users to plan transfers, seek alternative routes, and become familiar with the basic operational guidelines of the system while waiting for transit. Also, static information sources cannot address the operating issues of transit such as delays, route detours, or other notifications. On streets near to transit stops, trailblazer signs can be posted to inform and orient people. These signs should include the direction of the transit station, the walking distance from that point and which routes serve it. These signs will help existing passengers locate the bus stop and will also serve as advertisements for transit to non-riders. Areas with high numbers of tourists can benefit greatly from trailblazer signs. Another set of static information that can be provided at transit stops is oriented to arriving passengers. A map of the area, showing points of interest, walking routes, and other transit routes with transfer information, would be a valuable resource to passengers alighting. This information would not need to be updated as frequently and the information would serve all visitors to the neighborhood, not just transit passengers.

BusTransportSG

Dynamic Transit Stop Information

Dynamic information systems at transit stops answer the all-important question of "when". One of the biggest disadvantages to using transit is the uncertainty and anxiety that occurs when waiting for transit vehicles to arrive. Studies have shown that when transit customers are questioned about the duration of their wait time, they cited a time 50% higher then their actual wait time (Peng et al., 1999). This same phenomenon occurs when a transfer is required. By providing real-time information at transit stops, the worries of passengers can be minimized, greatly enhancing the pleasantness of the trip. Another benefit is the reduced

waiting time. If a passenger arrives at a station and notices a long time until arrival, they can spend their time productively at activities near the transit route (Mishalani et al., 2000).

Real-time information systems can affect passenger behavior. If more than one route serves their destination, passengers can make route choice decisions based on the real-time arrival information provided, thereby reducing their total travel time. When notified in advance of a delay, passengers can make important decisions, and avoid the frustrating situation of being late and uninformed. Passengers experiencing delay can switch to another mode, take an alternate transit route, or at least call ahead to notify a waiting party of the delay (Peng et al., 1999) In this way, real-time information systems can make transit more efficient and enjoyable.

One type of useful dynamic system is a display that counts down until the arrival of the next vehicle. This works well at stations served by one or two. Variable message signs and video monitors can be used at transit stations served by many routes. These message signs can show arrival times, and when delays occur they can suggest alternate routes. This simplifies route planning quite a bit. Whereas a static schedule shows an entire day worth of departures that must be sifted through in order to find the information desired, a variable message sign streamlines the process by showing only the next few departures and their current status. This is the relevant information at that moment for any transit rider. An announcer can also broadcast this information through a terminal or transit stop. Announcements help the visually impaired; however, announcements cannot often be heard due to background noise or insufficient speakers.

A study of riders in London after the countdown system was initiated showed that a majority of riders felt that waiting time passed more quickly, bus service was more reliable, waiting at night was safer and that passengers' general feelings towards bus travel improved. Of the people surveyed, 93% thought that the countdown system should be implemented at every bus stop in London (Transportation Research Board, 2003). One hundred percent of riders stated that they use Tri-Met's Transit Tracker always (82%) or sometimes (18%) and they placed significant value on the information (Transportation Research Board, 2003). The question is not whether real-time information is useful or valued by customers but how to efficiently implement it.

Kiosks at transit stops essentially provide the same services as a transit agency website, but are made publicly available at transit stops. Personalized route planning can be accomplished. Also, many kiosks can provide real-time information on arrivals and delays.

Perhaps the most dynamic of all information sources is live human interaction. Station attendants are an excellent way to provide information to customers. New transit riders especially appreciate the ability to ask their questions in person. In addition, attendants can serve as security monitors to help maintain passenger safety in transit stations. The major disadvantages to attendants are the relatively high labor costs they impose and the fact that they can only serve one customer at a time. However, the expense may be justified at busy stations or at stations with lots of new riders, such as airports and passenger rail stations.

Shek Kip Mei Station PIDS and clock 2

In-Vehicle Information Systems

The final type of information system is in-vehicle. Passengers need to know where to alight and what routes, if any, they are required to transfer to in order to complete their trip. If passengers are traveling through unfamiliar areas and stops are not clearly discernable, a great deal on anxiety can be generated. Passengers fear that they may have missed their stop. In-vehicle information systems can be classified as static and dynamic.

Static In-Vehicle Information

The simplest way to distribute information in-vehicle is to distribute printed schedules and maps. These can be identical to the format of the pretrip printed pamphlets and the posted transit stop information. While this type of information is better than nothing, it leaves much to be desired. Unless the route map has landmarks on it, it does not help the user figure out where they currently are in relation to their destination. Other forms of static information include route maps and system maps that are posted on transit vehicle walls. This usually works best on rail cars where people are free to move around.

One method to inform the user of their current location is through signage at transit stops. This is prevalent along rail lines where there are fewer stops and each is well defined. By looking out of the windows of a transit vehicle, users can read the name of the stop and compare it on the map to where they want to alight. Signage such as this does not cost very much to implement and is relatively permanent. Clearly labeled stations and stops help passengers judge how much longer their trip will be, and reduce

passenger anxiety because they know their location at all times.

Dynamic In-Vehicle Information

When exterior signage at stations and stops is impractical, it is possible to dynamically inform passengers of their location. Variable message boards mounted in transit vehicles and announcements are one way of informing passengers of their location. Alternatively, dynamic lighted maps can be used to indicate location. This has the advantage of including a map with the location so users know where they are relative to their stop. However, this type of system is impractical on a bus due to its configuration.

Cellular phones and other portable devices allow users to access the pretrip information sources en route. This can be useful for finding transfer information and general route information, especially with GPS units embedded in smart phones. Also, wireless communication and GPS is often limited on transit systems that operate underground.

Costs of Information Systems

Information systems range in price from very low (for posted laminated maps and schedules) to high ($10,000 per info kiosk and over $300,000 for website startup costs). Peng et al. (1999) found that transit websites and personal communication devices services were the least expensive to maintain. Once the startup costs of a website are covered, it costs very little to update the information and even less to distribute it. Unlike phone centers, there is never a busy signal (except rarely at extremely high peak usage times) and very little labor is required to maintain it. Additionally, websites are easy to implement, are versatile, and have almost unlimited information carrying capacity.

Call centers are cheaper to implement but much more expensive to run than websites. For users without internet access or capabilities, call centers are still important to maintain. Berlin Transport provides a 24-hour transit information telephone service that receives one million calls annually. The estimated cost per call is $7.50 (U.S. Department of Transportation, 2003).

The most effective pretrip information systems combine transit routes from multiple agencies, modes, and carriers into one easy-to-use service. This allows the call center to charge the agency for handling the call. The National Rail Enquiry Service in England provides non-automated transit information to callers. It is funded through charging train operators £0.55 per call. It has been estimated that each call generates £2.45 of revenue for train operators (U.S. Department of Transportation, 2003).

Capital costs for installing automatic vehicle location (AVL) systems varies significantly, but a study of seven US transit agencies found that the cost ranged from $3000-$10000 per vehicle (Transportation Research Board, 2003). The annual operations and maintenance cost for AVL systems range from $100-$700 per vehicle. These types of systems are relatively new and the costs are likely to decrease as technology improves. Also, because real-time vehicle location systems have been implemented in numerous places already, fewer mistakes will occur in future installations.

Variable message signs, countdown timers, and other real-time displays are relatively easy to implement. The cost is certainly higher than stationary signs ($500-$1500 per unit) but they are versatile, easy to implement, user friendly, and they deal with one of the fundamental problems in transit: anxiety (Peng et al., 1999).

Kiosks are expensive to implement and maintain, but provide personalized service and can be very user friendly. One way kiosks have been partially funded in Madrid is through advertising and co-location rights with the Savings Bank of Madrid ATM machines (U.S. Department of Transportation, 2003). When implemented together, Kiosks and variable message signs provide nearly all the transit information required at a terminal or transit stop.

Variable message signs also are the best option for in-vehicle information systems. Variable message signs and CCTV can partially funded through advertising revenue.

Funding

Transit information systems can be funded through a variety of sources. Users are willing to pay more for transit service that includes real-time information, ads can be shown on message boards, and increases in transit use can pay for the capital costs and operating expenses of transit information systems.

With many additional variable message signs and CCTVs, ads can very easily be cycled in between transit information. On busier systems, this can translate into a lot of revenue and can likely pay for the cost of the display (but not the AVL). Transport for London

cites that revenue increased 1.5% on lines equipped with countdown real-time information systems (Transportation Research Board, 2003). This certainly will help defray the cost of implementation.

Finally, users are willing to pay for information. This can be done either directly through fees (such as a transit phone information system) or through higher fares. Previous studies have found ranges from $0.01 to $0.41 per trip or up to $0.50 per request (Transportation Research Board, 2003).

Benefits of Transit Information Systems

Improving transit information systems removes one of the largest barriers to riding transit. In a study of California commuters, Abdel-Aty et al. (1996) found that 38 percent of commuters who do not use transit might consider transit if more information were easily available. Because over 80 percent of users surveyed did not currently use transit, this is a significant number of people. This study also found that the information that mattered most to non-transit users was waiting times, frequency of service, and operating hours. Another study from Harris County (METRO) in 1992 found that non-riders primary information need was bus schedules and connection information (Transportation Research Board, 1999). All of these items relate to time, reflecting people's dislike of the uncertain waits and delays associated with transit. Real-time transit information systems can make waiting seem shorter, and in some cases decrease waiting time by allowing passengers to arrive "just in time". A survey by the Puget Sound Regional Council found that agreement with the statement "When waiting, I'm happier if I know when the bus will come" on a scale of –5 to 5 was 3.9 (Transportation Research Board, 2003). This shows that people really value this piece of information.

Once the barrier of uncertain wait times is removed, customers will be encouraged to use transit more often. A survey by the transit cooperative research program estimated that transit use increased 1% to 3% after deploying real-time bus arrival systems (Transportation Research Board, 2003). Data from Europe attributes ridership increases of 5.8% in Brussels, 5% in Liverpool, and 3% in Turin to the implementation of real-time information systems at bus stops (Transportation Research Board, 2003). The primary reason that agencies in this survey chose to implement these systems was to improve the level of customer service. This translates into ridership increases when a significant investment, like a real-time arrival system, is implemented.

Besides luring choice riders, improved transit information systems give disabled riders increased mobility. Information systems help them negotiate transit systems through better pretrip planning and information along the route.

This page titled 4.5: Traveler Information for Transit is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

5: TRAFFIC

5.1: Queueing
5.2: Traffic Flow
5.3: Queueing and Traffic Flow

This page titled 5: Traffic is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

5.1: Queueing

A Schematic of a Queue System

Queueing is the study of traffic behavior near a certain section where demand exceeds available capacity. Queues can be seen in many common situations: boarding a bus or train or plane, freeway bottlenecks, shopping checkout, exiting a doorway at the end of class, waiting for a computer in the lab, a hamburger at McDonald's, or a haircut at the barber. In transportation engineering, queueing can occur at red lights, stop signs, bottlenecks, or any design-based or traffic-based flow constriction. When not dealt with properly, queues can result in severe network congestion or "gridlock" conditions, therefore making them something important to be studied and understood by engineers.

Cumulative Input-Output Diagram (Newell Curve)

Newell Curve showing the cumulative inputs and outputs from a queue

Based on the departure rate and arrival rate pair data, the delay of every individual vehicle can be obtained. Using the input-output (I/O) queueing diagram shown in the side figure, it is possible to find the delay for every individual vehicle: the delay of the i^{th} vehicle is time of departure - time of arrival ($t_2 - t_1$). Total delay is the sum of the delays of each vehicle, which is the area in the triangle between the arrival (*A(t)*) and departure (*D(t)*) curves.

Distributions

Arrival Distribution - Deterministic (uniform) OR Random (e.g. Poisson)

Service Distribution - Deterministic OR Random

Service Method:

- First Come First Served (FCFS) or First In First Out (FIFO)
- Last Come First Served (LCFS) or Last In First Out (LIFO)
- Priority (e.g. HOV bypasses at ramp meters)

Characterizing Queues

Queue Length Characteristics - Finite or Infinite

Number of Channels - Number of Waiting Lines (e.g. Ramps = 2, Supermarket = 12)

We use the following notation:

- Arrival Rate = λ
- Departure Rate = μ

Utilization Rate ρ = $\rho = \frac{\lambda}{\mu}$

Degree of Saturation

Oversaturated: $\lambda > \mu$

Undersaturated: $\lambda < mu$

Saturated: $\lambda = \mu$

Little's Formula

Little's Formula:

$$E(n) = \lambda E(v) \tag{5.1.1}$$

his means that the average queue size (measured in vehicles) equals the arrival rate (vehicles per unit time) multiplied by the average waiting time (both delay time in queue plus service time) (in units of time). This result is independent of particular arrival distributions and, while perhaps obvious, is an important fundamental principle that was not proven until 1961.

A traffic queue

Uncapacitated Queues (M/D/1) and (M/M/1)

It has been shown that queue sizes, waiting times, and delays differ between M/D/1 and M/M/1 queueing. These differences are represented in formulas and shown below. Note the minor differences between the two.

Comparison of M/D/1 and M/M/1 queue properties

'	**M/D/1**	**M/M/1**
E(w) (average waiting time)	$E(w) = \frac{\rho}{2\mu(1-\rho)}$	$E(w) = \frac{\lambda}{\mu(\mu-\lambda)}$
E(v) (average total delay)	$E(v) = \frac{2-\rho}{2\mu(1-\rho)}$	$E(v) = \frac{1}{(\mu-\lambda)}$
E(n) (expected number of units in the system (including vehicles being served))	$E(n) = \frac{(2-\rho)\rho}{2\mu(1-\rho)}$	$E(n) = \frac{\rho}{(1-\rho)}$

Notes:

- Average wait time (E(w)) excludes service time.
- Average total delay (E(v)) is (wait time + service time). This is sometimes referred to *average delay time* due to the existence of the bottleneck.

- Expected number of units in the system (E(n)) includes customers currently being served (in number of units). According to Little's Law this is arrival rate multiplied by the average time spent in the system (E(n)=\lambda*E(v)\).
- Sometimes Expected number of units in the queue (E(m)) is requested, excluding customers being served, which is a different formula (arrival rate multiplied by the average waiting time $E(m) = \lambda * E(w)$), and obviously results in a small number.

Stochastic Queueing Queue Length

Comparison Of Stochastic And Deterministic Queueing And BPR.

Comparison Of Stochastic Queueing With BPR And Akcelik Formula.

Uncapacitated queues (M/M/1) (random arrival and random service)

In addition to the properties stated before, M/M/1 queueing have a few additional ones of which to take note.

Additional M/M/1 queue properties

'	**M/M/1**
Probability of n units in the system	$P(n) = \rho^n(1-\rho)$
Average waiting time of arrival, including queue and service	$E(v) = \frac{1}{(\mu-\lambda)}$

Average waiting time in queue (excluding vehicles being served)	$E(w) = \frac{\lambda}{\mu(\mu-\lambda)}$
Expected number of units in the system (including vehicles being served)	$E(n) = \lambdaE(v) = \frac{\lambda}{\mu-\lambda}$
Mean queue length (excluding vehicles being served)	$E(m) = \lambdaE(w) = \frac{\lambda^2}{\mu(\mu-\lambda}$)
Probability of spending time t or less in system	$P(v \leq t) = 1 - e^{-(1-\rho)\mut}$
Probability of spending time t or less in the queue	$P(w \leq t) = 1 - \rhoe^{-(1-\rho)\mut}$
Probability of spending more than time t in the queue, given the queue is not empty	$P(w > t \mid w > 0) = 1 - e^{-(1-\rho)\mut}$
Probability of more than N vehicles in queue	$P(n > N) = \rho^{N+1}$

Capacitated Queues (Finite)

Capacitated queues permit a maximum number of vehicles to wait, and thus have different properties than uncapacitated queues. For single channel undersaturated finite queues where the maximum number of units in system is specified as N and with random arrivals and departures (λ, μ) we have:

Additional M/M/1 queue properties

'	**M/M/1**
Probability of n Units in System	$P(n) = \frac{(1-\rho)}{1-\rho^(N+1)}(\rho)^n$
Expected Number of Units in System	\(E(n)=\frac{(\rho)}{(1-\rho)}\frac{1-(N+1)(\rho)^N+N\rho^N+1}{1-\rho^{N+1}\)

Real Life Causes of Queue Generation

For Roads:

- Geometric Bottlenecks (lane drops, hard curves, hills)
- Accidents and Incidents
- Traffic Signals and Intersection Controls
- At-Grade Crossings with other Modes (Railroad crossings, drawbridges, etc.)
- Toll Booths
- Ramp Meters
- "Gawker" Effect
- Inclement Weather

For Trains and Transit:

- Platform Capacities
- Fare Gates
- Ticket Windows/Ticket Machines
- Minimum Safe Separation for Trains
- Security Checkpoints
- Efficiency of Trains Entering and Leaving Station (number of tracks, switches, etc.)

For Aviation and Airports:

- Runways
- Designated Minimum Safe Following Distances for Planes (by government)
- Physical Minimum Safe Following Distance for Planes (creation of turbulence, etc.)
- Available Airspace for Approaches and Departures
- Ticketing Counters/Check-in Procedures
- Security Checkpoints
- Baggage Systems

- Terminal Capacity for Planes
- Internal Terminal Capacity for Passengers
- Inclement Weather

For Water and Maritime:

- Locks and Dams
- Port Capacities
- Minimum "Safe" Distances (determined by government and physics)
- Inclement Weather

For Space Flight:

- Launch Capacity
- Minimum Spacings between Orbital Vehicles
- Inclement Weather on Earth
- Unfavorable Celestial Conditions

Examples

Example 1

At the Krusty-Burger, if the arrival rate is 1 customer every minute and the service rate is 1 customer every 45 seconds, find the average queue size, the average waiting time, and average total delay. Assume an M/M/1 process.

Solution

To proceed, we convert everything to minutes.

Service time:

45/60 = 0.75 minutes per customer. Alternatively, you can say the server can handle 60/45 = 1/0.75 = 1.33 customers per minute.

The arrival rate is 1 customer per minute.

The utilization rate $\rho = (60/60)/(60/45) = 0.75$ Meaning the server is busy on average 75% of the time.

Average queue size (E(n)):

$E(n) = \frac{\rho}{1-\rho} = \frac{0.75}{(1-0.75)} = 3$

$E(n) = \lambda E(v) = 1 * 3.00$ Little's Formula

Average wait time:

$E(w) = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{1}{1.33(1.33-1)} = 2.25$

Average delay time:

$E(v) = \frac{1}{(\mu-\lambda)} = \frac{1}{(1.33-1)} = 3 = E(w) + servicetime = 2.25 + 0.75$

Comparison:

We can compute the same results using the M/D/1 equations, the results are shown in the Table below.

Comparison of M/D/1 and M/M/1 queue properties

'	M/D/1	M/M/1
E(n) (average queue size (#))	1.875	3
E(w) (average waiting time)	1.125	2.25
E(v) (average total delay)	1.875	3

As can be seen, the delay associated with the more random case (M/M/1, which has both random arrivals and random service) is greater than the less random case (M/D/1), which is to be expected.

Example 2

How likely was it that Homer got his pile of hamburgers in less than 1, 2, or 3 minutes?

Solution

$P(t \leq 1) = 1 - e^{-(1-\rho)\backslash \mathrm{muT}} = 1 - e^{-(1-0.75)1.333*1} = 1 - e^{-0.333} = 0.283$

$P(t \leq 2) = 1 - e^{-(1-\rho)\backslash \mathrm{muT}} = 1 - e^{-(1-0.75)1.333*2} = 1 - e^{-0.667} = 0.487$

$P(t \leq 3) = 1 - e^{-(1-\rho)\backslash \mathrm{muT}} = 1 - e^{-(1-0.75)1.333*3} = 1 - e^{-1} = 0.631$

Example 3

Before he encounters the “pimply faced teen” who serves burgers, what is the likelihood that Homer waited more than 3 minutes?

Solution

$P(w \leq 3) = 1 - \backslash \mathrm{rho}^{-(1-\rho)\backslash \mathrm{muT}} = 1 - 0.75e^{-(1-0.75)1.333*3}$

$P(w > 3) = 1 - P(w \leq 3)$

Example 4

How likely is it that there were more than 5 customers in front of Homer?

Solution

$P(n > 5) = \rho^{N+1} = 0.75^{6} = 0.178$

Thought Question

Problem

How does one minimize wait time at a queue?

Solution

Cutting in line always helps, but this problem will be answered without breaking any rules. Think about going out to dinner, only to find a long line at your favorite restaurant. How do you deal with that? Maybe nothing can be done at that time, but the next time you go to that restaurant, you might pick a new time. Perhaps an earlier one to avoid the lunch or dinner rush. Similar decisions can be seen in traffic. People that are tired of being in network queues on their way to work may attempt to leave earlier or (if possible) later than rush hour to decrease their own travel time. This typically works well until all the other drivers figure out the

same thing and shift congestion to a different time.

To minimize wait time at a queue on the road, some places allow lane filtering and even lane splitting under certain circumstances. Some other places have considered these unsuccessfully.

> "There is evidence (Hurt, 1981) that traveling between lanes of stopped or slow-moving cars (i.e., lane splitting) on multiple-lane roads (such as interstate highways) slightly reduces crash frequency compared with staying within the lane and moving with other traffic."

Two motorcycle riders lane splitting in California, USA

Sample Problems

Sample Problem 1: Queueing at a Tollbooth

Application of Single-Channel Undersaturated Infinite Queue Theory to Tollbooth Operation. Poisson Arrival, Negative Exponential Service Time

- Arrival Rate = 500 vph,
- Service Rate = 700 vph

Determine

- Percent of Time operator will be free
- Average queue size in system
- Average wait time for vehicles that wait

Note: For operator to be free, vehicles must be 0

Answer available at the end of this section.

$P(n) = \rho^n(1-\rho) = (0.714)^0(1-0.714) = 28.6$

$Q = \frac{\rho}{1-\rho} = \frac{0.714}{0.286} = 2.5$

$w = \frac{\lambda}{\mu-(\mu-\lambda)} = \frac{500}{700(700-500)} = 0.003571\ hours = 12.8571\ sec$

$Q = \lambdaw = 500 * 0.00357 = 1.785$

Service Time = $\frac{1}{\mu} = \frac{1}{700} = 0.001429\ hours = 5.142\ sec$

$t = \frac{1}{(\mu-\lambda)} = \frac{1}{(700-500)} = 0.005\ hours = 18\ seconds$

Sample Problem 2: Queueing at a Ramp Meter

A ramp has an arrival rate of 200 cars an hour and the ramp meter only permits 250 cars per hour, while the ramp can store 40 cars before spilling over.

(A) What is the probability that it is half-full, empty, full?

(B) How many cars do we expect on the ramp?

Answer available at the end of this section.

Sample Problem 3: Queueing at a Ramp Meter

In this problem we apply the properties of capacitated queues to an expressway ramp. Note the following

- Ramp will hold 15 vehicles
- Vehicles can enter expressway at 1 vehicle every 6 seconds
- Vehicles arrive at ramp at 1 vehicle every 8 seconds

Determine:

(A) Probability of 5 cars,

(B) Percent of Time Ramp is Full,

(C) Expected number of vehicles on ramp in peak hour.

Answer available at the end of this section.

Variables

- A(t) = λ - Arrival Rate
- D(t) = μ - Departure Rate
- 1/μ - service time
- ρ = λ/ μ - Utilization
- Q - average queue size including customers currently being served (in number of units)
- w - average wait time
- t - average delay time (queue time + service time)

Key Terms

- Queueing theory
- Cumulative input-output diagram (Newell diagram)
- average queue length
- average waiting time
- average total delay time in system
- arrival rate, departure rate
- undersaturated, oversaturated
- D/D/1, M/D/1, M/M/1
- Channels
- Poisson distribution,
- service rate
- finite (capacitated) queues, infinite (uncapacitated) queues

External Exercises

The assignment seeks to provide students with the opportunity to gain a better understanding of two queuing theories: M/D/1 and M/M/1. Two preformatted spreadsheets have been made available for assistance in computing the values sought after in this exercise. While these spreadsheets provide the computations for these results, the formula is listed below for reference:

\[WT_q=(\dfrac{C_{\lambda}^2+C_{\mu}^2}{2C_{\lambda}^2})(\dfrac{\rho}{1-\rho})\frac{1}{\mu}

Where:

- $WT_q)$: Average customer delay in the queue
- C_λ : Coefficient of variation (CV) of the arrival distribution

- C_μ : Coefficient of variation of the departure distribution
- CV : Standard deviation/mean; CV = (1/SqRt (mean)) for Poisson process and CV = 0 for constant distribution
- μ : Average departure rate
- λ : Average arrival rate
- ρ : Utilization = Arrival rate/service rate $(\rho = \lambda/\mu$

M/D/1 Queueing

Download the file for M/D/1 Queueing from the University of Minnesota's STREET website: M/D/1 Queue Spreadsheet

With this spreadsheet, run 5 simulations for each of the 10 scenarios, using the arrival and departure information listed in the table below. In other words, program the same data into the spreadsheet 5 different times to capture a changing seed and, thus, produce slightly different answers because of the model's sensitivity. A total of 50 simulations will be run.

Scenario	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Arrival Rate	0.01	0.025	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Service Rate	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

Also, find the utilization for all ten scenarios. Based on the utilization and the distribution variability, use the above equation to compute the average delays for all scenarios with utilization values of less than 1.

Finally, summarize the average-delays obtained both from the simulation and from the WTq equation in the same delay-utilization plot. Interpret your results. How does the average user delay change as utilization increases? Does the above equation provide a satisfactory approximation of the average delays?

M/M/1 Queueing

Download the file for M/M/1 Queueing from the University of Minnesota's STREET website: M/M/1 Queue Spreadsheet

With this spreadsheet, run 5 simulations for each of the 10 scenarios, using the arrival and departure information listed in the table below. In other words, program the same data into the spreadsheet 5 different times to capture a changing seed and, thus, produce slightly different answers because of the model's sensitivity. A total of 50 simulations will be run.

Scenario	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Arrival Rate	0.01	0.025	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Service Rate	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Scenario	1	2	3	4	5	6	7	8	9	10

Again, you need to use five different random seeds for each scenario. Summarize the results in a delay-utilization plot. Interpret your results (You do NOT need to use the WTq equation given above to compute delays for the M/M/1 queue).

Additional Questions

- Finally compare the M/M/1 queue and the M/D/1 queue. What conclusion can you draw? (For the same average arrival rate, do users experience the same delays in the two queuing systems? Why or why not?).
- Provide a brief example where M/M/1 might be the appropriate model to use.
- Provide a brief example where M/D/1 might be the appropriate model to use.

Additional Questions

1. What is a queue?
2. Give examples of queueing in real life.
3. What important variables affect queue length.

4. How do you compute total delay from a Newell Curve? What is vehicle delay? How many vehicles are there in the system at a given time.
5. What are some difficulties with calculating queueing times?
6. Define over and under-saturated?
7. Characterize a queue.
8. Define arrival and service rates.
9. If arrival rates > departure rates, what does that imply.
10. Name three types of statistical distributions that describe the behavior of queues and explain them.
11. Can queueing in succession of a length of road mediate arrival demand which reduces delay?
12. Define congestion.
13. Explain the bus bunching idea in relation to queue.
14. How can you determine if a channel is saturated?
15. What is a loop detector?
    1. How do loop detectors communicate with stop lights?
    2. How do large vehicles affect queue detectors?
16. When does departure rate depend upon arrival rate?
17. What are examples of service methods?
18. What is the effect of controls systems in series?
19. What is the difference between over and undersaturated queues? Give an example of an oversaturated queue.
20. What is the difference between uncapacitated and capacitated queues?
21. What does the variable ρ mean? What is meant when $\lambda > \mu$, $\lambda < \mu$
22. Why is there random congestion if the hourly flow is less than hourly capacity ($\rho < 1$)?
23. How can constraints be included in predicting how many vehicles are expected in a queue? (i.e. capacity of queue)
24. What happens if the average number of cars exceeds ramp capacity?
25. Do equations change for uncapacitated to capacitated queues?
26. How do you calculate the expected number of units in the system?
27. Explain why $E(n) \neq E(m)$ in words. Why is the average number of units in the system not the same as the mean queue length?
28. Is the expected number of units in the system an average number of units?
29. When μ and λ are random, what kind of queue is it?
30. What does Poisson refer to?
31. What happens to travel time as arrivals approach capacity?
32. Are most systems under or oversaturated in general?
33. What is the average service time?
34. Graph average travel time vs. rho. How does this relate to the volume delay function in Route Choice.
35. How accurate are these formulas for probability of waiting time and queue length?

Additional Problems available at the end of this section.

1. What is probability you will wait 15 minutes or more if on average 15 cars/min arrive and 14 cars/min are serviced
2. What is the average waiting time if there is a 600 vph arrival rate, and a 500 vph service rate.
3. If there is an arrival rate of 100 vehicles per hour, what is the service time?
4. How would you solve for an average time waiting in queue if the arrival rate is 400 vph and service rate is 450 vph?
5. If the arrival rate is 250 vph and the service rate is 600 vph, what is the time for vehicles waiting to get on the system?
6. With the arrival rate of 250 vph and service rate of 275 vph, ho wmuch free time is there on a ramp and how long is the average wait time.
7. Calculate the average number of vehicles (wait) in the system with x arrivals and y departures.

Homework

1. Deterministic Queueing

A. Draw a typical queuing input-output diagram (Newell Curve) (for one lane) consistent with observed data on a freeway bottleneck with an uncongested capacity of 1800 vehicles per hour per lane and an arrival rate that starts at 1600 vehicles per hour for 15 minutes, rises to 2000 vehicles per hour for 30 minutes, and drops back to 1600 vehicles per hour for the final 15 minutes.

B. Show how to compute delay on the graph.

C. Show the delay for the 500th vehicle

D. How many vehicles are in the congested region at time the 500th vehicle enters the congested region.

E. How many vehicles are in the congested region at the time the 500th vehicle leaves the front of the queue.

2. You are asked to model a freeway ramp meter with one lane of SOV traffic, which waits at the meter, and an HOV bypass lane that has no delay. Assume vehicles arrive with a random Poisson distribution, and the service rate (the rate at which the light turns green) is stochastic (it is determined by freeway conditions, which from the point of view of travelers on the ramp is random, with a negative exponential distribution). In the peak hour, 800 vehicles arrive at the ramp, of which 100 are high occupancy vehicles with 2 passengers each. The traffic light is timed to turn green on average once every 4 seconds, and is only green long enough to let one vehicle through. Assume the ramp can hold as many vehicles as want to use it.

A. Determine the percent of time the ramp is empty, not empty

B. How much time will each HOV user save?

C. How much time does each SOV user save because the HOV users don't wait in the queue?

D. Determine the average queue length

E. How often does the queue exceed 15 cars?

3. For the M\/D\/1 queuing model,

A. What is the arrival distribution?

B. What is the departure distribution?

C. How many servers are there?

D. Provide two examples where the model might apply (one transportation and one non-transportation).

4. Under what conditions are ramp meters beneficial, detrimental?

5. Who wins and who loses when ramp meters are installed, illustrate your argument with an example.

6. How might ramp meters be made more popular without reducing their effectiveness?

Videos

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Answers

Sample Problem 1: Queueing at a Tollbooth

$\rho = \frac{\lambda}{\mu} = (\frac{500}{700}) = 0.714$

Sample Problem 2: Queueing at a Ramp Meter

$\rho = 200/250 = 0.8$

Part (A)

What is the probability that it is half-full, empty, full?

Half-full

\(P(n)=\frac{(1-\rho)}{1-\rho^{N+1}}(\rho)^n \to P(20)=\frac{(1-0.8}{1-0.8^{40+1}(0.8)^{20}=0.00023\)

Empty

$P(n) = \frac{(1-\rho)}{1-\rho^{N+1}}(\rho)^n \to P(0) = \frac{(1-0.8)}{1-0.8^{40+1}}(0.8)^0 = 0.20$

Full

$P(n) = \frac{(1-\rho)}{1-\rho^{N+1}}(\rho)^n \to P(40) = \frac{(1-0.8)}{1-0.8^{40+1}}(0.8)^{40} = 2.65 * 10^{-5}$

Part (B)

How many cars do we expect on the ramp?

$E(n) = \frac{(\rho)}{(1-\rho)}\frac{1-(N+1)(\rho)^N+N\rho^{N+1}}{1-\rho^{N+1}} = \frac{(0.8)}{(1-0.8)}\frac{1-(40+1)(0.8)^{40}+40(0.8)^{40+1}}{1-(0.8)^{40+1}} = 4$

Sample Problem 3: Queueing at a Ramp Meter

$\lambda = \frac{3600}{8} = 450$

$\mu = \frac{3600}{6} = 600$

$\rho = \frac{\lambda}{\mu} = 0.75$

Part A

Probability of 5 cars,

\(P(n)=\frac{(1-\rho)}{1-\rho^{N+1}}(\rho)^n=\frac{(1-0.75)}{1-0.75^{16}}(0.75)^{15}=0.0033=0.33%\)

Part C

Expected number of vehicles on ramp in peak hour.

$E(n) = \frac{(\rho)}{(1-\rho)}\frac{1-(N+1)(\rho)^N+N\rho^{N+1}}{1-\rho^{N+1}} = \frac{(0.75)}{(1-0.75)}\frac{1-(15+1)(\rho)^{15}+15\rho^{16}}{1-\rho^16} = 2.81 \approx 3$

Additional Questions

This page titled 5.1: Queueing is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

5.2: Traffic Flow

Traffic Flow is the study of the movement of individual drivers and vehicles between two points and the interactions they make with one another. Unfortunately, studying traffic flow is difficult because driver behavior cannot be predicted with one-hundred percent certainty. Fortunately, drivers tend to behave within a reasonably consistent range; thus, traffic streams tend to have some reasonable consistency and can be roughly represented mathematically. To better represent traffic flow, relationships have been established between the three main characteristics: (1) flow, (2) density, and (3) velocity. These relationships help in planning, design, and operations of roadway facilities.

Traffic flow theory

Time-Space Diagram

Traffic engineers represent the location of a specific vehicle at a certain time with a time-space diagram. This two-dimensional diagram shows the trajectory of a vehicle through time as it moves from a specific origin to a specific destination. Multiple vehicles can be represented on a diagram and, thus, certain characteristics, such as flow at a certain site for a certain time, can be determined.

Road Space Requirements.png

Flow and density

Flow (q) = the rate at which vehicles pass a fixed point (vehicles per hour) ,

\[t_{measured}=Average \text{ } measured \text{ } time \text{ } headway\)

\[q=\frac{3600 N}{t_{measured}}

Density (Concentration) (k) = number of vehicles (N) over a stretch of roadway (L) (in units of vehicles per kilometer)

$$k = \frac{N}{L} \tag{5.2.1}$$

where:

- N = number of vehicles occupying a highway segment of length {\displaystyle L}L
- q = equivalent hourly flow
- L = length of roadway
- k = density

Time-space diagram showing trajectories of vehicles over time and space

Speed

Measuring speed of traffic is not as obvious as it may seem; we can average the measurement of the speeds of individual vehicles over time or over space, and each produces slightly different results.

Time mean speed

Time mean speed ($\bar{t}$) = arithmetic mean of speeds of vehicles passing a point

$$\bar{v}_t = \frac{1}{N}\sum_{n=1}^{N} v_n \tag{5.2.2}$$

Space mean speed

Space mean speed ($\bar{v_s}$) is defined as the harmonic mean of speeds passing a point during a period of time. It also equals the average speeds over a length of roadway.

\[\bar v_t=\dfrac{N}{\sum_{n=1}^N \frac{1}{v_n}}\]

Relating time and space mean speed

Note that the time mean speed is average speed past a point as distinct from space mean speed which is average speed along a length.

The two speeds are related as

$$\bar{v}_t = \bar{v}_s + \frac{\sigma_s^2}{\bar{v}_s} \tag{5.2.3}$$

The time mean speed higher than the space mean speed, but the differences vary with the amount of variability within the speed of vehices. At high speeds (free flow), differences are minor, whereas in congested times, they might differ a factor 2.

Headway

The following definitions give what is referred to as the *brutto* gap (Asela) (Italian for *gross*), in contrast to *netto* gaps (Italian for *net*). *Netto* gaps give the distance or time between the rear bumper of a vehicle and the front bumper of the next.

Time headway

Time headway (h_t) = difference between the time when the front of a vehicle arrives at a point on the highway and the time the front of the next vehicle arrives at the same point (in seconds)

Average Time Headway ($\bar{h}_t$) = Average Travel Time per Unit Distance * Average Space Headway

$$\bar{h}_t = \bar{t} * \bar{h}_s \tag{5.2.4}$$

Space headway

Space headway (h_s) = difference in position between the front of a vehicle and the front of the next vehicle (in meters)

Average Space Headway ($\bar{h}_s$)= Space Mean Speed * Average Time Headway

$$\bar{h}_s = \bar{v}_s * \bar{h}_t \tag{5.2.5}$$

Note that density and space headway are related:

\[k=\frac{1}{\bar{h_s}\]

Fundamental Diagram of Traffic Flow

The variables of flow, density, and space mean speed are related definitionally as:

$$q = k\bar{v}_s \tag{5.2.6}$$

Traditional Model (Parabolic)

Properties of the traditional fundamental diagram.

- When density on the highway is zero, the flow is also zero because there are no vehicles on the highway
- As density increases, flow increases
- When the density reaches a maximum jam density (k_j), flow must be zero because vehicles will line up end to end
- Flow will also increase to a maximum value (q_m), increases in density beyond that point result in reductions of flow.
- Speed is space mean speed.
- At density = 0, speed is freeflow (v_f). The upper half of the flow curve is uncongested, the lower half is congested.
- The slope of the flow density curve gives speed. Rise/Run = Flow/Density = Vehicles per hour/ Vehicles per km = km / hour

Traditional fundamental diagram of traffic

Observation (Triangular or Truncated Triangular)

Actual traffic data is often much noisier than idealized models suggest. However, what we tend to see is that as density rises, speed is unchanged to a point (capacity) and then begins to drop if it is affected by downstream traffic (queue spillbacks). For a single link, the relationship between flow and density is thus more triangular than parabolic. When we aggregate multiple links together (e.g. a network), we see a more parabolic shape.

Flow varying by density on detector 688 (I-94 at 49th/53rd Avenue in Minneapolis) on November 1st 2000

Microscopic and Macroscopic Models

Models describing traffic flow can be classed into two categories: microscopic and macroscopic. Ideally, macroscopic models are aggregates of the behavior seen in microscopic models.

Microscopic Models

Microscopic models predict the following behavior of cars (their change in speed and position) as a function of the behavior of the leading vehicle.

Traffic phases in a the microscopic fundamental diagram (truncated triangular)

Macroscopic Models

Macroscopic traffic flow theory relates traffic flow, running speed, and density. Analogizing traffic to a stream, it has principally been developed for limited access roadways (Leutzbach 1988). The fundamental relationship “q=kv” (flow (q) equals density (k) multiplied by speed (v)) is illustrated by the fundamental diagram. Many empirical studies have quantified the component bivariate relationships (q vs. v, q vs. k, k vs. v), refining parameter estimates and functional forms (Gerlough and Huber 1975, Pensaud and Hurdle 1991; Ross 1991; Hall, Hurdle and Banks 1992; Banks 1992; Gilchrist and Hall 1992; Disbro and Frame 1992).

The most widely used model is the Greenshields model, which posited that the relationships between speed and density is linear. These were most appropriate before the advent of high-powered computers enabled the use of microscopic models. Macroscopic properties like flow and density are the product of individual (microscopic) decisions. Yet those microscopic decision-makers are affected by the environment around them, i.e. the macroscopic properties of traffic.

While traffic flow theorists represent traffic as if it were a fluid, queueing analysis essentially treats traffic as a set of discrete particles. These two representations are not-necessarily inconsistent. The figures to the right show the same 4 phases in the fundamental diagram and the queueing input-output diagram. This is discussed in more detail in the next section.

Traffic phases in the queueing cumulative input-output (Newell) diagram

Examples

Example 1: Time-Mean and Space-Mean Speeds

Given five observed velocities (60 km/hr, 35 km/hr, 45 km/hr, 20 km/hr, and 50 km/hr), what is the time-mean speed and space-mean speed?

Solution

Time-Mean Speed:

$\bar{v}_t = \frac{1}{5}(60 + 35 + 45 + 20 + 50) = 42$

Space-Mean Speed:

\(\bar v_s=\frac{N}{\sum_{n=1}^N\dfrac{1}{v_n}=\frac{5}{\dfrac{1}{60}+\dfrac{1}{35}+\dfrac{1}{45}+\dfrac{1}{20}+\dfrac{1}{50}}=36.37\)

The time-mean speed is 42 km/hr and the space-mean speed is 36.37 km/hr.

Example 2: Computing Traffic Flow Characteristics

Given that 40 vehicles pass a given point in 1 minute and traverse a length of 1 kilometer, what is the flow, density, and time headway?

Solution

Compute flow and density:

$q = \frac{3600(40)}{60s} = 2400\ veh/hr$

$k = \frac{40}{1} = 40\ veh/km$

Find space-mean speed:

$q = k\bar{v}_s = 2400 = 40\bar{v}_s$

$\bar{v}_s = 60km/hr$

Compute space headway:

$k = 40 = \frac{1}{\bar{h}_s}$

$\bar{h}_s = 0.025km = 25m$

Compute time headway:

$\bar{h}_s = \bar{v}_s * \bar{h}_t = 25 = (60 * 1000/3600)\bar{h}_t$

$\bar{h}_t = 1.5s$

The time headway is 1.5 seconds.

EXAMPLE 3: The spot speeds (expressed in km/hr) observed at a road section are 66, 62, 45, 79, 32, 51,56,60,53 and 49. The median speed (expressed in km/hr) is .

Solution: Median speed is the speed at the middle value in series of spot speeds that are arranged in ascending order. 50% of speed values will be greater than the median 50% will be less than the median. Ascending order of spot speed studies are 32,39,45,51,53,56,60,62,66,79

Median speed = (53 +56)/2=54.5 km/hr.

Thought Question

Problem

Microscopic traffic flow simulates the behaviors of individual vehicles while macroscopic traffic flow simulates the behaviors of the traffic stream overall. Conceptually, it would seem that microscopic traffic flow would be more accurate, as it would be based on driver behavior than simply flow characteristics. Assuming microscopic simulation could be calibrated to truly account for driver behaviors, what is the primary drawback to simulating a large network?

Solution

Computer power. To simulate a very large network with microscopic simulation, the number of vehicles that needed to be assessed is very large, requiring a lot of computer memory. Current computers have issues doing very large microscopic networks in a timely fashion, but perhaps future advances will do away with this issue.

Sample Problem

Four vehicles are traveling at constant speeds between sections X and Y (280 meters apart) with their positions and speeds observed at an instant in time. An observer at point X observes the four vehicles passing point X during a period of 15 seconds. The speeds of the vehicles are measured as 88, 80, 90, and 72 km/hr respectively. Calculate the flow, density, time mean speed, and space mean speed of the vehicles.

Answer available at the end of this section.

Variables

- d_n = distance of nth vehicle
- t_n = travel time of nth vehicle
- v_n = speed (velocity) of nth vehicle
- $h_{t,nm}$ = time headway between vehicles n and m
- $h_{s,nm}$ = space (distance) headway between vehicles n and m
- q = flow past a fixed point (vehicles per hour)
- N = number of vehicles
- $t_{measured}$ = time over which measurement takes place (number of seconds)
- t = travel time
- k = density (vehicles per km)
- L = length of roadway section (km)
- v_t = time mean speed
- v_s = space mean speed
- v_f = freeflow (uncongested speed)
- k_j = jam density
- q_m = maximum flow

Key Terms

- Time-space diagram
- Flow, speed, density
- Headway (space and time)

- Space mean speed, time mean speed
- Microscopic, Macroscopic

Supplementary Reading

- Revised Monograph on Traffic Flow Theory

Videos

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Answers

Sample Problem

Flow

$q = N(\frac{3600}{t_{measured}}) = 4(\frac{3600}{15}) = 960\ veh/hr$

Density

\(k=\frac{N}{L}=\frac{4*1000}{280}=14.2 \text{ } veh/km

Time Mean Speed

$\bar{v}_t = \frac{1}{N}\sum_{n=1}^{N} v_n = \frac{1}{4}(72 + 90 + 80 + 88) = 82.5\ km/hr$

Space Mean Speed

\(\bar v_s=\frac{N}{\sum_{n=1}^N \frac{1}{v_i}}=\frac{4}{\frac{1}{72} \frac{1}{90} \frac{1}{80} \frac{1}{88}}=81.86

$t_i = L/v_i$

$t_A = L/v_A = 0.28/88 = 0.00318hr$

$t_B = L/v_B = 0.28/80 = 0.00350hr$

$t_C = L/v_C = 0.28/90 = 0.00311hr$

$t_D = L/v_D = 0.28/72 = 0.00389hr$

$\bar{v}_s = \frac{NL}{\sum_{n=1}^{N} i_n} = \frac{4*0.28}{(0.00318+0.00350+0.00311+0.00389)} = 81.87\ km/hr$

This page titled 5.2: Traffic Flow is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

5.3: Queueing and Traffic Flow

Queueing and Traffic Flow is the study of traffic flow behaviors surrounding queueing events. It combines elements from the previous two sections, Queueing and Traffic Flow.

Queueing and Traffic Flow

The first side figure illustrates a traffic bottleneck that drops the roadway from two lanes to one. It allows us to illustrate the changes in capacity, the changes in lane flow, and the stability in the total section flow. Over an extended period of time, by laws of conservation, flow through the bottleneck (q) must equal flow through the upstream section (Q).

$$\sum_t q = \sum_t Q \tag{5.3.1}$$

A bottleneck causes lane flow (Ql) to drop, but not section flow (Q), where:

$$Q = \sum_t Q_l \tag{5.3.2}$$

When researchers observe a “backward-bending” flow-travel time curve, (which we will tackle later in the section) this is occasionally what they are seeing.

$$\sum_t q = \sum_t Q \geq \sum_t Q_l \tag{5.3.3}$$

An illustrative plan of a highway bottleneck

Traffic phases in the queueing cumulative input-output (Newell) diagram

The lane flow drops upstream of a bottleneck when demand for a bottleneck exceeds the capacity of the bottleneck. Thus, the same level of flow can be observed at two different speeds on an upstream lane. The first is at freeflow speed in uncongested conditions and the second is at a lower speed when the downstream bottleneck is at capacity. In other words, congestion (a queue) forms when Q > q_{max} for any period of time. Ultimately, what enters the queue must eventually exit; otherwise the queue will grow to infinity. This is illustrated with the queueing input-output (IO) diagram shown in the second side figure. We can identify four distinct phases in traffic.

- Phase 1 is the uncongested phase when there is no influence of the increasing density on the speeds of the vehicles. The speed does not drop with the introduction of newer vehicles onto the freeway.
- Phase 2 finds the freeway cannot sustain the speed with injection of newer vehicles into the traffic stream. The density increases while speed falls, maintaining the flow.
- Phase 3 shows decreased speed and decreased flows. This is caused by very low speeds cause the queue discharge to drop (slightly) at an active bottleneck, or a queue from a downstream bottleneck may be constraining the flow. As reported by many researchers (Banks, 1991; Hall and Agyemang-Duah, 1991; Persaud and Hurdle, 1991; Cassidy and Bertini, 1999), the reduced bottleneck capacity after breakdown ranges from 0% to 8%. A majority of papers conclude that the section capacity reduction a queue presents is negligible. (As noted above, lane flow upstream of a bottleneck does drop, but this is due to the downstream bottleneck). This is a second source of the “backward-bending” phenomenon.
- Phase 4 is the recovery phase. During this phase, the density of traffic starts decreasing and speed starts increasing.

Traffic phases in a the microscopic fundamental diagram (truncated triangular)

The circled region, ‘A’ (in the third side figure) is typically where we start observing ‘freeway breakdown’. In other words, this region occurs when upstream flow exceeds some critical downstream capacity at a specific point and there is a drop in speed. In the circled region, ‘B’, flow drops are associated with very low speeds. The phenomenon is particularly evidenced by the formation of queues upstream of where the breakdown occurs and a low discharge rate of vehicles due to sustained low speeds.

Traditional queues have “servers”. For example, the check-out line at the grocery store can serve one customer every 5 minutes, or one item every 10 seconds, etc. A conveyor belt can serve so many packages per hour. Capacity is often referred to as belonging

to the road. However when we talk about road capacity, it is really a misnomer, as capacity is located in the driver, more precisely in the driver's willingness and ability to drive behind the driver ahead. If drivers were willing and able to drive behind the vehicle ahead with no gap (spacing between vehicles), and that driver was driving behind the driver ahead of him with no gap, and so on, at high speeds, many more vehicles per hour could use the road. However, while some compression of vehicles occurs in heavy traffic, this situation is unstable because a driver will tap the brakes or even let-up on the accelerator for any number of reasons (to change lanes, to respond to someone else trying to change lanes, because he sees an object in the road, to limit the forces when rounding a corner, etc.). This by definition lowers his speed, which in turn will lower the flow. Risk-averse drivers behind him will slow down even more (the driver has established some unpredictability in behavior, it is reasonable for other drivers to establish an even larger gap to accommodate the unpredictable driver's behavior, especially given there is a reaction time between the lead driver's actions, and the following driver's perception (lead vehicle is slowing), decision (must brake), action (tap the brakes), the vehicle's response to the action (tighten brakes on wheel)). In this way, maximum flow possible, our capacity (\(q_{max\)), is a function of the drivers. The road shapes the driver's willingness to take risks. Drivers will slow down around curves, vehicles may have difficulty accelerating uphill, or even from a slower speed (and even if they don't, driver's may provide insufficient fuel before they realize they are going too slow by not giving the vehicle enough gas), merges take time to avoid collision, etc.

Clearly, different drivers and different vehicles (for instance racecars or taxis) could increase the maximum flow through the bottleneck (q_{max}). It is better to think of capacity as a maximum sustainable flow (over an extended period of time), given typical drivers' willingness to follow (subject to highway geometrics and environmental conditions) and their vehicles' ability to respond to decisions. A series of aggressive drivers may exceed this 'capacity' for a short period of time, but eventually more cautious drivers will even out the function.

The IO diagram lets us understand delay in a way that the fundamental diagrams of traffic flow don't easily allow. The first point to note about the IO diagram is that delay differs for each driver. The average delay can be measured easily (the total area in the triangle is the total delay, the average delay is just that triangle divided by the number of vehicles). The variation (or standard deviation) can also be measured with some more statistics. As the total number of vehicles increases, the average delay increases.

The second point to note about the IO diagram is that the total number of queued vehicles (the length of the queue) can also be easily measured. This also changes continuously; the length of the queue rises and falls with changes in the arrival rate (the tail of the queue edge of the shockwave, where travel speed changes suddenly). Shockwaves indicate a change in state, or speed that suddenly occurs. One such shockwave is found where vehicles reach the back of queues. The arrival rate, coupled with the queue clearance rate, will tell you where the back of the queue is. This is interesting and is where the traveler first suffers delay – but it is not the source of the delay.

Only if the arrival rate exactly equals the departure rate would we expect to see a fixed length of the queue. If the queue results from a management practice (such as ramp metering), we can control the departure rate to match the arrival rate, and ensure that the queue remains on the ramp and doesn't spill over to neighboring arterials. However that observation suggests that congested "steady state" is likely to be a rare phenomenon, since in general the arrival rate does not equal the service rate.

Queueing analysts often make a simplifying assumption that vehicles stack vertically (queues take place at a point). This is of course wrong, but not too wrong. The resulting travel time is almost the same as if the queue were measured over space. The difference is that the time required to cover distance is included when we make the better assumption, even under freeflow conditions it takes time to travel from the point where a vehicle entered the back of the queue to the point where it exits the front. We can make that correction, but when queueing is taking place, that time is often small compared to the time delayed by the queue. We also assume a first-in, first-out logic, though again this can be relaxed without distracting from the main point. Another assumption we will make for exposition is that this is a deterministic process; vehicles arrive in a regular fashion and depart in a regular fashion. However sometimes vehicles bunch up (drivers are not uniform), which leads to stochastic arrivals and departures. This stochastic queueing can be introduced, and will in general increase the measured delay.

Hypercongestion

It has been observed that the same flow can be achieved on many links at two different speeds. Some call this the "backward-bending" phenomenon (Hau 1998, Crozet and Marlot 2001). The queueing analysis framework also has implications for "hypercongestion" or "backward-bending" flow-travel time curves, such as shown in the figure to the right. Recall we identified two sources for "backward-bending" speed-flow relationships. The first has to do with the point of observation. Observing the lane flow upstream of a bottleneck gives the impression of a backward bending relationship, but this disappears at the bottleneck itself. Under any given demand pattern, flow and speed are a unique pair. When demand is below the downstream active bottleneck's capacity, a flow on an upstream link can be achieved at high speed. When demand is above the downstream active bottleneck's capacity, the same flow on the upstream link can only be achieved at a low speed because of queueing. The second has to do

with a capacity drop at the bottleneck itself under congested conditions. However, much research reports that this drop is slight to non-existent.

As in the bottleneck, we define lowercase q to be flow (vehicles per hour) departing the front of the bottleneck and uppercase Q to be flow arriving at the back of the bottleneck. We also define k to be density (vehicles per kilometer), v to be speed (kilometers per hour), and s to be service rate (seconds per vehicle). The fundamental diagrams of traffic flow (q-k-v curves) represent a model of traffic flow as stylized in the traditional textbook representation of the fundamental diagram of traffic flow. The reasons why q should drop as k increases beyond a certain point at an isolated bottleneck are unclear. In other words, why should flow past a point drop just because the number of vehicles behind that point increases? Why should leading traffic be influenced by the behavior of following traffic?

If traffic behaves as a queue through a bottleneck (illustrated above), we should consider reasons why traffic flow departing the queue would not stay at its maximum.

One reason is if the vehicles in the queue could not travel fast enough so that the front of the following car could not reach the point of the front of the leading car in the time allotted the service rate. If a lane serves 1800 vehicles per hour, it serves 1 vehicle every 2 seconds, we say there is a 2 second service rate. If there were a 10 meter spacing between the vehicles (including the vehicle length plus a physical gap), this implies that the service rate would be worse than 1 vehicle every 2 seconds only if it took longer than 2 seconds to travel 10 meters (i.e. speed < 18 km /hour). This is very slow traffic and may properly be called hypercongestion. Why would traffic get that slow if freeflow speed is 120 km / hour? In general traffic departing the front of the queue won't be traveling that slowly, as the shockwave that reduces speed (the wave of red brake lights) has moved to the back of the queue.

A second reason the bottleneck flow might drop is if the departure flow is affected by external sources, namely a bottleneck further downstream spilling backwards. If the queue at a downstream bottleneck becomes sufficiently long, it will reduce the number of vehicles that can depart the upstream bottleneck. This is because the first bottleneck is no longer controlling, the downstream bottleneck is.

A third reason that bottleneck flow might drop is if the bottleneck is not being fully served (slots for cars are going unfilled). If vehicles are separated by large time headways, then the bottleneck might lose capacity. Thus, if vehicles choose large gaps, the bottleneck flow might drop. However, in congested situations, vehicles tend to follow more closely, not less closely.

For an isolated bottleneck, the departure flow remains (essentially) a constant and arrival flow varies. As more vehicles arrive at the back of the queue, the expected wait at the queue increases. All vehicles will eventually be served. In general, there is no practical constraint on the number of vehicles arriving at the back of the queue, but there is a maximum output flow in vehicles per unit time. Examining traffic upstream of the bottleneck is interesting, but does not get to the root of the problem – the bottleneck itself. This view of hypercongestion is thus not inconsistent with the conclusion drawn by Small and Chu (1997) that the hypercongested region is unsuitable for use as a supply curve in congestion pricing analyses.

BackwardBending.png

Thought Question

Problem

Bottlenecks are shown here to be the main cause of traffic flow congestion and queueing. The examples given are of lane drops, which are not very common (some drivers think they're more common than they actually are). Is this the only way a bottleneck can form?

Solution

Absolutely not. A bottleneck is defined as a constriction of traffic flow, where demand exceeds available capacity. Infrastructure-wise, a bottleneck would be where a lane drop is present, where lane widths decrease, or a sight obstruction lowers the natural free-flow speed. However, bottlenecks can also form because of traffic. Chaotic traffic, primarily where lane changing is occurring, is a prime source of a bottleneck. On-ramps, off-ramps, and weaving areas are the most common examples, but there are many more. Think about how often you have been stuck in standstill traffic where a lane drop is not present.

Additional Questions

1. Why were ideal models of flow vs. density assumed for so long if data did not support it?
2. How do we estimate density of road (in planning or as an after-evaluation?)
3. Explain why the graph of density vs. speed at a bottleneck is not linear using queueing theory.
4. Define time headway and space headway.
5. Why do we need two different speed measures, define each, which is more appropriate? How are these measures related?
6. What are the 3 fundamental principles of traffic flow? What are their relationships?
7. What does the slope on the space time diagram tell you?
8. What is the difference between looking at traffic as a particle rather than a flow?
9. Define density.
10. As density increases, what happens to flow?
11. Where are time-space diagrams used?
12. What is space headway on a space-time diagram? What is time headway?
13. What does high density mean?
14. Draw a realistic speed vs. density graph? Compare it with the traditional model.
15. Describe headway speed?
16. How does a bottleneck affect traffic flow upstream of the point?
17. What might cause a falloff of flow as density rises?
18. Why does density reach a maximum when flow is zero?
19. How do these analysis of mathematical relationships between q,k,v help in planning, design, and operations of road?

Variables

- q - Flow through the bottleneck
- Q - Flow through the upstream section
- Q_l - Lane Flow
- t - time

Key Terms

- Bottleneck
- Queueing
- Traffic Flow
- Lane Flow
- Lane Flow Drops
- Flow-Travel Time Curve
- "Backwards-Bending"
- IO
- Hypercongestion

This page titled 5.3: Queueing and Traffic Flow is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

6: TRAFFICE CONTROL

6.1: Shockwaves
6.2: Traffic Signals
6.3: Traffic Control Devices
6.4: Metering and Analogs

This page titled 6: Traffice Control is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

6.1: Shockwaves

Shockwaves are byproducts of traffic congestion and queueing. They are transition zones between two traffic states that move through a traffic environment like, as their name states, a propagating wave. On the urban freeway, most drivers can identify them as a transition from a flowing, speedy state to a congested, standstill state. However, shockwaves are also present in the opposite case, where drivers who are idle in traffic suddenly are able to accelerate. Shockwaves are one of the major safety concerns for transportation agencies because the sudden change of conditions drivers experience as they pass through a shockwave often can cause accidents.

Visualization

While most people have probably experienced plenty of traffic congestion first hand, it is useful to see it systematically from three different perspectives: (1) That of the driver (with which most people are familiar), (2) a birdseye view, and (3) a helicopter view. Some excellent simulations are available here, please see the movies:

Visualization with the TU-Dresden 3D Traffic Simulator

This movie shows traffic jams without an "obvious source" such as an on-ramp, but instead due to randomness in driver behavior: Shockwave traffic jams recreated for first time

Analysis of shockwaves

Shockwaves can be seen by the cascading of brake lights upstream along a highway. They are often caused by a change in capacity on the roadways (a 4 lane road drops to 3), an incident, a traffic signal on an arterial, or a merge on freeway. As seen above, just heavy traffic flow alone (flow above capacity) can also induce shockwaves. In general, it must be remembered that capacity is a function of drivers rather than just being a property of the roadway and environment. As the capacity (maximum flow) drops from C_1 to C_2, optimum density also changes. Speeds of the vehicles passing the bottleneck will of course be reduced, but the drop in speed will cascade upstream as following vehicles also have to decelerate.

The figures illustrate the issues. On the main road, far upstream of the bottleneck, traffic moves at density k_1, below capacity (k_{opt}). At the bottleneck, density increases to accommodate most of the flow, but speed drops.

Shockwave Diagram

Shockwave Diagram

Shockwave Math

Shockwave speed

If the flow rates in the two sections are q_1 and q_2, then $q_1 = k_1 v_1$ and $q_2 = k_2 v_2$.

$$v_w = \frac{q_2 - q_1}{k_2 - k_1} \tag{6.1.1}$$

Relative speed

With v_1 equal to the space mean speed of vehicles in area 1, the speed relative to the line w is:

$$v_{r1} = v_1 - v_w \tag{6.1.2}$$

The speed of vehicles in area 2 relative to the line w is \(v_{r2}=v-v_w\]

Boundary crossing

The number of vehicles crossing line 2 from area 1 during time period t is

$$N_1 = v_{r1}k_1t = (v_1 - v_w)k_1t \tag{6.1.3}$$

and similarly

$$N_2 = v_{r2}k_2t = (v_2 - v_w)k_2t \tag{6.1.4}$$

By conservation of flow, the number of vehicles crossing from left equals the number that crossed on the right

$N_1 = N_2$

so:

$v_2k_2 - v_1k_1 = v_w(k_2 - k_1)$

or

$q_2 - q_1 = v_w(k_2 - k_1)$

which is equivalent to

$v_w = \frac{q_2 - q_1}{k_2 - k_1}$

Examples

Example 1

The traffic flow on a highway is $q_1 = 2000\ veh/hr$ with speed of $v_1 = 80\ km/hr$. As the result of an accident, the road is blocked. The density in the queue is $k_2 = 275\ veh/km$. (Jam density, vehicle length = 3.63 meters).

- (A) What is the wave speed (v_w)?
- (B) What is the rate at which the queue grows, in units of vehicles per hour (q)?

Solution

(A) At what rate does the queue increase?

1. Identify Unknowns:

$k_1 = q/v_1 = 2000/80 = 25\ veh/km$

$v_2 = 0,\ \ q_2 = k_2v_2 = 0$

2. Solve for wave speed (v_w)

$v_w = \frac{q_2 - q_1}{k_2 - k_1} - \frac{0 - 2000}{275 - 25} = -8km/hr$

Conclusion: the queue grows against traffic

(B) What is the rate at which the queue grows, in units of vehicles per hour?

$N_1 = (v_1 - v_w)k_1t = (v_2 - v_w)k_2t = N_2$

dropping t (let t = 1)

$v_1k_1 - v_wk_1 = v_2k_2 - v_wk_2$

$q_1 - v_wk_1 = q_2 - v_wk_2$

$2000 - (-8) * 25 = 0 - (-8) * 275$

$2200\ veh/hr = 2200\ veh/hr$

Thought Question

Problem

Shockwaves are generally something that transportation agencies would like to minimize on their respective corridor. Shockwaves are considered a safety concern, as the transition of conditions can often lead to accidents, sometimes serious ones. Generally, these transition zones are problems because of the inherent fallibility of human beings. That is, people are not always giving full attention to the road around them, as they get distracted by a colorful billboard, screaming kids in the backseat, or a flashy sports car in the adjacent lanes. If people were able to give full attention to the road, would these shockwaves still be causing accidents?

Solution

Yes, but not to the same extent. While accidents caused by driver inattentiveness would decrease nearly to zero, accidents would still be occurring between different vehicle types. For example, in a case where conditions change very dramatically, a small car (say, a Beetle) would be able to stop very quickly. A semi truck, however, is a much heavier vehicle and would require a longer distance to stop. If both were moving at the same speed when encountering the shockwave, the truck may not be able to stop in time before smashing into the vehicle ahead of them. That is why most trucks are seen creeping along through traffic with very big gaps ahead of them.

Sample Problem

Flow on a road is $q_1 = 1800veh/hr/lane$, and the density of $k_1 = 14.4veh/km/lane$. To reduce speeding on a section of highway, a police cruiser decides to implement a rolling roadblock, and to travel in the left lane at the speed limit ($v_2 = 88km/hr$) for 10 km. No one dares pass. After the police cruiser joins, the platoon density increases to 20 veh/km/lane and flow drops. How many vehicles (per lane) will be in the platoon when the police car leaves the highway?

How long will it take for the queue to dissipate?

Answer available at the end of this section.

Additional Questions

1. What example or explanation can be given for shockwaves on freeways?
2. Explain the 4 phases of traffic diagram and why each has the slope it has.
3. How do microscopic and macroscopic models of traffic flow differ?
4. Why does flow decrease in phase 3 of the traffic phase diagram?
5. What types of events can create rolling roadblocks?
6. How is speed calculated from a flow v. density diagram?
7. What is a shockwave and how do you compute its speed?
8. Can service rate decrease with increasing congestion?
9. What factors affect the wave speed?
10. What does the slope of the shockwave represent in the graph?
11. Is it possible to increase the density of the queue and still have the shockwave going downstream? What would be a an example of this?
12. Why is the bottleneck QK curve projected on the mainroad QB curve to find the speed?
13. What is Greenshields model, does it accurately predict traffic flow?

14. What is the phase in traffic flow, after density is at its max, called?
15. What happens to speed when the density increases?
16. What does the slope represent in a traffic phase diagram
17. Give an example of density rising and speed dropping
18. What is v_w?
19. What does it mean when the flow rate begins to decrease in the traffic phase diagram?
20. During the bottleneck phase, what happens to flow as density increases? (flow remains constant)
21. How can you tell if the queue grows with or against traffic?
22. Does the bottleneck always slow traffic?
23. How can traffic avoid the bottleneck?
24. How do incidents and rolling roadblocks affect shockwaves differently?
25. How do speed, density, and flow change during the peak hour
26. List six examples where gaps are important.

Homework

1. The upstream traffic flow on a road with two lanes in the relevant direction is q_1= 1800 veh/hr with freeflow speed of v_1=90 km/hr. As the result of a crash, one lane is blocked and flow is reduced to q_2 = 900 veh/hr. The density in the queue is k_2 = 70 veh/km. The value of time is $10/hour/vehicle

A. What is the shockwave speed (v_w)?

B. If this continues for 1 hour, before the lane is cleared, how many vehicles long will the queue be?

C. How much did this crash cost society in terms of delay (travel time in excess of freeflow travel time)?

D. If a freeway service patrol (tow truck service) could clear the crash in 30 minutes instead of an hour, how much should society be willing to pay for that service?

E. If a variable message sign placed 2 km upstream of the crash site instantly (at the moment of the crash) advises travelers about a diversion to a frontage road 1 km upstream of the crash site, how valuable is the freeway service patrol (how much should society be willing to pay under these circumstances)?

2. A fixed-time ramp meter with unlimited storage space can serve 1 vehicle every 5 seconds. Vehicles approach the back of the ramp queue at 80 km/hr at a rate of 1 vehicle every 4 seconds. The density in the queue is 275 veh/km.

A. What is the wave speed?

B. What is the rate at which the queue grows, in units of vehicles per hour (q)?

Variables

- q - flow
- c - capacity (maximum flow)
- k - density
- v - speed
- v_r - relative speed (travel speed minus wave speed)
- v_w - wave speed
- N - number of vehicles crossing wave boundary

Key Terms

- Shockwaves
- Time lag, space lag

Videos

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Answers

Sample Problem

Solve for Unknowns:

Original speed

$v_1 = \frac{q_1}{k_1} = \frac{1800}{14.4} = 125 km/hr$

Flow after police cruiser joins

$q_2 = k_2 v_2 = 88 * 20 = 1760 veh/h$

Step 1

Calculate the wave velocity:

$v_w = \frac{q_2 - q_1}{k_2 - k_1} = \frac{1760 - 1800}{20 - 14.4} = -7.14 km/hr$

Step 2

Determine the growth rate of the platoon (relative speed)

$v_{r2} = v_2 - v_w = 88 - (-7.14) = 95.1 km/hr$

Step 3

Determine the time spent by the policee cruiser on the highway

$t = d/v = 10 km / 88 km/hr = 0.11 hr = 6.8 minutes$

Step 4

Calculate the Length of platoon (not a standing queue)

Step 5

What is the rate at which the queue grows, in units of vehicles per hour?

$\Deltaq = q_1 - k_1 v_w = q_2 - k_2 v_w = 1800 - (14.4 * -7.14) = 1760 - (20 * -7.14) = 1902 veh/hr$

Step 6

The number of vehicles in platoon

$L_p k_2 = 10.46 km * 20 veh/km = 209.2 vehicles$

OR

$\Deltaqt = 1902 veh/hr * 0.11 hr = 209.2 veh$

How long will it take for the queue to dissipate?

(a) Where does the shockwave go after the police cruiser leaves? Just reverse everything?

$v_w = \frac{q_2 - q_1}{k_2 - k_1} = \frac{1800 - 1760}{14.4 - 20} = -7.14 km/hr$

Second wave never catches first.

(b) Return speed to 125, keep density @ 20? $q_2 = 20 * 125 = 2500$

$v_w = \frac{q_2 - q_1}{k_2 - k_1} = \frac{2500 - 1760}{20 - 20} = \infty km/hr$

Second wave instantaneously catches first

(c) Return speed to 125, keep density @ halfway between 14.4 and 20? $q_2 = 20 * 125 = 2500$

$v_w = \frac{q_2 - q_1}{k_2 - k_1} = \frac{2150 - 1760}{17.2 - 20} = 139 \ km/hr$

Second wave quickly catches first

$139 * t = 7.14 * (.11 + t)$

$(139 - 7.14)t = 7.14 * 11$

$t = 0.006h = .35\ min = 21\ sec$

(d) Drop upstream $q_u k$ to slow down formation curve.

$v_w = \frac{q_2 - q_1}{k_2 - k_1} = \frac{1780 - 1760}{14.24 - 20} = -3.47\ km/hr$

Second wave eventually catches first

(e) If $q_{upstream}$ falls below downstream capacity, wave dissipates

$v_w = \frac{q_2 - q_1}{k_2 - k_1} = \frac{1250 - 1760}{10 - 20} = +51\ km/hr$

Forward moving wave.

This page titled 6.1: Shockwaves is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

6.2: Traffic Signals

Traffic Signals are one of the more familiar types of intersection control. Using either a fixed or adaptive schedule, traffic signals allow certain parts of the intersection to move while forcing other parts to wait, delivering instructions to drivers through a set of colorful lights (generally, of the standard red-yellow (amber)-green format). Some purposes of traffic signals are to (1) improve overall safety, (2) decrease average travel time through an intersection, and (3) equalize the quality of services for all or most traffic streams. Traffic signals provide orderly movement of intersection traffic, have the ability to be flexible for changes in traffic flow, and can assign priority treatment to certain movements or vehicles, such as emergency services. However, they may increase delay during the off-peak period and increase the probability of certain accidents, such as rear-end collisions. Additionally, when improperly configured, driver irritation can become an issue. Traffic signals are generally a well-accepted form of traffic control for busy intersections and continue to be deployed. Other intersection control strategies include signs (stop and yield) and roundabouts. Intersections with high volumes may be grade separated.

Traffic signals can be pretimed, semi-actuated, or fully-actuated. Pretimed intersections have a fixed cycle length. This is easy to implement but can cause excessive delay at some intersections. Semi-actuated intersections have vehicle detectors on the minor roadway. When a vehicle approaches on the minor roadway, the detector receives a signal to change the light to green. In a fully-actuated intersection, all approaches have a detector. Each phase has an initial green light interval to provide time for standing vehicles to get through the intersection. This initial time is extended if the detector at the approach detects a car moving through the intersection. If there are no cars moving through the intersection for a given period of time, the light will change. This is called a "gap out". After the maximum amount of time has passed for the light to be green, the light will change even if there are still cars moving through the intersection. This is called a "max out".

A Traffic Light

Intersection Queuing

At an intersection where certain approaches are denied movement, queuing will inherently occur. Of the various queuing models, one of the more commons and simple ones is the D/D/1 Queuing Model. This model assumes that arrivals and departures are deterministic (D) and one departure channel exists. D/D/1 is quite intuitive and easily solvable. Using this form of queuing with an arrival rate λ and a departure rate μ, certain useful values regarding the consequences of queues can be computed.

One important piece of information is the duration of the queue for a given approach. This time value can be calculated through the following formula:

$$t_c = \frac{\rho r}{1-\rho} \tag{6.2.1}$$

Where:

- t_c = Time for queue to clear
- ρ = Arrival Rate divided by Departure Rate
- r = Red Time

With this, various proportions dealing with queues can be calculated. The first determines the proportion of cycle with a queue.

$$P_q = \frac{r + t_c}{C} \tag{6.2.2}$$

Where:

- P_q = Proportion of cycle with a queue
- C = Cycle Length

Similarly, the proportion of stopped vehicles can be calculated.

$$P_s = \frac{\lambda(r + t_C)}{\lambda(r+g)} = \frac{r + t_c}{C} = P_q \tag{6.2.3}$$

$$P_s = \frac{\lambda(r + t_C)}{\lambda(r + g)} = \frac{\mu t_C}{\lambda C} = \frac{t_C}{\rho C} \tag{6.2.4}$$

Where:

- P_s = Proportion of Stopped Vehicles
- g = Green Time

Therefore, the maximum number of vehicles in a queue can be found.

$Q_{max} = \lambda r$

Typical Signal Schedule and Traffic Flow Diagram, North-South across Market (1929) From Signal Timing Schedule for Traffic Control Plan, June 15, 1929. Attempted "green wave": 8.5mph on Market; 50 vara district: 10.5 mph north-south, 14.5 mph east-west; 100 vara district: 14.5mph north-south, 20.5mph east-west.

Intersection delay

Various models of intersection delay at isolated intersections have been put forward, combining queuing theory with empirical observations of various arrival rates and discharge times (Webster and Cobbe 1966; Hurdle 1985; Hagen and Courage 1992). Intersections on arterials are more complex phenomena, including factors such as signal progression and spillover of queues between adjacent intersections. Delay is broken into two parts: *uniform delay*, which is the delay that would occur if the arrival pattern were uniform, and *overflow delay*, caused by stochastic variations in the arrival patterns, which manifests itself when the arrival rate exceeds the service flow of the intersection for a time period.

Delay can be computed with knowledge of arrival rates, departure rates, and red times. Graphically, total delay is the product of all queues over the time period in which they are present.

$$D_t = \frac{\lambda r^2}{2(1 - \rho)} \tag{6.2.5}$$

Similarly, average vehicle delay per cycle can be computed.

$$d_{avg} = \frac{\lambda r^2}{2(1 - \rho)} \frac{1}{\lambda C} \tag{6.2.6}$$

$$d_{avg} = \frac{r^2}{2C(1 - \rho)} \tag{6.2.7}$$

From this, maximum delay for any vehicle can be found.

$d_{max} = r$

Level of Service

In order to assess the performance of a signalized intersection, a qualitative assessment called Level of Service (LOS) is assessed, based upon quantitative performance measures. For LOS, the performance measured used is average control delay per vehicle. The general procedure for determining LOS is to calculate lane group capacities, calculate delay, and then make a determination.

Lane group capacities can be calculated through the following equation:

$$c = s\frac{g}{C} \tag{6.2.8}$$

Where:

- c = Lane Group Capacity
- s = Adjusted Saturation Flow Rate

- g = Effective Green Length
- C = Cycle Length

Average control delay per vehicle, thus, can be calculated by summing the types of delay mentioned earlier.

$$d = (d_1(PF)) + d_2 + d_3 \tag{6.2.9}$$

If your intersection is D/D/X: $d = ((d_1(PF)) + d_3$

This is because there are no random arrivals.

If your intersection is M/D/X: $d = (d_1(PF)) + d_2 + d_3$

You might think that there would be no deterministic arrivals because the intersection is M/D/X, however, this is incorrect. d_1 can be thought of as the baseline for the intersection.

Where:

- d = Average Signal Delay per vehicle (sec)
- d_1 = Average Delay per vehicle due to uniform arrivals (sec) (equivalent to D_T in previous section)
- PF = Progression Adjustment Factor
- d_2 = Average Delay per vehicle due to random arrivals (sec)
- d_3 = Average delay per vehicle due to initial queue at start of analysis time period (sec)

Uniform delay can be calculated through the following formula:

$$d_1 = \frac{0.5C\left(1 - \frac{g}{C}\right)^2}{1 - \left[min(1, X)\frac{g}{C}\right]} \tag{6.2.10}$$

Where:

X = Volume/Capacity (v/c) ratio for lane group.

Similarly, random delay can be calculated:

$d_2 = 900T\left[(X - 1 + \sqrt{(X-1)^2 + \frac{8kIX}{cT}}\right]$

Where:

- T = Duration of Analysis Period (in hours)
- k = Delay Adjustment Factor that is dependent on signal controller mode
- I = Upstream filtering/metering adjustment factor

Overflow delay generally only applies to densely urban corridors, where queues can sometimes spill over into previous intersections. Since this is not very common (usually the consequence of a poorly timed intersection sequence, the rare increase of traffic demand, or an emergency vehicle passing through the area), it is generally not taken into account for simple problems.

Delay can be calculated for individual vehicles in certain approaches or lane groups. Average delay per vehicle for an approach A can be calculated using the following formula:

$$d_A = \frac{\sum_i d_i v_i}{\sum_i v_i} \tag{6.2.11}$$

Where:

- d_A = Average Delay per vehicle for approach A (sec)
- d_i = Average Delay per vehicle for lane group i on approach A (sec)
- v_i = Analysis flow rate for lane group i

Critical Lane Groups

For any combination of lane group movements, one lane group will dictate the necessary green time during a particular phase. This lane group is called the Critical Lane Group. This lane group has the highest traffic intensity (v/s) and the allocation of green time for each phase is based on this ratio.

The sum of the flow ratios for the critical lane groups can be used to calculate a suitable cycle length.

$$Y_c = \sum_{i=1}^{n} (\frac{v}{s})_{ci} \tag{6.2.12}$$

Where:

- Y_c = Sum of Flow Ratios for Critical Lane Groups
- $(v/s)_{ci}$ = Flow Ratio for Critical Lane Group i
- n = Number of Critical Lane Groups

Similarly, the total lost time for the cycle is also an element that can be used in the calculation of cycle length.

$$L = \sum_{i=1}^{n} (t_L)_{ci} \tag{6.2.13}$$

Where:

- L = Total lost Time for Cycle
- $(t_L)_{ci}$ = Total Lost Time for Critical Lane Group i

Cycle Length Calculation

Cycle lengths are calculated by summing individual phase lengths. Using the previous formulas for assistance, the minimum cycle length necessary for the lane group volumes and phasing plan can be easily calculated.

$$C_{min} = \frac{L * X_C}{X_c - \sum_{i=1}^{n} Y_i} \tag{6.2.14}$$

Where:

- C_{min} = Minimum necessary cycle length
- X_c = Critical v/c ratio for the intersection
- $v/s)_{ci}$ = Flow Ratio for Critical Lane Group
- n = Number of Critical Lane Groups

This equation calculates the minimum cycle length necessary for the intersection to operate at an acceptable level, but it does not necessarily minimize the average vehicle delay. A more optimal cycle length usually exists that would minimize average delay. Webster (1958) proposed an equation for the calculation of cycle length that seeks to minimize vehicle delay. This optimum cycle length formula is listed below.

$$C_{opt} = \frac{[(1.5L) + 5]}{\left(1.0 - \sum_{i=1}^{n} Y_i\right)} \tag{6.2.15}$$

Where:

C_{opt} = Optimal Cycle Length for Minimizing Delay

Green Time Allocation

Once cycle length has been determined, the next step is to determine the allocation of green time to each phase. Several strategies for allocating green time exist. One of the more popular ones is to distribute green time such that v/c ratios are equalized over critical lane groups. Similarly, v/c ratios can be found with predetermined values for green time.

$$X_i = \frac{v_i}{c_i} = \frac{v_i}{s_i * g_i/C} \tag{6.2.16}$$

Where:

X_i = v/c ratio for lane group i

With knowledge of cycle lengths, lost times, and v/s ratios, the degree of saturation for an intersection can be found.

$$X_c = \sum \frac{v_i}{s_i} \frac{C}{C-L} \tag{6.2.17}$$

Where:

X_c = Degree of saturation for an intersection cycle

From this, the total effective green for all phases can be computed.

$$\sum g_i = \sum \frac{v_i}{s_i} \frac{C}{X_c} = C - L \tag{6.2.18}$$

Second Method For Green Time

Another method for calculating the effective red and green time for a given cycle is to minimize the total delay of the intersection. By assuming the intersection is controlled based on D/D/1 queuing the above equations for total delay can be used. Since the light will contain at least two or more directions the total delay must be calculated for each direction, and then added together to determine the total delay of the intersection. For a two way intersection with opposing lights a and b

$$D_t = \frac{\lambda_a r_a^2}{2(1-\rho_a)} + \frac{\lambda_b r_b^2}{2(1-\rho_b)} \tag{6.2.19}$$

Also the effective red time is the cycle length minus the effective green time for the other directions.

$$C = r_a + g_a \tag{6.2.20}$$

$$g_a = r_b \tag{6.2.21}$$

By substituting the two above equation for cycle length and effective red time into the total delay equation, it can then be written with only one variable red time.

$$D_t = \frac{\lambda_a r_a^2}{2(1-\rho_a)} + \frac{\lambda_b (C-r_a)^2}{2(1-\rho_b)} \tag{6.2.22}$$

By taking the derivative and setting it equal to zero, the minimum effective red time can be calculated. The other directions effective red time, and the effective green times for each directions can then be calculated by using the two above equations involving the cycle length.

Demonstrations

- Flash animation: Signal Phasing (by Karen Dixon and Thomas Wall of Oregon State University)
- Flash animation: Signal Progression (by Karen Dixon and Thomas Wall of Oregon State University)

Examples

Example 1: Intersection Queuing

An approach at a pretimed signalized intersection has an arrival rate of 0.1 veh/sec and a saturation flow rate of 0.7 veh/sec. 20 seconds of effective green are given in a 60-second cycle. Provide analysis of the intersection assuming D/D/1 queuing.

Solution

Traffic intensity, ρ, is the first value to calculate.

$\rho = \frac{\lambda}{\mu} = \frac{0.1}{0.7} = 0.14$

Red time is found to be 40 seconds (C - g = 60 - 20). The remaining values of interest can be easily found.

Time to queue clearance after the start of effective green:

$t_c = \frac{\rho r}{1-\rho} = \frac{0.14(40)}{1-0.14} = 6.51s$

Proportion of the cycle with a queue:

$P_q = \frac{r+t_c}{C} = \frac{40+6.51}{60} = 0.775$

Proportion of vehicles stopped:

$P_s = \frac{\lambda(r+t_C)}{\lambda(r+g)} = \frac{0.1(40+6.51)}{0.1(40+20)} = 0.775$

Maximum number of vehicles in the queue:

$Q_max = \lambda r = 0.1(40) = 4$

Total vehicle delay per cycle:

$D_t = \frac{\lambda r^2}{2(1-\rho)} = \frac{0.1(40^2)}{2-(1-0.14)} = 93veh-s$

Average delay per vehicle:

$d_{avg} = \frac{r^2}{2C(1-\rho)} = \frac{(40)^2}{2(60)(1-0.14)} = 15.5s$

Maximum delay of any vehicle:

$d_{max} = r = 40\ s$

Example 2: Total Decay

Compute the average approach delay given certain conditions for a 60-second cycle length intersection with 20 seconds of green, a v/c ratio of 0.7, a progression neutral state (PF=1.0), and no chance of intersection spillover delay (overflow delay). Assume the traffic flow accounts for the peak 15-minute period and a lane capacity of 840 veh/hr, and that the intersection is isolated.

Solution

Uniform Delay:

$d_1 = \frac{0.5C(1-\frac{g}{C})^2}{1-[min(1,X)\frac{g}{C}]} = \frac{0.5(60)(1-\frac{20}{60})^2}{1-[min(1,0.7)\frac{20}{60}]} = 17.39s$

Random Delay:

$T = 0.25$ (from problem statement)

$X = 0.7$

$k = 0.5$ (for pretimed control)

$I = 1.0$ (isolated intersection)

$c = 840$

$d_2 = 900T\left[(X-1) + \sqrt{(X-1)^2 + \frac{8kIX}{cT}}\right] = 900(0.25)\left[(0.7-1) + \sqrt{(0.7-1)^2 + \frac{8(0.5)(1)(0.7)}{840(0.25)}}\right] = 4.83s$

Overflow Delay:

Overflow delay is zero because it is assumed that there is no overflow.

$d_3 = 0$

Total Delay:

$d = d_1(PF) + d_2 + d_3 = 17.39(1) + 4.83 + 0 = 22.22s$

The average total delay is 22.22 seconds.

Example 3: Cycle Length Calculation

Calculate the minimum and optimal cycle lengths for the intersection of Oak Street and Washington Avenue, given that the critical v/c ratio is 0.9, the two critical approaches have a v/s ratio of 0.3, and the Lost Time equals 15 seconds.

Solution

Minimum Cycle Length:

$C_{min} = \frac{L*X_c}{X_c - \sum_{i=1}^{n} Y_i} = \frac{15*0.9}{[0.9-(2(0.3))]} = 45s$

Optimal Cycle Length:

$C_{opt} = \frac{[(1.5L)+5]}{(1.0-\sum_{i=1}^{n} Y_i)} = \frac{1.5(15)+5}{1.0-2(0.3)} = 68.75s$

The minimum cycle length is 45 seconds and the optimal cycle length is 68.75 seconds.

Thought Question

Problem

Why don't signalized intersections perform more efficiently than uncontrolled intersections?

Solution

The inherent lost time that comes from each signal change is wasted time that does not occur when intersections are uncontrolled. It comes at quite a surprise to most of the Western World, where traffic signals are plentiful, but there are intersections that perform quite well without any form of control. There is an infamous video on YouTube that shows an uncontrolled intersection in India where drivers somehow navigate through a busy, chaotic environment smoothly and efficiently. The video is humorous to watch, but it shows a valid point that uncontrolled intersections can indeed work and are quite efficient.

However, the placement of traffic signals is for safety, as drivers entering an uncontrolled intersection have a higher likelihood of being involved in a dangerous accident, such as a T-bone or head-on collision, particularly at high speed.

Sample Problem

An approach at a pretimed signalized intersection has an arrival rate of 500 veh/hr and a saturation flow rate of 3000 veh/hr. 30 seconds of effective green are given in a 100-second cycle. Analyze the intersection assuming D/D/1 queueing by describing the proportion of the cycle with a queue, the maximum number of vehicles in the queue, the total and average delay, and the maximum delay.

Answer available at the end of this section.

Additional Questions

Homework

1. Define the following terms

A. Signal phase Answers: A signal phase is the signal combinations in a signal controller that controls a group of traffic movement off and on. B. Signal cycle Answers: The loop time for a specific traffic phase. The time length between the starting of the phase to it apears again. C. Max-out Answers: the maximum flow of the approach D. Gap-out Answers: the flow of approach accounting for the time of green light period E. Lost time Answers: sources of "lost time" at a signalized intersection are: Start-up lost time, clearance lost time and the yield the pedestrians.

2. A pretimed signal controls a four-way intersection with no turning permitted and zero lost time. The EB and WB traffic volumes are 900 and 800veh/h, and the two movements share the same effective green and effective red portions of cycles. The NB and SB directions also share cycle times, with volumes of 300 and 250 veh/h, respectively. If the saturation flow of all approaches is 1800veh/h, the cycle length is 75 seconds, and D/D/1 queuing applies, determine the effective red and green times that must be allocated to each directions combination (N-S, E-W) to minimize the total vehicle delay and compute the total delay per cycle.

3 AADT (Average Annual Daily Traffic) data on a single intersection between an arterial street (East-West) and a minor street (North-South) has been collected for the current year. Approximate saturation flow rates and collected traffic volumes are given in the table below.

Approach	**Volume (veh/hr)**	**Saturation Flow (veh/hr)**
North Bound (NB)	380	1200
South Bound (SB)	295	1200
East Bound Left Turn (EBLT)	260	1800
East Bound (EB)	1160	3000
West Bound Left Turn (WBLT)	250	1800
West Bound (WB)	1050	3000

Based on the traffic demand forecast, it is expected demand for the EB and WB directions (including left turn traffic) will increase by 10% each in the next year, and that for the NB and SB directions by 5% each, while the saturated flow rates remain unchanged. The traffic demand has changed so significantly that, as a city traffic engineer, you are asked to develop a new signal design at this intersection for the NEXT YEAR.

A. The design is required by the city to be a three-phase cycle with leading left turn and no overlap. Draw a sketch of the intersection with arrows indicating the directions of traffic to illustrate your design.

B. The signal at this intersection is traffic-actuated, but you are going to work out a pre-timed signal design in your proposal. Describe the differences between a pre-time signal and a traffic-actuated signal. Under which circumstance a pre-timed signal design can be used for a traffic-actuated signal?

C. You are also asked to develop an alternative design for this intersection because the city received complaints of too much travel delay at this intersection. Based on the calculated cycle length, how are you going to develop this plan?

Variables

- t_c - Time for queue to clear
- ρ - Arrival Rate divided by Departure Rate
- r - Red Time
- P_q - Proportion of cycle with a queue
- P_s - Proportion of Stopped Vehicles
- c - Lane Group Capacity
- s - Adjusted Saturation Flow Rate
- g - Effective Green Length
- C - Cycle Length
- d - Average Signal Delay per vehicle (sec)
- d_1 - Average Delay per vehicle due to uniform arrivals (sec)
- PF - Progression Adjustment Factor
- d_2 - Average Delay per vehicle due to random arrivals (sec)
- d_3 - Average delay per vehicle due to initial queue at start of analysis time period (sec)
- X - Volume/Capacity (v/c) ratio for lane group.
- T - Duration of Analysis Period (in hours)
- k - Delay Adjustment Factor that is dependent on signal controller mode
- I - Upstream filtering/metering adjustment factor
- d_A - Average Delay per vehicle for approach A (sec)
- d_i - Average Delay per vehicle for lane group i on approach A (sec)
- v_i - Analysis flow rate for lane group i
- d_I - Average Delay per vehicle for the intersection (sec)
- v_A - Analysis flow rate for approach A
- Y_c - Sum of Flow Ratios for Critical Lane Groups
- $(v/c)_{ci}$ - Flow Ratio for Critical Lane Group i
- n - Number of Critical Lane Groups
- C_{min} - Minimum necessary cycle length
- X_c - Critical v/c ratio for the intersection
- $(v/s)_{ci}$ - Flow Ratio for Critical Lane Group
- C_{opt} - Optimal Cycle Length for Minimizing Delay
- X_i - v/c ratio for lane group i
- X_c - Degree of saturation for an intersection cycle

Key Terms

- Delay
- Total Delay
- Average Delay
- Uniform Delay
- Random Delay
- Overflow Delay
- Cycle Length
- v/c ratio
- v/s ratio
- Saturation Flow Rate
- Red Time
- Effective Green

- Minimum Cycle Length
- Optimal Cycle Length
- Critical Lane Group
- Degree of Saturation
- Progression Adjustment Factor
- Lost Time
- Queue

Videos

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

External Exercises

Use the GAME software at the STREET website to learn how to coordinate traffic signals to reduce delay.

Use the OASIS software at the STREET website to study how signals change when given information about time-dependent vehicle arrivals.

Answers

Sample Problem

With the statements in the problem, we know:
Green Time = 30 seconds
Red Time = 70 seconds
Cycle Length = 100 seconds
Arrival Rate = 500 veh/hr (0.138 veh/sec)
Departure Rate = 3000 veh/hr (0.833 veh/sec)
Traffic intensity, ρ, is the first value to calculate.
$\rho = \frac{\lambda}{\mu} = \frac{500}{3000} = 0.167$
Time to queue clearance after the start of effective green:
$t_c = \frac{\rho r}{1-\rho} = \frac{0.167(70)}{1-0.167} = 14.03s$
Proportion of the cycle with a queue:
$P_q = \frac{r+t_c}{C} = \frac{70+14.03}{100} = 0.84$
Proportion of vehicles stopped:
$P_s = \frac{\lambda(r+t_C)}{\lambda(r+g)} = \frac{0.138(70+14.03)}{0.138(70+30)} = 0.84$
Maximum number of vehicles in the queue:
$Q_max = \lambda r = 0.138(70) = 9.66$

Total vehicle delay per cycle:

$D_t = \frac{\lambda r^2}{2(1-\rho)} = \frac{0.138(70^2)}{2-(1-0.157)} = 406veh - s$

Average delay per vehicle:

$d_{avg} = \frac{r^2}{2C(1-\rho)} = \frac{(70)^2}{2(100)(1-0.167)} = 29.41s$

Maximum delay of any vehicle:

$d_{max} = r = 70\ s$

Thus, the solution can be determined:

Proportion of the cycle with a queue = 0.84 Maximum number of vehicles in the queue = 9.66 Total Delay = 406 veh-sec Average Delay = 29.41 sec Maximum Delay = 70 sec

This page titled 6.2: Traffic Signals is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

6.3: Traffic Control Devices

Traffic Control Devices include street signs, traffic signals, and road markings. These signs, signals, and stripes guide drivers in navigation and control of their vehicles. Traffic signals are treated above in the section Traffic Signals. Signs and stripings, the other controls, have usage described in the *Manual of Uniform Traffic Control Devices* (MUTCD).

Road signs in the United States are standardized in the *Manual on Uniform Traffic Control Devices*, the history of the evolution of this document is given in.Signs evolved from local practice, cities and states copying neighbors, and inventing what they needed, and then later standardizing (first for rural and urban areas separately, and then jointly) after the value of coordination became apparent when automobile travelers crossed jurisdictional boundaries. From the first center line in Michigan in 1911, stop sign in 1915, a 1923 established the basis of the shapes used for road signs today. The Mississippi Valley Association of State Highway Departments proposed the following. Shapes which were most complex to make (circle, octagon) would indicate most danger and be used least often.

- Round: Used to warn of RR crossing;
- Octagon: Used only to signify a stop;
- Diamond: used to indicate ordinary condition of danger requiring precaution at all times;
- Square: Used to indicate intermittent danger conditions requiring little more than ordinary care;
- Rectangular: Used to indicate regulatory or directional information;
- Cut-Out; Used a distinctive shape different from the above to identify highway routes.

This system was improved over time. In 1924 the Minnesota Department of Highways published its *Manual of Markers and Signs* with the same shapes, but the white background was made yellow. In 1924, the American Association of State Highway Officials adopted the MVASHD plan (with black on yellow), however Red and green on signs rejected because of visibility at night.

Separately, standards were being developed for cities. Traffic signals are largely an urban phenomenon. While the date of the first traffic signal is contested, the electric traffic signal appeared in Cleveland in 1914, and the first three-color traffic signal in 1920. In 1924 the National Conference on Street and Highway Safety developed urban sign recommendations:

- Stop = Red,
- Proceed = Green,
- Caution = Yellow,
- Cross-Roads = Purple or other distinctive color.
- Centerlines = White, but these were to be used only where it was unsafe to be on the left side of the road (curves, hill crests, intersections, railroad crossings)

The objective of AASHO in these early years was first to inventory all of the sign characteristics that had been locally deployed, and then to standardized various aspects: Shape, Word, Color, Symbol, Uniformity of Erection and Application. Even as late as 1930, the third National Conference on Street and Highway Safety, designed for urban published a *Manual on Street Traffic Signs, Signals and Markings* which had either white or black paint for concrete, and white or yellow paint for bituminous. A red border and legend on yellow was suggested for stop signs.

Finally, in 1932, a Joint Committee on Uniform Traffic Control Devices met to rectify and combine the separate AASHO and NCSHS manuals for rural and urban traffic into a complete manual. Main initial points were color codes, signs at night, and reduced sign sizes in urban areas. Visibility research was undertaken, sponsored by the Bureau of Public Roads. Minor changes continued after this date, though a modern driver would certainly understand the road at this point. For instance, in the 1954 MUTCD the stop sign changes from black on yellow to white on red; yield sign introduced as triangle (black on yellow), emulating European standards.

This table reviews some of the key decisions in the creation of the *Manual on Uniform Traffic Control Devices*.

Timeline of Traffic Control Devices

- 1911 First center line in Michigan
- 1914 First electric traffic signal installation in Cleveland

- 1915 First stop sign Detroit
- 1920 First three-color traffic signal
- 1923 Shapes agreed to at Mississippi Valley Association of State Highway Departments (Shapes which were most complex to make (circle, octagon) would indicate most danger and be used least often.
- 1924 Minnesota Department of Highways publishes Manual of Markers and Signs (same shapes, white background -> yellow background)
- 1924 AASHO adopts MVASHD (with black on yellow). Red and green on signs rejected because of visibility at night
- 1924 National Conference on Street and Highway Safety developed urban sign recommendations
- 1925: AASHO Manual and Specifications for the Manufacture, Display, and Erection of U.S. Standard Road Markers and Signs ... Tried to build inventory of sign characteristics: Goal of signs: Shape, Word, Color, Symbol, Uniformity of Erection and Application
- 1926 Second National Conference on Street and Highway Safety. Authorized survey from American Engineering Council to collate national practice and make recommendations.
- 1927 Bureau of Public Roads (part of USDA) first national signing manual.
- 1930 Third National Conference on Street and Highway Safety. Publishes Manual on Street Traffic Signs, Signals and Markings. White or black paint for concrete, white or yellow paint for bituminous. Red border and legend on yellow for stop sign. Designed for urban areas.
- 1932 Joint Committee on Uniform Traffic Control Devices met to rectify and combine AASHO and NCSHS separate manuals for rural and urban traffic into a complete manual. Main initial points were color codes, signs at night, and reduced sign sizes in urban areas. Visibility research was undertaken, sponsored by BPR.
- 1935 First Manual on Uniform Traffic Control Devices
- 1942 - War Emergency Edition
- 1948 MUTCD
- 1954 MUTCD - stop sign changes from black on yellow to white on red; yield sign introduced as triangle (black on yellow), emulating European standards.
- 1961 MUTCD
- 1971 MUTCD
- 1978 MUTCD
- 1988 MUTCD
- 2003 MUTCD

Roundabouts

History of Roundabouts

The use of the modern roundabouts began in the 1980s which saw a widespread use in Europe and Australia, and while the United States began implementing the modern roundabout in the early 1990s, their original conception came from the rotaries and traffic circles which were used during early 1900s to 1940s. The modern roundabout is defined as an island in the center traffic where cars enter that area and travel counterclockwise before choosing their exit.

There are many designs which are used for roundabouts depending on circumstances such as traffic congestion, crash and fatality rate, and geographical area. The different designs are as follows:

- Mini-roundabouts
- Urban compact roundabouts
- Urban single-lane roundabouts
- Urban double-lane roundabouts
- Rural single-lane roundabouts
- Rural double-lane roundabouts

Reasons for using Roundabouts

The main purpose for using roundabouts is to ease the flow of traffic in certain areas, but there are also many advantages to having roundabouts rather than intersections ranging from safety matters to environmental issues. The drawbacks of installing roundabouts are mainly cleanup duty and the cost. States which experience heavy snowfall and severe weather conditions will be forced to do extra cleanup work as the area taken by roundabouts are generally larger than intersections. Also, the cost of

constructing roundabouts can be expensive in some areas, were it is estimated roundabouts cost on average $317,000, while the cost of installing an intersection is pegged at an average of $300,000. However, installing roundabouts offers suitable solutions to city planners in many ways. The major concern for city planners is the safety of all people involved in vehicle conflicts. Two-lane Intersections generate 32 conflict points while roundabouts generate 8 conflict points between vehicles only, which results in 75% less possible crashes that rise from these conflicts. In addition, roundabouts provide less pedestrian to vehicle conflicts as the crosswalks in the roundabouts are farther away from the flow of traffic, and its placement shows potential conflicts with oncoming vehicles that are planning to enter the roundabout.

Such measures that are implicitly built in the roundabout design provide greater safety to the vehicles as well as the pedestrians. Also, the nature of the roundabout forces oncoming traffic to slow down to enter it, thereby eliminating traffic congestion and the use of stop signs. Yield signs are used instead of stop signs, which during heavy traffic may cause congestion, and yield sign allows vehicles to drive freely if the space in the roundabout is available. The environmental aspect of roundabouts can be considerable depending on the traffic flow and number of vehicles that use the roundabout. The Insurance institute for highway safety states that studies have shown roundabouts reduce the consumption of fuel by 235,000 gallons in one year as opposed to intersection when the study was conducted on intersection that were to be converted to roundabouts. There reasons are simple, continuous flow of traffic reduces traffic delays and vehicles stops, which in turn reduces the amount of fuel burned by the vehicle as opposed to when it is still at an intersection waiting for a green light. Therefore, having roundabouts can be environmentally friendly and economically sound for vehicle operators.

Safety Statistics

Roundabouts offer less conflicts among vehicles and pedestrians, therefore naturally crashes tend to be reduced. Based on 10 samples, the average crash rate in roundabouts is 0.13 crashes per million vehicles, which present 40% less crashes, 75% less crashes with injuries and almost no fatalities when compared to intersection due to the fact that roundabouts eliminate the possibility of vehicles colliding head on. The only type of roundabout collision experienced would be a side collision which in general is much safer than a head on collision.

Concerns about Roundabouts

While roundabouts reduce conflict possibilities, they do not eliminate them. The disregard for stop signs allows reckless drivers to drive through roundabouts without paying attention to traffic within the roundabout, which may cause problems. In addition, the size of the island which vehicles drive around may block the line of sight of vehicles trying to enter the roundabout, which under some circumstances may cause crashes. Finally, the location and size of the roundabout has to be considered carefully before being implemented as certain vehicles (large trucks, firefighter trucks, emergency vehicles etc) may have difficulties driving around the roundabouts if they were improperly placed. Finally, in 1972, the town Swindon in the United Kingdom saw the installation of the first "magic roundabout" which consisted of six mini-roundabouts which are installed around one central island where drivers drive counterclockwise around the central island and clockwise around the mini-roundabouts. Although Swindon residents find the roundabout quite useful, tourists as well as outsiders have consistently complained about the dangers that might rise from such a roundabout since it is one of a kind, and not all drivers are accustomed to operating vehicles around such a roundabout. A poll conducted by the Britannia Rescue places the Magic Roundabout at the fourth scariest road in 2009.

Pedestrian Access

The research findings on pedestrian safety at roundabouts are somewhat unclear. There have been relatively few studies, mostly conducted in Europe, concerning pedestrians and roundabouts. Pedestrian-vehicle crashes, the most commonly used dependent measure in pedestrian safety studies, tend to occur infrequently both before and after an intersection is converted to a roundabout. As a result, it is difficult to draw firm conclusions from the literature regarding pedestrian safety and roundabouts. One issue that is often not considered in pedestrian research is the degree to which pedestrian volume changes when intersections with signal or stop-sign control are converted to roundabouts. There is a need for research on this topic as well as a broad range of other pedestrian-related concerns at roundabouts. Little is known about the effect of roundabouts on older pedestrians, children, and pedestrians with disabilities. Anecdotal evidence indicates that many Australian engineers (who have extensive experience with roundabouts) consider these intersections to be unsuitable if large numbers of pedestrians are present, however some improvements can be helpful especially when considering pedestrians with disabilities.

Improvements for wayfinding

- Well-defined walkway edges

- Separated walkways, with landscaping at street edge to preclude prohibited crossings to center island
- Tactile markings across sidewalk to identify crossing locations
- Bollards or architectural features to indicate crossing locations
- Detectable warnings (separate at splitter islands) at street edge
- Perpendicular crossings ; where angled, use curbing for alignment cues
- High-contrast markings
- Pedestrian lighting

Orientation and mobility techniques used by blind individuals at traditional intersections rely heavily on traffic sounds. When traffic signals and stop signs regulate traffic movements at intersections, the resulting breaks in traffic flow provide identifiable and predictable periods – gaps – during which pedestrians can cross. Such predictable breaks do not usually occur at roundabouts, and so pedestrians must make judgments about the speed and travel paths of approaching vehicles (and the duration of gaps between vehicles). It appears that sighted adults are generally able to safely make such judgments, although some pedestrians may have difficulty doing so. Research suggests that the selection of appropriate gaps at roundabouts is problematic for blind pedestrians at some roundabouts.

This page titled 6.3: Traffic Control Devices is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

6.4: Metering and Analogs

Metering

Metering (or Ramp Metering) is the application of traffic control to freeway on-ramps to limit the rate of flow entering freeway sections.

Metering has several purposes. The primary purpose is to optimize traffic flow. Metering keeps traffic flowing at or near freeflow speed and at or near maximum flow. Since it is just below the maximum flow, it can be thought of as risk averse. How far below the maximum you set the metered flow rate is an indication of risk aversion (how important it is to avoid breakdown). However, by maximizing total output flow, while avoiding throughput-reducing traffic breakdowns, metering flow should maximize traffic flow on other facilities as well.

Metering has other objectives, including the break-up of platoons entering freeways. Limiting the number of vehicles that enter the freeway all at once increases the likelihood of each vehicle finding a suitable gap into which it can merge, thus decreasing the likelihood of shockwaves being generated at freeway on-ramps.

Metering similarly improves safety by reducing the number of stop-and-start maneuvers required, and decreasing the likelihood of crashes.

Finally, metering provides a tool to manage incident conditions dynamically. Incidents may block lanes, and metering restricts inflow so that congestion on the freeway is diminished. If ramp queues are long enough, vehicles may divert to other routes.

Metering does not come for free, it imposes costs, which are increased delays on ramps. When determining the appropriate metering rate, ramp delays created must be considered along with freeway delays reduced.

Ramp metering works by keeping traffic just below capacity, at the sweet spot of freeflow speed and maximum throughput

Metering to control bottleneck flow

The Minnesota Algorithm is one strategy for implementing ramp metering. The objective is to meet target flow for each "zone", which is the set of on-ramps upstream of a bottleneck.

The flow balance equation used in ramp metering is given below, with M and F being the control variables. The method is to adjust inflow to meet targeted outflow:

$$A + U + M + F = X + B + S$$

Where:

- A upstream mainline volume (measured variable);
- U sum of unmetered entrance ramp volumes (measured variable);
- M sum of metered local access ramp volumes (controlled variable);
- F sum of metered freeway to freeway access ramp volumes (controlled variable)
- X sum of exit ramp volumes (measured variables);
- B downstream bottleneck volume at capacity (constant);
- S space available within the zone (volume based on a measured variable).

Twin Cities Ramp Metering Holiday

Timeline

- Late 1990s: Long delays at some ramps (up to 20 minutes);
- 1999 Republican State Senator Dick Day, who “drives over 70,000 miles per year” from Owatonna, Minnesota pushes

"Freedom to Drive" package: shut-off meters, revert HOV lanes, left lane is passing-only lane.
- Nov. 28, 1999 : Star-Tribune does large Sunday, Front Page piece on Ramp Meters.
- Early 2000: MnDOT commission's 3 University of Minnesota studies (Levinson, Michalopoulos, Kwon) to evaluate meters. MnDOT engineers believe results would be ``catastrophic *for traffic*.
- May 2000 Legislature insists on shut-down experiment, at least 4 weeks.
- Cambridge Systematics hired to conduct study (outsiders).
- Meters shut off for 8 weeks from Oct. to Dec., 2000.

Consequence of ramp metering shutoff

The absence of metering resulted in increased freeway system (freeway plus ramp) travel time (reduced speeds) for most travelers. However short trips were advantaged as they did not need to wait at ramp meters to use the freeway system. Travel time variability on the freeway system increased, as freeway congestion was more frequent (though ramp delays were now nonexistent). However the equity of the system increased, as delays were more uniformly distributed.

The number of trips using the freeway system tended to increase. Average weekday peak period trip lengths on the freeway system dropped. We can infer that more non-work trips used the freeway during this period.

Control Logics

It turns out, the most efficient ramp metering control logic is the one metering the nearest entrance ramp(s) to a critical freeway section so as to keep the flow of this section strictly below capacity. This is also the least equitable, as it implies travelers from one ramp may receive all of the delay while others don't.

In determining a control strategy, equity and efficiency are traded-off. There are a number of implementation considerations:

- Control – how to "keep the flow of this section strictly below capacity".
- Control variables: Flow vs. Density,
- Feedback vs. Feed-forward,
- Linear vs. Nonlinear
- Selection of the threshold values for the controller
- Reliability: Risk averse vs. Risk seeking
- Equity considerations – Can one only meter the "nearest entrance ramp(s)"

There are a set of practical coordination mechanisms:

- Maximum delay/queue length restrictions
- Coordination group: zones, helper groups, no coordination
- Theoretical approach: minimize non-linear weighted travel time
- Optional constraints: smooth of operation etc.

Analogs

Transportation networks have analogs with network processes in other systems, such as water networks, structures, and electrical networks. Some of the relationships are outlined below.

'	**Transportation**	**Water: Hydrostatics**	**Structures**	**Electrical**
Node Conservation Law	Flow (q)			Current (Kirchoff's Current Law)
Fundamental Law	q = kv	P = ρgh	F=δ (mv)/ δ (T)	V=IR
	k = q/v		F= v δ (m)/ δ (T)	V=I/G

	v=q/k	Bernoulli's Equation:		Ohm's Law on resistor
		Constant=p+1/2ρ V^2+ ρgh		
		P=F/A (area)		
		F=ma		
Analogs	flow (q)	Pressure (P*A)	δm/δT	Current (I)
	density (k)	Density (ρ)	Force (F)	Voltage (V)
	velocity (v)	velocity (v)	velocity	Conductance (G)
Equilibrium Conditions	Wardrop (time equal on used pairs in parallel)		Sum of horizontal (and sum of vertical) forces on a structure = 0, sum of moments = 0.	Voltage drop across two components in parallel are equal

Structures

- F= force
- m = Mass
- a = acceleration
- T = Time

Transportation

- q = flow
- k = density
- v = velocity

Electricity

- V= Voltage
- I = Current
- R = Resistance
- G = Conductance = 1 / Resistance

Water

- P = hydrostatic pressure
- ρ = fluid density =mV = mass *volume
- g = acceleration due to gravity
- h = height
- c= constant
- A = area

This page titled 6.4: Metering and Analogs is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

7: GEOMETRIC DESIGN

7.1: Sight Distance
7.2: Grade
7.3: Earthwork
7.4: Horizontal Curves
7.5: Vertical Curves

This page titled 7: Geometric Design is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

7.1: Sight Distance

Sight Distance is a length of road surface which a particular driver can see with an acceptable level of clarity. Sight distance plays an important role in geometric highway design because it establishes an acceptable design speed, based on a driver's ability to visually identify and stop for a particular, unforeseen roadway hazard or pass a slower vehicle without being in conflict with opposing traffic. As velocities on a roadway are increased, the design must be catered to allowing additional viewing distances to allow for adequate time to stop. The two types of sight distance are (1) stopping sight distance and (2) passing sight distance.

Derivations

Stopping Sight Distance

Stopping Sight Distance (SSD) is the viewable distance required for a driver to see so that he or she can make a complete stop in the event of an unforeseen hazard. SSD is made up of two components: (1) Braking Distance and (2) Perception-Reaction Time.

Forces acting on a vehicle that is braking

For highway design, analysis of braking is simplified by assuming that deceleration is caused by the resisting force of friction against skidding tires. This is applicable to both an uphill or a downhill situation. A vehicle can be modeled as an object with mass m sliding on a surface inclined at angle θ.

While the force of gravity pulls the vehicle down, the force of friction resists that movement. The forces acting this vehicle can be simplified to:

$$F = W(sin(\theta) - fcos(\theta)) \tag{7.1.1}$$

where

- $W = mg$ = object's weight,
- f = coefficient of friction.

Using Newton's second law we can conclude then that the acceleration (a) of the object is

$$a = g(sin(\theta)) - fcos(\theta)) \tag{7.1.2}$$

Using our basic equations to solve for braking distance (d_b) in terms of initial speed (v_i) and ending speed (v_e) gives

$$d_b = \frac{v_i^2 - v_e^2}{-2a} \tag{7.1.3}$$

and substituting for the acceleration yields

$$d_b = \frac{v_i^2 - v_e^2}{2g(fcos(\theta) - sin(\theta))} \tag{7.1.4}$$

For angles commonly encountered on roads, $cos(\theta) \approx 1$ and $\mathbf{sin(\theta) \approx tan(\theta) = G}$ $sin(\theta) \approx tan(\theta) = G$, where G is called the road's grade. This gives

\[d_b=\frac{v_i^2-v_e^2}{2g(f \pm G)\]

Using simply the braking formula assumes that a driver reacts instantaneously to a hazard. However, there is an inherent delay between the time a driver identifies a hazard and when he or she mentally determines an appropriate reaction. This amount of time is called perception-reaction time. For a vehicle in motion, this inherent delay translates to a distance covered in the meanwhile. This extra distance must be accounted for.

For a vehicle traveling at a constant rate, distance d_r covered by a specific velocity v and a certain perception-reaction time t_r can be computed using simple dynamics:

$$d_r = (vt_r) \tag{7.1.5}$$

Finally, combining these two elements together and incorporating unit conversion, the AASHTO stopping sight distance formula is produced. The unit conversions convert the problem to metric, with v_i in kilometers per hour and d_s in meters.

$$d_s = d_r + d_b = 0.278t_r v_i + \frac{(0.278v_i)^2}{19.6(f \pm G)} \tag{7.1.6}$$

A Note on Sign Conventions

We said \(d_b=\frac{v_i^2-v_e^2}{2g(f \pm G)\)

Use: $(f - G)$ if going downhill and $(f + G)$ if going uphill, where G is the absolute value of the grade

Passing Sight Distance

Passing Sight Distance (PSD) is the minimum sight distance that is required on a highway, generally a two-lane, two-directional one, that will allow a driver to pass another vehicle without colliding with a vehicle in the opposing lane. This distance also allows the driver to abort the passing maneuver if desired. AASHTO defines PSD as having three main distance components: (1) Distance traveled during perception-reaction time and accleration into the opposing lane, (2) Distance required to pass in the opposing lane, (3) Distance necessary to clear the slower vehicle.

The first distance component d_1 is defined as:

$$d_1 = 1000t_1\left(u - m + \frac{at_1}{2}\right) \tag{7.1.7}$$

where

- t_1 = time for initial maneuver,
- a = acceleration (km/h/sec),
- u = average speed of passing vehicle (km/hr),
- m = difference in speeds of passing and impeder vehicles (km/hr).

The second distance component d_2 is defined as:

$$d_2 = (1000ut_2) \tag{7.1.8}$$

where

- t_2 = time passing vehicle is traveling in opposing lane,
- u = average speed of passing vehicle (km/hr).

The third distance component d_3 is more of a rule of thumb than a calculation. Lengths to complete this maneuver vary between 30 and 90 meters.

With these values, the total passing sight distance (PSD) can be calculated by simply taking the summation of all three distances.

$$d_p = (d_1 + d_2 + d_3) \tag{7.1.9}$$

Demonstrations

- GIF animation: Stopping Sight Distance on Flat Surface (contributed by Oregon State University faculty and students)

- GIF animation: Stopping Sight Distance on Downhill Grade (contributed by Oregon State University faculty and students)
- Flash animation: Bicycle Crash Type (contributed by Oregon State University faculty and students)

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Examples

Example 1: Stopping Distance

A vehicle initially traveling at 66 km/h skids to a stop on a 3% downgrade, where the pavement surface provides a coefficient of friction equal to 0.3. How far does the vehicle travel before coming to a stop?

Solution

$$d_b = \frac{(66*(\frac{1000}{3600}))^2-(0)^2}{2*(9.8)*(0.3-0.03)} = 63.5m$$

Example 2: Coefficient of Friction

A vehicle initially traveling at 150 km/hr skids to a stop on a 3% downgrade, taking 200 m to do so. What is the coefficient of friction on this surface?

Solution

$$d_b = \frac{(150*(\frac{1000}{3600}))^2-(0)^2}{2*(9.8)*(f-0.03)} = 200m$$

$$f-0.03) = \frac{(150*(\frac{1000}{3600}))^2-(0)^2}{2*(9.8)*200}$$

$$f = 0.47$$

Example 3: Grade

What should the grade be for the previous example if the coefficient of friction is 0.40?

Solution

$$d_b = \frac{(150*(\frac{1000}{3600}))^2-(0)^2}{2*(9.8)*(0.40-G)} = 200m$$

$$(0.40-G) = \frac{(150*(\frac{1000}{3600}))^2-(0)^2}{2*(9.8)*200}$$

$$G = 0.44-0.40 = 0.04$$

Thus the road needs to be a 4 percent uphill grade if the vehicles are going that speed on that surface and can stop that quickly.

Example 4: Crash Reconstruction

You are shown an crash scene with a vehicle and a light pole. The vehicle was estimated to hit the light pole at 50 km/hr. The skid marks are measured to be 210, 205, 190, and 195 meters. A trial run that is conducted to help measure the coefficient of friction reveals that a car traveling at 60 km/hr can stop in 100 meters under conditions present at the time of the accident. How fast was the vehicle traveling to begin with?

Solution

First, Average the Skid Marks.

$(210 + 205 + 190 + 195)/4 = 200$

Estimate the coefficient of friction.

$d_b = \frac{(60*(\frac{1000}{3600}))^2-(0)^2}{2*(9.8)*(f-0)} = 100m$

$f = \frac{(60*(\frac{1000}{3600}))^2-(0)^2}{2*(9.8)*100} = 0.14$

Third, estimate the unknown velocity

$d_b = \frac{(v*(\frac{1000}{3600}))^2-(50*(\frac{1000}{3600}))^2}{2*(9.8)*(0.14-0)} = 200m$

$\left(v*(\frac{1000}{3600})\right)^2 - \left(50*(\frac{1000}{3600})\right)^2 = 200m*(2*(9.8)*(0.14))$

$548.8 + 192.9 = v^2(\frac{1000}{3600})^2$

$v^2 = \frac{741.7}{0.077} = 9612.43$

$v = 98km/h$

Example 5: Compute Stopping Sight Distance

Determine the Stopping Sight Distance from Example 4, assuming an AASHTO recommended perception-reaction time of 2.5 seconds.

Solution available at the end of this section.

Thought Question

Problem

If the coefficient of friction is 0 (zero) and the grade is 0, how long does it take a moving vehicle to stop?

Solution

Forever

Note, the design conditions for roads are wet, i.e. a lower coefficient of friction

Sample Problem

You see a body lying across the road and need to stop. If your vehicle was initially traveling at 100 km/h and skids to a stop on a 2.5% upgrade, taking 75 m to do so, what was the coefficient of friction on this surface?

Solution

$d_b = \frac{(100*(\frac{1000}{3600})^2-(0)^2}{2*(9.8)*(f+0.025)} = 75m$

$(f + 0.025) = \frac{(27.78)^2}{2*(9.8)*75}$

$f = 0.50$

Homework

1. Name five principal characteristics of visual reception important in driving.

2. You are shown an accident scene with a vehicle and a tree on uphill grade of 3%. The vehicle was estimated to hit the tree at 120 km*h^{-1} . The average length of skid marks was 20 meters. You have found that a car traveling that section under similar weather conditions at 60 km*h^{-1} can stop in 60 m. What was its initial speed?

3. Draw a road's cross-section and label the elements.

4. How do the calculations of stopping sight distance and passing sight distance differ?

5. What is the recommended value used for perception reaction time according to AASHTO?

6. What is a standard lane width used in new highway designs (to nearest foot or centimeter)?

Additional Questions

1. What happens if an accident is caused by poorly designed roads? → Lawsuits
2. Explain Stopping Sight Distance again → = perception reaction distance + braking distance
3. Isn't 200 m long distance for braking → Yes unless very high speed on very slick surface (or going downhill).
4. Is higher coefficient of friction used in road design? → Not too often, more often taken as a function of materials and construction, and wear and tear on road (older roads have less friction).
5. Does coefficient of friction properly account for the ways cars brakes work and the manner in which drivers apply the brakes?
6. Are stopping distance (related to accident reconstruction) admissible as evidence in court? Would this be by a licensed professional engineer? → Yes
7. How close are accident reconstruction calculations to be actual accidents? → Should be on average correct ...
8. For stopping distance, why don't we have a factor to include vehicle size and weight?
9. Why would the grade "G" be positive in the stopping distance equation? What would the sign be in the Stopping Distance Equation. What does a negative grade mean if you are were assuming to be going downhill?
10. Describe the interaction between gravity and friction when going up or downhill. Is friction helped or hindered?
11. Why can cosine and sine be ignored in calculations? --> Small angle approximations.
12. What are the steps in accident reconstruction
13. If you have an accident reconstruction problem, do you incorporate grade when finding friction? Yes, but the grade is known.
14. How are skid marks useful in determining initial speed of vehicle?
15. What affects the friction force?
16. What can stopping distance measure be used for?
17. What effect does grade have on stopping distance?
18. What if more or less skid-marks found at an accident scene. How are averages computed when distances are far apart?

19. What is average perception reaction time given by AASHTO
20. What is the maximum grade of a driveway?
21. What type of braking is assumed in the stopping distance equation?
22. When an object is sliding on an inclined surface, what two forces are operating on it?
23. When is the road grade equal to 0.
24. Why is accident reconstruction performed? How does it work?

Variables

- d_s - stopping (sight) distance (m)
- d_r - perception reaction distance (m)
- d_b - braking distance (m)
- d_p - passing distance (m)
- v_i - initial speed (km/h)
- t_r - perception/reaction time (seconds)
- f - AASHTO stopping friction coefficient (dimensionless)
- G - roadway grade (dimensionless)

Key Terms

- SSD : Stopping Sight Distance
- PSD : Passing Sight Distance
- PRT : Perception-Reaction Time

Standards and Practices

AASHTO Recommended Friction Coefficients

(km/hr)	Coefficient of Skidding Friction (f)
30	0.40
40	0.38
50	0.35
60	0.33
70	0.31
80	0.30
90	0.30
100	0.29
110	0.28
120	0.28

Answers

Example 5: Compute Stopping Sight Distance

$d_s = ((1000/3600) * 98 * 2.5) + (98 * 0.278)^2/(2 * 9.8 * 0.14) = 338$

This page titled 7.1: Sight Distance is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

7.2: Grade

Road Vehicle Performance forms the basics that make up highway design guidelines and traffic analysis. Roads need to designed so that the vehicles traveling upon them are accommodated. For example, the Interstate Highway system in the United States has a maximum allowable grade that can be used, so that the semitrucks that frequently use these roads would be able to travel them without encountering grade problems. While this can pose quite a design challenge in mountainous regions like Colorado Rockies, without it those trucks would be forced to find better alternative routes, which costs time and money. The ability of a certain vehicle type to use a road is dependent on the power produced by its motor, as well as other road-based and environmental characteristics. When choosing a road design, these elements need to be considered.

Filbert Street in San Francisco (North Beach)

Illustration

How steep is the grade in San Francisco? The famous Lombard Street, the curviest street in the world only has a grade of 14.3%. Sources vary on the steepest street in San Francisco. Filbert Street between Hyde and Leavenworth has a grade of 31.5%, Prentiss between Chapman and Powhattan reportedly has a grade of 37%.

Driving in San Francisco at times is like a roller coaster, you can't see below the nose of your car sitting at the top of the hill, though you can usually see the base of the hill.

What is max grade allowed by law?

Maximum grade is not regulated so much by law as by engineering standards. Maximum grade varies by type of road, and expected speed. In practice, it depends on the alternatives: Is the alternative no road at all? Max grade in the relatively flat Minnesota might be lower than Max grade in mountainous Colorado where there are fewer alternatives. Some typical values for illustration:

- An interstate is out of standard if it has a grade > 7%.
- The National Road (built in 1806) had a maximum grade of 8.75%.
- Local roads are much higher (12% or 15% are sometimes allowed)
- Otter Tail MN County roads 6%, alleys 8%
- Driveways can be as much as 30% for a short distance

Lombard Street in San Francisco

Fundamental Characteristics

Tractive effort and resistance are the two main forces that oppose one another and determine the performance of roadway vehicles. Tractive effort is the force exerted against the roadway surface to allow a vehicle to move forward. Resistance encompasses all forces that push back and impede motion. Both of these are in units of force. The general formula for this is outlined below:

$$F_t = ma + R_a + R_{rl} + R_g \tag{7.2.1}$$

Where:

- F_t = Tractive Effort
- m = Vehicle Mass
- a = Acceleration
- R_a = Aerodynamic Resistance
- R_{rl} = Rolling Resistance
- R_g = Grade Resistance

These components are discussed in greater detail in the following sections.

Aerodynamic Resistance

Aerodynamic resistance is a force that is produced by turbulent air flow around the vehicle body. This turbulence is dependent on the shape of the vehicle, as well as the friction of air passing over the vehicle's surface. A small portion of this resistance comes from air flow through vehicle components, such as interior ventilation. This resistance can be estimated through the following formula:

$$R_a = \frac{\rho}{2} A C_D V^2 \tag{7.2.2}$$

Where:

- ρ = Air Density
- A = Frontal area of the vehicle
- C_D = Coefficient of Drag
- V = Speed of the vehicle

Air density is a function of elevation and temperature. Frontal area and coefficient of drag are generally unique to each vehicle or type of vehicle.

Rolling Resistance

Rolling resistance is caused by the interaction of tires with the roadway surface. Three main causes exist that create this resistance. The first is the rigidity of the tire and the roadway surface. The second is tire pressure and temperature. The third is vehicular operating speed. This value of rolling friction can be calculated from a very simplified formula, given here in metric. V is in meters per second.

$$f_{rl} = 0.01(1 + \frac{V}{44.73}) \tag{7.2.3}$$

The resistance caused by this friction will increase as weight is added to the vehicle. Therefore, rolling resistance can be calculated.

$$R_{rl} = f_{rl} W \tag{7.2.4}$$

Where:

- W = Vehicle Weight
- f_{rl} = Rolling Friction

Grade Resistance

Grade resistance is the simplest form of resistance. It is the gravitational force acting on the vehicle. This force may not be exactly perpendicular to the roadway surface, especially in situations when a grade is present. Thus, grade resistance can be calculated in the following formula:

$$R_g = WG \tag{7.2.5}$$

Where:

- W = Vehicle Weight
- G = Grade (length/length)

Tractive Effort

Tractive effort is the force that allows the vehicle to move forward, subject to the resistances of the previous three forces.

The derivation of the formula comes from understanding the forces and moments that on around the various tires. It can be summarized into a simple concept, illustrated here.

For a rear-wheel drive car:

$$F_{max} = \frac{\mu W(l_f - f_{rl}h)/L}{1 - \rho h/L} \tag{7.2.6}$$

For a front-wheel drive car:

$$F_{max} = \frac{\mu W(l_f + f_{rl}h)/L}{1 + \rho h/L} \tag{7.2.7}$$

Where:

- F_{max} = Maximum Tractive Effort
- μ = Coefficient of road adhesion
- W = Vehicle Weight
- l_r = Distance from rear axle to vehicle's center of gravity
- l_f = Distance from front axle to vehicle's center of gravity
- f_{rl} = Coefficient of rolling friction
- h = Height of the center of gravity above the roadway surface
- L = Length of wheelbase

Grade Computation

Most of the work surrounding tractive effort is geared toward determining the allowable grade of a given roadway. With a certain known vehicle type using this road, the grade can be easily calculated. Using the force balance equation for tractive effort, a value for grade can be separating, producing the formula below:

$$G = \frac{F_t - F_a - F_{rl}}{W} \tag{7.2.8}$$

This calculation produces the maximum grade allowed for a given vehicle type. It assumes that the vehicle is operating at optimal engine capacity and, thus, no acceleration can occur, dropping that element from the overall equation.

Examples

Example 1: Racecar Acceleration

A racecar is speeding down a level straightaway at 100 km/hr. The car has a coefficient of drag of 0.3, a frontal area of 1.5 m^2, a weight of 10 kN, a wheelbase of 3 meters, and a center of gravity 0.5 meters above the roadway surface, which is 1 meter behind the front axle. The air density is 1.054 kg/m^3 and the coefficient of road adhesion is 0.6. What is the maximum possible rate of acceleration for the vehicle?

Solution

Use the force balancing equation to solve for a.

$F_t = ma + R_a + R_{rl} + R_g$

Since the straightaway is a level one, the grade is zero, thus removing grade resistance from the general problem.

$R_g = 0$

Aerodynamic resistance is computed:

$R_a = \frac{\rho}{2} A C_d V^2 = \frac{1.054}{2}(1.5)(0.3)(100 * 1000/3600)^2 = 183/textN$

Rolling resistance is computed:

$f_{rl} = 0.01(1 + \frac{V}{44.73}) = 0.01(1 + \frac{100*1000/3600}{44.73}) = 0.016$

$R_{rl} = f_{rl} W = 0.016(10000) = 160\ N$

Tractive Effort is computed:

$F_{max} = \frac{\mu W(l_f - f_{rl} h)/L}{1 + \muh/L} = \frac{0.6(10000)(1 - 0.016(0.5))/3}{1 - 0.6(0.5)/3} = 2204\ N$

Looking back to the force balancing equation:

$ma = F_t - R_a - R_{rl} - R_g = 2204 - 183 - 160 - 0 = 1861\ N$

Divide out mass, which can be computed from weight by dividing out gravity.

$W = mg$

$m = \frac{W}{g} = 100009.8 = 1020kg$

Thus, divide mass from the force and acceleration can be found.

$a = \frac{F}{m} = 18611020 = 1.82\ m/s^2$

Thus, the vehicle is accelerating at a rate of 1.82 meters per second squared.

Example 2: Going Up a Hill

Using the same case from Example 1, assume that instead the racecar encounters a steep hill that it must travel up. It is desired that the driver maintain the 100 km/h velocity at a very minimum. With that being said, what would be the maximum grade that the hill could be?

Solution

t the steepest eligible hill, the racecar would be able to maintain 100 km/h without any room for acceleration or deceleration. Therefore, acceleration goes to zero. All other values would stay the same from Example 1. Using the grade formula, the maximum grade can be calculated.

$G = \frac{F_t - F_a - F_{rl}}{W} = \frac{2204 - 183 - 160}{10000} = 0.1861$

The maximum allowable grade is 18.61%.

Thought Question

Problem

Why is it that, in mountainous country, trucks and cars have different speed limits?

Answer

Tractive effort is one of the leading reasons, as trucks have a harder time going up steep hills than typical passenger cars, but it is not the only one. Safety is another leading reason, surprisingly, as big rig trucks are obviously more difficult to control in a harsh environment, such as a mountain pass.

Sample Problem

It has been estimated that a Tour-de-France champion could generate a sustained 510 Watts of power while a healthy young human male (HYHM) can generate about 310 Watts of power. The bicycling champion and HYHM are going to race (on bicycles) up a hill with a 6% upgrade, that is five miles long, and the elevation at the top of the hill is 5000 feet. Both rider/bicycle combinations weigh 170 lbs, with frontal area 0.4m2 and coefficient of drag 0.9 (values being typical of bicyclists in crouched racing positions). The coefficient of rolling resistance for both bicycles is 0.01. Assume {\displaystyle \rho } ρ is 1.0567 kg/cubic-m. Remember, power equals the product of force and velocity. (1) Who gets to the top first? (2) How much longer does it take the loser to make it to the top?

Answer available at the end of this section.

Additional Questions

Homework

1. Determine the maximum grade an empty tractor-trailer with a mass of 10000 kg can climb at a sustained speed of 75 km/h.

The frontal area (A) of the truck is 10 m^2, the drag coefficient = 1.00 and the available power is 280 kW, ρ=1.2 $kg*m^{-3}$, g = 9.81 $m*sec^{-2}$

2. What is the maximum grade a bicycle can climb at a sustained speed of 12 km/h assuming the following:

- Rho = density of air (1.20 $kg*m^{-3}$)
- g = 9.81 $m*sec^{-2}$
- Mass (with bicyclist) = 100 kg
- Nominal Power = 1 kW
- Frontal Area (A) = 1 m^2
- Drag Coefficient = 0.5
- Available power = 100 percent of nominal power
- Fr = 0.020 W

3. Bicyclists get tired. Assume the bicyclist slowly loses power (and thus velocity) going up the hill of the grade you just determined in question 2 above. If available power drops continuously from 100% to 0% over the course of the hour, to the nearest kilometer, how far does the bicyclist travel? (Use appropriate approximations.)

Additional Questions

1. What is maximum grade a vehicle can travel, what effects it? Explain how to find maximum grade at a given speed. Explain what each variable means and how it contributes to figuring max grade.
2. Why are A and C_D the only things need to calculate air resistance. → Because it is a crude approximation of aerodynamics. Real aerodynamics are of course complicated, but the additional detail does not sufficiently improve accuracy.
3. If there were no heavy vehicles, would planners design roads with steeper grades? → Yes, they should. There has been discussion for a number of years for separate car and truck highways, since car-only highways would be cheaper and truck-only highways could avoid congestion. However additional rights of way would be required
4. Define rolling resistance. Why is rolling resistance different for cars and trucks? What is the difference between rolling resistance and maximum grade?
5. What is the relationship between available power and nominal power? Why do trucks use more of their nominal power than cars.
6. What is the mass/power ratio used for?

7. What happens if air resistance is larger than tractive effort force.
8. What is maximum grade a road can have?
9. What forces affect vehicle motion?
10. What are two relevant parameters for passenger car, type {\displaystyle P}P?
11. Define F_r, F_a, F_t, F_g.
12. If given a grade of 3%, should you use 0.03 or 3 in the equation?
13. How does coefficient of rolling resistance vary between vehicles.
14. What kind of safety factors are in these equations.

Variables

- F_t - Tractive Effort Force
- F_a - Aerodynamic resistance Force
- F_{rl} - Rolling Resistance Force
- F_g - Grade Resistance Force
- W - Vehicle Weight
- m - Vehicle Mass
- a - Acceleration
- ρ - Air Density
- A - Frontal area of the vehicle
- C_D - Coefficient of Drag
- V - Speed of the vehicle
- f_{rl} - Rolling Friction
- G - Grade
- μ - Coefficient of road adhesion
- l_r - Distance from rear axle to vehicle's center of gravity
- l_f - Distance from front axle to vehicle's center of gravity
- h - Height of the center of gravity above the roadway surface
- L - Length of wheelbase

Key Terms

- Tractive Effort
- Aerodynamic Resistance
- Rolling Resistance
- Grade Resistance

Answers

Sample Problem

The winner is obvious the Tour-de-France champion, as everything between the two is equal with the exception of power. Since the champion produces more power, he, by default, would win.

We need to find the maximum steady state speed for each racer to compute the differences in arrival time. At the steady state speed we must have:

$F_t = R_a + R_{rl} + R_g$

Where:

- F_t = Tractive Effort
- m = Vehicle Mass
- a = Acceleration
- R_a = Aerodynamic Resistance
- R_{rl} = Rolling Resistance
- R_g = Grade Resistance

We also know:

- rho = 1.0567 kg/cubic-m
- \(W|) = 756 N
- f_{rl} = 0.01
- G = 0.06
- A = 0.4 m^2
- C_D = 0.9

To estimate available tractive effort we can use the definition of power as time rate of work, P = FV, to get:

$F = \frac{P}{v}$

Substituting in the components of the individual formulas to the general one, we get:

$\frac{P}{v} = \frac{\rho}{2} A C_D V^2 + f_{rl} W + WG$

If we compute rolling and grade resistance, we get:

$R_{rl} = f_{rl} W = 0.01(756) = 7.56\ N$

$R_g = WG = 756(0.06) = 45.36\ N$

Together, they add up to 52.92 N.

Aerodynamic Resistance can be found to be:

$R_a = \frac{1.0567}{2}(0.4)(0.9)v^2 = 0.19v^2$

With everything substituted into the general formula, the end result is the following formulation:

$v = \frac{P}{[0.19v^2 + 52.92]}$

This problem can be solved iteratively (setting a default value for v and then computing through iterations) or graphically. Either way, when plugging in 510 watts for the Tour-de-France champion, the resulting velocity is 7.88 meters/second. Similarly, when plugging in 310 watts for the HYHM, the resulting velocity is 5.32 meters/second.

The hill is five miles in length, which translates to 8.123 kilometers, 8,123 meters. It will take the champion 1030 seconds (or 17.1 minutes) to complete this link, whereas the HYHM will take 1527 seconds (or 25.4 minutes). The resulting difference is 8.3 minutes.

This page titled 7.2: Grade is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

7.3: Earthwork

Earthwork is something that transportation projects seldom avoid. In order to establish a properly functional road, the terrain must often be adjusted. In many situations, geometric design will often involve minimizing the cost of earthwork movement. Earthwork is expressed in units of volumes (cubic meters in metric). Increases in such volumes require additional trucks (or more runs of the same truck), which cost money. Thus, it is important for designers to engineer roads that require very little earthwork.

The dump truck is amongst the equipment necessary for earthwork to occur

Cross Sections and Volume Computation

A Roadway Cross Section on otherwise Level Ground

To determine the amount of earthwork to occur on a given site, one must calculate the volume. For linear facilities, which include highways, railways, runways, etc., volumes can easily be calculated by integrating the areas of the cross sections (slices that go perpendicular to the centerline) for the entire length of the corridor. More simply, several cross sections can be selected along the corridor and an average can be taken for the entire length. Several different procedures exist for calculating areas of earthwork cross sections. In the past, the popular method was to draw cross sections by hand and use a planimeter to measure area. In modern times, computers use a coordinate method to assess earthwork calculations. To perform this task, points with known elevations need to be identified around the cross section. These points are considered in the (X, Y) coordinate plane, where X represents the horizontal axis paralleling the ground and Y represents the vertical axis that is elevation. Area can be computed with the following formula:

$$|A = \frac{1}{2}\sum_{i=1}^{n} X_i(Y_{i+1} - Y_{i-1}| \tag{7.3.1}$$

Where:

- A = Area of Cross-Section
- n = Number of Points on Cross Section (Note: n+1 = 1 and 1-1=n, for indexing)
- X = X-Coordinate
- Y = Y-Coordinate

With this, earthwork volumes can be calculated. The easiest means to do so would by using the average end area method, where the two end areas are averaged over the entire length between them.

$$V = \frac{A_1 + A_2}{2}L \tag{7.3.2}$$

Where:

- V = Volume
- A_1 = Cross section area of first side
- A_2 = Cross section area of second side
- L = Length between the two areas

If one end area has a value of zero, the earthwork volume can be considered a pyramid and the correct formula would be:

$$V = \frac{AL}{3} \tag{7.3.3}$$

A more accurate formula would the prismoidal formula, which takes out most of the error accrued by the average end area method.

$$V_p = \frac{L(A_1 + 4A_m + A_2)}{6} \tag{7.3.4}$$

Where:

- V_p = Volume given by the prismoidal formula
- A_m = Area of a plane surface midway between the two cross sections

Cut and Fill

Various sections of a roadway design will require bringing in earth. Other sections will require earth to be removed. Earth that is brought in is considered **Fill** while earth that is removed is considered **Cut**. Generally, designers generate drawings called Cut and Fill Diagrams, which illustrate the cut or fill present at any given site. This drawing is quite standard, being no more than a graph with site location on the X-axis and fill being the positive range of the Y-axis while cut is the negative range of the Y-axis.

A Typical Cut/Fill Diagram

Mass Balance

Using the data for cut and fill, an overall mass balance can be computed. The mass balance represents the total amount of leftover (if positive) or needed (if negative) earth at a given site based on the design up until that point. It is a useful piece of information because it can identify how much remaining or needed earth will be present at the completion of a project, thus allowing designers to calculate how much expense will be incurred to haul out excess dirt or haul in needed additional. Additionally, a mass balance diagram, represented graphically, can aid designers in moving dirt internally to save money.

Similar to the cut and fill diagram, the mass balance diagram is illustrated on two axes. The X-axis represents site location along the roadway corridor and the Y-axis represents the amount of earth, either in excess (positive) or needed (negative).

A Typical Mass Diagram (Note: Additional Dirt is Needed in this Example)

Demonstrations

- Flash animation: Shrinkage (contributed by Oregon State University faculty and students)
- Flash animation: Swell (contributed by Oregon State University faculty and students)

Examples

Example 1: Computing Volume

A roadway is to be designed on a level terrain. This roadway is 150 meters in length. Four cross sections have been selected, one at 0 meters, one at 50 meters, one at 100 meters, and one at 150 meters. The cross sections, respectively, have areas of 40 square meters, 42 square meters, 19 square meters, and 34 square meters. What is the volume of earthwork needed along this road?

Solution

Three sections exist between all of these cross sections. Since none of the sections end with an area of zero, the average end area method can be used. The volumes can be computed for respective sections and then summed together.

Section between 0 and 50 meters:

$V = \frac{A_1+A_2}{2} L = \frac{40\ m^2+42\ m^2}{2} \cdot 50\ m = 2050\ m^3$

Section between 50 and 100 meters:

$V = \frac{A_1+A_2}{2} L = \frac{42\ m^2+19\ m^2}{2} \cdot 50\ m = 1525\ m^3$

Section between 100 and 150 meters:

$V = \frac{A_1+A_2}{2} L = \frac{19\ m^2+34\ m^2}{2} \cdot 50\ m = 1325\ m^3$

Total volume is found to be:

$2050\ m^3 + 1525\ m^3 + 1325\ m^3 = 4900\ m^3$

Example 2: Mass Balance

Given the following cut/fill profile for each meter along a 10-meter strip of road built on very, very hilly terrain, estimate the amount of dirt left over or needed for the project.

- 0 meters: 3 meters of fill
- 1 meter: 1 meter of fill
- 2 meters: 2 meters of cut
- 3 meters: 5 meters of cut
- 4 meters: 7 meters of cut
- 5 meters: 8 meters of cut
- 6 meters: 2 meters of cut
- 7 meters: 1 meter of fill
- 8 meters: 3 meters of fill
- 9 meters: 6 meters of fill
- 10 meters: 7 meters of fill

Solution

If 'cut' is considered an excess of available earth and 'fill' is considered a reduction of available earth, the problem becomes one of simple addition and subtraction.

$[(-3\ m) + (-1\ m) + 2\ m + 5\ m + 7\ m + 8\ m + 2\ m + (-1\ m) + (-3\ m) + (-6\ m) + (-7\ m)] \cdot 1\ m^2 = 3\ m^3$

3 m^3 of dirt remain in excess.

Thought Question

Problem

If it is found that the mass balance is indeed balanced (end value of zero), does that automatically mean that no dirt transport, either out of or into the site, is needed?

Solution

No. Any soil scientist will eagerly state that dirt type can change with location quite quickly, depending on the region. So, if half a highway cuts from the earth and the other half needs fill, the dirt pulled from the first half cannot be simply dumped into the second half, even if mathematically it balances. If the soil types are different, the exact numbers of volume needed may be different, as different soil types have different properties (settling, water storage, etc.). In the worse case, not consulting a soil scientist could result in your road being washed out!

Sample Problem

Given the end areas below, calculate the volumes of cut (in cubic meters) and fill between stations 0+00 and 2+50. Determine the true amount of excess cut or fill to be removed.

- 0+00: Fill = 60
- 0+50: Fill = 50

- 0+75: Cut = 0, Fill = 25
- 1+00: Cut = 10, Fill = 5
- 1+15: Cut = 15, Fill = 0
- 1+50: Cut = 30

Answer available at the end of this section.

Homework

1. What is a Mass Diagram and why is it important? What are the units on the vertical axis of a Mass Diagram (Metric)?

2. Looking at a mass diagram, suppose you see that at Station 1+750 the ordinate equals -1000, at Station 1+780 the ordinate is -2750, and at Station 1+900 the ordinate is zero. A) Describe what happens between stations 1+750 and 1+780. B) If the project ends at Station 1+900, what can you conclude about the earthwork required for this project?

Demonstrations

- Road construction and earthmoving

Variables

- A - Area of Cross-Section
- n - Number of Points on Cross Section (Note: n+1 = 1 and 1-1=n, for indexing)
- X - X-Coordinate
- Y - Y-Coordinate
- V - Volume
- A_1 - Cross section area of first side
- A_2 - Cross section area of second side
- L - Length between the two areas
- V_p - Volume given by the prismoidal formula
- A_m - Area of a plane surface midway between the two cross sections

Key Terms

- Cut
- Fill
- Mass Balance
- Area
- Volume
- Earthwork
- Prismoidal Volume

Answers

Sample Problem

wo different methods need to be used here to compute earthwork volumes along the five strips. The average end area method can be used for non-zero sections. The pyramid method needs to be used for areas with zero ends.

For 0+00 to 0+50, use average end area:

$Fill = \frac{60+50}{2}(50) = 2750$

For 0+50 to 0+75, use average end area:

$Fill = \frac{50+25}{2}(25) = 937.5$

For 0+75 to 1+00, use the average end area method for the fill section and the pyramid method for the cut section:

$Fill = \frac{25+5}{2}(25) = 375$

$Cut = \frac{10(25)}{3} = 83.3$

For 1+00 to 1+15, use the pyramid method for the fill section and the average end area method for the cut section:

$Fill = \frac{5(15)}{3} = 25$

$Cut = \frac{10+15}{2}(15) = 187.5$

For 1+15 to 1+50, use the average end area method:

$Cut = \frac{15+30}{2}(35) = 787.5$

The sums of both cut and fill can be found:

Fill = 4087.5 cubic-meters

Cut = 1058.3 cubic-meters

Thus, 3029.2 cubic-meters of dirt are needed to meet the earthwork requirement for this project.

This page titled 7.3: Earthwork is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

7.4: Horizontal Curves

Horizontal Curves are one of the two important transition elements in geometric design for highways (along with Vertical Curves). A horizontal curve provides a transition between two tangent strips of roadway, allowing a vehicle to negotiate a turn at a gradual rate rather than a sharp cut. The design of the curve is dependent on the intended design speed for the roadway, as well as other factors including drainage and friction. These curves are semicircles as to provide the driver with a constant turning rate with radii determined by the laws of physics surrounding centripetal force.

Fundamental Horizontal Curve Properties

A Horizontal Curve in Oslo, Norway

Physics Properties

Aside from momentum, when a vehicle makes a turn, two forces are acting upon it. The first is gravity, which pulls the vehicle toward the ground. The second is centrifugal force, for which its opposite, centripetal acceleration is required to keep the vehicle on a curved path. For any given velocity, the centripetal force needs to be greater for a tighter turn (one with a smaller radius) than a broader one (one with a larger radius). On a level surface, side friction f_s serves as a countering force to the centrifugal force, but it generally provides very little resistance/force. Thus, a vehicle has to make a very wide circle in order to make a turn on the level.

Given that road designs usually are limited by very narrow design areas, wide turns are generally discouraged. To deal with this issue, designers of horizontal curves incorporate roads that are tilted at a slight angle. This tilt is defined as superelevation, or e, which is the amount of rise seen on an angled cross-section of a road given a certain run, otherwise known as slope. The presence of superelevation on a curve allows some of the centripetal force to be countered by the ground, thus allowing the turn to be executed at a faster rate than would be allowed on a flat surface. Superelevation also plays another important role by aiding in drainage during precipitation events, as water runs off the road rather than collecting on it. Generally, superelevation is limited to being less than 14 percent, as engineers need to account for stopped vehicles on the curve, where centripetal force is not present.

The allowable radius R for a horizontal curve can then be determined by knowing the intended design velocity V, the coefficient of friction, and the allowed superelevation on the curve.

$$R = \frac{v^2}{g(e + f_s)} \tag{7.4.1}$$

With this radius, practitioners can determine the degree of curve to see if it falls within acceptable standards. Degree of curve, D_a, can be computed through the following formula, which is given in Metric.

(7.4.2)

Where:

- D_a= Degree of curve [angle subtended by a 30.5-m (100 ft) arc along the horizontal curve

Application of Superelevation

Bristol Motor Speedway

One place you will see steep banking is at automobile racetracks. These tracks do not operate in winter, and so can avoid the problems of banking in winter weather. Drivers are also especially skilled, though crashes are not infrequent. For NASCAR fans, the following table may be of interest.

Table: Banking at US Racetracks

Track	Length (miles)	Banking (degrees)
Chicago Motor Speedway	1	0.00
Infineon Raceway	1.949	
Watkins Glen International	2.45	
Pocono Raceway	2.5	6.00
Homestead-Miami Speedway	1.5	8.00
Indianapolis Motor Speedway	2.5	9.00
Memphis Motorsports Park	0.75	11.00
Phoenix International Raceway	1	11.00
Las Vegas Motor Speedway	1.5	12.00
Martinsville Speedway	0.526	12.00
New Hampshire Int'l Speedway	1.058	12.00
California Speedway	2	14.00
Kentucky Speedway	1.5	14.00
Richmond International Raceway	0.75	14.00
Kansas Speedway	1.5	15.00
Michigan International Speedway	2	18.00
Nashville Speedway USA	0.596	18.00
North Carolina Speedway	1.017	22.00
Darlington Raceway	1.366	23.00
Atlanta Motor Speedway	1.54	24.00
Dover Downs Int'l Speedway	1	24.00
Lowe's Motor Speedway	1.5	24.00
Texas Motor Speedway	1.5	24.00
Daytona International Speedway	2.5	31.00
Talladega Superspeedway	2.66	33.00
Bristol Motor Speedway	0.533	36.00

Geometric Properties

Horizontal curves occur at locations where two roadways intersect, providing a gradual transition between the two. The intersection point of the two roads is defined as the **Point of Tangent Intersection (PI)**. The location of the curve's start point is defined as the **Point of Curve (PC)** while the location of the curve's end point is defined as the **Point of Tangent (PT)**. The PC is a distance T from the PI, where T is defined as Tangent Length. Tangent Length can be calculated by finding the central angle of the curve, in degrees. This angle is equal to the supplement of the interior angle between the two road tangents.

A Typical Horizontal Curve (Plan View)

$$T = Rtan\left(\frac{\Delta}{2}\right) \tag{7.4.3}$$

Where:

- T= tangent length (in length units)

- Δ= central angle of the curve, in degrees
- R= curve radius (in length units)

The PT is a distance L from the PC where L is defined as Curve Length. Curve length can be determined using the formula for semicircle length:

$$L = \frac{R\Delta\pi}{180} \tag{7.4.4}$$

The distance between the PI and the vertex of the curve can be easily calculated by using the property of right triangles with T and R. Taking this distance and subtracting off the curve radius R, the external distance E, which is the smallest distance between the curve and PI, can be found.

$$E = R\left(\frac{1}{cos(\frac{\Delta}{2})} - 1\right) \tag{7.4.5}$$

Where:

- E= external distance (in length units)

Similarly, the middle ordinate M can be found. The middle ordinate is the maximum distance between a line drawn between PC and PT and the curve. It falls along the line between the curve's vertex and the PI.

$$M = R\left(1 - cos\left(\frac{\Delta}{2}\right)\right) \tag{7.4.6}$$

Where:

- M= middle ordinate (in length units)

Similarly, the geometric formula for chord length can find C, which represents the chord length for this curve.

$$C = 2Rsin\left(\frac{\Delta}{2}\right) \tag{7.4.7}$$

Sight Distance Properties

Limited Curve Sight Distance Ahead

Unlike straight, level roads that would have a clear line of sight for a great distance, horizontal curves pose a unique challenge. Natural terrain within the inside of the curve, such as trees, cliffs, or buildings, can potentially block a driver's view of the upcoming road if placed too close to the road. As a result, the acceptable design speed is often reduced to account for sight distance restrictions.

Two scenarios exist when computing the acceptable sight distance for a given curve. The first is where the sight distance is determined to be less than the curve length. The second is where the sight distance exceeds the curve length. Each scenario has a respective formula that produces sight distance based on geometric properties. Determining which scenario is the correct one often requires testing both to find out which is true.

Given a certain sight distance S and a known curve length L and inner lane centerline radius R_v, the distance a sight obstruction can be from the interior edge of the road, M_s can be computed in the following formulas.

$$S < l : m_s = r_v\left(=""1 - cos\frac{28.65s}{r_v} =""\right) \tag{7.4.8}$$

<="" p=""> </l:m_s=r_v>

$$S > L : M_s = R_v\left(1 - cos\frac{28.65L}{R_v}\right) + \left(\frac{S-L}{2}\right)sin\left(\frac{28.56L}{R_v}\right) \tag{7.4.9}$$

Demonstrations

- Flash animation: Roadside Clear Zone (by Karen Dixon and Thomas Wall)
- Flash animation: Superelevation (by Karen Dixon and Thomas Wall)

Examples

Example 1: Curve Radius

A curving roadway has a design speed of 110 km/hr. At one horizontal curve, the superelevation has been set at 6.0% and the coefficient of side friction is found to be 0.10. Determine the minimum radius of the curve that will provide safe vehicle operation.

Solution

$R = \frac{v^2}{g(e+f_s)} = \frac{(110*(1000/3600))^2}{9.8(.06+0.10)} = 595/textmeters$

Example 2: Determining Stationing

A horizontal curve is designed with a 600 m radius and is known to have a tangent length of 52 m. The PI is at station 200+00. Determine the stationing of the PT.

Solution

What is known for this problem

$T = Rtan\left(\frac{\Delta}{2}\right)$

$52 = 600tan\left(\frac{\Delta}{2}\right)$

$\Delta = 9.9\ deg$

$L = \frac{R\pi\Delta}{180} = \frac{600\pi 9.9}{180} = 104$

$PC = PI - T = 200 + 00 - 0 + 52 = 199 + 48$

$PT = PC + L = 199 + 48 + 1 + 04 = 200 + 52$

Example 3: Stopping Distance

A very long horizontal curve on a one-directional racetrack has 1750-meter centerline radius, two 4-meter lanes, and a 200 km/hr design speed. Determine the closest distance from the inside edge of the track that spectators can park without impeding the necessary sight distance of the drivers. Assume that the sight distance is less than the length of the curve, a coefficient of friction of 0.3, and a perception-reaction time of 2.5 seconds.

Solution

With a centerline radius of 1750 meters, the centerline of the interior lane is 1748 meters from the vertex (1750 - (4/2)). Using the stopping sight distance formula (See Sight Distance), SSD is computed to be 664 meters. With this, the distance from the track that spectators can be parked can easily be found.

Extra \left or missing \right <="" p=""> </l:m_s=r_v>

This gives the distance (31.43 m) to the center of the inside lane. Subtracting half the lane width (2m in this case) would give the distance to the edge of the track, 29.43 m.

Sample Problem

A given curve was very poorly designed. The two-lane road used has a lower-than-average coefficient of friction (0.05), no superelevation to speak of, and 4-meter lanes. 900 kg vehicles tend to go around this curve and are stylistically top heavy. County engineers have warned that this curve cannot be traversed as safely as other curves in the area, but politicians want to keep the speed up to boost tourism in the area. The curves have a radius of 500 feet and a design speed of 80 km/hr. Because the vehicles using the curve are top heavy, they have a tendency to roll over if too much side force is exerted on them (the local kids often race around the curve at night to get the thrill of "two-wheeling"). As an engineer, you need to prove that this curve is infeasible before an accident occurs. How can you show this?

Answer available at the end of this section.

Demonstrations

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Additional Questions

Homework

1. Why might maximum superelevation be higher in South Texas than in Northern Minnesota?

2. Which conic section forms the basis of horizontal curves?

3. When a vehicle is traveling around a horizontal curve, it is subject to two forces. What are these forces, and how do they operate on the vehicle (draw a clear diagram illustrating the forces).

4. An existing horizontal curve has a radius of 100 meters, which restricts the maximum speed on this section of road. Highway officials want a maximum design speed of 150 km/hr.

Assume the coefficient of side friction is 0.15 and rate of superelevation on both the original and rebuilt sections is 0.06.

Compute the existing speed and find the new radius of curvature.

5. A flat horizontal curve on a 2-lane highway is designed with a 609.600 m radius, the curve has a tangent length of 121.920 m and the PI is at station 3+139.440

Design Speed (km/h)	Brake reaction distance (m)	Braking distance on level (m)	Calculated Stopping Sight Distance (m)	Design Stopping Sight Distance (m)
80	55.2	73.4	129.0	130
90	62.6	92.9	155.5	160

100	69.5	114.7	184.2	185
110	76.5	138.8	215.3	220

Source: AASHTO: A Policy on Geometric Design of Highways and Streets

The road has 3.6 m lanes and a 96 km/h design speed.

a. Determine the stationing of the PT. Draw your solution. b. Determine the distance that must be cleared from the inside edge of the inside lane to provide sufficient stopping sight distance. Draw your solution.

6. A horizontal curve is designed with a 600 m radius. The curve has a tangent of 125 m and the PI is at metric station 10+000 (10 kilometers and 0 meters). Determine the stationing of the PT. Draw a diagram showing your answer.

Additional Questions

1. Name 4 types of horizontal curves.
2. When are the non-simple types most used? Why are reverse curves so bad? → Special applications, including mountains, restricted right-of-way, or anywhere that a simple curve cannot be feasibly used.
3. How is weather accounted for (e.g. ice) → superelevation (e)
4. Explain m? When would m and M be equal? Are the equations for “m” used in determining the placement of buildings or billboards on existing roadways? → Used as a justification to keep stuff off the side of the road. May be used in roadway construction to avoid unmovable objects, or when placing objects to avoid unmovable road.
5. When should you design speed for curves rather than the curve for speed? → If it is new construction, design the radius to serve the desired speed. If you are setting speed limits, set the speed based on the existing curve.
6. When the angle increases, does the tangent length increase or decrease
7. Write the constraints on the calculation of horizontal curves
8. What are the characteristics of horizontal curves
9. Draw a simple horizontal curve and its components.
10. Can one add T to the PI station to get the station of the PC? (No)
11. How common is a horizontal curve on a vertical curve? → Depends on where you are, very common in mountainous areas. Not especially uncommon.
12. Are most horizontal curves designed using circles? → (In US, almost uniformly yes. It is also easier for driver, who just needs to set the steering wheel, constant readjustment is not required)
13. In English Units what does station 10+25 mean? It indicates the station is 1025 feet from 0+00.
14. On concrete roads there are grooves perpendicular to the direction of traffic to increase stopping road friction. Do these grooves (rumblestrips) reduce side friction?
15. What is maximum superelevation when a road is really icy? → superelevation cannot change seasonally (it would be too expensive to jack up the road). So Max e is annual.
16. Why are side friction factors for urban and rural roads different? What are different factors in rural vs. urban areas?
17. Why is Radius of curvature important when building roads?
18. What are some of the highest values of e and f_s around the country?
19. Why do the standards (maximums/minimums) for e and f_s change with location?
20. Label all of the forces acting on a vehicle traveling up a hill (and around a corner)
21. What is curve resistance? What affects is? Why does it matter?
22. What are the main variables used in curvature problems, superelevation?
23. Why do engineers bank curves?
24. What is centrifugal force?
25. Why would an engineer want to increase or decrease the radius of curvature?
26. What does superelevation represent
27. What shapes do horizontal and vertical curves have. Why is this helpful for the driver
28. If you want to decrease R, what strategies do you have?
29. Is side friction the same as static friction? What does it mean?
30. When the max value of e is given, can a smaller number be used in the final answer → yes
31. What is superelevation? How is it affected by centrifugal force?
32. How does f_s affect design speeds

Variables

- R - Centerline Curve Radius
- D_a - Degree of curve [angle subtended by a 30.5-m (100 ft) arc along the horizontal curve
- T - tangent length (in length units)
- Δ - Deflection angle of curve tangents. Also central angle of the curve, in degrees.
- E - External. Smallest distance between the curve and PI
- M - Middle ordinate
- L - Curve Length
- C - Chord Length
- S - Sight Distance
- M_s - Acceptable distance from inner edge of road for a sight obstruction to be placed without impeding sight distance
- R_v - Radius of innermost lane centerline

Key Terms

- PC: Point of Curve
- PI: Point of Tangent Intersect
- PT: Point of Tangent

Answers

Sample Problem

The stated speed is 80 km/hr. The easiest way would be to prove that this is too high. We will look at the innermost lane, since forces will be greater there. Using the general curve radius formula and solving for v, we find:

\(R=\frac{v^2}{g(e+f_s)}=\frac{v^2}{9.8(0.05+0)=500-(4/2)\)

$v = 15.62\ m/s = 56.23\ km/hr$

80 km/hr is much greater than 56.23 km/hr, which by default means that more force is being exerted on the vehicle than the road can counter. Thus, the curve's speed limit is dangerous and needs to be changed.

This page titled 7.4: Horizontal Curves is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

7.5: Vertical Curves

Vertical Curves are the second of the two important transition elements in geometric design for highways, the first being Horizontal Curves. A vertical curve provides a transition between two sloped roadways, allowing a vehicle to negotiate the elevation rate change at a gradual rate rather than a sharp cut. The design of the curve is dependent on the intended design speed for the roadway, as well as other factors including drainage, slope, acceptable rate of change, and friction. These curves are parabolic and are assigned stationing based on a horizontal axis.

Fundamental Curve Properties

Parabolic Formulation

A Road Through Hilly Terrain with Vertical Curves in New Hampshire

Two types of vertical curves exist: (1) Sag Curves and (2) Crest Curves. Sag curves are used where the change in grade is positive, such as valleys, while crest curves are used when the change in grade is negative, such as hills. Both types of curves have three defined points: PVC (Point of Vertical Curve), PVI (Point of Vertical Intersection), and PVT (Point of Vertical Tangency). PVC is the start point of the curve while the PVT is the end point. The elevation at either of these points can be computed as e_{PVC} and e_{PVT} for PVC and PVT respectively. The roadway grade that approaches the PVC is defined as g_1 and the roadway grade that leaves the PVT is defined as g_2. These grades are generally described as being in units of (m/m) or (ft/ft), depending on unit type chosen.

Both types of curves are in parabolic form. Parabolic functions have been found suitable for this case because they provide a constant rate of change of slope and imply equal curve tangents, which will be discussed shortly. The general form of the parabolic equation is defined below, where y is the elevation for the parabola.

$$y = ax^2 + bx + c \tag{7.5.1}$$

A Typical Crest Vertical Curve (Profile View)

At x = 0, which refers to the position along the curve that corresponds to the PVC, the elevation equals the elevation of the PVC. Thus, the value of c equals e_{PVC}. Similarly, the slope of the curve at x = 0 equals the incoming slope at the PVC, or g_1. Thus, the value of b equals g_1. When looking at the second derivative, which equals the rate of slope change, a value for a can be determined.

$$a = \frac{g_2 - g_1}{2L} \tag{7.5.2}$$

Thus, the parabolic formula for a vertical curve can be illustrated.

\[y=e_{PVC}+g_1x+\frac{(g_2-g_1)x^2}{2L}\]

Where:

- e_{pvc}: elevation of the PVC
- g_1: Initial Roadway Grade (m/m)
- g_2: Final Roadway Grade (m/m)
- L: Length of Curve (m)

Most vertical curves are designed to be Equal Tangent Curves. For an Equal Tangent Curve, the horizontal length between the PVC and PVI equals the horizontal length between the PVI and the PVT. These curves are generally easier to design.

Offset

Some additional properties of vertical curves exist. Offsets, which are vertical distances from the initial tangent to the curve, play a significant role in vertical curve design. The formula for determining offset is listed below.

$$Y = \frac{Ax^2}{200L} \tag{7.5.3}$$

Where:

- A: The absolute difference between g_2 and g_1, multiplied by 100 to translate to a percentage
- L: Curve Length
- x: Horizontal distance from PVC along curve

Stopping Sight Distance

Sight distance is dependent on the type of curve used and the design speed. For crest curves, sight distance is limited by the curve itself, as the curve is the obstruction. For sag curves, sight distance is generally only limited by headlight range. AASHTO has several tables for sag and crest curves that recommend rates of curvature, K, given a design speed or stopping sight distance. These rates of curvature can then be multiplied by the absolute slope change percentage, A to find the recommended curve length, L_m.

$$L_m = KA \tag{7.5.4}$$

Without the aid of tables, curve length can still be calculated. Formulas have been derived to determine the minimum curve length for required sight distance for an equal tangent curve, depending on whether the curve is a sag or a crest. Sight distance can be computed from formulas in other sections (See Sight Distance).

Crest Vertical Curves

The correct equation is dependent on the design speed. If the sight distance is found to be less than the curve length, the first formula below is used, whereas the second is used for sight distances that are greater than the curve length. Generally, this requires computation of both to see which is true if curve length cannot be estimated beforehand.

$$S < l : l_m = \frac{as^2}{200\left(\sqrt{h_1} + \sqrt{h_2} =""\right)^2} \tag{7.5.5}$$

<="" p=""> </l:l_m=\frac{as^2}{200>

$$S > L : L_m = 2s - \frac{200\left(\sqrt{h_1} + \sqrt{h_2}\right)^2}{A} \tag{7.5.6}$$

Where:

- L_m: Minimum Curve Length (m)
- A: The absolute difference between g_2 and g_1, multiplied by 100 to translate to a percentage
- S: Sight Distance (m)
- h_1: Height of driver's eye above roadway surface (m)
- h_2: Height of objective above roadway surface (m)

Sag Vertical Curves

Just like with crest curves, the correct equation is dependent on the design speed. If the sight distance is found to be less than the curve length, the first formula below is used, whereas the second is used for sight distances that are greater than the curve length. Generally, this requires computation of both to see which is true if curve length cannot be estimated beforehand.

$$S < l : l_m = \frac{as^2}{200(h + stan\beta)} \tag{7.5.7}$$

<="" p=""> </l:l_m=\frac{as^2}{200(h+stan>

$$S > L : L_m = 2s = \frac{200(H + Stan\beta)}{A} \tag{7.5.8}$$

Where:

- A: The absolute difference between g_2 and g_1, multiplied by 100 to translate to a percentage
- S: Sight Distance (m)
- H: Height of headlight (m)
- β: Inclined angle of headlight beam, in degrees

To find the position of the low point on a SAG vertical curve: x is the horizontal distance between the PVC and Low Point

$$x = \frac{G1L}{G1 - G2} \tag{7.5.9}$$

- G_1: Grade Down (%)
- G_2: Grade Up (%)
- L: Length of Vertical Curve (station) ei. 600 ft =6

Passing Sight Distance

In addition to stopping sight distance, there may be instances where passing may be allowed on vertical curves. For sag curves, this is not an issue, as even at night, a vehicle in the opposing can be seen from quite a distance (with the aid of the vehicle's headlights). For crest curves, however, it is still necessary to take into account. Like with the stopping sight distance, two formulas are available to answer the minimum length question, depending on whether the passing sight distance is greater than or less than the curve length. These formulas use units that are in metric.

$$S < l : l_m = \frac{a(psd^2)}{864} \tag{7.5.10}$$

< p=""> </l:l_m=\frac{a(psd^2)}{864}\]<>

$$S > L : L_m = 2PSD - \frac{864}{A} \tag{7.5.11}$$

Where:

- A: The absolute difference between g_2 and g_1, multiplied by 100 to translate to a percentage
- PSD: Passing Sight Distance (m)
- L_m: Minimum curve length (m)

Demonstrations

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Examples

Example 1: Basic Curve Information

A 500-meter equal-tangent sag vertical curve has the PVC at station 100+00 with an elevation of 1000 m. The initial grade is -4% and the final grade is +2%. Determine the stationing and elevation of the PVI, the PVT, and the lowest point on the curve.

Solution

The curve length is stated to be 500 meters. Therefore, the PVT is at station 105+00 (100+00 + 5+00) and the PVI is in the very middle at 102+50, since it is an equal tangent curve. For the parabolic formulation, c equals the elevation at the PVC, which is stated as 1000 m. The value of b equals the initial grade, which in decimal is -0.04. The value of a can then be found as 0.00006.

Using the general parabolic formula, the elevation of the PVT can be found:

$y = 0.00006x^2 + -.04x + 1000 = 0.00006(500)^2 + -0.4(500) + 1000 = 995\ m$

Since the PVI is the intersect of the two tangents, the slope of either tangent and the elevation of the PVC or PVT, depending, can be used as reference. The elevation of the PVI can then be found:

$y = -0.04x + 1000 = -0.04(250) + 1000 = 990$

To find the lowest part of the curve, the first derivative of the parabolic formula can be found. The lowest point has a slope of zero, and thus the low point location can be found:

$dy/dx = 0.00012x + -.04 = 0.00012x + -0.4 = 0$

$x = 333.33\ m$

Using the parabolic formula, the elevation can be computed for that location. It turns out to be at an elevation of 993.33 m, which is the lowest point along the curve.

Example 2: Adjustment for Obstacles

A current roadway is climbing a hill at an angle of +3.0%. The roadway starts at station 100+00 and elevation of 1000 m. At station 110+00, there is an at-grade railroad crossing that goes over the sloped road. Since designers are concerned for the safety of drivers crossing the tracks, it has been proposed to cut a level tunnel through the hill to pass beneath the railroad tracks and come out on the opposite side. A vertical crest curve would connect the existing roadway to the proposed tunnel with a grade of (-0.5)%. The prospective curve would start at station 100+00 and have a length of 2000 meters. Engineers have stated that there must be at least 10 meters of separation between the railroad tracks and the road to build a safe tunnel. Assume an equal tangent curve. With the current design, is this criteria met?

Solution

One way to solve this problem would be to compute the elevation of the curve at station 110+00 and then see if it is at least 10 meters from the tangent. Another way would be to use the Offset Formula. Since A, L, and x are all known, this problem can be easily solved. Set x to 1000 meters to represent station 110+00.

$Y = \frac{Ax^2}{200L} = \frac{(3.5)(1000)^2}{200(2000)} = 8.75\ m$

The design DOES NOT meet the criteria.

Vertical Curve Example 2.JPG

Example 3: Stopping Sight Distance

A current roadway has a design speed of 100 km/hr, a coefficient of friction of 0.1, and carries drivers with perception-reaction times of 2.5 seconds. The drivers use cars that allows their eyes to be 1 meter above the road. Because of ample roadkill in the area, the road has been designed for carcasses that are 0.5 meters in height. All curves along that road have been designed accordingly.

The local government, seeing the potential of tourism in the area and the boost to the local economy, wants to increase the speed limit to 110 km/hr to attract summer drivers. Residents along the route claim that this is a horrible idea, as a particular curve called "Dead Man's Hill" would earn its name because of sight distance problems. "Dead Man's Hill" is a crest curve that is roughly 600 meters in length. It starts with a grade of +1.0% and ends with (-1.0)%. There has never been an accident on "Dead Man's Hill" as of yet, but residents truly believe one will come about in the near future.

A local politician who knows little to nothing about engineering (but thinks he does) states that the 600-meter length is a long distance and more than sufficient to handle the transition of eager big-city drivers. Still, the residents push back, saying that 600 meters is not nearly the distance required for the speed. The politician begins a lengthy campaign to "Bring Tourism to Town", saying that the residents are trying to stop "progress". As an engineer, determine if these residents are indeed making a valid point or if they are simply trying to stop progress?

Solution

Using sight distance formulas from other sections, it is found that 100 km/hr has an SSD of 465 meters and 110 km/hr has an SSD of 555 meters, given the criteria stated above. Since both 465 meters and 555 meters are less than the 600-meter curve length, the correct formula to use would be:

\(S<l:L_m=\frac{as^2}{200 \left(\sqrt{h_1}_\sqrt{h_2}="" \right)^2="\frac{2(555)^2}{200(\sqrt{1.0}+\sqrt{0.5})^2=1055" \text{="" }="" m\)<="" p=""> </l:L_m=\frac{as^2}{200>

Since the 1055-meter minimum curve length is greater than the current 600-meter length on "Dead Man's Hill", this curve would not meet the sight distance requirements for 110 km/hr.

This seems like a very large gap. The question becomes, was the curve even good enough at 100 km/hr? Using the same formula, the result is:

\(S<l:L_m=\frac{as^2}{200 \left(="" \sqrt{h_1}+\sqrt{h_2}="" \right)^2="\frac{2(465)^2}{200(\sqrt{1.0}+\sqrt{0.5})^2=740" \text{="" }m\)<="" p=""> </l:L_m=\frac{as^2}{200>

740 meters for a minimum curve length is far greater than the existing 600-meter curve. Therefore, the residents are correct in saying that "Dead Man's Hill" is a disaster waiting to happen. As a result, the politician, unable to hold public confidence by his "progress" comment, was forced to resign.

Thought Question

Problem

Sag curves have sight distance requirements because of nighttime sight distance constraints. The headlights on cars have a limited angle at which they can shine with bright enough intensity to see objects far off in the distance. If the government were to allow a wider angle of light to be cast out on standard car headlights, would this successfully provide more stopping sight distance?

Solution

Yes, of course. For a single car traveling on a road with many sag curves, the design speed could be increased since more road could be seen. However, when additional cars were added to that same road, problems would begin to appear. With a greater angle of light being cast from headlights, drivers in opposing lanes would be severely blinded, forcing them to slow down to avoid causing an accident. Just think of the last time somebody drove by with their 'brights' on and blinded you. This problem could cause more accidents and force people to slow down, thus producing a net loss overall.

Sample Problem

To help prevent future collisions between cars and trains, an at-grade crossing of a rail road by a country road is being redesigned so that the county road will pass underneath the tracks. Currently the vertical alignment of the county road consists of an equal tangents crest vertical curve joining a 4% upgrade to a 3% downgrade. The existing vertical curve is 450 feet long, the PVC of this curve is at station 48+24.00, and the elevation of the PVC is 1591.00 feet. The centerline of the train tracks is at station 51+50.00. Your job is to find the shortest vertical curve that provides 20 feet of clearance between the new county road and the train tracks, and to make a preliminary estimate of the cut that will be needed to construct the new curve.

Answer available at the end of this section.

Homework

1. Which conic section forms the basis of vertical curves?

2. A section of a roadway goes up a 6% incline and includes no horizontal curves. Point A, which lies along this incline, has a station marker. Using a wheel measure, you measure the distance along the centerline from point A to the next PVC and find it is 880 meters. If point A is at Station 0+700, what is the station of the PVC?

3. Identify 4 criteria used to set the minimum length of a sag vertical curve (4 points), place them in order from most to least important (1 point):

4. An equal tangent sag vertical curve joins a –0.5 % grade with a +3.0% grade. Assume H=0.6, headlights are aimed upward at an angle of 1 degree

If the PVI of the grades is at metric station 10+100 and has an elevation of 300 meters, using the headlight criterion determine the station elevation of the PVC and PVT for a design speed of 100 km/hr and a coefficient of friction of 0.30. Draw a diagram with the values of your solution.

5. A crest vertical curve on a 2-way, 2-lane road is designed for 100 km/hr. The grades are 1.6% and – 1.4%. If the tangents intersect at metric station 2 + 000.00 (where 2 represents 2 km and 000.00 is 0 meters) and at an elevation of 100 meters above sea level, determine the stopping sight distance, the length of the curve, and the stations and elevations for the Beginning and End of the Vertical Curve. Assume h1 = 1.1 m and h2 = 0.15 m, perception reaction time is 2.5 seconds, f=0.30.

Draw a diagram with the values of your solution.

6. An equal tangent sag vertical curve joins a -2.5% grade with a +3.0% grade. Assume H=0.67, headlights are aimed upward at an angle of 1 degree. If the PVI of the grades is at metric station 5+750 and has an elevation of 0 meters, using the headlight criterion determine the station elevation of the PVC and PVT for a design speed of 120 km/hr and a coefficient of friction of 0.31. Draw a diagram with the values of your solution.

7. You have been asked to design an equal-tangent vertical curve to connect grades of the +1.0% and -2.0%. The design speed is 113 km/h. Assume a value of 3.4 m*s-2 for deceleration rate, a value of 2.5s for reaction time, and a value of 9.807m*s for gravitational constant.

A. Calculate the Stopping Sight Distance (SSD) if we assume vehicles travel at the design speed.

B. If we take this SSD as the required minimum Sight Distance, determine the minimum length of curve. (According to AASHTO, we assume a driver eye height, H1, of 1.08m and a roadway object height, H2, of 0.6m)

Additional Questions

Sag Vertical Curves

1. What are the considerations when designing sag vertical curves? Which is most important?

2. What does the comfort criteria in a sag vertical curve describe? Is there a time when comfort criterion is the determining factor? Is there a time when the appearance criteria is determining?
3. What values are used for H and b in determining L for sag vertical curve?
4. When calculating SSD, which grade should be used?
5. Is it true that whether a curve is a crest or sag vertical curve is determined by the initial and final grade?
6. Are properties of headlights different for cars and trucks? → design for worst case (shorter car). Headlights are positioned by vehicle. Some vehicles have driver adjustments. Design is for normal headlights, not highbeams.
7. Why are there different equations for L depending on S>L, S
8. Is there any trick to knowing whether SL before starting?
9. How does rain/fog affect sight distance? → reduces it, so drivers should reduce speed and be more alert. Design is for normal conditions.
10. What is the drainage criterion? How is water dealt with on sag curves?
11. Why are headlights angled?

Crest Vertical Curves

1. Is there a limit on hill crests??
2. What is the difference between crest and sag curves
3. How can you determine if S > L or S < L when doing SSD calculations? --> If don't know, trial and check. If use S > L equation and S is greater than L, great, otherwise, must solve again. Why are equations different? Think about the relationships. Solve them at S = L.
4. If SSD or PSD is different in opposite directions over a crest vertical curve, is it common to have different speed limits/ warning signs. Usually speed is same in both directions, not always.
5. Is darkness considered in speed/stopping calculations (since can't see as far)? For Sag curves, there is a headlight criterion.
6. What height is used in equations for L?
7. Can the highest point on the curve be higher than the PVI? →:(No, but can be higher than the point on the curve under the PVI). How is it possible for the PVI elevation to be above the actual elevation
8. When is SSD different from AASHTO guidelines? Do engineers assume different stopping sight distances under certain conditions?
9. Why compute SSD for downgrade only?
10. Why is it not always possible to allow passing on vertical curve two lane roads. Is passing sight regularly used ? → (on 2 lane rural roads with good sight lines, yes).
11. Explain idea of stopping sign as it pertains to hills with (horizontal) curves → a bit complex for intro to transportation, covered in 4xxx level curve. Think about which governs, the horizontal SSD or the vertical SSD (i.e. whichever is lower, which depends on local conditions). Is passing sight distance based on big trucks or average sized cars? → cars (see h1 and compare it with where you are sitting in the vehicle). (Do trucks and small cars have enough sight to pass). → It is not the size of the car you are in so much as the size of the car you see (h2). You are more likely to see a big truck, so the equation is conservative for larger vehicles. When do you plan for each? (Always plan for SSD, consider PSD when there is possibility of dashed yellow line).
12. What are 3 parameters that affect the design of vertical curves
13. Explain crest vertical curve graphically
14. Determine the minimum length of the curve.
15. What type of curve is used in the design of crest vertical curves? (Ans: Parabola)
16. Why is it necessary to curve roads that have different slopes on either side of the curve?
17. Is L Arc length or horizontal displacement?
18. Why is stopping distance important?
19. Label PVC, PVI and PVT on a graph. What is the difference between PVC and PVT?
20. What is an equal tangent cuve
21. What does an aesthetically unappealing curve look like?
22. What does SSD depend on?
23. How does variable "a" used in elevation, differ from variable "A" used in length of curve.
24. In what two ways can the SSD and L be related
25. What is the first step in developing the profile of a curve -> Find the center of the curve.

Additional Problem

1. Compute length of crest vertical curve with +2.5% and –2.% grade, speed = 90 km/hr &S < L, f=0.25, PRT = 2.5 sec

1. Find the elevation along the curve?
2. Find the highest point on a crest vertical curve

Variables

- L - Curve Length
- e - Elevation of designated point, such as PVC, PVT, etc.
- g - Grade
- A - Absolute difference of grade percentages for a certain curve, in percent
- y - Elevation of curve
- Y - Offset between grade tangent from PVC and curve elevation for a specific station
- h_1 - Height of driver's eye above roadway surface
- h_2 - Height of object above roadway surface
- H - Height of headlight
- β - Inclined angle of headlight beam, in degrees
- S - Sight Distance in question
- K - Rate of curvature
- L_m - Minimum Curve Length

Key Terms

- PVC: Point of Vertical Curves
- PVI: Point of Vertical Intersection
- PVT: Point of Vertical Tangent
- Crest Curve: A curve with a negative grade change (like on a hill)
- Sag Curve: A curve with a positive grade change (like in a valley)

Answers

Sample Problem

Solution:

A.) What is the curve length?

The curves are equal tangent, so we know that: PVI - PVC = PVT - PVI = L/2

From this, we know:

PVC of First Curve: 48+24 PVI of Both Curves (this is a constant): (48+24.00 + (450/2) = 50+49)

Use the offset formula for the first curve to find the vertical distance between the tangent and the curve:

- x_{rr}: Distance between PVI and railroad tracks: (51+50 - 50+49 = 101 feet)
- PVI Elevation: 1591 + (0.04*225) = 1600 feet

Use the offset formula for the first curve to find the vertical distance between the tangent and the curve:

$Y = \frac{Ax^2}{200L}$

Where:

A = 7 (Change in Grade)

L = 450 feet

x = 51+50 - 48+24 = 326 feet

The offset, Y, at the railroad tracks's station is computed to be 8.27 feet.

The road is to be lowered an additional 20 feet. Therefore, the new offset at that sight would become 28.27 feet. The equation for the second curve becomes:

$Y = 28.27 = \frac{7x^2}{200L}$

Neither L (length of curve) or x (distance of offset from PVC) are known. However, we know this is an equal tangent curve, meaning the distance from PVC to PVI is L/2 for the curve in question. Also, the distance between PVI and the railroad tracks is x_{rr}, which is 101 feet. Therefore, (L/2) + x_{rr} equals the distance from PVC to the railroad tracks, which is what we want for x. Thus, we are left with one unknown and one equation.

$Y = 28.27 = \frac{7(L+101)^2}{200L}$

The new curve length is found to be 2812 feet.

B.) How deep is the cut at the PVI?

To find this, the offset formula can be used again, using the length L/2 as the distance from the PVC on any curve.

For the old curve, the offset is:

$Y = \frac{A(L/2)^2}{200L} = \frac{7(450/2)^2}{200(450)} = 3.935\ feet$

For the new curve, using data found from before, the offset is:

$Y = \frac{A(L/2)^2}{200L} = \frac{7(2812/2)^2}{200(2812)} = 24.605\ feet$

The difference between the two offsets is 20.67 feet. This is the depth of the cut at the PVI.

This page titled 7.5: Vertical Curves is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

INDEX

L

Logit model
3.2: Choice Modeling

M

mode choice analysis
3.2: Choice Modeling

This page titled Index is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.

GLOSSARY

S

Sample Word 1 | Sample Definition 1

This page titled Glossary is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by David Levinson et al. (Wikipedia) via source content that was edited to the style and standards of the LibreTexts platform.