T

Tableau

► Simplex Tableau

Tabu Search

Fred W. Glover
OptTek Systems, Inc., Boulder, CO, USA

Introduction

Tabu Search (TS) is a metaheuristic that guides a local heuristic search procedure to explore the solution space beyond local optimality. Widespread successes in practical applications of optimization include finding better solutions to problems in scheduling, sequencing, resource allocation, investment planning, telecommunications and many other areas. Some of the diversity of tabu search applications is shown in Table 1. (For a more comprehensive list of applications, see the book by Glover and Laguna 1997.)

Tabu search is based on the premise that methods for complex optimization problems, particularly those arising in real world applications, can function more effectively if they incorporate flexible and responsive memory. Accompanying this premise is the corollary that such memory is employed together with strategies expressly designed for exploiting it. More broadly, tabu search embodies the following principle: If a problem has exploitable features, but contains a structure sufficiently complex to prevent these features from being known in advance, then a method can derive advantages by monitoring its behavior in relation to the space in which it operates. The purpose of the monitoring is effectively to generate a map of the regions the method has visited as a foundation for modifying its behavior, where this map can take multiple forms that ultimately become expressed in the decision rules employed to negotiate the solution space. The hallmark of a TS method is therefore a capacity to guide its progress by reference to its own unfolding history. Such a method evidently is implicitly or explicitly structured to employ learning. Based on this perspective, methods that incorporate a significant portion of the tabu search framework are sometimes called Adaptive Memory Programming (AMP) methods.

The emphasis on responsive exploration (and hence purpose) in tabu search, whether in a deterministic or probabilistic implementation, derives from the supposition that a bad strategic choice can yield more information than a good random choice. In a system that uses memory, a bad choice based on strategy can provide useful clues about how the strategy may profitably be changed. Even in a space with significant randomness – which fortunately is not pervasive enough to extinguish all remnants of order in most real-world problems – a purposeful design can be more adept at uncovering the imprint of structure, and thereby at affording a chance to exploit the conditions where randomness is not all-encompassing.

These basic elements of tabu search have several important features that are summarized in Table 2.

Tabu search is concerned with finding new and more effective ways of taking advantage of the concepts embodied in Table 2, and with identifying associated principles that can expand the foundations

S.I. Gass, M.C. Fu (eds.), *Encyclopedia of Operations Research and Management Science*,
DOI 10.1007/978-1-4419-1153-7, © Springer Science+Business Media New York 2013

Tabu Search, Table 1 Illustrative tabu search applications

Scheduling	**Telecommunications**
Flow-Time Cell Manufacturing	Call Routing
Heterogeneous Processor	Bandwidth Packing
Scheduling	Hub Facility Location
Workforce Planning	Path Assignment
Classroom Scheduling	Network Design for Services
Machine Scheduling	Customer Discount Planning
Flow Shop Scheduling	Failure Immune Architecture
Job Shop Scheduling	Synchronous Optical Networks
Sequencing and Batching	
Design	**Production, Inventory and Investment**
Computer-Aided Design	Flexible Manufacturing
Fault Tolerant Networks	Just-in-Time Production
Transport Network Design	Capacitated MRP
Architectural Space Planning	Part Selection
Diagram Coherency	Multi-item Inventory Planning
Fixed Charge Network Design	Volume Discount Acquisition
Irregular Cutting Problems	Fixed Mix Investment
Lay-Out Planning	
Location and Allocation	**Routing**
Multicommodity Location/ Allocation	Vehicle Routing
Quadratic Assignment	Capacitated Routing
Quadratic Semi-Assignment	Time Window Routing
Multilevel Generalized Assignment	Multi-Mode Routing
	Mixed Fleet Routing
	Traveling Salesman
	Traveling Purchaser
	Convoy Scheduling
Logic and Artificial Intelligence	**Graph Optimization**
Maximum Satisfiability	Graph Partitioning
Probabilistic Logic	Graph Coloring
Clustering	Clique Partitioning
Pattern Recognition/ Classification	Maximum Clique Problems
Data Integrity	Maximum Planner Graphs
Neural Network Trainings	P-Median Problems
Neural Network Design	
Technology	**General Combinational Optimization**
Seismic Inversion	Zero–one Programming
Electrical Power Distribution	Fixed Charge Optimization
Engineering Structural Design	Nonconvex Nonlinear Programming
Minimum Volume Ellipsoids	All-or-None Networks
Space Station Construction	Bilevel Programming
Circuit Cell Placement	General Mixed Integer Optimization
Off-Shore Oil Exploration	

Tabu Search, Table 2 Principal tabu search features

Adaptive Memory
Selectivity (including strategic forgetting)
Abstraction and decomposition (through explicit and attributive memory)
Timing:
recency of events
frequency of events
differentiation between short term and long term
Quality and impact:
relative attractiveness of alternative choices
magnitude of changes in structure or constraining relationships
Context:
regional interdependence
structural interdependence
sequential interdependence
Responsive Exploration
Strategically imposed restraints and inducements (tabu conditions and aspiration levels)
Concentrated focus on good regions and good solution features (intensification processes)
Characterizing and exploring promising new regions (diversification processes)
Non-montonic search patterns (strategic oscillation)
Integrating and extending solutions (path relinking)

of intelligent search. As this occurs, new strategic mixes of the basic ideas emerge, leading to improved solutions and better practical implementations.

Tabu Search Foundations

The basis for tabu search may be described as follows. Given a function $f(x)$ to be optimized over a set X, TS begins in the same way as ordinary local search, proceeding iteratively from one point (solution) to another until a chosen termination criterion is satisfied. Each $x \in X$ has an associated neighborhood $N(x) \subset X$, and each solution $x' \in N(x)$ is reached from x by an operation called a move.

TS goes beyond local search by employing a strategy of modifying $N(x)$ as the search progresses, effectively replacing it by another neighborhood $N^*(x)$. As the previous discussion intimates, a key aspect of tabu search is the use of special memory structures which serve to determine $N^*(x)$, and hence to organize the way in which the space is explored.

The solutions admitted to $N^*(x)$ by these memory structures are determined in several ways. One of these, which gives tabu search its name, identifies solutions encountered over a specified horizon (and implicitly, additional related solutions), and forbids them to belong to $N^*(x)$ by classifying them tabu (The tabu terminology is intended to convey a type of restraint that embodies a cultural connotation – i.e., one that is subject to the influence of history and context, and capable of being surmounted under appropriate conditions).

The process by which solutions acquire a tabu status has several facets, designed to promote a judiciously aggressive examination of new points. A useful way of viewing and implementing this process is to conceive of replacing original evaluations of solutions by tabu evaluations, which introduce penalties to significantly discourage the choice of tabu solutions (i.e., those preferably to be excluded from $N^*(x)$, according to their dependence on the elements that compose tabu status). In addition, tabu evaluations also periodically include inducements to encourage the choice of other types of solutions, as a result of aspiration levels and longer term influences. The following subsections describe how tabu search takes advantage of memory (and hence learning processes) to carry out these functions.

Explicit and *Attributive Memory* – The memory used in TS is both explicit and attributive. Explicit memory records complete solutions, typically consisting of elite solutions visited during the search (or highly attractive but unexplored neighbors of such solutions). These special solutions are introduced at strategic intervals to enlarge $N^*(x)$, and thereby provide useful options not in $N(x)$.

TS memory is also designed to exert a more subtle effect on the search through the use of attributive memory, which records information about solution attributes that change in moving from one solution to another. For example, in a graph or network setting, attributes can consist of nodes or arcs that are added, dropped or repositioned by the moves executed. In more abstract problem formulations, attributes may correspond to values of variables or functions. Sometimes attributes are also strategically combined to create other attributes by using vocabulary building methods (Glover and Laguna 1993; Glover 1999; Glover et al. 2000).

Short-Term Memory and its Accompaniments – An important distinction in TS arises by differentiating between short-term memory and longer-term memory. Each type of memory is accompanied by its own special strategies. The most commonly used short-term memory keeps track of solution attributes that have changed during the recent past, and is called recency-based memory. To exploit this memory, selected attributes that occur in solutions recently visited are designated tabu-active, and solutions that contain tabu-active elements, or particular combinations of these attributes, are those that become tabu. This prevents certain solutions from the recent past from belonging to $N^*(x)$ and hence from being revisited. Other solutions that share such tabu-active attributes are also similarly prevented from being revisited. The use of tabu evaluations, with large penalties assigned to appropriate sets of tabu-active attributes, can allow tabu status to vary by degrees.

Managing Recency-Based Memory – The process is managed by creating one or several tabu lists, which record the tabu-active attributes and implicitly or explicitly identify their current status. The duration that an attribute remains tabu-active (measured in numbers of iterations) is called its tabu tenure. Tabu tenure can vary for different types or combinations of attributes, and can also vary over different intervals of time or stages of search. This varying tenure makes it possible to create different kinds of tradeoffs between short-term and longer-term strategies. It also provides a dynamic and robust form of search. (See, e.g., Glover 1990; Taillard 1991, Glover and Laguna 1993, 1997.)

Aspiration Levels – An important element of flexibility in tabu search is introduced by means of aspiration criteria. The tabu status of a solution (or a move) can be overruled if certain conditions are met, expressed in the form of aspiration levels. In effect, these aspiration levels provide thresholds of attractiveness that govern whether the solutions may be considered admissible in spite of being classified tabu. Clearly a solution better than any previously seen deserves to be considered admissible. Similar criteria of solution quality provide aspiration criteria over subsets of solutions that belong to common regions or that share specified features (such as a particular functional value or level of infeasibility). Additional examples of aspiration criteria are provided later.

Candidate List Strategies – The aggressive aspect of TS is reinforced by seeking the best available move

that can be determined with an appropriate amount of effort. It should be kept in mind that the meaning of best is not limited to the objective function evaluation. (As already noted, tabu evaluations are affected by penalties and inducements determined by the search history.) For situations where $N^*(x)$ is large or its elements are expensive to evaluate, candidate list strategies are used to restrict the number of solutions examined on a given iteration.

Because of the importance TS attaches to selecting elements judiciously, efficient rules for generating and evaluating good candidates are critical to the search process. Even where candidate list strategies are not used explicitly, memory structures to give efficient updates of move evaluations from one iteration to another, and to reduce the effort of finding best or near best moves, are often integral to TS implementations. Intelligent updating can appreciably reduce solution times, and the inclusion of explicit candidate list strategies, for problems that are large, can significantly magnify the resulting benefits.

The operation of these short-term elements is illustrated in Fig. 1. The representation of penalties in Fig. 1 either as large or very small expresses a thresholding effect: either the tabu status yields a greatly deteriorated evaluation or else it chiefly serves to break ties among solutions with highest evaluations. Such an effect of course can be modulated to shift evaluations across levels other than these extremes. If all moves currently available lead to solutions that are tabu (with evaluations that normally would exclude them from being selected), the penalties result in choosing a "least tabu" solution.

The TS variant called probabilistic tabu search follows a corresponding design, with a short-term component that can be represented by the same diagram. The approach additionally keeps track of tabu evaluations generated during the process that results in selecting a move. Based on this record, the move is chosen probabilistically from the pool of those evaluated (or from a subset of the best members of this pool), weighting the moves so that those with higher evaluations are especially favored. Fuller discussions of probabilistic tabu search are found in Glover (1989), Glover and Laguna (1997), Soriano and Gendreau (1993) and Crainic et al. (1993).

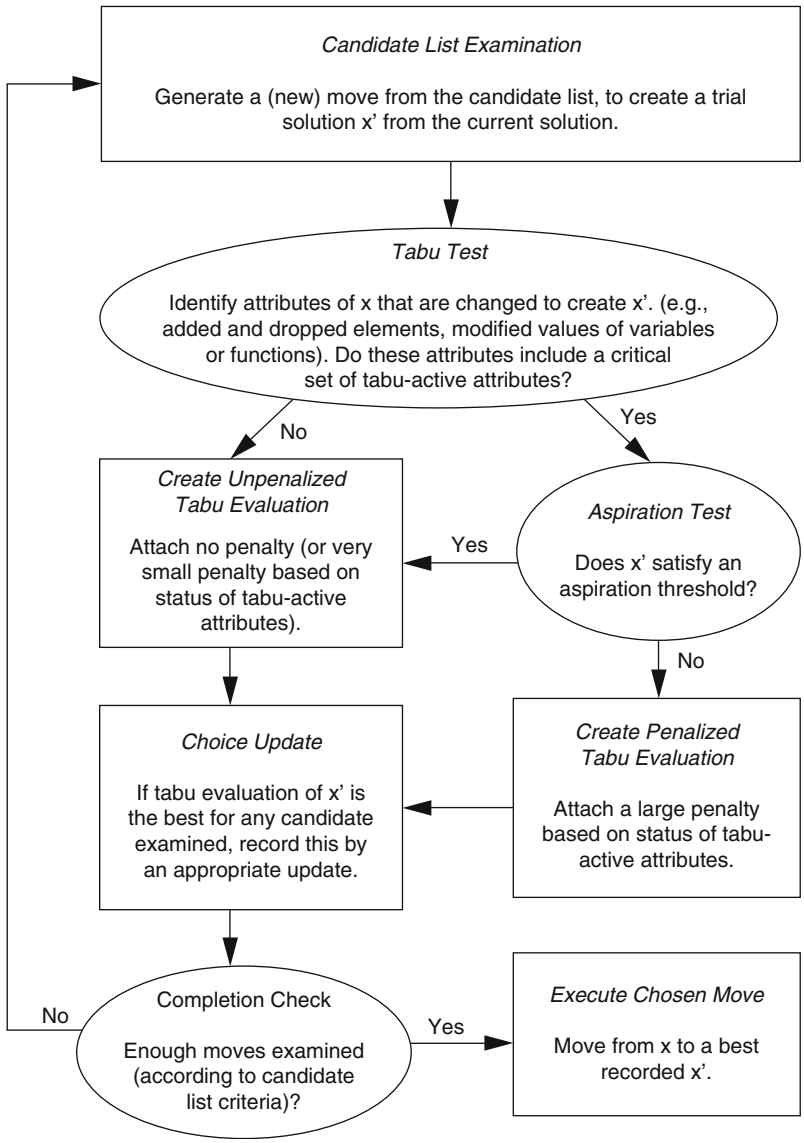
Longer-Term Memory – In some applications, the short-term TS memory components are sufficient to produce very high quality solutions. However, in general, TS becomes significantly stronger by including longer-term memory and its associated strategies.

Special types of frequency-based memory are fundamental to longer-term considerations. These operate by introducing penalties and inducements determined by the relative span of time that attributes have belonged to solutions visited by the search, allowing for regional differentiation.

Perhaps surprisingly, the use of longer-term memory does not require long solution runs before its benefits become visible. Often its improvements begin to be manifest in a relatively modest length of time, and can allow solution efforts to be terminated somewhat earlier than otherwise possible, due to finding very high quality solutions within an economical time span. The fastest methods for job shop and flow shop scheduling problems, for example, are based on including longer-term TS memory. On the other hand, it is also true that the chance of finding still better solutions as time grows – in the case where an optimal solution is not already found – is enhanced by using longer-term TS memory in addition to short-term memory.

Intensification and Diversification – Two highly important longer-term components of tabu search are intensification strategies and diversification strategies. Intensification strategies are based on modifying choice rules to encourage move combinations and solution features historically found good. They may also initiate a return to attractive regions to search them more thoroughly. A simple instance of this second type of intensification strategy is shown in Fig. 2.

The strategy for selecting elite solutions is italicized in Fig. 2 due to its importance. Two variants have proved quite successful. One, due to, introduces a diversification measure to assure the solutions recorded differ from each other by a desired degree, and then erases all short-term memory before resuming from the best of the recorded solutions. The other variant, due to Nowicki and Smutnicki (1993), keeps a bounded length sequential list that adds a new solution at the end only if it is better than any previously seen. The current last member of the list is always the one chosen (and removed) as a basis for resuming search. However, TS short-term memory that accompanied this solution also is saved, and the first move also forbids the move previously taken from this solution, so that a new solution path will be launched.

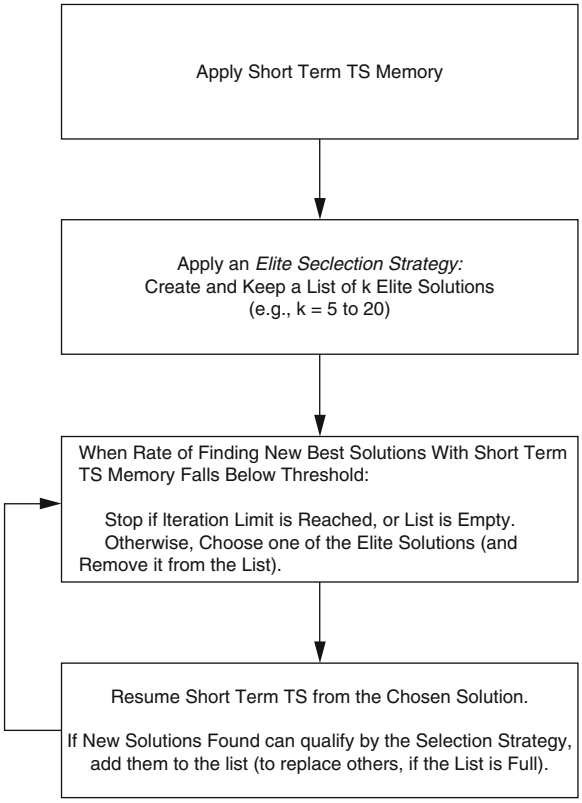


Tabu Search, Fig. 1 Tabu evaluation (short term memory)

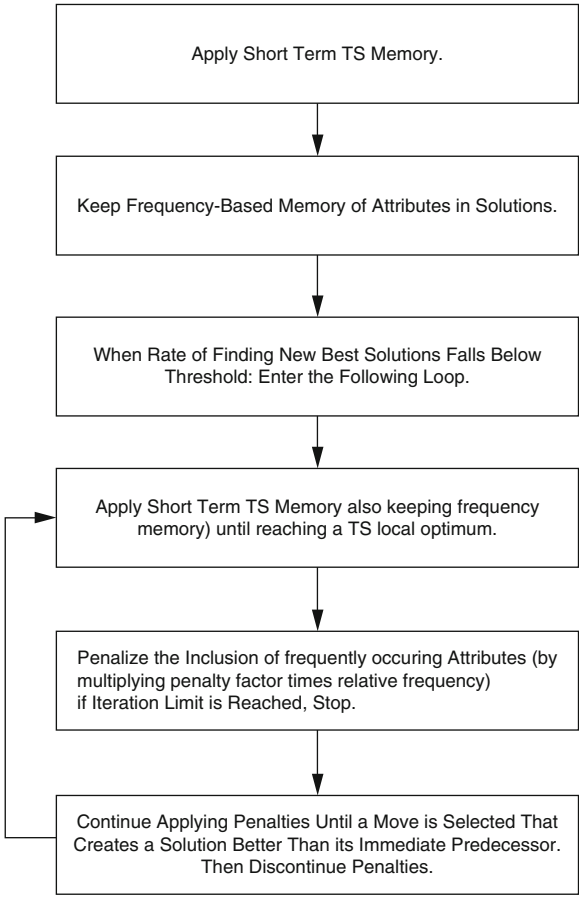
This second variant is related to a strategy that resumes the search from unvisited neighbors of solutions previously generated (Glover 1990). Such a strategy keeps track of the quality of these neighbors to select an elite set, and restricts attention to specific types of solutions, such as neighbors of local optima or neighbors of solutions visited on steps immediately before reaching such local optima. This type of unvisited neighbor strategy has been little examined. It is noteworthy, however, that the two variants previously indicated have provided solutions of remarkably high quality.

Diversification Strategies – TS diversification strategies, as their name suggests, are designed to drive the search into new regions. Often they are based on modifying choice rules to bring attributes into the solution that are infrequently used. Alternatively, they may introduce such attributes by partially or fully re-starting the solution process.

The same types of memories previously described are useful as a foundation for such procedures, although these memories are maintained over different (generally larger) subsets of solutions than those maintained by intensification strategies. A simple diversification



Tabu Search, Fig. 2 Simple TS intensification approach



Tabu Search, Fig. 3 Simple TS diversification approach

approach that keeps a frequency-based memory over all solutions previously generated, and that has proved very successful for machine scheduling problems, is shown in Fig. 3. Significant improvements over the application of short term TS memory have been achieved by this procedure.

Diversification strategies that create partial or full restarts are important for problems and neighborhood structures where a solution trajectory can become isolated from worthwhile new alternatives unless a radical change is introduced. Diversification strategies can also utilize a long-term form of recency-based memory, which results by increasing the tabu tenure of solution attributes.

The two special TS strategies called path relinking and strategic oscillation embody aspects of both intensification and diversification and have proved highly effective in a variety of contexts (Glover and Laguna 1993; Yagiura et al. 2006). The determination of effective ways to balance the concerns of intensification and diversification represents a promising research area. These concerns also lie at the heart of effective parallel processing implementations. The goal from the TS perspective is to design patterns of communication and information sharing across subsets of processors in order to achieve the best tradeoffs between intensification and diversification functions. General analyses and studies of parallel processing with tabu search are given in Taillard (1991, 1993), Battiti and Tecchiolli (1992), Chakrapani and Skorin-Kapov (1993), and Crainic et al. (1993a, 1993b).

Concluding Remarks

Complementarities among the perspectives of tabu search and those favored by the artificial intelligence and neural network communities raise the possibility of creating systems that integrate their fundamental

concerns. Examples are provided by the creation of tabu training and learning models (de Werra and Hertz 1989; Beyer and Ogier 1991; Battiti and Tecchioli 1993; Gee and Prager 1994) and tabu machines (Chakrapani and Skorin-Kapov 1993). The outcomes from this work have shown promising consequences for supplementing customary connectionist models – as by yielding levels of performance notably superior to that of models based on Boltzmann machines, and by yielding processes for modifying network linkages that give more reliable mappings of inputs to outputs.

The practical successes of tabu search have promoted useful research into ways to exploit its underlying ideas more fully. At the same time, many facets of these ideas remain to be explored. The issues of identifying best combinations of short- and long-term memory and best balances of intensification and diversification strategies still contain many unexamined corners (Glover 2007), and some of them undoubtedly harbor important discoveries for developing more powerful solution methods in the future.

Fundamental advances in applications of tabu search have been assembled in a collection of "Tabu Search Vignettes" accessible via the Internet at the author's Web site. These include summaries of key developments in a variety of areas, including:

Constraint Solving and Its Applications (Resource Assignment, Planning and Timetabling, Integer Programming Feasibility, Satisfiability, Mobile Network Frequency Assignment)
Chemical Industry Applications (Computer Aided Molecular Design (CAMD), Heat Exchanger Network (HEN) Synthesis, Phase Equilibrium Calculations, Gibbs Free Energy Minimization, Optimal Component Lumping Problems)
Classification
Feature Selection
Satellite Range Scheduling
Maritime Transportation for International Trade
Conservation Area Network Design
High Level Synthesis
Graph Coloring
Delivery
Routing with Loading and Inventory Constraints
Heterogeneous Routing and Scheduling
Capacitated Facility Location
Multi-period Forest Harvesting
Manpower Scheduling
DNA Sequencing
Airline Disruption Management
Internet Traffic Engineering
Matrix Bandwidth Minimization
Generalized Assignment
Constraint Satisfaction (Work Shift Scheduling, Set-Covering and Nurse Scheduling)
Resource-Constrained Project Scheduling
Dynamic Optimization (Trade Market Prediction, Meteorological Forecast, Robotics Motion Control)

See

- ▶ Artificial Intelligence
- ▶ Heuristics
- ▶ Metaheuristics
- ▶ Neural Networks

References

Battiti, R., & Tecchioli, G. (1993). *Training neural nets with the reactive tabu search*. Technical Report UTM 421, University of Trento, Italy, November.

Battiti, R., & Tecchiolli, G. (1992). Parallel biased search for combinatorial optimization: Genetic algorithms and TABU. *Microprocessors and Micro-Systems, 16*, 351–367.

Battiti, R., & Tecchiolli, G. (1994). The reactive tabu search. *ORSA Journal on Computing, 6*, 126–140.

Beyer, D., & Ogier, R. (1991). Tabu learning: A neural network search method for solving nonconvex optimization problems. *Proceedings of the International Joint Conference on Neural Networks*, IEEE and INNS, Singapore.

Chakrapani, J., & Skorin-Kapov, J. (1993). Connection machine implementation of a tabu search algorithm for the traveling salesman problem. *Journal of Computing and Information Technology (CIT), 1*(1), 29–36.

Crainic, T. G., Gendreau, M., Soriano, P., & Toulouse, M. (1993). A tabu search procedure for multi-commodity location/allocation with balancing requirements. *Annals of Operations Research, 41*(1–4), 359–383.

Crainic, T. G., Toulouse, M., & Gendreau, M. (1993a). *A study of synchronous parallelization strategies for tabu search*. Publication 934, Centre de recherche sur les transports, Université de Montréal.

Crainic, T. G., Toulouse, M., & Gendreau, M. (1993b). *Appraisal of asynchronous parallelization approaches for tabu search algorithms*. Publication 935, Centre de recherche sur les transports, Université de Montréal.

de Werra, D., & Hertz, A. (1989). Tabu search techniques: A tutorial and an applications to neural networks. *OR Spectrum, 11*, 131–141.

Gee, A. H., & Prager, R. W. (1994). Polyhedral combinatorics and neural networks. *Neural Computation, 6*, 161–180.

Gendreau, M., Soriano, P., & Salvail, L. (1993). Solving the maximum clique problem using a tabu search approach. *Annals of Operations Research, 41*, 385–404.
Glover, F. (1989). Tabu search-part I. *ORSA Journal on Computing, 1*, 190–206.
Glover, F. (1990). Tabu search-part II. *ORSA Journal on Computing, 2*, 4–32.
Glover, F. (1995). Tabu thresholding: Improved search by nonmonotonic. *ORSA Journal on Computing, 7*, 426–442.
Glover, F. (1999). Scatter search and path relinking. In D. Corne, M. Dorigo, & F. Glover (Eds.), *New ideas in optimization* (pp. 297–316). UK: McGraw Hill.
Glover, F. (2007). Tabu search – uncharted domains. *Annals of Operations Research, 149*(1), 89–98.
Glover, F., & Laguna, M. (1993). Tabu search. In C. Reeves (Ed.), *Modern heuristic techniques for combinatorial problems* (pp. 70–141). Oxford: Blackwell.
Glover, F., & Laguna, M. (1997). *Tabu search*. Norwell, MA: Kluwer.
Glover, F., Laguna, M., & Marti, R. (2000). Fundamentals of scatter search and path relinking. *Control and Cybernetics, 29*(3), 653–684.
Glover, F., Laguna, M., Taillard, E., & de Werra, D. (Eds.) (1993). *Tabu search*. Special issue of the Annals of Operations Research (Vol. 41). J.C. Baltzer.
Hansen, P., & Jaumard, B. (1990). Algorithms for the maximum satisfiability problem. *Computing, 44*, 279–303.
Hertz, A., & de Werra, D. (1991). The tabu search metaheuristic: How we used it. *Annals of Mathematics and Artificial Intelligence, 1*, 111–121.
Nowicki, E., & Smutnicki, C. (1993). *A fast taboo search algorithm for the job shop problem*. Report 8/93, Institute of Engineering Cybernetics, Technical University of Wroclaw.
Soriano, P., & Gendreau, M. (1993). *Diversification strategies in tabu search algorithms for the maximum clique problem*. Publication #940, Centre de Recherche sur les Transports, Université de Montréal.
Taillard, E. (1991). *Parallel tabu search technique for the job shop scheduling problem*. Research Report ORWP 91/10, Departement de Mathematiques, Ecole Polytechnique Federale de Lausanne.
Taillard, E. (1993). Parallel iterative search methods for vehicle routing problems. *Networks, 23*, 661–673.
Yagiura, M., Ibaraki, T., & Glover, F. (2006). A path relinking approach with ejection chains for the generalized assignment problem. *European Journal of Operational Research, 169*, 548–569.

Taguchi Loss Function

► Total Quality Management

Tail Distribution Function

For a random variable X, $\Pr\{X>x\}$. For a c.d.f. F, $F^c = 1 - F$, also known as the complementary CDF.

Tandem Queues

Queues in series.

See

► Networks of Queues

Technological Coefficients

The generic name given to the a_{ij} coefficients of the constraint set of a linear-programming problem.

Telecommunication Networks

► Communications Networks
► Queueing Theory

Terminal

A location used by a carrier for freight consolidation, break-bulk, interchange, and shipment and vehicle service.

See

► Logistics and Supply Chain Management

The Institute of Management Sciences (TIMS)

Founded in 1953, The Institute of Management Sciences (TIMS) was an international organization for management science professionals and academics. It was merged with the Operations Research Society of America (ORSA) into the Institute for Operations Research and the Management Sciences (INFORMS) effective January 1, 1995. The objectives of TIMS were

(1) to identify, extend and unify scientific knowledge contributing to the understanding and practice of management, (2) to promote the development of the management sciences and the free interchange of information about the practice of management among managers, scientists, scholars, students, and practitioners of the management sciences within private and public institutions, (3) to promote the dissemination of information on such topics to the general public, and (4) to encourage and develop educational programs in the management sciences. TIMS published the journal *Management Sciences* (in 40 volumes) and other publications (some jointly with ORSA). It held national meetings (jointly with ORSA), sponsored meetings by its technical colleges and geographic sections, and held international meetings in various countries.

See

- ▶ Institute for Operations Research and the Management Sciences (INFORMS)
- ▶ Operations Research Society of America (ORSA)

Theorem of Alternatives

Many such theorems exist, with a typical one being: either $\boldsymbol{Ax} = \boldsymbol{b}$ has a solution or $\boldsymbol{yA} = \boldsymbol{0}$, $\boldsymbol{yb} \neq \boldsymbol{0}$ has a solution. They can be shown to be equivalent to the strong duality theorem of linear programming.

See

- ▶ Farkas' Lemma
- ▶ Gordan's Theorem
- ▶ Strong Duality Theorem
- ▶ Transposition Theorems

Theory of Constraints

Graham K. Rand
Lancaster University, Lancaster, UK

In the early 1980s, a novel was published which has subsequently been read all over the world by many executives, production planners and shop floor workers. *The Goal* sets out Eli Goldratt's ideas on how production should be planned (Goldratt and Cox 2004). The ideas were developed in the production planning system OPT (Optimized Production Technology) which was marketed by Creative Technology, Inc. (Rand 1990). These ideas were later broadened to encompass other areas such as marketing, distribution and project management in two further novels, *It's Not Luck* (Goldratt 1994) and *Critical Chain* (Goldratt 1997), and the theory widened to become the Theory of Constraints. In the novel, *Necessary but Not Sufficient* (Goldratt et al. 2000), set in the computer software industry, it is argued that although new technology may be necessary for major improvements, it is not sufficient. The theory has been applied to retailing through two further books, first by means of a conversation between Goldratt and his daughter, *The Choice* (Goldratt 2008), and in the novel, *Isn't it Obvious?* (Goldratt et al. 2009). Among the methods in his approach, Evaporating Clouds and Current Reality Tree have become widely used. Technical details are found in Goldratt (1990a, b).

See

- ▶ Production Management

References

Goldratt, E. M. (1990a). *What is this thing called the Theory of Constraints?* Great Barrington, MA: North River Press.
Goldratt, E. M. (1990b). *The haystack syndrome*. Great Barrington, MA: North River Press.
Goldratt, E. M. (1994). *It's not luck*. Great Barrington, MA: North River Press.
Goldratt, E. M. (1997). *Critical chain*. Great Barrington, MA: North River Press.
Goldratt, E. M. (2008). *The choice*. Great Barrington, MA: North River Press.
Goldratt, E.M., & Cox, J. (2004). *The goal* (3rd Rev. Ed.). Great Barrington, MA: North River Press.
Goldratt, E. M., Eshkoli, I., & Brownleer, J. (2009). *Isn't it obvious?* Great Barrington, MA: North River Press.
Goldratt, E. M., Schragenheim, E., & Ptak, C. A. (2000). *Necessary but not sufficient*. Great Barrington, MA: North River Press.
Rand, G. K. (1990). RP, JIT and OPT. In L. C. Hendry & R. W. Eglese (Eds.), *Operational research tutorial papers, 1990*. Birmingham: Operational Research Society.

Thickness

The minimum number of edge-disjoint planar subgraphs into which a graph can be decomposed.

See

▶ Graph Theory

Time Series Analysis

Christina M. Mastrangelo[1], James R. Simpson[2] and Douglas C. Montgomery[3]
[1]University of Virginia, Charlottesville, VA, USA
[2]Florida State University, Tallahassee, FL, USA
[3]Arizona State University, Tempe, AZ, USA

Introduction

A time series is an ordered sequence of observations. This ordering is usually through time, although other dimensions, such as spatial ordering, are sometimes encountered. A time series can be continuous, as when an electrical signal such as voltage is recorded. Typically, however, most industrial time series are observed and recorded at specific time intervals and are said to be discrete time series. If only one variable is observed, the time series is said to be univariate. However, some time series involve simultaneous observations on several variables. These are called multivariate time series.

There are three general objectives for studying time series: 1) understanding and modeling of the underlying mechanism that generates the time series, 2) prediction of future values, and 3) control of some system for which the time series is a performance measure. Examples of the third application occur frequently in industry. Almost all time series exhibit some structural dependency. That is, the successive observations are correlated over time, or autocorrelated. Special classes of statistical methods that take this autocorrelative structure into account are required.

Figure 1 shows examples of time series with distinctly different features. In Fig. 1a, the time series x_t appears to vary around a constant level. Such a time series is said to be stationary in the mean. In Fig. 1b, non-stationary behavior can be observed, i.e., the time series x_t drifts with no obvious fixed level. Some nonstationary time series may exhibit trends, or the variance of the series may increase as the level of the time series increases. Seasonal variation is illustrated in Fig. 1c.

The autocorrelation function is a very useful tool in characterizing time series behavior. The autocorrelation between x_t and x_{t+k} is defined as

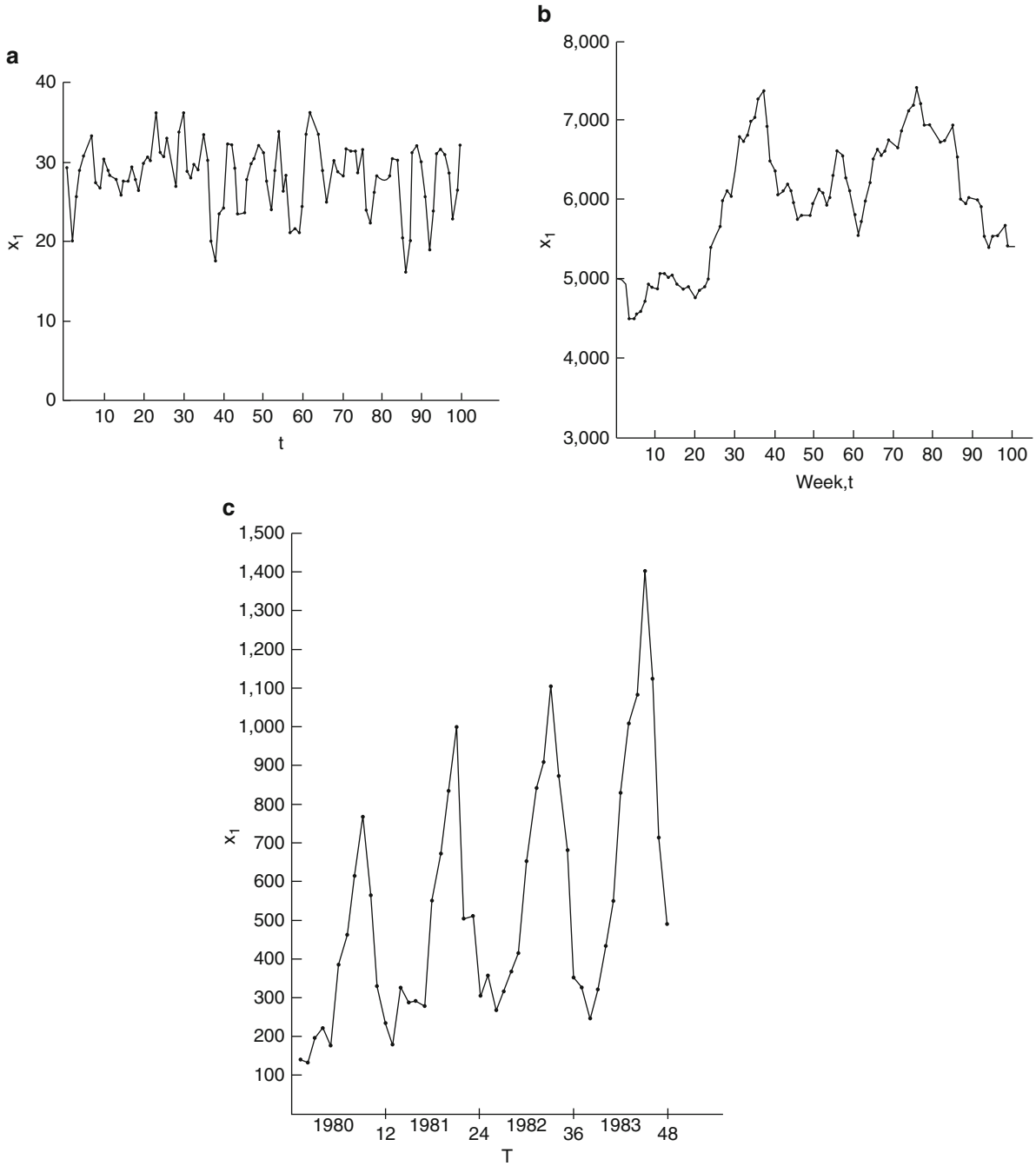
$$\rho_k = \frac{\mathrm{cov}(x_t, x_{t+k})}{\sqrt{V(x_t)V(x_{t+k})}} = \frac{\gamma_k}{\gamma_0}$$

where $\mathrm{cov}(x_t, x_{t+k}) = E[(x_t - m)(x_{t+k} - m)]$. This is called the autocorrelation at lag k. The usual estimate of ρ_k, $k = 1, 2, \ldots, K$, is the sample autocorrelation function

$$r_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} = \frac{\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}$$

Figure 2 shows the sample autocorrelation function for the time series in Fig. 1a. The dotted lines are two standard error limits. Notice that there is a large positive value or spike at lag 1 and the sample autocorrelation function decays as a damped sine wave from lag 1. The sample autocorrelation function is very useful in the identification of an appropriate time series model.

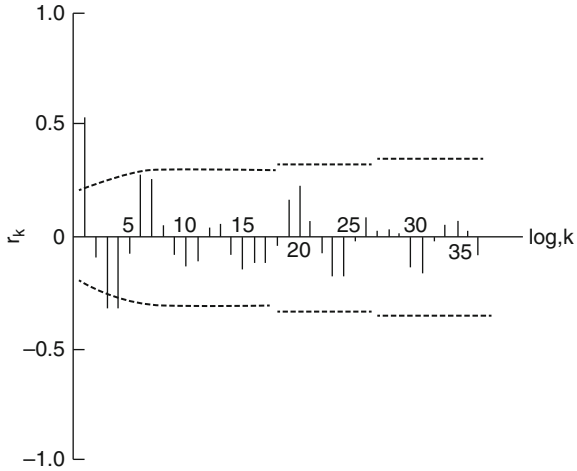
The partial autocorrelation function, denoted by ϕ_{kk}, is also useful in the identification process. It can be interpreted as the simple correlation between two random variables x_t and x_{t-k} after adjusting for the intermediate variables x_{t-1}, $x_{t-2}, \ldots, x_{t-k+1}$. Once the sample autocorrelation and partial autocorrelation functions are estimated, they may be plotted. A tentative model is then identified by comparing the observed patterns with the theoretical function patterns. For an autoregressive process of order p, ϕ is nonzero when k is less than or equal to p and greater than zero for k greater than p. In other words, while



Time Series Analysis, Fig. 1 (**a**) Viscosity of a chemical product. (**b**) Demand for a plastic container. (**c**) Monthly demand for a 48-oz soft drink in hundreds of cases

the autocorrelation function of an autoregressive process decays in an exponential fashion, the partial autocorrelation function cuts off after lag p.

In addition, the inverse autocorrelation function and the extended sample autocorrelation function are useful in time series model identification. See Fuller (1996), Montgomery, Johnson, and Gardiner (1990), Cleveland (1972), and Abraham and Ledolter (1983) for definitions of these functions and more details.



Time Series Analysis, Fig. 2 Sample autocorrelation function

Time Series Modeling Methods

There are several widely-used approaches for modeling and analysis of time series data. Regression methods play a fundamental role. If y_t represents the time series of interest and $x_{jt}, j = 1, 2, \ldots, k$ are a collection of other time series thought to be related to y_t, then it is possible to fit a regression model of the form

$$y_t = \beta_0 + \sum_{t=1}^{n} \beta_j x_{jt} + \varepsilon_t, \qquad t = 1, \ldots, n$$

using least squares or some suitable variation. Usually, however, the errors e_t are autocorrelated and more complex estimation schemes are needed. Several estimation methods are available which result in estimates similar to least squares estimates, but the standard errors may be very different. Yule-Walker estimation uses the Yule-Walker equations to estimate the autoregressive parameters of the errors and generalized least squares to estimate $\boldsymbol{\beta}$. Harvey (1990) gives a full description of this and other methods.

Smoothing methods are frequently used in time series analysis. In particular, exponential smoothing is widely used for producing short-term forecasts of many types of industrial time series. Much of the original work in this area is by Brown (1962), Holt (1957), and Winters (1960). Exponential smoothing is often developed heuristically starting with a simple model such as $x_t = b + e_t$, where e_t are independent random variables and b is an unknown constant. Simple or first-order exponential smoothing is defined as

$$S_t = \alpha x_t + (1 - \alpha) S_{t-1}$$

where $0 \le \alpha \le 1$. The smoothed statistic S_t estimates the constant b, so the forecast for any future observation $X_{t+\tau}$ made at the end of period t is

$$\hat{x}_{t+\tau}(t) = S_t$$

Extensions of this methodology to forecasting linear and quadratic trend and incorporating seasonal behavior are described in Montgomery, Johnson and Gardiner (1990). Goodman (1974) and Cogger (1974) showed that exponential smoothing for a kth order polynomial results in forecasts that are optimal in a mean square error sense for certain classes of non-stationary time series. McKenzie (1978) extended these results to models that may include transcendental terms.

The class of autoregressive integrated moving averages (ARIMA) models proposed by Box, Jenkins and Reinsel (2008) and Jenkins (1979) have been very successful for time series modeling and forecasting. The general form for this family of models is

$$\begin{aligned}(1 - \phi_1\beta - \phi_2 B^2 - \cdots - \phi_p B^p)(1 - B)^d x_t \\ = \theta_0 + (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)\varepsilon_t\end{aligned}$$

where ϕ_i are the autoregressive parameters, θ_j are the moving average parameters, B is a backshift operator defined such that $B^r x_t = x_{t-r}$,$(1 - B)^d = \Delta^d$ is the backward difference operator, and $\in_t$ is an uncorrelated sequence of random disturbances with mean zero and variance σ^2. This model can also be extended to incorporate seasonal behavior (see Box et al. 2008; Montgomery et al. 1990). One chooses a model by specifying the integers p, d, and q, resulting in an ARIMA(p, d, q) model. This is usually done by examining the sample autocorrelation and partial autocorrelation function. For example, if the sample autocorrelation function decays as a damped sine wave and the partial autocorrelation function has large spikes only at lags 1 and 2, a tentative ARIMA model estimation with $p = 2$ and $q = 0$ might be considered.

Nonlinear regression methods are used to estimate the parameters ϕ_i and θ_j. The approach requires initial point estimates of the parameters and then uses an iterative search technique to minimize the residual sums of squares. Most computer packages implement a modification of the Gauss-Newton method suggested by Marquardt (1963). The Gauss-Newton method first linearizes the nonlinear function with a Taylor series expansion and then iterates to find improved parameter estimates. Unfortunately, the original Gauss-Newton approach will not always converge. So Marquardt proposed a modified search procedure that adds a small bias to the parameter estimates to ensure convergence to the minimum residual sums of squares. Computer packages provide reasonable initial point estimates making the estimation routine transparent to the user.

Finally, the residuals from the fitted model are studied to test model adequacy. Generally, one should examine the autocorrelation function of the residuals, for if the model is adequate, the residuals should be approximately uncorrelated. The tests on residual autocorrelations suggested by Box and Pierce (1970) and Ljung and Box (1978) are useful in this regard. Residual plots, such as a plot of residuals verses the fitted x_t, and a normal probability plot of the residuals, are useful in detecting model inadequacy. Thus model estimation is typically iterative involving cycles of tentative model identification, estimation, and residual analysis.

To illustrate, consider the container demand data from Fig. 1b. It can be shown that an appropriate choice of p, d, and q is $p = 0$, $d = 1$, and $q = 1$, resulting in the ARIMA(0,1,1) = IMA(1,1) model

$$(1 - B)x_t = (1 - \theta B)\varepsilon_t.$$

The least squares estimate of the parameter θ in this model is $\bar{\theta} = -0.70$. Therefore, the final model is

$$x_t = x_{t-1} + \varepsilon_t + 0.7\varepsilon_{t-1}.$$

This model is satisfactory with respect to the adequacy criteria cited above.

Forecasting

An important objective of any time series model is forecasting future values. The term forecasting is used in the time series analysis literature although most results are based on the general theory of linear prediction developed by Kalman (1960), Whittle (1963), Box, Jenkins and Reinsel (2008), and many others. The objective is to produce minimum mean square error forecasts.

Minimum mean square error forecasts for ARIMA models are obtained by taking the conditional expectation $E(X_{t+\tau}|X_t,\ X_{t-1},\ldots)$. For example, the minimum mean square error forecast for the ARIMA (0,1,1) = IMA(1,1) model shown earlier for the container data is

$$E(x_{t+\tau}|x_t, x_{t-1}, \ldots) \equiv \hat{x}_{t+\tau}(t) = x_t + 0.7\varepsilon_t \qquad (1)$$

where $e_t(1) = x_t - \hat{x}_t(t-1)$ is the one-step ahead forecast error. Figure 3 shows the forecasts obtained from this model. It is usually necessary to provide prediction intervals for forecasts as well as point estimates. Figure 3 shows the 50% and 95% prediction limits for the forecast of future container demand. For details of the construction of these limits, see Box, Jenkins and Reinsel (2008) and Montgomery, Johnson and Gardiner (1990).

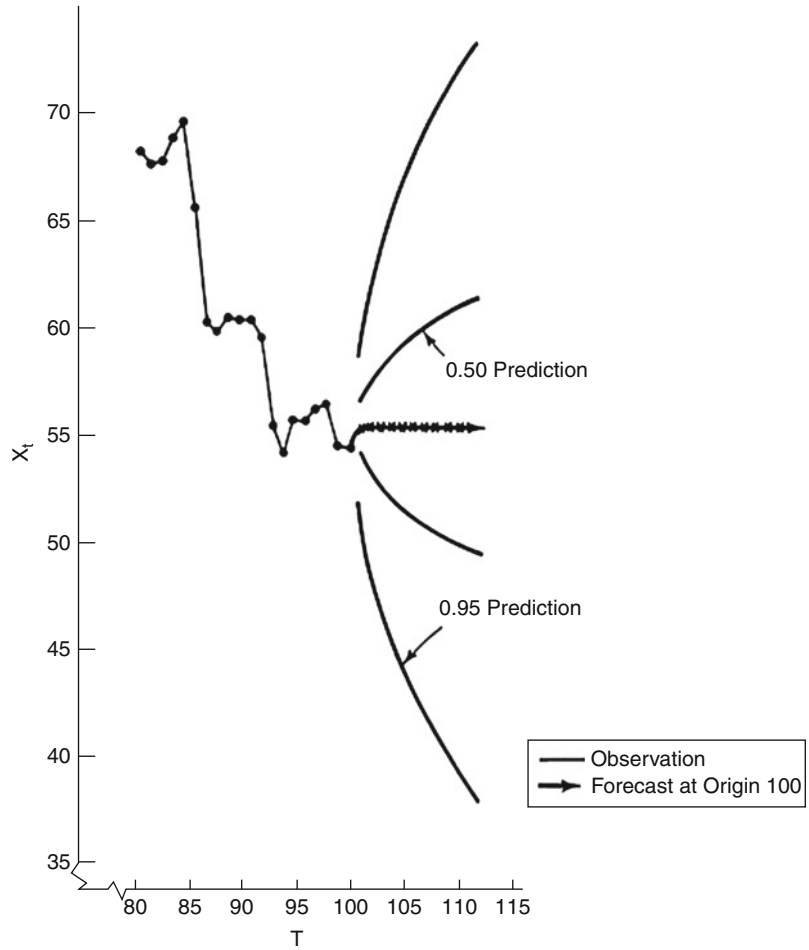
Forecasts from ARIMA models are equivalent to forecasts produced by other methods in certain cases. For example, the forecasts from an IMA(1,1) model, such as that given above for the container demand data, are identical to those produced by simple first-order exponential smoothing. Other relationships between exponential smoothing and ARIMA models are given by Box, Jenkins and Reinsel (2008) and Pandit and Wu (1974).

The form of the eventual forecast function for ARIMA models is also of interest, because it leads in some cases to efficient methods for forecast generation and updating. The form of the forecast function or several common ARIMA models is given in Box, Jenkins and Reinsel (2008).

Transfer Functions and Related Topics

If y_t and x_t are two stationary time series related through the mean filter

$$y_t = V(B)x_t + \varepsilon_t$$



Time Series Analysis, Fig. 3 Forecast of plastic container demand at origin 100, with 0.50 and 0.95 percent prediction limits

then $V(B) = \sum_{j=-\infty}^{\infty} v_j B^j$ is called the transfer function of the filter and e_t is called the noise series of the system. Typically, x_t and e_t are assumed to follow ARMA = ARIMA(p,0,q) models. It is customary to write

$$V(B) = \frac{\omega_S(B)B^b}{\delta_r(B)}$$

where $\omega_S(B) - \omega_0 - \omega_1 B - \omega_2 B^2 - \ldots - \omega_S B^S$, $\delta_r(B) - \delta_0 - \delta_1 B - \delta_2 B^2 - \ldots - \delta_r B^r$ and b is a delay representing the time before the input at time t produced an effect on the output. A transfer function model is identified by choosing appropriate values of s, r, b, and a model for the noise e_t. Usually s, r, and b will be no larger than 2. The cross-correlation function is useful in model identification.

Once a suitable transfer function model is identified, the parameters are estimated by nonlinear regression methods, and diagnostics checks are applied, much like in classical univariate ARIMA modeling. Minimum mean square error forecasts are generated using a similar approach, based on conditional expectation at time t of $y_{t\tau}$. For detailed examples of identification, estimation, diagnostic checking, and forecasting with transfer functions, see Box, Jenkins and Reinsel (2008) and Montgomery, Johnson and Gardiner (1990). The latter authors presented an example showing that for relatively short forecast lead times, the forecasts from a transfer function model will usually be superior to those produced by a univariate ARIMA model.

An important special case of the transfer function occurs when the input series x_t is a sequence of indicator variables that represent the occurrence

of identifiable, unique events that are thought to influence the output y_t. These events are called interventions and the resulting models are called intervention models. An intervention model is often used to provide a statistical basis for concluding that the identifiable event has resulted in a change in the time series.

Box and Tiao (1975) developed the basic intervention analysis methodology and applied it to photo chemical pollution data from the Los Angeles basin. They showed that the opening of the Golden State Freeway and the adoption of a new law, that reduced the proportion of reactive hydrocarbons in local gasoline, reduced ozone levels, and that required changes in automobile engines reduced ozone levels only in warm weather months. Other intervention studies were reported by Montgomery and Weatherby (1980) and Wichern and Jones (1977).

Intervention models are also useful in the study of time series outliers. Fox (1972) proposed two types of outliers, additive and innovational. Other useful references on this topic are Tsay (1986) and Chang, Tiao and Chen (1988).

In some time series problems, one observes m different variables x_{1t}, x_{2t},..., x_{mt} in a multivariate framework. One way to model this structure is with a multivariate ARIMA model of the form

$$\Phi_p(B)X_t = \Theta_q(B)\varepsilon_t$$

where $x_r' = [x_{1t}, x_{2t}, \ldots, x_{mt}]$, Φ_p (B) and Θ_q (B) are matrix polynomials of autoregressive and moving average parameters, respectively, and ε_t is a sequence of independent multivariate random vectors each with mean zero and covariance matrix $\sum$. These are sometimes called vector time series models. Basic references for these models include Jenkins (1979), Granger and Newbold (1977), and Hannan (1970). The state space modeling approach is also useful for representing multiple series. See Hannan (1970) and Akaike (1976) for a complete description of state space modeling.

Computing

A number of software packages perform the time series modeling and forecasting functions previously described, including some spreadsheet statistical analysis add-ins. The two high-end software support tools commonly used by researchers and practitioners are SAS and S-Plus. Both programs provide a wide range of modeling options including various smoothing alternatives and extensive ARIMA modeling features. SAS is also capable of developing transfer function and intervention models. S-Plus provides the capability to model time series in the presence of outliers. More advanced procedures are also available from SAS and S-Plus. Several other PC-based software programs, including MINITAB, STATGRAPHICS, R, JMP, Autobox, and EViews, provide high-quality time series modeling and forecasting support. For ARIMA modeling, the software programs provide the plots, nonlinear estimation, and forecasting tools necessary to develop successful models.

See

- ▶ Exponential Smoothing
- ▶ Forecasting
- ▶ Quality Control
- ▶ Regression Analysis

References

Abraham, B., & Ledolter, J. (1983). *Statistical methods for forecasting*. New York: John Wiley.

Akaike, H. (1976). Canonical correlations analysis of time series and the use of an information criterion. In R. Mehra & D. G. Lainiotis (Eds.), *Advances and case studies in system identification*. New York: Academic Press.

Box, G. E. P., & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of American Statistical Association, 64*.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). *Time series analysis, forecasting and control* (4th ed.). New York: Wiley.

Box, G. E. P., & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association, 70*, 70–79.

Brown, R. G. (1962). *Smoothing, forecasting and prediction of discrete time series*. Englewood Cliffs, NJ: Prentice-Hall.

Chang, I., Tiao, G. C., & Chen, C. (1988). Estimations of time series parameters in the presence of outliers. *Technometrics, 30*, 193–204.

Cleveland, W. S. (1972). The inverse autocorrelations of a time series and their applications. *Technometrics, 14*, 277–293.

Cogger, K. O. (1974). The optimality of general-order exponential smoothing. *Operations Research, 22*, 858–867.

Fox, A. J. (1972). Outliers in time series. *Journal of the Royal Statistical Society, Series B, 43*, 350–363.
Fuller, W. A. (1996). *Introduction to statistical time series*. New York: John Wiley.
Goodman, J. L. (1974). A new look at higher-order exponential smoothing for forecasting. *Operations Research, 22*, 880–888.
Granger, G. W. C., & Newbold, P. (1977). *Forecasting economic time series*. New York: Academic Press.
Hanan, E. J. (1970). *Multiple time series*. New York: John Wiley.
Harvey, A. C. (1990). *The econometric analysis of time series* (2nd ed.). Cambridge, MA: MIT Press.
Holt, C. C. (1957). *Forecasting trends and seasonal by exponentially weighted moving averages*. ONR Memorandum No. 52, Carnegie Institute of Technology.
Jenkins, G. M. (1979). *Practical experiences with modeling and forecasting time series*. Lancaster, England: GJM Publications.
Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *ASME Journal of Basic Engineering for Industry, Series D, 82*, 35–45.
Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika, 65*, 297–303.
Marquardt, D. W. (1963). An algorithm for least squares estimation of nonlinear parameters. *Journal of the Society of Industrial and Applied Mathematics, 2*, 431–441.
McKenzie, E. (1978). The monitoring of exponentially weighted forecasts. *Journalof the Operational Research Society, 29*.
Montgomery, D. C., Johnson, L. A., & Gardiner, J. S. (1990). *Forecasting and time series analysis* (2nd ed.). New York: McGraw-Hill.
Montgomery, D. C., & Weatherby, G. (1980). Modeling and forecasting time series using transfer function and intervention methods. *AIIE Transactions, 12*, 289–307.
Pandit, S. M., & Wu, S. M. (1974). Exponential smoothing as a special case of a linear stochastic system. *Operations Research, 22*, 868–879.
Tsay, R. S. (1986). Nonlinearity tests for time series. *Biometrika, 73*, 461–466.
Whittle, P. (1963). *Prediction and regulation by linear least-square methods*. Princeton, NJ: Van Nostrand.
Wichern, D. W., & Jones, R. H. (1977). Assessing the input of market disturbances using intervention analysis. *Management Science, 21*, 329–337.
Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Operations Research, 22*, 858–867.

Time/Cost Trade-offs

An approach to scheduling where the project duration is shortened with a minimum of added costs.

See

► Network Planning

Time-stepped Simulation

A computer model in which time is incremented by a simulated clock. Each appropriate function is recomputed after the clock is incremented in a cyclic manner. A model may be linearly coded and entirely time-stepped or an event-driven simulation may use time-stepping for some critical function with a cycle of sub-functions.

See

► Event-driven Simulation
► Simulation of Stochastic Discrete-Event Systems

Timetabling

Michael W. Carter
University of Toronto, Toronto, Ontario, Canada

Introduction

Most dictionaries do not include the word timetabling as a single word. It is often listed as either two words (time table) or hyphenated (as time-table). The Oxford English Dictionary defines a timetable as:

> A tabular list or schedule of times at which successive things are to be done or happen, or of the times occupied in the parts of some process. spec. **a.** A printed table or book of tables showing the times of arrival and departure of railway trains at and from the stations; also a similar table of times of arrival and departure of passenger boats or other public conveyances. **b.** A chart used in railway traffic offices, showing by means of cross lines, in one direction representing hours and minutes and in the other miles, the position of the various trains at any given moment. **c.** A time-sheet on which a record is kept of the time worked by each employee. **d.** A table showing how the schedule of a school or other educational institution, for any day, or for a week, is allotted to the various classes and subjects. **e.** Mus. A table of notes showing their relative time value.

The Oxford dictionary also defines the verb time-table as "To schedule, to plan or arrange according to a timetable, to include in a timetable. Hence time-tabled and timetabling."

Professor Anthony Wren, at the first Practice and Theory of Automated Timetabling (PATAT) conference in Edinburgh, 1995, defined timetabling as "the allocation, subject to constraints, of resources to objects being placed in space-time, in such a way as to satisfy as nearly as possible a set of desirable objectives. Examples are class and examination time-tabling and some forms of personnel allocation, for example manning of toll booths subject to a given number of personnel." In the latter case, the process is defined in terms of developing timetables for each individual employee.

In other words: timetabling involves deciding when events/activities will take place in time; but it does not involve assigning resources to those activities. For example, a bus timetable for a particular metropolitan bus route may require "one bus to leave the main terminal every 30 minutes between 6:00 am and 11:00 p.m.; and every 10 minutes during rush hours 7:00 am to 9:00 am and 4:00 p.m. to 6:00 p.m.". The time table does not specify which buses or drivers should be allocated to each trip. In course timetabling, the objective is to decide what day and time each section of each course should be held. It does not specify which students will be assigned to each section.

Normally, when one sees the word timetabling in an operations research context, people are referring to problems relating to timetabling of courses or examinations in a school. Furthermore, it refers to the concept of developing algorithms, usually computer programs, for the automatic construction of time-tables. There are a number of other related problems in timetabling which will be described; but they are often referred to under different titles. As described by McCollum and Burke (2010) in the Preface to the Proceedings for PATAT 2010, "computer-aided timetable generation . . . includes personnel rostering, school timetabling, sports scheduling, transportation timetabling and university timetabling."

Timetabling can also be described as a subset of the larger discipline called scheduling. One can define scheduling as the more general problem of determining the times for activities and assigning the necessary resources. In some cases, for example in Sports Time tabling, once it is decided when a match will occur between a pair of teams, (and who the home team is), all major resources have already implicitly been assigned (the two teams and the stadium). Hence Sports Time tabling is commonly referred to as Sports Scheduling. In this case, the terms are justifiably interchangeable.

It will be frequently distinguished between feasibility and optimality. A feasible solution is any solution that satisfies all of the constraints. An optimal solution is the (possibly unique) solution among all feasible answers which maximizes (minimizes) some objective function. In some timetabling problems, it is sufficient to find a feasible solution.

Examination Timetabling

Examination Timetabling is the simplest timetabling problem to describe, although it is not always easy to solve. The basic problem is to assign examinations to a limited number of available periods in such a way that there are no conflicts or clashes. That is, no student is required to write two examinations at the same time. The problem is closely related to the graph coloring problem. Each examination is represented by a node. Two nodes are connected by an edge if there is at least one student who is required to write the two corresponding exams. The graph coloring problem asks the question: Can the nodes of this graph be colored using p colors such that no two nodes with the same color are connected by an edge? If each color represents an examination period, and if p is the number of periods available, then coloring the graph is equivalent to finding a conflict free assignment of exams to the available periods.

In practice, the basic feasibility issue may be the critical problem. In particular, for any given problem instance, there is a minimum number of periods required to allow a feasible solution. In graph theory terminology, this is called the chromatic number of a graph. If the number of periods provided is close to the theoretical minimum, then you need an algorithm that concentrates on finding a feasible solution. There has been considerable research on good coloring algorithms. Given plenty of periods, it is easy to find a conflict free timetable. The coloring problem is trivial, and efforts can be focused on searching for a good answer using some secondary objectives. Without enough periods, it is not possible to find a feasible solution, and the objective must be changed to something like minimize the number of student conflicts.

The most common secondary objective is to try to spread each student's exams as evenly as possible. Each institution will impose a variety of additional constraints on the basic model such as:

- Some exams may have precedence constraints (e.g., "exam A must precede exam B");
- Some exams must be consecutive (e.g., "exam C must immediately precede exam D");
- Some exams are excluded from certain periods;
- Limited available rooms and/or seats; and
- There may be special resource requirements.

For a more comprehensive description of the exam timetabling problem and a survey of practical approaches, refer to Carter (1986), and Qu et al. (2009).

School Timetabling

Class-Teacher timetabling is normally associated with high schools or elementary level schools where the students are grouped into a set of classes and each class has a set of courses that it must take. Professor Dominique de Werra (1985) defines the basic class-teacher model in the following terms. Let $C = \{c_1, c_2, \ldots, c_m\}$ be a set of classes and $T = \{t_1, t_2, \ldots, t_n\}$ be a set of teachers. An $n \times m$ requirements matrix, $R = \{r_{ij}\}$ is given where r_{ij} is the required number of times class c_i must meet with teacher t_j. In the basic model, it is assumed that all lectures are the same length (say one period). Given a set of p periods, the problem is to assign each meeting to some period such that no teacher (and no class) is involved in more than one meeting at a time. The basic problem has no objective function, so the issue is simply to find a feasible solution.

It can be shown that this problem is easy to solve (in the computational complexity sense) in that there exists a polynomial algorithm to find a solution (using a matching algorithm) under the simple and obvious conditions that no teacher (or class) is required to attend more than p periods. The problem remains easy if the basic model is extended to include assigning meetings over a week, where limits are imposed on the number of times each class-teacher pair can meet on any one day.

Unfortunately, most practical problems will have a few extra conditions, and the problem quickly becomes computationally intractable (NP-Complete). For example, if it is assumed that some of the teachers (and/or classes) are not available in every period, then the problem is no longer easy. This is also true if the teachers and classes are available every period, but some of the meetings have been preassigned to specific periods. Another common complication is that some meetings are for more than one period. For example, some meetings may require two or three consecutive periods.

The problem is also often complicated by adding room availability constraints. For example, there may be certain meetings (science, physical education, music, etc.) which require specific rooms. This problem can be expressed using a three dimensional requirements matrix that specifies the number of meetings between class i and teacher j in a room of type k, where there are a limited number of each type of room. This problem is also NP-Complete. Refer to Kingston (2008) for more details.

Course Timetabling

Course timetabling is normally associated with universities, and involves the assignment of sections of courses (lectures, laboratories, tutorials, seminars, etc.) to specific days of the week and times of day. In the course-timetabling problem, unlike the class concept, each student selects a set of courses personally tailored to their own needs. (In practice, many students will have very similar selection patterns.) The primary objective is often to find a timetable that minimizes the expected number of student conflicts.

Strictly speaking, based on the definition given here, course timetabling does not include the assignment of resources (teachers, rooms, special equipment, or even students). In many practical instances, most teachers will be assigned to teach specific course sections before timetabling, while rooms, special equipment and students are assigned after time-tabling. In large schools, many of the courses will be offered in more than one section. Students must be divided up into (roughly equal) groups and assigned to separate sections. This problem is referred to as sectioning or student scheduling. Some packages have been designed to attack all of these problems simultaneously. However, due to the large number of variables involved, most practical methods approach the

problems sequentially. The basic course-timetabling problem will be described here. The interested reader can refer to Lewis (2008) for a more detailed discussion of each of the subproblems, and references to practical applications.

The basic course-timetabling problem usually includes a number of side constraints. Courses and course sections should be spread in a particular way throughout the week. For example, an institution may require that all sections of the same course be timetabled at the same times. A course may be divided into multiple meetings (two or three times per week), and there may be restrictions on the meeting patterns that can be used (e.g., Mon., Wed., Fri. at 9:00 a.m.). Some schedules include an allowance for lunch periods, travel time between classes, and the number of hours per day for students and teachers.

In practice, there are two main variations of the course-timetabling problem: the master timetable approach, and the demand driven system. Practitioners typically feel very strongly about their preference for one or the other. Under a master-time-tabling system, the institution will first create a course timetable, and then students register for courses (after consulting a list that describes when each class is offered). The term master timetable refers to the common practice of starting this year's timetable based on the previous year, and making any required changes based on revisions to course offerings. With a demand driven timetable, the institution posts a list of (proposed) course offerings without any times, and students pre-register for courses before timetabling is performed.

The main advantage of a demand driven system is that the timetable can be constructed using actual student course requests. With a master-timetable system, the timetable must be developed without knowing what the students really want or need. Individual department timetable representatives try to build a timetable that will work for students in their own program in each year. This is very difficult unless the programs are highly structured. In more flexible environments, students often have difficulty selecting the credits that they need without conflicts. A major problem in the U.S. today is that students in many institutions find it impossible to complete their program in the nominal program length due to timetable issues.

There are several disadvantages of a demand-driven system. It requires additional data collection effort, since students must pre-register for courses (typically 4–5 months before term starts) and then, when they get the results of their requests, they start making changes in a second round. In a master-timetabling system, students should be able to construct a conflict free timetable on the first attempt. A demand-driven system also puts fairly tight time constraints on the timetabling process. In a master timetable system, the institution can construct the timetable a year in advance, and some schools publish the times in the course calendar. In a demand-driven system, the students submit course requests a few months before the term starts, and all of the timetabling activity is compressed.

One of the curious issues in the timetabling problem creates a bit of a paradox in the demand-driven system when courses are taught in multiple sections. You cannot assign students to sections (conflict-free) until you have timetabled the sections; but, you cannot timetable the sections until you know which students are in each section. One solution is to assign students to a specific section in advance of timetabling, for the purpose of finding good times. These assignments can be re-evaluated in the student scheduling phase at the end.

Anyone interested in timetabling should refer to the Web site maintained by the University of Nottingham, on automated scheduling, optimisation and planning.

There are a number of other (less common) problems that share the basic timetabling structure. Sports timetabling is the problem of trying to find a rotation for a set of teams such that each team can play every other team twice (once at home and once away). If there are no side constraints, there are some elegant solutions related to tournaments, including a mathematical construction based on permutations (see survey by Kendall et al. 2010). There has also been some research on Employee Timetabling/Rostering, where you want to determine shift work patterns for employees in order to meet a given demand pattern. A particular well-studied variation on this problem is the nurse-rostering problem (see review by Burke et al. 2004).

See

► Computational Complexity
► Graph Theory

- ▶ Higher Education
- ▶ Scheduling and Sequencing
- ▶ Sports

References

Burke, E. K., De Causmaecker, P., Vanden Berghe, G., & Van Landeghem, H. (2004). The state of the art of nurse rostering. *Journal of Scheduling, 7*, 441–499.

Carter, M. W. (1986). A survey of practical applications of examination timetabling algorithms. *Operations Research, 34*, 193–202.

de Werra, D. (1985). An introduction to timetabling. European *Journal of Operational Research, 19*, 151–162.

Kendall, G., Knust, S., Ribeiro, C. C., & Urrutia, S. (2010). Scheduling in sports: An annotated bibliography. *Computers and Operations Research, 37*, 1–19.

Kingston, J. H. (2008). Resource assignment in high school timetabling. *Proceedings of the Seventh International Conference on the Practice and Theory of Automated Timetabling*, August 2008.

Lewis, R. (2008). A survey of metaheuristic-based techniques for university timetabling problems. *OR Spectrum, 30*, 167–190.

McCollum, B., & Burke, E. (2010). Preface. *Proceedings of the 8th International Conference on the Practice and Theory of Automated Timetabling*, 10–13 August 2010 (Queen's University of Belfast).

Oxford English Dictionary, On-line edition as of March 2011.

Qu, R., Burke, E. K., McCollum, B., Merlot, L. T. G., & Lee, S. Y. (2009). A survey of search methodologies and automated system development for examination timetabling. *Journal of Scheduling, 12*, 55–89.

Wren, A. (1996). Scheduling, timetabling and rostering – A special relationship? In Burke & Ross (Eds.), *Practice and theory of automated timetabling: Vol. 1153. Lecture notes in computer science*. Springer.

TIMS

- ▶ The Institute of Management Sciences (TIMS)

Tolerance Analysis

A sensitivity analysis procedure applied to a linear-programming problem that allows for simultaneous changes of the objective function cost coefficients and/or right-hand-sides of the constraints.

See

- ▶ Hundred Percent Rule
- ▶ Sensitivity Analysis

Total Float

The amount of time a project work time can be delayed without affecting the duration of the project. Total float can be used in only one activity in a path. If no schedule times are specified for starting and finishing the various activities, then the float is calculated as the difference between the latest start time and the earliest start time, or the difference between the latest finish time and the earliest finish time. Float can be positive, negative or zero.

See

- ▶ Network Planning

Total Quality Management

John S. Ramberg
Pagosa Springs, CO, USA

Introduction

During the decade of the 1980s, U.S. corporations recognized the quality achievements of their Japanese counterparts and began to understand the messages being delivered by Deming, Juran and others on the importance of quality (Deming 2000; Defeo and Juran 2010). They devised methods for obtaining, understanding and communicating customer needs and requirements within their organizations, developed strategies for improving their engineering design, development, manufacturing and delivery processes, and created new corporate cultures that included the formation of self-directed working groups and encouragement of employee participation. Through this focus on quality and the development and adaptation of techniques for achieving customer

satisfaction, some of these corporations have demonstrated improvement in achieving high quality, timely deliveries at low costs and ultimately improved their business performance. Many of these firms called this new management and operations philosophy Total Quality Management or TQM.

At the outset, many TQM programs were simply copies of Japanese efforts. As cultural differences between the Japanese and western world were better understood, and as other quality contributions were recognized, many U.S. firms developed their own unique quality programs. See Prybutok and Zhang (2010) and Vol. 4 of *Quality Management Journal* for health care agency examples.

Other firms, frustrated by false starts and questionable implementations, began to question the value of total quality management, and some have given up, regarding it as just another fad (Senge 1993). In many of these latter situations, quality efforts have been misdirected or unfocused. In some cases, quality improvement activities were simply knee-jerk reactions to the customers who complained most vehemently to the highest level of the organization. Ramberg (1994) described some of the scurrilous characters who proclaim TQM, while delivering just another program; he raised the question, "TQM: Thought Revolution or Trojan Horse." Three decades later, many organizations, especially nonprofits and governmental, are still not aware of total quality management.

While TQM connotes much more than simply the three words total, quality and management, nevertheless, definitions of each of the three words seem an appropriate place to begin. A typical dictionary definition of total is: all or whole, that is constituting the whole; complete. The definition of quality is a bit more difficult to comprehend as U.S. firms have come to understand. A formal definition, as given by the American Society for Quality (ASQ). "The totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs." Finally, management is the act, be it a science, an art or manner, of planning, directing, organizing and controlling a firm's decisions and actions. As an aside, it is interesting to note that the phrase "to manage" originated as "to train (a horse) in his paces, or to cause to do the exercises of the manage (Merriam-Webster, 2004)!"

A Profound Understanding of Quality

Quality is the pivotal word in TQM. A fundamental reason for the U.S. losing world leadership in manufacturing during the 1960s and 1970s was its lack of a profound understanding of the Q word. The gurus of quality, in the interest of developing a better understanding of the importance of quality, created shorter, more explicit, operationally oriented definitions such as "fitness for use" (Juran, 1988), "conformance to specifications" (Crosby 1989), "long term loss to society" (Taguchi 1986), and "a predictable degree of uniformity and dependability, at a low cost and suited to the market" (Deming, as paraphrased by Gitlow, Oppenheim, and Oppenheim, 1995).

Some have made light of the differences in these operational definitions of quality. A few have concluded that even the quality gurus cannot agree on the definition of quality. They should be viewed as being complementary, each definition emphasizing its definer's experience base in relation to the customer in question. "Fitness for use" is an appropriate operational definition of quality in the creation and marketing of a product or service on the production floor, where an employee may be far removed from the customer, the translation of quality performance measures into specific dimensions having specified targets and specification limits seems a necessity. Finally, if "loss to society" is thought of as "long-term business loss," then its relation to the other two operational definitions becomes clearer. Deming's definition exhibits his emphasis on variability and its reduction as a fundamental step in improving quality.

A first step in attaining a profound understanding of quality is the realization that it is customer-driven. It not only begins with the customer, in the end, it is judged by the customer. While the "voice of the customer" is imperative, a customer may not be able to fully articulate his needs and desires. Even the most sophisticated customers are not likely to be able to envision all of the characteristics of a product that will satisfy and "delight" them. Expert panels can serve an important role, but they too have their limitations. Obtaining this information is a complex task. Based on this input, product creators, developers and deliverers must envision these dimensions of quality that will satisfy and delight their customer.

Furthermore, they must maintain a dialog with the customer so that they will continue to understand and respond to this dynamic "voice of the customer."

Traditionally managers have viewed quality and cost as a zero-sum game. That is, any improvement in quality will occur only at a substantial additional cost. The following quote from Vaughn (1967) illustrates this "conventional wisdom:" "The trade off is between the effects of less emphasis on quality and the cost of more of it–." The following is a counter example. A citizens' group discovered that their water company was losing 40% of its treated water, prior to delivery. This meant that they were treating 167% more water than demanded, and hence 167% additional treatment facilities were required. Rather than making any attempt to reduce losses, the leadership had committed its users to millions of dollars of debt for a land purchase to build an un-needed reservoir.

Juran (1988) categorized quality associated costs (and estimates of the associated percentage for one industry) into four broad groups, those due to Internal Failures (30%), External Failures (40%), Appraisal (25%), and Detection and Prevention (5%). He also discussed how these percentages are dependent upon the maturity of the product line and effort expended on quality improvement. Juran's classic model for optimum quality levels also emphasized that there is a tradeoff between quality and cost. He stated that "failure costs decline until they are over-taken by the increasing costs associated with appraisal and prevention. At this point total costs increase." Juran also made clear the cost of quality through the phrase, hidden factory, where he exhibited the additional resources necessary to deliver products and services.

Cole (1992) made an excellent case for a fundamental paradigm shift regarding quality and costs and timeliness, based on the achievements of the Japanese. His conclusions are given in Table 1. Compare Cole's views with the old quality paradigm, "you get what you pay for." The truth is that high cost, alone, is not a guarantee that a product will be of high quality. Indeed, some times the contrary is true, resulting in what *Consumer Reports* refers to as "best buys." The six achievements of the Japanese, cited by Cole, have an important impact on conclusions drawn from quality cost models. Specifically, they indicate that the point at which it is no longer cost effective to improve quality is at a much lower defective rate than previously thought.

Total Quality Management, Table 1 Cole's underlying reasons for Japanese achievements in quality

"-realized that the costs of poor quality were far larger than had been recognized."
"-recognized that focusing on quality improvement as a firm-wide effort improved a wide range of performance measures."
"-established a system that moved toward quality improvement and toward low-cost solutions simultaneously."
"-focused on preventing error at the source, thereby dramatically reducing appraisal costs."
"-shifted the focus of quality improvement from product attributes to operational procedures."
"-evolved a dynamic model in which customer demands for quality rise (along with their willingness to pay for these improvements)."

Juran (1988) noted that many disagreements about achieving quality result from the fact that there are two fundamentally different quality issues, one income oriented and the other cost oriented. Features that produce customer satisfaction are income oriented. They are the key to attracting new customers and through satisfaction of retaining them. Cost oriented quality issues are the defects and failures that incur. They cause dissatisfaction and the loss of customers. As customers become aware of a product and indeed a producers track record through publications such as *Consumer Reports*, they also impact the ability of attracting and retaining customers. Furthermore, they impact the profitability of the firm through the dollars lost internally in defectives and rework and externally through warranty costs and other required services.

Establishing, appraising or judging the quality of a product are far more difficult than simply defining it. In his highly acclaimed book *Management of Quality*, Garvin (1988) elaborated eight dimensions of quality, including performance, features, reliability, conformance, durability, serviceability, aesthetics, and perceived quality. Through his study on air conditioners, he illustrated the differences in the perception of quality of various constituencies, noting that customers, companies (as represented by first line supervisors), service personnel and *Consumer Reports* view quality quite differently and elaborated on the reasons for these different perceptions.

While top-level management communicates in dollars, operations level personnel must be bilingual,

communicating in both dollars and things, that is, in product units and performance measures. Taguchi popularized the use of loss functions to provide a link between these two languages. They provide a means for expressing the deviation of product characteristics from their targeted values in dollars. These loss functions can be determined through internal costs of a product at each stage of design, development, manufacture and delivery.

Quality is achieved by elaborating the important product characteristics, their targets and specifications. The ability of a product to meet these specifications depends upon its design, development and the processes employed in its manufacture. Product and process information is often gathered through capability studies, where measurements are obtained on important product characteristics, and control charts are employed to address the stability of the processes and the predictability of future performance. These product characteristics are frequently summarized by a statistical distribution, or even more succinctly, by the process capability, six sigma.

Process capability indices are typically employed to combine this voice of the customer with the voice of the product/process into a dimensionless measure. Pignatiello and Ramberg (1996) reviewed this approach, stressing the importance of an appropriate data collection scheme and the statistical analysis and summarization of results. These indices, which are dimensionless quantities, are then employed in quality improvement project selection.

Total Quality

The term quality has traditionally been associated with manufacturing, and more explicitly, with the products, processes, functions, and facilities associated with manufacturing. The modern total quality viewpoint extends this factory oriented view of quality to encompass all products, goods and services whether they are for sale or not.

Total quality proponents embrace training and education as universal, in direct contrast to the Taylor system, a system to which U.S. leadership in productivity has been attributed. Taylor made a strategic decision to separate planning and execution. This decision was based on his assessment that the then immigrant work force was uneducated and that it was not economically feasible to educate them in a timely manner. A more highly educated work force represents an untapped resource for improving quality and productivity. Total quality proponents recognized this improvement in the educational level, and the responsibility for not only utilizing this resource, but improving it through the continuing education of the work force. Furthermore, they recognized these workers as stakeholders, and that by empowering these stakeholders, productivity and quality can be further enhanced.

Total Quality Management

While neither embraced the term Total Quality Management, its origins can be traced to the work of W. Edwards Deming and Joseph J. Juran, and through the implementation of their quality philosophy, concepts and methods in Japanese industry. Kolesar (1994, 2008) discusses the contributions of Deming and Juran to the Japanese quality revolution following WWII. The importance of TQM became fully recognized in the U.S. only after its successful Japanese implementation. The domination of their products, as a direct result of their outstanding quality, especially in the auto industry, could not go unrecognized. With this recognition, Deming and Juran gained the attention of enlightened U.S. corporate and government leaders.

Deming is perhaps best known for the Shewhart/Deming PDCA cycle, and his 14 point manifesto, which is fundamental to TQM philosophy. The PDCA cycle, now called the PDSA cycle, meaning Plan, Do, Study, and Act, provides a fundamental structure for achieving quality. Gitlow et al. (1995) give an excellent discussion of Deming's 14 points and employed the PDSA approach for achieving quality improvement. Scherkenbach (1986, 1991) provides a balanced view of the key characteristics of the philosophy of Deming given in Table 2. For example, one of Deming 14 points is "reduce waste," which Scherkenbach has balanced with "add value."

Kolesar (2008) states, "Juran's 1954 lectures have been credited with being seminal contributions to the Japanese quality control movement." Juran (1988) recognized the importance of including quality in the management game plan, as well as the

Total Quality Management, Table 2 Key characteristics of the Deming philosophy, from W.W. Scherkenbach (1991)

Reduce waste	Add value
Constancy of purpose	Continual improvement
Improvement	Innovation
Team	Individual
Long-term	Short-term
Inputs	Outputs
Synthesis	Analysis
Knowledge	Action

Total Quality Management, Table 3 Quality tools — the magnificent seven plus one

Control Charts
Check Sheets
Histograms
Pareto Diagrams
Ishikawa Fishbone Diagrams
Scatter Plots
Flow Charts or Process Diagrams
Multi-Variate Charts

need for developing managerial processes in managing quality. He noted financial management included three processes: financial planning (producing the budget), financial control (assuring that the budget will be met), and financial improvement (ways of increasing income and decreasing costs). Translated to quality, these are known as the Juran Trilogy: Quality Planning, Quality Control and Quality Improvement. A major advantage facilitating the implementation of these ideas is that senior management already understands them in the financial arena. Juran also stated "universal sequences for accomplishing these processes, the quality planning road map, quality control and the quality improvement processes." Fundamental to his methodology is the recognition of the presence of chronic quality wastes resulting from disconnected alarm systems.

Senge (1993) presents a TQM paradigm that is based on the three cornerstones: Guiding Ideas, Infrastructure, and Theory, Tools and Methods. He noted that guiding ideas are based on a vision. Without this vision, everything is mechanical and pedestrian. Leaders expressing this vision and these guiding ideas must practice them. When they make a decision differently, their colleagues and subordinates will know! However, these ideas and the behavior and actions of the leaders is not enough. An infrastructure is necessary for diffusing these ideas. Conflicts in goals must be resolved and this implies the importance of accountability and an appropriate reward structure. Finally, there is the theory, tools and methods cornerstone. Again, a necessary and important part of the structure, but certainly not sufficient on its own. OR/MS tends to be tool oriented.

Tables 3 and 4 list these essential tools, which seem so simple that they are frequently neglected in college courses. These tools of TQM are communications

Total Quality Management, Table 4 Quality management — the seven tools

Affinity Diagram
Interrelationship Digraph
Tree Diagram
Prioritization Matrices
Matrix Diagram
Process Decision Program Chart
Activity Network Diagram

enhancers that assist one in listening and talking to processes, products, systems and people. Smith (1998) described these tools and more advanced problem solving methods within the context of diagnostic disciplines.

Transformation to Quality Organizations

Implementation of total quality management in a firm requires a transformation of the organization, and any transformation of an organization is doomed to failure if it does not recognize the importance of the human aspect. Scherkenbach (1991) elaborated a theory of transformation that emphasizes this human aspect of quality. Scherkenbach notes how differently people view the world and why they are motivated by different means. Some, such as management scientists and operations researchers, live in the logical world. They tend to proceed on the basis of logical actions. Others, including many top-level managers and workers alike, live in a physical world. This is the world of policies, procedures, standards, rewards, and punishments. They do it by the book. Still others, such as sales personnel, marketing specialists and artists live in the emotional world, typified by the statement, "The force is with you."

Scherekenbach's point is not to create stereotypes, but to enable a better understanding of why arguments made in one of these domains often do not have a substantive impact on people living in another domain, i.e., when dealing with others, it is imperative to recognize that they may not be motivated by different forces. To make progress in relationships with others, one needs to be cognizant of their view and address them in an appropriate manner. As a point of exclamation to those of us who live in the logical world, Scherkenbach quotes Schopenhauer: "No one ever convinced anybody by logic; and even logicians use logic only as a source of income."

He goes on to describe transformation through three process relationships: one for each world view, and all given in terms of different mind states or attitudes dependent, independent and interdependent. Many people function solely in either the dependent or independent mode. An important aspect of the quality transformation is to facilitate the move to the interdependent mode.

TQM and Principle-Based Management

Each of us holds an important key to any quality transformation process in which we are involved. Covey (2004) suggested that we begin the quality transformation by taking action on ourselves first; then proceed through the four steps of his inside-out principle based management. He described these four steps as self, interpersonal, managerial and organizational. At the self level, he stresses the need to carefully develop our vision, decide what our life is about and develop those principles that will serve as our guidelines in making all of our decisions in life. Next is the need to act on this vision in a consistent manner that builds an internal source of security. Immediate or complete success should not be expected since this is a learning process. Incorporating and practicing the Shewhart/Deming PDSA cycle in our own work is an important method for improving the quality or our own work.

As we achieve some comfort with ourselves, and create a more positive opinion of ourselves, we will be able to move on to the interpersonal level. Covey stated that quality at the interpersonal level means that we live by the correct principles in our relationship with other people. Here Covey used the analogy of a bank account, that is, we make deposits to and withdrawals from an emotional bank account. He stated three important ground rules for achieving quality in interpersonal relationships. First, when we have a problem with a person, we should go directly to them and explain it. The second relates to the conduct of meetings. His ground rule is that no one is allowed to make a point in a meeting until they restate the point of their predecessor, and state it in a manner that is satisfactory to that person. He notes that this eliminates the majority of disagreements, since most of them are simply misunderstandings. Through this mechanism potential misunderstandings can be quickly clarified, avoiding arguments, further miscommunications and withdrawals from the emotional bank account. Furthermore, having greatly reduced the number of misunderstandings, there is a better chance to disagree agreeably when new disagreements take place. An important question is do we have the courage to practice this ground rule and continue to practice it even if the rest of group does not.

Finally, when we do make mistakes, we need to have the courage to say that we were wrong. No excuses. We must apologize to the person; we must also apologize to the other people involved. At the managerial level, quality means that we attempt to empower people. In this way they become increasingly independent of us. They supervise themselves, and we become a source of help, rather than a micromanager. Empowerment begins with self-control and self-inspection and extends to self-directing work teams. These teams plan processes, establish schedules, assign personnel and maintain discipline through peer pressure. They accomplish the work that was once limited to managers and specialists. Juran (1988) suggests that this system could be the successor to the Taylor system. It offers the opportunity to step off of the productivity and quality plateaus, which have been directly traced to the lack of involvement of the total work force, a result of not questioning the assumptions underlying Taylor's original separation of planning and execution. A craftsperson created a product from start to finish, and thus recognized the impact of each step on the following one. The production worker, as the execution of production was broken into individual components, had a smaller and decreasing opportunity to comprehend his role in achieving quality. As a result, inspection departments and later quality

departments emerged, acting as policing units in the goal to achieve quality.

At the organizational level, the key is in the structures and the leadership styles. Are the leaders in harmony with the mission statement? Was everyone involved in the development of the mission statement?

TQM and the Malcolm Baldrige Award

The Malcolm Baldrige Award framework provides an excellent road map for implementing TQM, as well a method for evaluating a firm's progress (NIST 1999). The framework emphasizes dynamic relationships between eleven categories of core values and concepts. These underlying core values and concepts are: customer-driven quality, leadership, continuous improvement and learning, employee participation and development, fast response, design quality and prevention, long-range view of the future, management by fact, partnership development, corporate responsibility and citizenship and results orientation.

The stated goals are customer satisfaction, customer satisfaction relative to competitors, customer retention and market share gain as measured by product and service quality, productivity improvement, waste reduction/elimination, supplier performance and financial results. Leadership is viewed as the "driver" category of core values and concepts, driving the two categories: business results and customer focus and satisfaction through a system of processes. The system of processes consists of four "well-defined and well-designed processes" for achieving the firm's performance requirements and the firm's customer requirements. These four system categories are information and analysis, strategic planning, human resource development and management, and process management. The criteria, which are updated annually, are disseminated by the American Society for Quality Control and the National Institute of Standards and Technology.

TQM and Six Sigma

Six Sigma is a relatively new program for accomplishing institutionalizing quality. The fundamental concept was created by a Motorola reliability engineer. Lean six sigma, a more recent development, incorporates fundamental industrial engineering and business "lean practices," with six sigma quality principles. Ramberg (2000) describes six sigma programs, and details its history in "Six Sigma: Fad or Fundamental."

The Status of TQM

One of the first evaluations of TQM was conducted by Senge. In his 1993 ASQ Annual Conference keynote address, titled "The Health and Well Being of the TQM Movement," he posed the following questions: "Are fundamental breakthroughs being made? Are they being made in your organization?" Following this opening, he summarized surveys by Arthur D. Little and McKinsey, and made the following conclusions. Out of 500 firms surveyed, less than a third were accomplishing anything! Two thirds of the TQM programs had ground to a halt! He went on to diagnose TQM failures and successes. Based on his case studies, he concluded that there were only a few major reasons for failure. The three major ones were: conflict between time and effort; wavering goals, and employee perception that their job was at risk.

Even where TQM has "succeeded," there are questions about the measures used to judge that success. That is, in many cases, even where the TQM indicators improved, the health of the company (e.g., as judged by its price) did not get any better, even over a reasonably long term. That is, TQM did not improve the health of the organization as judged by its stockholders. Reporting on the root cause of these problems, Senge concluded that a major reason was that most organizations viewed TQM as programmatic. Presented or implemented in this manner, TQM is certain to be DOA.

Comparative studies measuring the impact of TQM on a firm's business performance also began to appear. Jarrell and Easton (1994) reported some evidence that long-term performance of firms adopting TQM is improved. This result is consistent across the accounting and stock price performance measures examined. Similar, but overall stronger results, were found when the analysis was limited to a subsample of pilot firms identified as having more mature and well-integrated TQM systems. Hendricks and Singhal (1999) concluded that effective implementation of TQM "pays off in a big way." They made this

conclusion by comparing the business performances of firms judged to have successfully implemented TQM with a control group of firms.

van der Wiele et al. (2000) examined TQM through a "fad, fashion, and fit" analysis. Utilizing a range of research studies, which began in the late 1980s, they identified three stages in the evolution by which a fad can achiever a fit with previous management practice. In stage 1, the fad must be clearly defined and measurable. For TQM this clarification was ISO 9000 and the Baldridge Award. Stage 2 is the move to a fashion, which happens when major pressures toward widespread adoption of the fad are present. Again, ISO 9000 serves as an example, because suppliers experienced a pressure from major customers to achieve certification. van der Wiele et al. (2000) state, "As a consequence, the ISO 9000 series became a fast-spreading fashion." They elaborate that "Stage 3 is the move either from fad to fit or from fashion to fit. Fit into normal management practice means that the original fad will have effected the normal way of working within whole organizations and not just a small part such as would be the case in the adoption of a mere fashion." Their fieldwork shows that such a change will only occur when there is strong internal motivation and emotional involvement to implement TQM. They also point out that, "Should such a move take place from fad or fashion to fit, then the chances are that organizational performance will also be perceived to have been effected in a positive way."

Prajogo and Brown (2004) examining the relationship between TQM practices and quality performance in Australian organizations. They compared organizations that adopted formal TQM programs with those without a formal program. They concluded that the lack of a formal program did not necessarily mean TQM principles were not being practiced. Their findings also showed that the firms adopting formal TQM programs implemented several TQM practices at a higher level than those that did not have TQM programs. However, they did not observe a significant difference between organizations implementing formal TQM programs, and those organizations simply adopting TQM practices, suggesting that it is the adoption of quality practices that matters rather than formal programs per se.

While some researchers have given a rather pessimistic view on the future of the quality management movement, Kujala and Lillrank (2004) note that quality management has survived the failure of some of its success stories, such as those of Motorola and Xerox. They affirm that TQM remains to be properly defined, and that its scientific foundations are still not transparent.

Cheng (2007) explored a model for integrating TQM and Six Sigma with business strategy. He concluded that, "Implementing Six Sigma has become a common theme in organizations of all sizes, within a TQM infrastructure."

To summarize, it seems that TQM and its derivatives are fitting into management infrastructure. However, it is important for quality proponents that TQM is not the only thing. TQM will continue to require definition and structural development based on scientific foundations. Most recent of these has come from the Six Sigma movement, and more recently from Lean Six Sigma. Transformational leadership remains a requirement for continued success.

See

► Quality Control
► Reliability of Stochastic Systems

References

Cheng, J. L. (2007). Six sigma and TQM in Taiwan: An empirical study. *Quality Management Journal, 14*(2), 7–18.

Cole, R. E. (1992). The quality revolution. *Production and Operations Management, 1*, 118–120.

Covey, S. (2004). *The 7 habits of highly effective people*. New York: Fireside Books.

Crosby, P. B. (1989). What are requirements? *Quality Progress*, 47. ASQC.

Defeo, J. A., & Juran, J. M. (2010). *Juran's Quality handbook* (6th ed.). Southbury, CT: Juran Institute.

Deming, W. E. (2000). *Out of the crisis*. Cambridge, MA: MIT Press.

Garvin, D. A. (1988). *Managing quality*. New York: Free Press.

Gitlow, H., Oppenheim, A., & Oppenheim, R. (1995). *Quality management: Tools and methods for improvement*. Homewood, IL: Irwin.

Hendricks, K. B., & Singhal, V. R. (1999). Don't count TQM out. *Quality Progress*, 35–42.

Jarrell, S. L., & Easton, G. S. (1994). An exploratory empirical investigation of the effects of total quality management on corporate performance. In P. Lederer (Ed.), *The practice of quality management*. Cambridge, MA: Harvard Business School Press.
Juran, J. M. (1988). *Juran's quality control handbook* (4th ed.). New York: McGraw-Hill.
Kolesar, P. J. (1994). What Deming told the Japanese in 1950. *Quality Management Journal, 2*(1), 9–24.
Kolesar, P. J. (2008). Juran's lectures to Japanese executives in 1954: A perspective and some contemporary issues. *Quality Management Journal, 15*(3), 7–16.
Kujala, J., & Lillrank, P. (2004). Total quality management as a cultural phenomenon. *Quality Management Journal, 11*(4), 43–55.
Merriam Webster Staff. (2004). *The Merriam-Webster dictionary*. Springfield, MA: Merriam-Webster.
NIST (National Institute of Standards and Technology). (1999). *Ten years of business excellence for America*. Gaithersburg, MD: National government publication.
Pignatiello, J. J., Jr., & Ramberg, J. S. (1996). Process capability: Engineering and statistical issues. In J. B. Keats & D. C. Montgomery (Eds.), *Statistical applications in process control*. New York: Marcel Dekker.
Prajogo, D., & Brown, A. (2004). The relationship between TQM practices and quality performance and the role of formal TQM programs: An Australian empirical study. *Quality Management Journal, 15*(3), 32–42.
Prybutok, V. R., & Zhang, X. (2010). Introduction to the special issue on quality management in healthcare. *Quality Management Journal, 17*(4), 7.
Ramberg, J. S. (1994). TQM: Thought revolution or Trojan horse? *OR/MS Today, 21*(4), 18–24.
Ramberg, J. S. (2000). Six sigma: Fad or fundamental. *Quality Digest*, May 2000.
Scherkenbach, W. W. (1986). *The Deming route to quality and productivity: Road maps and road-blocks*. Washington, DC: ASQC Press/Washington CEE Press.
Scherkenbach, W. W. (1991). *Deming's road to continual improvement*. Knoxville, TN: SPC Press.
Scholtes, P. R., & Hacquebord, H. (1988). Six strategies for beginning the quality transformation (Part III). *Quality Progress*, 28–33.
Senge, P. (1990). *The fifth discipline: The Art and practice of the learning organization*. New York: Doubleday.
Senge, P. (1993). Quality management: Current state of the practice. *Keynote speech at the American Quality Congress*.
Smith, G. F. (1998). Determining the cause of quality problems: Lessons from diagnostic disciplines. *Quality Management Journal, 5*(2), 24–41.
Taguchi, G. (1986). *Introduction to quality engineering*. Tokyo: Asian Productivity Organization.
van der Wiele, A., Williams, A. R. T., & Dale, B. G. (2000). Total quality management: Is it a fad, fashion, or fit? *Quality Management Journal, 65*(2), 65–79.
Vaughn, R. C. (1967). *Introduction to industrial engineering*. Ames, IA: Iowa State University Press.
Wooden, J., & Yaeger, D. (2009). *A game plan for life: The power of mentoring*. New York: Bloomsbury Press.

TQC

Total quality control.

See

► Quality Control
► Total Quality Management

TQM

Total quality management.

See

► Quality Control
► Total Quality Management

Traffic Analysis

Denos C. Gazis
PASHA Industries, Katonah, NY, USA

Introduction

Traffic analysis has flourished since the 1950s, stimulated from the need to address the ever-growing traffic problems of cities around the world. In true scientific tradition, it has yielded an understanding of the fundamental characteristics of automobile traffic, which in turn spawned significant contributions in the management and optimization of traffic facilities. This article outlines some of the most important developments in one area of traffic analysis, that of traffic flow, including certain associated queueing phenomena. Aspects of control of traffic networks that are outside the scope of this article can be found in Gazis (1992).

A Kinematical Theory of Traffic Flow

One of the earliest, and most durable, contributions to the understanding of traffic flow was given by Lighthill and Whitham (1955). They viewed the traffic as a special fluid which obeys some basic laws consistent with the physical nature of traffic, such as its unidirectional influence of a vehicle only on the traffic behind it, the constraints on flow imposed by human limitations, etc. The Lighthill-Whitham theory is based on two basic postulates:

1. Traffic is conserved, in the sense that traffic units by and large are neither created nor annihilated; and
2. There is a fundamental relationship between traffic flow and traffic density, resulting from the physical characteristics of the traffic system.

The first postulate is expressed in the relationship

$$\frac{\partial k}{\partial t} + \frac{\partial q}{\partial x} = 0 \tag{1}$$

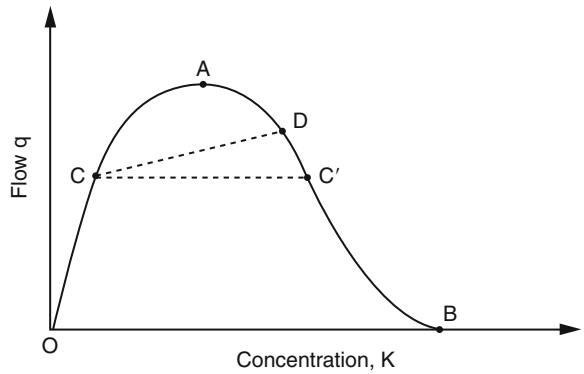
where q is the traffic flow in vehicles per unit of time t, k is the density of traffic in vehicles per unit of distance x, and v is the (average) speed of the traffic fluid. The second postulate is expressed by the relationship

$$q = f(k)$$

between flow q and density k such as that shown in Fig. 1. At zero density, there is zero flow. The flow is also zero at some jam density, k_j, because traffic grinds to a halt as vehicles are packed bumper to bumper. Between these two extremes, traffic flow builds up to a maximum and then decreases down to zero.

A number of interesting properties of traffic can be described on the basis of these two postulates. They relate to observable phenomena such as wave propagation, i.e., the movement along the traffic stream of a transition point corresponding to a change in traffic characteristics, the queueing caused by an obstruction of the traffic movement, etc.

Wave Propagation — Traffic moving at a steady-state flow rate q_1 and density k_1 may shift to a different flow rate q_2, and a corresponding density k_2, by a change in roadway quality, obstruction, or other external influence. When this happens, vehicles situated in a transition region undergo maneuvers adjusting their speed and inter-vehicle spacing, and



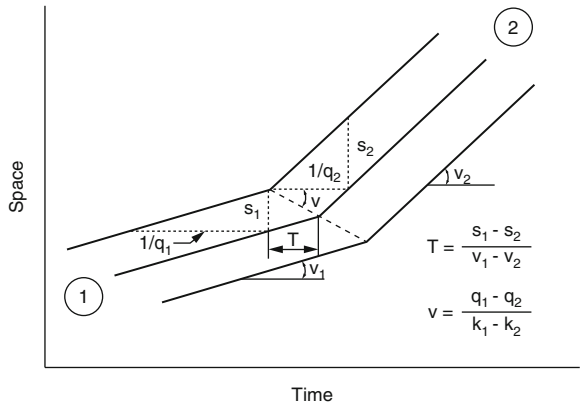
Traffic Analysis, Fig. 1 Flow vs. concentration relationship

this transition region generally moves either forward or backward in space depending on the nature of the change. The adjustments of speed and spacing are gradual, but for the purpose of deriving the characteristics of wave propagation may be assumed abrupt as suggested by Lighthill and Whitham (1955). This assumption leads to the conclusion that a change from one steady-state flow condition to another is associated with a "shock wave," an expression that pervades the traffic engineering literature.

The shock wave marks the transition from one speed to another, and moves always backwards with respect to the traffic stream, since vehicles exert an influence only on vehicles behind them. (The influence of an occasional tailgating vehicle pushing the vehicle in front is ignored as an unimportant aberration). The speed of movement of the shock wave along a roadway, may be obtained on the basis of Eq. (1), and is given by

$$\upsilon = \frac{q_1 - q_2}{k_1 - k_2} \tag{3}$$

It should be pointed out that the result given in Eq. (3) depends only on the postulate of conservation of traffic, and is totally independent of any specific relationship between flow and concentration, or even on the existence of such a relationship. It results from kinematical considerations shown in Fig. 2. The transition from one steady state flow situation to another results in a propagation of the change of the corresponding speed along the roadway. The phase velocity of this propagation depends only on the values of the initial and final pairs of flow, q, and



Traffic Analysis, Fig. 2 Transition from one steady-state-flow situation to another

concentration, k, and is given by Eq. (3). If, in addition, a relationship between flow and concentration is assumed (Fig. 1), different domains of traffic quality, and corresponding characteristics of wave propagation, can be defined as follows:

1. The range from zero flow at zero density to maximum flow (Section OA, Fig. 1) corresponds to relatively uncongested traffic flow. A small increase in density in this domain moves forward along the roadway;
2. The range from maximum flow to zero flow at "jam density" (Section AB, Fig. 1) corresponds to relatively congested, stop-and-go traffic. A small increase of density in this domain moves backwards along the roadway; and
3. Any transition from one steady state flow to another (as from point C to point D, Fig. 1) is associated with a wave propagation given by the slope of segment CD.

Queueing — Queueing may be caused by a reduction in roadway capacity at a fixed point on the roadway, or by an obstruction causing traffic to shift from the uncongested to the congested branches of the (q, k) curve, even without reduction in flow rate, (line CC′ in Fig. 1). The rate of growth of the queue can be estimated using the same methodology described above. For example, a total obstruction of flow q and density k causes a queue formation, with the tail-end of the queue moving backwards along the roadway with speed equal to

$$\upsilon = \frac{q}{k_j - k} \tag{4}$$

Additional results from the kinematic treatment of traffic — An extensive literature exists on applications of the Lighthill-Whitham model to various traffic phenomena. A word of caution is appropriate with regard to such applications. The Lighthill-Whitham model describes well only transitions from one steady state to another. Any attempt to apply the model to a sequence of traffic maneuvers that do not allow enough relaxation time between changes of speeds violates the basic spirit of the model.

An interesting extension of the above kinematical treatment of traffic was applied by Gazis and Herman (1992) for the treatment of a moving obstruction such as that caused by a vehicle moving more slowly than the other vehicles in the traffic stream. The character of this " moving bottleneck" is different from that of a fixed bottleneck, and the Gazis-Herman treatment derives the characteristic queueing behavior associated with it. Gazis and Herman obtain a description of the queueing caused by a slow vehicle on a two-lane highway. Both lanes are affected by such a vehicle, one by direct trapping of vehicles behind the slow one, and the other by interference from vehicles escaping from the queue behind this vehicle. The result is that queueing takes place in both lanes in the vicinity of the slow vehicle, with the affected vehicles moving at an average speed only marginally higher than that of the slow one, until they come abreast of this slow vehicle and are able to escape at their normal speed. Gazis and Herman also propose an explanation of the phenomenon of a phantom bottleneck, the seemingly unexplainable regions of congestion that drivers often traverse. Some of them may be caused by a moving bottleneck caused by a vehicle that slows down temporarily and then resumes its normal speed; for example, a heavily loaded truck temporarily slowing down along an uphill portion of the roadway. The Gazis-Herman treatment provides a rational way of estimating the minimum allowable speed on a highway, which would not affect its throughput.

A Boltzmann-like Model of Traffic Flow

In 1959, Prigogine suggested a model of traffic flow founded on statistical mechanics, analogous to the Boltzmann model of gases (Prigogine 1961). The Prigogine model was subsequently developed extensively by Herman, Prigogine and their

collaborators (Prigogine, Herman and Anderson 1962, 1965; Prigogine and Herman 1971). They considered a stream of traffic as an ensemble of units associated with certain statistical properties. In particular, a vehicle was associated with a desired speed which it would follow as long as it was not constrained by another vehicle in front with a lower desired speed.

Thus, traffic is described in terms of a probability density for the speed, v, of an individual car, $f(x, v, t)$. This density may vary as a function of time, t, and a coordinate x along the highway. The basic equation for this function f is assumed to be

$$\frac{\partial f}{\partial t} + \upsilon \frac{\partial f}{\partial x} = \left(\frac{\partial f}{\partial t}\right)_{relaxation} + \left(\frac{\partial f}{\partial t}\right)_{interaction} \tag{5}$$

The first term of the righthand side of Eq. (5) is a consequence of the fact that $f(x, v, t)$ differs from some desired speed distribution $f^0(v)$. A car tries to "relax" to its desired speed as soon as it finds an opportunity to do so. The second term of the righthand side corresponds to the slowing down of a fast vehicle by a slow one. True to his tradition as a leading expert in statistical mechanics, Prigogine frequently referred to this second term as the collision term — a rather unsettling choice of words in this context!

The form for these two terms was chosen for mathematical convenience and plausibility, leading to the equation

$$\frac{\partial f}{\partial t} + \upsilon \frac{\partial f}{\partial x} = \frac{f - f_0}{\tau} + (1-p)k(V-\upsilon)f \tag{6}$$

where τ is a characteristic relaxation time, p is the probability of a car's passing another car, and V is the average speed of the stream of traffic. The second term of the right-hand side of Eq. (6) corresponds to the interaction term, and tends to zero at very light traffic concentration when the probability of passing is close to unity, in which case the relaxation term is dominant. If, in addition, a highway with constant properties along its length is assumed, then $\partial f/\partial x = 0$ and the solution of Eq. (6) is

$$f(\upsilon, t) = f^0(\upsilon) + \left[f(\upsilon, 0) - f^0(\upsilon)\right]e^{-t/\tau} \tag{7}$$

If interested only in solutions of Eq. (6) that are independent of time and space, then the lefthand side of this equation is zero. The equation may then be solved to yield an equation of state whose general form, for small values of the concentration, corresponds to an approximately linear increase of flow with concentration, e.g.,

$$q = V^0 k \tag{8}$$

where V^0 is the average of the desired speed. As k increases, the flow q falls below the straight line (8) due to the increasing influence of interactions.

In the range of high concentrations, q is independent of f^0 and depends only on τ and p, according to the equation

$$q = \frac{1}{\tau(1-p)}. \tag{9}$$

The complete solution of Eq. (6) for steady-state flow, independent of time and space, is shown in Fig. 3. For any given f^0, the flow q rises with k, reaches a maximum, and then decreases until it intersects a curve corresponding to Eq. (9). This curve may be viewed as a universal curve of collective flow, characterized by high densities and very little passing. One very realistic feature of this theory is the fact that it predicts probable stoppage of some vehicles in the domain of collective flow, in agreement with the common experience of stop-and-go traffic at high concentrations.

It is appropriate to make an observation concerning the linkage of the Herman-Prigogine and Lighthill-Whitham theories in the range of very high densities. Since traffic at those densities is of a stop-and-go nature, it is not really steady-state traffic in the sense of being associated with constant speed and density. Rather, it is associated with alternating states of following slow platoons and escaping from them. Given this fact, it becomes clear that one should not try to apply the Lighthill-Whitham method in describing shock waves and wave propagation involving transitions into this domain of traffic movement, since the L-W theory describes well only clean transitions between two steady-state situations.

Herman and Prigogine (1979), together with several collaborators, went on to use the results of their model to develop a two-fluid approach to town traffic. This approach postulates that traffic in towns is a mixture of

two fluids, one that moves and one that is stopped. Any individual vehicle traverses a network in a stop-and-go fashion, moving part of the time and being stopped part of the time. The quality of service in a particular urban network can be described in terms of two parameters that can be determined by circulating a test vehicle through the network and measuring the percentages of time during which the vehicle is moving, or is being stopped. Thus the two-fluid model yields a simple description of the system-wide traffic quality in congested urban networks. It allows comparison between different urban networks, and it offers the potential of identifying important elements of the network, related to its geometry or control features, which may be targeted for improvement of the service quality.

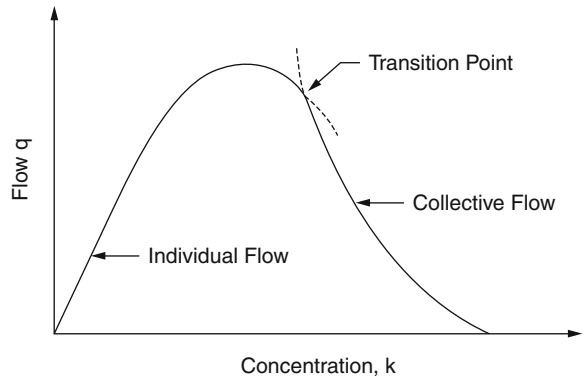
A Car-following Theory of Traffic Flow

Reuschel (1950) and Pipes (1953) proposed models to describe the detailed motion of cars proceeding close together in a single lane. This microscopic, car-following theory of traffic flow was extensively developed by Herman et al. (1959). The theory is based on the fact that when drivers do not have the freedom to pass a vehicle in front, they follow it in a way that is controlled by the overriding need to avoid coinciding with the leader in space and time. In trying to achieve this reasonable objective, drivers react to a limited set of inputs. The postulate of the car-following theory, confirmed by experiments, was that drivers reacted mostly to the relative speed between their car and that of the one in front. Experiments showed a high correlation between the acceleration of a car and its speed relative to that of a leader, after a time-lag of the order of 1 second. This led to the linear car-following model

$$\frac{d^2x_n(t+T)}{dt^2} = \lambda\left[\frac{dx_{n-1}(t)}{dt} - \frac{dx_n(t)}{dt}\right] \qquad (10)$$

in which n denotes the position of a car in a line of cars (a platoon), λ is a constant gain factor, T is the reaction time-lag, and x_n is the position of the nth car on the highway.

This model was used to investigate the stability of a traffic platoon when a perturbation in its movement is introduced. The movement of the platoon is said to be locally stable if the amplitude of a perturbation, for any given car in the platoon, decreases in time. It is asymptotically stable if the amplitude of the perturbation decreases as it propagates upstream. The value of the product λT is the determinant of stability or instability, local or asymptotic. When $\lambda T < 1/e$, a perturbation is damped exponentially as it is passed on to the following car, signifying a very stable situation. For λT between the values of $1/e$ and $\pi/2$, the perturbation produces oscillations of decreasing amplitude between pairs of cars, signifying still a locally stable situation. For $\lambda T > \pi/2$, a perturbation produces oscillations of increasing amplitude, signifying a locally unstable situation.



Traffic Analysis, Fig. 3 Flow vs. concentration relationship according to the Boltzmann-like model of traffic flow

With regard to asymptotic stability, the dividing line is at $\lambda T = 1/2$. For values of λT below 1/2, the amplitude of a perturbation decreases as it propagates backwards; for values of λT greater than 1/2, it increases. This means that between $1/e(\sim 0.368)$ and 1/2 is a situation that is locally stable but asymptotically unstable. Any pair of cars in a platoon is able to absorb a perturbation, but it amplifies it as it passes it backwards, until the perturbation is so large that it causes a collision.

The linear car-following model may be satisfactory in describing fluctuations around a steady-state, constant speed situation. It cannot be expected to describe equally well transitions from one steady state to another involving large changes of speed. For this reason, Gazis et al. (1961) proposed a nonlinear model in which the gain factor is not constant but depends on the speed of the follower and the relative spacing between leader and follower according to the relationship

$$\lambda = \frac{[v_n(t+T)]^l}{[x_{n+1}(t) - x_n(t)]^m} \tag{11}$$

where c is a constant, $v = dx/dt$ is the speed, and (l, m) are integer exponents identifying particular nonlinear models.

Various values of pairs (l, m) were used to define car-following models and investigate their predictions concerning transitions between one steady-state flow situation and another. Integrating over time Eq. (10), with λ described by Eq. (11), leads to the functional relationship between changes of speed and concentration. Together with appropriate boundary conditions, for example the condition of zero speed at jam density, bumper-to-bumper concentration, one can then obtain a phenomenological relationship between flow and concentration such as that shown in Fig. 1. Various pairs (l, m) have been used which yielded quite plausible relationships, consistent with observations.

The preceding discussion outlines most of the key contributions in the car-following treatment of traffic flow. Additional studies have been contributed by Gazis (1965) within the framework of control theory to account for physical constraints on the system, such as limited acceleration or deceleration capability of cars.

Concluding Remarks

As is the case for every scientific endeavor, much can be done to improve the theories of traffic analysis. For example, car-following theories ignore interaction of cars with other than the car just in front, whereas there is evidence that drivers are very much conscious of happenings several cars in front of them, and this consciousness tends to improve the stability of traffic. Another observation that must be made about virtually all traffic models described here is that they effectively correspond to flat, straight, and infinitely long highways. It is clear that the geometry of highways, including curves and inclination, has a strong effect on the behavior of traffic. A systematic study of such effects would greatly advance understanding of traffic movement, and produce necessary tools for future improvements in traffic management.

The analytical description of traffic flow has already had a profound influence on traffic engineering practice, and the advent of activities in the area of Intelligent Transportation Systems (ITS) points to an increasing reliance on analytical investigations of traffic systems toward improvement of their operation. One needs an improved understanding, and an improved analytical description of traffic phenomena, such as the onset of congestion, queueing, and inter-vehicle signal propagation, in order to create the theoretical under-pinning toward the use of high technology for the improvement of traffic systems, which is the central thrust of ITS. Some improvement will come from direct application of analytical results. For example, the development of automatic highways will undoubtedly draw from knowledge based on car-following models. Other improvements may come from the improved understanding of traffic phenomena that traffic analysis provides, leading to improved heuristic schemes for the control and optimization of traffic systems.

See

- ▶ Network Optimization
- ▶ Networks of Queues
- ▶ Queueing Theory
- ▶ Simulation of Stochastic Discrete-Event Systems

References

Anderson, R. L., Herman, R., & Prigogine, I. (1962). On the statistical distribution function theory of traffic flow. *Operations Research, 10*, 180–196.

Ardekani, S. A., & Herman, R. (1985). A comparison of the quality of traffic service in downtown networks of various cities around the world. *Traffic Engineering and Control, 26*, 574–581.

Ardekani, S. A., & Herman, R. (1987). Urban network-wide traffic variables and their relations. *Transportation Science, 21*, 1–16.

Bick, J. H., & Newell, G. F. (1960). A continuum model for two-directional traffic flow. *Quarterly of Applied Mathematics, 18*, 191–204.

Chandler, R. E., Herman, R., & Montroll, E. W. (1958). Traffic dynamics: Studies in car-following. *Operations Research, 6*, 165–184.

Chang, M.-F., & Herman, R. (1981). Trip time versus stop time and fuel consumption characteristics in cities. *Transportation Science, 15*, 183–209.

Edie, L. C., & Foote, R. S. (1960). Effect of shock waves on tunnel traffic flow. *Proceedings of Highway Research Board, 39*, 492–505.

Edie, L. C., Herman, R., & Lam, T. N. (1980). Observed multilane speed distribution and the kinetic theory of vehicular traffic. *Transportation Science, 14*, 55–76.

Foster, J. (1962). An investigation of the hydrodynamic model for traffic flow with particular reference to the effect of various speed-density relationships. *Proceedings of Australian Road Research Board, 1*, 229–257.

Gazis, D. C. (1965). Control problems in automobile traffic. *Proceedings of IBM scientific symposium on control theory and applications*, *IBM Yorktown Heights, New York*, 171–185.

Gazis, D. C. (1992). Traffic modelling and control: Store and forward approach. In M. Papageorgiou (Ed.), *Concise encyclopedia on traffic and transportation* (pp. 278–284). New York: Pergamon Press.

Gazis, D. C., & Herman, R. (1992). The moving and phantom bottlenecks. *Transportation Science, 6*, 223–229.

Gazis, D. C., Herman, R., & Potts, R. B. (1959). Car-following theory of steady-state traffic flow. *Operations Research, 7*, 499–505.

Gazis, D. C., Herman, R., & Rothery, R. W. (1961). Nonlinear follow-the-leader models of traffic flow. *Operations Research, 9*, 546–567.

Greenberg, H. (1959). An analysis of traffic flow. *Operations Research, 7*, 79–85.

Herman, R., & Potts, R. B. (1961). Single-lane traffic theory and experiment. In Herman, R. (Ed.), *Proceedings of the 1st international symposium on the theory of traffic flow*. Elsevier, 120–146.

Herman, R., & Ardekani, S. A. (1984). Characterizing traffic conditions in urban areas. *Transportation Science, 18*, 101–140.

Herman, R., Montroll, E. W., Potts, R. B., & Rothery, R. W. (1959). Traffic dynamics: Analysis of stability in car following. *Operations Research, 7*, 86–106.

Herman, R., & Prigogine, I. (1979). A two-fluid approach to town traffic. *Science, 204*, 148–151.

Leutzbach, W. (1967). Testing the applicability of the theory of continuity on traffic flow at bottle-necks. In Edie, L. C., Herman, R., & Rothery, R. W. (Eds.), *Proceedings of the 3rd international symposium on theory of traffic flow*. Elsevier, 1–13.

Lighthill, M. J., & Whitham, G. B. (1955). On kinematic waves: II. A theory of traffic flow on long crowded roads. *Proceedings of Royal Society (London), A229*, 317–345.

Makigami, Y., Newell, G. F., & Rothery, R. W. (1971). Three-dimensional representations of traffic flow. *Transportation Science, 5*, 302–313.

Newell, G. F. (1965). Instability in dense highway traffic, a review. In Almond, J. (Ed.), *Proceedings of the 2nd international symposium on theory of traffic flow*. OECD, 73–83.

Newell, G. F. (1991). *A simplified theory of kinematic waves*. Research report UCB-ITS-RR-91-12, University of California at Berkeley.

Pipes, L. A. (1953). An operational analysis of traffic dynamics. *Journal of Applied Physics, 24*, 274–281.

Prigogine, I. (1961). A Boltzmann-like approach to the statistical theory of traffic flow. In Herman, R. (Ed.), *Proceedings of the 1st international symposium on the theory of traffic flow*. Elsevier, 158–164.

Prigogine, I., & Andrews, F. C. (1960). A Boltzmann-like approach for traffic flow. *Operations Research, 8*, 789–797.

Prigogine, I., & Herman, R. (1971). *Kinetic theory of vehicular traffic*. New York: American Elsevier.

Prigogine, I., Herman, R., & Anderson, R. L. (1965). Further developments in the Boltzmann-like theory of traffic flow. In Almond, J. (Ed.), *Proceedings of the 2nd international symposium on the theory of traffic flow*. OECD, 129–138.

Prigogine, I., Herman, R., & Anderson, R. L. (1962). On individual and collective flow. *Académie Royale de Belgique — Bulletin de la Classe des Sciences, 48*, 792–804.

Prigogine, I., Resibois, P., Herman, R., & Anderson, R. L. (1962). On a generalized Boltzmann-like approach for traffic flow. *Académie Royale de Belgique — Bulletin de la Classe des Sciences, 48*, 805–814.

Reuschel, A. (1950). Fahrzeugbewegungen in der Kolonne bei gleichfoermig beschleunigtem oder verzoegertem Leitfahrzeug. *Zeit. d. oesterreichischen Ing. u. Arch. Vereins, 95*, 73–77.

Underwood, R. T. (1962). Some aspects of the theory of traffic flow. *Proceedings of Australian Road Research Board, 1*, 35.

Underwood, R. T. (1964). Traffic flow models. *Traffic Engineering and Control, 5*, 699–701.

Traffic Equations

In a queueing network, the set of linear equations that results from balancing flow into each node with the flow out. These traffic equations are derived by recognizing that the total input seen at a node comes from summing the flow of new arrivals from outside the network with the flow of arrivals that are due to departures from service completions at nodes within the network:

$$\lambda_i = \gamma_i + \sum_j \lambda_j r_{ij}$$

where λ_i is the total input flow rate seen at node i, γ_i is the external input rate to node i, r_{ij} is the probability that a service completion at node i is routed to node j, and the summation is taken over all nodes in the network.

See

► Conservation of Flow
► Networks of Queues
► Queueing Theory

Traffic Intensity

The average load offered to each server in a queueing system.

See

- ▶ Offered Load
- ▶ Queueing Theory

Traffic Process

A stochastic point or marked point process representing the flow of customers on the arcs of a queueing network. Marks represent some aspect of the customer or the state of the network and the points represent the epoch of the event.

See

- ▶ Arrival Process
- ▶ Departure Process
- ▶ Input Process
- ▶ Networks of Queues
- ▶ Output Process

Transfer Function

- ▶ Time Series Analysis

Transient Analysis

The time-dependent solution of a stochastic system (such as a queueing network), as contrasted with a steady-state solution.

See

- ▶ Queueing Theory

Transition Function

A function describing the transition probabilities of a Markov process $\{X(t), t \in T\}$ into a subset A of the state space as $p(s, x; t, A) = \Pr\{X(t) \in A | X(s) = x\}$, for state x and times $s < t$ in the time domain T.

See

- ▶ Markov Chains
- ▶ Markov Processes

Transition Matrix

The matrix of (single-step) stationary transition probabilities of a Markov chain $\{X_n\}$, $\boldsymbol{P} = [p_{ij}]$, where $p_{ij} = \Pr\{X_{n+1} = j | X_n = i\}$ is the conditional probability that the chain moves to state j from state i in one step.

See

- ▶ Markov Chains
- ▶ Markov Processes

Transition Probabilities

The conditional probabilities describing the movement from state to state of a Markov process $\{X(t), t \in T\}$. In general, the transition probabilities are written as $\Pr\{X(t) \in A | X(s) = x\}$ for times $s < t$ in the time domain T and state x and event (set) A in the state space. For a homogeneous discrete-time Markov chain (DTMC) $\{X_n, \mathrm{n} \geq 0\}$, the stationary transition probabilities are $\Pr\{X_{n+1} = j | X_n = i\} = p_{ij}$, for states i and j in the space state.

See

- ▶ Markov Chains
- ▶ Markov Processes

Transportation Problem

A linear-programming problem of the following form is called a transportation problem:

$$\text{Minimize} \sum_i \sum_j c_{ij} x_{ij}$$

subject to

$$\sum_j x_{ij} = a_j \quad i = 1, \ldots, m \quad (\text{origins/supply})$$
$$\sum_i x_{ij} = b_j \quad j = 1, \ldots, n \quad (\text{destinations/demand})$$
$$x_{ij} \geq 0.$$

The variables $\{x_{ij}\}$ represent a shipment of a homogeneous product from origin i to destination j, where the $\{a_i\}$ are the amounts of the product to be shipped from the origins i, and the $\{b_j\}$ are the amounts demanded by the destinations j. The form presented here assumes $\sum_i a_i = \sum_j b_j$, but the problem can also be formulated with the origin constraints as $\geq$ inequalities and the destination constraints as $\leq$ inequalities, without the restriction that the total supply equal the total demand. It can be shown that if the $\{a_i\}$ and $\{b_j\}$ are integers, than an optimal basic feasible solution exists that is all integer. The transportation problem is a special network problem whose network representation is called a bipartite graph. The special case with $m = n$ and all $\{a_i\}$ and $\{b_j\}$ equal to 1 is the assignment problem. A transportation problem can be solved by direct application of the simplex method, but due to its mathematical structure, the problem can be solved by an efficient modification of the simplex method called the transportation (primal-dual) simplex method. It can also be solved by specialized network algorithms.

See

- ▶ Assignment Problem
- ▶ Network Optimization
- ▶ Northwest-Corner Solution
- ▶ Transportation Simplex (Primal-Dual) Method
- ▶ Unbalanced Transportation Problem

Transportation Problem Paradox

Some transportation problems exhibit the paradox that an optimal solution can be improved if the total amount of units shipped is more than the total amount shipped by the optimal solution. In other words, one can ship more for less.

Transportation Simplex (Primal-Dual) Method

The dual problem to the primal equation form of the transportation problem can be stated as follows:

$$\text{Maximize} \sum_i a_i u_i + \sum_j b_j v_j$$

subject to

$$u_i + v_j \leq c_{ij} \quad \text{for all } (i, j).$$

Here the $(m + n)$ set of dual variables u_i and v_j are unrestricted (free) variables. Note that the primal has a redundant equation due to the equality of the total supply and demand. Thus, a feasible basis matrix to the transportation problem is of dimension $(m + n - 1) \times (m + n - 1)$. It can be shown that any feasible basis matrix can be arranged into a triangular form. For a given basis, the simplex method requires that the corresponding dual constraints must hold at equality, i.e., $u_i + v_j = c_{ij}$ for all variables x_{ij} in the basis. This $(m + n - 1) \times (m + n)$ set of dual equations can be reduced to an $(m + n - 1) \times (m + n - 1)$ system by arbitrarily setting one of the dual variables, say $u_1 = 0$. This corresponds to removing, as a redundant constraint, the first equation of the transportation problem. The resulting dual square set of equations also has a triangular form that allows for the efficient calculation of the $\{u_i\}$ and $\{v_j\}$ that correspond to the current basic solution. These values of u_i and v_j are used to calculate the $(u_i + v_j)$ terms for the nonbasic

variables, and if each one is less than or equal to its corresponding c_{ij}, then by duality theory and complementary slackness, the current basis is optimal. If the latter condition does not hold, the usual simplex criterion is used to select a variable to enter the basis and a new basic feasible solution is generated by simple adjustments to the flows in the network that describe the current basic feasible solution. This network is a tree that connects all origins and destinations, and the addition of the new variable (or arc to the tree) enables the new solution to be calculated readily. This primal-dual process is repeated until an optimal solution is found. Such a solution exists because the transportation problem always has feasible solutions and the solution set is bounded.

See

▶ Network Optimization
▶ Transportation Problem

Transposition Theorems

Transposition theorems deal with disjoint alternatives of solvability of linear systems. For example, Stiemke's transposition theorem is the following: For a matrix $\boldsymbol{A} \neq \mathbf{0}$, the following statements are equivalent: (1) $\boldsymbol{Ax} = \mathbf{0}$, $\boldsymbol{x} > \mathbf{0}$, has no solution, and (2) $\boldsymbol{\mu A} \leq \mathbf{0}$, $\boldsymbol{\mu A} \neq \mathbf{0}$ has a solution.

See

▶ Farkas' Lemma
▶ Gordan's Theorem
▶ Strong Duality Theorem
▶ Theorem of Alternatives

Transshipment Problem

▶ Minimum-Cost Network-Flow Problem
▶ Network Optimization

Traveling Salesman Problem

Karla L. Hoffman[1], Manfred Padberg[2] and Giovanni Rinaldi[3]
[1]George Mason University, Fairfax, VA, USA
[2]New York University, New York, NY, USA
[3]CNR – Istituto di Analisi dei Sistemi ed Informatica (IASI), Rome, Italy

Introduction

The traveling salesman problem (TSP) has commanded much attention from mathematicians and computer scientists specifically because it is so easy to describe and so difficult to solve. The problem can simply be stated as: if a traveling salesman wishes to visit exactly once each of a list of m cities (where the cost of traveling from city i to city j is c_{ij}) and then return to the home city, what is the least costly route the traveling salesman can take? A complete historical development of this and related problems can be found in Hoffman and Wolfe (1985), Applegate et al. (2006), and Cook (2011).

The importance of the TSP is that it is representative of a larger class of problems known as combinatorial optimization problems. The TSP problem belongs in the class of such problems known as *NP*-complete. Specifically, if one can find an efficient (i.e., polynomial-time) algorithm for the traveling salesman problem, then efficient algorithms could be found for all other problems in the *NP*-complete class. To date, however, no one has found a polynomial-time algorithm for the TSP. Does that mean that it is impossible to solve *any* large instances of such problems? To the contrary, nowadays many practical optimization problems of truly large scale are solved to optimality routinely. From 1992 to 2006, Concorde, a software created by D. Applegate, R.E. Bixby, V. Chvátal, and W.J. Cook (Applegate et al. 1995, 2006), solved (among many others) a traveling salesman problem that models the production of printed circuit boards having 7,397 holes (cities), a problem over the 13,509 largest cities in the U.S., one over the 24,978 cities of Sweden, and, finally, a 85,900 city problem arising from a VLSI

application. So, although the question of what it is that makes a problem difficult may remain open, the computational record of specific instances of TSP problems coming from practical applications is optimistic.

How are such problems tackled? Obviously, one cannot consider a brute-force approach. For example, for a 16-city traveling salesman problem, there are 653,837,184,000 distinct routes that would need to be evaluated. Rather than enumerating all possibilities, successful algorithms for solving the TSP problem eliminate most of the routes without ever explicitly considering them.

Formulations

The first step to solving instances of large TSPs must be to find a good mathematical formulation of the problem. In the case of the traveling salesman problem, the mathematical structure is a graph where each city is denoted by a point (or node) and lines are drawn connecting every two nodes (called arcs or edges). Associated with every line is a distance (or cost). When the salesman can get from every city to every other city directly, then the graph is said to be complete. A round-trip (route) of the cities corresponds to some subset of the lines, and is called a tour or a Hamiltonian cycle in graph theory. The length of a tour is the sum of the lengths of the lines in the round-trip.

Depending upon whether or not the direction in which an edge of the graph is traversed matters, one distinguishes the asymmetric from the symmetric traveling salesman problem. To formulate the asymmetric TSP on m cities, one introduces zero-one variables

$$x_{ij} = \begin{cases} 1 \text{ if the edge } i \to j \text{ is in the tour} \\ 0 \text{ otherwise} \end{cases}$$

and, given the fact that every node of the graph must have exactly one edge pointing towards it and one pointing away from it, one obtains the classic assignment problem. These constraints alone are not enough since this formulation would allow subtours, i.e., it would allow disjoint loops to occur. For this reason, a proper formulation of the asymmetric traveling salesman problem must remove these subtours from consideration by the addition of subtour-elimination constraints. The problem then becomes

$$\begin{aligned} \min \quad & \sum_{j=1}^{m} \sum_{i=1}^{m} c_{ij} x_{ij} \\ \text{s.t.} \quad & \sum_{j=1}^{m} x_{ij} = 1 \ \text{ for } i = 1, \ldots, m \\ & \sum_{j=1}^{m} x_{ij} = 1 \ \text{ for } j = 1, \ldots, m \\ & \sum_{i \in K} \sum_{j \in K} x_{ij} \le |K| - 1 \ \text{ for all } K \subset \{1, \ldots, m\} \end{aligned}$$

where K is any nonempty proper subset of the cities $1, \ldots, m$. The cost c_{ij} is allowed to be different from the cost c_{ji}. Note that there are $m(m-1)$ 0–1 variables in this formulation.

To formulate the symmetric traveling salesman problem, one notes that the direction traversed is immaterial, so that $c_{ij} = c_{ji}$. Since direction does not now matter, one can consider the graph where there is only one arc (undirected) between every two nodes. Thus, let $x_j \in \{0,1\}$ be the decision variable where j runs through all edges E of the undirected graph and c_j is the cost of traveling that edge. To find a tour in this graph, one must select a subset of edges such that every node is contained in exactly two of the edges selected. Thus, the problem can be formulated as a 2-matching problem in a graph having $m(m-1)/2$ 0–1 variables, that is, half of the number of the previous formulation. As in the asymmetric case, subtours must be eliminated through subtour elimination constraints. The problem can therefore be formulated as

$$\begin{aligned} \min \quad & (1/2) \sum_{j=1}^{m} \sum_{k \in J_{(j)}} c_k x_k \\ \text{s.t.} \quad & \sum_{k \in J_{(j)}} x_k = 2 \ \text{ for all } j = 1, \ldots, m \\ & \sum_{j \in E(K)} x_j \le |K| - 1 \ \text{ for all } K \subset \{1, \ldots, m\} \\ & x_j = 0 \, or \, 1 \ \text{ for all } j \in E, \end{aligned}$$

where $J(j)$ is the set of all undirected edges connected to node j and $E(K)$ is the subset of all undirected edges

connecting the cities in any proper, nonempty subset K of all cities. Of course, the symmetric problem is a special case of the asymmetric one, but practical experience has shown that algorithms for the asymmetric problem perform, in general, badly on symmetric problems. Thus, the latter need a special formulation and solution treatment. In addition, as an ATSP instance can be easily turned into a symmetric one with twice the number of nodes, any algorithm for STSP can be used to solve an ATSP.

Algorithms

Exact approaches to solving such problems require algorithms that generate both a lower bound and an upper bound on the true minimum value of the problem instance. Any round-trip tour that goes through every city exactly once is a feasible solution with a given cost that cannot be smaller than the minimum cost tour. Algorithms that construct feasible solutions, and thus upper bounds for the optimum value, are called heuristics. These solution strategies produce answers but often without any quality guarantee as to how far off they may be from the optimal answer. Heuristic algorithms that find a feasible solution in a single attempt are called constructive heuristics, while algorithms that iteratively modify and try to improve some given starting solution are called improvement heuristics. When the solution one obtains is dependent on the initial starting point of the algorithm, the same algorithm can be used multiple times from various (random) starting points. Often, if one needs a solution quickly, one may settle for a well-designed heuristic algorithm that has been shown empirically to find near-optimal tours to many TSP problems. Research by Golden and Stewart (1985), Jünger, Reinelt and Rinaldi (1994), Johnson and McGeoch (2002), and Applegate et al. (2006) describes algorithms that find solutions to extremely large TSPs (problems with hundreds of thousands, or even millions of variables) to within 1 or 2% of optimality in very reasonable times. The heuristic algorithm of Lin and Kernighan appears so far to be the most effective in term of solution quality, in particular with the variant proposed by Helsgaun (2000), which was able to find, for the first time, the optimal solution (although without a quality guarantee) of several instances of TSPLIB, a well known library of TSP problems described in Reinelt (1991). For genetic algorithmic approaches to the TSP, see Potvin (1996); for simulated annealing approaches see Aarts, Korst and Laarhoven (1988); for neural net approaches, see Potvin (1993); for tabu search approaches, see Fiechter (1990); and for a very effective evolutionary algorithm, *see* Nagata (2006). Probabilistic analysis of heuristics are discussed in Karp and Steele (1985); performance guarantees for heuristics are given in Johnson and Papadimitriou (1985) and Arora (2002), where an amazing result concerning the polynomial-time approximability is described for Euclidean TSP instances (where the nodes are points in the plane and the traveling costs are the Euclidean distances between the points). For an analysis of the heuristics for the ATSP, see Johnson et al. (2002).

In order to know about the closeness of the upper bound to the optimum value, one must also know a lower bound on the optimum value. If the upper and lower bound coincide, a proof of optimality is achieved. If not, a conservative estimate of the true relative error of the upper bound is provided by the difference of the upper and the lower bound divided by the lower bound. Thus, one needs both upper and lower bounding techniques to find provably optimal solutions to hard combinatorial problems or even to obtain solutions meeting a quality guarantee.

So how does one obtain and improve the lower bound? A relaxation of an optimization problem is another optimization problem whose set of feasible solutions properly contains all feasible solution of the original problem and whose objective function value is less than or equal to the true objective function value for points feasible to the original problem. Thus, the true problem is replaced by one with a larger feasible region but that is more easily solvable. This relaxation is continually refined so as to tighten the feasible region so that it more closely represents the true problem. The standard technique for obtaining lower bounds on the TSP problem is to use a relaxation that is easier to solve than the original problem. These relaxations can have either discrete or continuous feasible sets. Several relaxations have been considered for the TSP. Among them are the n-path relaxation, the assignment relaxation, the 2-matching relaxation, the 1-tree relaxation, and the linear programming relaxation. For randomly generated asymmetric TSPs, problems having up to 7,500 cities

have been solved, in the early 1990s, using an assignment relaxation which adds subtours within a branch and bound framework and which uses an upper bounding heuristic based on subtour patching, (Miller and Pekny 1991). For the symmetric TSP, the 1-tree relaxation and the 2-matching relaxations have been most successful. These relaxations have been embedded into a branch-and-bound framework.

The process of finding constraints that are violated by a given relaxation is called a cutting plane technique and all successes for large TSP problems have used cutting planes to continuously tighten the formulation of the problem. To obtain a tight relaxation the inequalities utilized as cutting planes in many computational approaches to the TSP are often facet-defining inequalities.

One of the simplest classes of cuts that have been shown to define facets of the underlying TSP polytope is the subtour elimination cut. Besides these constraints, comb inequalities, clique tree inequalities, path, wheelbarrow and bicycle inequalities, ladder, crown, domino and many other inequalities have also been shown to define facets of this polytope. The underlying theory of facet generation for the symmetric traveling salesman problem is provided in Grötschel and Padberg (1985), Jünger, Reinelt and Rinaldi (1994) *and* Naddef (2002); analogous results for the ATSP polytope are provided in Balas and Fischetti (2002). The algorithmic descriptions of how these inequalities are used in cutting plane approaches are discussed in Padberg and Rinaldi (1991), in Jünger, Reinelt and Rinaldi (1994), *and in* Applegate et al. (2006) where it is also shown how the polynomial-time equivalence between optimization and separation can be turned into a powerful algorithmic tool to generate inequalities not necessarily belonging to one of the known types.

Cutting plane procedures can then be embedded into a tree search in an algorithmic framework referred to as branch and cut and proposed in Padberg and Rinaldi (1991), where it is shown how such approach made it possible to solve some still unsolved instances of sizes up to 2,392 nodes. Some of the largest TSP problems solved have used parallel processing to assist in the search for optimality. This is the case of the software Concorde, where all the known algorithmic ideas for the TSP (and many new ones) have been carefully implemented. With this code, Applegate et al. (2006) managed to solve all problems of the TSPLIB to optimality; for the largest one, of 85,900 nodes, they used 96 workstations for a total of 139 years of CPU time.

As understanding of the underlying mathematical structure of the TSP problem improves, and with the continuing advancement in computer technology, it is likely that many difficult and important combinatorial optimization problems will be solved using a combination of cutting plane generation procedures, heuristics, variable fixing through logical implications and reduced costs, and tree search.

Applications

One might ask, however, whether the TSP problem is important enough to have received all of the attention it has. Much of the attention that the problem has received is because it is a relatively simple problem to describe and yet a difficult (from a complexity viewpoint) optimization problem to solve. However, there are important cases of practical problems that can be formulated as TSP problems and many other problems are generalizations of this problem. Besides the drilling of printed circuits boards described above, problems having the TSP structure occur in the analysis of the structure of crystals (Bland and Shallcross 1987), in the overhauling of gas turbine engines (Pante et al. 1987), in material handling in a warehouse (Ratliff and Rosenthal 1981), in cutting stock problems (Garfinkel, 1977), in the clustering of data arrays (Lenstra and Rinooy Kan 1975), in the sequencing of jobs on a single machine (Gilmore and Gomory 1964), in the assignment of routes for planes of a specified fleet (Boland et al. 1994) and in genome sequencing (Ben-Dor and Chor 1997; Ben-Dor et al. 2000). Related variations on the traveling salesman problem include the resource-constrained traveling salesman problem, which has applications in scheduling with an aggregate deadline (Pekny and Miller 1991). This paper also shows how the prize collecting traveling salesman problem (Balas 2002) and the orienteering problem (Golden et al. 1987; Fischetti et al. 2002) are special cases of the resource constrained TSP. Most importantly, the traveling salesman problem often comes up as a subproblem in more complex combinatorial problems, perhaps

the best-known application being the vehicle routing problem. This is the problem of determining for a fleet of vehicles which customers should be served by each vehicle and in what order each vehicle should visit the customers assigned to it. For relevant surveys, see Christofides (1985), Fisher (1987), and the book *The Vehicle Routing Problem*, edited by Toth and Vigo (2001).

Concluding Remarks

The seminal paper on the TSP is Dantzig, Fulkerson and Johnson (1954). Books by Lawler et al. (1985), Reinelt (1994) and Gutin and Punnen (2002), and the survey and annotated bibliography by Jünger, Reinelt and Rinaldi (1994, 1997), summarize most of the research up through 2002 and provide extensive references. For a deep understanding of how algorithms for TSP work, see the book by Applegate et al. (2006), which besides providing a wide overview on TSP history and on its applications, also gives a detailed description of how all the components of the Concorde software are built: a valuable source for algorithm designers. Finally, the book by Cook (2011) is for a more general audience, requiring almost no mathematical background to read, but very nicely and completely describing the TSP from several interesting viewpoints. The computer program Concorde, the TSPLIB, and many other sources of information on the TSP are available electronically at a Web site that can be easily located through Web search.

See

- ▶ Assignment Problem
- ▶ Branch and Bound
- ▶ Chinese Postman Problem
- ▶ Integer and Combinatorial Optimization
- ▶ Combinatorics
- ▶ Computational Complexity
- ▶ Graph Theory
- ▶ Heuristics
- ▶ Linear Programming
- ▶ Network
- ▶ *NP*, *NP*-Complete, *NP*-Hard
- ▶ Tabu Search

References

Aarts, E. H. L., Korst, J. H. M., & Laarhoven, P. J. M. (1988). A quantitative analysis of the simulated annealing algorithm: A case study for the traveling salesman problem. *Journal of Statistical Physics, 50*, 189–206.

Applegate, D., Bixby, R. E., Chvátal, V., & Cook, W. (1995). *Finding cuts in the TSP(A preliminary report) DIMACS* (Technical Report 95–05). New Brunswick, USA: Rutgers University.

Applegate, D., Bixby, R. E., Chvátal, V., & Cook, W. (2006). *The traveling salesman problem: A computational study*. Princeton: Princeton University Press.

Arora, S. (2002). Approximation algorithms for geometric TSP. In G. Gutin & A. P. Punnen (Eds.), *The traveling salesman problem and its variations* (pp. 207–222). Dordrecht, The Netherlands: Kluwer.

Balas, E. (2002). The prize collecting traveling salesman problem and its applications. In G. Gutin & A. P. Punnen (Eds.), *The traveling salesman problem and its variations* (pp. 663–696). Dordrecht, The Netherlands: Kluwer.

Balas, E., & Fischetti, M. (2002). Polyhedral theory for the asymmetric traveling salesman problem. In G. Gutin & A. P. Punnen (Eds.), *The traveling salesman problem and its variations* (pp. 117–168). Dordrecht, The Netherlands: Kluwer.

Ben-Dor, A., & Chor, B. (1997). On constructing radiation hybrid maps. *Journal of Computational Biology, 4*, 517–533.

Ben-Dor, A., Chor, B., & Pelleg, D. (2000). RHO-radiation hybrid ordering. *Genome Research, 10*, 365–378.

Bland, R. E., & Shallcross, D. F. (1987). *Large traveling salesman problem arising from experiments in X-ray crystallography: A preliminary report on computation* (Technical Report No. 730). Ithaca, New York: School of OR/ IE, Cornell University.

Burkard, R. E., Deineko, V. G., van Dal, R., van der Veen, J. A. A., & Woeginger, G. J. (1998). Well-solvable cases of the traveling salesman problem: A survey. *SIAM Review, 40*, 496–546.

Cook, W. (2011). *In pursuit of the salesman: Mathematics at the limits of computation*. Princeton: Princeton University Press.

Dantzig, G. B., Fulkerson, D. R., & Johnson, S. M. (1954). Solution of a large-scale traveling salesman problem. *Operations Research, 2*, 393–410.

Fiechter, C. N. (1990). *A parallel tabu search algorithm for large scale traveling salesman problems* (Working Paper 90/1). Switzerland: Department of Mathematics, Ecole Polytechnique Federale de Lausanne.

Fisher, M. L. (1988). Lagrangian optimization algorithms for vehicle routing problems. In G. K. Rand (Ed.), *Operational research '87*, pp. 635–649.

Golden, B. L., Levy, L., & Vohra, R. (1987). The orienteering problem. *Naval Research Logistics, 34*, 307–318.

Golden, B. L., & Stewart, W. R. (1985). Empirical analysis of heuristics. In E. L. Lawler, J. K. Lenstra, A. H. G. Rinooy Kan, & D. B. Shmoys (Eds.), *The traveling salesman problem* (pp. 207–250). Chichester: John Wiley.

Grötschel, M., & Holland, O. (1991). Solution of large scale symmetric traveling salesman problems. *Mathematical Programming, 51*, 141–202.

Grötschel, M., & Padberg, M. W. (1985). Polyhedral theory. In E. L. Lawler, J. K. Lenstra, A. H. G. Rinooy Kan, & D. B. Shmoys (Eds.), *The traveling salesman problem* (pp. 251–306). Chichester: John Wiley.

Gutin, G., & Punnen, A. P. (Eds.). (2002). *The traveling salesman problem and its variations*. Dordrecht, The Netherlands: Kluwer.

Helsgun, K. (2000). An effective implementation of the Lin-Kernighan traveling salesman heuristic. *European Journal of Operational Research, 126*, 106–130.

Hoffman, A. J., & Wolfe, P. (1985). History. In E. L. Lawler, J. K. Lenstra, A. H. G. Rinooy Kan, & D. B. Shmoys (Eds.), *The traveling salesman problem* (pp. 1–16). Chichester: John Wiley.

Johnson, D. S., & McGeoch, L. A. (2002). Experimental analysis of heuristics for the STSP. In G. Gutin & A. P. Punnen (Eds.), *The traveling salesman problem and its variations* (pp. 369–444). Dordrecht: Kluwer.

Johnson, D. S., & Papadimitriou, C. H. (1985). Performance guarantees for heuristics. In E. L. Lawler, J. K. Lenstra, A. H. G. Rinooy Kan, & D. B. Shmoys (Eds.), *The traveling salesman problem* (pp. 145–180). Chichester: John Wiley.

Johnson, D. S., Gutin, G., McGeoch, L. A., Yeo, A., Zhang, W., & Zverovitch, A. (2002). Experimental analysis of heuristics for the ATSP. In G. Gutin & A. P. Punnen (Eds.), *The traveling salesman problem and its variations* (pp. 485–488). Dordrecht, The Netherlands: Kluwer.

Jünger, M., Reinelt, G., & Rinaldi, G. (1994). The traveling salesman problem. In M. Ball, T. Magnanti, C. Monma, & G. Nemhauser (Eds.), *Handbook on operations research and the management sciences* (pp. 225–330). Amsterdam: North Holland.

Jünger, M., Reinelt, G., & Rinaldi, G. (1997). The traveling salesman problem. In M. Dell'Amico, F. Maffioli, & S. Martello (Eds.), *Annotated bibliographies in combinatorial optimization* (pp. 199–221). New York: Wiley.

Karp, R., & Steele, J. M. (1985). Probabilistic analysis of heuristics. In E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy Kan, & D. B. Shmoys (Eds.), *The traveling salesman problem* (pp. 181–205). Chichester: John Wiley.

Lawler, E. L., Lenstra, J. K., Rinnooy Kan, A. H. G., & Shmoys, D. B. (Eds.). (1985). *The traveling salesman problem*. Chichester, UK: John Wiley.

Miller, D., & Pekny, J. (1991). Exact solution of large asymmetric traveling salesman problems. *Science, 251*, 754–761.

Naddef, D. (2002). Polyhedral theory and branch-and-cut algorithm for the symmetric TSP. In G. Gutin & A. P. Punnen (Eds.), *The traveling salesman problem and its variations* (pp. 21–116). Dordrecht, The Netherlands: Kluwer.

Nagata, Y. (2006). New EAX crossover for large TSP instances. *Lecture Notes in Computer Science, 4193*, 372–381.

Padberg, M. W., & Grötschel, M. (1985). Polyhedral computations. In E. L. Lawler, J. K. Lenstra, A. H. G. Rinooy Kan, & D. B. Shmoys (Eds.), *The traveling salesman problem* (pp. 307–360). Chichester: John Wiley.

Padberg, M. W., & Rinaldi, G. (1991). A branch and cut algorithm for the resolution of large-scale symmetric traveling salesmen problems. *SIAM Review, 33*, 60–100.

Potvin, J. V. (1993). The traveling salesman problem: A neural network perspective. *INFORMS Journal on Computing, 5*, 328–348.

Potvin, J. V. (1996). Genetic algorithms for the traveling salesman problem. *Annals of Operations Research, 63*, 339–370.

Ratliff, H. D., & Rosenthal, A. S. (1981). *Order-picking in a rectangular warehouse: A solvable case for the traveling salesman problem* (PDRC Report Series No. 81–10). Atlanta: Georgia Institute of Technology.

Reinelt, G. (1991). TSPLIB–A traveling salesman library. *ORSA Journal on Computing, 3*, 376–384.

Reinelt, G. (1994). *The traveling salesman: Computational solutions for TSP applications*. Berlin: Springer-Verlag.

Toth, P., & Vigo, D. (2001). *The vehicle routing problem*. Philadelphia: SIAM.

Tree

In a network, a tree is a subnetwork (graph) that has no cycles and connects all nodes of a subnetwork, that is, a unique path exists between each node. A tree that connects all n nodes of a network is called a spanning tree and has $(n - 1)$ arcs.

See

► Minimum Spanning Tree Problem
► Network Optimization

Triangular Matrix

A square matrix $\boldsymbol{A} = (a_{ij})$ such that either all the elements a_{ij} above the diagonal are 0 or all the elements below the diagonal are 0. The former is called a lower triangular matrix and the latter an upper triangular matrix.

Trim Problem

Problem of determining how rolls or sheets of material should be cut to minimize the amount of wasted material (trim) while meeting the demand for different sizes of cuts. The problem originally arose in the context of cutting large rolls of newsprint into desired smaller sizes. The trim problem can be formulated and solved as a linear or

integer program. It was the problem that motivated column generation procedures.

See

▶ Column Generation
▶ Cutting Stock Problems

Trivial Solution

For the homogeneous linear equations $\boldsymbol{Ax} = 0$, the solution $\boldsymbol{x} = 0$ is called a trivial solution.

See

▶ Nontrivial Solution
▶ Null Space

Truck Dispatching

The dynamic assignment of trucks (drivers) to loads and/or customers.

See

▶ Logistics and Supply Chain Management
▶ Vehicle Routing

Truckload (TL) Shipment

A shipment weighing at least the minimum weight to qualify for a TL-size rate reduction.

See

▶ Logistics and Supply Chain Management

TS

▶ Tabu Search

TSP

▶ Traveling Salesman Problem

Tucker Tableau

A reduced simplex tableau of a linear-programming problem that considers the tableau as representation of both the primal and dual problems.

Two-Phase Simplex Method

Any version of the simplex method that requires the finding of a first basic feasible solution using artificial variables (Phase I) and then the finding of an optimal feasible solution (Phase II).

See

▶ Artificial Variables
▶ Phase I Procedure
▶ Phase II Procedure