Chapter 2

LITERATURE REVIEW

2.1 Introduction

This chapter provides a brief review of the extensive literature that exists in the area of network optimization problems, as relevant to our study. In the area of network routing, a review of time-dependent shortest path problems is provided first, followed by a review of hazmat routing problems. Subsequently, we focus on network distribution and location, and discuss emergency response problems and related location-allocation models on networks. Thereafter, in the area of network design, a review of existing approaches to water distribution system design is described. Finally, we examine continuous location-allocation problems such as the rectilinear distance, Euclidean distance, and l_p distance location-allocation problems, and also provide a brief review of the pure location counterparts of these location-allocation problems.

2.2 The Time-Dependent Shortest Path Problem.

Deterministic, time-dependent shortest path (TDSP) problems have been widely studied for the case of determining a single shortest path (TD-1SP). Cook and Halsey (1958) have extended Bellman's principle of optimality for dynamic programming (1958) to this case and Dreyfus (1969) has suggested the use of Dijkstra's algorithm (1959) for determining time-dependent shortest paths. Halpern (1977) first noted the limitations of the approach of Dreyfus (1969), and showed that if there exists a $y > 0$ such that $y+d_{ij}(t+y) < d_{ij}(t)$, where $d_{ij}(t)$ is the travel time on arc (i, j) as a function of time t, then the departure

from node *i* must be delayed, or the optimal path might include cycles. Kaufman and Smith (1990) subsequently studied the assumptions under which the existing TDSP algorithms would work, and showed that if the link-delays follow the first-in-first-out (FIFO) rule or consistency assumption, then one could use an expanded static (time-space) network to obtain optimal paths. Orda and Rom (1990), on the other hand, studied various types of waiting-at-nodes scenarios, and proposed algorithms for these different cases. They showed that if waiting is allowed at nodes, then the consistency assumption is not required, and they prescribed an algorithm for identifying optimal waiting times at the source node if waiting is not allowed elsewhere in the network. Furthermore, they demonstrated that for the forbidden waiting case, the paths obtained without the consistency assumption may not be simple, and showed that the continuous-time version of the problem is NP-Hard (1989). Malandraki (1993) analyzed the TDSP problem and extended Halpern's result for the special case of differentiable link delay functions and showed that the consistency assumption would be satisfied by verifying that the first derivative of the link delay function did not exceed negative unity. Friesz et al. (1986) analyzed the source waiting case in the context of the traffic equilibrium problem. They postulated that route choice is not independent of departure times and they calculated optimal source waiting times based on minimizing the total cost to the user. Koutsopoulos and Xu (1994) noted the need for realistic traffic link delay functions in order for TDSP algorithms to be effective and prescribed the use of a link-delay function that gives weightage for both real-time and historic delay components.

Ziliaskopoulos and Mahmassani (1993) provided an efficient solution approach to the problem by discretizing time into τ time periods and developed a pseudo-polynomial

time algorithm for the all source - single sink case having a complexity of $O(m^3\tau^2)$, for a network having m source nodes. This algorithm can identify least cost shortest paths in networks having general link travel costs that do not necessarily satisfy the FIFO rule, for a given set of starting times (multi-states). The authors have also developed an efficient implementation procedure for Intelligent Transportation Systems applications (Ziliaskopoulos et al., 1997). Mahmassani and Ziliaskopoulos (1996) noted that turning movements of vehicles in congested urban networks contribute significantly to the travel time. The authors have prescribed an efficient label-correcting procedure that uses an extended forward-star structure to represent the network including intersection movements and movement prohibitions. Miller (1994) noted the time-dependent nature of risk associated with links in a transportation networks and proposed the application of time-dependent shortest paths to finding dynamic, minimum risk routes for the shipment of hazardous materials. Cai et al. (1997) have analyzed the problem of finding a least cost path on a network with time-dependent delays and costs such that the total delay is less than or equal to a pre-specified value. They have also studied various waiting models, similar to the considerations of Orda and Rom (1990). Chen and Tang (1998) have analyzed a shortest path problem on a mixed-schedule network, subject to side constraints. This network has a subset of nodes that admit waiting and a discrete, finite set of feasible arrival times. Note that such a problem can be transformed to the TDSP case (subject to side constraints) by suitably redefining the link delays for the network based on the restricted set of arrival times on the mixed schedule network. Haquari and Dejax (1997) have analyzed a similar problem, considering time-varying costs and knapsack-like constraints. However, no deterministic TDSP algorithm has been reported in the literature

that prescribes an effective implementation for cases other than the single shortest path problem.

The more general case of the stochastic, time-dependent shortest path problem was first studied by Hall (1986). He showed that the replacement of the probability distribution for link delays by their expected values would yield sub-optimal results and prescribed a dynamic programming algorithm to solve the problem using conditional probability theory. Kaufman and Smith (1990) subsequently showed that the time-space network formulation and expected link delays could be used to solve the problem if the consistency assumption is satisfied. Miller et al. (1994) have prescribed an efficient label-correcting algorithm to obtain Pareto-optimal paths by discretizing the probability distribution of the link delays.

The *k*-shortest disjoint paths problem for the case of $k = 2$, and for constant delays is called the shortest pairs of arc-disjoint paths (SPDP) problem. (This is the static counterpart of the time-dependent TD-2SP problem.) Specialized polynomial-time algorithms for this problem (SPDP) were first developed based on the fact that the problem reduces to finding shortest augmenting paths in a specially constructed residual network. For a network having n arcs and m nodes, Suurballe (1974) presented an algorithm for the single source - single sink case having a time complexity of $O(n \log_{(1+n/m)} m)$. This procedure solved the problem as a special case of a minimum cost network flow problem using two efficient implementations (Tarjan, 1984) of Dijkstra's single source shortest path algorithm. Suurballe (1982) also developed an algorithm to solve the SPDP problem for the single source - all sinks node-disjoint case, by effectively combining the SPDP calculations for various sinks, and obtained an algorithm having a

time complexity of $O(m^2 \log_2 m)$ for an m sink problem. An efficient implementation of this algorithm is given in Suurballe and Tarjan (1984). Here, the disjoint pairs of paths from the origin node to all the other nodes in the network are determined using a single Dijkstra-like calculation to derive an algorithm having a time complexity of $O(n \log_{(1+n/m)} m)$. Orda and Rom (1988) expounded on the use of node-disjoint paths to route duplicate packets in computer networks in order to achieve a more balanced utilization of network resources. They also prescribed several algorithms that can be used to generate such sets of disjoint paths. An algorithm that solves the SPDP problem for the all sources - single sink case has been developed by Ogier et al. (1993). This method makes use of a modified distributed Bellman-Ford type of a procedure. Bhandari (1994) has studied the SPDP problem in the context of network planning for telecommunication fiber networks. Chen and Tang (1995) have studied the problem of routing aircraft in a three-dimensional space around an airport to minimize noise exposure to the population. They proposed using a dynamic network flow model (Ford and Fulkerson, 1958; Ahuja et al., 1993) to generate a time-space network and to find a given number of time-disjoint paths on this network such that the total cost is minimized.

2.3 Minimum Risk Routing of Hazardous Materials on Transportation Networks

Research has primarily addressed route-finding techniques that minimize either the total travel time (Brogan and Cashwell, 1985), the expected number of accidents (fatal or otherwise), the accident probability, the residential population within a given distance from the route (Glickman, 1983), the risk of spill (Patel and Horowitz, 1994), or some combination of these factors. This research has been motivated by Talcott (1992), and

surveyed by List et al. (1991), and Erkut and Verter (1995). It is immediately apparent that such single objective models cannot take into account conflicting criteria such as truck operating costs and expected damage. Abkowitz and Cheng (1988), Batta and Chiu (1988), Gopalan et al. (1990), Erkut and Verter (1994, 1997) among others have considered such problems. Zografos and Davis (1989) note that uncapacitated formulations of the hazmat routing problem may fail to capture the objective of the equitable distribution of risk and proposed a capacitated, multi-objective mathematical programming model for the systemwide routing of hazmat. List et al. (1991) have presented a comprehensive survey of such routing problems. Helander and Melachrinoudis (1997) and List et al. (1991) have pointed out that it is difficult to separate siting from routing decisions since location implies the selection of routes and conversely, suitable paths are required for evaluating locations. However, only a few analytical models have been devoted to this important problem. Erkut and Verter (1995) and List et al. (1991) have presented a survey of such models. The earliest model was due to Shobrys (1981) who solved the routing (shortest path) problem and siting (p-median) problem separately. Zografos and Samara (1990) have developed a combined location and routing goal programming model for hazmat transportation and disposal that minimizes travel time, routing risk, and disposal risk, while Helander and Melachrinoudis (1997) have proposed integrated models that minimize the expected number of accidents.

An important consideration in hazmat routing is that, although a route may have a very low probability of an accident occurring, or a low associated expected consequence, if the potential consequence that is incurred given that an accident occurs is high, then we may not be wise in choosing such a route. This viewpoint is shared by several authors who

point out that decision-making which affects the public's safety should not be based on traditional mathematical expectation that falls short of incorporating extreme events such as low probability-high consequence (LPHC) situations, where the latter remain concealed in the analysis. Karlsson and Haimes (1988 a, b) have embedded the statistics of extremes within their Partitioned Multiobjective Risk Method (PMRM). Glickman (1983) also considered the tradeoff between expected values and the extreme values of risk in deciding between alternative routings of hazmat. A related network-based model for transporting hazmat developed by Sivakumar et al. (1993) determines a path that minimizes the risk, given that an accident occurs. The model continues routing shipments on this particular path until the first accident occurs. An alternative model is suggested in this paper that adds a constraint on the maximum permissible accident probability of a selected path. For both these models, various heuristic procedures are developed and tested. Sivakumar and Batta (1994) formulated and solved a variance-constrained shortest path problem. This model finds application in obtaining a safe route for the transportation of liquefied gas, a hazmat whose dispersal rate and extent is highly unpredictable. Interestingly, the solution procedure involves the use of a linearization procedure that can be directly obtained using the Reformulation Linearization Technique (RLT) of Sherali and Adams (1990) and Sherali and Tuncbilek (1992), which is also used in this dissertation study. In another approach, Sivakumar et al. (1995) used an objective function that minimizes the expected risk of the first accident, and incorporated additional side-constraints of the type we include herein, while also considering the equity of risk among the different zones of the geographical region being studied. Jin et al. (1996) and Jin and Batta (1995) have treated various other objective functions based on minimizing the expected consequence given a

number of trips to be made, and given a threshold on the related number of accidents tolerated before shipments cease. Boffey and Karkazis (1995) have showed that nonlinear risk minimization models are more appropriate for very high risk routing situations. The present study also treats LPHC events, and is a more complete study of a model described in a conference proceedings by Glickman and Sherali (1991). The hazmat shipment problem is formulated as a network optimization model that minimizes the conditional expectation of a catastrophic outcome, i.e., the expected consequence given that a catastrophic accident has occurred, subject to the side-constraints that the expected value of the consequence (expected risk) is lesser than or equal to a predetermined value *v*, and that the probability of an accident on the selected path is no more than some specified value η. The need for such side-constraints is motivated by Glickman and Sherali (1991), and also by several insightful examples given by Erkut (1995), where it is shown that a simple minimization of the conditional objective function can lead to illogical path choices. Recently, Erkut and Ingolfsson (1998) noted that the FHWA definition of risk (FHWA Handbook, 1994) ignores the risk averse attitudes of many decision-makers when dealing with LPHC events and have proposed several catastrophe avoidance models for hazmat route planning.

2.4 Distribution, Location-Allocation and Simulation Models on Networks for Emergency Response and Risk Management

Traffic incidents annually account for nearly sixty percent of the delay (in vehicle-hours) on the roads. These incidents are non-recurrent random events that cause disruptions and reductions in road capacities.

Church and Roberts (1983) have presented a model that addresses the correlation of quality of service and response time. In general, the benefit of an emergency response decreases with the incident response time. A cover story in the *Engineering News Record* (October 30, 1995) that features the newly constructed 217 million dollar emergency response center for the city of Chicago, Illinois, lends further credence to this belief. The new emergency response center houses an advanced fiber-optic based digital communication network that helps reduce the connection time from 4-6 seconds to 1.2 seconds. The emergency response personnel were able to answer distress calls within 10 seconds at least 99% of time, a significant improvement over the previous service performance of answering calls within 12 seconds at least 70% of the time. An important goal of the response center was to reduce the arrival time (currently five to ten minutes) of response vehicles at the scene by at least two minutes. Van Aerde et al. (1995) developed and tested the popular simulation model INTEGRATION to evaluate the operations of traffic signal networks in the *Burlington Skyway Freeway Traffic Management System*. They concluded from empirical simulation tests that effective and quick incident response strategies can significantly reduce traffic delay. Furthermore, the authors observed that providing real-time information to vehicles during incidents has a similar impact on traffic delay. These researchers have made an observation in their study of emergency vehicle routing that such response vehicles tend to "amplify" the stochastic nature of the network, and that current travel data rather than average data is more useful (Marcus and Krechmer, 1995; Zhang and Ritchie, 1994). Substantial attention has been paid in the past to the reduction of the incident detection time, and in predicting the expected time of clearing an incident, but there has been relatively little research effort devoted to

minimizing the traffic incident response time. On the other hand, considerable attention has been devoted to the problem of determining optimal location and dispatching strategies that minimize expected response times for various other critical situations such as fires and medical emergencies. Some of these approaches are discussed below.

Most methodologies in the literature for minimizing emergency response times utilize location-allocation models. There has been an enormous amount of research work carried out in the field of location-allocation modeling since the formulation of the first location-allocation problem by Cooper in 1963. The simplest location-allocation problem is the Weber problem addressed by Friedrich in 1929, which involves locating a production center so as to minimize the aggregate weighted distance from various raw material sources. The seminal work in this area was on the *p*-median problem, initially formulated by Hakimi (1964, 1965). The *p*-median model has been used to solve service facility location problems that are analogous to the location of switching centers in electrical networks. Given a graph $G(N, A)$, the *p*-median problem requires us to locate a set of p supply points in G, so that the total (or average) transportation cost of the supply from the nearest supply point to the demand nodes (having known demands) that are located on the nodes (N) of G is minimized. Hakimi showed that at least one such optimal solution locates the supply points on the nodes of the network. This result helps us to restrict our search for the optimal locations to the nodes N of G. For $p > 1$, this problem can be viewed as a location-allocation problem, since the location of the facilities governs its allocation of services to meet nodal demands. Hakimi also defined a set of m points X^*_m in a deterministic undirected graph G as a set of *absolute m-medians* if for every $X_m \subseteq G$, the sum of the weighted response times from X_m to the demand nodes is greater than or

equal to the sum of the weighted response times from X^*_m. He showed that when the utility attributes for travel costs are convex, there exists at least one *m*-node subset of *G* which is a set of absolute *m*-medians.

Torgeas et al. (1971) proposed a set covering model to locate emergency service facilities. One problem with this formulation is that the optimal solution may yield a very high value for the number of supply depots to be located. Also, the model does not assign any priority to the location of supply depots at pre-existing facilities. Toward this end, Hendrick et al. (1974) provided a hierarchical model whose objective function is formulated such that facilities are assigned to pre-existing sites, provided their use does not increase the value of the objective function value of the set covering problem. An alternative model for locating emergency vehicle depots is the maximal covering location problem (Eaton et al., 1985) that has been used to locate medical rescue vehicles in the city of Austin, Texas. This model attempts to provide maximum response coverage to the network subject to a constraint on the maximum number of response vehicles available.

Handler and Mirchandani (1979) have analyzed the case of solving location-allocation problems on probabilistic networks. The formulation here considers the travel times to be stochastic. The authors have shown that for convex travel cost functions, Hakimi's result would continue to hold true, but that the optimal solution obtained by replacing the travel cost functions by their expected values and solving the resultant deterministic *p*-median problem would not be the same solving the stochastic case. Daskin (1983) proposed the use of a maximal expected covering location model for locating emergency response vehicles based on the idea that not all vehicles allocated to serve a particular zone in the network would actually be available during times of emergency.

Cavalier and Sherali (1985) have analyzed static and sequential location-allocation problems on undirected networks in which the demands can occur on links having uniform probability distributions. The authors showed that except for the 1-median case, the problem is generally nonconvex.

Dynamic location problems arise in situations where current supply centers may have to be re-evaluated with time due to time-varying characteristics of the network. Orda and Rom (1990) have prescribed an algorithm that myopically determines the location sequence over a given horizon by solving 1-median problems on a network having time-varying edge weights, given a starting optimal location. Cavalier and Sherali (1983) have studied sequential location problems on chain tree graphs based on myopic, long-range, and discounted present worth policies. Using a Markovian assumption to describe the dynamic nature of the network, Berman (1977) developed a model for the repositioning of emergency service vehicles, in order to minimize long-run expected costs and showed that at least one set of optimal locations of m units exists on the nodes of the network.

Several computational complexity results for the p-median and p-center problems are available in the literature. A survey of the p-median and p-center problems is presented in Tansel et al. (1983). Hakimi et al. (1978) showed that Hakimi's method (1964) for finding the absolute 1-center (Hakimi, 1965) can be implemented in $O(|E|n^2 \log n)$ effort, and prescribe a computational refinement which reduces the effort to $O(|E|n \log n)$ for the unweighted case. Further refinements were obtained by Kariv and Hakimi (1979), resulting in an $O(|E|n \log n)$ algorithm for the weighted case and to $O(|E|n)$ for the unweighted case, where $|E|$ is the number of arcs, and n is the number of nodes in the network, respectively. Minieka (1981) developed an $O(n^3)$ algorithm for solving the

unweighted 1-center problem. Kariv and Hakimi (1979) showed that the *p*-center and *p*-median problems on a general network are NP-Hard. They also showed that the weighted case can be reduced to a computationally finite one, and presented an algorithm whose complexity is $O(|E|^p(n^{2p-1})\log n/(p\text{-}1)!)$. In addition, they prescribed an algorithm for the unweighted case whose complexity is $O(|E|^p(n^{2p-1})/(p\text{-}1)!)$. Hakimi (1964) solved the absolute 1-median problem for a general network by prescribing an $O(n^3)$ algorithm. In the special case of tree-graphs, Goldman (1972) solved the unweighted 1-center problem, based on the repeated application of the "trichotomy theorem" that either determines the edge on which the absolute center lies, or reduces the search to the two subtrees obtained by removing all interior points on that edge. Halfin (1974) refined Goldman's algorithm, improving it from $O(n^2)$ to $O(n)$. An algorithm of complexity $O(n^2 \log n)$ is described by Kariv and Hakimi (1979) for finding the absolute *p*-center of a vertex weighted tree network. These complexity results have been improved since, and several algorithms have been presented in the literature having comparable orders of complexity. An efficient $O(n)$ "tree-trimming" algorithm to find the 1-median was presented by Hua Lo-Keng et al. (1962). Matula and Kolde (1976) suggested an $O(n^3p^2)$ algorithm for finding the *p*-median on a tree network. Kariv and Hakimi (1979) proposed an $O(n^2p^2)$ algorithm for the same problem.

In the specific case of traffic incident response models, Zografos et al. (1993) used a districting model to obtain optimal locations of vehicles that minimize the total average incident response workload per vehicle on freeways, subject to a constraint on the maximum number of available vehicles. Daskin (1987) constructed a mixed-integer programming (MIP) model for the simultaneous location, dispatching, and routing of

incident response vehicles. Pal and Sinha (1997) also used an MIP model to determine optimal locations for response vehicles that minimizes annual response vehicle costs, given the frequencies of incidents at potential sites in the network, and subject to a constraint on the maximum number of vehicles. A stochastic emergency response model presented by Jarvis (1975) employs a hypercube model (Larson, 1974) in developing vehicle dispatching strategies and location plans. Among dynamic models for emergency response, the fire-engine relocation model proposed and tested by Kolesar et al. (1974) offers an interesting approach. Upon the occurrence of a fire-related incident, the unassigned vehicles in depots are optimally repositioned to offset the loss in coverage with respect to future fire mishaps due to currently dispatched vehicles. Nathanail and Zografos (1995) have developed a simulation tool for evaluating the effectiveness of freeway incident response operations. In the case where multiple incidents need to be responded to, they created a priority list of incidents based on the nature of the incident and generated response plans based on this priority and under various dispatch policies. The authors did not attempt to prescribe policies that can offset the effect of loss in coverage due to unavailable servers on the system. This study also noted that using multiple emergency vehicles to respond to incidents can significantly reduce response time. Most of these models assume that the closest available vehicle is dispatched to the site of the current incident. Several authors (for example, see Carter et al., 1972) have pointed out the shortcomings accompanying this assumption.

Mirchandani and Odoni (1979) noted that if there is a high probability that response units will not be available, a policy of dispatching the closest response unit may not be optimal for certain performance criteria. The hypercube queuing model proposed

by Larson (1974) estimates the probability that each unit will be busy under various dispatch policies. Daskin and Haghani (1984) demonstrated that for stochastic travel times, dispatching multiple response vehicles using the shortest path from the vehicle depot to the current incident site may not be optimal with respect to minimizing the expected arrival time of the first response vehicle, and they proposed an iterative method to obtain optimal multiple routes. Another important observation on routing was made by Carter et al. (1972). They noted that when service to anticipated future demands is considered, dispatching the nearest available vehicle may not be an optimal choice.

Vehicle inspection is an important tool that can be used to regulate hazardous traffic flow and to minimize the occurrence of hazmat-related incidents. While most of the focus in the literature has been on strategies to optimally respond to such incidents (Sherali et al., 1997), and on the routing and scheduling of hazmat carriers (List et al., 1991), the inspection strategy is based on the premise that it is better to prevent such an accident rather than respond to it. The main task in such a strategy is to decide on the location of the inspection stations, so that a measure of risk associated with the flow of uninspected hazmat vehicles through the network is minimized.

The Inspection Station Location Problem (ISLP) is closely related to a general class of flow-intercepting location-allocation models (Berman et al., 1995). Similar models have been used to locate advertisement billboards with maximum exposure to traffic (Hodgson, 1990) and for locating discretionary service facilities (Berman et al., 1992) to intercept the maximum number of potential customers traveling past them (Mirchandani et al., 1995 and Hodgson et al., 1996).

While the application of the ISLP model to traffic management is a fairly new concept, inspection station location models have received far greater attention in the area of quality control in multi-stage, serial and non-serial production processes. The problem here is to optimally locate product inspection stations and to detect defective parts early in the production process, so as to minimize the total cost per unit produced. Such models are generally stochastic optimization models and represent a more general case of the ISLP model. This is due to the fact that inspected parts can potentially become defective upstream of the inspection station. Efficient dynamic programming procedures have been prescribed for the optimal location of inspection stations in serial production systems, and Raz (1995) has surveyed the use of such models in multi-stage production systems. In general, these models may not be directly applicable to the ISLP problem for transportation networks due to differences in the network topology and the nature of network flows.

2.5 Optimal Design of Water Distribution Systems

This class of problems seeks an optimal design of the links of the water distribution network along with elevated heads at sources so as to satisfy flow requirements at demand nodes within specified ranges of water pressure. For the case of a single network pattern, Swamee et al. (1973) showed that the optimal design is a branched network. However, such a network has no built-in redundancy and could degrade the network performance and reliability. Hence, the focus in the literature has been on networks having cycles built in to introduce redundancy in the design. Comprehensive reviews of work on designing water distribution systems are given in Lansey and Mays (1985) and Sherali and Smith

(1993). The techniques employed range from the traditional nonlinear Hardy-Cross solver and Newton-Raphson methods, to more complicated hierarchical decomposition methods based on linear programming approximations. Initial approaches mainly involved the use of gradient-based search. A Lagrangian penalty function formulation was presented by Jacoby (1968), Watanatada (1973) and Shamir (1974). Bhave (1978) used a nonlinear optimization model with continuous pipe diameters and solved the problem via a procedure wherein the pressure heads are adjusted in an iterative manner. Collins et al. (1978) used linear approximation and projection techniques for nonlinear optimization such as the Frank-Wolfe method and the Convex Simplex Method to solve the problem, but the algorithm proved to be non-robust.

Several decomposition approaches to solve the looped network problem are available in the literature. Rowell and Barnes (1982) developed a method that alternates between solving the pipe diameter problem and the network flow problem separably, while Goulter and Morgan (1983) incorporated a feedback mechanism between these subproblems to reduce the error induced by the process. Bhave (1983) developed a two-stage linear programming approach for solving the multi-source, multi-loop problem. Morgan and Goulter (1985) developed a heuristic procedure to analyze multi-loop systems for various loading scenarios. Taher and Labadie (1996) have developed an integrated GIS and optimization module for solving water distribution system design problems in urban settings. In general, a variety of research efforts over the last two decades have focused on the least cost pipe sizing decision, most of them generating improved local optimum solutions for several standard test problems from the literature, with no adequate lower bounds to evaluate the prescribed solutions. Two notable

exceptions are Eiger et al. (1994) and Sherali and Smith (1995) who developed the first global optimization procedures for a loop and path based formulation, and for an arc-based formulation of the problem, respectively.

The most general water distribution design problem is to build new networks, or to replace old and deteriorating sections of existing networks with new configurations having enhanced capacity, while integrating network reliability and redundancy issues, network expansion and pipe sizing decisions, and multi-period economic analysis, all within a holistic framework. Traditionally, most of the work on the design of water distribution networks has focused on developing optimization procedures for the least cost pipe sizing problem. An exception to this is a paper by Sherali and Smith (1993) in which the authors have developed an integrated pipe-reliability-and-cost, and network optimization approach, that provides replacement recommendations along with the design and sizing of expanded and replaced sections of the network.

One of the most significant approaches for solving the water distribution system design problem has been proposed by Alperovits and Shamir (1977), who developed the successive linear programming gradient (LPG) method. Here, hydraulic loops are formulated for each basic loop in the network and for each path from a source to a demand node. The problem is then projected onto the space of the flow variables. For each (initially assumed) flow distribution, the other decision variables are optimized via a linear program. The gradient of the total cost with respect to changes in the flow distribution is used to modify the flows, and the procedure is repeated until convergence to a local optimum is obtained. Quindry et al. (1979) showed that Alperovits and Shamir had missed certain terms in their gradient expressions. They modified the LPG method to

incorporate these terms and demonstrated an improvement in solving Alperovits and Shamir's test problem using the same starting solution (Quindry, 1981). However, the inclusion of these terms renders the procedure ineffective for large-scale problems. Lansey and Mays (1985) proposed an alternative method to solve the nonlinear single-stage model of the LPGT method. Fujiwara et al. (1987) and Kessler and Shamir (1989) have proposed alternative derivations of the linear programming based gradient expressions along with other algorithmic enhancements to improve the computational efficiency of the LPG approach. Fujiwara and Khang (1990) have also extended the LPG method to generate a sequence of improving local solutions using a two phase decomposition approach. Another type of decomposition approach has been developed by Sherali and Smith (1993), where the problem is projected onto the space of the network design variables (pipe lengths of various diameters and source elevation heads), and an auxiliary convex cost network flow subproblem of the type analyzed by Collins et al. (1978) is used to guide the variations in the design variables.

A first global optimization approach to the least cost pipe sizing decision was proposed by Eiger et al. (1994). This procedure enforces hydraulic consistency requirements via an enumeration of all possible basic loops and source-to-demand node paths in the network, as opposed to a link-wise formulation of these constraints. A branch-and-bound algorithm is developed based on partitioning the hyperrectangle restricting the flows into several subrectangles. At each node of the branch-and-bound tree, a subgradient-based heuristic is applied to determine an upper bound via the nonsmooth, nonconvex, projection of the problem onto the space of the flow variables. An

independent relaxed, duality-based linear programming formulation is used to compute lower bounds.

Sherali and Smith (1995) presented another global optimization approach that employs a Reformulation-Linearization Technique (RLT) to construct tight linear programming relaxations for the given problem in order to compute lower bounds. The procedure is embedded in a branch-and-bound scheme. Convergence to an optimal solution is induced by coordinating this process with an appropriate partitioning scheme. The authors also proposed several algorithmic refinements and enhancements that need to be made in order to make the algorithm practical and effective for large-scale problems. Several test problems from the literature are solved to exact global optimality for the first time using this approach. In particular, these results indicate that some of the solutions reported by Eiger et al. (1994) are in error due to a degree of infeasibility in the flow conservation constraints. In another global optimization approach, Sherali et al. (1998) transformed the nonlinear, nonconvex network problem into the space of certain design variables. By relaxing the nonlinear constraints in the transformed space via suitable polyhedral outer approximations, the authors derived a linear lower bounding problem that was embedded in a branch-and-bound algorithm. Improved computational results were reported using this approach.

The use of stochastic optimization techniques such as simulated annealing and genetic algorithms have also been employed with considerable success. The least cost pipe design problem is a hard nonconvex optimization problem having a number of local optima, and has hence proven difficult to solve. Randomized enumeration methods such as simulated annealing and genetic algorithms have recently become very popular as

optimization tools for solving complex problems (Frey et al., 1996). Loganathan et al. (1995) used an outer flow search and an inner optimization procedure to identify improved local minima and obtained the best solution to the New York network test problem known prior to the present study. The outer search scheme selects alternative flow configurations to find an optimal flow division among pipes, and an inner linear program is used for the design of least cost pipe diameters. Two search schemes are used to permit a local-optimum-seeking method to migrate among various local minima. Dandy et al. (1996) have developed an enhanced genetic algorithmic approach for the pipe network optimization problem, using an adjacency or creeping mutation operator and gray codes rather than the traditional binary codes. The authors demonstrated a significant improvement in the quality of solution and computational time over previous research conducted by Murphy and Simpson (1992) and Simpson et al. (1994), based on the use of genetic algorithms for pipe network optimization using only bitwise mutation operators. However, as evident from the results in Sherali and Smith (1995) and Sherali et al. (1998), the solutions computed using these methods for some standard test problems are suboptimal.

2.6 Euclidean Distance and l_p Distance Location and Location-Allocation Problems

The Euclidean distance location-allocation (EDLAP) can be decomposed into two distinct problems by fixing certain variables. For a fixed set of allocations, the problem reduces to a pure location problem (see Francis et al., 1991), whereas for a fixed set of locations, the problem reduces to the ordinary transportation/allocation problem (see Bazaraa et al., 1990). As shown in Sherali and Nordai (1988), this class of problems is NP-Hard, even if

all demand points are located on a straight line. Its objective function is nonconvex, resulting in multiple local minima for the problem. Moreover, the objective function is nondifferentiable at points where the location of any source coincides with the location of any destination, and this precludes a direct application of gradient-based algorithms for finding even locally minimizing solutions. However, two known useful properties are that an optimal flow solution occurs at an extreme point of the transportation constraint set *W* (Cooper, 1972), while an optimal set of locations for the sources lies in the convex hull of the locations of the customers (Wendell and Hurter, 1973).

Cooper (1972) was the first to study this class of problems. He proposed an exact solution approach based on a total enumeration of the vertices of *W*, which is prohibitive except for small-sized instances. However, he also suggested a useful heuristic scheme that is quite natural and popular, and is known as the alternating procedure. As its name suggests, this procedure exploits the structure of the problem by alternatively solving the location and the allocation subproblems. Starting with some initial locations for the sources, the resulting transportation problem is solved in order to find the corresponding optimum allocations. With this known allocation, the new optimal locations are determined. This is continued until no further improvement occurs in the objective value. Murtagh and Niwattisyawong (1982) suggested that a local search procedure such as MINOS could be applied to a good starting solution determined in this fashion, using a perturbation technique to overcome the nondifferentiability of the objective function. As an alternative approach, Avriel (1980) proposed a geometric programming technique based on the special structures of this location-allocation problem, and Cavalier and Sherali (1986) have explored heuristics for the case of area demand distributions. Aside

from the total enumeration approach of Cooper, the only other exact procedure developed for Problem EDLAP appears in the (unpublished) dissertation of Selim (1979), where a biconvex programming cutting plane procedure is developed. While this is more tractable than total enumeration, it was found to be effective for problems having only about $m = 5$ customers and $n = 5$ facilities. In fact, as we shall see in the sequel, for the two test problems of size $(m, n) = (5,5)$ solved by Selim, our procedure finds significantly improved solutions while discovering the global optima for these problems.

The uncapacitated version of Problem EDLAP, on the other hand, has received relatively more attention in the literature. Since there is no restriction on the capacity of the sources, each customer is served by a single closest source. Cooper (1964, 1967) has proposed similar heuristic solution procedures for this problem that are based on alternatively solving the location and the allocation subproblems, and Eilon et al. (1971) have suggested some improvements to make these procedures computationally more effective. Following this, Kuenne and Soland (1972) have proposed another heuristic procedure along with a branch-and-bound type of exact solution procedure to optimally solve the problem, where partial solutions are constructed by assigning a subset of destinations to the sources. Similar to the hyberboloid approximation procedure (HAP) of Eyster et al. (1973), Chen (1983) has used a differentiable approximation to the objective function of the problem and solved it using a quasi-Newton based method. However, the resulting solution is not necessarily a global minimum. For $n = 2$, Ostresh (1975) has utilized the property of nonoverlapping convex hulls to find a global minimum (also see Drezner (1984) for a similar technique), and Rosing (1992) has generalized the procedure of Ostresh (1975) to solve the multifacility version of this problem. More recently, Hansen

et al. (1998) have proposed an effective heuristic method based on the analogous *p*-median problem, and Chen et al. (1998) have developed an exact approach using D.C. (difference of convex functions) and concave programming constructs. Impressive computational results are presented for a large number of customers (up to 10, 000) when there exist only two new facilities to be located. However, their procedure was unable to solve some problems having $n = 3$ new facilities and $m = 30$ customers due to memory requirements of their combinatorial, enumerative approach.

Sherali and Tuncbilek (1992a) have solved the capacitated version of the location-allocation problem when the separation penalty is assumed to be proportional to the square of the Euclidean distance. Using calculus, they showed that this problem can be transformed to a quadratic convex maximization problem by projecting it onto the space of allocation variables, and derived a branch-and-bound enumeration algorithm in this allocation space. A variant of Problem EDLAP based on the use of rectilinear distance has been variously treated in the literature. Vaish (1974) showed that this case reduces to solving the bilinear programming problem. Love and Morris (1975) have prescribed an exact solution procedure using a set reduction algorithm for the uncapacitated case, and showed that the problem is equivalent to the *p*-median problem on a weighted connected graph. Sherali and Shetty (1977) have addressed the capacitated case and developed a cutting plane algorithm to solve the problem. This was subsequently improved by Shetty and Sherali (1979) for an even more general version of the problem by providing deeper cutting planes and a better computational implementation. A detailed description of results for different variants of Problem EDLAP can be seen in Sherali et al. (1994) and the references cited therein. While all the foregoing variants offer special structures that make

them more amenable to solution procedures, Problem EDLAP has posed an open challenge due to its difficult nonconvex and nondifferentiable structure.

We also consider in this paper the l_p distance location-allocation problem where the separation penalty in the objective function is based on the l_p distance $\left\{\left|x_i - a_j\right|^p + \left|y_i - b_j\right|^p\right\}^{1/p}$ between source i and customer j, $\forall\,(i, j)$, where (x_i, y_i) is the coordinates of the source i, and (a_j, b_j) is the coordinates of customer j. As shown empirically by Love and Juel (1982), a value of $1 < p < 2$ (i.e., between the rectilinear and the Euclidean distance measures) more accurately represents actual travel distances over road networks.

Several results regarding the convergence of solution procedures for this problem are available in the literature. Kuhn (1973) first proved the global convergence of Wieszfeld's algorithm for case of $p = 2$, under the assumption that none of the iterates in the sequence of points generated by the algorithm coincides with the locations of any of the existing facilities. However, Chandrasekaran and Tamir (1989) showed that convergence may not occur for a continuous set of initial points if the locations are contained within an affine subspace of R^n. Ostresh (1978) suggested a modification to the stepsize chosen when an vertex iterate coincides with an existing facility in order to guarantee global convergence for any set of starting points. Katz (1974) showed that local convergence for Weiszfeld's algorithm is always linear if the optimal location does not coincide with the location of an existing facility, and in the complementary case, it can be quadratic or sublinear under certain conditions. Brimberg and Love (1992) showed that these results can be extended to the generalization of Weiszfeld's procedure for l_p

distances, where $p \in (1, 2)$. Morris (1981) has proven global convergence for the Hyperboloid Approximation Procedure for l_p distances using a smoothing approximation for the absolute value of a real number. Calamai and Conn (1987) have also prescribed an algorithm for the l_p distance location problem. Brimberg and Love (1992a, 1993a) have shown that while the use of a smoothing function eliminates any existing singularities, the final locations generated via such a procedure may not coincide with the true optimal locations. The authors then demonstrated the global convergence for Wieszfeld's procedure for l_p distances where $p \in (1, 2)$ and showed that for any starting point in R^n, the algorithm generates a sequence of points that converges to an unique limit point, and if the sequence is regular, this limit point will always be an optimal solution. However, when $p > 2$, the directions generated by the procedure need not always be descent directions, and consequently, the sequence may have more than one limit point. Based on the empirical studies conducted by Love and Morris (1972, 1979), the use of the l_p norm for estimating travel distances in transportation networks was observed to be statistically superior over the Rectilinear norm and the Euclidean norm. Brimberg and Love (1992b) showed that if the coordinate axes are suitably rotated with respect to the geographical area under study, the value of p used to represent road distances will always lie in (1, 2). This result was subsequently verified in an empirical study conducted by Love and Walker (1993) in which the values of p for eight different countries and nine large urban centers were empirically determined and found to lie in the interval (1, 2).

While the pure location problem has been variously analyzed, there does not exist any algorithm (other than total enumeration) that has been developed for the capacitated location-allocation problem. However, for the uncapacitated version of this location-

allocation problem, Love and Juel (1982) showed that the problem can be equivalently transformed to a concave minimization problem, and accordingly, developed five heuristic strategies that differ from each other in the manner in which they perturb a given local optimum solution. More recently, Bongartz et al. (1994) have developed a local search methodology for the general l_p distance location-allocation problem, where the {0,1} constraints on the allocations are relaxed. The algorithm involves both a step to compute a good starting solution and a specialized linesearch procedure that retains the feasibility of the allocation and recognizes the discontinuity of the first derivative along the search direction. A set of necessary and sufficient conditions for a local minimum of the relaxed problem are then given, which in turns leads to an efficient algorithm involving the use of an active set strategy and orthogonal projections. In this dissertation, we will develop a first global optimization approach for the capacitated general l_p distance location-allocation problem.