Available online at www.sciencedirect.com

ScienceDirect

Applied Mathematical Modelling 31 (2007) 817–831

APPLIED MATHEMATICAL MODELLING

www.elsevier.com/locate/apm

On solving the discrete location problems when the facilities are prone to failure

Shine-Der Lee [*], Wen-Tin Chang

Graduate School of Industrial Management Science, National Cheng Kung University, 1 University Road, Tainan 70101, Taiwan, ROC

Received 1 October 2004; received in revised form 1 March 2005; accepted 14 December 2005
Available online 7 February 2006

Abstract

The classical discrete location problem is extended here, where the candidate facilities are subject to failure. The unreliable location problem is defined by introducing the probability that a facility may become inactive. The formulation and the solution procedure have been motivated by an application to model and solve a large size problem for locating base stations in a cellular communication network. We formulate the unreliable discrete location problems as 0–1 integer programming models, and implement an enhanced dual-based solution method to determine locations of these facilities to minimize the sum of fixed cost and expected operating (transportation) cost. Computational tests of some well-known problems have shown that the heuristic is efficient and effective for solving these unreliable location problems.

Keywords: Discrete location model; Unreliable facilities; Heuristic; Site selection

1. Introduction

Discrete location problems are usually formulated as resource allocation models that are concerned with the selection of facilities to accommodate or serve a given set of customers, so as to optimize the associated fixed and operational cost. The facility can be fire station, emergency shelter, service center, telecommunication post, and logistic center. The facility may provide service to one or several demand points. The classical model assumed that the facility is always available; it will provide service under any circumstance. In reality, these facilities can be unreliable; they will not provide service due to maintenance, capacity limit, breakdown, or shutdown of unknown causes. It is hence of theoretical and practical interests to consider the location problem when facilities are unreliable.

The facilities in the location models are often assumed fully reliable in the literature. Among many approaches for solving these problems, the branch and bound methods [1,2] are widely used for solving the uncapacitated location problems. Schrage [3] developed a tight linear programming formulation of the location problem, which is different from the one used in the earlier models. Later, Erlenkotter [4] developed a

[*] Corresponding author. Tel.: +886 6 275 7575x53146; fax: +886 6 236 2162.
E-mail address: shindlee@mail.ncku.edu.tw (S.-D. Lee).

doi:10.1016/j.apm.2005.12.012

dual-based procedure for solving uncapacitated location problems. The computational test in [4] has shown that the dual-based approach is superior to several other heuristic methods. Brandeau and Chiu [5] gave an updated review of the literature body in location research. Some perspectives are described in [6]. An uncapacitated location problem with primary and secondary service requirements is considered in [7], where the facilities are assumed reliable and each customer requires service from two different sources. At the time of this writing, it appears that there are no formulation and procedure to determine the locations of these discrete facilities when they are unreliable.

On the other hand, when facilities are on the plane, Drezner [8] considered a continuous location problem when the facility is unreliable. More recently, Lee [9] developed efficient procedures to determine the locations of these facilities when they are located on the plane. In parallel with the unreliable planar location problem, variations of the classical p-median problems can be found in [14], where the conditional location problem is addressed. Current et al. [10] considered the dynamic location problem when the total number of facilities is uncertain. The randomness of availability in the covering problem, an extension of the p-median problem, can also be found in [11].

The purposes of this paper are twofolds. We formulate the discrete location problems when the facilities are unreliable as 0–1 integer programming model. An enhanced dual-based procedure, which is based on [4] for solving the classical location model, is then developed and tested. The formulation and the solution procedure have been motivated by an application to model and solve a large size problem for locating over two thousand base stations in a large cellular network. The major differences of the enhanced procedure from that of Erlenkotter [4] are the sequential application of dual processes to both the primary facility (the nearest facility to serve a customer when it is available) and the secondary facility (the second nearest facility to serve a customer when the nearest facility is closed or failed), and the simple branch and bound heuristic; both are more efficient than that in the original approach. The computational study has indicated that the procedure is both efficient and effective. The remaining of this paper is organized as follows. Section 2 briefly describes the application background; and the mathematical model for the unreliable discrete location problem is formulated. In Section 3, the proposed threefold procedure is described. Computational results and relevant findings are reported in Section 4. A brief concluding summary is given in the last section.

2. Motivations and problem formulation

The location of base stations or cellular towers is of paramount importance for the performance of cellular communication network. Cellular providers are driven by the goal to enhance performance, particularly as it related to the receipt and the transmission of signals generated by the mobile communication system. Given a potential market (and an area) to serve the mobile phone users, cellular providers have to make the following key decisions:

- The number of base stations or cellular towers to be located, which is budget dependent.
- Optimal position or location of the base stations to maximize the coverage (service availability and quality) and/or the level of penetration in the region given.
- Maximal transmitting power (capacity) for each base station.
- Antenna height.

The first two aspects are major problems in the design of the cellular communication network. The last two issues are regulated by both building codes and electro-magnetic emission restrictions, and often are resolved after the primary network design has been specified. The capacity limit is then specified to meet the total assigned demand of users and to reduce the fixed installation cost per station.

Unlike the classical location problems where the location of customer or demand is fixed, the mobile user (subscriber) accesses the communication system within the region in the cellular network, which implies that the location of the demand (the receipt and the transmission of phone calls to/from the user) is usually nondeterministic; i.e., user dynamically roams around to access the base stations. This ‘randomness’ of user’s demand is of major concerns at the network design stage within the metropolitan area, which consists of 70–85% of the market. To optimize the service performance, only high-rise buildings are selected as candidate base stations (or cellular towers) to reduce the obstacle interference and to maximize the level of penetration. There is only a limited number of buildings can be selected and configured, and the fixed cost to lease the space and install the equipment is high.

The objective of the network design is to determine the number of base stations and their locations such that all users of the cellular network are served with adequate service level. However, for an economic design of competitive cellular network, a tradeoff between the cost of coverage and service availability is desired. The cost of coverage for a given number of base stations is the total fixed cost to lease, install, and operate these stations. Service availability cannot be measured deterministically and directly due to the randomness and dynamics of the user's demand. The expected level of service or penetration is a better measure than the traditional signal strength in typical communication network, since it captures the randomness of customer's demand. Due to the 'roaming' of users, a base station becomes inactive (equivalent to breakdown) when it capacity reaches the upper limit, and the relay of service has to be processed by the next nearest base station. In this sense, the facility becomes unreliable due to random demand of the users. Transmitting quality is maximized when user is in the vicinity of a base station. Hence, a good surrogate to measure the service quality is the distance from the user to the base station, which is inversely proportional to the transmitting quality. Note that this measure is satisfactory only if high-rise buildings are selected for installing the cellular towers, since the obstacle interference has been minimized. The measure is also accurate in the rural area since obstacle interference is diminishing and demand of user is low in comparison with that in the metropolitan area.

In the configuration of a cellular communication system, demand of nearly a million users in the network is aggregated into a reduced number of clusters to reduce the problem complexity. Users access the nearest base station to maximize the service quality, unless this base station becomes unavailable due to capacity limit or breakdown. In this case, the next nearest base station will be used to transmit the signal from/to the user. This implies that service availability and quality is maximized when the expected transmitting distance from each user to the base stations is minimized. This cellular location problem can now be restated in a formal way below.

In the unreliable discrete location problem, n unreliable, candidate facilities (base stations) are to be selected to serve the demand of m customers (clusters). The objective function to be optimized consists of two parts, the fixed charge to erect a candidate facility, and the expected weighted sum of the transportation cost (can be distance dependent or defined by other measure such as signal loss in the communication system) from each customer to different facilities when the facilities are unreliable. We should assume that the probability of a facility becomes inactive is specified or known. To serve a customer, in the case that the nearest facility (with the best transmission quality, i.e., the lowest transportation cost in the network) is unavailable, another facility with the next lowest transportation cost will provide service, etc. Thus, each customer should be at least served by two facilities; the one with the best quality (or the lowest cost) is called the primary facility, and the one with the second lowest cost is the secondary facility. The capacity of each facility is assumed sufficient to accommodate the largest demand from any customer. Notations that are used for the formulation are given below:

I is the set of demand point, with $i = 1,\ldots,m$ (cluster of subscribers).
J is the set of unreliable candidate facilities, with $j = 1,\ldots,n$ (base stations).
f_j is the fixed charge when facility j is selected (installation and operating cost).
P_j is the probability that facility j is reliable or active (the percentage of time that a subscriber can access the nearest base station), hence, $1 - P_j$ is the roaming rate of the user.
C_{ij} is the transportation cost from customer i to facility j (a quantitative measure of service quality).
P_jX_{ij} is the proportion of total demand of customer i, which is supplied from facility j, when the facility is active. For customer i, j is the primary facility, which has the lowest transportation cost or the best service quality.
$(1 - P_j)Y_{ik}$ is the proportion of total demand of customer i, which is supplied from facility k, $k \neq j$, when the primary facility failed to provide service. For customer i, k is the secondary facility, which has the second lowest transportation cost.
K_j is 1 if facility j is established and 0 otherwise.

L_{ij} is 1 if facility j is the primary facility for customer i, and it is 0 otherwise.
M_{ij} is 1 if facility j is the secondary facility for customer i, i.e., when the primary facility failed, and it is 0 otherwise.

The objective function to be optimized is

$$\text{Minimize} \quad Z_p = \sum_{j=1}^{n} f_j K_j + \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} \sum_{\substack{k=1 \\ k \neq j}}^{n} \left[C_{ij} P_j X_{ij} + C_{ik}(1 - P_j) Y_{ik} \right] \right\}$$

$$\text{Subject to} \quad \sum_{j=1}^{n} \sum_{\substack{k=1 \\ k \neq j}}^{n} \left[P_j X_{ij} + (1 - P_j) Y_{ik} \right] = 1, \quad \forall i \in I, \tag{1}$$

$$X_{ij} \leqslant K_j, \quad \forall i \in I, \ j \in J, \tag{2}$$

$$Y_{ij} \leqslant K_j, \quad \forall i \in I, \ j \in J, \tag{3}$$

$$X_{ij} \leqslant L_{ij}, \quad \forall i \in I, \ j \in J, \tag{4}$$

$$Y_{ij} \leqslant M_{ij}, \quad \forall i \in I, \ j \in J, \tag{5}$$

$$L_{ij} + M_{ij} = 1, \quad \forall i \in I, \ j \in J, \tag{6}$$

$$X_{ij}, Y_{ij} \geqslant 0, \quad \forall i \in I, \ j \in J, \tag{7}$$

$$K_j \in \{0, 1\}, \quad \forall j \in J, \tag{8}$$

$$L_{ij} \in \{0, 1\}, \quad \forall i \in I, \ j \in J, \tag{9}$$

$$M_{ij} \in \{0, 1\}, \quad \forall i \in I, \ j \in J. \tag{10}$$

Constraint set (1) shows that each customer should be served by either the primary facility or the secondary facility (when the primary facility is inactive) at any instant. Constraints sets (2) and (3) indicate the flow feasibility between each customer and the facility. Constraint sets (4)–(6) enforce the flow feasibility so that each facility is used as either the primary or the secondary facility, but not both; and the demand of each customer can not be satisfied by using only one facility, due to the unreliability of the facility.

Compared with the classical discrete location model, which has only n integer variables and $2mn + m + n$ constraints, we have $2mn + n$ integer variables and $9mn + m + n$ constraints. The problem size of the unreliable model has increased significantly. In particular, the number of discrete variables has increased from n to $2mn$. However, the number of integer variables to model the location problem can be significantly reduced by the introduction of the following property.

Property (The single sourcing property). *For the unreliable discrete location problems with uncapacitated facilities, each demand point will be served by only one facility at any instant, either the primary facility or the secondary facility, but not both.*

Proof. We will extend the single sourcing property in the classical model to this new problem. Note that the classical model is a special case (by setting $P_j = 1$) of the described model here. We consider any customer i at any instant, when the primary facility (the nearest station with the cost C_{ij}) is active, the demand will be met by this facility, which is obvious. When this facility is inactive at a given instant, the demand will be satisfied by the secondary facility, which has the second lowest transportation cost, since the facility is uncapacitated. Thus, only one facility is required to serve each customer at any instant. □

Based on the single sourcing property, the discrete variables sets of (9) and (10) can be dropped, and the constraints of (2) and (3) combined. A compact reformulation is given below:

$$\text{Minimize} \quad Z_p = \sum_{j=1}^{n} f_j K_j + \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} \sum_{\substack{k=1 \\ k \neq j}}^{n} \left[C_{ij} P_j X_{ij} + C_{ik}(1-P_j) Y_{ik} \right] \right\}$$

Subject to Constraint sets (1), (7), (8), and

$$X_{ij} + Y_{ij} \leqslant K_j \quad \forall i \in I, \ j \in J. \tag{11}$$

The new constraint set (11) can be interpreted as follows. A base station j can be used to serve customer i in two different ways. In the first case, facility j must be the one with the lowest transportation cost (the best service quality) for customer i, i.e., facility j is the primary facility for this customer, which implies that $X_{ij} = 1$. Alternatively, facility j must be the one with the second lowest transportation cost to customer i. That is, facility j is the secondary facility for this customer, which implies that $Y_{ij} = 1$. The constraint set also implies that each facility j can only be used to serve one customer as either the primary or the secondary facility, but not both. Thus, the single sourcing property has been included in the model.

The number of discrete variables in the compact model is comparable to that of the classical location model, while the number of continuous variables has increased by twofolds. The problem size has been reduced to some significant extent. We now explore the dual of the relaxed compact model, i.e., its linear programming model with the integer variables (8) replaced by

$$0 \leqslant K_j \leqslant 1, \quad \forall j \in J. \tag{12}$$

Let V_i be the dual variable corresponding to (1), and U_{ij} be the dual variable corresponding to (11), the dual of the relaxed linear programming model can be formulated as follows:

$$\text{Maximize} \quad Z_{\mathrm{D}} = \sum_{i=1}^{m} V_i$$

$$\text{Subject to} \quad \sum_{i=1}^{m} U_{ij} \leqslant f_j, \quad \forall j \in J, \tag{13}$$

$$P_j V_i - U_{ij} \leqslant P_j C_{ij}, \quad \forall i \in I, \ j \in J, \tag{14}$$

$$(1-P_j) V_i - U_{ij} \leqslant (1-P_j) C_{\substack{ik \\ k \neq j}}, \quad \forall i \in I, \ j \in J, \tag{15}$$

$$U_{ij} \geqslant 0, \quad \forall i \in I, \ j \in J, \tag{16}$$

$$V_i \text{ unrestricted}, \quad \forall i \in I.$$

From (14) and (15), we have $U_{ij} \geqslant P_j(V_i - C_{ij})$, and $U_{ij} \geqslant (1-P_j)\left(V_i - C_{\substack{ik \\ k \neq j}}\right)$, for $\forall i \in I, j \in J$. Thus, $U_{ij} = \max\left\{0, P_j(V_i - C_{ij}), (1-P_j)\left(V_i - C_{\substack{ik \\ k \neq j}}\right)\right\}$, $\forall i \in I, j \in J$. The simplified dual formulation is

$$\text{Maximize} \quad Z_{\mathrm{D}} = \sum_{i=1}^{m} V_i$$

$$\text{Subject to} \quad \sum_{i=1}^{m} \max\left\{0, P_j(V_i - C_{ij}), (1-P_j)\left(V_i - C_{\substack{ik \\ k \neq j}}\right)\right\} \leqslant f_j, \quad \forall j \in J. \tag{17}$$

The complementary slackness conditions are

$$K_j^* \left\{ f_j - \sum_{i=1}^{m} \max\left[0, P_j(V_i^* - C_{ij}), (1-P_j)\left(V_i^* - C_{\substack{ik \\ k \neq j}}\right)\right] \right\} = 0, \quad \forall j \in J, \tag{18}$$

$$(K_j^* - X_{ij}^* - Y_{ij}^*) \max\left\{0, P_j(V_i^* - C_{ij}), (1-P_j)\left(V_i^* - C_{\substack{ik \\ k \neq j}}\right)\right\} = 0, \quad \forall i \in I, \ j \in J. \tag{19}$$

Let $\{V_i^+\}$ be the set of dual feasible solution and the selected facilities be J^+ in the corresponding primal solution. For every customer i, and $V_i^+ \geqslant C_{ij}$, all $j \in J^+$, this solution satisfies $\sum_{i=1}^{m} \max\left\{0, P_j(V_i - C_{ij}), (1-P_j)\left(V_i - C_{\substack{ik\\ k\neq j}}\right)\right\} = f_j$, for every $j \in J^+$. A dual solution that satisfies both (18) and (19) corresponds to a primal solution, which is the optimum. In general, a solution may satisfy the constraints except (19). That is, for some customer i, which satisfies $V_i^+ \geqslant C_{ij}$, all $j \in J^+$; this condition has been met by more that two j's. The duality gap for the primal solution and dual solution is given by

$$Z_p^+ - Z_{\mathrm{D}}^+ = \sum_{i=1}^{m} \sum_{\substack{j\in J^+\\ k\neq J^+(i)}} \max\left\{0, P_j(V_i^+ - C_{ij}), (1-P_j)\left(V_i^+ - C_{\substack{ik\\ k\neq j}}\right)\right\},$$

where $J^+(i) \in J^+$. This can be easily established by extending the primal-dual relationship.

3. The threefold heuristics

When the number of base stations is modest, commercial software can be used to find solutions of the unreliable location problem. This approach is unrealistic and infeasible since the application case includes more than two thousand base stations to be selected to serve about one million users. After some aggregation of demand points, the problem size ($n = 2000$, $m > 50{,}000$) is too large to be managed by the regular integer programming software. Fig. 1 depicts geographical locations of the potential 2000 base stations in a metropolitan area. This motivated the development of an efficient heuristic that can be used to handle both moderate and very large problem sizes.

The proposed procedure is essentially based on Erlenkotter's [4] dual procedure, which deals with the discrete location problems when the facilities are reliable. It extends and enhances the dual procedure so that it can be used to solve the unreliable problems. The major differences of the enhanced procedure from that of Erlenkotter are the sequential application of dual processes to both the primary facility (the nearest facility to a customer, when is open) and the secondary facility (when it is closed), and the simple branch and bound heuristic; both are more efficient than that in the original process. The first component of our solution procedure is a dual ascent procedure, which cyclically increases the value of each dual variable until it is blocked by the dual feasibility. In general, the probability P_j is large. Thus, starting with any dual feasible solution, values of the dual variable with respect to the primary facility are increased with the highest priority. Later, the same procedure is applied to increase the dual values associated with the secondary facilities, based on the constructed dual solution from the primary facilities in the first pass. The second component of the approach begins when all values of the dual feasible variables are blocked from the further increase. If the dual solution satisfies the complementary slackness conditions, the solutions are optimal, and the primal solutions are determined by the primal-dual conditions. If not, the dual variables are adjusted to improve the solution. The dual variables that violated the complementary slackness conditions are selected to improve the feasibility. This is to maintain high level of dual improvement to provide better bound for further search such as the branch and bound procedure. If the complementary slackness can be satisfied, we terminate the dual procedures. Otherwise, a simple branch and bound procedure is initiated to find the optimal solution, using the objective of the dual problem and the corresponding candidate facilities in the primal solution at the second stage to determine the search.

The dual ascent procedure is described in Section 3.1, which is used to compute the dual value of solutions for both the primary and the secondary facilities. The solution obtained at the first phase is then refined in the dual adjustment procedure, which is described in Section 3.2. Integer solutions are obtained as the by-product of the dual solution. If complementary slackness conditions cannot be satisfied, a simple branch and bound procedure is employed to complete the search process, which is described in Section 3.3.

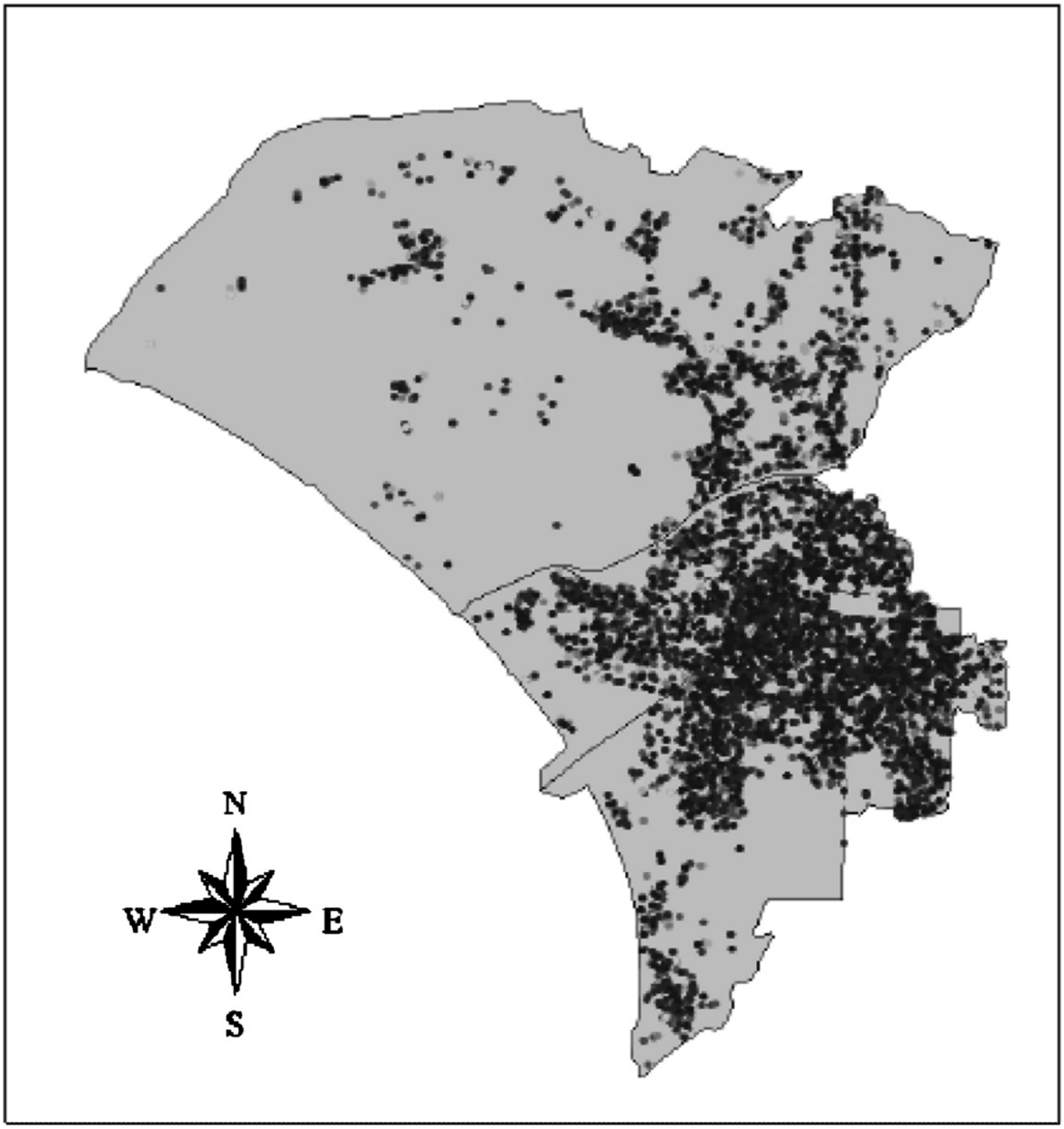

Fig. 1. Geographical locations of the potential base stations in a metropolitan area. Population of the city: 0.8 million. Number of subscribers within the coverage: 0.92 million.

3.1. The dual ascent procedure

Additional notations that are used in the solution procedure are given below:

S_j is the value of the dual slack solution.
C_i^k for each customer i, re-index C_{ij} in non-decreasing order, $k = 1, 2 \ldots, n$.
$k(i)$ is the facility which offers service to a customer with the kth non-decreasing transportation cost, for $i \in I$; i.e., $k(i) = \min\left\{k, P_jV_i \leqslant P_jC_i^{k(i)}\right\}$.
J^+ is the set of candidate facilities in the dual solution, initially $J^+ = J$.
J_i^+ is the set of facilities that satisfied the condition $P_jV_i > P_jC_{ij}$, $j \in J^+$, for $i \in I$.

$J^* = \left\{ j : \sum_{i=1}^m \max\left\{0, P_j(V_i - C_{ij}), (1-P_j)\left(V_i - C_{\substack{ik\\k\neq j}}\right)\right\} = f_j \right\}$ is the optimal set of facilities.

J_i^* is the set of facilities that satisfied the condition $P_jV_i \geqslant P_jC_{ij}$, $j \in J^*$, for $i \in I$.

$I^+ = \{i : V_i \text{ is eligible for change}\}$, initially $I^+ = I$.

$I_j^+ = \{i : J_i^* = (j)\}$.

$J^+(i)$: facility j provides service to customer i with the lowest transportation cost, i.e., j is the primary facility for customer i.

$J'^+(i)$: facility j provides service to customer i with the second lowest transportation cost, i.e., j is the secondary facility for customer i.

$C_i^- = \max_{j\in J}\{C_{ij} : P_jV_i > P_jC_{ij}\}$, for $i \in I$.

The objective of this procedure is to construct a dual solution, which satisfies constraint (18), but may violate (19). This is implemented by two phases of the dual ascent process.

Procedure (*Dual ascent*). Phase one—Ascent procedure for the primary facilities: (In general, $P_j \geqslant 0.5$, the dual solution corresponding to the primary facilities is constructed first. When $P_j < 0.5$, the two-phase procedure is implemented in the reversed order.)

1. To obtain the initial dual feasible value, set $P_jV_i = P_jC_i^1$ for each i. For every $i \in I$, it satisfies: $S_j = f_j - \sum_{i=1}^m \max\{0, P_j(V_i - C_{ij})\} \geqslant 0, j \in J$. If $P_jV_i = P_jC_i^{k(i)}$, increase $k(i)$ by 1.
2. Set $i = 1$, and flag $F = 0$.
3. If $i \notin I^+$, go to step 6.
4. Let $\varDelta_i = \min_{j\in J}\{S_j : P_jV_i \geqslant P_jC_{ij}\}$. If $\varDelta_i > P_jC_i^{k(i)} - P_jV_i$, then $\varDelta_i = P_jC_i^{k(i)} - P_jV_i$; and $F = 1$. Set $k(i) = k(i) + 1$.
5. For every $j \in J$, when $P_jV_i \geqslant P_jC_{ij}$, increase P_jV_i by $\varDelta_i$ and decrease S_j by $\varDelta_i$.
6. If $i < m$, then increase i by 1 and return to step 3.
7. If $F = 1$, return to step 2. Else, terminate this phase and go to the next phase.

Phase two—Ascent procedure for the secondary facilities:

1. Use the dual value (V_i) obtained in phase one as the initial dual solution. When $P_jV_i = P_jC_i^{k(i)-1}$, set $(1-P_j)V_i = (1-P_j)C_i^{k(i)-1}$. If $P_jV_i > P_jC_i^{k(i)-1}$, then let $(1-P_j)V_i = (1-P_j)C_i^{k(i)}$. Also let $ku(i) = k(i)$, for $i \in I$, and $S_j = S_j$ (the values at the end of the first phase), for $j \in J$.
2. Set $i = 1$, and flag $F = 0$.
3. If $i \notin I^+$, go to step 6.
4. If $P_jC_i^{k(i)-1} \leqslant P_jC_{ij}$, Let $\varDelta_i = \min_{j\in J}\{S_j : (1-P_j)V_i \geqslant (1-P_j)C_{ij}\}$. If $\varDelta_i > (1-P_j)C_i^{ku(i)} - (1-P_j)V_i$, then $\varDelta_i = (1-P_j)C_i^{ku(i)} - (1-P_j)V_i$; and $F = 1$, $ku(i) = ku(i) + 1$.
5. For every $j \in J$, when $(1-P_j)V_i \geqslant (1-P_j)C_{ij}$ and $P_jC_i^{k(i)-1} \leqslant P_jC_{ij}$, increase $(1-P_j)V_i$ by $\varDelta_i$ and decrease S_j by $\varDelta_i$.
6. If $i < m$, then increase i by 1 and return to step 3.
7. If $F = 1$, return to step 2. Else, terminate the dual ascent process.

At the end of the ascent procedure, the sum of P_jV_i (at the end of phase one) and $(1-P_j)V_i$ (at the end of phase two) is treated as the dual solution V_i, for $i \in I$. The dual ascent procedure yields a candidate dual solution and a candidate primal solution through $K_j = 1, j \in J^+$, and $X_{ij} = 1$, $i \in i^+(j), j \in J$. If these solutions satisfy (19), the solutions are optimal by the complementary slackness conditions. Otherwise, the dual solutions are improved through a dual adjustment procedure, which is described below.

3.2. The dual adjustment procedure

After the dual ascent procedure is implemented to increase the objective value of the dual problem, the complementary slackness conditions (19) of the primal-dual models: $(K_j^* - X_{ij}^* - Y_{ij}^*)$

$\max\left\{0, P_j(V_i^* - C_{ij}), (1-P_j)\left(V_i^* - C_{\substack{ik \\ k\neq j}}\right)\right\} = 0$, may be violated. This procedure is used to improve the dual values so that the conditions can be met as much as possible. To initiate this process, some customer i for which (19) is violated is selected. If this dual value is decreased, it creates slack on at least two binding constraints of (17). Hence, other dual variables can be increased that are limited by these constraints. Due to the unreliability of the facilities, the dual objective is not guaranteed to increase.

Procedure (*Dual adjustment*)

1. Set $i = 1$, i.e., initiate the dual adjustment from the first customer.
2. If $|J_i^+| \leqslant 2$, go to step 7.
3. If $I^+_{j^+(i)} = \phi$ and $I^+_{j'^+(i)} = \phi$, go to step 7.
4. For every $j \in J$, which satisfies $P_jV_i \geqslant P_jC_{ij}$, decrease P_jV_i to C_i^- and increase S_j by $P_jV_i - C_i^-$.
5. Let $I^+ = I^+_{j^+(i)} \cup I^+_{j'^+(i)}$ and perform the dual ascent procedure. Augment the set I^+ by including the customer $i \in I\backslash I^+$, and perform the dual ascent procedure. Repeat this augmentation until $I^+ = I$.
6. If P_jV_i (adjusted in the step 5) $\neq P_jV_i$ (adjusted in step 4), return to step 2. Else, go to the next step.
7. If $i < m$, then increase i by 1 and return to step 2. Else, stop.

At the end of the dual adjustment procedure, we have a dual feasible solution on hand. If it satisfies the complementary slackness conditions of (18) and (19), a primal solution can be derived by the primal dual relationship. The heuristic is thus terminated, and the primal solution is the optimum. On the other hand, duality gap exists and a simple branch and bound procedure is followed to find the optimal solution.

3.3. The simple branch and bound procedure

In general, the duality gap produced from the dual procedure in the previous section is very small (see also [4]). Hence, we use a heuristic version of the branch and bound procedure, instead of the sophisticated one. It is necessary to adapt this simple strategy for solving very large problems, in order to maintain the computational feasibility.

Procedure (*Heuristic branch and bound*)

1. Initialize the procedure with only essential facilities in J^+, deleting non-essential facilities to obtain the best primal integer solution so far. The solution and the objective obtained so far are used as the basis, i.e., the lower/upper bound value of the objective function.
2. Count the number of times each facility j is open in the dual procedures, i.e., in the set J^+. Sequence these facilities with non-increasing order. Tie-breaking by the sequence it enters this set.
3. Sequentially, starting from the first two facilities, compute the primal objective value when the facilities in the ordered sequence are open. For unreliable location problems, we require at least two facilities open.
4. If the primal objective is increased when a new facility in the sequence enters the set J^+ (i.e., branching on this binary variable), the primal objective value of the last solution is used as the upper bound of objective value of the optimum solution.
5. Continue until the ordered list is emptied.

Other sophisticated schemes have not been investigated, due to the high level of effectiveness of the dual solution. This is verified through numerical experiments, which are reported later. In addition, an exact branch and bound procedure, which is modified from that of Erlenkotter, is used if the optimum solution is a must, or the problem size is manageable.

Procedure (*Exact branch and bound*)

1. Initialize the procedure with only essential facilities in J^+, deleting non-essential facilities to obtain the best primal integer solution so far. The solution and the objective obtained so far are used as the basis.

2. For branching, some facility contributing to the violation of a complementary slackness condition is selected. The corresponding lowest-cost source $J^{+}(i)$ is branched.
3. To fix facility j closed, the fixed charge is replaced by $+\infty$, so that the current solution is still feasible. Apply the dual ascent and adjustment procedures to improve the dual objective value.
4. To fix facility j open, the fixed charge is replaced by 0. The dual feasibility is restored by reducing the dual value for each $P_jV_i \geqslant P_jC_{ij}$ to P_jC_{ij}. (Here we assume that $P_j > 0.5$.) Since (17) is still enforced, reduction of these V_i does not change the dual objective value. Other dual values may be increased due to this creation of slackness. Apply the dual ascent and adjustment procedures to improve the dual objective value.
5. Fathoming of nodes in the procedure is either by bounding or by obtaining a primal integer solution that satisfies (19).

Due to the impact of unreliability on the dual solution, the branch and bound procedure can not be expected to perform as well as those described in [4].

4. Computational studies

Due to the overwhelmingly large problem size of the described application, a detailed description for the implementation of the heuristic is out of the scope of the current paper. Instead, computational results for some reasonable large problems are reported in this section, to illustrate the potential applicability of the proposed heuristic. Computational tests on these well-known problems are conducted on VAX9420 mainframe platform. The solution procedures are coded with ANSI C. The tested problems are adapted from the literature, including (5 × 8) of Khumawala [2] with 5 facilities and 8 customers, (33 × 33) and (57 × 57) of Karg and Thompson [12] with 33 (and 57) facilities and 33 (and 57) customers, and (100 × 100) of Krolak et al. [13] with 100 facilities and 100 customers. The probability and the fixed charges of the facility are varied to generate alternative problem sets.

In the next two sections, we first give an example to illustrate the computational details of the proposed heuristic. The performance of the heuristic is then reported in Section 4.2.

4.1. An illustration

We assume that, for each user, the probability of requesting service from the nearest facility is 0.9 (P_j), and from the next nearest facility is 0.1. That is, the roaming rate of users is ten percent in the cellular network. This example is adapted from [2] with 5 facilities and 8 customers. The fixed charges are (100, 70, 60, 110, 80), respectively; and the transportation cost matrix is

$$C_{ij} = \begin{vmatrix} 120 & 180 & 100 & \infty & 60 & \infty & 180 & \infty \\ 210 & \infty & 150 & 240 & 55 & 210 & 110 & 165 \\ 180 & 190 & 110 & 195 & 50 & \infty & \infty & 195 \\ 210 & 190 & 150 & 180 & 65 & 120 & 160 & 120 \\ 170 & 150 & 110 & 150 & 70 & 195 & 200 & \infty \end{vmatrix}.$$

To begin with, we have $(P_jV_i) = (108, 135, 90, 135, 45, 108, 99, 108)$, with dual objective value $Z_{\mathrm{D}} = 828$, and the dual slack variables $(S_j) = (f_j) = (100, 70, 60, 110, 80)$. After one iteration of the first phase in the dual ascent procedure, we have $(P_jV_i) = (153, 162, 99, 162, 49.5, 175.5, 144, 148.5)$, $Z_{\mathrm{D}} = 1093.5$, and $(S_j) = (46, 25, 55.5, 2, 26)$. After two more iterations, all dual variables have been blocked by the constraints, and the process is terminated. We have $(P_jV_i) = (162, 171, 107, 162, 58.5, 175.5, 146, 148.5)$, $Z_{\mathrm{D}} = 1130.5$, and $(S_j) = (15.5, 14, 38.5, 0, 0)$. We now proceed to the second phase of the dual ascent procedure. For the secondary facilities, the initial dual solution is $((1 - P_j)V_i) = (18, 19, 15, 18, 6.5, 19.5, 18, 16.5)$, $Z_{\mathrm{D}} = 1130.5 + \sum(1 - P_j)V_i = 1261$, and $(S_j) = (15.5, 14, 38.5, 0, 0)$. At the end of the first iteration, we have $((1 - P_j)V_i) = (21, 19, 15, 18, 6.5, 19.5, 18, 19.5)$, $Z_{\mathrm{D}} = 1130.5 + \sum(1 - P_j)V_i = 1267$, and $(S_j) = (15.5, 11, 35.5, 0, 0)$. The detail of the dual ascent procedure is given in Table 1.

Table 1
The detail of dual ascent procedure in the illustration

Phase 1	Dual values P_jV_i, $i = 1,\ldots,8$ (dual objective Z_D)	S_j, $j = 1,\ldots,5$
Iteration 0	108, 135, 90, 135, 45, 108, 99, 108 (828)	100, 70, 60, 110, 80
Iteration 1	153, 162, 99, 162, 49.5, 175.5, 144, 148.5 (1093.5)	46, 25, 55.5, 2, 26
Iteration 2	162, 171, 107, 162, 54, 175.5, 146, 148.5 (1126)	20, 18.5, 43, 0, 0
Iteration 3	162, 171, 107, 162, 58.5, 175.5, 146, 148.5 (1130.5)	15.5, 14, 38.5, 0, 0
Phase 2	Dual values $(1 - P_j)V_i$, $i = 1,\ldots,8$ (dual objective Z_D)	S_j, $j = 1,\ldots,5$
Iteration 0	18, 19, 15, 18, 6.5, 19.5, 18, 16.5 (1261)	15.5, 14, 38.5, 0, 0
Iteration 1	21, 19, 15, 18, 6.5, 19.5, 18, 19.5 (1267)	15.5, 11, 35.5, 0, 0
Iteration 2	21, 19, 15, 18, 6.5, 19.5, 18, 30.5 (1278)	15.5, 0, 24.5, 0, 0
Termination	V_i, $i = 1,\ldots,8$ (dual objective Z_D)	S_j, $j = 1,\ldots,5$
	183, 190, 122, 180, 65, 195, 164, 1769 (1278)	15.5, 0, 24.5, 0, 0

We now initiate the dual adjustment procedure. Since for every i, we have $|J_i^+| \leqslant 2$, the procedure is terminated with $Z_D = 1278$. This dual solution satisfies the complementary slackness conditions; and $J^+ = \{2,4,5\}$, where $X_{15} = X_{25} = X_{35} = X_{45} = X_{52} = X_{64} = X_{72} = X_{84} = 1$, and $Y_{12} = Y_{24} = Y_{32} = Y_{44} = Y_{54} = Y_{65} = Y_{74} = Y_{82} = 1$. All constraints have been satisfied, and the objective values of the primal and the dual are the same, we have the optimal integer solution.

For a more challenging problem, we change the fixed charges of the candidate facilities to $(f_j) = (200, 200, 200, 400, 300)$. At the end of the dual ascent procedure, we have $(V_i) = (210, 255, 153, 287.5, 60, 242.5, 180, 217.5)$, $Z_D = \sum V_i = 1605.5$, and $(S_j) = (0, 0, 0, 56.5, 0)$. The corresponding facilities to be selected are $J^+ = \{1, 2, 3, 5\}$, where the complementary slackness conditions (19) are violated. We now proceed to the dual adjustment procedure. Since $|J_1^+| > 2$, there are more than two facilities which provide service to customer 1. The dual slack values of S_1, S_3, S_5 are increased by 27, respectively, and the dual value P_jV_1 is reduced from 189 to 162. The dual procedure is applied, we have $(P_jV_i) = (162, 236, 135, 216, 54, 248.5, 180, 203.5)$, and $((1 - P_j)V_i) = (21, 19, 15, 24, 6.5, 21, 20, 19.5)$, $Z_D = 1581$, and $(S_j) = (0, 0, \ 0, 0, 0)$. At the next iteration, since $|J_2^+| > 2$, the dual slack values of S_1,S_3, S_4, S_5 are increased by 65, respectively, and the dual value P_jV_2 is reduced from 236 to 171. The dual procedure is applied again, we have $(P_jV_i) = (189, 209, 135, 216, 54, 248.5, 180, 203.5)$, and $((1 - P_j)V_i) = (21, 40.5, 15, 24, 7, 21, 20, 19.5)$, $Z_D = 1603$, and $(S_j) = (0, 0, 0, 14, 0)$. No more adjustment can be made; and the adjustment procedure is terminated.

Since the duality gap is greater than zero, we proceed to the heuristic branch and bound procedure with $J^+ = \{1, 2, 3, 5\}$, and dual objective $Z_D = 1603$, primal objective $Z_P = 1823.5$. The number of times each facility enters the candidate set are $|J_5| = |J_2| = |J_3| = 20$, $|J_4| = 14$, and $|J_1| = 2$. Hence we have $J^+ = \{5, 2, 3\}$, where the upper bound is 1823.5 and the lower bound is 1603. We first fix the facility 3 closed, the primal solution is infeasible. To fix facility 3 open, a primal feasible solution is derived with objective value 1823.5. Further branching does not give any improvement. We terminate the heuristic with $J^+ = \{5, 2, 3\}$.

4.2. The computational study

Totally we solved 80 problems, where the problem sizes are (5×8), (33×33), (57×57), and (100×100). For each problem set, the probabilities P_j are 1.0 (fully reliable), 0.95, 0.9, 0.85, 0.8, and 0.7, respectively. Due to the high efficiency of the heuristic, we first solve the location problem with $P_j = 1.0$ for each data set. The primal solution is then derived to obtain a candidate solution for the unreliable problem, if it is feasible. These computational results are summarized and discussed below.

For the smallest problem of Khumawala [2] with 5 facilities and 8 customers, it takes less than 0.1 s to find the solution. The heuristic finds optimum solutions in most of the cases. Both the solution quality and the computational efficiency appear good. These statistics are summarized in Table 2.

For the (33×33) problem of Karg and Thompson [12], similar experiments are performed with fixed charges of facilities, 1000, 2000, 3000, and 4000, respectively. The average computational time to find the

Table 2
Computational results for the (5 × 8) unreliable location problem

Probability, P_j	Fixed charge, f_j	Objective value, Z_{P}	Deviation of objective from the true optimum %	CPU time in seconds
1.00	Illustration 1	1235	0.0	<0.1
0.95	Illustration 1	1261.5	0.0	<0.1
0.90	Illustration 1	1278	0.0	<0.1
0.85	Illustration 1	1294.5	0.0	<0.1
0.80	Illustration 1	1311.0	0.0	<0.1
0.70	Illustration 1	1344	0.0	<0.1
1.0	Illustration 2	1625	2.8	<0.1
0.9	Illustration 2	1823.5	0.0	<0.1

Table 3
Computational results for the (33 × 33) unreliable location problem

Probability, P_j	Fixed charge, f_j	Objective value, Z_{P}	Deviation of objective from the true optimum %	CPU time in seconds
1.00	1000	14,832	0.0	0.01
0.95	1000	15,497	0.0	0.4
0.90	1000	16,163	0.0	0.3
0.85	1000	16,828	0.0	0.27
0.80	1000	17,494	0.10	0.45
0.7	1000	18,823	1.00	0.42
1.00	2000	20,627	1.29	0.05
0.95	2000	21,149	0.01	0.70
0.90	2000	21,935	0.18	0.65
0.85	2000	22,721	1.01	0.93
0.80	2000	23,507	1.84	0.85
0.7	2000	24,707	1.99	0.85
1.00	3000	23,747	0.0	0.02
0.95	3000	25,397	2.64	1.02
0.90	3000	27,077	4.69	1.80
0.85	3000	28,105	5.30	1.56
0.80	3000	28,420	4.37	1.92
0.7	3000	29,865	5.28	1.85
1.00	4000	25,474	0.0	0.02
0.95	4000	27,398	0.0	0.15
0.90	4000	28,929	0.23	1.27
0.85	4000	29,913	0.02	0.63
0.80	4000	30,898	0.0	1.80
0.7	4000	33,499	3.50	2.30
		Average	1.39	0.84

solution is less than 1 s. The average deviation of the objective values from those of the true optimum is less than two percents. These statistics are summarized in Table 3.

For the (57 × 57) problem of Karg and Thompson [12], similar experiments are performed with fixed charges 3000, 4000, 5000, and 6000, respectively. The average computational time to find the solution is about 14 s. The average deviation of the objective values from those of the true optimum is about 2%. These statistics are summarized in Table 4.

For the (100 × 100) problem of Krolak et al. [13], computational experiments are performed with fixed charges 7000, 8000, 9000, and 10,000, respectively. The average computational time to find the solution is about 64 s. The average deviation of the objective values from those of the true optimum is about four percents. These statistics are summarized in Table 5.

As a summary, the performance of the heuristic appears good. Though most of the times, the heuristic does not find the true optimum solution, the average deviation of the objective values from those of the true optimum is less than three percents. The statistics in Tables 2–5 also indicate that the solution quality of the heu-

Table 4
Computational results for the (57 × 57) unreliable location problem

Probability, P_j	Fixed charge, f_j	Objective value, Z_P	Deviation of objective from the true optimum %	CPU time in seconds
1.00	3000	32,136	0.0	0.07
0.95	3000	33,440	0.51	10.93
0.90	3000	34,744	1.04	14.80
0.85	3000	36,048	0.76	15.25
0.80	3000	37,363	2.21	14.17
0.7	3000	39,961	2.85	7.95
1.00	4000	35,547	0.0	0.33
0.95	4000	36,962	0.60	7.98
0.90	4000	38,320	1.52	10.65
0.85	4000	39,598	1.52	13.40
0.80	4000	41,207	3.55	12.31
0.7	4000	44,037	3.48	15.73
1.00	5000	38,982	1.28	0.18
0.95	5000	40,141	0.55	12.53
0.90	5000	41,319	1.6	12.51
0.85	5000	42,498	0.86	11.15
0.80	5000	43,677	0.84	22.28
0.7	5000	46,035	0.58	24.25
1.00	6000	40,472	0.0	0.02
0.95	6000	43,971	2.55	24.80
0.90	6000	46,375	5.11	23.51
0.85	6000	47,813	4.96	24.60
0.80	6000	48,208	1.95	31.25
0.7	6000	50,393	1.65	29.85
		Average	1.78	14.19

Table 5
Computational results for the (100 × 100) unreliable location problem

Probability, P_j	Fixed charge, f_j	Objective value, Z_P	Deviation of objective from the true optimum %	CPU time in seconds
1.00	7000	84,828	1.48	9.8
0.95	7000	87,906	2.02	72.6
0.90	7000	90,983	2.71	64.2
0.85	7000	94,060	3.71	73.8
0.80	7000	97,138	4.22	88.2
0.7	7000	102,294	3.69	67.2
1.00	8000	95,563	8.76	7.30
0.95	8000	96,754	7.12	58.2
0.90	8000	98,674	6.27	67.2
0.85	8000	99,062	3.60	61.8
0.80	8000	101,284	2.91	73.2
0.7	8000	107,946	3.53	81.0
1.00	9000	97,637	6.77	6.50
0.95	9000	98,266	4.13	49.8
0.90	9000	106,325	9.38	57.0
0.85	9000	103,062	3.40	48.0
0.80	9000	105,340	2.57	58.2
0.7	9000	121,801	11.48	45.0
1.00	10,000	94,455	0.0	8.2
0.95	10,000	98,982	1.03	58.2
0.90	10,000	103,569	2.45	74.4
0.85	10,000	108,035	4.09	84.0
0.80	10,000	112,302	5.12	69.0
0.7	10,000	121,616	6.75	80.4
		Average	4.03	63.6

ristic seems degenerate when the problem size increases. Nevertheless, this degeneracy appears small, i.e., increasing the problem size by twofolds only increases the average deviation by about 1%. Large variations of such deviation can be found in Table 5.

In terms of computational effort, the longest time to solve the largest problem of (100×100) is less than 1.5 min. However, it takes approximately 4–6 h to solve the application case, for a given set of parameter setting such as different roaming rates. As expected, the computational time increases as the problem size increases. However, the increase of CPU time is much faster than the increase of problem sizes. In addition, the increase of computational effort is much faster when the number of new facilities increases than when the number of users increases. It indicates the inherently computational difficulty to solve such combinatorial optimization problems.

5. Concluding summary

The unreliable discrete location problem has been studied, where the facilities are subject to failure. The unreliability of the facility is modeled by introducing the probability that a facility may fail to provide service. We formulated this unreliable problem and extended dual based heuristic to solve the new problem. The three-phase procedure utilizes the dual formulation and the primal-dual relationship to devise solutions. A primal solution is generated as a by-product of the process. If complementary slackness conditions are satisfied, the solution on hand can be shown to be the optimum. Otherwise, dual adjustment procedure is used to improve the dual solution; and finally the simplest branch and bound routine is used to find the optimal solution for the unreliable problem.

Computational tests have indicated that the enhanced heuristic is efficient and gives excellent solution quality. The heuristic is also flexible in terms of trade-off both solution quality and computational effort by augmenting the process with simple enumerative procedure. Additional computational improvements are possible if different strategies are used to adjust the dual solutions, and more efficient branch and bound procedure can be developed by considering the special property of these unreliable location problems.

The formulation and solution procedure have been motivated by modeling and solving a large size problem for locating over two thousand base stations in a large cellular network, where nearly one million users are considered in the system. The roaming of users (equivalent to randomness of demand, both the transmission of calls and the location of users) is modeled with a new approach, where a base station becomes inactive or unavailable when users could not access this facility. The solution procedure is used to generate alternative solutions for decision-making in the network design stages. One important feature of the approach is the capability and flexibility to generate different network configurations, to allow for further qualitative and quantitative analyses. These features include the pre-selection of some base stations in the configuration process, and the sensitivity analysis of solutions when the roaming rate of users and the availability of base stations are varied. The details of the implementation is out of the scope of the current paper.

Acknowledgment

This work was supported in part by the National Science Council of Taiwan under grant NSC87-2416-H006-027.

References

[1] M. Efroymson, T. Ray, A branch-and-bound algorithm for plant location, Operat. Res. 14 (1966) 361–368.
[2] B. Khumawala, An efficient heuristic procedure for the uncapacitated warehouse location problem, Naval Res. Logist. Quart. 20 (1973) 109–121.
[3] L. Schrage, Implicit representation of variable upper bounds in linear programming, Math. Program. Study 4 (1975) 118–132.
[4] D. Erlenkotter, A dual-based procedure for uncapacitated facility location, Operat. Res. 26 (6) (1978) 992–1009.
[5] M.L. Brandeau, S.S. Chiu, An overview of representative problems in location research, Manage. Sci. 35 (1989) 645–674.
[6] C. ReVelle, G. Laporte, Plant location problem: new models and research prospects, Operat. Res. 44 (6) (1996) 864–874.
[7] H. Pirkul, Uncapacitated facility location problem with primary and secondary facility requirements, IIE Trans. 21 (4) (1989) 337–348.

[8] Z. Drezner, Heuristic solution methods for two location problems with unreliable facility, J. Operat. Res. Soc. 38 (6) (1987) 509–514.

[9] S.-D. Lee, On solving unreliable planar location problems, Comput. Operat. Res. 28 (4) (2001) 329–344.

[10] J. Current, S. Ratick, C. ReVelle, Dynamic facility location when the total number of facilities is uncertain: a decision analysis approach, Eur. J. Operat. Res. 110 (3) (1998) 597–609.

[11] V. Marianov, C. ReVelle, The queuing maximal availability location problem—A model for the sitting of emergency vehicles, Eur. J. Operat. Res. 93 (1) (1996) 110–120.

[12] R. Karg, G. Thompson, A heuristic approach to solving traveling salesman problems, Manage. Sci. 10 (1964) 225–248.

[13] P. Krolak, W. Felts, G. Marble, A man–machine approach toward solving the traveling salesman problems, Commun. ACM 14 (1971) 327–334.

[14] Z. Drezner, On the conditional *p*-median problem, Comput. Operat. Res. 22 (5) (1995) 525–530.