7-2008

A multiple server location–allocation model for service system design

Siddhartha Syam
Marquette University, siddhartha.syam@marquette.edu

Computers & Operations Research, Vol. 35, No. 7 (July 2008): 2248-2265. DOI. This article is © Elsevier and permission has been granted for this version to appear in e-Publications@Marquette. Elsevier does not grant permission for this article to be further copied/distributed or hosted elsewhere without the express permission from Elsevier.

A Multiple Server Location–Allocation Model for Service System Design

Siddhartha S. Syam
Department of Management, College of Business Administration, Marquette University, Milwaukee, WI

Abstract

Service systems are endemic in a service economy, and effective system design is fundamental to the competitiveness of service organizations such as retailers, distributors, and healthcare providers. This is because system design may significantly facilitate (or hinder) the attainment of important organizational objectives such as minimizing system cost and maximizing service level. This paper develops and solves a comprehensive nonlinear location–allocation model for service system design that incorporates several relevant costs and considerations. These include, for instance, transportation, facility, and waiting costs, queuing considerations, multiple servers, multiple order priority levels, multiple service sites, and service distance limits. The model is first converted to an equivalent linear form and then solved using Lagrangian relaxation. A computational study shows problems with 250 service districts, 60 service sites, and 250 candidate locations are solved in about two and a half minutes. An extensive managerial experiment is conducted that evaluates alternative system designs from a number of important perspectives including centralization versus decentralization, system configuration, and service distance limit. Each scenario is evaluated with respect to two fundamental criteria, namely, total cost and service level. The analysis provides insights into important tradeoffs that must be taken into consideration in designing an effective service system.

Keywords

Service system design, Location–allocation, Lagrangian relaxation

1. Introduction

Service systems are of fundamental importance in a developed service economy for various reasons. Service industries including retailing, healthcare, and distribution depend on well-designed service systems to meet demand in a timely fashion. Manufacturers also rely on service operations to provide after-sales services for complex products. When catastrophic events occur, such as 9/11 in New York City and Hurricane Katrina in the Gulf Coast, the critical nature of service systems in the public domain, such as emergency management systems, becomes apparent.

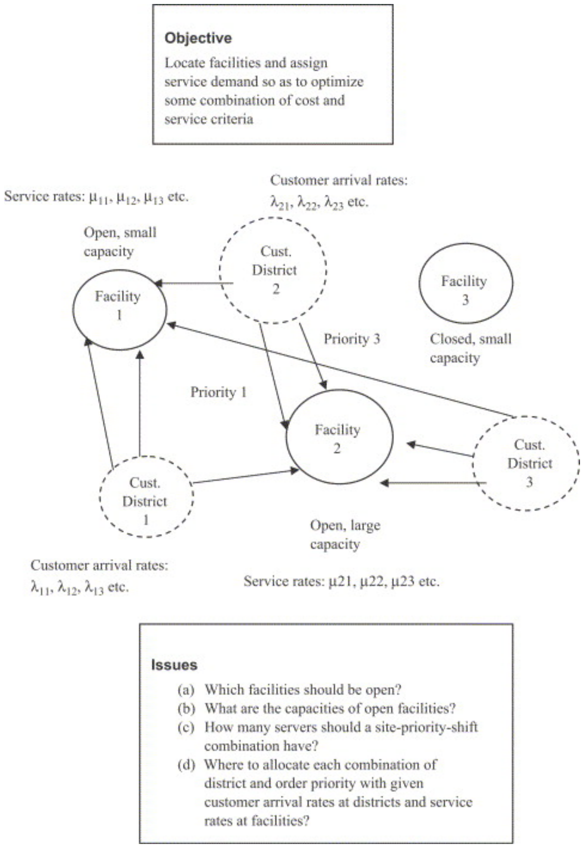
The focus of this paper is the general service system design problem (GSDP), which differs from the problem of designing an emergency services system (ESDP) in one primary way: the key issue in GSDP is the capacity of service facilities to process incoming service demand, which is not limited to emergencies, while the focus of ESDP is the capability of the field units (police patrols, fire engines, ambulances, etc.) to quickly move to emergency sites. With that caveat, GSDP and ESDP share some important characteristics. These include the configuration of the system including the location of key facilities and the assignment of service districts to open facilities. They also include knowledge of the key parameters of the system such as the rate of occurrence of events and incidents that lead to demand for service. Instances of these events could be patients arriving at a clinic or customers arriving at a retail outlet center, depending on the nature and objectives of the system.

The configuration of the service system has an overwhelming impact on its efficiency, effectiveness and productivity. This has made service system analysis a fairly active area of research for both academics and practitioners. The focus of this paper is service system design, with the inclusion of some features that are relevant to emergency response systems. Among these features are the incorporation of multiple levels of order priority (often referred to as severity or acuity in a healthcare context) and multiple modes of transportation. An example of the application of priority levels is the use of helicopters to transport critically injured patients to a hospital, and ambulances to transport those whose injury or illness is not considered life-threatening. Further, the paper incorporates differentials in cost rates depending on event priority, and also allows for multiple servers, multiple shifts, and differences in service capacities between service sites.

The objective of this paper is to model and solve the service site design problem with the features discussed above. The original motivation for the research came from our participation in a project involving the design of the Veterans' Health Administration (VHA) service system for specialized treatment services. The scope of the paper, however, goes beyond the VHA system or indeed healthcare systems in general. The model developed here may be applied, with minor modifications, to a broad variety of services. These include commercial applications in telecommunications (call centers), retailing, marketing, and applications in the public sector such as law enforcement, healthcare, and emergency management.

The problem is modeled as a 0–1 binary integer programming model which selects a pre-specified number of open facilities (referred to in this paper as service sites) from a pool of candidate service centers. The model's goal is to assign each combination of order priority and service district to a service site so as to minimize the total cost of assignment. It incorporates several relevant costs including fixed and overhead costs of service sites, waiting costs, service costs, and travel costs. Service capacity at each service site at the time of occurrence is taken into account when making assignments, as is the maximum permissible service distance. These features allow the paper to investigate an important aspect of system design, which is the degree of capacity centralization. Highly centralized systems have relatively few service sites with concentrated capacity, while decentralized systems have many open facilities with relatively distributed capacity. The model computes available service site capacity as the product of the proportion of capacity assigned to the service and total center capacity. The problem is solved using the Lagrangian relaxation [1] methodology.

A wide variety of problems is solved for which computational results are provided and managerial implications are discussed. Fig. 1 that follows provides a simplified schematic of the problem environment of the service system design problem investigated in this paper. The figure indicates that the environment involves a set of potential service centers and a set of customer demand districts. Customer arrivals and service times are probabilistic. However, customer arrival rates by priority level and shift at the various districts are known as are service rates by priority level and shift at the potential centers. The objective of the decision-maker is to optimize a combination of cost and service criteria. Important issues include the determination of which service centers to open, corresponding center capacities, and the allocation of customers to appropriate centers.



Objective

Locate facilities and assign service demand so as to optimize some combination of cost and service criteria

Service rates: $\mu_{11}, \mu_{12}, \mu_{13}$ etc.
Customer arrival rates: $\lambda_{21}, \lambda_{22}, \lambda_{23}$ etc.
Open, small capacity
Facility 1
Cust. District 2
Facility 3
Priority 3
Closed, small capacity
Priority 1
Facility 2
Cust. District 3
Cust. District 1
Open, large capacity
Customer arrival rates: $\lambda_{11}, \lambda_{12}, \lambda_{13}$ etc.
Service rates: µ21, µ22, µ23 etc.

Issues

(a) Which facilities should be open?
(b) What are the capacities of open facilities?
(c) How many servers should a site-priority-shift combination have?
(d) Where to allocate each combination of district and order priority with given customer arrival rates at districts and service rates at facilities?

Fig. 1. Service system with multiple servers, order priority levels, and time shifts.

The following sections are concerned, respectively, with (i) a brief summary of the relevant literature, (ii) presentation and discussion of the binary programming model, (iii) the solution methodology, (iv) computational results and managerial implications and (v) concluding thoughts and possibilities for future research.

2. Background

The history of academic research in emergency response systems research starts in the 1960s when the City of New York collaborated with the Rand Institute [2] to study the optimal design for the city's firefighting and police patrol systems. The techniques used were OR/MS tools such as simulation and mathematical programming, and the collaboration resulted in several outstanding tools that were implemented by the City of New York and published in more than 15 refereed journal articles in top-tier academic journals. A general description of many of these projects may be found in a survey of management science applications in the area of emergency responsiveness that is provided in [3].

Deterministic location models are frequently found in the GSDP and ESDP literature. ReVelle and Swain [4] did early work on GSDP with a model that minimizes the average travel time between service site-node assignments, subject to a restriction on the number of open facilities. This restriction is the hallmark of the p-median model [5] which is a well-known formulation of the location–allocation problem. Toregas et al. [6] introduced a maximal travel time constraint, and formulated the problem as a set covering problem, where each node is ‘covered’ by at least one member of a set of facilities that meet the travel time restriction for the particular node. Church and ReVelle [7] formulated the maximal covering location problem which relaxes the requirement that all demand nodes should be covered. Galvao and ReVelle [8] provide a Lagrangian heuristic to solve the maximal covering problem. The service system model developed in this paper, while including non-deterministic elements, has commonalities with the p-median and set covering models. A review of location–allocation modeling applications for health services in developing countries is found in Rahman and Smith [9].

Probabilistic location models include the hypercube model discussed in Larson [10]. The model incorporated several performance measures including travel time to incidents. The assumptions include a Poisson process for the service calls, and exponentially distributed service times, which are also followed in this paper. Numerous modifications of the hypercube model, as well probabilistic set covering models that extend the purview of deterministic models to include stochastic environments are reviewed in Swersey [11].

Next, we briefly note some recent representative instances of the continuing research in this area, without claiming to provide comprehensive coverage. Berman et al. [12] develop a model involving multiple transfer points and one service site, where a transfer point is a location where a group of people are loaded into a (usually) faster mode of transportation. Sinreich and Marmor [13] apply simulation to the analysis of emergency room services, while Beraldi et al. [14] use stochastic programming to hedge against uncertain conditions in environments where service sites must be identified and the number of emergency vehicles to be assigned to each site must be determined. D’Amico et al. [15] apply simulated annealing and graph partitioning to the design of police districts, including issues pertaining to the quality of emergency services such as limits to the response time to calls for service. Service standards, such as service availability within a specified time or a specified distance, are also investigated in Marianov and ReVelle [16]. The study contains an analysis of the maximal availability location problem with a model for the situating of emergency vehicles. Similarly, resource allocation in emergency evacuation networks is studied in Bakuli and Smith [17] and Rajan and Mannur [18] investigate set covering-location models for emergency situations that need multiple response units. Pirkul and Schilling [19] consider the location of emergency facilities with workload capacity limits and the need for backup service.

Amiri [20], [21] consider service system design issues that have similarity to those considered in this paper. These studies, like the current paper, determine service site location and user node allocation, and employ mathematical techniques of solution based on Lagrangian relaxation. The main area of application of the problem studied in Amiri [20] is the design of telecommunication networks, although other applications are possible. The goal is to minimize total costs which consist of access costs, waiting or queuing costs, setup costs, and operating costs. Computational results are provided for an integer programming model solved using two heuristic approaches based on Lagrangian relaxation. Amiri [21] extends the general approach to the case where backup or secondary service is provided at user nodes.

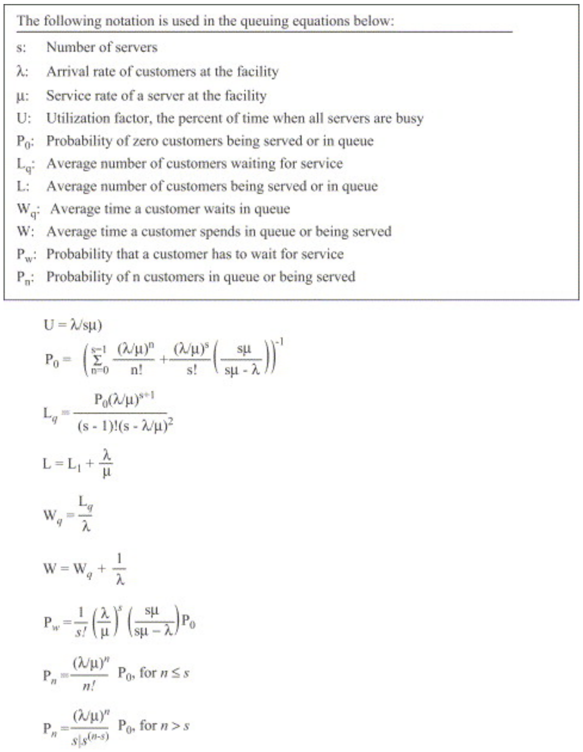
The research described in this paper incorporates several important features not found in Amiri [20], [21]. In addition to multiple service districts and queuing considerations, the model developed here includes (i) a pre-specified number of service sites as in the p-median problem, which have to be chosen from the set of candidate locations, (ii) multiple levels of order priority (for instance, a heart attack might represent the highest priority while the common cold could represent the lowest level), (iii) multiple level staffing, processing, and transportation costs, (iv) multiple work shifts, (v) multiple and varying numbers of servers and (vi) varying waiting times and associated waiting costs that depend on the number of servers available and demand rates.

In addition, this paper draws on general ideas associated with emergency response systems described in Green and Kolesar [3]. For instance, it imposes a service distance limit for each level of order priority (or severity), and service center capacity limits for every combination of priority and time shift. It also includes fixed and overhead costs at service sites for each level of priority. The model developed is quite general and variants may be useful in the design of telecommunications and other networks where queuing considerations are relevant.

3. Service model formulation

The model is based on assumptions which are typical of queuing behavior in service systems. It is assumed that user arrivals are Poisson distributed, and that service time is exponentially distributed. It is also assumed that facilities have sufficiently large buffers relative to the demand that infinite buffer size may be assumed. The model allows for multiple servers at each service site and multiple classes of service corresponding to the priority category of the incoming demand. A service limit, equal to a pre-specified proportion of service site capacity, is imposed at each service site. The rationale for this is the fact that certain types of capacity are shared by different specialties (for instance, physicians and nursing staff on duty have to be shared by programs in hospitals). Thus, capacity constraints on individual programs are needed to avoid excessive allocation of resources to any single program.

The service system is modeled as a set of independent M/M/s queues [22], [23] in which service rates depend on the average service rate of an individual server and the available number of servers. Using these assumptions, and the Poisson distributed demand rates, service times, waiting times, and associated waiting costs can be computed for each combination of service site, priority class, and work shift. The specific formulae used to calculate the service and waiting times are standard and shown in Fig. 2. The index sets, variables, and parameters of the service system model are shown next, followed by the model itself. A discussion follows the model, which is subsequently referred to as Model (P).



The following notation is used in the queuing equations below:

s: Number of servers

λ: Arrival rate of customers at the facility

μ: Service rate of a server at the facility

U: Utilization factor, the percent of time when all servers are busy

P_0: Probability of zero customers being served or in queue

L_q: Average number of customers waiting for service

L: Average number of customers being served or in queue

W_q: Average time a customer waits in queue

W: Average time a customer spends in queue or being served

P_w: Probability that a customer has to wait for service

P_n: Probability of n customers in queue or being served

$$U = \lambda/s\mu)$$

$$P_0 = \left(\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!}\left(\frac{s\mu}{s\mu - \lambda}\right)\right)^{-1}$$

$$L_q = \frac{P_0(\lambda/\mu)^{s+1}}{(s-1)!(s-\lambda/\mu)^2}$$

$$L = L_1 + \frac{\lambda}{\mu}$$

$$W_q = \frac{L_q}{\lambda}$$

$$W = W_q + \frac{1}{\lambda}$$

$$P_w = \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s\left(\frac{s\mu}{s\mu - \lambda}\right)P_0$$

$$P_n = \frac{(\lambda/\mu)^n}{n!}\ P_0,\ \text{for } n \leq s$$

$$P_n = \frac{(\lambda/\mu)^n}{s|s^{(n-s)}}\ P_0,\ \text{for } n > s$$

Fig. 2. Queuing equations for M/M/s queues.

Model index sets:

I	set of service districts, indexed by i
J	set of candidate service centers indexed by j
K	set of order priority levels, indexed by k
L	set of work shifts, indexed by l
S	set of server categories, indexed by s

Model variables:

θ_{ijkl}binary, = 1ifdistrict$i$, priority$k$, shift$l$isassignedtoservicesite$j$, 0otherwise
$\psi_j$binary, = 1ifcandidatecenter$j$isopen, 0otherwise
β_{jkls}binary, = 1ifservercategory$s$appliestocenter$j$, priority$k$, shift$l$, 0otherwise

Model parameters:

μ_{jkl}	service rate for a single server at center j, priority k, shift l
λ_{ikl}	user arrival rate at district i, priority k, shift l
D_{ij}	distance in miles between district i and center j
Sc$_{jkl}$	unit staffing cost at center j, priority k, shift l
Mc_{jkl}	unit processing/assistance cost at center j, priority k, shift l
Tc$_{ijk}$	transportation cost per customer between district i and center j for priority k
Wc$_k$	waiting cost per customer per period for order priority k
Wt$_{jkls}$	average waiting time per customer at center j, priority k, shift l, server category s
α_k	fixed cost for order priority k
σ_k	variable overhead cost for order priority k
Δ_{jkl}	total service capacity at center j, priority k, shift l
ς_{jkl}	capacity proportion allocated at center j, priority k, shift l
Θ	number of service sites (i.e., open candidate centers)

χ_{jkls}	number of servers at center j, priority k, shift l, server category s
ω_k	service distance limit for order priority k
ξ	scale economy parameter

Model (P):

Minimize:

The sum of fixed costs, overhead costs, transportation costs, processing costs, staffing costs, and waiting costs which is equal to

$$\sum_{j\in J}\sum_{k\in K}\sum_{l\in L}(\alpha_k\varsigma_{jkl}\Delta_{jkl})^{\xi}\psi_j+\sum_{i\in I}\sum_{j\in J}\sum_{k\in K}\sum_{l\in L}\{(\sigma_k\lambda_{ikl})^{\xi}+\lambda_{ikl}D_{ij}\mathrm{Tc}_{ijk}+\lambda_{ikl}\mathrm{Mc}_{jkl}\}\theta_{ijkl}$$

$$+\sum_{j\in J}\sum_{k\in K}\sum_{l\in L}\sum_{s\in S}(\chi_{jkls}\mathrm{Sc}_{jkl})\beta_{jkls}+\sum_{i\in I}\sum_{j\in J}\sum_{k\in K}\sum_{l\in L}\sum_{s\in S}(Wt_{jkls}\mathrm{Wc}_k)\lambda_{ikl}\theta_{ijkl}\beta_{jkls}$$

subject to:

(1)

$$\sum_{j\in J}\theta_{ijkl}=1\forall i,k,l,$$

(2)

$$\theta_{ijkl}\leqslant\psi_j\forall i,j,k,l,$$

(3)

$$\sum_{s\in S}\beta_{jkls}\leqslant 1\forall j,k,l,$$

(4)

$$\sum_{s\in S}\beta_{jkls}\geqslant\theta_{ijkl}\forall i,j,k,l,$$

(5)

$$\sum_{i\in I}\lambda_{ikl}\theta_{ijkl}\leqslant\chi_{jkls}\mu_{jkl}\beta_{jkls}\forall j,k,l,s,$$

(6)

$$\chi_{jkls}\mu_{jkl}\beta_{jkls}\leqslant\varsigma_{jkl}\Delta_{jkl}\forall j,k,l,s,$$

(7)

$$D_{ij}\theta_{ijkl}\leqslant\omega_k\forall i,j,k,l,$$

(8)

$$\sum_{j\in J}\psi_j=\Theta,$$

(9)

$$\theta_{ijkl}, \beta_{jkls}, \psi_j \in \{0,1\}.$$

The first term in the objective function represents the fixed costs of service sites. These costs depend on the level of priority, since the processing equipment and operating environment costs more for severe problems or injuries. The scale economy parameter captures savings due to economies of scale (if any). The second term is the sum of service site overhead costs (also incorporating scale economies), transportation costs between service sites and districts (which depend on distance and the level of priority) and processing costs. The third term in the objective function represents server staffing costs. The final term represents waiting costs. It is nonlinear because it contains the product of the variable representing the assignment of a district to a service site for a particular priority and shift combination (i.e., θ_{ijkl})) and the variable representing the number of servers available for the service with the particular priority and shift combination (i.e., β_{jkls})). This term is necessary because the waiting time in a queue, and therefore the associated waiting cost, are dependent on both the rate of arrival of demand at a service site and the number of servers available to meet the demand. The nonlinearity makes the model significantly more difficult to solve than a linear model would be. The model is therefore 'linearized' (i.e., converted to an equivalent linear form) by means of a modeling artifact described in the next section.

The constraints of the model reflect the general operating environment of the service system. Constraints (1) say that, for every combination of priority and shift, a district has to be assigned to exactly one service site. Constraints (2) impose the requirement that an assignment can be made only to an open service center. Constraints (3) limit each combination of service site, priority, and shift, to, at most, one server category (and, therefore, the corresponding number of servers). Constraints (4) indicate that a combination of service site, priority, and shift will have to select a server category and corresponding positive number of servers if any patient or user district is assigned to it. Constraints (5) ensure that, service capacity is greater than the demand rate for any combination of service site, priority, shift, and server category. Since each service site has a limited capacity to process a particular service. Constraints (6) ensure that the service rate for any combination of service site, priority, shift, and server category does not exceed the service capacity assigned to the service. This service capacity is the product of the total capacity at the site and the proportion of capacity assigned to the particular service. Constraints (7) impose the requirement that no assignment should exceed the service distance upper limit of the corresponding level of priority. Finally, Constraints (8) indicate that the number of service sites has to equal a pre-specified number, and Constraints (9) impose binary requirements on the variables of the model.

4. Solution methodology

The linearization of Model (P), the Lagrangian approach to solution of combinatorial problems, the dual Lagrangian model, and the details of the solution methodology are provided next.

4.1. Linearization of Model (P)

As stated earlier, the product of variables in the fourth term of the objective function, $\sum_{i\in I} \sum_{j\in J} \sum_{k\in K} \sum_{l\in L} \sum_{s\in S} (\mathrm{Wt}_{jkls}\mathrm{Wc}_k)\lambda_{ikl}\theta_{ijkl}\beta_{jkls}$, makes the model nonlinear and difficult to solve. The model is therefore converted to an equivalent linear form in the following two steps:

(i) A new binary variable, δ_{ijkls} equal to $\theta_{ijkl}\beta_{jkls}$ is introduced. It replaces the product of variables in the fourth term which therefore becomes $\sum_{i\in I} \sum_{j\in J} \sum_{k\in K} \sum_{l\in L} \sum_{s\in S} (Wt_{jkls}Wc_k)\lambda_{ikl}\delta_{ijkls}$.

(ii) In order to have equivalency between the new and original forms of Model (P), δ_{ijkls} must equal 0 whenever either θ_{ijkl} or β_{jkls} is equal to 0, and it must equal 1 whenever both θ_{ijkl} and β_{jkls} are both equal to 1. These conditions are imposed by introducing a set of four constraints shown below:

(10)

$$\delta_{ijkls} \leqslant \theta_{ijkl} \forall i, j, k, l, s,$$

(11)

$$\delta_{ijkls} \leqslant \beta_{jkls} \forall i, j, k, l, s,$$

(12)

$$\delta_{ijkls} \geqslant \theta_{ijkl} + \beta_{jkls} - 1 \forall i, j, k, l, s,$$

(13)

$$\delta_{ijkls} \in \{0,1\} \forall i, j, k, l, s.$$

The truth table shown in Table 1 establishes the equivalency of δ_{ijkls} to the product of variables, $\theta_{ijkl}\beta_{jkls}$.

Table 1. Truth table for variable δ_{ijkls}

θ_{ijkl}	β_{jkls}	$\theta_{ijkl}\beta_{jkls}$	$\theta_{ijkl} + \beta_{jkls} - 1$	δ_{ijkls}
0	0	0	−1	0
0	1	0	0	0
1	0	0	0	0
1	1	1	1	1

4.2. Lagrangian approach

Model (P) shown in the previous section is a fairly complicated 0–1 integer programming variant of the p-median facility location model (in which exactly p facilities are open). The p-median model is an instance of difficult combinatorial problems known as NP-complete [24], implying that problem difficulty increases rapidly with size. A methodology known to perform well on these kinds of problems is Lagrangian relaxation [25], even though it is heuristic in nature, and has to be highly customized to the specific problem being solved.

The essence of the Lagrangian relaxation methodology is the relaxation of complicating constraints in order to create a problem that is amenable to solution by relatively straightforward methods. The objective function is penalized in proportion to the degree of relaxation by adding the relaxed constraints to the objective function multiplied by coefficients called Lagrangian multipliers. This process results in a problem whose optimal objective value provides a lower bound on the optimal objective value of the original primal problem (P). The Lagrangian dual problem is to maximize the value of the lower bound by making iterative adjustments to the values of the multipliers. While heuristic, an effective method for doing this is the subgradient method (1) for which convergence results are found in Polyak [26]. The subgradient method does not guarantee monotonic improvement in the lower bound, but frequently progresses, over a number of iterations, to improved (i.e., higher) lower bounds.

An upper bound on the optimal objective value of (P) is necessary in order to check for the degree of convergence of the solutions. An upper bound is available from a feasible solution, and a Lagrangian methodology will often generate a feasible solution either at every iteration, or at the end of a pre-specified number of iterations. In the case of the upper bound, it is evident that improvement follows from a lower objective function value. The updated upper and lower bounds are checked frequently for degree of convergence, and the overall procedure is terminated when acceptable convergence is attained. In the case of the (modified) service Model (P), the complicating constraints are (2), (4), (5), (10)–(12) since each of these has more than one variable associated with it. Accordingly, these constraints are relaxed in order to create a relatively amenable lower-bounding model. The details of the lower bounding and upper bounding routines are provided in the subsections that follow.

4.3. Dual problem

The complicating constraints (2), (4), (5), (10)–(12) are relaxed and 'dualized' (i.e., added to the objective function) after associating dual multipliers with each of them as follows:
multipliers ρ_{ijkl}, γ_{ijkl}, φ_{jkls}, εa_{ijkls}, εb_{ijkls}, εc_{ijkls} are associated with Constraints (2), (4), (5), (10)–(12), respectively. The resulting Lagrangian dual problem is the following:
Maximize: $\pi(\rho_{ijkl}, \gamma_{ijkl}, \varphi_{jkls}, \varepsilon a_{ijkls}, \varepsilon b_{ijkls}, \varepsilon c_{ijkls})$ with dual variables ρ_{ijkl}, γ_{ijkl}, φ_{jkls}, εa_{ijkls}, εb_{ijkls}, and εc_{ijkls} non-negative, where $\pi(\rho_{ijkl}, \gamma_{ijkl}, \varphi_{jkls}, \varepsilon a_{ijkls}, \varepsilon b_{ijkls}, \varepsilon c_{ijkls})$ is Model (D1) shown below. In this model, $\boldsymbol{OF}(\boldsymbol{P})$ denotes the objective function of Model (P):

Model (D1): Minimize: $\boldsymbol{OF}(\boldsymbol{P}) + \sum_{i\in I} \sum_{j\in J} \sum_{k\in K} \sum_{l\in L} \{\rho_{ijkl}(\theta_{ijkl} - \psi_j) + \gamma_{ijkl}(\theta_{ijkl} - \sum_{s\in S} \beta_{jkls})\} + \sum_{j\in J} \sum_{k\in K} \sum_{l\in L} \sum_{s\in S} \varphi_{jkls}(\sum_{i\in I} \lambda_{ikl}\theta_{ijkl} - \chi_{jkls}\mu_{jkl}\beta_{jkls}) + \sum_{i\in I} \sum_{j\in J} \sum_{k\in K} \sum_{l\in L} \sum_{s\in S} \varepsilon a_{ijkls}(\delta_{ijkls} - \theta_{ijkl}) + \sum_{i\in I} \sum_{j\in J} \sum_{k\in K} \sum_{l\in L} \sum_{s\in S} \varepsilon b_{ijkls}(\delta_{ijkls} - \beta_{jkls}) + \sum_{i\in I} \sum_{j\in J} \sum_{k\in K} \sum_{l\in L} \sum_{s\in S} \varepsilon c_{ijkls}(\theta_{ijkl} + \beta_{jkls} - 1 - \delta_{ijkls})$

subject to:

(14)

$$\sum_{j\in J} \theta_{ijkl} = 1 \forall i, k, l,$$

(15)

$$\sum_{s\in S} \beta_{jkls} \leqslant 1 \forall j, k, l,$$

(16)

$$\chi_{jkls}\mu_{jkl}\beta_{jkls} \leqslant \varsigma_{jkl}\mathit{\Delta}_{jkl} \forall j, k, l, s,$$

(17)

$$D_{ij}\theta_{ijkl} \leqslant \omega_k \forall i, j, k, l,$$

(18)

$$\sum_{j\in J} \psi_j = \mathit{\Theta},$$

(19)

$$\theta_{ijkl}, \beta_{jkls}, \psi_j, \delta_{ijkls} \in \{0,1\}.$$

It may be observed that Model (D1) is separable by variable, and decomposes into the following four models minus the term $\sum_{i\in I} \sum_{j\in J} \sum_{k\in K} \sum_{l\in L} \sum_{s\in S} \varepsilon c_{ijkls}$. The lower bound is initialized to a very low value and stored in a counter called '$\mathrm{LB_{inc}}$' (i.e., incumbent lower bound) prior to the first iteration of the solution procedure. The lower bound calculated at each iteration will be computed by solving the models below and stored in counter '$\mathrm{LB_{iter}}$'. This counter is initialized to zero at the beginning of each iteration.

Model (D2):

$$\text{Minimize:} \sum_{j\in J}\sum_{k\in K}\sum_{l\in L}\left\{(\alpha_k \varsigma_{jkl}\mathit{\Delta}_{jkl})^{\xi} - \sum_{i\in I}\rho_{ijkl}\right\}\psi_j$$

subject to:

(20)

$$\sum_{j\in J}\psi_j = \mathit{\Theta},$$

(21)

$$\psi_j \in \{0,1\}\forall j.$$

Model (D3):

$$\text{Minimize:} \sum_{j\in J}\sum_{k\in K}\sum_{l\in L}\sum_{s\in S}\left\{\chi_{jkls}\mathrm{Sc}_{jkl} - \sum_{i\in I}(\gamma_{ijkl} - \varepsilon b_{ijkls} + \varepsilon c_{ijkls}) - \phi_{jkls}\chi_{jkls}\mu_{jkl}\right\}\beta_{jkls}$$

subject to:

(22)

$$\sum_{s\in S}\beta_{jkls} \leqslant 1\forall j,k,l,$$

(23)

$$\chi_{jkls}\mu_{jkl}\beta_{jkls} \leqslant \varsigma_{jkl}\mathit{\Delta}_{jkl}\forall j,k,l,s,$$

(24)

$$\beta_{jkls} \in \{0,1\}\forall j,k,l,s.$$

Model (D4):

$$\text{Minimize:} \sum_{i\in I}\sum_{j\in J}\sum_{k\in K}\sum_{l\in L}\left\{(\sigma_k\lambda_{ikl})^{\xi} + (\lambda_{ikl}D_{ij}\mathrm{Tc}_{ijk} + \lambda_{ikl}\mathrm{Mc}_{jkl}) + \rho_{ijkl} + \gamma_{ijkl} + \left(\sum_{s\in S}\phi_{jkls}\right)\lambda_{ikl} + \sum_{s\in S}(\varepsilon c_{ijkls} - \varepsilon a_{ijkls})\right\}\theta_{ijkl}$$

subject to:

(25)

$$\sum_{j\in J}\theta_{ijkl} = 1\forall i,k,l,$$

(26)

$$D_{ij}\theta_{ijkl} \leqslant \omega_k\forall i,j,k,l,$$

(27)

$$\theta_{ijkl} \in \{0,1\} \forall i,j,k,l.$$

Model (D5):

$$\text{Minimize:} \sum_{i \in I} \sum_{j \in J} \sum_{k \in K} \sum_{l \in L} \sum_{s \in S} ((Wt_{jkls} Wc_k)\lambda_{ikl} + \varepsilon a_{ijkls} + \varepsilon b_{ijkls} - \varepsilon c_{ijkls})\delta_{ijkls}$$

subject to:

(28)

$$\delta_{ijkls} \in \{0,1\} \forall i,j,k,l,s.$$

The solution procedures for the subproblems are based on greedy algorithms [29] and described in procedures PD2, PD3, PD4, and PD5 below corresponding to models (D2), (D3), (D4), and (D5), respectively.

Procedure PD2.

(a) Initialize a List LD2 to {null}, and a counter D2T to zero.

(b) For each candidate center (indexed by j), form the coefficient $\mathrm{C2}_j = \sum_{k \in K} \sum_{l \in L} \{(\alpha_k \varsigma_{jkl} \varDelta_{jkl})^{\xi} - \sum_{i \in I} \rho_{ijkl}\}$, and it to List LD2.

(c) Sort List LD2 in ascending order.

(d) Add the first $\varTheta$ coefficients in List LD2 to D2T and set the corresponding ψ_j variables to 1. Set the remaining ψ_j variables to 0.

(e) Add D2T to $\mathrm{LB}_{\mathrm{iter}}$. End the procedure.

Procedure PD3.

(a) Initialize a counter D3T to zero. Then, for each combination of service site (indexed by j), priority (indexed by k), and shift (indexed by l) do the following.

(b) Initialize a list LD3 to {null}.

(c) For each server category (indexed by s), form the coefficient $\mathrm{C3}_{jkls} = \chi_{jkls}\mathrm{Sc}_{jkl} - \sum_{i \in I} (\gamma_{ijkl} - \varepsilon b_{ijkls} + \varepsilon c_{ijkls}) - \phi_{jkls}\chi_{jkls}\mu_{jkl}$.

(d) If $\mathrm{C3}_{jkls} < 0$ and constraint (23) is not violated, add $\mathrm{C3}_{jkls}$ to list LD3.

(e) If list LD3={null}, set $\beta_{jkls} = 0$ for all s and go to the next combination at step (a). Otherwise: identify s^* corresponding to Min{$\mathrm{C3}_{jkls}$} in list LD3, set $\beta_{jkls} = 1$ for $s = s^*$, add the corresponding coefficient to D3T, and set $\beta_{jkls} = 0$ for all remaining s.

(f) If every combination has been processed, add D3T to $\mathrm{LB}_{\mathrm{iter}}$ and end the procedure. Otherwise, go to the next combination at step (a).

Procedure PD4.

(a) Initialize a counter D4T to zero. Next, for each combination of district (indexed by i), priority (indexed by k) and shift (indexed by l) do the following.

(b) Initialize a list LD4 to {null}.

(c) For each service site (indexed by j), form the coefficient $\mathrm{C4}_{ijkl} = (\sigma_k \lambda_{ikl})^{\xi} + (\lambda_{ikl} D_{ij} \mathrm{Tc}_{ijk} + \lambda_{ikl} \mathrm{Mc}_{jkl}) + \rho_{ijkl} + \gamma_{ijkl} + (\sum_{s \in S} \ \phi_{jkls})\lambda_{ikl} + \sum_{s \in S} \ (\varepsilon c_{ijkls} - \varepsilon a_{ijkls})$.

(d) If constraint (26) is not violated, add $\mathrm{C4}_{jkls}$ to the list LD4.

(e) Identify j^* corresponding to $\mathrm{Min}\{\mathrm{C4}_{ijkl}\}$ in list LD4. Set θ_{ijkl} corresponding to $j^* = 1$, and $\theta_{ijkl} = 0$ for all remaining j. Add the coefficient corresponding to j^* to D4T.

(f) If every combination has been processed, add D4T to $\mathrm{LB}_{\mathrm{iter}}$ and end the procedure. Otherwise, go to the next combination at step (a).

Procedure PD5.

(a) Initialize a counter D5T to zero. Next, for each combination of district (indexed by i), service site (indexed by j), priority (indexed by k), shift (indexed by l), and server category (indexed by s) do the following.

(b) Form the coefficient $\mathrm{C5}_{ijkls} = (\mathrm{Wt}_{jkls} \mathrm{Wc}_k)\lambda_{ikl} + \varepsilon a_{ijkls} + \varepsilon b_{ijkls} - \varepsilon c_{ijkls}$. If $\mathrm{C5}_{ijkls} < 0$, add it to D5T, and set $\delta_{ijkls} = 1$, otherwise set $\delta_{ijkls} = 0$.

(c) If every combination has been processed, add D5T to $\mathrm{LB}_{\mathrm{iter}}$, and end the procedure. Otherwise, return to step (a) for the next combination.

The greedy procedures described above make it clear that the continuous versions of the subproblems (in which the binary requirements on variables are replaced by continuous ranges between 0 and 1) would naturally have integer solutions in which each variable has either the value 0 or the value 1. It follows that the model has the integrality property [27], [28] which means that the lower bound found by solving its Lagrangian relaxation will not be tighter than the lower bound found by solving its linear programming relaxation. Nevertheless, the partitioning of Model (D1) leads to subproblems that can be rapidly solved using relatively fast greedy algorithms outlined above.

At the end of each iteration, $\mathrm{LB}_{\mathrm{iter}}$ is compared to $\mathrm{LB}_{\mathrm{inc}}$. If $\mathrm{LB}_{\mathrm{iter}} > \mathrm{LB}_{\mathrm{inc}}$, then the incumbent lower bound is updated, i.e., $\mathrm{LB}_{\mathrm{inc}} = \mathrm{LB}_{\mathrm{iter}}$.

4.4. Updating of dual multipliers

The subgradients [1], [28] of the dual function $\pi(\rho_{ijkl}, \gamma_{ijkl}, \varphi_{jkls}, \varepsilon a_{ijkls}, \varepsilon b_{ijkls}, \varepsilon c_{ijkls})$ are the following:

$$\xi_1(\rho_{ijkl}) = \theta_{ijkl} - \psi_j, \xi_2(\gamma_{ijkl}) = \theta_{ijkl} - \sum_{s \in S} \ \beta_{jkls} \text{and}\ \xi_3(\varphi_{jkls}) \\ = \left(\sum_{i \in I} \ \lambda_{ikl}\theta_{ijkl}\right) - \chi_{jkls}\mu_{jkl}\beta_{jkls}, \xi_4(\varepsilon a_{ijkls}) = (\delta_{ijkls} - \theta_{ijkl}), \xi_5(\varepsilon b_{ijkls}) \\ = (\delta_{ijkls} - \beta_{jkls}) \text{and} \xi_6(\varepsilon c_{ijkls}) = (\theta_{ijkl} + \beta_{jkls} - 1 - \delta_{ijkls}).$$

The dual multipliers are updated in a systematic manner at each iteration using these subgradients and a step size for which convergence results are provided in Polyak [26]. This step size is the following: $\Gamma^i = \delta^i (Iub - Ilb)/\| \eta^i \|^2$, where δ^i is a scalar between 0 and 2, and η^i is the vector of subgradients at iteration i. In the current implementation, δ^i starts at 2.0, and is progressively reduced at each iteration, until a lower limit of 0.10 is reached, when δ^i is restored to 2.0. Iub and Ilb are, respectively, the incumbent upper and lower bounds.

The identification of the optimal values of multipliers is, in general, a difficult task [29] which is why, in practice, most implementations resort to iterative schemes such as the one described above. Accordingly, let Φ represent, in a general way, any of the six dual multipliers, with ϑ as the corresponding subgradient. Then, the value of the multiplier at iteration $i+1$ is the following: $\Phi^{i+1} = \Phi^{i} + \Gamma^{i}(\vartheta)$. The lower bound problem in the next iteration is solved by applying the updated dual multipliers.

4.5. Primal heuristic

The upper bound is initialized at the beginning of the solution procedure to a very high value, $\mathrm{UB}_{\mathrm{inc}}$ (incumbent upper bound). After each iteration of the lower bounding routine, a heuristic is used to generate a (primal) feasible solution from the solution to the lower bound problem, noting that the lower bound solution is generally not feasible for the primal Model (P). This heuristic routine has four steps, described below.

Step 1: (a) Initialize a counter, $\mathrm{UB}_{\mathrm{iter}}$ to zero. This counter will contain the objective function value for the solution generated by the upper bound heuristic. Let V be the set of open service sites at the end of the lower bound routine, that is, $V = \{j: \psi_j = 1\}$.

(b) According to Constraints (1) and (2) of the primal model, for each combination of district, priority, and shift, exactly one θ_{ijkl} binary variable has to equal 1, and the service center has to be from the set of open service sites i.e., j∈V.

(c) Also, Constraint (7) imposes a distance requirement on the choice of the θ_{ijkl} variable. A θ_{ijkl} variable that corresponds to an open service site and does not violate the distance requirement is deemed to be an eligible variable.

(d) For each combination of district, priority, and shift, the upper bound routine loops through all eligible variables finally selecting the variable which has the smallest objective function contribution. This contribution is added to the counter, $\mathrm{UB}_{\mathrm{iter}}$.

Step 2: (a) Constraints (3) and (4) imply that for each θ_{ijkl} variable equal to 1, exactly one β_{jkls} binary variable has to equal 1. For each combination of center, priority, and shift, a number of β_{jkls} variables may be eligible, with eligibility defined as not violating Constraints (5) and (6).

(b) The upper bound routine first determines, for each such combination, whether or not a θ_{ijkl} variable is equal to 1. If so, it identifies the set of eligible β_{jkls} variables. Finally, it loops through all the eligible β_{jkls} variables, choosing exactly one depending on which contributes the least service cost to the primal objective function. The contribution is added to the counter $\mathrm{UB}_{\mathrm{iter}}$.

Step 3: The heuristic loops through each combination of district, site, priority, shift, and server category. For each combination, if θ_{ijkl} and β_{jkls} are both equal to 1, then the corresponding waiting cost is added to the counter $\mathrm{UB}_{\mathrm{iter}}$.

Step 4: The objective function contribution corresponding to all the open service centers, that is, corresponding to $j \in V$, are added to the heuristic objective function counter, $\mathrm{UB}_{\mathrm{iter}}$. At this point, no primal constraint is violated, and the objective function value of the heuristic provides an upper bound to the optimal objective function value of the primal problem. If $\mathrm{UB}_{\mathrm{iter}} < \mathrm{UB}_{\mathrm{inc}}$, then $\mathrm{UB}_{\mathrm{inc}} = \mathrm{UB}_{\mathrm{iter}}$.

The convergence between the incumbent upper and lower bounds is conducted at the end of each iteration. If either (i) the convergence is less than a pre-specified value or (ii) the number of iterations is equal to the maximum permitted number, the solution procedure is terminated. Otherwise, a new iteration is started, which generates a new set of lower and upper bounds. Fig. 3 summarizes the Lagrangian methodology deployed in this paper.

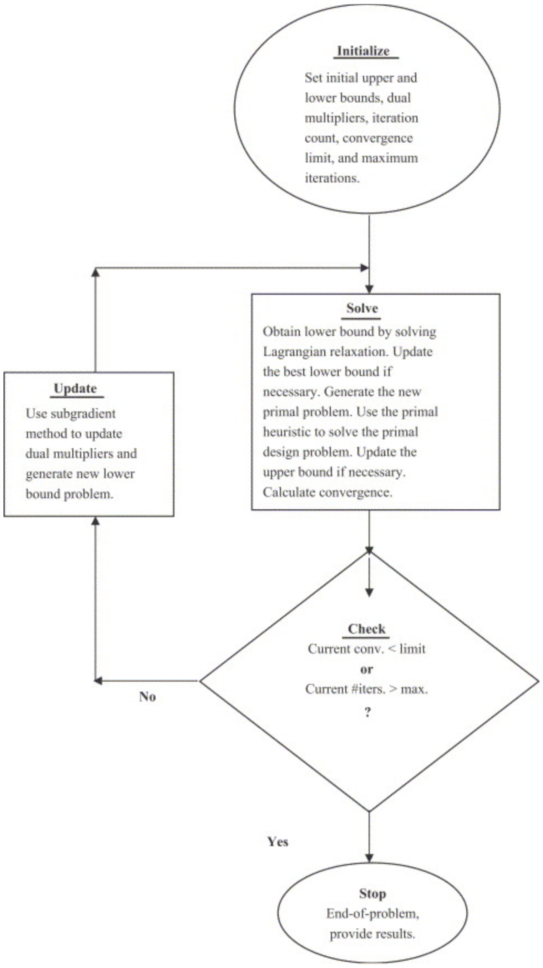

Initialize
Set initial upper and lower bounds, dual multipliers, iteration count, convergence limit, and maximum iterations.
Solve
Obtain lower bound by solving Lagrangian relaxation. Update the best lower bound if necessary. Generate the new primal problem. Use the primal heuristic to solve the primal design problem. Update the upper bound if necessary. Calculate convergence.
Update
Use subgradient method to update dual multipliers and generate new lower bound problem.
Check
Current conv. < limit
or
Current #iters. > max.
?
No
Yes
Stop
End-of-problem, provide results.

Fig. 3. Lagrangian methodology for system design.

5. Computational study and managerial implications

The focus of the computational study conducted for this paper is understanding the policy implications of key managerial and environmental parameters. A managerial parameter is within management's control, while an environmental parameter is determined in the short run by exogenous factors. These implications are discussed after a description of the general parameters of the study. The data for the study are simulated, based on uniform distributions between assumed minimum and maximum values of the relevant parameters. For instance, at the base level of capacity, the number of available servers, for each combination of service site, shift, and order priority, is distributed uniformly between two and six servers. The assumed values for data parameters are provided in the Appendix.

The managerial implications of the results found in this study are analyzed after a brief discussion of the solution times and the computational gaps (between upper and lower bounds) associated with problems of varying size. This establishes an approximate correspondence between the size of a problem and the expected solution time and precision. The computational times and gaps reported in Table 2 are the averages for 10 problems for each configuration, solved on a Windows XP personal computer with 1 Gbyte of memory, and a processing clock speed of 3.2 GHz. The problem size is varied by changing (i) the number of service districts, (ii) the required number of service sites, and (iii) the number of candidate centers while keeping the number of shifts per day fixed at three, the number of server categories fixed at three (high, medium, and low), and the number of order priority levels also fixed at three for each problem. The service distance limit for this set of problems is 40 miles.

Table 2. Computational times and bounds

Number of districts	**Required number of facilities**	**Number of candidate centers**	**Average gap % between lower and upper bounds**	**Average comp. time (s)**
40	10	40	0.825	0
60	15	60	0.900	0
100	25	100	0.946	4
150	40	150	0.887	10
200	50	200	1.152	72
250	60	250	1.506	145

The average computational times shown provide a measure of confidence that large versions of the service system problem can be solved in reasonable time. The largest problems solved in this study, with 250 service districts, 60 service sites, and 250 candidate centers were solved with average time about two and a half minutes corresponding to a maximum of 20 iterations in the Lagrangian methodology. The average gaps between upper and lower bounds are also reasonable, averaging 1.5% for the largest problem category investigated and close to 1% for other categories.

From a managerial perspective, it is more interesting and often more important to evaluate alternative scenarios that are dependent on key managerial policies and the service environment. The system design factors investigated in this paper are: (i) the service distance limit between sites and districts; (ii) system configuration, particularly the number of service sites relative to the numbers of service districts and candidate centers; (iii) concentration of capacity and (iv) economies of scale. Concentration and centralization represent quandaries that underlie the design of all production systems, whether in services or manufacturing. The service distance limit is important when service time (in particular, response time) is critical—these include emergency response services such as police, fire-fighting, and ambulatory services. Service response times are usually strongly correlated with the distances between service districts and the sites to which they are assigned. The level of the scale economies parameter is an important environmental parameter. It is dependent on the system technology and is not generally fully within management's short-run control. The managerial experiment examines the implications of alternative values of the scale economies parameter.

Alternative scenarios are evaluated using two broad criteria: the service proportion as measured by the proportion of demand met and total cost, made up of (i) primary cost elements such as fixed costs, service costs, and waiting costs, and (ii) the cost of lost service, discussed below. Alternative centralization/decentralization policies, which correspond to different numbers of service sites and varying average site capacity, generally exhibit variation in the service proportion. This is because a greater number of sites and/or higher site capacity could allow a larger number of users to access the system, given a particular service distance limit. At the same time, smaller numbers of service sites and/or smaller average site capacity are usually associated with lower primary costs because of lower fixed costs, service costs, etc. This study examines the effects of both forms of centralization: (i) reducing the number of service sites while keeping average site capacity constant and (ii) increasing the average capacity of a smaller number of service sites. Similarly, a larger service distance limit is expected to meet a higher proportion of demand, given a fixed number of open facilities. Finally, varying values of the scale economies parameter corresponds to different technologies deployed by the firm, which in turn should have implications for centralization policy. This issue is investigated using three alternative values of the scale economies parameter.

In order to investigate these managerial issues with the model and methodology developed in this paper, a modeling artifact has to be introduced. This is necessitated by Constraints (1)–(6) of Model (P) which, as explained next, force the system to provide sufficient capacity to meet all demand. First, Constraints (1) and (2) require that any combination of district, priority, and shift should be assigned to one open center. Constraints (3) and (4) imply that one server category (for the corresponding priority and shift) in the selected center should serve the assigned demand combination. Finally, Constraints (5) and (6) require the server category chosen to have sufficient capacity to meet the demand corresponding to the assigned demand combination. Since this is true for every demand combination, capacity has to suffice to meet demand for the model as a whole.

In order to examine the impact of shortfalls in capacity corresponding to the alternative system design policies without violating Constraints (1)–(6), all demand that cannot be met by the set of 'real' service sites is allocated to a 'dummy' or 'artificial' service site. This artifact is similar in concept to the 'dummy' supply node idea often used to solve 'unbalanced' transportation problems with the 'balanced' transportation model. The only cost corresponding to these allocations is the cost of lost service, simply the product of a specified lost service penalty and the number of users whose demand for service is not satisfied. The lost service penalty per user is sometimes a precise figure such as lost federal government funding per user (in the case of the VHA) or an estimated figure that serves as a proxy for lost goodwill, lost sales, etc.

The effects of alternative centralization and service distance policies are investigated for a service configuration involving 60 service districts, 60 candidate centers, three priority levels, three shifts, and three server categories. The total demand for service is the same in all scenarios (corresponding to the demand in the 60 districts), which allows the comparison of alternative policies using service proportion and cost criteria. The experiment investigates several combinations of scenarios in which the service distance limit is progressively reduced from 80 to 10 miles in increments of 20 miles and the number of open service sites is progressively reduced from 60 to five depending on the centralization level. The cost and service effects of alternative levels of the scale economies parameter and centralized capacity are also evaluated.

The results reported in Table 3, Table 4, Table 5, Table 6 are the averages obtained from all the scenarios investigated in this study as follows. The scenarios correspond to four factors which are distance limit, system configuration, centralization of capacity, and scale economies. This results in a total of 3000 solved problems obtained as follows: 5 (distance limits) ×5 (configurations) ×4 (capacity levels) ×3 (scale parameters) ×10 (problems/combination). The averages for any particular level of a given factor are generated by holding the factor level constant and including all possible combinations of the other factors. For instance, the average results for the 80 mile distance limit reported in Table 3 are the averages from the 'slice' of 600 problems that are found in the master set of 3000 solved problems when only problems with the distance limit of 80 miles are selected. The intent is to avoid any bias which might occur if results for a factor level are confined to a smaller subset of solved problems. Unfortunately, considerations of space and clarity preclude the reporting of results for every possible combination.

Table 3. Distance limit results

Dist. limit	**Primary cost**	**Lost serv. cost**	**Total cost**	**Relative cost (%)**	**Service (%)**	**Relative serv. (%)**
80	37 281	8709	45 990	100.00	78.91	100.00
60	36 853	10 879	47 731	103.79	73.66	93.35
40	35 661	12 414	48 075	104.53	69.92	88.60
20	34 179	18 859	53 038	115.32	54.33	68.85
10	34 716	28 683	63 399	137.85	30.48	38.62

Table 4. System configuration results

Config.	**Primary cost**	**Lost serv. cost**	**Total cost**	**Relative cost (%)**	**Service (%)**	**Relative serv. (%)**
60–60–60	60 975	9180	70 155	176.79	77.67	100.00
60–30–60	49 300	12 359	61 659	155.38	69.94	90.05
60–20–60	37 908	14 578	52 486	132.27	64.62	83.20
60–10–60	23 468	19 061	42 529	107.18	53.94	69.45

60–05–60	17 139	22 543	39 682	100.00	45.50	58.58

Table 5. Centralization results

Cap. level	**Primary cost**	**Lost serv. cost**	**Total cost**	**Relative cost (%)**	**Service (%)**	**Relative serv. (%)**
1x	60 975	9180	70 155	164.87	77.67	100.00
2x	40 603	13 463	54 066	127.06	67.30	86.65
3x	32 375	16 693	49 069	115.32	59.58	76.71
4x	21 618	20 934	42 551	100.00	49.40	63.60

Table 6. Scale economies results

Scale parm.	**Primary cost**	**Lost serv. cost**	**Total cost**	**Relative cost (%)**	**Service (%)**	**Relative serv. (%)**
0.80	25 676	16 427	42 102	100.00	60.18	95.92
0.85	35 825	15 391	51 216	121.65	62.74	100.00
0.90	45 714	15 908	61 622	146.36	61.47	97.98

In these tables, the first column indicates the problem category, the second column shows the primary cost for the category, the third column provides the lost service cost, the total cost and relative (to the minimum) total cost are shown in columns four and five, and the sixth and seventh columns contain the service proportion and relative (to the maximum) service proportion, respectively.

Table 3 summarizes the effects of alternative service distance limits, ranging from 80 to 10 miles. It shows that a service distance limit of 10 miles corresponds to the highest total cost and the lowest proportion of demand met. In contrast, a service limit of 80 miles corresponds to the lowest cost and the highest service proportion. Other service limits show a progressive increase of cost and a progressive decrease of service as the service limit is reduced from 80 miles. While the results appear to favor relatively longer service distance limits, two important caveats should be kept in mind. First, specialized service systems such as emergency systems often need to impose additional service response time restrictions that favor shorter distance limits. Model (P) provides a very useful device for evaluating the likely cost and service impact of time restrictions. Secondly, the results for a given scenario are significantly impacted by the level of the lost service penalty relative to other costs. Again, Model (P) provides the decision-maker with the capability to evaluate alternative scenarios involving varying relative levels of the lost service penalty.

The results corresponding to alternative system configurations are provided in Table 4. It is seen that the 60 service site system (all systems have 60 service districts and 60 candidate centers) is associated with the highest service proportion, about 78%. However, the total cost is also the highest for this configuration. The table also shows that the five service site configuration provides the lowest total cost and the lowest service level. Overall, the service–cost tradeoff appears to tilt in favor of the 20 center system. This follows from the observations that the 20 site system is (i) about 32% costlier than a five center system and serves about 25% more users and (ii) about 44% cheaper than a 60 site system while serving about 17% more users.

Table 5 shows the results corresponding to problems which have different levels of average service site capacity combined with a varying number of service sites. In effect, each problem category represents a level of capacity concentration. The first category corresponds to the base capacity level, with the number of servers varying between two and five for each combination of site, priority, and shift, with corresponding levels of site capacity (the Appendix provides details). The base category is tested on configurations with 60 service sites only. The second category doubles the number of servers and the site capacity and it is applied to configurations with 30 and 20 service sites. The third and fourth category represent high degrees of capacity concentration, three and four times, respectively, as the base level. The third category is tested with configurations with 20 and 10 service sites, while the fourth is applied to configurations with 10 and five service sites. The results in Table 5 indicate that category one is best from a service proportion standpoint, while category four performs best with respect to cost. Categories two and three provide intermediate values for both cost and service and appear to be good choices if extremes with respect to either criteria are to be avoided.

The results with respect to varying levels of the scale economies parameter are presented in Table 6. Three levels of the scale economies parameter are compared, representing various levels of increasing returns to scale. This reflects situations where higher levels of centralization lead to fixed and overhead cost savings to a greater or smaller degree, at least for the capacity ranges being considered. As noted earlier, the scale economy parameter is viewed as an environmental parameter (at least in the short run) that reflects the technology deployed. Technologies corresponding to decreasing or constant returns to scale can be captured by a negative or unitary value of the scale economies parameter. Table 6 shows that the lowest level of the parameter corresponds to the lowest total cost by a significant margin. It also indicates that service proportions are about the same for the three parameter values investigated. The cost results are subject to a caveat though—because this study views technology as constant in the short run, the differences in initial investment costs between technologies are not incorporated in the model. The fixed and overhead costs in the model reflect the costs corresponding to various levels of priority or severity rather than technology. Thus, Model (P) would have to be modified to a small degree in order to capture differences between the costs of sophisticated versus simpler technologies.

Table 3, Table 4, Table 5, Table 6 represent different analytical perspectives that assist managers in decision-making. Managers can focus on perspectives that are of high importance in a particular context. As is generally the case with decision support systems, management may have to incorporate additional information that is not contained in Model (P), depending on the particular service environment. For instance, budgetary limitations and/or service response time restrictions may limit the choices available to management. The analysis summarized in Table 3, Table 4, Table 5, Table 6 provides the decision-maker with information regarding the tradeoffs that are an inevitable part of system design, and helps him/her to have a better understanding of the cost/service ramifications associated with alternative designs.

6. Conclusions

This paper has presented a comprehensive model for formulating the service system design problem and a methodology for solving the model based on Lagrangian relaxation. This model is a variant of both the p-median and the location set covering models. It is shown that the methodology can solve fairly large instances of the problem in reasonable computing times.

The model incorporates several features that are commonly encountered in service networks. These include multiple levels of servers, order priority levels, service districts, work shifts, service sites, and candidate centers. Service demand is based on the Poisson distribution, and service times are exponentially distributed. The model includes a wide range of relevant costs such as fixed and overhead costs, waiting costs, traveling costs, service costs, and processing costs. It incorporates service capacity limits at service sites and economies of scale that are applied to fixed and overhead costs. The cost of lost sales (or lost service) is included in the analysis by means of a modeling artifact.

An extensive managerial experiment is conducted to study the implications of alternative policies with respect to key managerial parameters such as the maximum service distance limit and the levels of system centralization and concentration. The implications of varying levels of economies of scale are also discussed. This experiment is described in detail in Section 5. It is seen that the system parameters have significant implications for the key evaluation criteria used in this study which are system cost and the system service proportion which is the proportion of service demand met.

While the model developed in this study provides important insights for managerial decision-making, its limitations provide some direction for future research in service system design. Ideas for extending the current research include incorporation of (i) service response time restrictions; (ii) modeling of dispersed capacity systems such as police patrols; (iii) general service time distributions; (iv) finite queues at service facilities and (v) initial investment costs of alternative technologies. It is believed that the current model provides a solid foundation for a managerial decision support system for analyzing service system design problems, and also for the inclusion of additional complexities and nuances as listed above.

Appendix A. Data parameters

Data for this study are simulated on the basis of uniform distributions. The ranges for these distributions are as follows:

•Number of servers at center/priority/shift: U(2,6) for the base case, multiples of two, three, and four of the base case for cases 2x,3x, and 4x in Table 5.

•Service rate per server at center/priority/shift: U(6,9) for priority one, 90% and 81% of U(6,9) for priority two and three, respectively.

•Arrival rate at district/priority/shift: U(3,5) for priority one, 90% and 81% of U(3,5) for priority two and three, respectively.

•Processing cost per unit at center/priority/shift: U($20,$40).

•Staffing cost per unit at center/priority/shift: U($10,$20).

•Transportation cost per customer for district/center/priority: U($10,$20) for priority one, 90% and 81% of U($10,$20) for priority two and three, respectively.

•Lost service cost per unit: $25 000 per lost unit.

•Fixed cost per unit capacity: U($4,$10) for priority one, 90% and 81% of U($4,$10) for priority two and three, respectively.

•Overhead cost per unit capacity: U($2,$8) for priority one, 90% and 81% of U($2,$8) for priority two and three, respectively.

References

[1] M.L. Fisher. **An applications oriented guide to Lagrangian relaxation.** Interfaces, 15 (1985), pp. 10-21

[2] R.W. Archibald, R.B. Hoffman. **Introducing technological change in a bureaucratic structure.** New York City-RAND Institute, P-4025 (1) (1969), pp. 1-20

[3] L.V. Green, P.J. Kolesar. **Improving emergency responsiveness with management science.** Management Science, 50 (8) (2004), pp. 1001-1014

[4] C.S. ReVelle, R. Swain. **Central facilities location.** Geographical Analysis, 2 (1970), pp. 30-42

[5] C.S. ReVelle, D. Marks, L.C. Liebman. **An analysis of private and public sector location models.** Management Science, 16 (1) (1970), pp. 692-707

[6] C. Toregas, R. Swain, C.S. ReVelle, L. Bergman. **The location of emergency services facilities.** Operations Research, 19 (1971), pp. 1363-1373

[7] R.L. Church, C.S. ReVelle. **The maximal covering location problem.** Papers of the Regional Science Association, 32 (1974), pp. 101-118

[8] R.D. Galvao, C.S. ReVelle. **A Lagrangean heuristic for the maximal covering location problem.** Location Science, 88 (1996), pp. 114-123

[9] S. Rahman, D.K. Smith. **Use of location–allocation models in health service development planning in developing nations.** European Journal of Operational Research, 123 (1) (2000), pp. 437-452

[10] R.C. Larson. **Hypercube queueing model.** Encyclopedia of operations research and management science, Kluwer, Boston (2001)

[11] Swersey AJ. The deployment of police, fire and emergency medical units. In: Pollock SM, Rothkopf M, Barnett A, editors, Handbooks in operations research and management science, vol. 6. New York: North-Holland; 1994. p. 151–90.

[12] O. Berman, Z. Drezner, G.O. Wesolowsky. **The service site and transfer points location problem.** International Transactions in Operational Research, 12 (4) (2005), pp. 387-402

[13] D. Sinreich, Y. Marmor. **Emergency department operations: the basis for developing a simulation tool.** IIE Transactions, 37 (3) (2005), pp. 233-245

[14] P. Beraldi, M.E. Bruni, D. Conforti. **Designing robust emergency medical service via stochastic programming.** European Journal of Operational Research, 158 (1) (2004), pp. 183-193

[15] S.J. D'Amico, S.J. Wang, R. Batta, C.M. Rump. **A simulated annealing approach to police district design.** Computers & Operations Research, 29 (6) (2002), pp. 667-684

[16] V. Marianov, C. ReVelle. **The queueing maximal availability location problem: a model for the siting of emergency vehicles.** European Journal of Operational Research, 93 (1) (1996), pp. 110-120

[17] D.L. Bakuli, J.M. Smith. **Resource allocation in state-dependent emergency evacuation networks.** European Journal of Operational Research, 89 (3) (1996), pp. 543-555

[18] B. Rajan, N.R. Mannur. **Covering-location models for emergency situations that require multiple response units.** Management Science, 36 (1) (1990), pp. 16-23

[19] H. Pirkul, D.A. Schilling. **The siting of emergency service facilities with workload capacities and backup service.** Management Science, 34 (7) (1988), pp. 896-908

[20] A. Amiri. **Solution procedures for the service system design problem.** Computers & Operations Research, 24 (1) (1997), pp. 49-61

[21] A. Amiri. **The design of service systems with queueing time cost, workload capacities and backup service.** European Journal of Operational Research, 104 (1) (1998), pp. 201-217

[22] L. Kleinrock. **Queuing systems.** Wiley Interscience, New York (1976)

[23] C.T. Ragsdale. **Spreadsheet modeling & decision analysis.** (4th ed.), South-Western Thomson Learning, Ohio (2004)

[24] M.R. Garey, D.S. Johnson. **Computers and intractability: a guide to the theory of NP-completeness.** Freeman, San Fransisco (1979)

[25] A.M. Geoffrion. **Lagrangian relaxation for integer programming.** Mathematical Programming, 2 (1974), pp. 2-82-114

[26] B. Polyak. **A general method for solving extremum problems.** Soviet Mathematics Doklady, 1 (1967), pp. 593-597

[27] M.S. Bazaraa, J.J. Goode. **A survey of various tactics for generating Lagrangean multipliers in the context of Lagrangean duality.** European Journal of Operational Research, 3 (1) (1979), pp. 322-328

[28] B. Gavish. **On obtaining the 'best' multipliers for a Lagrangean relaxation for integer programming.** Computers & Operations Research, 5 (1) (1978), pp. 55-71

[29] G.L. Nemhauser, L.A. Wolsey. **Integer and combinatorial optimization.** Wiley, New York (1988)