Computers & Operations Research 48 (2014) 20–30

Contents lists available at ScienceDirect

Computers & Operations Research

journal homepage: www.elsevier.com/locate/caor

Efficient solution of a class of location–allocation problems with stochastic demand and congestion

Navneet Vidyarthi [a,*], Sachin Jayaswal [b]

[a] Department of Supply Chain and Business Technology Management, John Molson School of Business, Concordia University, Montreal, QC H3G 1M8, Canada
[b] Production and Quantitative Methods, Indian Institute of Management, Vastrapur, Ahmedabad, Gujarat 380 015, India

ARTICLE INFO

Available online 12 March 2014

Keywords:
Location–allocation
Service system design
Queueing
Stochastic demand
Congestion
Constraint generation method

ABSTRACT

We consider a class of location–allocation problems with immobile servers, stochastic demand and congestion that arises in several planning contexts: location of emergency medical clinics; preventive healthcare centers; refuse collection and disposal centers; stores and service centers; bank branches and automated banking machines; internet mirror sites; web service providers (servers); and distribution centers in supply chains. The problem seeks to simultaneously locate service facilities, equip them with appropriate capacities, and allocate user demand to these facilities such that the total cost, which consists of the fixed cost of opening facilities with sufficient capacities, the access cost of users' travel to facilities, and the queuing delay cost, is minimized. Under Poisson user demand arrivals and general service time distributions, the problem is set up as a network of independent M/G/1 queues, whose locations, capacities and service zones need to be determined. The resulting mathematical model is a non-linear integer program. Using simple transformation and piecewise linear approximation, the model is linearized and solved to ϵ-optimality using a constraint generation method. Computational results are presented for instances up to 400 users, 25 potential service facilities, and 5 capacity levels with different coefficients of variation of service times and average queueing delay costs per customer. The results indicate that the proposed solution method is efficient in solving a wide range of problem instances.

1. Introduction

Problems arising in several planning contexts require deciding: (i) the location of service facilities and their capacities; and (ii) allocation of service zones to the located service facilities. Examples include location of emergency service facilities such as medical clinics and preventive health care facilities [25,26,27]; stores and service centers; bank branches and automated banking machines [1,11,23]; automobile emission testing stations [11]; web service providers' facilities [3]; proxy/mirror servers in communication networks [24] and distribution centers in supply chains [17,22]. All the above examples are characterized by servers (medical clinics, bank branches, distribution centers, etc.) that are immobile in that the customers need to travel to the service facilities to avail of their services, as opposed to the servers traveling (mobile servers) to the customers' site in response to calls for their services. Such problems are generally also characterized by random nature of service calls (demand arrivals) and their service requirements (service times). These problems are commonly known in the literature as facility location problems with immobile servers, stochastic demand and congestion [8]. They are also termed as service system design problems with stochastic demand and congestion [4–6,13]. Literature review for this class of problems is provided by Berman and Krass [8] and Boffey et al. [10].

For facility location problems with stochastic demand and congestion, the following two factors are important: (i) the costs of providing service; and (ii) the quality of service, with an objective generally requiring a balance between the two. The costs of providing service are related to the fixed cost of opening/operating the service facilities and the cost of accessing these facilities by the users. The service quality, on the other hand, is often measured in terms of: (i) the average number of users waiting for service; (ii) average waiting time per user; or (iii) the probability of serving a user within a time limit [13]. Balance between service costs and service quality is commonly achieved in the literature using a combination of the total cost of opening and accessing facilities and the cost associated with waiting customers, which is minimized in the objective function [4,5,11,13,23]. Others in the literature minimize the cost of providing service subject to a minimum threshold on the service quality, where the service quality may be defined in one of the ways described above [18,19,21].

[*] Corresponding author. Tel.: +1 514 848 2424x2990; fax: +1 514 848 2824.
E-mail addresses: navneetv@jmsb.concordia.ca (N. Vidyarthi), sachin@iimahd.ernet.in (S. Jayaswal).

http://dx.doi.org/10.1016/j.cor.2014.02.014
0305-0548 © 2014 Elsevier Ltd. All rights reserved.

In this paper, we use the former of the two approaches described above, i.e., we consider minimization of the total cost, which includes the cost of opening and accessing facilities and the cost associated with waiting customers. It is worth noting that due to the complexity of the underlying problem, most papers in this category make assumptions such as (i) either the number or capacity of the facilities (or both) is fixed; (ii) the demand arrival process is Poisson; and (iii) the service times follow an exponential distribution (see [1,4,13,19,23] and references therein). Despite these simplifying assumptions, the techniques proposed to date to solve the problem, with the exception of Elhedhli [13], are either approximate or heuristic based.

The contribution of this paper is twofold. First, by assuming a general distribution for the service times at facilities, as opposed to exponential distribution, we present a more generalized model of the problem than available in the extant literature. More specifically, our proposed model seeks to determine the minimum cost system-optimal configuration (location of service facilities and their capacity levels as well as the allocation of service zones to these facilities) of a service system under Poisson arrivals and general service time distribution, where the total cost consists of the costs of opening and accessing service facilities and the cost associated with waiting customers. As discussed above, the problem, even with the simplifying assumption of exponential service time distribution, is too difficult to solve using exact methods. The proposed model, with general service time distribution, is even more challenging to solve. So, our second contribution lies in the exact (ϵ-optimal) solution method that we propose to solve our model. Our proposed solution method is based on a simple transformation and piecewise linearization of our non-linear integer programming (IP) model, which is solved to optimality (or ϵ-optimality) using a constraint generation algorithm.

The remainder of the paper is organized as follows. In Section 2, we describe the problem setting, followed by its non-linear IP model. Section 3 describes the transformation and the piecewise linearization approach for the non-linear IP model. To solve the linearized model, we present a constraint generation based solution approach in Section 4. Computational results are reported in Section 5. Section 6 concludes with some directions for future research.

2. Problem formulation

Consider a set of user nodes, each indexed by $i \in I$ whose demand for service occurs continuously over time according to an independent Poisson process with rate λ_i. We consider a *directed choice* environment, where users are assigned to facilities, each indexed by $j \in J$, by a central decision maker. This is applicable, for example, in the case of a "virtual call center" consisting of geographically dispersed telephone call centers, routing of calls to which is centrally determined (see [11], and references therein). The directed choice model is also applicable in the case of medical clinics and preventive health care facilities; automobile emission testing stations; and distribution centers in supply chains, if users' choice can be influenced through imposition of tolls or differential service fees. Later, we show how our model can be adapted to the *user choice* environment where the choice of service facility is not dictated or influenced by the central authority but exercised solely by the users. Recent studies of models with directed choice settings include Aboolian et al. [2] whereas models with user choice settings can be found in Baron et al. [7,27], and references therein.

We assume that users from any node are entirely assigned to a single service facility, where each facility operates with an infinite buffer to accommodate users waiting for service. If x_{ij} is a binary variable that equals 1 if the demand for service from user node i is satisfied by facility j, and 0 otherwise, then the aggregate demand arrival rate at facility j, as a result of the superposition of Poisson processes, also follows a Poisson process with mean $\Lambda_j = \sum_{i \in I} \lambda_i x_{ij}$ [15].

There are two approaches to model the capacity of a service facility [2,7]. One is to model the given service facility as a single server with flexible service capacity μ, which can be adjusted either continuously or in discrete steps. The second approach is to assume multiple parallel servers, each with a given single capacity level μ. In this case, the decision variable is the appropriate number of servers to be installed at the given service facility. In the case of call centers, automated banking machines , automobile emission testing stations, or distribution centers, where adding capacity would imply adding a call center employee, a banking machine, a testing station, or a loading/unloading dock respectively, the multiple server model is more appropriate. However, in cases where it is not clear what a "server" represents (e.g. hospitals or emergency medical clinics), and the capacity can be increased in a variety of ways (by improving patient flow or technology; adding nurses, doctors, support staff or examination rooms, etc.), single server model would be suitable. In this paper, we adopt the former approach, and model each facility as a single server with multiple capacity levels, from which one capacity level is to be selected, if the facility is opened. We take this approach primarily for tractability of the resulting model. However, a single server model may still be a good approximation of a multi-sever facility if the utilization of the service facility is reasonably high. This is because under reasonably high system utilization, a system with s parallel servers, each with capacity μ, is known to perform similar to a single server with capacity $s\mu$.

For each service facility, we allow the option of selecting one of the several capacity levels $\mu_{jk}, k \in K$ with fixed cost f_{jk} (amortized over the planning period). Let y_{jk} be a binary variable that equals 1 if facility at site j is open and equipped with a capacity level $k \in K$, 0 otherwise. Further, assume that the service times at any facility j are independent and identically distributed with a mean $1/\mu_{jk}$ and variance σ^2_{jk} if it is equipped with a capacity level k. Each facility j is thus modeled as an $M/G/1$ queue with a service rate $\mu_j = \sum_{k \in K} \mu_{jk} y_{jk}$ and variance in service times given by $\sigma^2_j = \sum_{k \in K} \sigma^2_{jk} y_{jk}$. Thus, the service system design problem is modeled as a network of independent M/G/1 queues.

Under steady state conditions ($\Lambda_j/\mu_j < 1$), first-come-first-serve (FCFS) queuing discipline, and infinite buffers to accommodate users waiting for service, the expected waiting time (including the time spent in service) of users at facility j is given, by the Pollaczek–Khintchine formula [15], as

$$E[w_j] = \left(\frac{1+Cv_j^2}{2}\right)\frac{\tau_j \rho_j}{1-\rho_j} + \tau_j = \left(\frac{1+Cv_j^2}{2}\right)\frac{\Lambda_j}{\mu_j(\mu_j - \Lambda_j)} + \frac{1}{\mu_j} \qquad (1)$$

where $\tau_j = 1/\mu_j$ is the average service time at facility j, $\rho_j = \Lambda_j/\mu_j$ is the average utilization of facility j, and $Cv_j = \sigma_j \mu_j$ is the coefficient of variation of service times at facility j. $E[w_j]$ can be written in terms of location and allocation variables (y_{jk} and x_{ij}) as

$$E[w_j(\mathbf{x},\mathbf{y})] = \frac{\left(1+\sum_{k \in K} Cv_{jk}^2 y_{jk}\right)\sum_{i \in I}\lambda_i x_{ij}}{2\sum_{k \in K}\mu_{jk} y_{jk}\left(\sum_{k \in K}\mu_{jk} y_{jk} - \sum_{i \in I}\lambda_i x_{ij}\right)} + \frac{1}{\sum_{k \in K}\mu_{jk} y_{jk}} \qquad (2)$$

The expected number of users in service or waiting for service at facility j is given, using Little's law, as $\Lambda_j E[w_j]$. If d denotes the average waiting time cost per customer (henceforth called unit queuing delay cost), then the *total delay/congestion cost* in the network can be expressed as $d\sum_{j \in J}\Lambda_j E[w_j(\mathbf{x},\mathbf{y})] = d\sum_{j \in J}\sum_{i \in I}\lambda_i x_{ij} E[w_j(\mathbf{x},\mathbf{y})]$. We assume there is a variable access cost c_{ij} of providing service to users at node i from facility at site j. The problem is to