PERGAMON

Computers & Operations Research 30 (2002) 87–96

computers & operations research

www.elsevier.com/locate/dsw

A note on solutions to the maximal expected covering location problem

Fernando Y. Chiyoshi[a], Roberto D. Galvão[a], Reinaldo Morabito[b,*]

[a] *Programa de Engenharia de Produção, COPPE/UFRJ, Caixa Postal 68507, 21945-970 Rio de Janeiro, RJ, Brazil*

[b] *Departamento de Engenharia de Produção, Universidade Federal de Sao Carlos, Caixa Postal 676, 13565-905 São Carlos, SP, Brazil*

Received 1 September 2000; received in revised form 1 May 2001

Abstract

The maximal expected covering location problem (MEXCLP) and its adjusted counterpart (AMEXCLP) compute expected coverage arising only from unqueued calls, whereas the interactive use of the hypercube queueing model (HQM) considers both unqueued and queued calls in this computation. In this note we show that the three models are not strictly comparable because of the structural differences in their objective functions and that, when using HQM, it is important to state clearly the factor being used to express traveling time in terms of service time units.

Scope and purpose

The maximal expected covering location problem (MEXCLP) addresses, under three simplifying assumptions, the problem of optimally locating servers so as to maximize the expected coverage of demand, considering the possibility of server unavailability when a call enters the service system. These simplifying assumptions were later relaxed by Batta et al. (Transport. Sci. 23 (1989) 277) through the use of an adjusted model (AMEXCLP) and the interactive use of the hypercube queueing model (HQM); these authors compared the expected coverage computed using the three corresponding objective functions. In the present note we address two points in this previous work which, in our opinion, deserve a more detailed discussion: (i) the models are not strictly comparable because of structural differences in their objective functions; (ii) when using HQM it is important to state clearly the factor being used to express traveling time in terms of service time units.

Keywords: Covering location problems; Hypercube queueing model; Maximal expected coverage

* Corresponding author. Fax: +55-16-2608240.

E-mail address: morabito@power.ufscar.br (R. Morabito).

0305-0548/02/$ - see front matter
PII: S0305-0548(01)00083-1

1. Introduction

The maximal expected covering location problem (MEXCLP) was proposed by Daskin [2] as an extension of the maximal covering location problem (MCLP) formulated by Church and ReVelle [3], to account for the possibility of server unavailability due to a congested system. MEXCLP addresses the problem of optimally locating servers so as to maximize the expected coverage of demand, considering the possibility of server unavailability when a call enters the service system. When formulating MEXCLP Daskin [2] makes 3 simplifying assumptions: servers operate independently, each server has the same busy probability and server busy probabilities are invariant with respect to their location.

Batta et al. [1] question the validity of these assumptions and attempt to relax all three of them simultaneously by using an iterative version of the hypercube queueing model (HQM), which we call the hypercube location model (HLM). As expected coverage is not one of the outputs of the queueing model, they start by developing an expression for expected coverage given the standard outputs of HQM. Their next objective is to optimize server locations so as to maximize the system-wide expected coverage.

To fulfill their goal they use a single node substitution heuristic procedure which uses HQM to compute expected coverage at each iteration; the iterative process is repeated until no further improvement in the expected coverage is possible by making single node substitutions from the current set of server locations. In order to solve the equilibrium equations of the hypercube queueing model at each iteration of the heuristic procedure, Batta et al. [1] use an approximate procedure (based upon the methodology developed by Larson [4]), as opposed to the exact methodology originally developed by Larson [5].

These authors also propose an adjusted MEXCLP model (which they call AMEXCLP), where they relax the assumption that servers are independent by using the correction factors proposed by Larson [4]. Batta et al. compare the expected coverage obtained by the MEXCLP model of Daskin with corresponding values obtained through AMEXCLP and their proposed hypercube optimization (single node substitution) procedure, using a 55-node and 3 server test problem, with nodal locations and weights taken from Swain [6].

This note addresses two points in the work of Batta et al. [1] which, in our opinion, deserve more detailed discussion. The first point is the fact that the models are not strictly comparable: by analyzing their objective functions it can be seen that, while both MEXCLP and AMEXCLP are restricted to coverage arising from unqueued calls, HLM takes into account unqueued as well as queued calls to predict the expected coverage. Unless the system is operating at a very low overall workload or under very restrictive cover constraints, in which case no significant contribution to coverage is expected from queued calls, the models should produce different expected coverages due to the very nature of their objective functions. The second point is related to the fact that the process of evaluating the contribution of queued calls to expected coverage in the hypercube location model involves two time variables, traveling times and service times. Since different assumptions concerning the relationship between these two variables lead to different predictions of expected coverage, it seems important to state clearly the factor being used to express traveling time in terms of service time units.

2. Comparison of the objective functions

In the models discussed by Batta et al. there are differences among the objective functions, resulting from the different sets of underlying hypotheses and from the extended capability of HLM. In order to identify precisely where the differences in coverage predicted by three models lie, we use a 3-server system and begin by writing the expression of the expected coverage for node i in each case. Except for the introduction of variable v_{ik} in Eq. (3) below, so that the differences between the three models are seen more clearly, we generally use the notation of Batta et al. [1], which is repeated here for the sake of convenience. Let:

E_i	expected coverage of node i,
p	server busy probability,
y_{ki}	1 if node i is covered by at least k servers, 0 otherwise,
h_i	weight of demand generated at node i,
M	total number of servers to be located ($M=3$ in the example),
$Q(M,p,j)$	correction factors defined by Larson [4] to relax the assumption that servers are independent,
P_s	probability that a randomly arriving call finds all servers busy (probability of saturation),
x_{ik}	fraction of *unqueued* calls from node i that are assigned to server k,
v_{ik}	1 if node i is covered by server k, 0 otherwise,
Z_{ik}	travel time from node i to server's k current location,
d	*critical covering time*, the time beyond which a node i is considered *not covered.*

The objective functions are:

(i) **MEXCLP**

$$E_i = \{(1-p)y_{1i} + (1-p)py_{2i} + (1-p)p^2y_{3i}\}h_i. \tag{1}$$

(ii) **AMEXCLP**

$$E_i = \{(1-p)y_{1i} + (1-p)pQ(3,p,1)y_{2i} + (1-p)p^2Q(3,p,2)y_{3i}\}h_i. \tag{2}$$

(iii) **HLM**

$$E_i = \{(1-P_s)x_{i1}v_{i1} + (1-P_s)x_{i2}v_{i2} + (1-P_s)x_{i3}v_{i3}\}h_i + h_iP_s \sum_{kstZ_{ik} \leqslant d} \text{Prob}\{\text{Queueing delay} \leqslant d - Z_{ik}\}/(M=3). \tag{3}$$

Note that in (3) the first term refers to unqueued calls, whereas the second term is related to queued calls.

It can be seen that, in contrast with HLM, both MEXCLP and AMEXCLP are assuming no contribution from queued calls to total expected coverage. It seems therefore more appropriate to compare the expected coverage predicted by MEXCLP and AMEXCLP with the expected coverage predicted by HLM *only* for unqueued calls (first term of (3)). This also raises the

Table 1
Calculation of the fraction of unqueued covered calls for the hypercube location model (nodes covered by 3 servers)

Dispatch vector	Server 1	Server 2	Server 3	Total
123	0.1730	0.0814	0.0636	0.3180
132	0.1730	0.0424	0.1026	0.3180
213	0.0931	0.1613	0.0636	0.3180
231	0.0425	0.1613	0.1142	0.3180
312	0.0735	0.0424	0.2021	0.3180
321	0.0425	0.0735	0.2021	0.3180

question on whether it would be more appropriate to use the zero-line-capacity hypercube queueing model for comparison with results produced by MEXCLP and AMEXCLP.

For a given node i, the expected coverage depends on the number of servers whose traveling time to node i does not exceed the critical covering time d. If node i is covered by all three servers, the expected fraction of covered calls will be:

- $(1-p)+(1-p)p+(1-p)p^2$ for MEXCLP;
- $(1-p)+(1-p)pQ(3,p,1)+(1-p)p^2Q(3,p,2)$ for AMEXCLP, and
- $(1-P_s)x_{i1}+(1-P_s)x_{i2}+(1-P_s)x_{i3}$ for HLM.

For illustration purposes we take the 55 node, 3 server test problem of Batta et al. [1], with system-wide workload of $p=0.8212$ and consider the 3 servers located at nodes 7, 9 and 19. The hypercube queueing model associated with this system was solved exactly, using the Gauss–Siedel method to solve the linear system of equations related to the state probabilities. In evaluating the expected coverage under HLM for unqueued calls, we note that for a given dispatch vector, say $\{123\}$, the term $(1-P_s)x_{i1}$ is just the probability that server 1 is free

$$(1-P_s)x_{i1}=P\{000\}+P\{010\}+P\{100\}+P\{110\}.$$

The next term, $(1-P_s)x_{i2}$, the probability that the server 1 is busy and server 2 is free, can be evaluated as

$$(1-P_s)x_{i2}=P\{001\}+P\{101\}.$$

Finally the last term, $(1-P_s)x_{i3}$, the probability that servers 1 and 2 are both busy and server 3 is free, equals the probability $P\{011\}$. The terms of the expected coverage can be evaluated for other dispatch vectors in a similar way. The results are shown in Table 1.

It can be seen that, when all servers are within the critical covering time d from a given node, the expected fraction of unqueued covered calls from that node does not depend on the dispatch vector, since it equals the probability that at least one server is free, which is $(1-P_s)=0.3180$.

The fraction of covered calls in this case is 0.4462 and 0.3189 for MEXCLP and AMEXCLP, respectively. The MEXCLP model predicts a much higher expected coverage than the hypercube location model. In fact, the MEXCLP prediction is 40.3% higher than that of HLM. Its implied saturation probability is correspondingly lower: 0.5538 compared to 0.6820 for HLM. The MEXCLP prediction looks somewhat unrealistic in that, given the system's overall

Table 2
Calculation of the fraction of unqueued covered calls for the hypercube location model (nodes covered by 2 servers)

Dispatch vector	Server 1	Server 2	Server 3	Total
123	0.1730	0.0814		0.2544
132	0.1730		0.1026	0.2756
213	0.0931	0.1613		0.2544
231		0.1613	0.1142	0.2755
312	0.0735		0.2021	0.2756
321		0.0735	0.2021	0.2755

Table 3
Fraction of unqueued covered calls, two server coverage

Covering servers	HLM	MEXCLP	% from HLM	AMEXCLP	% from HLM
1,2	0.2544	0.3256	28.02	0.2694	5.92
1,3	0.2756	0.3256	18.17	0.2694	−2.23
2,3	0.2755	0.3256	18.19	0.2694	−2.22

workload (0.8212), it is unlikely that the response units are servicing that many calls within critical covering time d.

The prediction of AMEXCLP matches closely that of HLM. We recall that AMEXCLP is derived from MEXCLP by relaxing the simplifying assumption that servers operate independently from one another, through the use of the correction factors $Q(M, p, j)$. The overestimation of MEXCLP can therefore be attributed to the actual non-independence of the servers. The closeness between the predictions by AMEXCLP and HLM points, on the other hand, to the soundness of the corresponding underlying assumptions.

When MEXCLP and AMEXCLP are used, for nodes that are not covered by all servers, the terms of the expected coverage that exceed the number of servers within critical covering time d must be dropped. When HLM is used, the dispatch of servers located farther than d must be dropped; the expected coverage given by HLM becomes dependent on the servers in the dispatch vector associated with the node which are within critical covering time d from the node. For nodes covered by two servers, for example, the terms of expected coverage for HLM are given in Table 2. The comparative data for the three models are shown in Table 3.

For two-server coverage MEXCLP is overestimating the expected coverage from 18.17% to 28.02%, when compared with the coverage predicted by HLM. In relation to AMEXCLP, depending on the dispatch vector, the difference in expected coverage can be from −2.22% to +5.92%. Consequently, the system-wide coverage predicted by AMEXCLP does not greatly diverge from that predicted by HLM.

Table 4
Fraction of unqueued covered calls, one server coverage

Preferential server	HLM	MEXCLP, AMEXCLP	% from HLM
1	0.1730	0.1788	3.35
2	0.1613	0.1788	10.84
3	0.2021	0.1788	−11.52

For nodes with single server coverage, the expected coverage predictions are the same for MEXCLP and AMEXCLP, since no term with the Q factor is included in the latter model. As for HLM, the expected coverage depends on the preferential server of the node; the corresponding fractions of covered calls are available from Table 2. The relevant coverage data are shown in Table 4.

Note that, for nodes covered by one server, the fraction of covered calls is just the complement of the workload of the server. In fact, by assuming that all servers have the same workload, both MEXCLP and AMEXCLP are overestimating the expected coverage by 10.84% for the busiest server (located at node 9), and underestimating that of the least busy server (located at node 19) by 11.52%.

According to Larson and Odoni [7], the factor $Q(M, p, j)$ is a monotonically decreasing function of j for systems with a high workload (in fact, for $p > 1 - 2/M$). Consequently, by assuming that the servers operate independently, MEXCLP will overestimate the expected coverage most of the time, the exception being the nodes with single server coverage, in which case the non-independence factor is not in effect.

Larson and Odoni [7] also remark that workload sharing among servers causes the workloads of the units to be more evenly distributed than the workloads of the primary response areas. This supports the conjecture that the assumption that all servers have the same workload (MEXCLP, AMEXCLP) will not significantly affect the quality of the corresponding expected coverage. Closeness between the AMEXCLP and HLM predictions of expected coverage is therefore to be expected, as far as unqueued calls are concerned. In fact, the corresponding expected coverages are 2244.26 for MEXCLP, 1679.85 for AMEXCLP and 1672.28 for HLM (these values were obtained by summing up Eq. (1)–(3), respectively, over all 55 nodes of the example). This brings us to the issue of queued calls.

Since queued calls are not taken into account by MEXCLP and AMEXCLP, we may ask whether the zero-line-capacity hypercube queueing model (instead of the infinite-line-capacity model used by Batta et al.) is not more appropriate for comparison purposes. It should be noted that for the zero-line-capacity model the ratio between arrival rate and service rate is no longer the workload of the system; for this model the workload is given by $(1 - P_S)\lambda/\mu$. This is so because the calls that find all servers busy are assumed to be lost and will not add to the workload of the system. In fact, the saturation probability of the system in this case (measured by Prob{111}) is 0.2772, which is also the fraction of lost calls. The corresponding system workload is therefore $(1 - 0.2772) \times 0.8212 = 0.5936$. The expected coverage then rises to 3801.28, compared with 1956.00 (see Table 5) predicted by the infinite-line-capacity model

Table 5
Expected coverage for different values of T_f, hypercube location model

T_f	Unqueued calls	Queued calls	Total coverage
30	1672.28	371.03	2043.31
40	1672.28	283.72	1956.00
50	1672.28	229.65	1901.93
60	1672.28	192.88	1865.16
70	1672.28	166.26	1838.54
80	1672.28	146.09	1818.37
90	1672.28	130.29	1802.57
100	1672.28	117.57	1789.85

(for a *time factor* of 40, see Section 3). The degree to which these predictions differ suggests that we are dealing with different systems altogether.

The difference above arises from the fact that the infinite-line-capacity HQM assumes that the calls for which travel plus waiting time exceeds the critical covering time d, though not covered, are serviced, adding to the workload of the system. If the same assumption is in effect for MEXCLP and AMEXCLP, these models can be compared to HLM only for unqueued calls.

3. The time factor

In MEXCLP and AMEXCLP there is no apparent way of dealing with the calls that are held in the waiting queue. In HLM, however, under the usual assumptions of Poisson arrivals, exponential service times and a queueing discipline that does not depend on actual service times, the contribution of queued calls to the expected coverage can be easily evaluated. If the FCFS queueing discipline is in effect, the queue waiting time is known to be exponentially distributed. For homogeneous servers, the waiting time distribution has the mean $(M\mu - \lambda)^{-1}$, where μ is the service rate common to all servers and λ is the system arrival rate. If unit service rate is assumed, we have $(M - \lambda)^{-1}$ for the mean waiting time and

$$P\{W \leqslant d - Z_{jk}\} = 1 - \exp\{-(M - \lambda)(d - Z_{jk})\}.$$

In evaluating this probability, the travel times and the critical covering time d must be expressed in units of mean service time. To this end a *time factor* T_f must be introduced

$$P\{W \leqslant d - Z_{ik}\} = 1 - \exp\{-(M - \lambda)(d - Z_{ik})/T_f\}.$$

For the 55 node, 3-server test problem used by Batta et al. [1], with system-wide workload of 0.8212, the expected coverages for different values of T_f are shown in Table 5.

It can be seen that, depending on the assumption regarding the *time factor*, different values of expected coverage are obtained. If expected coverage arising only from queued calls is considered, the differences may be very significant. These results suggest that, when reporting expected coverage evaluated by HLM, it is important to state the *time factor* used. The results reported by Batta et al. [1] suggest that a time factor of 40 was used by the authors.

We note that, for $\Sigma_i h_i = 6400$ used in the test problem, the number of queued calls is $0.6820 \times 6400 = 4364.80$, where 0.6820 is the system's saturation probability. Of these, only 283.72 are covered calls. If a less stringent critical covering time d were used, a more substantial contribution to expected coverage would come from queued calls. In this case there would be less agreement between the expected coverage calculated by AMEXCLP and HLM than in the current case.

4. Factors that may further affect expected coverage

In addition to the model used and, in the case of HLM, to the value used for T_f, as discussed in Sections 2 and 3 above, there are other factors that may influence the computation of expected coverage. For example, we may consider: (i) use of fixed preference schemes which are not based on shortest distances; (ii) use of preference schemes that are not fixed "a priori" (which will be referred to as non-fixed preference schemes); (iii) consideration of the probabilistic nature of travel times. Each of these three issues will be dealt with in some detail in the following paragraphs.

Carter et al. [8] observe that under certain conditions, when the objective is to minimize mean travel times, it may be better not to dispatch the closest available server under a fixed preference scheme. The same observation holds when the focus shifts from travel times to coverage of demand, as it can be seen from the following example.

Let us imagine a two-atom system, with one server located in each atom, and a fixed preference scheme based on shortest distances. Suppose a call arrives from a point closer to server 1 but also within covering time (distance) from server 2. This call is answered by server 1 under the prevailing preference scheme. Suppose now that, while server 1 is busy, a second call arrives from a point within covering time from the location of server 1, but beyond covering time from server 2. This second call may be answered by server 2, but this is not a "covered" call. If, due to a different preference scheme, the first call were answered by server 2 and the second call by server 1, both calls would be considered "covered", resulting in improved expected coverage.

In problems which have as their focus coverage of demand, fixed preference schemes that dispatch the closest available server only if it is within the critical covering time of the calling point would probably outperform, in terms of expected coverage, a preference scheme solely based on shortest distances.

A second point we would like to address is non-fixed preference schemes. Preference schemes may in fact be made dependent on the state of the system when a call arrives; to the best of our knowledge Jarvis [9] was the only author to study non-fixed preference schemes, through the optimization of a Markovian decision process.

The third point we would like to emphasize is the probabilistic nature of travel times. As a matter of fact, in practical situations, coverage of a point by a server is a probabilistic event: the closer the base location of the server is to the demand area, the greater the probability that the call will be answered within the critical covering time.

Goldberg et al. [10] developed a model that addresses some of the limitations of the Daskin/Batta et al. models. This is a non-linear integer programming model that allows probabilistic

travel times and computes "optimal" fixed preference schemes. The authors achieve this by the use of two specific variables. The first is a variable P_{ij} that gives the probability that a server located at base j, when available, can reach demand area i within critical covering time. The second is a decision variable x_{ijk} intended to determine "optimal" fixed preference schemes: $x_{ijk} = 1$ if base j is the kth preferred base for demand area i, $x_{ijk} = 0$ otherwise.

The authors recognize that it is not practical to use the model to determine "optimal" fixed preference schemes. They show, however, that under a fixed preference scheme with variables x_{ijk} determined outside the model, a candidate set of locations for the servers can be evaluated with respect to expected coverage by using the P_{ij} probabilities given by the model. This is therefore an alternative to the objective function used in HLM (see Section 2), that can find heuristic solutions to the location problem under the assumption of probabilistic travel times.

Finally, there is a conceptual issue which refers to the way uncovered calls are dealt with. The underlying assumption of Daskin's expected coverage model is that dispatches involving travel times that are greater than the critical covering time are dealt with in the same way as "covered" calls, adding to the workload of the system. Were dispatches of this type forbidden, one would expect an increased availability of servers to answer calls within the critical covering time, resulting in increased expected coverage.

We could reason in a similar manner with respect to calls that find the system saturated. If queuing is not allowed, these are lost calls in the sense that they do not use the resources of the system, leaving room for more calls to be successfully answered from an expected coverage point of view (see Section 2).

5. Conclusions

Of the two issues discussed in this note, namely the structural differences in the objective functions of the three models considered for server location and the effect of the *time factor* when calculating the expected coverage by HLM, the former is by far more important. Factors that may further affect the computation of expected coverage are discussed in Section 4.

From our analysis of the objective functions it can be seen that AMEXCLP, by relaxing the assumption that the servers operate independently, improves on the results produced by MEXCLP. On the other hand HLM, in addition to allowing the relaxation of the other two simplifying assumptions of MEXCLP, provides a way of dealing with queued calls which AMEXCLP does not. HLM is therefore more than just an improvement over the MEXCLP/AMEXCLP models: it has the capability of providing a more accurate description of the system.

When dealing with location problems for systems in which the contribution of queued calls to expected coverage is important, HLM emerges in fact as the only alternative.

Acknowledgements

The authors thank the anonymous referees for their helpful comments and suggestions.

References

[1] Batta R, Dolan JM, Krishnamurthy NP. The maximal expected covering location problem: revisited. Transportation Science 1989;23:277–87.
[2] Daskin MS. A maximal expected covering location model: formulation, properties and heuristic solution. Transportation Science 1983;17:48–70.
[3] Church RL, ReVelle CS. The maximal covering location problem. Papers of the Regional Science Association 1974;32:101–18.
[4] Larson RC. Approximating the performance of urban emergency service systems. Operations Research 1975;23:845–68.
[5] Larson RC. A hypercube queueing model for facility location and redistricting in urban emergency services. Computers and Operations Research 1974;1:67–95.
[6] Swain R. A decomposition algorithm for a class of facility location problems, unpublished Ph.D. thesis, Cornell University, Ithaca, NY, 1971.
[7] Larson RC, Odoni AR. Urban operations research. New Jersey: Prentice–Hall, Inc., 1981.
[8] Carter GM, Chaiken JM, Ignall E. Response areas for two emergency units. Operations Research 1972;20: 571–94.
[9] Jarvis JP. Optimization in stochastic service systems with distinguishable servers, Internal Report TR-19-75, MIT Operations Research Center, Cambridge, MA, 1975.
[10] Goldberg J, Dietrich R, Chen J, Mitwasi M, Valenzuela T, Criss E. Validating and applying a model for locating emergency medical vehicles in Tucson, AZ. European Journal of Operational Research 1990;49: 308–24.

Fernando Yassuo Chiyoshi is an Associate Professor in the Post-Graduate Department of Production Engineering at the Federal University of Rio de Janeiro. His research interests include queueing theory, simulation in meta heuristics, particularly simulated annealing. He has publications in these areas in the recent years.

Roberto Diéguez Galvão is a Full Professor in the Post-Graduate Department of Production Engineering at the Federal University of Rio de Janeiro. His research interests include optimization of discrete location and distribution models and the embedding of these models into Spatial Decision Support Systems. He has several publications in these areas in the last two decades.

Reinaldo Morabito is an Associate Professor in the Post-Graduate Department of Production Engineering at the Federal University of São Carlos, Brazil. His research interests are in the fields of operations management and operations research. He has published articles in cutting and packing problems, queuing networks applied to manufacturing systems, combinatorial optimization, logistics and transportation planning.