ON WORST-CASE AGGREGATION ANALYSIS FOR NETWORK LOCATION PROBLEMS*

Richard L. FRANCIS
Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA

Timothy J. LOWE
Department of Management Science, University of Iowa, Iowa City, IA 52242, USA

Abstract

Network location problems occur when new facilities must be located on a network, and the network distances between new and existing facilities are important. In urban, regional, or geographic contexts, there may be hundreds of thousands (or more) of existing facilities, in which case it is common to aggregate existing facilities, e.g. represent all the existing facility locations in a zip code area by a centroid. This aggregation makes the size of the problem more manageable for data collection and data processing purposes, as well as for purposes of analysis; at the same time, it introduces errors, and results in an approximating location problem being solved. There seems to be relatively little theory for doing aggregation, or evaluating the results of aggregation; most approaches are based on experimentation or computational studies. We propose a theory that has the potential to improve the means available for doing aggregation.

1. Introduction

To motivate the work to follow, it is helpful to first briefly relate some experiences one of us had in working on an actual location problem.

In 1989 the state of Florida, like several other large states in the USA in recent years, requested proposals from contractors to operate a motor vehicle inspection program in major metropolitan areas in the state. The program, which is mandated by Florida law, will involve stations where vehicles will be inspected to determine whether or not they satisfy exhaust pollution standards. For each contract zone where National Ambient Air Quality Standards are exceeded, contractors submitted bids which included the locations of the stations, the number of lanes of each of the stations, financial information, etc. Chosen contractors will charge the owner of each inspected vehicle a fee, a portion of which will be rebated to the state to

*This research was supported in part by the National Science Foundation, Grant No. DDM-9023392.

fund an agency to monitor operation of the stations. Contracts are intended to be given for seven years, with an option to extend for two additional one-year periods.

Contractors were required to submit, with their bids, estimates of average travel distances of vehicle registrants, assuming each registrant visits a closest station. More than four million registrants existed in the contract zones throughout the state. Subsequent to, and in part as a consequence of our involvement, the agency provided information to the contractors on the number of registrants in each contract zone by zip code area, as well as the latitude and longitude of 345 centroids of zip code areas in the contract zones. This information was used both by the contractors, and by us, for estimating average distances travelled, assuming each registrant used a closest station. The following sentence was included in the request for proposals to address the travel distance question: "The average driving distance from residences to inspection stations shall be no more than six (6) miles for at least ninety percent (90%) of the affected registered motor vehicle owners in the designated problem areas." Further, the agency listed five areas on which proposals would be graded, with a possible maximum grade of 200 points. The following area, worth 40 points out of 200, addressed the driving distance question. "The public convenience of the inspection stations, including a calculation of the maximum average driving distance of six (6) miles to an inspection facility applicable to ninety percent of motor vehicles within each program area for each contract zone and a calculation of the maximum waiting time."

We provided the agency with a simple, custom-made PC code with graphics which it used to help evaluate the contractor-proposed locations, as well as the results of some location-allocation analysis for one contract zone. Our program had the capability of using either Euclidean, rectilinear, or network travel distances, and estimated average and total travel distances of vehicle registrants, assuming each registrant visits a station closest to their zip code area centroid.

Since contractors had the right to appeal the agency awards, it was important to the agency that they be in a position to claim they had treated proposals impartially. One attraction of having a program, from the point of view of the agency, was that the program was impartial; it provided a common basis for evaluating all proposals, and also gave the agency a way to double-check the contractor's distance measure calculations.

The agency used our code as part of their process of evaluating the locations of all bids received, and this use was documented for purposes of an appeal (which occurred and was turned down).

This experience has identified several approximation questions of interest which, as best we can determine, are not well resolved. The principal question involves approximating each zip code area by its centroid, so that the total distance between registrants in the zip code area and the station was approximated by the product of the number of people in the zip code area and the distance between the station and the centroid; this kind of approximation, called aggregation, is *pervasive* in applied location work (see Rushton [32]). It would simply have been impossible to represent the location of every one of the more than four million individual registrants within the time (and budget constraints) available. A second approximation issue involved the choice of a distance measure: Euclidean, rectilinear, or network. (The choice the agency eventually made was Euclidean, because all of the contractors used Euclidean distance.) A related issue, unresolved, was how much effort should be made to estimate distances accurately when other sources of error in the model also affected its accuracy, including the use of current registration data to predict a future situation, inaccuracies in the given registration data, and the assumption that each registrant would use a closest station to their (registrant) address (instead, say, of one closest to their place of work, or one on the way between their home and place of work).

Approximating locations in a zip code area by a zip code area centroid is a type of aggregation. There are a number of reasons for doing aggregation, involving the cost of obtaining and processing the data, the form in which the data is available (it may already be aggregated), the size of the resulting location models, and the computational and theoretical tractability of the resulting location models. Note that this last reason is not necessarily the same as the previous one, since a large model might still be tractable.

2. Literature

There is a substantial literature on doing aggregation, a striking aspect of which, from our point of view, is that it is almost all experimental. Various schemes for doing aggregation are devised and then tested computationally, either by using randomly generated problems or actual data. Relatively recent papers with extensive literature discussions on aggregation include those by Current and Schilling [5,6], Daskin et al. [7], Mirchandani and Reilly [27], and Rushton [32]. These papers document the widespread use of aggregation, and the difficulty of doing it well. It is well recognized that aggregation introduces inaccuracy into the model, but what to do about it seems unclear. Indeed, Daskin et al. point out that "Conclusions from the literature are often contradictory and are based on subjective assessments of whether the aggregate locations differ significantly from the disaggregate sites."

For representative aggregation computational studies, see Brown and Masser [3], Hillsman and Rhoda [19], Goodchild [14], Bach [1], Mirchandani and Reilly [27], Current and Schilling [5,6], and Casillas [4].

Rushton [32], in an insightful discussion of applications of location models, discusses aggregation. He states that "The effect of employing discrete spatial structures to represent data that is distributed continuously is largely unknown, though recent research has shown that the effects on the validity of results from location-allocation models can be considerable. In most cases, the decision to use a particular data structure is a matter of convenience and most analysts do not discuss the potential consequences of the data system they use or the alternatives they reject . . . . Clearly, optimum locations must be sensitive, to an important degree, at some level of spatial aggregation of data. *We do not know how to identify this level in advance for any given application*" (emphasis added). Later, he points out that "As for capturing geographical complexity, the history of applications shows little development: decisions to use crude distance functions and spatially aggregated data units continue to be the rule and the lack of any recognized method for evaluating the consequences of these choices on the quality of the results has encouraged analysts to simply dismiss the problem as intractable."

According to Goodchild, "Many of the fields to which location-allocation models have been applied take the data base as some aggregation of a geographically dispersed demand." Later, he states that "The minimax problem is similar to the median, then, in that in both cases the effects of aggregation are unique to particular solutions, and therefore in that no general rules for aggregation can be found."

We conclude from our literature search that there is very little *theoretical* basis for doing aggregation, that aggregation is regularly done in applied location modelling, and that too much aggregation can destroy the accuracy of a location model.

It is the purpose of this paper to contribute to a theoretical basis for doing aggregation. We believe we have a good way of measuring the relative quality of aggregation, and can study the aggregation problem *analytically*. We will develop in the sequel an error bound for aggregation, and show that minimizing the error bound is a provably difficult and well-studied location problem. We shall see that this problem depends on the structure of the original problem, which substantiates Goodchild's conclusion that "no general rules for aggregation can be found." Indeed, we believe the best that can be done in terms of aggregation is to find an algorithm for doing aggregation which exploits problem structure. We believe our approach also explains some of the difficulties with prior work on aggregation that Rushton discusses.

3. Planar aggregation

At this point, it helps to introduce a (planar) location model related to our inspection station study, in order to begin to discuss evaluating the quality of aggregations. Suppose the vector X represents the location of a station, and the points P_i, $1 \leq i \leq m$, are the m registrant locations in a zip code area. Thus, with

$$f(X) = \sum \{d(X, P_i) : 1 \leq i \leq m\},$$

$f(X)$ is the total distance between the station location and the registrant locations, where $d(X, Y)$ is either the Euclidean or rectilinear distance between any two points X and Y. Let C denote the centroid of the zip code area:

$$C = \sum \{(1/m)P_i : 1 \leq i \leq m\}.$$

A well-known result [10, chapter 7] is that

$$md(X,C) \le f(X) \quad \text{for all} \quad X.$$

Thus, the approach in the Florida study underestimated the total distance, a fact which we pointed out to the agency, but were not too concerned about, since our code was being used for *relative* rather than absolute purposes. In particular, if our underestimate was off by an additive constant, then this would have had no effect on *relative* comparisons of travel distances.

Subsequent to our work with the agency, we were able to find a better approximation than $md(X, C)$ to $f(X)$. From the triangle inequality,

$$f(X) \le md(X,C) + f(C) \qquad \text{for all } X.$$

Thus,

$$0 \le f(X) - md(X,C) \le f(C),$$

from which it follows, with $f'(X) \equiv [md(X, C) + f(C)/2]$, that

$$-f(C)/2 \le f(X) - f'(X) \le f(C)/2 \quad \text{for all} \quad X,$$

or

$$|f(X) - f'(X)| \le f(C)/2 \qquad \text{for all } X.$$

Note that when $X = C$, the bound is tight. Hence, no smaller bound on $|f(X) - f'(X)|$ than $f(C)/2$ exists which holds for all X.

We can see, in the sense of the last three displayed lines of inequalities, that $f'(X)$ is a better approximation to $f(X)$ than $md(X, C)$ (and differs from our underestimate by an additive constant); in one case, the maximum "error" is $f(C)/2$, while in the other case it is $f(C)$. Of course, we could not have computed $f(C)/2$ for the same reason we could not have computed $f(X)$, but we do believe it would have been possible to estimate $f(C)$ with reasonable accuracy. We believe the above analysis to be new.

4. Aggregation for network location models: Basic results

The above discussion leads to our proposed research on approximation in network location models. For general discussions of network location models, see Handler and Mirchandani [35], Halpern and Maimon [17], Tansel et al. [33], Francis et al. [12], Dearing [8], Brandeau and Chiu [2], Mirchandani and Francis [28], and Francis et al. [13]. We remark that while we shall concentrate on network models, most of the results are also true for analogous planar models, including lemmas 1 through 4 and the immediately following results for error bound location functions.

We suppose we have a given connected network G with a finite number of vertices, undirected and rectifiable arcs of positive length, with $d(x, y)$ denoting the (nonnegative) length of any shortest path in G from x to y, having the customary distance properties that $d(x, y) = d(y, x)$ for any x and y (symmetry), $d(x, y) \geq 0$, with $d(x, y) = 0$ meaning $x = y$ (nonnegativity), and $d(x, y) \leq d(x, z) + d(z, y)$ for any x, y and z (the triangle inequality). For any x and $y \in G$, we denote by $P(x, y)$ the set of all points on shortest paths joining x and y. (For planar distances, interpret $P(x, u)$ as $\{z: d(x, z) + d(z, y) = d(x, y)\}$, e.g. the line segment joining x and y for Euclidean distances.)

Given a set of p new facility locations, $X = \{x_1, \ldots, x_p\}$ in the network, m existing facility locations at $a_1, \ldots, a_m$, respectively, together with corresponding "weights" (e.g. populations) $w_1, \ldots, w_m$, with $D(X, a_i)$ denoting the distance between a_i and a *closest* new facility, we now define p-median and p-center functions f and g, respectively, by

$$f(X) = \sum \{w_i D(X, a_i) : 1 \leq i \leq m\},$$

$$g(X) = \max \{w_i D(X, a_i) : 1 \leq i \leq m\}.$$

These functions were introduced by Hakimi [15,16], and the problem of minimizing each function is known to be NP-hard (Kariv and Hakimi [21,22], Hsu and Nemhauser [20]).

Let us consider the approximations to the functions f and g we obtain when we replace each a_i by some a_i'. Note that while the a_i will typically be distinct, the a_i' will not, since we will be approximating at least two a_i by the same point; that is, we aggregate many a_i into less a_i'. To illustrate this *idea* with the p-median function $f(X)$, suppose $m = 4$, and we take $a_1' = a_2' = \alpha_1$, $a_3' = a_4' = \alpha_2$ for some α_1, α_2. Replacing each a_i by a_i' then gives the approximating function $\sum\{w_i D(X, a_i') : 1 \leq i \leq 4\}$ $= \omega_1 D(X, \alpha_1) + \omega_2 D(X, \alpha_2)$, where $\omega_1 = w_1 + w_2$, $\omega_2 = w_3 + w_4$.

The following quite simple result, based on the triangle inequality and the distance symmetry property stated above, is basic to our approach:

LEMMA 1

If $d(a_i, a_i') \leq \varepsilon_i$, $1 \leq i \leq m$, then for every x and for $1 \leq i \leq m$,

$$-\varepsilon_i \leq d(x, a_i) - d(x, a_i') \leq \varepsilon_i .$$

(a) The leftmost inequality holds as an equality if and only if $a_i \in P(x, a_i')$ and $d(a_i, a_i') = \varepsilon_i$.

(b) The rightmost inequality holds as an equality if and only if $a_i' \in P(x, a_i)$ and $d(a_i, a_i') = \varepsilon_i$.

In other words, if a_i is approximated by a_i', which is no farther than ε_i from a_i, then $d(x, a_i)$ "changes" by at most $\pm\varepsilon_i$. Note *the conditions in (a) and (b) indicate situations to avoid* if possible in doing aggregation, since in these cases we have a maximum difference between $d(x, a_i)$ and $d(x, a_i')$.

An immediate consequence of lemma 1 is

LEMMA 2

If $d(a_i, a_i') \le \varepsilon_i$, $1 \le i \le m$, then for every X and for $1 \le i \le m$,

$$-\varepsilon_i \le D(X, a_i) - D(X, a_i') \le \varepsilon_i .$$

We remark it should be posible to develop necessary and sufficient conditions for each of the two inequalities in lemma 2 to hold as an equality. In particular, consider the special case where we suppose $x_{j(i)}$ is an element in X closest to *both* a_i and a_i': then the two inequalities become $-\varepsilon_i \le d(x_{j(i)}, a_i) - d(x_{j(i)}, a_i') \le \varepsilon_i$, and we can use (a) and (b) of lemma 1, with $x_{j(i)}$ replacing x, to obtain such conditions.

Now define

$$\delta_f = \sum \{w_i\varepsilon_i : 1 \le i \le m\}, \quad \delta_g = \max \{w_i\varepsilon_i : 1 \le i \le m\}.$$

We can use lemma 2 to obtain

LEMMA 3

Suppose $d(a_i, a_i') \le \varepsilon_i$, $1 \le i \le m$. Let $h(X)$ be either the p-center function $g(X)$ or the p-median function $f(X)$, and let $h'(X)$ be the approximation to $h(X)$ obtained by replacing each a_i by a_i'.

(a) For all X,

$$-\delta \le h(X) - h'(X) \le \delta \quad \text{or} \quad |h(X) - h'(X)| \le \delta,$$

where $\delta = \delta_f$ for the p-median function, $\delta = \delta_g$ for the p-center function.

(b) Further (from (a)), if h^* and h'^* denote the minimum values of h and h', respectively, then

$$-\delta \le h^* - h'^* \le \delta \quad \text{or} \quad |h^* - h'^*| \le \delta.$$

It is convenient to call the absolute value of the difference between the true function h and the approximating function h' the *error function* (error, for short) and to call δ the *error bound*. Note that δ is a *worst-case* error bound, in the sense that it is an error bound for *all* X, so that it is *a bound on the maximum error*; it may make an aggregation look worse than it actually is. An aggregration with a small error bound will certainly be a very good aggregation, but an aggregation with a large error bound may still be acceptable. However, for *comparing* aggregations, we believe it can be helpful. We remark that what we call the error is called, in the terminology of Hillsman and Rhoda [19], "Source A" error.

Lemma 3 enables us to measure how big our error can be, regardless of the choice of X. Further, the error bound δ may be as small as can be found on the maximum error, since one can easily construct examples where the bound is attained. That is, using (a) and (b) of lemma 1, one can construct a_i', and choose X, so that $h(X) = h'(X) \pm \delta$. This is the fundamental reason for characterizing our error bounds as worst-case. Our bounds are also reasonable in the sense that the error goes to zero as δ goes to zero. We can also see that the value of delta we obtain depends upon the structure of the problem; delta is much smaller for the p-center than for the p-median problem, for example (assuming the same data).

For aggregation problems, we can – conceptually at least – consider choosing the a_i', the approximations to the a_i, so as to try to minimize the error bound. Consider, for example, the p-median function, and let $U = \{u_1, \ldots, u_q\}$ denote a set of q points from which the a_i' will be chosen (e.g. midpoints of sides of city blocks). A common approach in aggregation is to choose a closest point in U to a_i as a_i', so that $d(a_i, a_i') = D(U, a_i)$, where $D(U, a_i)$ is the distance between a_i and a closest point in U (e.g. replace every existing facility on one side of a city block by the center of the side of the city block). Thus, in our error bound analysis, if we choose $\varepsilon_i = d(a_i, a_i')$, then (see lemma 3) the error bound $\delta = \delta_f$ is given by the following error bound function:

$$\beta_f(U) \equiv \sum \{w_i D(U, a_i) : 1 \le i \le m\}. \tag{1}$$

Hence, *finding U to minimize the error bound involves minimizing a q-median function.* Similarly, we conclude that *finding U to minimize the error bound for a p-center function involves minimizing a q-center error bound function,* namely,

$$\beta_g(U) \equiv \max \{w_i D(U, a_i) : 1 \le i \le m\}.$$

Although (1) is a worst-case measure of the error bound, we can use (1) to draw some conclusions about doing aggregation, all of which seem quite reasonable. We note that all the w_i and $D(U, a_i)$ for which $D(U, a_i) > 0$ affect the quality of the aggregation, and do so in an additive sense, so that aggregating a_i with both large w_i and large $D(U, a_i)$ will lead to more error than aggregating a_i with small w_i and small $D(U, a_i)$. Choosing as a_i' a point in U other than a closest point in U to a_i will cause the error bound to increase. Further, the larger U is the better, in the sense that, if U contains $\hat{U}$, then $\beta_f(U) \le \beta_f(\hat{U})$, while for the error bound to be zero, *every* a_i must be an element of U.

Note, by comparison, that the error bound for $\beta_g(U)$, the p-center problem, is much less sensitive to the way aggregation is done, since it is only the largest of the $w_i D(U, a_i)$ that affects the error bound $\beta_g(U)$; otherwise, one can draw conclusions similar to those of the previous paragraph. This latter observation reinforces and quantifies comments of Goodchild [14], and of Mirchandani and Reilly [27].

The previous analysis can be used in a different way as follows. Suppose a modeler is willing to tolerate an error in terms of objective function value of at most B, so that the objective function values for the aggregated and original problems, at optimality, can differ in absolute value by at most B. To formulate an aggregated problem with this property it is enough, for the p-median problem, to choose a vector U such that $D(U, a_i) \leq B/W$ for $1 \leq i \leq m$, where $W = \sum\{w_i: 1 \leq i \leq m\}$, and for the p-center problem, to choose a vector U such that $D(U, a_i) \leq B/w_i$ for $1 \leq i \leq m$. In either case, *finding such a vector U involves finding a feasible solution to a covering problem*, i.e. at least one element of U is in a neighborhood of a_i of radius B/W in the first case and B/w_i in the second case. From the standpoint of model size, it would be of interest to find a minimum cardinality set U intersecting each such neighborhood. For tree network location problems, we note that Tansel et al. [34] have solved this problem. For other literature on covering for network location problems, see Mirchandani and Francis [28]. We give a more detailed analysis of the effect of aggregation on covering constraints in section 5.

The expression (1) explains, to some extent, why doing aggregation well can be difficult, since we must solve a difficult (NP-hard) problem in order to obtain a minimal bound. We can thus view typical aggregation approaches as described above and as given in the literature as being heuristic approaches (with unproven properties) to solving p-median problems; we believe this point of view to be new. We remark here that both Mulvey and Crowder [29] and Current and Schilling [5] have pointed out that one can do aggregation by solving a p-median problem, but gave no error bound analysis.

We observe also that we can now apply theory – and algorithms – developed for solving p-median problems (and p-center problems) to the problem of doing "good" aggregation; this may result in better rationales for doing aggregation. For example, if we want to find an aggregation to minimize the p-median error bound, there is an optimal U, U^*, where every element of U^* is a vertex; if we want to minimize the p-center error bound, there is an optimal U, U^*, where each element of U^* is a vertex or a "bottleneck point" or "intersection point", namely, a point y so that $w_i d(y, a_i) = w_j d(y, a_j)$ for some i and j, with $y \in P(a_i, a_j)$. We believe these observations are new.

With $|U| = q$, we note that q is independent of p; thus it is difficult to do aggregation even for a 1-median problem. On the other hand, if we want to minimize f repeatedly treating p as a parameter, we would only need to do a single aggregation.

Note that the error bound functions still involve all the a_i, so if our reason for doing aggregation is because of the difficulties in dealing with all of the a_i, then these difficulties will still occur in working with the error bound function; one can view this as an *aggregation paradox*. This suggests in some cases that we may need to be satisfied with some upper bound on the error bound instead of the error bound function itself, or instead of its minimum value. (For example, with W the sum of the w_i, and $\varepsilon_{\max}$ at least as large as every ε_i, $\delta_f \leq W\varepsilon_{\max}$.) Alternatively, we may need to *estimate* the error bound function rather than compute it exactly. It should also be clear that the reasons for having to do aggregation in the first place will have a major impact on what can be done in terms of error-bound analysis.

In order to illustrate the fact that our error-bounding approach can be improved if we can exploit geometric structure, let us make a comparison with the approximation problem discussed earlier for evaluating inspection station locations. Consider the 1-median function on a network, $f(x) = \Sigma\{d(x, a_i) : 1 \leq i \leq m\}$. For some point c, suppose we take $a_i' = c$ for $1 \leq i \leq m$. We obtain the approximating function $f'(x) = \Sigma\{d(x, a_i') : 1 \leq i \leq m\} = md(x, c)$. Lemma 3 then gives $-f(c) \leq f(x) - f'(x) \leq f(c)$ for all x, and the error bound $f(c)$ cannot be reduced provided c is the 1-median. By comparison, for the analogous planar problem we discussed earlier, we obtained a smaller error bound $(f(c)/2)$ when c is the centroid. We conclude that we have a better approximating 1-median function for the planar case than for the network case. Part of the difficulty here is that, for a general network, a centroid is not a well-defined concept. Indeed, it is not difficult to show that there may be *no* point c such that $md(x, c) \leq f(x)$ for all x when we have a network. Consider a network consisting of a single cycle with $m = 2$ nodes a_1, a_2, and distinct arcs (a_1, a_2) and (a_2, a_1) each of length 1. In this case, $f(x) = 1$ for all x. *Suppose* there is a point c in the network so that $md(x, c) \leq f(x)$ for all x, so that $2d(x, c) \leq 1$ for all x. However, if we choose x to be the "opposite" point from c in the network, so that $d(x, c) = 1$, we then conclude $2 \leq 1$.

The foregoing deals mostly with comparison of objective function values, but is also of some potential help in comparing optimal solutions. We can readily use lemma 3(a) to obtain lemma 4 as follows:

LEMMA 4

Suppose $d(a_i, a_i') \leq \varepsilon_i$, $1 \leq i \leq m$. Let $h(X)$ be either the p-center function $g(X)$ or the p-median function $f(X)$, and let $h'(X)$ be the approximation to $h(X)$ obtained by replacing each a_i by a_i'. Let $S(k) = \{X: h(X) \leq k\}$, $S'(k) = \{X: h'(X) \leq k\}$, that is, $S(k)$ and $S'(k)$ are level sets of h and h', respectively, of any arbitrary value k. Then, with δ defined in lemma 3,

$$S'(k - \delta) \subset S(k) \subset S'(k + \delta),$$

$$S(k - \delta) \subset S'(k) \subset S(k + \delta).$$

Further,

$$S'(h'^*) \subset S(h^* + 2\delta), \quad S(h^*) \subset S'(h'^* + 2\delta). \tag{2}$$

Note that $S(h^*)$ and $S'(h'^*)$ are the sets of all optimal solutions to our true problem and our approximating problem, respectively. The two containment propertie

(2) are the potentially important ones; supposing we can construct the S' sets, then we may be able to get "close" to $S(h^*)$, particularly if delta is small, or if the containment is "almost" an equality. In any case, these containment properties may help us to restrict our search for an optimal solution to the true problem, particularly if we can combine them with additional properties characterizing optimal solutions.

5. Covering constraints for network location problems

There is an application of some of the foregoing to covering constraints, which have the form

$$D(X, a_i) \le r_i, \quad 1 \le i \le m,$$

where again X is a set of the locations of p centers, and $D(X, a_i)$ is the distance between a_i and a *closest* center. Such constraints occur with covering problems, where we want to satisfy the constraints while minimizing the total number of new facilities that we locate.

We can again replace each a_i by a_i', with $d(a_i, a_i') \le \varepsilon$. Lemma 2 immediately implies the following: for any constant k with

$$L(k) \equiv \{X : D(X, a_i) \le r_i + k, \ 1 \le i \le m\},$$

$$L'(k) \equiv \{X : D(X, a_i') \le r_i + k, \ 1 \le i \le m\},$$

we have

$$L'(-\varepsilon) \subset L(0) \subset L'(\varepsilon),$$

$$L(-\varepsilon) \subset L'(0) \subset L(\varepsilon).$$

Thus, we can relate the set of all feasible solutions to the covering constraints $L(0)$, to sets of all feasible solutions to approximating covering constraints, and observe that the approximation improves as ε goes to zero.

In principle, at least, it is possible to consider the effect of aggregation on both objective function and constraints. For example, instead of studying the constrained p-median problem

$$\text{minimize} \quad f(X)$$

$$\text{subject to} \quad D(X, a_i) \le r_i, \quad 1 \le i \le m,$$

one could focus on the approximating problem

$$\text{minimize} \quad f'(X)$$

$$\text{subject to} \quad D(X, a_i') \le r_i, \quad 1 \le i \le m.$$

Many of the foregoing results can be modified to apply to such constrained problems. For example, consider the 2-parameter level set $\{X: f(X) \le k_1, D(X, a_i) \le r_i + k_2, i = 1, \ldots, m\}$, which we note is $S(k_1) \cap L(k_2)$. Likewise, the set we obtain by replacing f and each a_i by f' and a_i', respectively, in this 2-parameter set is $S'(k_1) \cap L'(k_2)$. Suppose $d(a_i, a_i') \le \varepsilon$, $i = 1, \ldots, m$, with δ as defined in lemma 3. Then, using our previous containment results, we have

$$S'(k_1 - \delta) \cap L'(-\varepsilon) \subset S(k_1) \cap L(0) \subset S'(k_1 + \delta) \cap L'(\varepsilon),$$

$$S(k_1 - \delta) \cap L(-\varepsilon) \subset S'(k_1) \cap L'(0) \subset S(k_1 + \delta) \cap L(\varepsilon).$$

Hence, each 2-parameter level set contains another 2-parameter level set and is in turn contained in yet another 2-parameter level set.

Another approach to approximating covering constraints involves a related minmax problem. Define the "modified" p-center function $H(X)$ by $H(X) \equiv \max\{D(X, a_i) - r_i: 1 \le i \le m\}$, and let H^* be the minimum value of $H(X)$, say $H^* = H(X^*)$. Note that there exists a feasible solution to the covering constraint if and only if $H^* \le 0$. Also, with $L(k) = \{X: D(X, a_i) \le r_i + k, 1 \le i \le m\}$, note that $L(k) = \{X: H(X) \le k\}$ and $L(0) = \{X: H(X) \le 0\}$.

For any two p-centers X, Y, let us define the "bottleneck matching" distance $\Delta(X, Y)$. With the indexing of, say Y fixed as $Y = \{y_1, y_2, \ldots, y_p\}$, find a matching $x_{[j]}$ to y_j, $j = 1, \ldots, p$, where the matching solves the problem

$$\min \max\{d(x_{[j]}, y_j) : 1 \le j \le p\},$$

with the minimization being over all matchings. We define $\Delta(X, Y)$ as the minimum objective function value for this problem. This problem is known as a minmax matching problem, and can be solved in $O(p^3)$ (Lawler [24]). We remark that whenever the function $d(x, y)$ is a distance (satisfies the nonnegativity, symmetry and triangle inequality properties), then it can be shown that the function $\Delta(X, Y)$ is also a distance.

The following is a kind of triangle inequality involving covering constraints and the foregoing "bottleneck matching" distance:

Remark

For any a_i, X, Y,

$$D(X, a_i) \le \Delta(X, Y) + D(Y, a_i).$$

This remark, together with basic definitions, results in the following:

CLAIM

If $\Delta(X, X^*) \le (k - H^*)$, then $X \in L(k)$.

Consider the case $k = 0$, and suppose $H^* \leq 0$, so that X^* satisfies the covering constraints. Intuitively, what the claim says is that if X is "close" to X^* $(\Delta(X, X^*) \leq -H^*)$, then X also satisfies the covering constraints. Note $\Delta(X, X^*) \leq k - H^*$ means that $d(x_{[j]}, x_j^*) \leq k - H^*$, $1 \leq j \leq p$.

Considering X^* fixed, one can show that $\Delta(X, X^*) \leq -H^*$ is not a necessary condition for X to satisfy the covering constraints. Consider the case with $p = 1$ and G a single cycle, consisting of arcs $[a_1, a_2]$, $[a_2, a_3]$ and $[a_3, a_1]$, each of length 4. Let only a_1 and a_2 be existing facilities, with $r_1 = r_2 = 4$. Let x^* be the midpoint of $[a_1, a_2]$; x^* minimizes $H(x)$, and $H^* = -2$. Note that a_3 satisfies the covering constraints $(a_3 \in L(0))$ but $d(a_3, x^*) = 6 > -H^*$. Thus, with cyclic networks one cannot expect $\Delta(X, X^*) \leq -H^*$ to be equivalent to the covering constraints. However, there are cases when equivalence can be proven.

Let $N(Y, \rho)$ denote the set of all X for which $\Delta(X, Y) \leq \rho$, a neighborhood with center Y and radius ρ. With this notation, the claim establishes that $N(X^*, k - H^*) \subset L(k)$.

For the case where G is a tree and $p = 1$, Francis et al. [11] proved that equality holds: $N(X^*, k - H^*) = L(k)$. Thus, for this case, the m covering constraints $d(x, a_i) \leq r_i + k$ are equivalent to the *single* constraint $d(x, x^*) \leq k - H^*$. One can see the point of this equivalence for approximation or aggregation: a single distance constraint can approximate (or aggregate) m constraints exactly. Thus, there may exist cases where the constraint $D(X, X^*) \leq k - H^*$ is a "reasonable" approximation to $D(X, a_i) \leq r_i + k$ for $1 \leq i \leq m$. The characterization of problem structures (beyond trees with $p = 1$) for which the approximation is reasonable is an open question.

We observe that $N(X^*, k - H^*) \subset L(k)$ implies $N(X^*, -H^*) \subset L(0)$, $N(X^*, -\varepsilon - H^*) \subset L(-\varepsilon)$ and $N(X^*, \varepsilon - H^*) \subset L(\varepsilon)$; the latter two neighborhoods can be constructed from the former with less effort than that required to obtain $N(X^*, -H^*)$. Analogous to $L(-\varepsilon) \subset L(0) \subset L(\varepsilon)$, we would then have $N(X^*, -\varepsilon - H^*) \subset N(X^*, -H^*) \subset N(X^*, \varepsilon - H^*)$, and could consider the neighborhoods as approximations to the three level sets. Similarly, recalling the definition of $L'(k)$ and its potential use with covering constraints, we could consider finding neighborhoods with the same center and successive radii different by ε that would be contained, respectively, in $L'(-\varepsilon)$, $L'(0)$, and $L'(\varepsilon)$; these neighborhoods would approximate the latter sets.

6. Aggregation for network location problems: A research agenda

Since the error bound problems we obtain are location problems, for which theory has been developed, it will be important to explore the impact of this theory on ways to do aggregation; for example, as mentioned above, the vertex optimality principle is relevant for p-median aggregation. Likewise, it is well known that if the allocation of existing facility locations (the a_i) to new facility locations is fixed, the p-center and p-median problems reduce, respectively, to collections of independent 1-center and 1-median problems. We believe this fact will be helpful for minimizing the error bound functions in some cases. Imagine, for example, a network showing major cities in the USA with interconnecting roads. Suppose we wish to solve a p-median problem on this network, and have decided to aggregate the various locations in each city into a single point; in this case, the error bound function becomes a sum of independent 1-median functions, one function for each city, and the resulting 1-median is the best point to represent that city by.

It seems interesting to note for the extreme case $|X| = |U|$, i.e. $p = q$, that *if* we could solve the error bounding problem to optimality, then we would also have an optimal solution, say X^*, to the original problem. Also, since the approximating problem would only have p (aggregated) existing facilities, we would have $h'(X^*) = 0$, so that X^* would be optimal to the approximating problem as well. (In addition, the error bound would be tight, i.e. $\beta_f(X^*) = |h(X^*) - h'(X^*)|$.) Thus, this extreme case illustrates a situation where solving the approximating problem could give a very poor objective function estimate but entirely satisfactory locations. While we believe that usually we would have p much less than q, it would be interesting to examine how early this situation begins to occur as the ratio of p to q approaches 1 from below. We might wonder, conversely, for p much less than q whether h'^* is a very good approximation to h^*, but the locations it gives could be "far" from the optimal ones to the true problem.

Particularly for the p-median problem, we have seen that the error bound function can be large. However, this may still be acceptable provided the *ratio* of it to the original p-median function is small (and likewise for the p-center case). Recognizing that the two functions have the same form, it appears to us that some theoretical analysis of this ratio may be possible. For example, if $|X| = |U|$, the ratio (for an optimal solution to each problem) will be one, which is surely unacceptable. When $|U|$ is much greater than $|X|$, the ratio will be closer to zero, which is better, since a ratio of zero means there is no aggregation at all (U includes every a_i). Being able to estimate this ratio would thus be helpful in evaluating the acceptability of an aggregation. A rough surrogate measure of this ratio might be $(1/q)/(1/p) = p/q$, since the minimum objective function value of f (of β_f) is directly proportional to $1/p$ (to $1/q$).

In the construction of a mathematical model, one usually faces some sort of a trade-off between realism and tractability, and this situation is particularly acute with the p-center and p-median models. Rushton [32] points out that "the optimal degree of aggregation is not known . . . ." A *highly* speculative model may possibly give some insight into the trade-off problem we encounter in aggregation: doing little aggregation (large q) results in a more accurate model, but one more difficult to deal with, while doing much aggregation (small q) results in a less accurate but more tractable model. Given $p \leq q \leq m$, *suppose* we measure the inaccuracy of an aggregation by $[p/q - p/m]$, and know a constant, say c_1, so that the inaccuracy "cost" is $c_1[p/q - p/m]$. *Suppose* also that the aggregation intractability "cost" is linearly proportional to q/m, with the constant of proportionality being c_2, so that the aggregation intractability cost is $c_2 q/m$. We note the inaccuracy cost is zero wher

$q = m$ (no aggregation), while c_2 is the intractability cost we incur when $q = m$. Then the total cost is given by $TC(q) \equiv c_1[p/q - p/m] + c_2q/m$, where $p \leq q \leq m$. This resulting model has essentially the same mathematical form as the famous EOQ model of inventory theory. For most problems, p is much less than m, so that the term c_1p/m is almost zero. Therefore, in what follows we shall ignore this term and concentrate on the model $C(q) \equiv c_1p/q + c_2q/m$ we obtain by deleting c_1p/m from $TC(q)$.

To find a "best" value q^* of q (one minimizing $C(q)$), take the derivative of $C(q)$ and set it to zero to obtain $\hat{q} = [(c_1 p\mathrm{m})/c_2]^{1/2}$, and then use as q^* either $\hat{q}$ rounded up or rounded down to the nearest integer, whichever has a lower cost. We assume $p \leq \hat{q} \leq m$; otherwise we would have to use as q^* either p or m depending on whether $\hat{q} < p$ or $\hat{q} > m$, respectively. We note that $p \leq \hat{q} \leq m$ is equivalent to $p/m \leq c_1/c_2 \leq m/p$, which is an indication of the importance of the ratios p/m and m/p. We see that choosing q is entirely dependent on the ratio c_1/c_2, which we can consider as the "trade-off" between the two considerations, inaccuracy and intractability.

Assuming $q^* = \hat{q}$, we have $C(q^*) = 2[(c_1c_2p)/m]^{1/2}$. Because of the square root, $C(q^*)$ is *relatively insensitive to the choice of* c_1 *and* c_2, which means that inaccurate estimates of c_1 and c_2 may not have a large effect. For example, if instead of the true value of c_1 we use, by mistake, either $c_1/2$ or $2c_1$ in the equations for $q^* = \hat{q}$ and $C(q^*)$, the result will be to increase $C(q^*)$ above its true value by only 6%. Lowe and Schwarz [25] analyze the sensitivity of the EOQ model to parameter estimates. We expect their results to be useful in the study of $C(q)$.

Of course, the aggregation cost would probably not be linear in q, but nonlinear, strictly increasing with a strictly positive first derivative (and thus strictly convex), in which case the result might be to give a smaller value of q^* than the one we compute. In any case, we can see that having some way to quantify the trade-off between the two issues of importance here would be extremely helpful.

Finding a way to exploit planar structure, as discussed briefly above, will be helpful. Perhaps for planar p-median and p-center problems we can do better than our worst-case approach. For example, our worst-case approach to minimize the error bound for the planar 1-median problem would give the planar 1-median – *not the centroid*, and without an additive constant such as $f(C)/2$ – as the best single aggregating point. Clearly, there are opportunities for research along these lines. Likewise, if we can capture some planar structure in our networks, it should be helpful.

While our error bound analysis has concentrated on the p-center and p-median problems, we remark it can also be used with other location problems, including the multifacility minisum and minimax location problems (Francis and White [10]), again resulting in certain q-median and q-center error bound functions, respectively. Thus, finding additional problems to which our analysis applies will also be of interest.

Computational testing of our error bound approach is certainly in order. Our approach measures the worst-case error, but average-case error would be more meaningful. We can obtain some insight into how closely these two errors are related through computational experimentation. Similarly, we can explore how good the error bound approach is for comparing two aggregations. We believe it will be more satisfactory for making relative than absolute decisions about aggregations. This computational testing could possibly lead to some theoretical average-case analysis, although it is clear that doing such an analysis will be much more difficult than doing worst-case analysis. In some cases, there can be compensating errors; as X becomes farther from one a_i, it becomes closer to anothe a_i, and our results to date do not capture this, whereas an average-case analysis might. We draw an analogy here with the analysis of heuristic algorithms for solving traveling salesmen problems (Fisher [9], Papadimitriou and Steiglitz [30]). Most of the analysis is of a worst-case nature, and heuristics that may have a frightening worst-case error bound often perform quite well in practice.

Conceptually at least, we can view our approach to aggregation as approximating one function, say h, by a second, say h'. This is a common practice in numerical analysis (although there, the functions are usually of a single real variable), so that some of the fundamental concepts in numerical analysis may be helpful. Hamming [18], for example, identifies three important questions for numerical function approximation that seem relevant in our context. (1) What class of functions are we approximating over? (2) What criterion do we use as a goodness of fit? (3) What accuracy is needed? Concerning the first two questions, for our error bound approach we are approximating over those functions we can obtain from the original ones by replacing each a_i by some a_i', and we are using the maximum error as a measure of goodness of fit, but there could certainly be other approaches. Concerning the third question, the answer is not as clear, as discussed above; what is involved is, on the one hand, a trade-off between data availability and data collection costs, analytical and computational tractability and, on the other hand, the benefits of having an accurate model. However, as mentioned for the Florida study, we can accept less accuracy for relative purposes than for absolute purposes.

The error bound analysis we have discussed is essentially *a priori* analysis in that it is done prior to solving the aggregation problem or the location problem *If* we knew optimal solutions to these two problems, we could do *a posterior* analysis, which should give us additional information. We can draw analogies witl aggregation done in various branches of mathematical programming where *a posterioi* analysis typically gives better bounds than *a priori* analysis. Rogers et al. [31 provide an excellent, up-to-date survey of aggregation in optimization. Alternativel if we knew enough about the *properties* of optimal solutions, we could make us of these properties to do what would amount to *a posteriori* analysis. Since o fundamental location problems of interest are NP-hard, it may only be possible t do approximate analysis here. For example, we could look at the case of the p center problem where the graph is a tree [28], which is well-solved, or at the median problem for a tree [26]. Alternatively, we could look at linear programmir relaxations of integer programming formulations of the location problem of interes for which theory is well-developed.

While the approximation problem we have concentrated on here is that of aggregation, there are other important approximation issues for network location theory. For example, it is relevant to ask in how much detail a road network should be represented in a network model. So far as we know, there is no theoretical analysis of this question, although it is clear that the accuracy of representation has a large effect on distance measure, and on locations provided by network location models. Perhaps our aggregation approach can shed some light on this question, if we can find a way to interpret it in terms of how many vertices are needed to give an accurate representation of the problem.

Acknowledgement

We are happy to thank an anonymous referee for suggesting an improvement in the definition of the function $\Delta(X, Y)$.

References

[1] L. Bach, The problem of aggregation and distance for analyses of accessibility and access opportunity in location-allocation models, Environ. Planning A13(1981)955–978.

[2] M.L. Brandeau and S.S. Chiu, An overview of representative problems in location research, Manag. Sci. 35(1989)645–673.

[3] P.J.B. Brown and I. Masser, An empirical investigation of the use of Broadbent's rule in spatial system design, in: *Spatial Representation and Spatial Interaction*, ed. I. Masser and P.J.B. Brown (Martinus Nijhoff, Boston, MA, 1978).

[4] P.A. Casillas, Data aggregation and the p-median problem in continuous space, in: *Spatial Analysis and Location-Allocation Models*, ed. A. Ghosh and G. Rushton (Van Nostrand Reinhold, New York, 1987), pp. 327–344.

[5] J.R. Current and D.A. Schilling, Elimination of source A and B errors in p-median location problems, Geograph. Anal. 19(1987)95–110.

[6] J.R. Current and D.A. Schilling, Analysis of errors due to demand data aggregation in the set covering and maximal covering location problems, Geograph. Anal. 22(1990)116–126.

[7] M.S. Daskin, A.E. Haghani, M. Khanal and C. Malandraki, Aggregation effects in maximum covering models, Ann. Oper. Res. 18(1989)115–140.

[8] P.M. Dearing, Location problems, Oper. Res. Lett. 4(1985)95–98.

[9] M.L. Fisher, Worst-case analysis of heuristic algorithms, Manag. Sci. 26(1980)1–17.

[10] R.L. Francis and J.A. White, *Facility Layout and Location: An Analytical Approach* (Prentice–Hall, Englewood Cliffs, NJ, 1974).

[11] R.L. Francis, T.J. Lowe and H.D. Ratcliff, Distance constraints for tree network multifacility location problems, Oper. Res. 26(1978)570–596.

[12] R.L. Francis, L.F. McGinnis and J.A. White, Locational analysis, Eur. J. Oper. Res. 12(1983) 220–252.

[13] R.L. Francis, L.F. McGinnis and J.A. White, *Facility Layout and Location: An Analytical Approach*, 2nd ed. (Prentice–Hall, Englewood Cliffs, NJ, 1992).

[14] M.F. Goodchild, The aggregation problem in location-allocation, Geograph. Anal. 11(1979) 240–254.

[15] S.L. Hakimi, Optimal locations of switching centers and the absolute centers and medians of a graph, Oper. Res. 12(1964)450–459.

[16] S.L. Hakimi, Optimal distribution of switching centers in a communication network and some related graph theoretic problems, Oper. Res. 13(1965)462–475.

[17] J. Halpern and O. Maimon, Algorithms for the m-center problem: a survey, Eur. J. Oper. Res. 10(1982)90–99.

[18] R.W. Hamming, *Numerical Methods for Scientists and Engineers* (McGraw Hill, New York, 1962).

[19] E.L. Hillsman and R. Rhoda, Errors in measuring distances from populations to service centres, Ann. Regional Sci. 12(1978)74–88.

[20] W.L. Hsu and G.L. Nemhauser, Easy and hard bottleneck location problems, Discr. Appl. Math. 1(1979)209–215.

[21] O. Kariv and S.L. Hakimi, An algorithmic approach to network location problems, I: The p-centers, SIAM J. Appl. Math. 37(1979)513–537.

[22] O. Kariv and S.L. Hakimi, An algorithmic approach to network location problems, II: The p-medians, SIAM J. Appl. Math. 37(1979)539–560.

[23] J. Krarup and P. Pruzan, Selected families of location problems, Ann. Discr. Math. 5(1979)327–387.

[24] E.L. Lawler, *Combinatorial Optimization: Networks and Matroids* (Holt, Rinehart and Winston, New York, 1976).

[25] T.J. Lowe and L.B. Schwarz, Parameter estimation for the EOQ lot-size model: Minimax and expected value choices, Naval Res. Logist. Quart. 30(1983)367–376.

[26] P.B. Mirchandani and A. Oudjit, Localizing 2-medians on probabilistic and determinisitic tree networks, Networks 10(1980)329–350.

[27] P.B. Mirchandani and J.M. Reilly, Spatial nodes in discrete location problems, Ann. Oper. Res. 6(1986)203–222.

[28] P.B. Mirchandani and R.L. Francis (eds.), *Discrete Location Theory* (Wiley, New York, 1990).

[29] J.M. Mulvey and H.P. Crowder, Cluster analysis: an application of Lagrangian relaxation, Manag. Sci. 25(1979)329–340.

[30] C.H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity* (Prentice–Hall, Englewood Cliffs, NJ, 1982).

[31] D.F. Rogers, R.D. Plante, R.T. Wong and J.R. Evans, Aggregation and disaggregation techniques and methodology in optimization, Oper. Res. 39(1991)553–582.

[32] G. Rushton, Applications of location models, Ann. Oper. Res. 18(1989)25–42.

[33] B.C. Tansel, R.L. Francis and T.J. Lowe, Location on networks, a survey, Parts I and 2, Manag. Sci. 29(1983)482–511.

[34] B.C. Tansel, R.L. Francis, T.J. Lowe and M.-L. Chen. Duality and distance constraints for the nonlinear p-center and covering problem on a tree network, Oper. Res. 30(1982)725–744.

[35] G. Handler and P.B. Mirchandani, *Location on Networks: Theory and Algorithms* (The MIT Press, Massachusetts, 1979).