Ann Oper Res (2007) 150:205–230
DOI 10.1007/s10479-006-0154-0

A fast swap-based local search procedure for location problems

Mauricio G. C. Resende · Renato F. Werneck

Published online: 9 January 2007

Abstract We present a new implementation of a widely used swap-based local search procedure for the p-median problem, proposed in 1968 by Teitz and Bart. Our method produces the same output as the best alternatives described in the literature and, even though its worst-case complexity is similar, it can be significantly faster in practice: speedups of up to three orders of magnitude were observed. We also show that our method can be easily adapted to handle the facility location problem and to implement related procedures, such as path-relinking and tabu search.

Keywords Local search · p-Median · Facility location · Experimental analysis · Reordering problem

1 Introduction

The p-median problem is defined as follows. Given a set F of m facilities, a set U of n users (or customers), a distance function $d : U \times F \to \mathcal{R}_+$, and an integer $p \leq m$, determine which p facilities to open so as to minimize the sum of the distances from each user to the closest open facility. In other words, given p, we want to minimize the cost of serving all customers.

Since this problem is NP-hard (Kariv and Hakimi, 1979), a polynomial-time algorithm to solve it exactly is unlikely to exist. The most effective algorithms proposed in the literature (Avella, Sassano, and Vasil'ev, 2003; Beasley, 1985; Briant and Naddef, 2004; Cornuéjols, Fisher, and Nemhauser, 1977; Galvão, 1980; Rosing, ReVelle, and Rosing-Vogelaar, 1979; Senne, Lorena, and Pereira, 2005) use branch-and-bound, with lower bounds

R. F. Werneck: The results presented in this paper were obtained while this author was a summer intern at AT&T Labs Research.

M. G. C. Resende (✉)
AT&T Labs Research, 180 Park Avenue, Florham Park, NJ 07932
e-mail: mgcr@research.att.com

R. F. Werneck
Department of Computer Science, Princeton University, 35 Olden Street, Princeton, NJ 08544
e-mail: rwerneck@cs.princeton.edu

obtained from some linear programming relaxation of the problem. In the worst case, all these methods are exponential, but they can be quite fast in practice (the recent algorithm by Avella, Sassano, and Vasil'ev (2003) is particularly effective). Also, similar techniques can be made to work as heuristics only, producing close-to-optimal solutions in reasonable time (Avella, Sassano, and Vasil'ev, 2003; du Merle et al., 1999; Senne and Lorena, 2000, 2002).

There are also simpler heuristics that use no duality (linear programming) information at all. The most natural options are constructive heuristics, methods that build solutions from scratch, usually in a greedy fashion (Cornuéjols, Fisher, and Nemhauser, 1977; Kuehn and Hamburger, 1963; Whitaker, 1983). A step further is to use a *local search procedure*, which takes an existing solution as input and tries to improve it (Goodchild and Noronha, 1983; Hodgson, 1978; Maranzana, 1964; Rosing, 1997; Taillard, 2003; Teitz and Bart, 1968). It does so in an iterative fashion, examining *neighboring solutions*, those that differ from the original one by a small (problem- and algorithm-specific) modification. Finally, there are metaheuristics, procedures that aim at exploring a large portion of the search space in an organized fashion to obtain close-to-optimal solutions (possibly using constructive algorithms and local search as subroutines). Recent examples in the literature include variable neighborhood search (Hansen and Mladenović, 1997), variable neighborhood decomposition search (Hansen, Mladenović, and Perez-Brito, 2001), tabu search (Rolland, Schilling, and Current, 1996; Voß, 1996), heuristic concentration (Rosing and ReVelle, 1997), scatter search (García-López et al., 2003), and a GRASP-based hybrid algorithm (Resende and Werneck, 2004).

This study concerns the local search proposed by Teitz and Bart (1968), based on swapping facilities. In each iteration, the algorithm looks for a pair of facilities (one to be inserted into the current solution, another to be removed) that would lead to an improved solution if swapped. If such a pair exists, the swap is made and the procedure is repeated.

Arya et al. have shown (Arya et al., 2001) that, in a metric setting, this algorithm always finds solutions that are within a factor of at most 5 from the optimum. However, for practical, non-pathological instances the gap is usually much smaller, just a few percentage points (Rosing, 1997; Whitaker, 1983). This has made the algorithm very popular among practioners, often appearing as a key subroutine of more elaborate metaheuristics (García-López et al., 2003; Hansen and Mladenović, 1997; Resende and Werneck, 2004; Rolland, Schilling, and Current, 1996; Rosing and ReVelle, 1997; Voß, 1996).

Our concern in this paper is not solution quality—the reader is referred to Rosing (1997) and Whitaker (1983) for insights on that matter. Our goal is to obtain the same solutions Teitz and Bart would, only in less time. We present an implementation that is significantly (often asymptotically) faster in practice than previously known alternatives.

The paper is organized as follows. In Section 2, we give a precise description of the local search procedure and a trivial implementation. In Section 3, we describe the best alternative implementation described in the literature, proposed by Whitaker (1983). Our own implementation is described in Section 4. We show how it can be adapted to handle the facility location problem and to handle related operations (such as path-relinking and tabu search) in Section 5. Experimental evidence to the efficiency of our method is presented in Section 6. Final remarks are made in Section 7.

Notation and assumptions. Before proceeding to the study of the algorithms themselves, let us establish some notation. As already mentioned, F is the set of potential facilities and U the set of users that must be served. The basic parameters of the problem are $n = |U|$, $m = |F|$, and p, the number of facilities to open. Although $1 \leq p \leq m$ by definition, we will ignore

trivial cases and assume that $1 < p < m$ and that $p < n$ (if $p \geq n$, we just open the facility that is closest to each user). We assume nothing about the relationship between n and m.

We use u to denote a generic user, and f a generic facility. The cost of serving u with f is $d(u, f)$, the *distance* between them, which is always nonnegative. (We do not make any other assumption about the distance function; in particular, we do not assume that the triangle inequality is valid.) A *solution* S is any subset of F with p elements, and represents the set of open facilities. Every user u is assigned to the closest facility $f \in S$ (the one that minimizes $d(u, f)$). This facility will be denoted by $\phi_1(u)$. Our algorithm often needs to access the second closest facility to u in S as well; it will be denoted by $\phi_2(u)$. To simplify notation, we will abbreviate $d(u, \phi_1(u))$ as $d_1(u)$, and $d(u, \phi_2(u))$ as $d_2(u)$.[1] We often deal specifically with a facility that is a candidate for insertion; it will be referred to as f_i (by definition $f_i \notin S$); similarly, a candidate for removal will be denoted by f_r ($f_r \in S$, also by definition).

Throughout this paper, we assume the *distance oracle* model, in which the distance between any customer and any facility can be determined in $O(1)$ time. In this model, all values of ϕ_1 and ϕ_2 for a given solution S can be straighforwardly computed in $O(pn)$ total time: for each of the n customers, we explicitly find the distances to the p open facilities and pick the smallest. Problems defined by a distance matrix clearly fall into the distance oracle model, but an explicit matrix is not always necessary. If users and facilities are points on the plane, for example, distances can also be computed in constant time. There are cases, however, in which that does not happen, such as when the input is given as a sparse graph, with distances determined by shortest paths. In such situations, one must precompute the corresponding distance matrix in order to apply our method with the same worst-case running time.

2 The swap-based local search

Introduced by Teitz and Bart (1968), the standard local search procedure for the p-median problem is based on swapping facilities. For each facility $f_i \notin S$ (the current solution), the procedure determines which facility $f_r \in S$ (if any) would improve the solution the most if f_i and f_r were interchanged (i.e., if f_i were inserted and f_r removed from the solution). If any such "improving" swap exists, the best one is performed, and the procedure is repeated from the new solution. Otherwise we stop, having reached a *local minimum* (or *local optimum*). Arya et al. have recently proven (Arya et al., 2001) that this procedure is guaranteed to produce a solution whose value is at most 5 times the optimum in the metric setting (i.e., when the triangle inequality holds). On non-pathological instances (those more likely to appear in practice), empirical evidence shows that the algorithm is often within a few percentage points of optimality (and often does find the optimal solution), being especially successful when both p and n are small (Rosing, 1997).

Our main concern is not solution quality, but the time it takes to run each iteration of the algorithm. Given a solution S, we want to find an improving neighbor S' (if it exists) as fast as possible.

A straighforward implementation takes $O(pmn)$ time per iteration. Start by determining the closest and second closest open facilities for each user; this takes $O(pn)$ time. Then, for each candidate pair (f_i, f_r), compute the profit that would result from replacing f_r with f_i.

[1] More accurate representations of $\phi_1(u)$, $\phi_2(u)$, $d_1(u)$, and $d_2(u)$ would be $\phi_1^S(u)$, $\phi_2^S(u)$, $d_1^S(u)$, and $d_2^S(u)$, respectively, since each value is a function of S as well. Since the solution will be clear from context, we prefer the simpler representation in the interest of readability.

To do that, one can reason about each user u independently. If the facility that currently serves u is not f_r (the facility to be removed), the user will switch to f_i only if this facility is closer, otherwise it will remain where it is. If u is currently assigned to f_r, the user will have to be reassigned, either to $\phi_2(u)$ (the second closest facility) or to f_i (the facility to be inserted), whichever is closest. The net effect is summarized by following expression:

$$profit(f_i, f_r) = \sum_{u:\phi_1(u)\neq f_r} \max\{0, [d_1(u) - d(u, f_i)]\} - \sum_{u:\phi_1(u)=f_r} [\min\{d_2(u), d(u, f_i)\} - d_1(u)].$$

The first summation accounts for users that are not currently assigned to f_r (these can only gain from the swap), and the second for users that are (they can gain or lose something with the swap). In the distance oracle model, the entire expression can be computed in $O(n)$ time for each candidate pair of facilites. There are p candidates for removal and $m - p$ for insertion, so the total number of moves to consider is $p(m - p) = O(pm)$. Each iteration therefore takes $O(pmn)$ time.

Several papers in the literature use this basic implementation, and others avoid using the swap-based local search altogether mentioning its intolerable running time (Rolland, Schilling, and Current, 1996; Rosing and ReVelle, 1997; Voß, 1996). These methods would greatly benefit from asymptotically faster implementations, such as Whitaker's or ours.

3 Whitaker's implementation

In Whitaker (1983), describes the so-called *fast interchange* heuristic, an efficient implementation of the local search procedure defined above. Even though it was published in 1983, Whitaker's implementation was not widely used until 1997, when Hansen and Mladenović (1997) applied it as a subroutine of a Variable Neighborhood Search (VNS) procedure. A minor difference between the implementations is that Whitaker prefers a *first improvement* strategy (a swap is made as soon as a profitable one is found), while Hansen and Mladenović prefer *best improvement* (all swaps are evaluated and the most profitable executed). In our analysis, we assume best improvement is used, even in references to "Whitaker's algorithm."

The key aspect of this implementation is its ability to find in $\Theta(n)$ time the best possible candidate for removal, given a certain candidate for insertion. The pseudocode for the function that does that, adapted from Hansen and Mladenović (1997), is presented in Fig. 1.[2] Function `findOut` takes as input a candidate for insertion (f_i) and returns f_r, the most profitable facility to be swapped out, as well as the profit itself (*profit*).

Given a certain candidate for insertion f_i, the function implicitly computes $profit(f_i, f_r)$ for all possible candidates f_r. What makes this procedure fast is the observation (due to Whitaker) that the profit can be decomposed into two components, which we call *gain* and *netloss*.

Component *gain* accounts for all users who would benefit from the insertion of f_i into the solution. Each is closer to f_i than to the facility it is currently assigned to. The difference between the distances is the amount by which the cost of serving that particular user will be reduced if f_i is inserted. Lines 4 and 5 of the pseudocode compute *gain*.

The second component, *netloss*, accounts for all other customers, those that would not benefit from the insertion of f_i into the solution. If the facility that is closest to u is removed,

[2] In the code, an expression of the form $a \xleftarrow{+} b$ means that the value of a is incremented by b units.

```
function findOut (S, f_i, φ_1, φ_2)
1     gain ← 0; /* gain resulting from the addition of f_i */
2     forall (f ∈ S) do netloss(f) ← 0; /* loss resulting from removal of f */
3     forall (u ∈ U) do
4           if (d(u, f_i) ≤ d_1(u)) then /* gain if f_i is close enough to u */
5                 gain +← [d_1(u) − d(u, f_i)];
6           else /* loss if facility that is closest to u is removed */
7                 netloss(φ_1(u)) +← min{d(u,f_i), d_2(u)} − d_1(u);
8           endif
9     endforall
10    f_r ← argmin_{f∈S}{netloss(f)};
11    profit ← gain − netloss(f_r);
12    return (f_r, profit);
end findOut
```

Fig. 1 Function to determine, given a candidate for insertion (f_i), the best candidate for removal (f_r). Adapted from Hansen and Mladenović (1997)

u would have to be reassigned either to $\phi_2(u)$ (its current second closest facility) or to f_i (the new facility), whichever is closest. In both cases, the cost of serving u will either increase or remain constant. Of course, this reassignment will only be necessary if $\phi_1(u)$ is the facility removed to make room for f_i. This explains why *netloss* is an array, not a scalar value: there is one value associated with each candidate for removal. All values are initially set to zero in line 2; line 7 adds the contributions of the relevant users.

Given this $O(n)$-time function, it is trivial to implement the swap-based local search procedure in $O(mn)$ time per iteration: simply call `findOut` once for each of the $m - p$ candidates for insertion and pick the most profitable one. If the best profit is positive, perform the swap, update the values of ϕ_1 and ϕ_2, and proceed to the next iteration. Updating ϕ_1 and ϕ_2 requires $O(pn)$ time in the worst case, but the procedure can be made faster in practice, as mentioned in Whitaker (1983). Since our implementation uses the same technique, its description is deferred to the next section (Section 4.3.1).

4 An alternative implementation

Our implementation has some similarity with Whitaker's, in the sense that both methods perform the same basic operations. However, the order in which they are performed is different, and in our case partial results are stored in auxiliary data structures. As we will see, with this approach we can use values computed in early iterations of the local search procedure to speed up later ones.

4.1 Additional structures

Before we present our algorithm, let us analyze Whitaker's algorithm from a broader perspective. Its ultimate goal is to determine the pair (f_i, f_r) of facilities that maximizes $profit(f_i, f_r)$. To do so, it computes $gain(f_i)$ for every candidate for insertion, and $netloss(f_i, f_r)$ for every pair of candidates. (In the description in Section 3, *gain* is a scalar and *netloss* takes as input only the facility to be removed; however, both are computed inside a function that is called for each f_i, which accounts for the additional dimension.) Implicitly, what the algorithm

does is to compute profits as

$$profit(f_i, f_r) = gain(f_i) - netloss(f_i, f_r).$$

Our algorithm defines $gain(f_i)$ precisely as in Whitaker's algorithm: it represents the total amount gained if f_i is added to S, regardless of which facility is removed:

$$gain(f_i) = \sum_{u \in U} \max\{0, d_1(u) - d(u, f_i)\}. \tag{1}$$

Our method differs from Whitaker's in the computation of *netloss*. While Whitaker's algorithm computes it explicitly, we do it in an indirect fashion. For every facility f_r in the solution, we define $loss(f_r)$ as the increase in solution value that results from the removal of f_r from the solution (assuming that no facility is inserted). This is the cost of transferring every customer assigned to f_r to its second closest facility:

$$loss(f_r) = \sum_{u:\phi_1(u)=f_r} [d_2(u) - d_1(u)]. \tag{2}$$

As defined, *gain* and *loss* are capable of determining the net effect of a single insertion or a single deletion, but not of a swap, which is nothing but an insertion and a deletion that occur simultaneously. Whitaker's algorithm can handle swaps because it computes *netloss* instead of *loss*. To compute *netloss* from *loss*, we use yet another function, $extra(f_i, f_r)$, defined so that the following is true for all pairs (f_i, f_r):

$$netloss(f_i, f_r) = loss(f_r) - extra(f_i, f_r). \tag{3}$$

From the pseudocode in Fig. 1, it is clear that $netloss(f_i, f_r)$ is actually defined as

$$netloss(f_i, f_r) = \sum_{\substack{u:[\phi_1(u)=f_r]\wedge \\ [d(u,f_i)>d_1(u)]}} [\min\{d(u, f_i), d_2(u)\} - d_1(u)]. \tag{4}$$

Substituting the values in Eqs. (2) and (4) into Eq. (3), we obtain an expression for *extra*:

$$extra(f_i, f_r) = \sum_{u:\phi_1(u)=f_r} [d_2(u) - d_1(u)] - \sum_{\substack{u:[\phi_1(u)=f_r]\wedge \\ [d(u,f_i)>d_1(u)]}} [\min\{d(u, f_i), d_2(u)\} - d_1(u)].$$

It is possible to simplify this expression. First, consider a user u for which $\min\{d(u, f_i), d_2(u)\} = d_2(u)$. It has no net contribution to *extra*: whatever is added in the first summation is subtracted in the second. Therefore, we can write

$$extra(f_i, f_r) = \sum_{\substack{u:[\phi_1(u)=f_r]\wedge \\ [d(u,f_i)<d_2(u)]}} [d_2(u) - d_1(u)] - \sum_{\substack{u:[\phi_1(u)=f_r]\wedge \\ [d_1(u)<d(u,f_i)<d_2(u)]}} [d(u, f_i) - d_1(u)].$$

Note that the range of the first summation contains that of the second. We can join both into a single summation,

$$extra(f_i, f_r) = \sum_{\substack{u:[\phi_1(u)=f_r]\wedge \\ [d(u,f_i)<d_2(u)]}} [d_2(u) - d_1(u) - \max\{0, d(u, f_i) - d_1(u)\}],$$

which can be further simplified to

$$extra(f_i, f_r) = \sum_{\substack{u:[\phi_1(u)=f_r]\wedge \\ [d(u,f_i)<d_2(u)]}} [d_2(u) - \max\{d(u, f_i), d_1(u)\}]. \tag{5}$$

This is our final expression for *extra*. We derived it algebraically from simpler expressions, but it is possible to get it directly with a bit of case analysis. This alternative approach was used in an earlier version of our paper (Resende and Werneck, 2003).

Given the expressions of *gain*, *loss*, and *extra* (Eqs. (1), (2), and (5)), we can find the profit associated with each move in a very simple manner:

$$profit(f_i, f_r) = gain(f_i) - loss(f_r) + extra(f_i, f_r). \tag{6}$$

The interesting aspect of this decomposition of *profit* is that the only term that depends on both the facility to be inserted and the one to be removed is *extra*. Moreover, this term is always nonnegative (see Eq. 5). This will be relevant in the implementation of the local search itself, as the next section will make clear.

4.2 Local search

Our implementation of the local search procedure assumes that all necessary values (*loss*, *gain*, and *extra*) are stored in appropriate data structures: one-dimensional vectors for *loss* and *gain*, and a two-dimensional matrix for *extra*.[3] Once these structures are computed, one can easily find the best swap in $O(pm)$ time: just use Eq. (6) to determine the profit for each candidate pair of facilities and pick the minimum.

To compute *gain*, *loss*, and *extra*, we note that every entry in these structures is a summation over some subset of users (see Eqs. (1), (2), and (5)). The contribution of each user can therefore be computed independently. Function `updateStructures`, shown in Fig. 2, does exactly that. Given a user u and its two closest facilities in solution S (given by ϕ_1 and ϕ_2), it adds u's contribution to *loss*, *gain*, and *extra*. The total running time of the procedure is $O(m - p) = O(m)$, since it is essentially a loop through all the facilities that do not belong to the solution. Given this function, computing *gain*, *loss*, and *extra* from scratch is straightforward: first reset all entries in these structures, then call `updateStructures` once for each user. Together, these n calls perform precisely the summations defined in Eqs. (1), (2), and (5).

We now have all the elements necessary to build the local search procedure with $O(mn)$ operations. In $O(pn)$ time, compute $\phi_1(\cdot)$ and $\phi_2(\cdot)$ for all users. In $O(pm)$ time, reset *loss*,

[3] Note that *gain* and *loss* could actually share the same m-sized vector, since they are defined for disjoint sets of facilities.

Fig. 2 Pseudocode for updating arrays in the local search procedure

```
function updateStructures (S, u, loss, gain, extra, φ1, φ2)
1    f_r ← φ1(u);
2    loss(f_r) +← [d2(u) − d1(u)];
3    forall (f_i ∉ S) do
4        if (d(u, f_i) < d2(u)) then
5            gain(f_i) +← max{0, d1(u) − d(u, f_i)};
6            extra(f_i, f_r) +← [d2(u) − max{d(u, f_i), d1(u)}];
7        endif
8    endfor
end updateStructures
```

gain, and *extra*. With n calls to `updateStructures`, each taking in $O(m)$ time, determine their actual values. Finally, in $O(pm)$ time, find the best swap using Eq. (6).

4.3 Acceleration

At first, our implementation seems to be merely a complicated alternative to Whitaker's; after all, both have the same worst-case complexity. Furthermore, our implementation has the clear disadvantage of requiring an $O(pm)$-sized matrix, whereas $\Theta(n+m)$ memory positions are enough for Whitaker's. The additional memory, however, allows for significant accelerations, as this section will show.

When a facility f_r is replaced by a new facility f_i, certain entries in *gain*, *loss*, *extra*, ϕ_1, and ϕ_2 become inaccurate. The straighforward way to update them for the next local search iteration is to recompute ϕ_1 and ϕ_2, reset the other arrays, and then call `updateStructures` again for all users.

A downside of this approach is that no information gathered in one iteration is used in subsequent ones. As a result, unnecessary, repeated computations are bound to occur. In fact, the actions performed by `updateStructures` depend only on u, $\phi_1(u)$, and $\phi_2(u)$; no value is read from other structures. If $\phi_1(u)$ and $\phi_2(u)$ do not change from one iteration to another, u's contribution to *gain*, *loss*, and *extra* will not change either. This means there is no need to call `updateStructures` again for u.

To deal with such cases, we keep track of *affected users*. A user u is *affected* if there is a change in either $\phi_1(u)$ or $\phi_2(u)$ (or both) after a swap is made. Sufficient conditions for u to be affected after a swap between f_i and f_r are:

1. either $\phi_1(u)$ or $\phi_2(u)$ is f_r, the facility removed; or
2. f_i (the facility inserted) is closer to u than the original $\phi_2(u)$ is.

Contributions to *loss*, *gain*, and *extra* need only be updated for affected users. If there happens to be few of them (which is often the case, as Section 6.2.1 shows) significant gains can be obtained.

Note, however, that updating the contributions of an affected user u requires more than a call to `updateStructures`. This function simply adds new contributions, so we must first subtract the old contributions made by u. To acomplish this, we use a function similar to `updateStructures`, with subtractions instead of additions.[4] This function (`undoUpdateStructures`) must be called for all affected users *before* ϕ_1 and ϕ_2 are recomputed.

[4] This function is identical to the one shown in Fig. 2, with all occurrences of $\overset{+}{\leftarrow}$ replaced with $\overset{-}{\leftarrow}$: instead of incrementing values, we decrement them.

```
procedure localSearch (S, φ1, φ2)
1   A ← U; /* A is the set of affected users */
2   resetStructures (gain, loss, extra);
3   while (TRUE) do
4       forall (u ∈ A) do updateStructures (S, u, gain, loss, extra, φ1, φ2);
5       (fr, fi, profit) ← findBestNeighbor (gain, loss, extra);
6       if (profit ≤ 0) then break; /* no improvement, we are done */
7       A ← ∅;
8       forall (u ∈ U) do /* find out which users will be affected */
9           if ((φ1(u) = fr) or (φ2(u) = fr) or (d(u, fi) < d(u, φ2(u)))) then
10              A ← A ∪ {u}
11          endif
12      endforall;
13      forall (u ∈ A) do undoUpdateStructures (S, u, gain, loss, extra, φ1, φ2);
14      insert(S, fi);
15      remove(S, fr);
16      updateClosest(S, fi, fr, φ1, φ2);
17  endwhile
end localSearch
```

Fig. 3 Pseudocode for the local search procedure

Figure 3 contains the pseudocode for the entire local search procedure, already taking into account the observations just made. Apart from the functions already discussed, three other nontrivial ones appear in the code. Function `resetStructures` sets all entries in *gain*, *loss*, and *extra* to zero. Function `findBestNeighbor` runs through these structures and finds the most profitable swap using Eq. (6). It returns which facility to remove (f_r), the one to replace it (f_i), and the profit itself (*profit*). Finally, `updateClosest` updates ϕ_1 and ϕ_2, possibly using the fact that the facility recently opened was f_i and the one closed was f_r (Section 4.3.1 explains how this is done).

Restricting updates to affected users can result in significant speedups in the algorithm, as Section 6.2.1 shows. There are, however, other accelerations to exploit. The pseudocode reveals that all operations in the main loop run in linear time, with three exceptions:

- updating closeness information (calls to `updateClosest`);
- finding the best swap to be made (calls to `findBestNeighbor`);
- updating the auxiliary data structures (calls to `updateStructures` and `undoUpdateStructures`).

These are the potential bottlenecks of the algorithm, since they all run in quadratic time in the worst case. The next three subsections analyze how each of them can be dealt with.

4.3.1 Closeness

Updating closeness information, in our experience, has proven to be a relatively cheap operation. Deciding whether the newly inserted facility f_i becomes either the closest or the second closest facility to each user is trivial and can be done in $O(n)$ total time. A more costly operation is updating closeness information for customers who had f_r (the facility removed) as either the closest or the second closest element. With a straighforward implementation, updating each such affected user takes $O(p)$ time. Since there are usually few of them, the total time spent tends to be a small fraction of the entire local search procedure.

The whole update procedure could actually be performed in $O(n \log p)$ worst-case time. It suffices to keep, for each user u, the set of open facilities in a heap with priorities given by their distances to u. Since this solution requires $O(np)$ additional memory positions and is not significantly faster, we opted for using the straighforward implementation in our code.

It is also important to mention that finding the set of closest and second closest elements from scratch is itself a cheap operation in some settings, even in the worst case. For example, when distances between customers and facilities are given by shortest paths on an underlying graph, this can be accomplished in $\tilde{O}(|E|)$ time (Thorup, 2001), where $|E|$ is the number of edges in the graph.[5]

In practice, the generic approach above seems to be good enough. Section 6.2.5 shows that there is not much to gain from accelerating this part of the algorithm; together, other procedures already dominate the running time of the local search. We therefore do not use specialized routines in this paper; we always assume we are dealing with arbitrary distance matrices.

4.3.2 Best neighbor

Given a solution, the straighforward way to find the most profitable swap is to compute $profit(f_i, f_r)$ (as defined in Eq. (6)) for all candidate pairs of facilities and pick the best. Since each *profit* computation takes constant time and there are $p(m - p)$ potential swaps, the entire procedure requires $\Theta(pm)$ operations. In practice, however, the best move can be found in less time.

It is convenient to think of $extra(f_i, f_r)$ as a measure of the interaction between the neighborhoods of f_r and f_i. After all, Eq. (5) shows that only users that have f_r as their current closest facility and are also close to f_i (i.e., have f_i closer than the second closest open facility) contribute to $extra(f_i, f_r)$. In particular, if there are no users in this situation, $extra(f_i, f_r)$ will be zero. Section 6.2.2 shows that this occurs rather frequently in practice, especially when p is large (and hence the average number of users assigned to each f_r is small).

Therefore, instead of storing *extra* as a full matrix, one may consider representing only nonzero elements explicitly: each row becomes a linked list sorted by column number. A drawback of this sparse representation is the impossibility to make random accesses in $O(1)$ time. Fortunately, this is not necessary for our purposes. All three functions that access the matrix (`updateStructures`, `undoUpdateStructures`, and `bestNeighbor`) can be implemented so as to go through each row sequentially.

In particular, consider the implementation of `bestNeighbor`. First, it determines the facility f_i^* that maximizes $gain(f_i)$ and the facility f_r^* that minimizes $loss(f_r)$. Since all values in *extra* are nonnegative, the pair (f_i^*, f_r^*) is at least as profitable as any pair (f_i, f_r) for which $extra(f_i, f_r)$ is zero. Then, the procedure computes the exact profits (given by Eq. (6)) for all nonzero elements in *extra*.

The whole procedure takes $O(m + \lambda pm)$ time, where λ is the fraction of pairs whose *extra* value is nonzero. As already mentioned, this value tends to be smaller as p increases, thus making the algorithm not only faster, but also more memory-efficient (when compared to the "full matrix" representation).

[5] The $\tilde{O}(\cdot)$ notation hides polylogarithmic terms.

4.3.3 Updates

As we have seen, keeping track of affected users can reduce the number of calls to `updateStructures`. We now study how to reduce the time spent in each of these calls.

Consider the pseudocode in Fig. 2. Line 5 represents a loop through all $m - p$ facilities outside the solution, but line 6 shows that we can actually restrict ourselves to facilities that are closer to u than $\phi_2(u)$ is. This is often a small subset of the facilities, especially when p is large.

This suggests a preprocessing step that builds, for each user u, a list of all facilities sorted by increasing distance to u. During the local search, whenever we need the set of facilities whose distance to u is less than $d_2(u)$, we just take the appropriate prefix of the precomputed list, potentially with much fewer than $m - p$ elements.

Building these lists takes $O(nm \log m)$ time, but it is done only once, not in every iteration of the local search procedure. This is true even if local search is applied several times within a metaheuristic (such as in Hansen and Mladenović (1997), Resende and Werneck (2003), and Rosing and ReVelle (1997)): a single preprocessing step is enough.

A more serious drawback of this approach is memory usage. Keeping n lists of size m in memory requires $\Theta(mn)$ space, which may be prohibitive. An alternative is to keep only relatively small prefixes, not the full list. They would act as a cache: when $d_2(u)$ is small enough, we just take a prefix of the candidate list; when $d_2(u)$ is larger than the largest distance represented, we explicitly look at all possible neighbors (each in constant time).

In some circumstances, the "cached" version may be faster than the "full" version of the algorithm, since preprocessing is cheaper. After all, instead of creating sorted lists of size m, we create smaller ones of size k (for some $k < m$). Each list can be created in $O(m + k \log k)$ time: first we find the k smallest elements among all m in $O(m)$ time (Cormen et al., 2001), then we sort them in $O(k \log k)$ time. For small values of k, this is an asymptotic improvement over the $O(m \log m)$ time required (per list) in the "full" case.

4.3.4 The reordering problem

There is a slight incompatibility between the accelerations proposed in Sections 4.3.2 and 4.3.3. On the one hand, the sparse matrix data structure proposed in Section 4.3.2 guarantees efficient queries only when each row is accessed sequentially by column number (facility label). Section 4.3.3, on the other hand, assumes that facilites are accessed in nondecreasing order of *distance* from the user. Functions `updateStructures` and `undoUpdateStructures` use both data structures: they take a list of facilities sorted by distance, but must process them in nondecreasing order of label. We need to make these two operations compatible.

The simplest solution is to take the list of facilities sorted by distance and sort it again by label. If the list has size k, this takes $O(k \log k)$ time. In the worst case k is $O(m)$, so this introduces an extra $\log m$ factor in the complexity of the algorithm. In practice, however, k is rather small, and the overhead hardly noticeable. In fact, we used this approach in a preliminary version of our paper (Resende and Werneck, 2003).

Even so, one would like to do better. Recall that the original list is actually a prefix of the list of all facilities (sorted by distance). Even though the prefix varies in size, the underlying sorted list does not: it is a fixed permutation of facility labels. This means we need to solve the following generic problem:

Let π be a fixed permutation of the labels $\{1, 2, \ldots, m\}$, and let π_k be the size-k prefix of π, for $1 \leq k \leq n$ ($\pi_n = \pi$, by definition). Given any k, sort π_k by label in $O(k)$ time. At most $O(m)$ preprocessing time is allowed.

To solve this, we use an algorithm that mimics insertion sort on a list, but takes advice from an "oracle" built during preprocessing. Assume we need to sort π_k, for some k. One way to do it is to take each element of π_k and insert it into a new list, ordered by label. With standard insertion sort, this would take $O(k^2)$ time. However, if we knew in advance where to insert each element, the procedure would take $O(k)$ time. The oracle will give us exactly that.

Let $\pi(i)$ be the i-th element of π. We define $pred(i)$ to be the *predecessor* of $\pi(i)$, the element after which $\pi(i)$ should be inserted during the algorithm above. The oracle will give us $pred(i)$ for every i.

The values of $pred(i)$ are set in the preprocessing step. Initially, it creates an auxiliary doubly-linked list L containing $0, 1, 2, \ldots, m$, in this order (element 0 will act as a sentinel). This can be trivially done in $O(m)$ time. Then, it removes elements from L one by one in *reverse order* with respect to π. In other words, the first element removed from L is $\pi(m)$, then $\pi(m-1)$, and so on, until $\pi(1)$ is removed; in the end, only 0 (the sentinel) will remain in L. Upon removing element $\pi(i)$ from L, the algorithm sets $pred(i)$ to be the predecessor of $\pi(i)$ (in L itself) at that particular moment. This procedure takes $O(m)$ time for each of the n lists.

Note that this procedure is in fact a simulation of insertion sort, but in reverse order. List L originally has all the elements of π_m; after one removal, we are left with π_{m-1}, and so on. At all times, L is sorted by label; if it has size k, it represents what the sequence looks like after the k-th element is inserted during insertion sort.

Given all the $pred(\cdot)$ values, sorting π_k is simple. We start with a list L' containing only a sentinel (0); it can be singly-linked, with forward pointers only. We then access the first i elements of π (following π's own order), inserting each element $\pi(i)$ into L' right after $pred(i)$. Eventually, L' will contain all the elements of $\pi(k)$ sorted by label, as desired. The running time is only $O(k)$.

5 Generalization

Section 4 presented our algorithm as a local search procedure for the p-median problem. In fact, with slight modifications, it can also be applied to the facility location problem. Moreover, the ideas suggested here are not limited to local search: they can also be used to accelerate other important routines, such as path-relinking and tabu search. This section details the adaptations that must be made in each case.

5.1 Facility location

The input of the *facility location problem* consists of a set of users U, a set of potential facilities F, a distance function $d : U \times F \to \mathcal{R}_+$, and a *setup cost function* $c : F \to \mathcal{R}_+$. The first three parameters are the same as in the p-median problem. The difference is that here the number of facilities to open is not fixed; there is, instead, a cost associated with opening each facilty, the setup cost. The more facilities are opened, the greater the setup cost will be. The objective is to minimize the total cost of serving all customers, considering the sum of the setup and service cost (distances).

Any valid solution to the p-median problem is a valid solution to the facility location problem. To use the local search procedure suggested here for this problem, we have to adjust the algorithm to compute the cost function correctly. As it is, the algorithm computes the service costs correctly, but assumes that the setup costs are zero. But including them is trivial: the service cost depends only on whether a facility is open or not; it does not depend on other facilities. Consider a facility f_i that is not in the solution; when evaluating whether it should be inserted or not, we must account for the fact that its setup cost will increase the solution value by $c(f_i)$. Similarly, simply closing a facility f_r that belongs to the solution saves us $c(f_r)$. To take these values into account, if suffices to initialize *gain* and *loss* with the symmetric of the corresponding setup costs, and not with zero as we do with the p-median problem. In other words, we initialize $gain(f_i)$ with $-c(f_i)$, and $loss(f_r)$ with $-c(f_r)$.

This is enough to implement a swap-based local search for the facility location problem. Note, however, that there is no reason to limit ourselves to swaps—we could allow individual insertions and deletions as well. This is not possible with the p-median problem because the number of facilities is fixed, but there is no such constraint in the facility location problem.

No major change to the algorithm is necessary to support individual insertions and deletions. As already mentioned, $gain(f_i)$ is exactly the amount that would be saved if facility f_i were inserted into the solution (with no corresponding removal). Similarly, $loss(f_r)$ represents how much would be lost if the facility were removed (with no corresponding insertion). Positive values of *gain* and negative values of *loss* indicate that the corresponding move is worth making. The greater the absolute value, the better, and we can find the maximum in $O(m)$ time. Furthermore, we can continue to compute the costs associated with swaps if we wish to. In every iteration of the local search, we could therefore choose the best move among all swaps, insertions, and deletions. So we essentially gain the ability to make insertions and deletions with barely any changes to the algorithm.

We observe that the idea of a swap-based local search for the facility location problem is, of course, not new; it was first suggested in the literature by Kuehn and Hamburger (1963).

5.2 Other applications

It is possible to adapt the algorithm to perform other routines, not only local search. (In this discussion, we will always deal with the p-median problem itself, although the algorithms suggested here also apply to facility location with minor adaptations.)

Consider the path-relinking operation (Glover, 1996; Glover, Laguna, and Martí, 2000; Laguna and Martí, 1999; Resende and Ribeiro, 2005). It takes two solutions as inputs, S_1 and S_2, and gradually transforms the first (the *starting solution*) into the second (the *guiding solution*). It does so by swapping out facilities that are in $S_1 \backslash S_2$ and swapping in facilities from $S_2 \backslash S_1$. In each iteration of the algorithm, the best available swap is made. The goal of this procedure is to discover some promising solutions on the path from S_1 to S_2. The precise use of these solutions varies depending on the metaheuristic using this procedure.

This function is remarkably similar to the swap-based local search procedure. Both are based on the same kind of move (swaps), and both make the cheapest move on each round. There are two main differences:

1. *Candidate moves*: In path-relinking, only a subset of the facilities in the solution are candidates for removal, and only a subset of those outside the solution are candidates for insertion—and these subsets change (i.e., get smaller) over time, as the algorithm advances into the path.

2. *Stopping criterion*: Whereas the local search procedure stops as soon as a local minimum is found, non-improving moves are allowed in path-relinking: it continues until the guiding solution is reached.

As long as we take these differences into account, the implementation of the local search procedure can also handle path-relinking. We need to define two functions: one to return the appropriate set of candidates for insertion and deletion, another to check if the move chosen by `bestNeighbor` should be made or not (i.e., to determine if the stopping criterion was met). In Section 4, these functions were defined implicitly: the candidates for insertion are all facilities outside the solution, the candidates for deletion are those in the solution, and the stopping criterion consists of testing whether the profit associated with a move is positive. Defining them explicitly is trivial for both local search and path-relinking.

In fact, by redefining these two functions appropriately, we can implement other routines, such as a simple version of tabu search. At all times, we could have two lists: one for elements that are forbidden to be inserted into the solution, another for elements that cannot be removed. The candidate lists would contain the remaining facilities, and the stopping criterion could be any one used for tabu search (number of iterations, for instance).

6 Empirical analysis

This section has two main goals. One is to present some empirical data to back up some of the claims we have made to guide our search for a faster algorithm. The other goal is to demonstrate that the algorithms suggested here are indeed faster than previously existing implementations of the local search procedure for the p-median problem. To keep the analysis focused, we will not deal with the extensions proposed in Section 5.

6.1 Instances and methodology

We tested our algorithm on four classes of problems. Three of them, TSP, ORLIB and ODM, have been previously studied in the literature for the p-median problem. The fourth, RW, is introduced here as a set of instances that benefit less from our methods.

Class TSP contains three sets of points on the plane (with cardinality 1400, 3038, and 5934), originally used in the context of the traveling salesman problem (Reinelt, 1991). In the p-median problem, each point is both a user to be served and a potential facility, and distances are Euclidean. Following (Hansen, Mladenović, and Perez-Brito, 2001), we tested several values of p for each instance, ranging from 10 to approximately $n/3$, when comparing our algorithm to Whitaker's.

Class ORLIB, originally introduced in Beasley (1985), contains 40 graphs with 100 to 900 nodes, each with a suggested value of p (ranging from 5 to 200). Each node is both a user and a potential facility, and distances are given by shortest paths in the graph.

The instances in class ODM, proposed by Briant and Naddef (2004), model the *optimal diversity management problem*. In this problem, one must assemble a certain product that appears in a large number of configurations, each defined by the presence or absence of a certain number of features. Briant and Naddef give as an example the electrical wiring in cars. Assuming that setting up an assembly line for every possible configuration is not economically viable, only p configurations are actually produced. Requests for other configurations will be fulfilled by the least costly alternative that is compatible (i.e., contains all the necessary features) among those produced. The goal is to decide which p configurations to produce,

given the demand and the unit cost for each existing configuration. To model this as a p-median problem, we make each configuration both a user and a facility. The cost of serving user u with facility f is the demand of u times the unit cost of f, as long as configuration f is compatible with configuration u; otherwise, the cost is infinity. We tested our algorithm on the four instances cited in Briant and Naddef (2004), with 535, 1284, 3773, and 5535 configurations. As in Briant and Naddef (2004), we tested values of p from 5 to 20 in each case.[6]

In class RW, each instance is a square matrix in which entry (u, f) is an integer taken uniformly at random from the interval $[1, n]$ and represents the cost of assigning user u to facility f. Four values of n were tested (100, 250, 500, and 1000), each with values of p ranging from 10 to $n/2$, totaling 27 combinations.[7] The random number generator we used when creating these instances (and in the algorithm itself) was Matsumoto and Nishimura's *Mersenne Twister* (Matsumoto and Nishimura, 1998).

Recall that the algorithms tested here use the distance oracle model, which assumes that retrieving the distance between any user and any facility takes $O(1)$ time. This can be trivially achieved for intances in RW (with a table look-up) and TSP (from the Euclidean coordinates). For ORLIB, we compute the all-pairs shortest paths in advance, as it is usually done in the literature (Hansen and Mladenović, 1997; Hansen, Mladenović, and Perez-Brito, 2001). These computations are not included in the running times reported in this section, since they are the same for all methods (including Whitaker's). For ODM, to compute the distance between a user and a facility we need to know whether the user is covered by that facility or not. To answer this question in $O(1)$ time, we precompute an $n \times m$ boolean incidence matrix with this information. The same expected complexity could be achieved with a hash table, which potentially uses less space but has higher overhead for accessing each element. The time to build the incidence matrix is also not included in the times reported here.

All tests were performed on an SGI Challenge with 28 196-MHz MIPS R10000 processors (with each execution of the program limited to one processor) and 7.6 GB of memory. All algorithms were coded in C++ and compiled with the SGI MIPSpro C++ compiler (v. 7.30) with flags `-O3 -OPT:Olimit=6586`. The source code is available from the authors upon request, as are the RW instances.

All running times shown in this paper are CPU times, measured with the `getrusage` function, whose precision is 1/60 second. In some cases, actual running times were too small for this precision, so each algorithm was repeatedly run for at least 5 seconds. Overall times were measured, and averages reported here.

When comparing different local search methods, we applied them to the same initial solutions. These were obtained by two different algorithms. The first is greedy (Whitaker, 1983): starting from an empty solution, we insert one facility at a time, always picking the one that reduces the solution cost the most. The second algorithm is random: we just pick a set of p facilities uniformly at random as the initial solution. All tests with random solutions were repeated five times for each method, using five different random seeds.

Running times mentioned in this paper refer to the local search only, and they do not include the cost of building initial solution (which is the same for all methods).

[6] In Briant and Naddef (2004), the authors do not show results for p greater than 16 in the instance with 3773 nodes. We include results for 17 to 20 as well, for symmetry.

[7] More precisely: for $n = 100$, we used $p = 10$, 20, 30, 40, and 50; for $n = 250$, $p = 10$, 25, 50, 75, 100, and 125; for $n = 500$, $p = 10$, 25, 50, 100, 150, 200, and 250; and for $n = 1000$, $p = 10$, 25, 50, 75, 100, 200, 300, 400, and 500.

6.2 Results

This section presents an experimental comparison of several variants of our implementation and Whitaker's method, *fast interchange* (we will use FI for short). We implemented FI based on the pseudocode in Hansen and Mladenović (1997) (obtaining comparable running times); the most important function was presented here in Fig. 1.

*6.2.1 Basic algorithm (*FM*)*

We start with the most basic version of our implementation, in which *extra* is represented as a full (non-sparse) matrix. This version (called FM, for *full matrix*) already incorporates some acceleration, since calls to `updateStructures` are limited to affected users only. However, it does *not* include the accelerations suggested in Sections 4.3.2 (sparse matrix) and 4.3.3 (preprocessing).

To demonstrate that keeping track of affected users can lead to significant speedups, we devised the following experiment. We took one instance from each class: odm1284 (class ODM, 1284 nodes), pmed40 (class ORLIB, 900 nodes), fl1400 (class TSP, 1400 nodes), and rw1000 (class RW, 1000 nodes). Note that they all have a similar number of nodes. Each instance was tested with 99 different values of p, from 1% to 99% of m. Since for very large values of p the greedy algorithm almost always finds local optima (thus rendering the local search useless), the initial solutions used in this experiment are random.

For each run, we computed how many calls to `updateStructures` and to `undoUpdateStructures` would have been made if we were not keeping track of affected users, and how many calls were actually made (in both cases, we did not count calls at the start of the first iteration, which is just the initialization). The ratio between these values, in percentage terms, is shown in Fig. 4 (each point is the average of five runs).

It is clear that the average number of affected users is only a fraction of the total number of users, even for small values of p, and drops significantly as the number of facilities to open increases. In all four instances, the average number of affected users eventually drops below 1% of n. By exploiting this fact, our implementation definitely has the potential to be faster than FI.

To test if this is indeed the case in practice, we ran an experiment with all instances from the four classes, with the values of p listed in Section 6.1. We used both greedy and random initial solutions. For each instance, we computed the speedup obtained by our method when compared to FI, i.e., the ratio between the running times of FI and FM. Table 1 shows the best, the (geometric) mean, and the worst speedups thus obtained considering all instances in each class.[8] Values greater than 1.0 favor our method, FM.

The table shows that even the basic acceleration scheme achieves speedups of up to more than 40. There are cases, however, in which FM is actually slower than Whitaker's method. This happens for instances in which the local search procedure performs very few iterations, insufficent to amortize the overhead of using a matrix. This is more common with the greedy constructive heuristic, which is more likely to find solutions that are close to being local optima, particularly when p is very large or very small (the worst case among all instances

[8] Since we are dealing with ratios, geometric (rather than arithmetic) means seem to be a more sensible choice; after all, if a method takes twice as much time for 50% of the instances and half as much for the other 50%, it should be considered roughly equivalent to the other method. Geometric means reflect that, whereas arithmetic means do not.

Table 1 Speedup obtained by FM (full matrix, no preprocessing) over Whitaker's FI

Solution	Class	Best	Mean	Worst
random	ODM	41.66	12.67	2.95
	ORLIB	21.19	5.76	1.64
	RW	20.96	7.62	2.51
	TSP	28.92	11.29	1.95
greedy	ODM	20.10	4.49	0.89
	ORLIB	14.20	3.76	1.07
	RW	13.99	5.50	1.47
	TSP	31.96	10.72	1.96

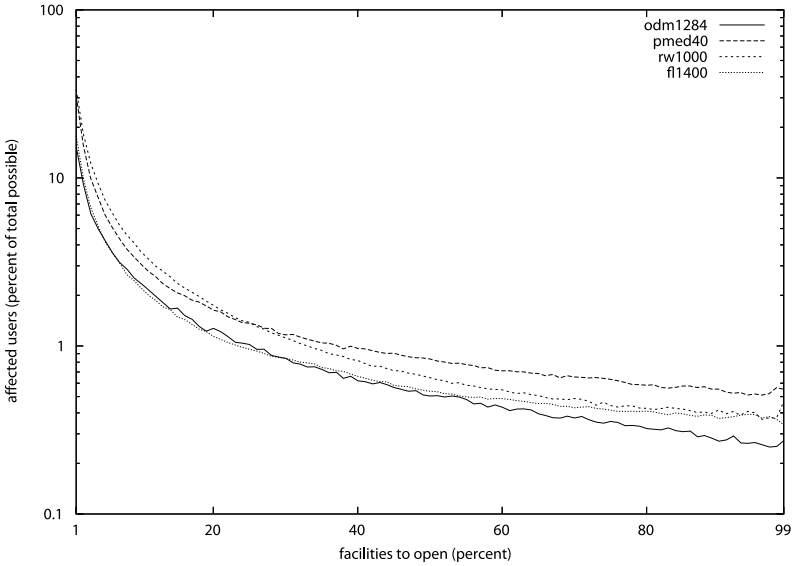

Fig. 4 Percentage of users affected during a run of the local search as a function of p (the percentage is taken over the set of all possible users that could have been affected, considering all iterations). One instance in each class is represented. Vertical axis is in logarithmic scale

happened with odm535 and $p = 6$). On average, however, FM has proven to be from three to more than ten times faster than FI.

6.2.2 Sparse matrix (SM)

We now analyze a second variant of our method. Instead of using a full matrix to represent *extra*, we use a sparse matrix, as described in Section 4.3.2. We call this variant SM. Recall that our rationale for using a sparse matrix was that the number of nonzero elements in the *extra* matrix is small. Figure 5 suggests that this is indeed true. For each of the four representative instances and each value of p (from 1 to 99% of m), it shows what fraction of the elements are nonzero (considering all iterations of the local search). The algorithm was run five times for each value of p, from five random solutions.

Although the percentage approaches 100% when the number of facilities to open is small, it drops very fast when p increases, approaching 0.1%. Note that rw1000, which is random, tends to have significantly more nonzeros for small values of p than other instances.

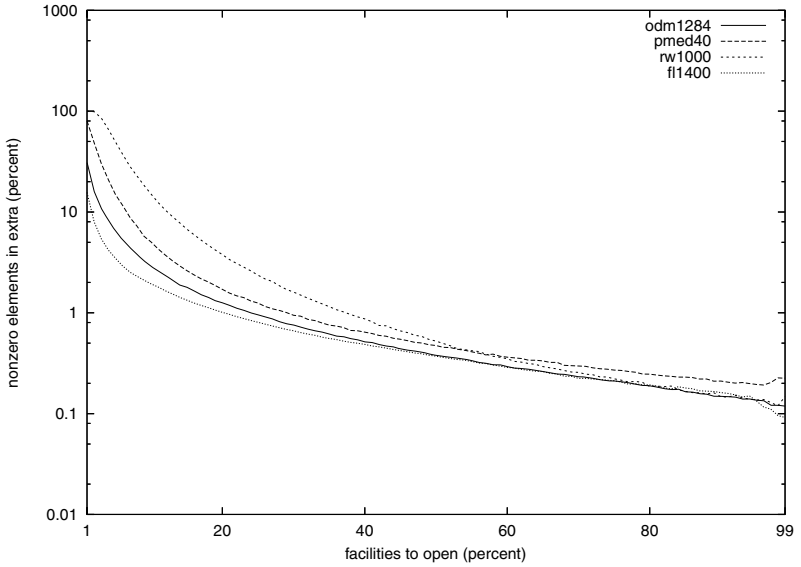

Fig. 5 Percentage of entries in the extra matrix that have nonzero values as a function of p. One instance of each class is represented. Vertical axis is in logarithmic scale

It is clear that the algorithm has a lot to benefit from representing only the nonzero elements of *extra*. However, the sparse matrix representation is much more involved than the array-based one, so some overhead is to be expected. Does it really reduce the running time of the algorithm in practice?

Table 2 shows that the answer to this question is "yes" most of the time. It represents the results obtained from all instances in the four classes, and contains the best, mean, and worst speedups obtained by SM over FI, for both types of initial solution (random and greedy).

As expected, SM has proven to be even faster than FM on average and in the best case (especially for the large instances with large values of p in the RW and TSP classes). However, some bad cases become slightly worse. This happens mostly for instances with small values of p: with a relatively large number of nonzero elements in the matrix, a sparse representation is not the best choice.

Table 2 Speedup obtained by SM (sparse matrix, no preprocessing) over Whitaker's FI

Solution	Class	Best	Mean	Worst
random	ODM	26.41	9.28	2.49
	ORLIB	46.88	6.66	1.19
	RW	114.36	12.47	1.95
	TSP	142.84	26.28	1.80
greedy	ODM	21.62	5.21	0.99
	ORLIB	24.88	4.36	1.00
	RW	49.35	8.36	1.22
	TSP	132.06	24.03	1.87

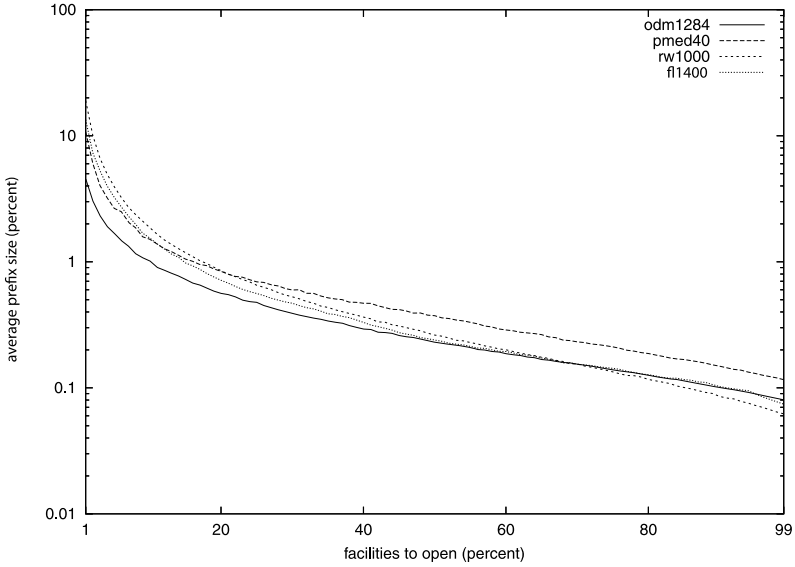

Fig. 6 Percentage of facilities actually visited when updating structures, for several values of p. One instance of each class is represented. Vertical axis is in logarithmic scale

6.2.3 Sparse matrix with preprocessing (SMP)

The last acceleration we study is the preprocessing step (Section 4.3.3), in which all potential facilities are sorted according to their distances from each of the users. We call this variant SMP, for *sparse matrix with preprocessing*. The goal of the acceleration is to avoid looping through all m facilities in each call to function `updateStructures` (and `undoUpdateStructures`). We just have to find the appropriate prefix of the ordered list.

Figure 6 shows the average size of the prefixes (as a percentage of m) that are actually checked by the algorithm, as a function of p (which varies from 1 to 99% of n). Initial solutions are random in this experiment.

As claimed before, the average prefix size is only tiny a fraction of m, for all but very small values of p. Considering only those prefixes instead of all facilities can potentially accelerate the local search. Of course, this does not come for free: the cost of preprocessing must be accounted for.

To determine the overall effect of these two conflicting factors, we tested SMP on all instances of our set. Table 3 shows the best, mean, and worst speedups obtained with respect to FI. Columns 3, 4, and 5 consider running times of the local search procedure only; columns 6, 7, and 8 also include preprocessing times.

The table shows that the entire SMP procedure (including preprocessing) is on average still much faster than Whitaker's FI, but often slightly slower than the other variants studied in this paper (FM and SM). However, as already mentioned, metaheuristics often need to run the local search procedure several times, starting from different solutions. Since preprocessing is run only once, its cost can be quickly amortized. Columns 3, 4, and 5 of the table show that once this happens, SMP can achieve truly remarkable speedups with respect not only to FI, but also to other variants studied in this paper. In the best case (instance rl5934 with $p = 800$), it is roughly 800 times faster than FI.

Table 3 Speedup obtained by SMP (sparse matrix, full preprocessing) over Whitaker's FI

		Local search only			Including preprocessing		
Solution	Class	Best	Mean	Worst	Best	Mean	Worst
random	ODM	46.18	13.77	3.42	8.26	3.00	0.87
	ORLIB	77.44	8.75	1.28	22.42	3.40	0.66
	RW	169.59	17.51	1.92	48.37	6.26	1.05
	TSP	812.80	186.81	4.63	128.03	31.92	1.89
greedy	ODM	33.16	7.21	1.33	3.30	0.67	0.15
	ORLIB	43.26	6.40	1.37	6.86	1.10	0.21
	RW	91.05	12.59	1.34	9.98	2.14	0.20
	TSP	695.57	161.86	5.11	71.42	18.92	1.45

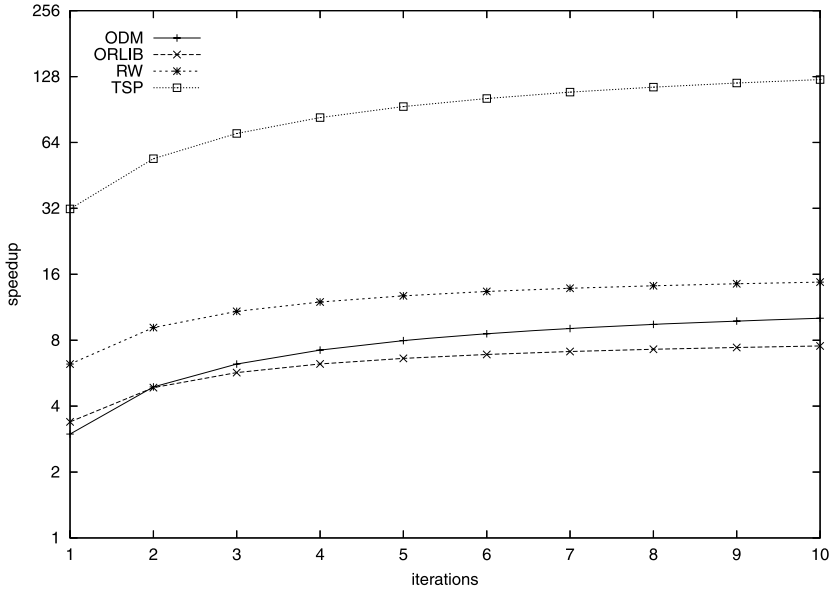

Fig. 7 Speedup of a multistart procedure implemented with SMP with respect to an implementation using Whitaker's method (FI)

To evaluate how fast the amortization is, consider what would happen in a simple multistart procedure. In each iteration, this algorithm generates a random solution and applies local search to it; the best solution found over all iterations is picked. We can predict the behavior of such a method (as far as running times are concerned) from the data used to build Table 3. After only one iteration, the mean speedups obtained when SMP is used instead of FI (Whitaker's method) will be those shown in the seventh column of the table. As the number of iterations increases, the mean speedups will gradually converge to the values in the fourth column. Figure 7 shows exactly what happens as a function of the number of iterations. After only ten iterations, the speedups are already close to those shown in the fourth column of Table 3: 10.1 for ODM, 7.5 for ORLIB, 14.7 for RW, and 124.0 for TSP.

Table 4 Speedup obtained by SM5 (sparse matrix, with preprocessing, cache size $5p/m$) over Whitaker's FI

		Local search only			Including preprocessing		
Solution	Class	Best	Mean	Worst	Best	Mean	Worst
random	ODM	46.12	13.68	3.42	14.48	4.04	0.86
	ORLIB	77.42	8.81	1.29	40.14	4.52	0.66
	RW	166.51	17.44	2.01	93.08	9.57	1.13
	TSP	774.96	176.42	4.49	283.71	62.97	2.20
greedy	ODM	32.65	7.16	1.30	6.23	0.96	0.14
	ORLIB	44.31	6.41	1.33	14.51	1.61	0.20
	RW	92.93	12.62	1.34	24.73	3.87	0.22
	TSP	747.72	160.93	5.07	177.62	40.65	1.73

Apart from the preprocessing time, another important downside of strategy SMP is memory usage: an array of size m is kept for each of the n customers. As mentioned in Section 4.3.3, one can use less memory by storing a vector with only a fraction of the m facilities for each customer. Table 4 shows what happens when we restrict the number of elements per vector to $5m/p$; we call this version of the local search SM5. In general, SMq is an algorithm that associates a list with qm/p facilities with each user. We use m/p as a parameter because this correlates well with the number of facilities each user has to look at to find an open one.

Tables 3 and 4 show that using restricted lists (as opposed to m-sized ones) can make the algorithm significantly faster when preprocessing times are considered. This is true especially for large instances. On average, SM5 is roughly twice as fast as SMP. The gains from a faster preprocessing more than offset the potential extra time incurred during the actual local search. In fact, the table also shows that the time spent on the main loop is barely distinguishable from SMP; the partial lists are almost always enough for the algorithm. Local search within SM5 can actually be slightly *faster* than within SMP. The possible cause here are cache effects; since less data is kept in memory, there is more locality to be exploited by the hardware.

6.2.4 Overall comparison

To get a better understanding of the performance of all variants proposed in this paper, we study in detail the largest instance in our set (rl5934, with almost 6000 customers and facilities). Figures 8 and 9 show the running times of several methods (FI, FM, SM, SM1, SM2, SM3, SM5, and SMP) for different values of p. Times are averages of five runs from different random solutions (the same set of initial solutions was given to each method). The first figure considers the local search only, whereas the second accounts for preprocessing times as well.

The figures show that for some methods, such as Whitaker's FI and the full-matrix variant of our implementation (FM), an increase in p leads to greater running times (although our method is still 10 times faster for $p = 1500$). For all other methods, which use sparse matrices, the time spent per iteration tends to decrease as p increases: the effect of swaps becomes more local, with fewer users affected and fewer neighboring facilities visited in each call to `updateStructures`. This latter effect explains why keeping even a relatively small list of neighboring facilities for each user seems to be worthwhile. The curves for variants SMP and SM5 are practically indistinguishable in Fig. 8, and both are much faster than SM (which keeps no list at all).

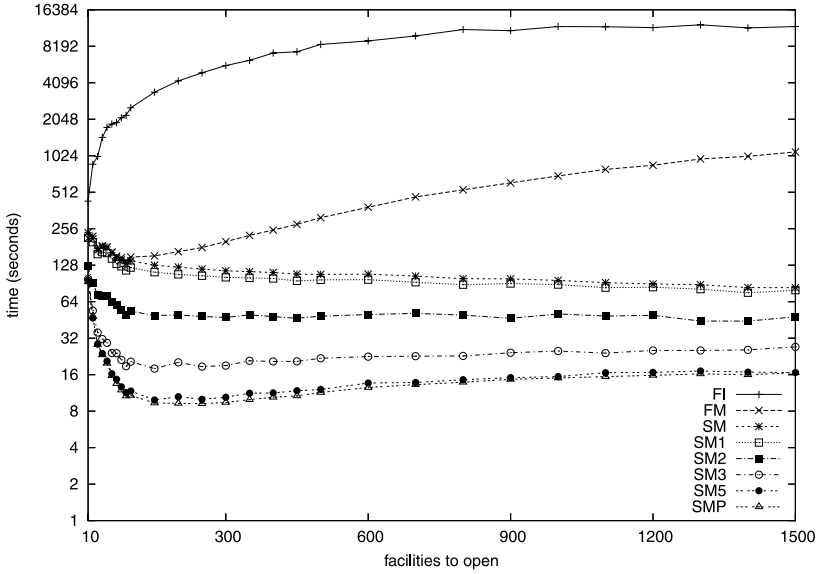

Fig. 8 Instance rl5934: dependency of running times on p for different methods. Times are in logarithmic scale and do not include preprocessing

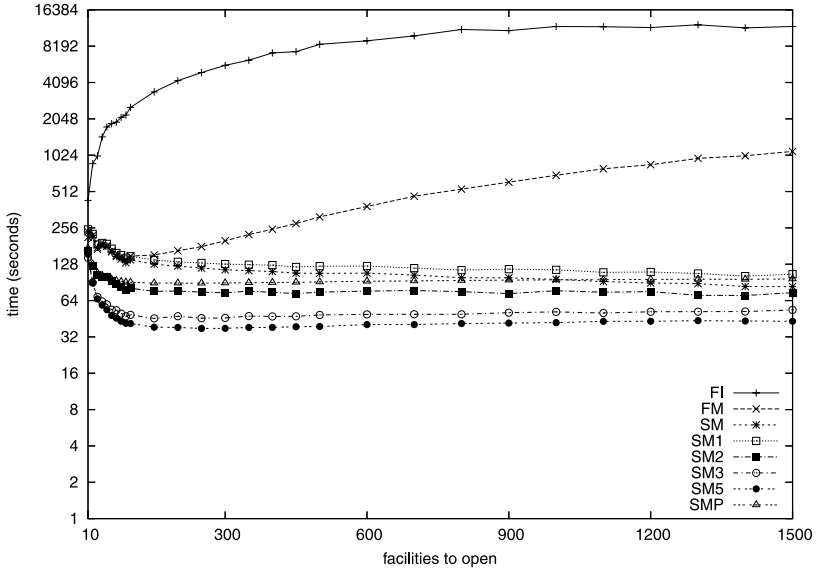

Fig. 9 Instance rl5934: dependency of running times on p for different methods. Times are in logarithmic scale and include preprocessing where applicable

As a final note, we observe that, because all methods discussed here implement the same algorithm, the number of iterations does not depend on the method itself. It does, however, depend on the value of p: in general, these two have a positive correlation for $p \leq m/2$, and negative from this point on, as Fig. 10 shows. This correlates well with the total number of solutions: there are $\binom{m}{p}$ solutions of size p, and this expression is maximized for $p = m/2$.

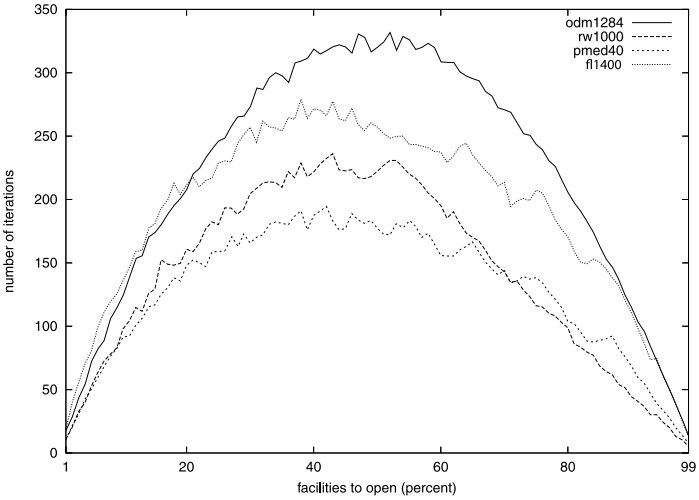

Fig. 10 Number of iterations of the local search procedure as a function of p, starting from random solutions. One instance from each class is represented

6.2.5 Profile

The results for SMP show that the modifications proposed in this paper can, together, result in significant acceleration. How much further can we go? Can additional modifications to the algorithm make it even faster?

These are open questions. However, we argue that small modifications are unlikely to lead to major gains, particularly when p is large. To support this claim, we devised the following experiment. For each class, we took the instance with the greatest number of users (n) and ran SMP with two values of p ($0.01n$ and $0.25n$), from five random solutions in each case. Table 5 shows the percentage of the total local search time (excluding preprocessing) spent in each section of the algorithm: initialization (which includes allocating the data structures), calls to `updateClosest`, calls to `updateStructures` (and `undoUpdate`

Table 5 Execution profile for method SMP: percentage of time spent on each of the potential bottlenecks (only the largest instance in each class is shown). Preprocessing times are not considered

Instance				Update	Update	Best	Other
Name	n, m	p	Init.	closest	struct.	neigh.	oper.
odm5535	5535	56	17.7	5.9	62.3	7.8	6.2
		1384	6.4	19.7	4.5	30.9	38.5
pmed40	900	9	6.7	1.7	89.8	0.6	1.2
		225	13.4	29.4	13.5	11.2	32.5
rw1000	1000	10	3.7	1.4	93.7	0.5	0.7
		250	12.1	26.7	15.1	14.5	31.6
rl5934	5934	60	12.2	5.7	74.0	5.0	3.1
		1484	10.7	41.0	4.6	22.7	21.0

`Structures`), calls to `bestNeighbor`, and other operations (such as determining which users are affected).

Note that calls to `updateStructures` and `undoUpdateStructures` dominate the running time for small values of p. This is to be expected: these functions run in $O(mn)$ time, while `bestNeighbor` and `updateClosest` run in $O(pn)$ and $O(pm)$ operations, respectively. When p increases, the running time for `updateStructures` and `undoUpdateStructures` actually decreases, since a larger fraction of the elements in the *extra* matrix will be zero (and therefore will not need to accessed). As a result, no component took more than 50% of the running time for $p = 0.25n$. In this case, even if we could make a component run in virtually no time, the algorithm would be at most twice as fast. A decent speedup, but not at all comparable to 800, the factor we were able to achieve in this paper. To obtain better factors, it seems necessary to work on all bottlenecks at once, or to come up with a different strategy altogether.

7 Concluding remarks

We have presented a new implementation of the swap-based local search for the p-median problem introduced by Teitz and Bart. We combine several techniques (using a matrix to store partial results, a compressed representation for this matrix, and preprocessing) to obtain speedups of up to three orders of magnitude with respect to the best previously known implementation, due to Whitaker. Our implementation is especially well suited to relatively large instances with moderate to large values of p and, due to the preprocessing step, to situations in which the local search procedure is run several times for the same instance (such as within a metaheuristic). When the local search has very few iterations, Whitaker's method can still be faster if the preprocessing time is considered.

An important test to the algorithms proposed here would be to apply them within more sophisticated metaheuristics. We have done that in (Resende and Werneck, 2004). That paper describes a multistart heuristic for the p-median problem that relies heavily on local search and path-relinking, both implemented according to the guidelines detailed in this paper. The algorithm has proved to be very effective in practice, obtaining remarkably good results (in terms of running times and solution quality) when compared to other methods in the literature.

A possible extension of our work presented would be to apply the methods and ideas presented here to problems beyond p-median and facility location. Swap-based local search is a natural procedure to be performed on problems such as maximum set cover, for example.

Acknowledgments We thank two anonymous referees for their helpful comments. We also thank Dennis Naddef for providing us with the instances in the `ODM` class.

References

Arya, V., N. Garg, R. Khandekar, A. Mayerson, K. Munagala, and V. Pandit. (2001). "Local Search Heuristics for k-Median and Facility Location Problems." In *Proc. 33rd ACM Symposium on the Theory of Computing*.

Avella, P., A. Sassano, and I. Vasil'ev. (2003). "Computational Study of Large-Scale p-Median Problems." Technical Report 08-03, DIS—Università di Roma "La Sapienza".

Avella, P., A. Sassano, and I. Vasil'ev. (2003). "A Heuristic for Large-Scale p-Median Instances." *Electronic Notes in Discrete Mathematics*, 13, 1–4.

Beasley, J.E. (1985). "A Note on Solving Large p-Median Problems." *European Journal of Operational Research*, 21, 270–273.

Briant, O. and D. Naddef. (2004). "The Optimal Diversity Management Problem." *Operations Research*, 52(4), 515–526.

Cormen, T., C. Leiserson, R. Rivest, and C. Stein. (2001). *Introduction to Algorithms*, 2nd edn. MIT Press.

Cornuéjols, G., M.L. Fisher, and G.L. Nemhauser. (1977). "Location of Bank Accounts to Optimize Float: An Analytical Study of Exact and Approximate Algorithms." *Management Science*, 23, 789–810.

du Merle, O., D. Villeneuve, J. Desrosiers, and P. Hansen. (1999). "Stabilized Column Generation." *Discrete Mathematics*, 194, 229–237.

Galvão, R.D. (1980). "A Dual-Bounded Algorithm for the *p*-Median Problem." *Operations Research*, 28, 1112–1121.

García-López, F., B. Melián-Batista, J.A. Moreno-Pérez, and J.M. Moreno-Vega. (2003). "Parallelization of the Scatter Search for the *p*-Median Problem." *Parallel Computing*, 29(5), 575–589.

Glover, F. (1996). "Tabu Search and Adaptive Memory Programming: Advances, Applications and Challenges." In R.S. Barr, R.V. Helgason, and J.L. Kennington (eds.), *Interfaces in Computer Science and Operations Research*, Kluwer, pp. 1–75.

Glover, F., M. Laguna, and R. Martí. (2000). "Fundamentals of Scatter Search and Path Relinking." *Control and Cybernetics*, 39, 653–684.

Goodchild, M.F. and V. Noronha. (1983). "Location-Allocation for Small Computers." Monograph 8, Department of Geography, University of Iowa.

Hansen, P. and N. Mladenović. (1997). "Variable Neighborhood Search for the *p*-Median." *Location Science*, 5, 207–226.

Hansen, P., N. Mladenović, and D. Perez-Brito. (2001). "Variable Neighborhood Decomposition Search." *Journal of Heuristics*, 7(3), 335–350.

Hodgson, M.J. (1978). "Toward More Realistic Allocation in Location-Allocation Models: An Interaction Approach." *Environment and Planning A*, 10, 1273–1285.

Kariv, O. and L. Hakimi. (1979). "An Algorithmic Approach to Nework Location Problems, Part II: The *p*-Medians." *SIAM Journal of Applied Mathematics*, 37(3), 539–560.

Kuehn, A.A. and M.J. Hamburger. (1963). "A Heuristic Program for Locating Warehouses." *Management Science*, 9(4), 643–666.

Laguna, M. and R. Martí. (1999). "GRASP and Path Relinking for 2-Layer Straight Line Crossing Minimization." *INFORMS Journal on Computing*, 11, 44–52.

Maranzana, F.E. (1964). "On the Location of Supply Points to Minimize Transportation Costs." *Operations Research Quarterly*, 15(3), 261–270.

Matsumoto, M. and T. Nishimura. (1998). "Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudorandom Number Generator." *ACM Transactions on Modeling and Computer Simulation*, 8(1), 3–30.

Reinelt, G. (1991). "TSPLIB: A Traveling Salesman Problem Library." *ORSA Journal on Computing*, 3, 376–384. *http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95/*.

Resende, M.G.C. and C.C. Ribeiro. (2005). "GRASP with Path-Relinking: Recent Advances and Applications." In T. Ibaraki, K. Nonobe, and M. Yagiura (eds.), *Metaheuristics: Progress as Real Problem Solvers*, Kluwer. In press.

Resende, M.G.C. and R.F. Werneck. (2003). "On the Implementation of a Swap-Based Local Search Procedure for the *p*-Median Problem." In R.E. Ladner (ed.), *Proceedings of the Fifth Workshop on Algorithm Engineering and Experiments (ALENEX'03)*, SIAM, pp. 119–127.

Resende, M.G.C. and R.F. Werneck. (2004). "A Hybrid Heuristic for the *p*-Median Problem." *Journal of Heuristics*, 10(1), 59–88.

Rolland. E., D.A. Schilling, and J.R. Current. (1996). "An Efficient Tabu Search Procedure for the *p*-Median Problem." *European Journal of Operational Research*, 96, 329–342.

Rosing, K.E. (1997). "An Empirical Investigation of the Effectiveness of a Vertex Substitution Heuristic." *Environment and Planning B*, 24, 59–67.

Rosing, K.E. and C.S. ReVelle. (1997). "Heuristic Concentration: Two Stage Solution Construction." *European Journal of Operational Research*, 97, 75–86.

Rosing, K.E., C.S. ReVelle, and H. Rosing-Vogelaar. (1979). "The *p*-Median and its Linear Programming Relaxation: An Approach to Large Problems." *Journal of the Operational Research Society*, 30(9), 815–823.

Senne, E.L.F. and L.A.N. Lorena. (2000). "Langrangean/Surrogate Heuristics for *p*-Median Problems." In M. Laguna and J.L. González-Velarde (eds.), *Computing Tools for Modeling, Optimization and Simulation: Interfaces in Computer Science and Operations Research*, Kluwer, pp. 115–130.

Senne, E.L.F. and L.A.N. Lorena. (2002). "Stabilizing Column Generation using Lagrangean/Surrogate Relaxation: An Application to *p*-Median Location Problems." *European Journal of Operational Research*. To appear.

Senne, E.L.F., L.A.N. Lorena, and M.A. Pereira. (2005). "A Branch-and-Price Approach to p-Median Location Problems." *Computers and Operations Research*, 32, 1655–1664.

Taillard, E.D. (2003). "Heuristic Methods for Large Centroid Clustering Problems." *Journal of Heuristics*, 9(1), 51–74.

Teitz, M.B. and P. Bart. (1968). "Heuristic Methods for Estimating the Generalized Vertex Median of a Weighted Graph." *Operations Research*, 16(5), 955–961.

Thorup M. (2001). "Quick k-Median, k-Center, and Facility Location for Sparse Graphs." In *Proceedings of the 28th International Colloquium on Automata, Languages and Programming (ICALP 2001)*, Volume 2076 of *Lecture Notes in Computer Science*, Springer, pp. 249–260.

Voß, S. (1996). "A Reverse Elimination Approach for the p-Median Problem." *Studies in Locational Analysis*, 8, 49–58.

Whitaker, R. (1983). "A Fast Algorithm for the Greedy Interchange of Large-Scale Clustering and Median Location Prolems." *INFOR*, 21, 95–108.