UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 12-19-(13)
Statistics and Econometrics Series
June 2012

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

A P-MEDIAN PROBLEM WITH DISTANCE SELECTION

Stefano Benati and Sergio García

Abstract

This paper introduces an extension of the p-median problem and its application to clustering, in which the distance/dissimilarity function between units is calculated as the distance sum on the q most important variables. These variables are to be chosen from a set of m elements, so a new combinatorial feature has been added to the problem, that we call the p-median model with distance selection. This problem has its origin in cluster analysis, often applied to sociological surveys, where it is common practice for a researcher to select the q statistical variables they predict will be the most important in discriminating the statistical units before applying the clustering algorithm. Here we show how this selection can be formulated as a non-linear mixed integer optimization mode and we show how this model can be linearized in several different ways. These linearizations are compared in a computational study and the results outline that the radius formulation of the p-median is the most efficient model for solving this problem.

Keywords: p-median problem, distance selection, radius formulation

Stefano Benati, Dipartimento di Sociologia e Ricerca Sociale, Universit`a degli Studi di Trento, Trento, Italy. Email address: Stefano.Benati@unitn.it.
Sergio García (corresponding author), Departamento de Estadística, Universidad Carlos III, Leganés (Madrid), Spain. Email address: sergio.garcia@uc3m.es.

A p-median problem with distance selection

Stefano Benati

Dipartimento di Sociologia e Ricerca Sociale, Università degli Studi di Trento, Trento, Italy

Sergio García*

Departamento de Estadística, Universidad Carlos III, Leganés (Madrid), Spain

Abstract

This paper introduces an extension of the p-median problem and its application to clustering, in which the distance/dissimilarity function between units is calculated as the distance sum on the q most important variables. These variables are to be chosen from a set of m elements, so a new combinatorial feature has been added to the problem, that we call the p-median model with distance selection. This problem has its origin in cluster analysis, often applied to sociological surveys, where it is common practice for a researcher to select the q statistical variables they predict will be the most important in discriminating the statistical units before applying the clustering algorithm. Here we show how this selection can be formulated as a non-linear mixed integer optimization mode and we show how this model can be linearized in several different ways. These linearizations are compared in a computational study and the results outline that the radius formulation of the p-median is

*Corresponding author

Email addresses: Stefano.Benati@unitn.it (Stefano Benati), sergio.garcia@uc3m.es (Sergio García)

the most efficient model for solving this problem.

Keywords: p-median problem, distance selection, radius formulation

1. Introduction

This paper studies the following clustering problem: suppose that we are given a set $\mathcal{U} = \{u_i\}_{i=1}^n$ of statistical units that are measured through a set of quantitative or qualitative features $\mathcal{F} = \{f_k\}_{k=1}^m$. These units and features are collected in a data matrix $V = [v_{ik}]$, where v_{ik} is the value that feature f_k takes for unit u_i. Our goal is to partition this set $\mathcal{U}$ into clusters such that the units classified in the same cluster are as similar as possible.

Clustering is one of the most common techniques of multivariate and exploratory data analysis and its principles can be found in many text books (for example, [9, 10]). The most common methods for clustering consist of two steps: in the first step, a dissimilarity measure or distance d_{ij} is established for every pair of units (u_i, u_j); in the second step, a clustering algorithm is applied to obtain the data partition. The most common methods are hierarchical partitions ([4]), or the application of the k-means method ([18]), and its variants (see [15]). One of these methods is based on the classical p-median problem ([14]), a model whose application to clustering is recorded in several studies ([20, 19, 16, 23, 1, 17]). More precisely, this model establishes the best clustering as the partition that after having chosen p representative elements (*medians*), then minimizes the total sum of the distances (equivalently, average distance) between each unit and its closest median. The model can be expressed in combinatorial terms as follows:

$$\min_{\substack{P \subseteq \mathcal{U} \\ |P| = p}} \sum_{u_i \in \mathcal{U}} \min\{d(u_i, u_j) \ / \ u_j \in P\}.$$

The model that we introduce in this paper assumes that the set $\mathcal{F}$ is large: not all statistical variables are relevant for clustering. In other words, some of them are useless or irrelevant to cluster membership, and including them in the distance function would only blur the distance/dissimilarity measure d_{ij} with redundant or irrelevant information. This situation is common when working with large databases as, for example, in large opinion polls or in census data, where the features $\mathcal{F}$ are more than hundreds. For example, the surveys contained in the World Value Survey repository ([24]) contain samples in which each statistical unit constitutes the answer to a questionnaire of more than 200 questions (i.e.,$m > 200$).

The most important aspects that describe the differences between units are often unclear. For example, we do not know a priori which social variables affect religious versus secular attitudes, or how righ-wing/left wing voters are characterized. As a consequence, the researcher reduces the database dimension considering small subsets of statistical variables: $Q \subseteq \mathcal{F}$. Then, for every chosen subset Q, the researcher defines the distance function d^Q, solves a p-median model and observes the clustering result. In practice, this choice is often arbitrary: Q is selected by rule of thumb (for example, with the support of some descriptive statistics). However, since the number of subsets Q grows exponentially with the size of $\mathcal{F}$, the researcher would benefit from a method of automatic pattern recognition.

In this paper, we propose a model that simultaneously selects the best set Q, the optimal medians P and the optimal partition. This model is an extension of the p-median problem, but in which the computation of the distances depends not only on the medians $P \subseteq \mathcal{U}$ but also on the selected features $Q \subseteq \mathcal{F}$. In combinatorial terms, the problem is formulated as follows:

$$\min_{\substack{P \subseteq \mathcal{U},\ |P| = p \\ Q \subseteq \mathcal{F},\ |Q| = q}} \sum_{u_i \in \mathcal{U}} \min\{d_{ij}^{Q} \ / \ u_j \in P\},$$

As can be seen, we have retained the p-median (and k-means) clustering approach that optimal partitions are characterized by the min-sum objective function and we call this problem *the p-median problem with distance selection*.

In some applications, the distance d_{ij}^{Q} is (or can be turned into) a linear function, so the problem is more tractable from a computational point of view. This is the case when we use the Manhattan distance (instead of the Euclidean, for example). In the applications that motivated this paper, we dealt with qualitative or ordinal data, coming from opinion polls or ranking assessments (see [2, 3, 24]). Qualitative data are expressed in the 0/1 binary scale, corresponding to the presence/absence of a feature, while ordinal data may represent an A-B-C quality assessment, or the 1-5 range of a Likert scale (1="Strongly disagree", 2="Disagree", 3="Neither agree nor disagree", 4="Agree", 5="Strongly agree"). For this kind of data, the most natural distance function is the Manhattan distance, for which $d_{ij} = \sum_k |v_{ik} - v_{jk}|$ and, therefore, $d_{ij}^{Q} = \sum_{k \in Q} |v_{ik} - v_{jk}|$.

As will be shown in Section 2, the p-median model can be directly extended to consider the decision variables $Q \subseteq \mathcal{F}$ too, but this extension leads to a quadratic non-convex problem. Instead of developing new solution tools for this particular non-linear model, our approach is to study different reformulations as a mixed integer linear problem and to determine which is the most efficient. The first formulation is a direct linearization of the initial quadratic model, and the second looks like an extension of the classic p-median formulation introduced in [21]. Two alternative formulations are then obtained through an arithmetic manipulation of these models, with the aim of reducing their size. Finally, the last model is the extension of the so-called radius formulation of the p-median problem: initially proposed in [5] for the uncapacitated facility location problem, this idea was employed in the following twenty-five years only in very few papers ([22, 6]), until recently, when it was used to solve the p-center problem ([8]). In recent years, the main advantage of this formulation (a reduced number of variables and constraints) has been exploited successfully to solve the p-median problem ([7, 11]) and the p-hub median problem ([12]). This paper follows this stream of reasearch, showing that the radius formulation is the most efficient way to solve the p-median problem with distance selection.

The rest of the paper is organized as follows. The p-median problem with distance selection is formulated in Section 2. The five different linear formulations are proposed in Section 3 and a computational study is carried out in Section 4. Finally, some conclusions are given in Section 5.

2. Model formulation

Assume that we are given a sample $\mathcal{U} = \{u_i\}_{i=1}^n$ of statistical units. For every unit i, the set $\mathcal{F} = \{f_k\}_{k=1}^m$ of statistical variables (i.e., features) is measured. We assume that, as is common in opinion polling or attitude surveys, variables f_k are represented by qualitative or ordinal data. If the data are qualitative, they are represented by 0-1. If the data are ordinal with g occurencies, or they are represented by a Likert scale with a finite number g of tiers, then we will refer to g as the dimension of the scale.

Let v_{ik} be the record of variable k for unit i. The distance, or difference, between unit u_i and unit u_j with respect to the feature f_k is $d_{ijk} = |v_{ik} - v_{jk}|$ and the overall distance between u_i and u_j is the 1-norm:

$$d_{ij} = \sum_{k=1}^{m} d_{ijk} = \sum_{k=1}^{m} |v_{ik} - v_{jk}|.$$

Suppose now that only a subset $Q \subseteq \mathcal{F}$ of statistical variables are considered relevant for the analysis and that, as a consequence, the differences between the units are calculated using Q only. The distance formula is thus expressed using the incidence vector z of subset Q:

$$d_{ij} = \sum_{k=1}^{m} d_{ijk} z_k,$$

where $z_k = 1$ if $f_k \in Q$ and $z_k = 0$ otherwise.

Data must be clustered using the p-median model and the min-sum criterion, so that its outcome consists of p clusters and its median is the most representative element, i.e., the cluster archetype. We define binary variables $y_j,\ j = 1, \ldots, n$, that take value 1 if unit j is a median (and 0 otherwise), and binary assignment variables x_{ij}, $i, j = 1, \ldots, n$, that take value 1 if unit i is assigned to the cluster j (and 0 otherwise). The model that we obtain is the following:

$$(F_0) \quad \min \quad \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\sum_{k=1}^{m} d_{ijk} \right) x_{ij}$$

$$
\begin{aligned}
\text{s.t.} \quad & x_{ij} \leq y_j, && i,j = 1,\ldots,n, \\
& \sum_{j=1}^{n} x_{ij} = 1, && i = 1,\ldots,n, \\
& \sum_{j=1}^{n} y_j = p, && \\
& x_{ij} \geq 0, && i,j = 1,\ldots,n, \\
& y_j = \{0,1\}\,, && j = 1,\ldots,n.
\end{aligned}
$$

Note that the assignment variables x_{ij} need not be declared binary because for fixed values of y there is always an optimal integral assignment.

If we now impose that only a subset $Q \subseteq \mathcal{F}$ of variables is to be used, we introduce new binary variables $z_k,\ k = 1,\ldots,m$, that stand for the incidence vector of Q, i.e., $z_k = 1$ if $f_k \in Q$ (and $z_k = 0$ otherwise). Then, the previous model is enlarged with the new variables z_k:

$$
\begin{aligned}
(F_1) \quad \min \quad & \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sum_{k=1}^{m} d_{ijk} z_k\right) x_{ij} && \\
\text{s.t.} \quad & x_{ij} \leq y_j, && i,j = 1,\ldots,n, \\
& \sum_{j=1}^{n} x_{ij} = 1, && i = 1,\ldots,n, \\
& \sum_{j=1}^{n} y_j = p, && \\
& \sum_{k=1}^{m} z_k = q, && \\
& x_{ij} \geq 0, && i,j = 1,\ldots,n, \\
& y_j \in \{0,1\}, && j = 1,\ldots,n, \\
& z_k \in \{0,1\}, && k = 1,\ldots,m.
\end{aligned}
$$

This formulation is non-linear because of the quadratic terms in the objective function. Moreover, the objective function is non-convex because the distance matrix is not positive semidefinite (the terms d_{ijk} can be arranged in such a way that the matrix is composed of zeros on the main diagonal and has positive or null terms elsewhere). Since the potential to solve to optimality large instances of a non-convex quadratic model is quite low, our goal is to linearize this formulation (F_1).

3. Integer linear formulations of p-median with distance selection

In this section, we propose five different ways to reformulate model (F_1) as a mixed integer linear problem.

3.1. Direct linearization of the quadratic variables

Formulation (F_1) is a mixed integer quadratic problem that can be turned into a linear mixed integer problem by introducing the new variables

$$w_{ijk} = x_{ij} z_k, \quad i, j = 1, \ldots, n, \ k = 1, \ldots, m.$$

Observe that variables w_{ijk} represent the 0-1 assignment of unit i to median j using variable k. Then, we need to add the inequalities

$$w_{ijk} \le x_{ij}, \qquad i, j = 1, \ldots, n, \ k = 1, \ldots, m, \tag{1}$$

$$w_{ijk} \le z_k, \qquad i, j = 1, \ldots, n, \ k = 1, \ldots, m, \tag{2}$$

$$w_{ijk} \ge x_{ij} + z_k - 1, \qquad i, j = 1, \ldots, n, \ k = 1, \ldots, m,$$

$$x_{ij} \in \{0, 1\}, \qquad i, j = 1, \ldots, n,$$

$$z_k \in \{0, 1\}, \qquad k = 1, \ldots, m.$$

Although we are imposing that variables x_{ij} must be binary, we can see that this condition can be replaced by $x_{ij} \geq 0$. Since we are considering a minimization problem, distances d_{ijk} are positive, and variables z_k and y_j are binary, there is a continuous optimal assignment x_{ij} that is binary. Besides, inequalities (1) and (2) are not necessary because they are satisfied at every optimal solution of the following model that linearizes formulation (F_1):

$$
\begin{array}{lll}
(F_2) \quad \min & \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m} d_{ijk} w_{ijk} & \\
\text{s.t.} & x_{ij} \leq y_j, & i,j = 1,\dots,n, \\
& \displaystyle\sum_{j=1}^{n} x_{ij} = 1, & i = 1,\dots,n, \\
& \displaystyle\sum_{j=1}^{n} y_j = p, & \\
& \displaystyle\sum_{k=1}^{m} z_k = q, & \\
& w_{ijk} \geq x_{ij} + z_k - 1, & i,j = 1,\dots,n,\ k = 1,\dots,m, \\
& w_{ijk} \geq 0, & i,j = 1,\dots,n,\ k = 1,\dots,m, \\
& x_{ij} \geq 0, & i,j = 1,\dots,n, \\
& y_j \in \{0,1\}, & j = 1,\dots,n, \\
& z_k \in \{0,1\}, & k = 1,\dots,m.
\end{array}
$$

This formulation contains $n^2(m+1)+n+2$ constraints and $n^2(m+1)+n+m$ variables, $n+m$ of which are binary. It must be remarked that, for fixed values of z, this is basically a p-median problem.

3.2. A p-median style formulation

An alternative formulation using only variables w_{ijk}, y_j and z_k is the following:

$$
\begin{aligned}
(F_3) \quad \min \quad & \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m} d_{ijk} w_{ijk} && \\
\text{s.t.} \quad & \sum_{k=1}^{m} w_{ijk} \leq q y_j && i,j = 1,\ldots,n, && (3)\\
& \sum_{j=1}^{n} w_{ijk} = z_k && i = 1,\ldots,n,\ k = 1,\ldots,m, && (4)\\
& \sum_{\substack{t=1,\\ t\neq k}}^{m} w_{ijt} \geq (q-1) w_{ijk} && i,j = 1,\ldots,n,\ k = 1,\ldots,m, && (5)\\
& \sum_{j=1}^{n} y_j = p, && &&\\
& \sum_{k=1}^{m} z_k = q, && &&\\
& w_{ijk} \geq 0, && i,j = 1,\ldots,n,\ k = 1,\ldots,m, &&\\
& y_j \in \{0,1\}, && j = 1,\ldots,n, &&\\
& z_k \in \{0,1\}, && k = 1,\ldots,m. &&
\end{aligned}
$$

Constraints (3) impose that unit i can be allocated to unit j only if this unit is chosen as a cluster median. Note that this is an aggregation of the constraints $w_{ijk} \leq y_k$: they are equivalent but the aggregated version (3) performed slightly better than the disaggregated form. Constraints (4) establish that, for fixed i and k, unit i must be allocated to some cluster median using variable k if, and only if, this variable k is selected. Finally, inequalities (5) guarantee the correct synchronization of the assignments: if unit i is allocated to median j using variable k, then it must be allocated to the same median j using the other $q-1$ selected variables.

This model has $n^2(m+1)+nm+2$ inequalities and n^2m+m+n variables, $n+m$ of which are binary. Observe that w_{ijk} need not be declared integer as, for fixed values of y and z, there is always an optimal binary assignment.

3.3. Arithmetic reformulations

The following two reformulations are two attempts to reduce the size of the problem. As will be seen, there is no combinatorial property that is used, but only some mathematics are involved in the computation of the allocation cost that allow some form of reduction. This is the reason why we call them arithmetic reformulation.

Both of the previous formulations have $\mathcal{O}(n^2m)$ variables. Now, we get a third formulation with only $\mathcal{O}(n^2)$ variables:

$$
\begin{aligned}
(F_4) \quad \min \quad & \sum_{i=2}^{n}\sum_{j=1}^{i-1} h_{ij} \\
\text{s.t.} \quad & x_{ij} \le y_j, \quad i,j=1,\ldots,n, \\
& \sum_{j=1}^{n} x_{ij} = 1, \quad i=1,\ldots,n, \\
& \sum_{j=1}^{n} y_j = p, \\
& \sum_{k=1}^{m} z_k = q, \\
& h_{ij} + M_{ij}(1 - x_{ij} - x_{ji}) \ge \sum_{k=1}^{m} d_{ijk} z_k, \quad 1 \le j < i \le n, \qquad (6) \\
& h_{ij} \ge 0, \quad i,j=1,\ldots,n,\ j<i,
\end{aligned}
$$

$$x_{ij} \in \{0,1\}, \quad i,j = 1,\dots,n,$$
$$0 \leq y_j \leq 1, \quad j = 1,\dots,n,$$
$$0 \leq z_k \leq 1, \quad k = 1,\dots,m.$$

In this formulation, h_{ij} represents the assignment cost of pair (i,j), either of assigning unit i to median j or of assigning unit j to median i (at most one of these two events will happen), and it takes value zero if none of these two assignments happens. We do not need h_{ij} if $j \geq i$ because $d_{ijk} = d_{jik}$ and $d_{iik} = 0$. In order to define correctly constraints (6), it is enough that $M_{ij} \geq \sum_{k=1}^{q} d_{ij(n-k+1)}$, where $d_{ij(t)}$ is the t-th largest d_{ijk} value (i.e., M_{ij} is at least the sum of the q largest distances associated to (i,j)).

Although it has only $\mathcal{O}(n^2)$ constraints and $\mathcal{O}(n^2)$ variables, this formulation has the disadvantage of having n^2 binary variables. Contrary to the other two formulations, F_4 does not require y_j or z_k to be binary: since variables x_{ij} are binary, then there is an optimal solution where variables y_j take values either zero or one. Concerning variables z_k, just note that the right hand side of constraints (6) will be as small as possible because we are in a minimization problem and, as a consequence, these variables will take value either zero or one at an optimal solution.

Finally, one further observation is that variables h_{ij} can be reduced to just h_i, that is, the measured distance for unit i. The new formulation is:

$$(F_5) \quad \min \quad \sum_{i=1}^{n} h_i$$
$$\text{s.t.} \quad x_{ij} \leq y_j, \quad i,j = 1,\dots,n,$$
$$\sum_{j=1}^{n} x_{ij} = 1, \quad i = 1,\dots,n,$$

$$\sum_{j=1}^{n} y_j = p,$$

$$\sum_{k=1}^{m} z_k = q,$$

$$h_i + M_{ij}(1 - x_{ij}) \geq \sum_{k=1}^{m} d_{ijk} z_k, \quad i, j = 1, \ldots, n,$$

$$h_i \geq 0, \quad i = 1, \ldots, n,$$

$$x_{ij} \in \{0, 1\}, \quad i, j = 1, \ldots, n,$$

$$0 \leq y_j \leq 1, \quad j = 1, \ldots,$$

$$0 \leq z_k \leq 1, \quad k = 1, \ldots, m.$$

3.4. Radius formulation

Last, we propose a totally different model: a radius formulation as introduced in [5] and recently applied very successfully to the p-median problem in [11]. Given a statistical unit u_i and a statistical feature f_k, we have that, since the statistical features are expressed in the Likert scale with g degrees, many units are located at the the same distance, i.e., within a given radius. In order to obtain the radius formulation, we first proceed as follows:

Step 1: Given a customer i, sort distances $\{d_{i1k}, d_{i2k}, \ldots, d_{ink}\}$ in increasing order and remove multiplicities to obtain G_{ik} different values. Let D_{ik1} be the smallest cost, D_{ik2} the second smallest cost and so on:

$$0 = D_{ik1} < D_{ik2} < \ldots < D_{ikG_{ik}}.$$

Step 2: Define binary variables r_{ikt} that take value one if, when feature k is selected, unit i is allocated at distance at least D_{ikt} (and zero otherwise).

It is quite easy to see that

$$\sum_{t=2}^{G_{ik}} (D_{ikt} - D_{ik,t-1}) r_{ikt} = \sum_{j=1}^{n} d_{ijk} w_{ijk},$$

and, as a consequence, the objective function is

$$\sum_{i=1}^{n} \sum_{k=1}^{m} \sum_{t=2}^{G_{ik}} (D_{ikt} - D_{ik,t-1}) r_{ikt}.$$

The resulting model uses also variables y_j, x_{ij} and z_k defined as in the previous models:

$$
\begin{aligned}
(F_6) \quad \min \quad & \sum_{i=1}^{n} \sum_{k=1}^{m} \sum_{t=2}^{G_{ik}} (D_{ikt} - D_{ik,t-1}) r_{ikt} \\
\text{s.t.} \quad & x_{ij} \le y_j, \quad i, j = 1, \dots, n, \\
& \sum_{j=1}^{n} x_{ij} = 1, \quad i = 1, \dots, n \\
& \sum_{j=1}^{n} y_j = p, \\
& \sum_{k=1}^{m} z_k = q, \\
& r_{ikt} + \sum_{\{j \ / \ d_{ijk} < D_{ikt}\}} x_{ij} \ge z_k, \quad i = 1, \dots, n, \\
& \qquad\qquad k = 1, \dots, m, \ t = 2, \dots, G_{ik}, \qquad (7) \\
& r_{ikt} \ge 0, \quad i = 1, \dots, n, \ k = 1, \dots, m, \ t = 2, \dots, G_{ik}, \\
& x_{ij} \ge 0, \quad i, j = 1, \dots, n, \\
& y_j \in \{0, 1\}, \quad j = 1, \dots, n, \\
& z_k \in \{0, 1\}, \quad k = 1, \dots, m.
\end{aligned}
$$

Given a unit i, a feature k and a distance level D_{ikt}, constraint (7) imposes that, if feature k is selected ($z_k = 1$), then either i is allocated to a median j such that this median is at distance $d_{ijk} < D_{ijt}$, or, if this is not possible, it is allocated to at least distance D_{ikt} ($r_{ikt} = 1$). Note also that, since we are minimizing and the distances are positive, there exists an optimal solution such that variables x_{ij} are zero/one. It is also immediate to see that we do not need to require variables r_{ikt} to be binary because they will have values either zero or one at an optimal solution.

Finally, observe that the number of variables of this model is $n+m+n^2+\sum_{i=1}^{n}\sum_{k=1}^{m} G_{ik}$, where $n+m$ variables are binary. Since $G_{ik} \leq g$, where g is the dimension of the scale that we are using for the answers (e.g., $g = 2$ if the answers are binary and $g = 5$ if we are using a 1-5 Likert scale), then the number of variables is upperly bounded by $n + m + n^2 + nmg$. Particularly, in our application, g is the number of tiers of the Likert scale and it is always a small number (i.e., $\mathcal{O}(1)$).

In Table 1 we can see a comparison of the number of constraints and variables of the different formulations (in the case of model F_6, the given value is an upper bound). In Table 2, we compare the order of these sizes. We can see that the radius formulation has always the best order in the number of constraints and variables (while the other models are always worse in either constraints or variables).

4. Computational experience

In this section we show an exhaustive computational study to analyze the performance of the formulations proposed in the previous section.

Formulation	Constraints	Variables	Binary
F_2	$n^2(m+1)+n+2$	$n^2(m+1)+n+m$	$n+m$
F_3	$n^2(m+1)+nm+2$	n^2m+m+n	$n+m$
F_4	$\frac{3n^2-n}{2}+2$	$\frac{3n^2+n}{2}+m$	n^2
F_5	$2n^2+n+2$	n^2+2n+m	n^2
F_6	$n^2+n(mg+1)+2$	$n^2+n(mg+1)+m$	$n+m$

Table 1: Number of variables and constraints.

Formulation	Constraints	Variables	Binary
F_2	$\mathcal{O}(n^2m)$	$\mathcal{O}(n^2m)$	$\mathcal{O}(n+m)$
F_3	$\mathcal{O}(n^2m)$	$\mathcal{O}(n^2m)$	$\mathcal{O}(n+m)$
F_4	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$
F_5	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$
F_6	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n+m)$

Table 2: Order of variables and constraints.

4.1. Instances generation

Models are tested on artificial data, with different values of n and m for the data dimension (they are the number of statistical units and statistical variables, respectively) and different values of p and q for the parameter specification (the number of clusters and statistical variables to select, respectively). Besides, we also simulated two kinds of answers, corresponding to two types of Likert scales: one scale has $g = 2$ and simulates a questionnaire composed of 0/1 binary variables, the second scale has $g = 5$ and represents agreement levels with tiers ranging from *strongly disagree* to *strongly agree*.

The simulated data are built following two steps. In the first step we fixed p^* (correct number of clusters), d^* (number of relevant variables), and the structure of the medians. In the second step we compiled the whole survey based on these hidden features.

In step 1, the ideal archetypes of each cluster, i.e., the cluster medians, are reported in a matrix $F \in \mathbb{R}^{p^*,d^*}$, where f_{ij} is the value of feature j in median i, p^* is the number of medians and d^* is the number of the relevant features. Particularly, when $g = 2$, instances of type L1A contain two medians with four relevant features with the following structure:

$$L1A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

and instances of type L1B contain four medians with six features as follows:

$$L1B = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

When $g = 5$, instances of type L5A contain two hidden medians with four features like this:

$$L5A = \begin{bmatrix} 4 & 4 & 4 & 4 \\ 2 & 2 & 2 & 2 \end{bmatrix},$$

and instances of type L5B have four medians and six features as in the following matrix:

$$L5B = \begin{bmatrix} 4 & 4 & 4 & 4 & 4 & 4 \\ 2 & 2 & 2 & 2 & 2 & 2 \\ 3 & 4 & 4 & 5 & 5 & 5 \\ 3 & 2 & 2 & 1 & 1 & 1 \end{bmatrix}.$$

The second step is to simulate artificial surveys, which take the form of a matrix $V \in \Re^{n,m}$ where v_{ij} is the answer of unit i to question j. We generate the instances in such a way that $m = 2d^*$, that is, half the features determine the cluster and the other half are useless. Then, two types of instances are considered. The first kind of instances, which we call *deterministic instances* and have the suffix "H" in their name, are compiled in such a way that units are the same as one median with respect to the correct d^* features. Every unit i is assigned randomly to a median j, and then $v_{ik} = f_{jk}$ for $k = 1, \ldots, d^*$, while the other values v_{ik} for $k = d^* + 1, \ldots, m$, are uniform random numbers. The second type of instances, that are called *probabilistic instances* and have the suffix "P" in their name, simulates surveys with more variability. In this case each unit i is assigned randomly to one median j, its feature $v_{ik} = f_{jk}$ with probability 0.8, uniform random with probability of 0.2 for $k = 1, \ldots, d^*$. The other values v_{ik} for $k = d^* + 1, \ldots, m$, are uniform random numbers.

4.2. Test results

Instances are generated and solved by combining the following values of the different parameters:

- $n = 30, 50, 80$.

- $m = 8$ (instances A) or $m = 12$ (instances B).
- Parameters p and q have the following values:
  - Instances A: $p = 2, 4, 6$, and $q = 2, 4, 6$.
  - Instances B: $p = 2, 4, 6$, and $q = 3, 6, 8$.
- Type of data: deterministic (H) or probabilistic (P).
- Number of medians: 2 or 4.
- Scale: binary (L1) or Likert (L5).

Therefore, we have $3 \times 2 \times 3 \times 3 \times 2 \times \times 2 = 216$ different instances, each of which we tested on the five different models introduced in the previous section (F_2, F_3, F_4, F_5, and F_6). We carried out the computational study on a Pentium IV computer with two processors (3.2 and 3.2 GHz) and 2 GB RAM. For each formulation and instance, the time limit was set to 3600 seconds of CPU time. The models were solved using the academic license of CPLEX 12.3.

In Tables 3a (deterministic instances) and 3b (probabilistic instances), we summarize the results of our computational study. In each table, for a given scale (L1/L5), type of data (A/B) and nature of data (deterministic/probabilistic) we have compacted the information that corresponds to the 27 instances that we have when we consider the different numbers of statistical units ($n = 30, 50, 80$) and the nine different pairs (p, q). For example, L1AH stands for the information of the 27 instances for scale L1, data A and deterministic case. The information that we provide is the following:

- Average: average time in CPU seconds. We consider solved instances and instances that reach the time limit, but not those instances that run out of memory.

- Median: median of the CPU times as before.

- SR60: number of instances solved to optimality in less than 60 seconds.

- Solved: number of instances solved to optimality before the time limit.

- Not solved: number of instances that reach the time limit without optimality.

- Out of memory: number of instances for which the computer runs out of memory.

What is clear from these tables, no matter which measure we consider, is that the radius formulation is by far the best model. Next, if we compare formulation F_2 (the direct linearization) with formulation F_3 (the p-median style one), we cannot see a clear superiority of one over the other: in some cases F_2 is better and in some other occasions F_3 is (although, in global, F_2 seems to be slightly better than F_3). Finally, if we look at formulations F_4 and F_5, then we see that their behaviour is more erratic: in some cases they reach optimality quite fast (as it can be seen from the medians of instances L1AH and L5AH), but in the other they perform extremely bad. Anyway, both of them are the worst models. From the tables, we can see that instances of type A are more difficult to solve than instances of type B, but this is is something natural because these last instances involve more features (i.e., the problems are larger).

In Table 4 we show the same information than before, but now it has been blocked in the three different sizes ($n = 30, 50, 80$), which means that there are 72 instances per block. In this new classification, the superiority of the radius formulation keeps consistent through the three sizes of instances. The main reason for this success is that model F_6 combines a reduced formulation with a small number of binary variables. Particularly, it is quite important the fact that, given a unit i and a feature k, there are very few different values d_{ijk} and, thus, there are not too many r_{ikt} variables.

Finally, detailed information for every instance is given in Tables 5-12. Here, we show the information of formulations F_2, F_3 and F_6. The other formulations (F_4 and F_5), performing extremely bad, are of no interest. The first three columns are the number of statistical units (n), number of medians (p) and number of relevant features used to measure the distances between units (q). Then, for each of the three shown formulations, we have the CPU time in seconds (T), best lower bound (BLB), best upper bound (BUB) and gap between these two values (Gap). A gap of zero means that the instance has been solved to optimality and this fact is highlighted by showing the best upper bound in boldface. We can see here more clearly than formulation F_6 performs extremely well when compared with formulations F_2 and F_3: the radius formulation is the best for all the considered instances. However, there are some particular hard instances, mainly probabilistic instances of type B with $n = 80$, that not even this best formulation can solve (but, even so, it is still the one that performs the best).

5. Conclusions and further research

Motivated by an application to clustering in sociological surveys, we have introduced in this paper a new extension of the p-median model where the distances between statistical units are calculated using only q out of m possible variables. Five different models are introduced to determine the optimal data partition and are compared: the computational experience shows that the radius formulation outperforms clearly any of the other proposed models.

Note that, although formulation F_6 performs the best, there are several instances that not even this formulation can solve. As a consequence, the next step of our research will be to analyze this formulation and look for some special properties that allow us to develop efficient solution algorithms to solve very large instances (for example, in the line of the methods developed in [11] and [12]). An improvement to the model may come from taking advantage of the equivalence between the radius formulation and the Hammer-Beresnev pseudo-boolean representation of the p-median problem. As noted in [13], the equivalence shows how many variables can be fixed to 0 and 1 before applying the optimization solver, reducing considerably the computational times. Whether those techniques can be extended to the case of variable selection will be an issue of future research.

Acknowledgments

Sergio García would like to acknowledge that research reported here was supported by *Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I+D+I)*, project MTM2009-14039-C06-04, RDEF funds and Fundación Séneca project 08716/PI/08.

References

[1] P. Avella, A. Sassano, and I. Vasil'ev. Computational study of large-scale p-median problems. *Mathematical Programming* 109 (1):89–114, 2007.

[2] S. Benati. Categorical data fuzzy clustering: An analysis of local search heuristics. *Computers & Operations Research* 35 (3):766–775, 2008.

[3] S. Benati and S. Stefani. The academic journal ranking problem: A fuzzy-clustering approach. *Journal of Classification*, 28 (1):7–20, 2011.

[4] A. Chevan and M. Sutherland. Hierarchical partitioning. *The American Statistician* 45 (2):90–96, 1991.

[5] G. Cornuéjols, G. Nemhauser, and L. Wolsey. A canonical representation of simple plant location-problems and its applications. *SIAM Journal On Algebraic And Discrete Methods* 1(3):261–272, 1980.

[6] G. Cornuéjols, G.L. Nemhauser, and L.A. Wolsey. *Discrete Location Theory*, chap. The uncapacitated facility location problem. Wiley-Interscience, 119–171, 1990.

[7] S. Elloumi. A tighter formulation of the p-median problem. *Journal Of Combinatorial Optimization* 19(1):69–83, 2010.

[8] S. Elloumi, M. Labbé, and Y. Pochet. A new formulation and resolution method for the p-center problem. *INFORMS Journal On Computing* 16 (1):84–94, 2004.

[9] B.S. Everitt, S. Landau, M. Leese, D. Stahl. *Cluster Analysis*, 5th ed. Wiley-Interscience, 2011.

[10] G. Gan, C. Ma, J. Wu. *Data Clustering: Theory, Algorithms and Applications* ASA-SIAM, 2007.

[11] S. García, M. Labbé, and A. Marín. Solving large p-median problems with a radius formulation. *INFORMS Journal on Computing* 23 (4):546–556, 2011.

[12] S. García, M. Landete, and A. Marín. New formulation and a branch-and-cut algorithm for the multiple allocation p-hub median problem. *European Journal Of Operational Research* 220 (1):48–57, 2012.

[13] B. Goldengorin. Data correcting approach for routing and location in networks. Forthcoming in *Handbook of Combinatorial Optimization*, D.Z. Du and P.M. Pardalos eds. Kluwer Academic Publishers, 2012.

[14] S.L. Hakimi. Optimum locations of switching centers and absolute centers and medians of a graph. *Operations Research* 1:450–459, 1964.

[15] P. Hansen and B. Jaumard. Cluster analysis and mathematical programming. *Mathematical Programming* 79:191–215, 1997.

[16] T.D. Klastorin. The p-median problem for cluster-analysis -a comparative test using the mixture model approach. *Management Science* 31 (1):84–95, 1985.

[17] H.F. Köhn, D. Steinley, and M.J. Brusco. The p-median model as a tool

for clustering psychological data. *Psychological Methods* 15 (1):87–95, 2010.

[18] J. MacQueen. Some methods for classification and analysis of multivariate observation. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press 1:281–297, 1967.

[19] J.M. Mulvey and H.P. Crowder. Cluster analysis: An application of lagrangian relaxation. *Management Science* 25 (4):329–340, 1979.

[20] M. Rao. Cluster analysis and mathematical programming. *Journal of the American Statistical Association* 66:622–626, 1971.

[21] C. ReVelle and R. Swain. Central facilities location. *Geographic Analysis* 2:30–42, 1970.

[22] H. Simao and J. Thizy. A dual simplex algorithm for the canonical representation of the uncapacitated facility location problem. *Operations Research Letters* 8 (5):279–286, 1989.

[23] E. Taillard. Heuristic methods for large centroid clustering problems. *Journal Of Heuristics* 9 (1):51–73, 2003.

[24] World Values Survey 1981-2008 Official Aggregate v.20090901. World Values Survey Association (www.worldvaluessurvey.org). Aggregate File Producer: ASEP/JDS, Madrid. 2009.

		F_2	F_3	F_4	F_5	F_6
L1AH	Average	501	513	929	1025	27
	Median	34	26	4	5	4
	SR60	18	17	17	16	24
	Solved	24	24	19	19	27
	Not solved	3	3	6	7	0
	Out of memory	0	0	2	1	0
L1BH	Average	1647	2095	2199	2056	488
	Median	1386	2720	3600	3600	18
	SR60	6	5	5	8	17
	Solved	18	15	9	12	25
	Not solved	9	12	12	13	2
	Out of memory	0	0	6	2	0
L5AH	Average	573	374	1031	1039	20
	Median	50	30	2	3	3
	SR60	15	20	17	16	24
	Solved	24	27	18	18	27
	Not solved	3	0	7	7	0
	Out of memory	0	0	2	2	0
L5BH	Average	1517	1316	2005	2047	161
	Median	526	489	3600	3329	20
	SR60	7	7	8	6	18
	Solved	20	21	11	12	27
	Not solved	7	6	12	11	0
	Out of memory	0	0	4	2	0

(a) Deterministic instances.

		F_2	F_3	F_4	F_5	F_6
L1AP	Average	1280	1893	2541	2854	192
	Median	218	1914	3600	3600	32
	SR60	6	4	4	4	16
	Solved	21	16	6	6	27
	Not solved	6	11	14	18	0
	Out of memory	0	0	7	3	0
L1BP	Average	2454	3443	3360	3322	1400
	Median	3600	3600	3600	3600	473
	SR60	1	0	0	0	5
	Solved	12	2	1	2	19
	Not solved	15	25	13	16	8
	Out of memory	0	0	13	9	0
L5AP	Average	1581	1657	3202	3271	309
	Median	1082	925	3600	3600	62
	SR60	3	1	0	0	13
	Solved	21	18	3	3	27
	Not solved	6	9	14	19	0
	Out of memory	0	0	10	5	0
L5BP	Average	2503	2938	3600	3600	1102
	Median	3600	3600	3600	3600	330
	SR60	0	0	0	0	5
	Solved	11	8	0	0	23
	Not solved	16	19	20	21	4
	Out of memory	0	0	7	6	0

(b) Probabilistic instances.

Table 3: Summary statistics.

		F_2	F_3	F_4	F_5	F_6
n=30	Average	275	899	2100	2236	20
	Median	96	158	3600	3600	9
	SR60	29	25	21	21	64
	Solved	72	62	28	28	72
	Not solved	0	10	34	40	0
	Out of memory	0	0	10	4	0
n=50	Average	1595	1868	1897	2016	171
	Median	1193	1600	3600	3600	60
	SR60	14	17	19	20	36
	Solved	54	43	20	23	72
	Not solved	18	29	22	28	0
	Out of memory	0	0	30	21	0
n=80	Average	2650	2569	2553	2647	1196
	Median	3600	3600	3600	3600	434
	SR60	13	12	11	9	22
	Solved	25	26	19	21	58
	Not solved	47	46	42	44	14
	Out of memory	0	0	11	7	0

Table 4: Summary statistics by size.

			F_2				F_3				F_6			
n	p	q	T	BLB	BUB	Gap	T	BLB	BUB	Gap	T	BLB	BUB	Gap
30	2	2	1	0.00	**0.00**	0.00	3	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	2	4	25	0.00	**0.00**	0.00	2	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	2	6	125	22.00	**22.00**	0.00	87	22.00	**22.00**	0.00	5	22.00	**22.00**	0.00
30	4	2	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	4	4	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	4	6	99	10.00	**10.00**	0.00	301	10.00	**10.00**	0.00	4	10.00	**10.00**	0.00
30	6	2	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	6	4	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	6	6	28	2.00	**2.00**	0.00	73	2.00	**2.00**	0.00	9	2.00	**2.00**	0.00
50	2	2	49	0.00	**0.00**	0.00	9	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
50	2	4	542	0.00	**0.00**	0.00	76	0.00	**0.00**	0.00	3	0.00	**0.00**	0.00
50	2	6	885	34.00	**34.00**	0.00	541	34.00	**34.00**	0.00	29	34.00	**34.00**	0.00
50	4	2	5	0.00	**0.00**	0.00	4	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
50	4	4	5	0.00	**0.00**	0.00	4	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
50	4	6	398	13.00	**13.00**	0.00	1166	13.00	**13.00**	0.00	37	13.00	**13.00**	0.00
50	6	2	5	0.00	**0.00**	0.00	4	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
50	6	4	5	0.00	**0.00**	0.00	4	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
50	6	6	341	5.00	**5.00**	0.00	618	5.00	**5.00**	0.00	29	5.00	**5.00**	0.00
80	2	2	39	0.00	**0.00**	0.00	26	0.00	**0.00**	0.00	15	0.00	**0.00**	0.00
80	2	4	43	0.00	**0.00**	0.00	26	0.00	**0.00**	0.00	18	0.00	**0.00**	0.00
80	2	6	3600	0.00	68.00	100.00	3600	25.20	72.00	65.00	130	68.00	**68.00**	0.00
80	4	2	32	0.00	**0.00**	0.00	25	0.00	**0.00**	0.00	5	0.00	**0.00**	0.00
80	4	4	36	0.00	**0.00**	0.00	26	0.00	**0.00**	0.00	7	0.00	**0.00**	0.00
80	4	6	3600	0.00	30.00	100.00	3600	5.00	30.00	83.33	185	27.00	**27.00**	0.00
80	6	2	30	0.00	**0.00**	0.00	25	0.00	**0.00**	0.00	3	0.00	**0.00**	0.00
80	6	4	34	0.00	**0.00**	0.00	26	0.00	**0.00**	0.00	6	0.00	**0.00**	0.00
80	6	6	3600	0.00	7.00	100.00	3600	0.00	19.00	100.00	222	7.00	**7.00**	0.00

Table 5: Instances L1AH ($m = 8$).

			F_2				F_3				F_6			
n	p	q	T	BLB	BUB	Gap	T	BLB	BUB	Gap	T	BLB	BUB	Gap
30	2	2	26	6.00	**6.00**	0.00	410	6.00	**6.00**	0.00	14	6.00	**6.00**	0.00
30	2	4	189	24.00	**24.00**	0.00	2167	24.00	**24.00**	0.00	18	24.00	**24.00**	0.00
30	2	6	201	41.00	**41.00**	0.00	244	41.00	**41.00**	0.00	7	41.00	**41.00**	0.00
30	4	2	2	0.00	**0.00**	0.00	3	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	4	4	81	12.00	**12.00**	0.00	513	12.00	**12.00**	0.00	28	12.00	**12.00**	0.00
30	4	6	155	28.00	**28.00**	0.00	443	28.00	**28.00**	0.00	7	28.00	**28.00**	0.00
30	6	2	3	0.00	**0.00**	0.00	2	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	6	4	63	6.00	**6.00**	0.00	344	6.00	**6.00**	0.00	27	6.00	**6.00**	0.00
30	6	6	76	20.00	**20.00**	0.00	199	20.00	**20.00**	0.00	5	20.00	**20.00**	0.00
50	2	2	218	12.00	**12.00**	0.00	3600	0.00	12.00	100.00	154	12.00	**12.00**	0.00
50	2	4	1845	37.00	**37.00**	0.00	3600	1.75	37.00	95.28	127	37.00	**37.00**	0.00
50	2	6	2202	73.00	**73.00**	0.00	1914	73.00	**73.00**	0.00	41	73.00	**73.00**	0.00
50	4	2	7	0.00	**0.00**	0.00	32	0.00	**0.00**	0.00	2	0.00	**0.00**	0.00
50	4	4	782	22.00	**22.00**	0.00	3600	4.84	24.00	0.00	154	22.00	**22.00**	0.00
50	4	6	1901	51.00	**51.00**	0.00	1782	51.00	**51.00**	0.00	41	51.00	**51.00**	0.00
50	6	2	9	0.00	**0.00**	0.00	11	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
50	6	4	732	13.00	**13.00**	0.00	3281	13.00	**13.00**	0.00	232	13.00	**13.00**	0.00
50	6	6	1145	39.00	**39.00**	0.00	3600	20.72	41.00	49.47	32	39.00	**39.00**	0.00
80	2	2	3208	22.00	**22.00**	0.00	3600	0.00	30.00	10.00	594	22.00	**22.00**	0.00
80	2	4	3600	0.00	118.00	100.00	3600	0.00	165.00	10.00	1398	61.00	**61.00**	0.00
80	2	6	3600	0.00	128.00	100.00	3600	70.46	116.00	39.26	227	113.00	**113.00**	0.00
80	4	2	84	0.00	**0.00**	0.00	94	0.00	**0.00**	0.00	13	0.00	**0.00**	0.00
80	4	4	3600	0.00	66.00	100.00	3600	0.00	213.00	10.00	1066	36.00	**36.00**	0.00
80	4	6	3600	0.00	79.00	100.00	3600	38.68	109.00	64.52	258	79.00	**79.00**	0.00
80	6	2	32	0.00	**0.00**	0.00	66	0.00	**0.00**	0.00	12	0.00	**0.00**	0.00
80	6	4	3600	0.00	23.00	100.00	3600	0.00	226.00	100.00	491	23.00	**23.00**	0.00
80	6	6	3600	0.00	66.00	100.00	3600	26.31	65.00	59.53	224	62.00	**62.00**	0.00

Table 6: Instances L1AP ($m = 8$).

			F_2				F_3				F_6			
n	p	q	T	BLB	BUB	Gap	T	BLB	BUB	Gap	T	BLB	BUB	Gap
30	2	3	93	6.00	**6.00**	0.00	2234	6.00	**6.00**	0.00	22	6.00	**6.00**	0.00
30	2	6	620	28.00	**28.00**	0.00	3600	7.77	28.00	72.26	49	28.00	**28.00**	0.00
30	2	8	1468	47.00	**47.00**	0.00	3070	47.00	**47.00**	0.00	35	47.00	**47.00**	0.00
30	4	3	1	0.00	**0.00**	0.00	14	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	4	6	6	0.00	**0.00**	0.00	40	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	4	8	291	18.00	**18.00**	0.00	3600	13.10	18.00	27.20	17	18.00	**18.00**	0.00
30	6	3	4	0.00	**0.00**	0.00	10	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	6	6	19	0.00	**0.00**	0.00	44	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	6	8	283	10.00	**10.00**	0.00	1512	10.00	**10.00**	0.00	18	10.00	**10.00**	0.00
50	2	3	1386	7.00	**7.00**	0.00	3600	0.00	15.00	100.00	258	7.00	**7.00**	0.00
50	2	6	3600	12.00	38.00	68.42	3600	0.00	198.00	100.00	237	38.00	**38.00**	0.00
50	2	8	3600	19.00	75.00	74.67	3600	31.91	245.00	86.97	142	75.00	**75.00**	0.00
50	4	3	113	0.00	**0.00**	0.00	60	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
50	4	6	2413	0.00	**0.00**	0.00	474	0.00	**0.00**	0.00	3	0.00	**0.00**	0.00
50	4	8	3600	22.33	39.00	42.74	3600	2.78	50.00	94.44	199	39.00	**39.00**	0.00
50	6	3	9	0.00	**0.00**	0.00	65	0.00	**0.00**	0.00	2	0.00	**0.00**	0.00
50	6	6	9	0.00	**0.00**	0.00	289	0.00	**0.00**	0.00	2	0.00	**0.00**	0.00
50	6	8	3600	12.00	23.00	47.83	3600	1.38	30.00	95.42	142	23.00	**23.00**	0.00
80	2	3	3600	0.00	62.00	100.00	3600	0.00	68.00	100.00	2023	16.00	**16.00**	0.00
80	2	6	3600	0.00	136.00	100.00	3600	0.00	136.00	100.00	3600	40.41	72.00	43.87
80	2	8	3600	0.00	193.00	100.00	3600	27.21	213.00	87.23	1382	132.00	**132.00**	0.00
80	4	3	1804	0.00	**0.00**	0.00	425	0.00	**0.00**	0.00	17	0.00	**0.00**	0.00
80	4	6	2678	0.00	**0.00**	0.00	2720	0.00	**0.00**	0.00	17	0.00	**0.00**	0.00
80	4	8	3600	0.00	165.00	100.00	3600	0.00	213.00	100.00	3600	45.62	63.00	27.59
80	6	3	128	0.00	**0.00**	0.00	202	0.00	**0.00**	0.00	7	0.00	**0.00**	0.00
80	6	6	735	0.00	**0.00**	0.00	2207	0.00	**0.00**	0.00	14	0.00	**0.00**	0.00
80	6	8	3600	0.00	80.00	100.00	3600	0.00	356.00	100.00	1398	37.00	**37.00**	0.00

Table 7: Instances L1BH ($m = 12$).

			F_2				F_3				F_6			
n	p	q	T	BLB	BUB	Gap	T	BLB	BUB	Gap	T	BLB	BUB	Gap
30	2	3	189	10.00	**10.00**	0.00	3600	0.00	10.00	100.00	35	10.00	**10.00**	0.00
30	2	6	1083	34.00	**34.00**	0.00	3600	6.57	34.00	80.67	95	34.00	**34.00**	0.00
30	2	8	2059	55.00	**55.00**	0.00	3600	53.18	55.00	3.31	50	55.00	**55.00**	0.00
30	4	3	170	3.00	**3.00**	0.00	2229	3.00	**3.00**	0.00	75	3.00	**3.00**	0.00
30	4	6	714	22.00	**22.00**	0.00	3600	10.83	23.00	52.91	73	22.00	**22.00**	0.00
30	4	8	1227	38.00	**38.00**	0.00	3600	24.34	38.00	35.94	42	38.00	**38.00**	0.00
30	6	3	54	0.00	**0.00**	0.00	740	0.00	**0.00**	0.00	5	0.00	**0.00**	0.00
30	6	6	567	15.00	**15.00**	0.00	3600	9.38	15.00	37.50	39	15.00	**15.00**	0.00
30	6	8	476	27.00	**27.00**	0.00	3600	14.23	27.00	47.30	61	27.00	**27.00**	0.00
50	2	3	2877	23.00	**23.00**	0.00	3600	0.00	43.00	100.00	319	23.00	**23.00**	0.00
50	2	6	3600	13.00	76.00	82.89	3600	1.62	82.00	98.03	1200	72.00	**72.00**	0.00
50	2	8	3600	17.50	108.00	83.80	3600	48.80	195.00	74.97	411	108.00	**108.00**	0.00
50	4	3	1386	8.00	**8.00**	0.00	3600	0.00	17.00	100.00	618	8.00	**8.00**	0.00
50	4	6	3600	0.00	47.00	0.00	3600	0.00	70.00	100.00	571	46.00	**46.00**	0.00
50	4	8	3600	0.00	75.00	100.00	3600	22.82	185.00	87.67	344	75.00	**75.00**	0.00
50	6	3	1444	1.00	**1.00**	0.00	3600	0.00	3.00	100.00	472	1.00	**1.00**	0.00
50	6	6	3600	4.63	30.00	84.58	3600	0.00	60.00	0.00	473	30.00	**30.00**	0.00
50	6	8	3600	6.00	57.00	89.47	3600	15.43	83.00	81.37	1390	56.00	**56.00**	0.00
80	2	3	3600	0.00	43.00	100.00	3600	0.00	104.00	100.00	2725	43.00	**43.00**	0.00
80	2	6	3600	0.00	130.00	100.00	3600	0.00	238.00	100.00	3600	26.49	134.00	80.23
80	2	8	3600	0.00	222.00	100.00	3600	52.31	322.00	87.23	3600	129.20	191.00	32.36
80	4	3	3600	0.00	36.00	100.00	3600	0.00	148.00	100.00	3600	0.00	20.00	100.00
80	4	6	3600	0.00	109.00	100.00	3600	0.00	316.00	100.00	3600	8.12	89.00	90.89
80	4	8	3600	0.00	180.00	100.00	3600	13.41	323.00	95.85	3600	46.74	140.00	66.62
80	6	3	3600	0.00	7.00	100.00	3600	0.00	148.00	0.00	3600	0.00	6.00	100.00
80	6	6	3600	0.00	100.00	100.00	3600	0.00	313.00	0.00	3600	34.00	68.00	50.00
80	6	8	3600	0.00	15.00	100.00	3600	2.72	312.00	99.13	3600	8.24	118.00	93.01

Table 8: Instances L1BP ($m = 12$).

			F_2				F_3				F_6			
n	p	q	T	BLB	BUB	Gap	T	BLB	BUB	Gap	T	BLB	BUB	Gap
30	2	2	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	2	4	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	2	6	124	59.00	**59.00**	0.00	59	59.00	**59.00**	0.00	4	59.00	**59.00**	0.00
30	4	2	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	4	4	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	4	6	198	38.00	**38.00**	0.00	70	38.00	**38.0**0	0.00	4	38.00	**38.00**	0.00
30	6	2	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	6	4	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	6	6	61	25.00	**25.00**	0.00	56	25.00	**25.00**	0.00	3	25.00	**25.00**	0.00
50	2	2	15	0.00	**0.00**	0.00	6	0.00	**0.00**	0.00	2	0.00	**0.00**	0.00
50	2	4	167	0.00	**0.00**	0.00	11	0.00	**0.00**	0.00	3	0.00	**0.00**	0.00
50	2	6	1207	109.00	**109.00**	0.00	402	109.00	**109.00**	0.00	18	109.00	**109.00**	0.00
50	4	2	8	0.00	**0.00**	0.00	5	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
50	4	4	224	0.00	**0.00**	0.00	15	0.00	**0.00**	0.00	2	0.00	**0.00**	0.00
50	4	6	1620	67.00	**67.00**	0.00	449	67.00	**67.00**	0.00	22	67.00	**67.00**	0.00
50	6	2	7	0.00	**0.00**	0.00	6	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
50	6	4	216	0.00	**0.00**	0.00	13	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
50	6	6	547	47.00	**47.00**	0.00	434	47.00	**47.00**	0.00	21	47.00	**47.00**	0.00
80	2	2	50	0.00	**0.00**	0.00	30	0.00	**0.00**	0.00	18	0.00	**0.00**	0.00
80	2	4	56	0.00	**0.00**	0.00	31	0.00	**0.00**	0.00	29	0.00	**0.00**	0.00
80	2	6	3600	101.00	193.00	47.67	3048	193.00	193.00	**0.00**	71	193.00	**193.00**	0.00
80	4	2	43	0.00	**0.00**	0.00	30	0.00	**0.00**	0.00	8	0.00	**0.00**	0.00
80	4	4	45	0.00	**0.00**	0.00	30	0.00	**0.00**	0.00	17	0.00	**0.00**	0.00
80	4	6	3600	9.00	121.00	92.56	2501	117.00	**117.00**	0.00	162	117.00	**117.00**	0.00
80	6	2	35	0.00	**0.00**	0.00	30	0.00	**0.00**	0.00	3	0.00	**0.00**	0.00
80	6	4	39	0.00	**0.00**	0.00	31	0.00	**0.00**	0.00	11	0.00	**0.00**	0.00
80	6	6	3600	54.50	82.00	33.54	2841	82.00	**82.00**	0.00	138	82.00	**82.00**	0.00

Table 9: Instances L5AH ($m = 8$).

			F_2				F_3				F_6			
n	p	q	T	BLB	BUB	Gap	T	BLB	BUB	Gap	T	BLB	BUB	Gap
30	2	2	31	8.00	**8.00**	0.00	70	8.00	**8.00**	0.00	12	8.00	**8.00**	0.00
30	2	4	103	37.00	**37.00**	0.00	98	37.00	**37.00**	0.00	8	37.00	**37.00**	0.00
30	2	6	212	93.00	**93.00**	0.00	72	93.00	**93.00**	0.00	6	93.00	**93.00**	0.00
30	4	2	17	2.00	**2.00**	0.00	96	2.00	**2.00**	0.00	6	2.00	**2.00**	0.00
30	4	4	79	21.00	**21.00**	0.00	85	21.00	**21.00**	0.00	14	21.00	**21.00**	0.00
30	4	6	257	69.00	**69.00**	0.00	89	69.00	**69.00**	0.00	9	69.00	**69.00**	0.00
30	6	2	7	0.00	**0.00**	0.00	3	0.00	**0.00**	0.00	3	0.00	**0.00**	0.00
30	6	4	80	13.00	**13.00**	0.00	117	13.00	**13.00**	0.00	19	13.00	**13.00**	0.00
30	6	6	120	52.00	**52.00**	0.00	76	52.00	**52.00**	0.00	7	52.00	**52.00**	0.00
50	2	2	432	24.00	**24.00**	0.00	2443	24.00	**24.00**	0.00	124	24.00	**24.00**	0.00
50	2	4	1178	62.00	**62.00**	0.00	863	62.00	**62.00**	0.00	50	62.00	**62.00**	0.00
50	2	6	3441	168.00	**168.00**	0.00	560	168.00	**168.00**	0.00	42	168.00	**168.00**	0.00
50	4	2	1082	11.00	**11.00**	0.00	803	11.00	**11.00**	0.00	64	11.00	**11.00**	0.00
50	4	4	1431	47.00	**47.00**	0.00	1044	47.00	**47.00**	0.00	159	47.00	**47.00**	0.00
50	4	6	3587	135.00	**135.00**	0.00	625	135.00	**135.00**	0.00	62	135.00	**135.00**	0.00
50	6	2	167	7.00	**7.00**	0.00	997	7.00	**7.00**	0.00	44	7.00	**7.00**	0.00
50	6	4	652	36.00	**36.00**	0.00	925	36.00	**36.00**	0.00	110	36.00	**36.00**	0.00
50	6	6	815	110.00	**110.00**	0.00	3600	96.00	110.00	12.73	38	110.00	**110.00**	0.00
80	2	2	2824	47.00	**47.00**	0.00	3600	4.34	47.00	90.78	2574	47.00	**47.00**	0.00
80	2	4	3600	0.00	157.00	100.00	3600	96.69	261.00	62.95	460	119.00	**119.00**	0.00
80	2	6	3600	45.00	286.00	84.27	3375	286.00	**286.00**	0.00	234	286.00	**286.00**	0.00
80	4	2	3600	0.00	71.00	100.00	3600	0.00	34.00	100.00	1377	22.00	**22.00**	0.00
80	4	4	3600	0.00	234.00	100.00	3600	53.26	163.00	67.32	1062	85.00	**85.00**	0.00
80	4	6	3600	6.33	235.00	97.30	3600	209.14	235.00	11.00	362	235.00	**235.00**	0.00
80	6	2	1308	14.00	**14.00**	0.00	3600	0.00	14.00	100.00	443	14.00	**14.00**	0.00
80	6	4	3264	71.00	**71.00**	0.00	3600	41.89	147.00	71.50	526	71.00	**71.00**	0.00
80	6	6	3600	31.00	211.00	85.31	3600	176.26	202.00	12.74	523	202.00	**202.00**	0.00

Table 10: Instances L5AP ($m = 8$).

			F_2				F_3				F_6			
n	p	q	T	BLB	BUB	Gap	T	BLB	BUB	Gap	T	BLB	BUB	Gap
30	2	3	57	13.00	**13.00**	0.00	570	13.00	**13.00**	0.00	10	13.00	**13.00**	0.00
30	2	6	225	52.00	**52.00**	0.00	322	52.00	**52.00**	0.00	16	52.00	**52.00**	0.00
30	2	8	526	113.00	**113.00**	0.00	489	113.00	**113.00**	0.00	15	113.00	**113.00**	0.00
30	4	3	6	0.00	**0.00**	0.00	9	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	4	6	28	0.00	**0.00**	0.00	3	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	4	8	343	60.00	**60.00**	0.00	328	60.00	**60.00**	0.00	10	60.00	**60.00**	0.00
30	6	3	1	0.00	**0.00**	0.00	3	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	6	6	15	0.00	**0.00**	0.00	2	0.00	**0.00**	0.00	1	0.00	**0.00**	0.00
30	6	8	261	40.00	**40.00**	0.00	506	40.00	**40.00**	0.00	16	40.00	**40.00**	0.00
50	2	3	639	20.00	**20.00**	0.00	3600	2.84	20.00	85.78	76	20.00	**20.00**	0.00
50	2	6	2662	80.00	**80.00**	0.00	3034	80.00	**80.00**	0.00	75	80.00	**80.00**	0.00
50	2	8	3600	42.50	192.00	77.86	2915	186.00	**186.00**	0.00	68	186.00	**186.00**	0.00
50	4	3	368	0.00	**0.00**	0.00	107	0.00	**0.00**	0.00	9	0.00	**0.00**	0.00
50	4	6	266	0.00	**0.00**	0.00	25	0.00	**0.00**	0.00	6	0.00	**0.00**	0.00
50	4	8	2806	104.00	**104.00**	0.00	2592	104.00	**104.00**	0.00	58	104.00	**104.00**	0.00
50	6	3	12	0.00	**0.00**	0.00	21	0.00	**0.00**	0.00	4	0.00	**0.00**	0.00
50	6	6	12	0.00	**0.00**	0.00	23	0.00	**0.00**	0.00	3	0.00	**0.00**	0.00
50	6	8	2455	77.00	**77.00**	0.00	1417	77.00	**77.00**	0.00	100	77.00	**77.00**	0.00
80	2	3	3600	0.00	161.00	100.00	3600	2.14	97.00	97.79	524	0.00	34.00	100.00
80	2	6	3600	0.00	542.00	100.00	3600	102.38	542.00	81.11	424	136.00	**136.00**	0.00
80	2	8	3600	0.00	692.00	100.00	3600	212.31	752.00	71.77	409	307.00	**307.00**	0.00
80	4	3	393	0.00	**0.00**	0.00	1079	0.00	**0.00**	0.00	28	0.00	**0.00**	0.00
80	4	6	3600	0.00	68.00	100.00	155	0.00	**0.00**	0.00	37	0.00	**0.00**	0.00
80	4	8	3600	0.00	278.00	100.00	3600	88.82	1240.00	92.84	276	163.00	**163.00**	0.00
80	6	3	2723	0.00	**0.00**	0.00	181	0.00	**0.00**	0.00	20	0.00	**0.00**	0.00
80	6	6	1969	0.00	**0.00**	0.00	159	0.00	**0.00**	0.00	24	0.00	**0.00**	0.00
80	6	8	3600	0.00	226.00	100.00	3600	45.87	146.00	68.58	2135	128.00	**128.00**	0.00

Table 11: Instances L5BH ($m = 12$).

			F_2				F_3				F_6			
n	p	q	T	BLB	BUB	Gap	T	BLB	BUB	Gap	T	BLB	BUB	Gap
30	2	3	240	36.00	**36.00**	0.00	888	36.00	**36.00**	0.00	66	36.00	**36.00**	0.00
30	2	6	1631	104.00	**104.00**	0.00	1614	104.00	**104.00**	0.00	64	104.00	**104.00**	0.00
30	2	8	1526	166.00	**166.00**	0.00	897	166.00	**166.00**	0.00	31	166.00	**166.00**	0.00
30	4	3	178	22.00	**22.00**	0.00	986	22.00	**22.00**	0.00	59	22.00	**22.00**	0.00
30	4	6	1220	74.00	**74.00**	0.00	1942	74.00	**74.00**	0.00	114	74.00	**74.00**	0.00
30	4	8	785	129.00	**129.00**	0.00	1659	129.00	**129.00**	0.00	53	129.00	**129.00**	0.00
30	6	3	119	13.00	**13.00**	0.00	1007	13.00	**13.00**	0.00	22	13.00	**13.00**	0.00
30	6	6	467	58.00	**58.00**	0.00	1932	58.00	**58.00**	0.00	76	58.00	**58.00**	0.00
30	6	8	507	109.00	**109.00**	0.00	3600	96.28	109.00	11.67	53	109.00	**109.00**	0.00
50	2	3	3600	17.00	62.00	72.58	3600	12.00	81.00	85.19	717	59.00	**59.00**	0.00
50	2	6	3600	29.75	206.00	85.56	3600	108.52	192.00	43.48	337	166.00	**166.00**	0.00
50	2	8	3600	27.40	296.00	90.74	3600	227.95	308.00	25.99	234	285.00	**285.00**	0.00
50	4	3	1838	33.00	**33.00**	0.00	3600	0.47	60.00	99.22	660	33.00	**33.00**	0.00
50	4	6	3600	4.75	118.00	95.97	3600	76.68	179.00	57.16	375	103.00	**103.00**	0.00
50	4	8	3600	7.20	220.00	96.73	3600	145.74	228.00	36.08	330	219.00	**219.00**	0.00
50	6	3	1466	21.00	**21.00**	0.00	3600	0.00	28.00	100.00	300	21.00	**21.00**	0.00
50	6	6	3600	27.00	99.00	72.73	3600	54.37	119.00	54.31	251	84.00	**84.00**	0.00
50	6	8	3600	62.73	181.00	65.34	3600	129.25	194.00	33.37	261	179.00	**179.00**	0.00
80	2	3	3600	0.00	163.00	100.00	3600	9.89	231.00	95.72	3600	1.16	107.00	98.91
80	2	6	3600	0.00	488.00	100.00	3600	147.55	786.00	81.23	2548	284.00	**284.00**	0.00
80	2	8	3600	0.00	713.00	100.00	3600	342.62	978.00	64.97	1146	465.00	**465.00**	0.00
80	4	3	3600	0.00	157.00	100.00	3600	0.00	331.00	100.00	3600	17.71	68.00	73.96
80	4	6	3600	0.00	340.00	100.00	3600	103.31	844.00	87.76	2163	190.00	**190.00**	0.00
80	4	8	3600	0.00	435.00	100.00	3600	268.37	1070.00	74.92	3194	369.00	**369.00**	0.00
80	6	3	3600	0.00	61.00	100.00	3600	0.00	331.00	100.00	3600	12.34	47.00	73.74
80	6	6	3600	0.00	270.00	100.00	3600	77.46	844.00	90.82	3600	130.52	167.00	21.84
80	6	8	3600	0.00	403.00	100.00	3600	209.44	1070.00	80.43	2302	330.00	**330.00**	0.00

Table 12: Instances L5BP ($m = 12$).